

Development and Evaluation of Machine Learning Models for Predicting Infant Mortality in Malawi Using 2024 Malawi Demographic and Health Survey Data

Asaye Mwisanyo Ngosi

asayengosi@gmail.com

University of Portsmouth

Research Article

Keywords: infant mortality, Malawi, machine learning, Random forest, Logistic regression, XGBoost, DHS, class imbalance, predictive modelling, SHAP

Posted Date: June 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9857787/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Development and Evaluation of Machine Learning Models for Predicting Infant Mortality in Malawi Using 2024 Malawi Demographic and Health Survey Data.

Author

Asaye Mwisanyo Ngosi

Affiliation

School of Computing, Faculty of Technology, University of
Portsmouth, Portsmouth, United Kingdom

Correspondence

Email: asayengosi@gmail.com / asaye.ngosi@myport.ac.uk

Abstract

Infant mortality remains an important public health concern in Malawi despite substantial progress in child survival over recent decades. The 2024 Malawi Demographic and Health Survey reported an infant mortality rate of 35 deaths per 1,000 live births, with neonatal deaths accounting for a large proportion of deaths during infancy. This study developed and evaluated machine learning models for predicting infant mortality in Malawi using the 2024 Malawi Demographic and Health Survey data.

A quantitative secondary data analysis design was used. The analytical sample included 50,261 child records with a valid infant survival outcome. Infant death was defined as death before the first birthday and coded as a binary outcome. Descriptive analysis, missing data assessment, chi-square tests, Cramer's V and multicollinearity assessment were conducted before model development. Logistic Regression was used as the baseline model, while Random Forest and XGBoost were developed as machine learning alternatives. Models were evaluated using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, PR-AUC and Brier score. Class weighting and SMOTE sensitivity analysis were applied to address class imbalance.

The analysis showed that infant death was a rare outcome, affecting 2,187 children, representing 4.35% of the unweighted sample and 4.72% of the weighted sample. Short gestation, birth interval, parity, maternal age at birth and multiple birth status were the most influential predictors. The improved main class-weighted XGBoost model was selected as the primary predictive model because it provided the best balance between predictive performance, data completeness and practical usefulness among the low-missingness models. The improved extended class-weighted XGBoost model achieved the highest overall PR-AUC, but it was treated as a sensitivity model because it included predictors with very high missingness.

The study concludes that machine learning can add value to infant mortality prediction in Malawi, although the improvement over Logistic Regression was modest. The findings highlight the importance of balancing predictive performance with interpretability, data quality and public health usefulness. The model should not be used as a stand-alone decision-making tool without external validation and further calibration, but it provides useful evidence on infant mortality risk patterns in Malawi.

Keywords: infant mortality; Malawi; machine learning; Random forest; Logistic regression; XGBoost; DHS; class imbalance; predictive modelling; SHAP.

List of Abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
ANC	Antenatal Care
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
DHS	Demographic and Health Survey
EDHS	Ethiopian Demographic and Health Survey
F1-score	Harmonic mean of precision and recall
ICF	Inner City Fund International
IM	Infant Mortality
LMICs	Low- and Middle-Income Countries
MDHS	Malawi Demographic and Health Survey
ML	Machine Learning
NSO	National Statistical Office
PR-AUC	Area Under the Precision-Recall Curve
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
U5M	Under-five Mortality
UNICEF	United Nations Children's Fund
VIF	Variance Inflation Factor
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

CHAPTER ONE

1.0 Introduction

Infant mortality remains a major public health concern globally, particularly in low- and middle-income countries where preventable deaths continue to occur despite improvements in service coverage, donor support and policy attention. Although child survival has improved substantially since 1990, sub-Saharan Africa continues to carry the highest burden of child mortality in the world, and infancy remains one of the most vulnerable periods of life (UNICEF, 2026).

Malawi has made considerable progress in improving child survival over the past three decades. According to the 2024 Malawi Demographic and Health Survey, under-five mortality fell from 234 deaths per 1,000 live births in the five years before the 1992 survey to 48 deaths per 1,000 live births in the five years before the 2024 survey. Over the same period, infant mortality dropped from 135 to 35 deaths per 1,000 live births, while neonatal mortality was estimated at 24 deaths per 1,000 live births. The report further shows that neonatal deaths account for 69% of infant deaths and 50% of under-five deaths in Malawi, highlighting infancy as a particularly critical stage for child survival (NSO and ICF, 2025).

According to Ntenda, a number of studies have identified socioeconomic, maternal, cultural, household, environmental, biological, and health service utilization factors as determinants of infant mortality. The factors include the child's characteristics, the mother's health and reproductive history, household socioeconomic conditions, environment, and access to maternal and newborn health services. Previous studies have also demonstrated that factors such as short birth intervals, maternal age, multiple births, prematurity, place of residence, and maternal education are linked to infant survival in low- and middle-income countries, including countries in sub-Saharan Africa like Malawi. (Manda, 1999; Baraki et al., 2020).

The increasing availability of nationally representative survey data and the growth of advanced analytical methods have created new opportunities to move beyond traditional descriptive and regression-based analyses towards predictive modelling. This study aims to build and evaluate machine learning prediction models for infant

mortality using Malawi Demographic and Health Survey data. Logistic regression will be used as a baseline model, while Random Forest and XGBoost will be applied as alternative machine learning approaches to assess whether predictive performance can be improved and to identify the most important predictors of infant death. Based on evidence from DHS based studies in low- and middle-income countries suggests that machine learning methods can improve mortality prediction by capturing nonlinear relationships and complex interactions that conventional models may not be able to establish (Adegbosin et al., 2020).

1.1 Background to the Study

Infant mortality is one of the most widely used indicators of population health, health system performance, and social development. Deaths during the first year of life often reflect broader weaknesses in maternal health, newborn care, nutrition, living conditions, environmental exposures, and access to preventive and curative services. Globally, although mortality levels have fallen significantly, progress has slowed and the remaining burden is heavily concentrated in sub-Saharan Africa (UNICEF, 2026).

In Malawi, the 2024 MDHS shows that child survival has improved greatly and infant mortality has declined. However, some important risks remain. Infant mortality is still higher among children born less than two years after a previous birth and among births classified as avoidable high-risk fertility. This shows that reproductive factors, especially short birth intervals, continue to play an important role in child survival (NSO and ICF, 2025). Earlier research from Malawi also found that birth interval and maternal age have a particularly strong influence on survival during infancy (Manda, 1999).

The predictors proposed for this study include sex of the child, birth order, preceding birth interval, multiple birth status, size at birth, gestational proxy, maternal age, education, parity, marital status, household wealth, place of residence, sanitation, water source, antenatal care, place of delivery, and skilled attendance at birth. These variables reflect both the literature and the available DHS data structure, and they provide a strong basis for predictive modelling of infant mortality in Malawi (Baraki et al., 2020; WHO, 2024).

1.2 Problem Statement

Despite the progress Malawi has made in reducing child mortality, infant mortality remains an important public health problem. The 2024 MDHS reports an infant mortality rate of 35 deaths per 1,000 live births, indicating that a substantial number of children still die before reaching their first birthday. The report further shows persistent reproductive and demographic risk patterns, including higher mortality among births following short intervals and among births in avoidable high-risk fertility categories (NSO and ICF, 2025).

Another challenge facing the country is that much of the existing evidence in Malawi and comparable settings has relied mainly on traditional statistical approaches to identify associated factors, with limited use of predictive approaches based on machine learning. While logistic regression remains useful and easy to interpret, it may not fully capture all the complex patterns in large datasets such as the DHS. Evidence from multi-country DHS analyses suggests that machine learning methods may provide improved predictive performance for child mortality outcomes (Adegbosin et al., 2020).

Therefore, this study addresses the lack of Malawi-specific evidence on the use of modern machine learning methods to predict infant mortality. Without this type of evidence, it may be difficult to identify infants at higher risk, understand which factors contribute most to infant deaths, and design more targeted public health interventions.

1.3 Justification of the Study

This study is important for several reasons. First, infant mortality remains a major public health concern in Malawi despite the progress reported in the 2024 MDHS. Second, the study offers a methodological contribution by comparing traditional regression with machine learning approaches.

From a public health point of view, infant mortality remains a key performance indicator of national health performance and equity. Although Malawi has made major gains in child survival, the continued burden of infant deaths indicates the need for better tools to identify risk and guide intervention. The 2024 MDHS provides a recent and

nationally representative dataset that is well suited for this study. The data were collected and cleaned through a well-designed survey process by trained and experienced personnel, which strengthens its reliability for analysing infant mortality in Malawi (NSO and ICF, 2025).

Machine learning offers several advantages over conventional regression-based approaches in prediction tasks, particularly where relationships among variables are nonlinear, interactive and complex. Adegbosin et al. (2020) demonstrated that deep learning methods could predict under-five mortality effectively using DHS data from 34 low- and middle-income countries, this underlines the relevance of machine learning for mortality prediction using survey data. However, logistic regression is still important because it is easy to interpret and widely used in epidemiology and public health research. Comparing logistic regression with machine learning methods is therefore useful because it helps assess both predictive performance and practical relevance.

From a policy point of view, identifying the main predictors of infant mortality in Malawi can help to identify risky areas and guide where interventions should be prioritised. This may include strengthening maternal and newborn care, promoting appropriate birth spacing, improving household welfare, and increasing access to essential health services. This is important because infancy still accounts for a large percentage of child deaths, and survival outcomes continue to be affected by health service gaps and existing inequalities (NSO and ICF, 2025; UNICEF, 2026).

Furthermore, this study contributes to the growing literature on the use of machine learning in mortality prediction and public health more broadly. It also addresses a problem that is directly relevant to Malawi by using recent, nationally representative data. In addition, the study responds to the need for more research applying predictive methods to mortality data in low-resource settings, especially for infant mortality, which remains an important public health and programme priority.

1.4 Aim of the Study

The aim of this study is to develop and evaluate machine learning prediction models for infant mortality using Malawi Demographic and Health Survey data.

1.5 Objectives of the Study

The specific objectives of the study are:

1. To identify key predictors of infant mortality using Malawi Demographic and Health Survey data.
2. To develop prediction models for infant mortality using logistic regression, Random Forest, and XGBoost.
3. To compare the predictive performance of the developed models using measures such as AUC-ROC, accuracy, sensitivity, specificity, precision, and F1-score.
4. To determine which modelling approach offers the best balance between predictive performance and interpretability in predicting infant mortality in the Malawian context.

1.6 Significance of the Study

The study is significant since it has a potential to contribute both practical and methodological evidence. Practically, it may help identify high-risk infants and inform more targeted maternal, newborn, and child health interventions. Methodologically, it may demonstrate whether machine learning approaches offer added value over traditional models in predicting infant mortality using survey data. In addition, by using recent Malawi data, the study contributes context-specific evidence that may be more relevant to local policy and programme planning than evidence borrowed from other settings.

1.7 Structure of the Thesis

This thesis is organised into five chapters. Chapter One presents the introduction, background, problem statement, justification, aim, objectives, and significance of the study. Chapter Two reviews the relevant literature on infant mortality and the use of

machine learning in mortality prediction. Chapter Three presents the methodology, including study design, data source, variables, model development, and evaluation strategy. Chapter Four presents the results of the analysis and model performance. Chapter Five discusses the findings, conclusions, limitations, and recommendations.

CHAPTER TWO

2.0 Literature Review

2.1 Introduction

This chapter reviews literature relevant to infant mortality and the use of machine learning in mortality prediction. It begins with an overview of the burden and determinants of infant mortality, followed by a discussion of the use of machine learning methods in public health prediction. The chapter then identifies gaps in the literature that justify the present study.

2.2 Overview of Infant Mortality

Infant mortality is widely recognised as an important indicator of health status, health system effectiveness, and social development. Deaths during the first year of life reflect the combined influence of maternal health, nutrition, living conditions, environmental exposures, and access to preventive and curative services. The infant period includes both the neonatal stage, which is closely tied to pregnancy, childbirth, and immediate newborn care, and the post-neonatal stage, which is influenced by infection, feeding, immunisation, and household conditions (UNICEF, 2026).

Globally, mortality among infants and young children has declined substantially over the past three decades, but disparities remain stark. Sub-Saharan Africa continues to have the highest under-five mortality rate in the world, and progress has slowed in recent years. Malawi reflects this broader regional pattern. Although mortality has fallen markedly, infant mortality remains high enough to justify continued analytical and policy attention (UNICEF, 2026; NSO and ICF, 2025).

2.3 Determinants of Infant Mortality

The literature consistently shows that infant mortality is influenced by factors operating at multiple levels, including the child, mother, household, and wider community. Child-level predictors include sex of the child, birth order, preceding birth interval, multiple birth status, size at birth, and prematurity or gestational duration. Maternal factors include age, education, marital status, employment, parity, and reproductive history.

Household and contextual factors include wealth, residence, region, sanitation, water source, and access to health services. These patterns have been reported in both DHS-based country studies and broader multi-country analyses (Manda, 1999; Baraki *et al.*, 2020).

Oduse (2024) found that infant mortality is shaped by socioeconomic, demographic, healthcare and biological factors. Key predictors included household wealth, maternal education, women's employment status, antenatal care attendance, place of delivery, skilled birth assistance, number of children ever born, region and birth size. These findings suggest that infant survival is influenced by both household conditions and access to maternal and newborn health services.

Similarly, Bizzego *et al.* (2021) used a machine learning approach to examine predictors of under-five mortality in low- and middle-income countries. Their study identified important predictors, including maternal age, education of the household head, household wealth, cooking fuel, toilet facilities, refrigeration and drinking water quality. This highlights that child mortality is influenced by wider social, environmental and household conditions, many of which are preventable or modifiable through improved education, healthcare access, sanitation and poverty reduction.

In Malawi specifically, the 2024 MDHS highlights the continued importance of birth interval and fertility-related risk. Children born less than two years after the preceding birth experience substantially higher infant mortality than those born after longer intervals, and births in high-risk fertility categories continue to show poorer survival outcomes (NSO and ICF, 2025). These findings suggest that reproductive patterns remain central to infant survival and should be incorporated carefully into predictive models.

Health-service variables also remain important. Skilled attendance at birth is widely recognised as a key component of safe childbirth and immediate newborn care, while antenatal care provides opportunities for risk identification, counselling, and linkage to skilled delivery services (WHO, 2023 UNICEF, 2024). For this reason, service-use factors such as antenatal care, place of delivery, and skilled birth attendance remain relevant predictors for infant mortality modelling.

2.4 Machine Learning in Infant Mortality Prediction

Machine learning refers to computational methods that enable algorithms to learn from data and generate predictions or classifications. In health research, machine learning is increasingly used for tasks such as diagnosis, risk stratification, and outcome prediction. Its key advantage lies in its ability to model nonlinear associations, interactions, and complex data structures more flexibly than many conventional statistical methods (Black et al., 2023).

In mortality research, machine learning has shown promising results. Adegbosin *et al.* (2020) demonstrated that deep learning methods could be used effectively to predict under-five mortality across multiple low- and middle-income countries using DHS data. Such evidence supports the rationale for applying machine learning methods to infant mortality prediction in Malawi, especially given the breadth of predictors available in nationally representative survey data.

Mfateneza et al. (2022) applied machine learning to predict infant mortality in Rwanda using Demographic and Health Survey data. Their study demonstrates how routinely collected population-based survey data can be used to develop predictive models for identifying infants at higher risk of death. The application of machine learning in this context is important because infant mortality is influenced by multiple interrelated factors, including maternal, household, socioeconomic and birth-related characteristics. Unlike traditional statistical approaches that mainly test predefined relationships, machine learning can explore complex patterns within large datasets and support more targeted public health planning.

Dessie and Adisu (2025) also applied machine learning to predict infant mortality in Ethiopia. Their study highlights the growing use of advanced computational methods in maternal and child health research across Africa. By applying machine learning to national health survey data, the study shows how predictive analytics can support early identification of vulnerable infants and inform interventions aimed at improving child survival. However, such models need to be interpreted carefully, especially in low-resource settings where data quality, model transparency and practical implementation within health systems remain important considerations.

In this study, logistic regression is proposed as the baseline model because it is widely accepted in epidemiology and public health and produces interpretable results. Random Forest and XGBoost are proposed as machine learning alternatives that may improve predictive performance by capturing more complex relationships in the data. Model performance will be evaluated using AUC-ROC, accuracy, sensitivity, specificity, precision, and F1-score. This approach is appropriate because prediction accuracy alone is not sufficient in public health research; models must also produce findings that are understandable, meaningful, and potentially actionable.

2.5 Literature Gap

Three main gaps emerge from the reviewed literature. First, much of the literature on infant and child mortality focuses on identifying associated factors rather than building predictive models for practical use. Second, while machine learning applications in mortality prediction are increasing, evidence from Malawi remains limited, especially using recent nationally representative DHS data. Third, relatively few studies focus specifically on infant mortality as a distinct outcome, despite its epidemiological importance and strong relevance to maternal and newborn care. These gaps justify a focused study on infant mortality prediction in Malawi using recent DHS data and comparative modelling approaches.

2.6 Chapter Summary

This chapter has reviewed literature on infant mortality and on the use of machine learning in mortality prediction. The evidence shows that infant mortality remains strongly shaped by demographic, socioeconomic, reproductive, environmental, and health service factors, and that machine learning methods may offer useful advantages for prediction. However, important gaps remain in Malawi-specific predictive evidence using nationally representative data. These gaps provide the basis for the present study.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter describe the methodology that will be used to develop and evaluate machine learning models for predicting infant mortality in Malawi using the 2024 Malawi Demographic and Health Survey data. The chapter explains the research design, data source, study population, sampling approach, variables, data preparation procedures, model development, model evaluation, ethical considerations, and methodological limitations.

Quantitative predictive modelling approach will be used, this approach is appropriate because the study is not only aimed at describing factors associated with infant mortality, but is also to develop models that would classify and predict whether an infant is at risk of dying before reaching the first birthday or not. Logistic regression will be used as the baseline model because of it simple, widely used and interpretability in public health research. Random Forest and XGBoost will then be b uilt as machine learning models to assess whether more flexible algorithms can improve predictive performance by capturing complex and non-linear relationships in the data (Breiman, 2001; Chen & Guestrin, 2016).

3.1 Research Philosophy and Approach

This study will follow a positivist research philosophy because it uses measurable variables, structured survey data, and statistical modelling to predict infant mortality. Infant mortality will be treated as a binary outcome, showing whether a child died before reaching 12 months or survived infancy.

A deductive approach will be used because the predictor variables will be guided by existing literature on infant mortality and child survival. Previous studies show that infant mortality is influenced by child, maternal, household, environmental, and health service factors (Manda, 1999; Baraki et al., 2020; NSO and ICF, 2025).

3.2 Research Design

This study will use a quantitative secondary data analysis design. It will analyse individual-level data from the 2024 Malawi Demographic and Health Survey. This design is suitable because the MDHS provides nationally representative data that can be used for statistical analysis and predictive modelling.

Although the DHS is cross-sectional, the infant mortality outcome will be derived from retrospective birth history data collected from women aged 15–49 years. These records include information on live births, child survival status, child age, and age at death for children who died. This makes it possible to define infant mortality as death before 12 months of age (The DHS Program, 2023; NSO & ICF, 2025).

The main steps will include selecting relevant variables, cleaning and recoding the data, defining the infant mortality outcome, conducting descriptive analysis, developing logistic regression, Random Forest and XGBoost models, evaluating model performance, and interpreting the best-performing model.

3.3 Study Setting

The study setting is Malawi, a low-income country in southern Africa. The country has made substantial progress in reducing child mortality over the past three decades, but infant mortality remains an important public health concern. The 2024 Malawi Demographic and Health Survey reported an infant mortality rate of 35 deaths per 1,000 live births. The report further shows that neonatal mortality accounted for a large proportion of infant deaths. (NSO & ICF, 2025).

Malawi has demonstrated to be a good setting for this study, especially considering the fact that infant mortality remains a key indicator for monitoring maternal, newborn and child health outcomes. The availability of recent nationally representative DHS data also provides a strong basis for developing and evaluating prediction models that are relevant to the country context.

3.4 Data Source

The analysis will use secondary data from the 2024 Malawi Demographic and Health Survey. The MDHS is part of the global DHS programme, which supports countries to collect nationally representative data on population, health, nutrition, fertility, mortality,

and service utilization indicators in a standardized manner. (The DHS Program, 2023). The survey used a nationally representative stratified two-stage cluster sampling design and the sampling frame was based on the 2018 Malawi Population and Housing Census. Firstly, clusters were selected and then households were selected from the listed households within each selected cluster. (NSO & ICF, 2025).

The dataset for this study will be the children's or birth recode file. This file contains information on children born to interviewed women, including survival status, age, age at death, maternal characteristics, household characteristics, reproductive history, and selected health service variables. The analysis will be conducted after obtaining permission to access the DHS dataset from the DHS Program.

3.5 Study Population

The target population for this study will be children born to women aged 15–49 years who participated in the 2024 MDHS. The analytical population will include children whose infant survival outcome can be clearly determined from the DHS birth history data.

The study will focus on live births because infant mortality is defined among live births. Stillbirths will not be included in the outcome definition because they are not part of the standard infant mortality denominator. Children who are alive but younger than 12 months at the time of the survey will be excluded from the main analysis because their final infant survival status is not yet fully observed. This is important because such children may still experience infant death before reaching their first birthday.

The main analytical sample will therefore include:

- children who were alive and aged at least 12 months at the time of the survey;
- children who died before reaching 12 months of age; and
- children who died after 12 months of age, who will be treated as having survived infancy for the purpose of the binary infant mortality outcome.

Records with missing information on survival status, age, or age at death will be excluded from the main analysis.

3.6 Inclusion and Exclusion Criteria

3.6.1 Inclusion Criteria

The study will include:

- live births recorded in the 2024 MDHS birth history data;
- children whose survival status is known;
- children with sufficient information to determine whether death occurred before 12 months of age; and
- records with relevant predictor variables available for modelling.

3.6.2 Exclusion Criteria

The study will exclude:

- stillbirths, because the outcome is infant mortality among live births.
- live children younger than 12 months at the time of the survey, because their infant survival status is not yet complete.
- records with missing survival status or missing age at death where the outcome cannot be determined.

3.7 Study Variables

3.7.1 Outcome Variable

The outcome variable will be infant mortality. It will be defined as death before reaching the first birthday. The outcome will be coded as a binary variable:

- **1 = infant death**, if the child died before 12 months of age;
- **0 = survived infancy**, if the child was alive at 12 months or died after reaching 12 months.

The outcome will be derived using DHS variables on survival status, current age of the child, and age at death. For children who died, age at death will be used to determine whether death occurred before 12 months. For children who are alive, only those aged at least 12 months will be included in the main analysis.

(The DHS Program, 2023; NSO & ICF, 2025).

3.7.2 Predictor Variables

Predictor variables will be selected based on the literature reviewed in Chapter Two, the study objectives, and the availability of variables in the 2024 MDHS dataset. They will be grouped into child, maternal, household, environmental, and health service factors.

These may include child sex, multiple birth status, birth order, preceding birth interval, size at birth, mother's age, education, marital status, parity, wealth index, residence, region, household size, water source, sanitation, antenatal care, place of delivery, and skilled birth attendance.

3.8 Data Management and Preparation

The DHS dataset will be downloaded after permission is granted by The DHS Program secretariate. The data will be imported into Python for cleaning, analysis and modelling. Python is appropriate for this study because it allows data management steps to be documented and repeated. In addition, it provides established libraries for statistical analysis and machine learning, including pandas, NumPy, scikit-learn and XGBoost (Pedregosa et al., 2011).

The data preparation process will include the following steps.

3.8.1 Data Inspection

The dataset will first be inspected to understand the number of records, variable names, labels, data types, missing values, and coding structure. Frequency tables will be generated for categorical variables, while summary statistics will be produced for continuous variables. This step will help identify outliers, missing values and variables that require recoding.

3.8.2 Outcome Variable Derivation

A new binary infant mortality variable will be created using survival status and age at death. Children who died before 12 months will be coded as infant deaths. Children

who survived beyond 12 months will be coded as non-infant deaths or survivors of infancy. Live children younger than 12 months will be excluded from the main analysis to reduce the risk of outcome misclassification.

3.8.3 Predictor Recoding

Predictor variables will be recoded into meaningful categories. For example, maternal education may be grouped as no education, primary, secondary, and higher education. Wealth index will be classified using the DHS wealth quintiles. Residence will be coded as rural or urban

Categorical variables will be converted into suitable formats for modelling.

3.8.4 Missing Data Handling

Missing data will be assessed before model development and the magnitude of missing values will be examined for each variable. Variables with very high levels of missingness may be excluded if they are unlikely to contribute reliable information to the models. For variables with low or moderate missingness, suitable imputation methods will be applied. For numerical variables, median imputation may be used because it is less affected by extreme values.

3.8.5 Survey Weights

Descriptive analysis will use DHS sampling weights to improve representativeness. For machine learning model development, the main focus will be predictive performance at the individual record level. Sample weights will also be considered during model fitting or sensitivity analysis, particularly for logistic regression and algorithms that allow weighted training.

3.9 Data Analysis Plan

The analysis will be conducted in two main stages: descriptive analysis and predictive modelling.

3.9.1 Descriptive Analysis

Descriptive analysis will be conducted to summarize the study sample and describe the distribution of the outcome and predictor variables. Frequencies and percentages will be used for categorical variables, while means, medians and standard deviations will be used for continuous variables where appropriate.

Cross-tabulations will also be produced to compare infant mortality across key predictors such as sex of child, birth interval, multiple birth status, maternal education, wealth quintile, residence, region, place of delivery, and skilled birth attendance.

The descriptive analysis will help provide context before model development and will support interpretation of the predictive modelling results.

3.9.2 Bivariate Analysis

Bivariate analysis may be conducted to explore crude relationships between each predictor and infant mortality. Chi-square tests will be used for categorical variables, while appropriate statistical tests will be used for continuous variables depending on their distribution.

3.10 Model Development

Three models will be compared: logistic regression, Random Forest, and XGBoost. These models were selected because they allow the study to compare a simple and interpretable baseline model with more flexible machine learning models that can capture complex patterns in the data.

3.10.1 Training and Testing Split

The cleaned analytical dataset will be divided into training and testing datasets. A stratified split will be used to ensure that the proportion of infant deaths and non-infant deaths is maintained in both datasets. 80 percent of the data may be used for training, while the remaining 20 percent will be reserved for testing. The training dataset will be used to build and tune the models, while the testing dataset will be used for final model evaluation.

3.10.2 Logistic Regression Model

Logistic regression will be used as the baseline model. It is widely used in public health because it is interpretable. In this study, logistic regression will provide a useful comparison against more complex machine learning models.

Before fitting the model, multicollinearity among predictors will be assessed using correlation analysis and variance inflation factors where appropriate. Variables with very high collinearity may be removed or combined. The logistic regression model will help identify predictors associated with higher or lower odds of infant mortality.

3.10.3 Random Forest Model

Random Forest is a machine learning method that builds multiple decision trees and combines their predictions to improve classification performance. It is useful because it can handle non-linear relationships, interactions between variables, and mixed types of predictors (Breiman, 2001).

In this study, the Random Forest model will be trained using the same predictor set as the logistic regression model. Hyperparameters such as the number of trees, maximum tree depth, minimum samples per split, minimum samples per leaf, and number of features considered at each split will be tuned using cross-validation. The model will be used to examine variable importance and this will help identify which predictors contribute most strongly to the model's classification decisions.

3.10.4 XGBoost Model

XGBoost will be used as the second machine learning model. XGBoost is a gradient boosting algorithm that builds trees sequentially, with each new tree attempting to correct errors made by previous trees. It is widely used in classification tasks because of its strong predictive performance and ability to handle complex relationships (Chen & Guestrin, 2016). The model is particularly useful in this study because infant mortality is expected to be a rare outcome, and the model allows adjustment for class imbalance.

3.11 Handling Class Imbalance

Infant mortality is expected to be a minority class in the dataset because most children survive beyond infancy. This creates a class imbalance problem, where a model may achieve high overall accuracy simply by predicting most children as survivors.

To address this issue, the study will apply several strategies. First, the training and testing split will be stratified by outcome class. Second, class weights may be used to give greater importance to the minority class during model training. Third, sensitivity analysis may be conducted using resampling methods such as the Synthetic Minority Oversampling Technique. (Chawla et al., 2002).

3.12 Model Evaluation

Model performance will be evaluated using several classification metrics. This is important because no single metric is sufficient to assess model performance, especially where the outcome is imbalanced.

The following metrics will be used:

Metric	Meaning
Accuracy	Proportion of all correctly classified cases
Sensitivity/Recall	Proportion of infant deaths correctly identified
Specificity	Proportion of survivors correctly identified
Precision	Proportion of predicted infant deaths that are true infant deaths
F1-score	Balance between precision and recall
ROC-AUC	Ability of the model to distinguish between infant deaths and survivors across thresholds
PR-AUC	Performance in identifying the minority class, especially useful in imbalanced datasets

Sensitivity will be important because the model should be able to identify children at risk of infant mortality. However, specificity and precision will also be considered to avoid wrongly classifying too many children as high-risk. The F1-score will balance precision and recall, while ROC-AUC and PR-AUC will assess the model's ability to distinguish between risk groups, especially where class imbalance exists. Confusion

matrices will be used to show true positives, false positives, true negatives, and false negatives.

3.13 Model Interpretation

Model interpretation will be important because the study aims not only to predict infant mortality, but also to understand the factors that influence the predictions. This will make the findings more useful for public health planning and decision-making.

For logistic regression, interpretation will be based on coefficients and odds ratios, which show the direction and strength of the relationship between each predictor and infant mortality. For Random Forest and XGBoost, feature importance and if necessary, SHAP values will be used to identify the variables that contribute most to the predictions.

3.14 Software and Tools

The analysis will be conducted using Python. The main libraries expected to be used include:

- pandas and NumPy for data management;
- pyreadstat for importing SPSS DHS files where necessary;
- scikit-learn for preprocessing, logistic regression, Random Forest, cross-validation and model evaluation;
- XGBoost for gradient boosting modelling;
- matplotlib and seaborn for visualisation; and
- SHAP for model explainability, if used.

The use of Python supports reproducibility because data processing steps can be documented in a single analysis workflow. The final code will be saved and organized to allow the analysis to be repeated.

3.15 Ethical Considerations

Secondary anonymized data from the 2024 Malawi Demographic and Health Survey. The DHS data are publicly available to registered users after approval from The DHS

Program. The dataset does not include direct personal identifiers such as names or exact household addresses.

The original 2024 MDHS received ethical approval through the relevant institutional review processes. Informed consent was obtained from survey participants during data collection, and the survey followed DHS ethical procedures for household surveys (NSO & ICF, 2025).

3.16 Validity, Reliability and Reproducibility

The validity of this study will be strengthened by using a nationally representative dataset collected through standard DHS procedures. The DHS uses structured questionnaires, trained fieldworkers, electronic data collection systems, and standardized data processing procedures, which improve data quality and comparability across countries and survey rounds (The DHS Program, 2023; NSO & ICF, 2025).

Reliability will be strengthened by clearly documenting all data processing steps. The use of Python code will allow the analysis to be reproduced.

Internal validity for prediction will be supported by separating the data into training and testing sets and by using cross-validation during model tuning. External validity will be limited because the model will be developed using Malawi DHS data only.

3.17 Methodological Limitations

Several methodological limitations may affect the study. First, some variables rely on mothers' recall, including birth history, size at birth, antenatal care, and age at death. This may lead to recall errors, especially for births that happened several years before the survey.

Second, some important clinical factors may not be included, such as birth weight, newborn complications, congenital conditions, quality of care, and detailed clinical history. These variables may be unavailable or may have high levels of missing data in the DHS dataset. Lastly, because infant mortality is a relatively rare outcome, class imbalance may affect how well the models identify infant deaths.

3.18 Chapter Summary

This chapter has described the methodology that will be used to develop and evaluate machine learning models for predicting infant mortality in Malawi. The study will use a quantitative secondary data analysis design based on the 2024 Malawi Demographic and Health Survey. Infant mortality will be defined as death before the first birthday, and predictors will be selected from child, maternal, household, environmental and health service variables.

The study will compare logistic regression, Random Forest and XGBoost models. Model performance will be evaluated using accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC and PR-AUC. Because infant mortality is expected to be an imbalanced outcome, the analysis will include strategies such as stratified splitting, class weighting and possible resampling. The chapter has also outlined ethical considerations, reproducibility measures and methodological limitations.

Overall, the methodology is designed to produce a clear, transparent and contextually relevant predictive modelling study that contributes to understanding infant mortality risk in Malawi.

CHAPTER FOUR: DATA ANALYSIS RESULTS

4.1 Introduction

This chapter presents the results of the data preparation, descriptive analysis, bivariate analysis, multicollinearity assessment, predictive model building and preliminary predictor interpretation for the infant mortality prediction study.

The outcome variable was infant death, coded as 1 for children who died before reaching 12 months and 0 for children who survived infancy. This aligns with the outcome definition in the methodology, where infant mortality was defined as death before the first birthday.

4.2 Analytical Sample and Outcome Distribution

The final analytical dataset contained 50,261 child records with a valid infant survival outcome. Of these, 48,074 children survived infancy, while 2,187 children died before reaching 12 months.

Outcome distribution, unweighted:

infant_death	frequency	percentage
0	48074	95.648714
1	2187	4.351286

Outcome distribution, weighted:

infant_death	weighted_frequency	weighted_percentage
0	47529.625131	95.280084
1	2354.488181	4.719916

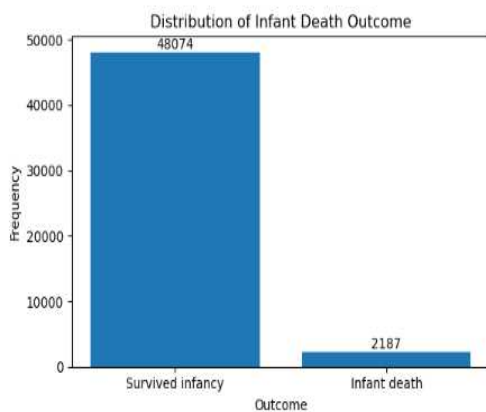


Figure 1: Distribution of Infant Mortality Outcome

Figure 1 shows the distribution of the infant mortality outcome variable. Infant death was relatively rare in the analytical sample. Of the 50,261 children included in the analysis, 2,187 died before reaching 12 months, representing 4.35% of the unweighted sample. The weighted estimate was slightly higher at 4.72%.

These results indicate a highly imbalanced outcome distribution, with the vast majority of children surviving infancy. This class imbalance was considered in subsequent model development, with emphasis placed on appropriate evaluation metrics beyond overall accuracy.

4.3 Missing Data Assessment

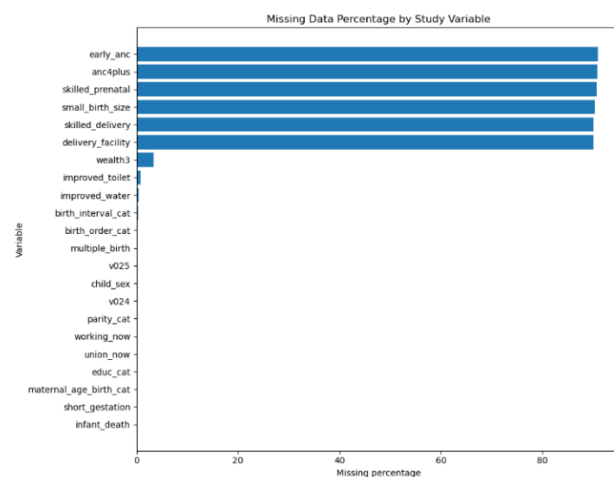


Figure 2: Missing Data Across Study Variables

Figure 2 presents the analysis of missing data across all study variables. Most demographic, maternal, and household variables were highly complete, with missingness below 1%. However, several health service-related variables, including antenatal care, skilled prenatal care, place of delivery, skilled delivery, and birth size, exhibited very high levels of missingness (approximately 90%).

Given the extent of missing data, these variables were excluded from the primary model. However, they were retained for inclusion in an extended sensitivity analysis to assess whether their incorporation improved predictive performance.

4.4 Bivariate Association Between Predictors and Infant Death

Chi-square tests were conducted to examine crude associations between categorical predictors and the outcome variable. The results showed that several predictors were statistically significantly associated with infant death.

Chi-square test results:

	predictor	chi_square	degrees_of_freedom	p_value
5	short_gestation	969.231797	1	8.755256e-213
1	multiple_birth	645.244872	1	2.416928e-142
3	birth_interval_cat	312.507597	3	1.952500e-67
6	maternal_age_birth_cat	197.663062	2	1.196772e-43
10	parity_cat	168.354774	2	2.768378e-37
2	birth_order_cat	87.862790	2	8.333693e-20
0	child_sex	32.759514	1	1.042946e-08
12	v024	35.556382	2	1.901208e-08
4	small_birth_size	16.993653	1	3.750498e-05
13	v025	16.375714	1	5.194657e-05
7	educ_cat	14.129730	2	8.546101e-04
11	wealth3	8.357898	2	1.531459e-02
15	improved_toilet	3.951371	1	4.683321e-02
8	union_now	2.891100	1	8.907007e-02
17	anc4plus	2.781206	1	9.537630e-02
9	working_now	2.526624	1	1.119395e-01
14	improved_water	2.393705	1	1.218246e-01
20	skilled_prenatal	0.458034	1	4.985443e-01
18	delivery_facility	0.430010	1	5.119840e-01
16	early_anc	0.153645	1	6.950763e-01
19	skilled_delivery	0.000000	1	1.000000e+00

Figure 3: Crude Associations Using Chi-Square Analysis

Figure 3 illustrates the results of the chi-square test assessing crude associations between categorical predictors and infant death. The bivariate analysis shows that infant mortality is significantly associated with several demographic and reproductive factors. Short gestation, multiple birth status, birth interval, maternal age at birth, parity, and birth order exhibited the strongest crude associations with infant death.

These findings indicate that reproductive and biological factors play a central role in infant survival. The results are consistent with existing literature showing that infant mortality is strongly influenced by child characteristics, maternal factors, and reproductive history. In contrast, some health service-related variables were not statistically significant at the bivariate level.

4.5 Strength of Association Using Cramer's V

The Cramer's V heatmap was used to show the strength of association between individual predictors and infant death.

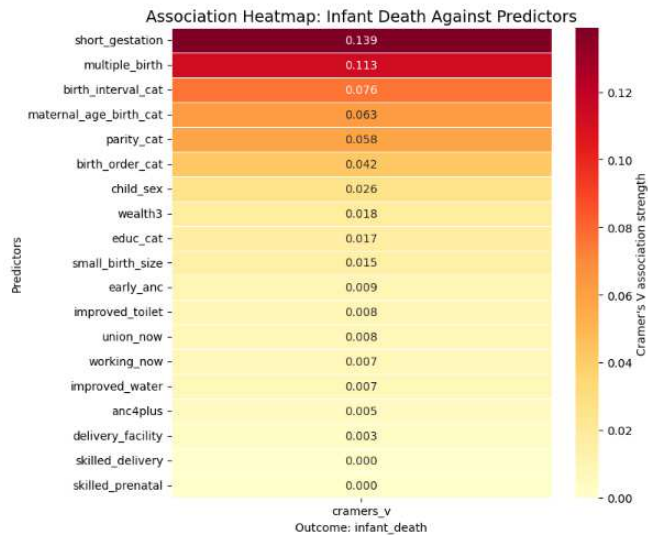


Figure 4: Heatmap of Associations Between Predictors and Infant Death

Figure 4 displays a heatmap of associations between individual predictors and infant death using Cramer's V. The strongest crude associations were observed for short gestation, followed by multiple birth status and birth interval. However, all Cramer's V values were below 0.15, indicating generally weak associations between individual predictors and infant mortality.

This suggests that no single predictor strongly explains infant death in isolation. Instead, infant mortality is influenced by the combined effect of multiple child, maternal, reproductive, and household factors. These findings support the use of multivariable logistic regression and machine learning approaches to capture complex relationships among predictors.

4.6 Adjusted Odds Ratio Results

	term	coefficient	adjusted_odds_ratio	ci_lower	ci_upper	p_value	significant
7	short_gestation_1.0	1.280639	3.598940	3.260487	3.972527	1.745565e-142	Yes
2	multiple_birth_1.0	1.514667	4.547907	3.917244	5.280105	5.284131e-88	Yes
15	parity_cat_3.0	1.039273	2.827162	2.425602	3.295200	2.425634e-40	Yes
4	birth_interval_cat_2.0	-0.731434	0.481218	0.418553	0.553266	9.106018e-25	Yes
5	birth_interval_cat_3.0	-0.807398	0.446017	0.377217	0.527365	3.535198e-21	Yes
8	maternal_age_birth_cat_2.0	-0.421268	0.656214	0.583322	0.738215	2.346796e-12	Yes
1	child_sex_2.0	-0.285405	0.751710	0.687876	0.821467	2.910059e-10	Yes
14	parity_cat_2.0	0.486944	1.627335	1.393539	1.900356	7.579121e-10	Yes
9	maternal_age_birth_cat_3.0	-0.594737	0.551708	0.434785	0.700073	9.865232e-07	Yes
6	small_birth_size_1.0	0.577018	1.780720	1.262152	2.512346	1.017313e-03	Yes

Figure 5: Adjusted Odds Ratios

Figure 5 depicts the adjusted odds ratios for predictors of infant mortality. Multiple birth status and short gestation were the strongest predictors, with odds ratios of 4.5 and 3.6, respectively. Infants from multiple births and those born with short gestation exhibited significantly higher odds of death compared to their counterparts. These findings highlight the importance of reproductive and biological factors in infant survival and provide interpretable epidemiological evidence complementing the machine learning models.

4.7 Multicollinearity Assessment

Before model development, multicollinearity was assessed using both the Cramer's V correlation matrix and variance inflation factors. This was in line with the methodology, which stated that multicollinearity would be assessed before fitting Logistic Regression and that variables with very high collinearity may be removed or combined.

	original_predictor	number_of_dummy_terms	maximum_vif	mean_vif	interpretation
2	birth_order_cat	2	inf	inf	Serious multicollinearity
3	birth_interval_cat	3	inf	inf	Serious multicollinearity
10	parity_cat	2	2.818384	2.468511	Acceptable
7	educ_cat	2	2.568187	2.389735	Acceptable
17	skilled_delivery	1	2.084605	2.084605	Acceptable
16	delivery_facility	1	2.081610	2.081610	Acceptable
6	maternal_age_birth_cat	2	1.922561	1.808946	Acceptable
11	wealth3	2	1.514925	1.387450	Acceptable
15	anc4plus	1	1.138172	1.138172	Acceptable
13	improved_toilet	1	1.113502	1.113502	Acceptable
4	small_birth_size	1	1.095280	1.095280	Acceptable
14	early_anc	1	1.078356	1.078356	Acceptable
18	skilled_prenatal	1	1.060705	1.060705	Acceptable
1	multiple_birth	1	1.027172	1.027172	Acceptable
12	improved_water	1	1.024199	1.024199	Acceptable
8	union_now	1	1.023614	1.023614	Acceptable
9	working_now	1	1.020964	1.020964	Acceptable
5	short_gestation	1	1.016134	1.016134	Acceptable
0	child_sex	1	1.001081	1.001081	Acceptable

Figure 6: Multicollinearity Assessment results.

Figure 6 shows the assessment of multicollinearity among predictors using variance inflation factors (VIF). Two highly correlated variable pairs were identified: birth order and birth interval, and delivery facility and skilled delivery. To reduce redundancy, birth order and skilled delivery were excluded due to strong correlations with birth interval and delivery facility, respectively. The remaining predictors exhibited acceptable VIF values, indicating that multicollinearity was adequately addressed for model development.

4.8 Implications for Model Development

The results of the data preparation and preliminary analysis informed the final modelling strategy. Variables with high missingness were excluded from the main model and retained only for the extended model. Highly collinear variables were reviewed and reduced to avoid redundancy. The final main predictor set therefore included variables with acceptable completeness and acceptable VIF values. Since infant death was a rare outcome, the subsequent model development used stratified train-test splitting and class imbalance strategies, and model performance was evaluated using sensitivity, precision, F1-score, ROC-AUC and PR-AUC rather than accuracy alone.

4.9 Model Development and Evaluation Results

Following the methodology, three models were developed and evaluated namely; Logistic Regression, Random Forest and XGBoost. Logistic Regression was used as the baseline model because of its interpretability in public health research, while Random Forest and XGBoost were used to assess whether machine learning models could improve predictive performance by capturing more complex and non-linear relationships in the data. The models were evaluated using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC and PR-AUC. Since infant death was a minority outcome, greater attention was given to sensitivity, F1-score, ROC-AUC and PR-AUC rather than accuracy alone.

4.10 Performance of the Main Low-Missingness Models

The main low-missingness model included predictors with acceptable completeness and excluded variables with very high missingness. This model was treated as the primary model because it was less affected by incomplete data. The predictors used in this model included household wealth, improved toilet, improved water, birth interval, child sex, parity, multiple birth, marital status, maternal education, maternal age at birth, short gestation and mother's working status.

	model_set	model	threshold_type	threshold	accuracy	precision	recall_sensitivity	specificity	f1_score	roc_auc	pr_auc	true_negative	false_positive	false_negative	true_positive
0	Main low-missingness model	Logistic Regression	Default 0.50	0.500000	0.719089	0.09747	0.661327	0.721714	0.169900	0.732214	0.145444	6940	2676	148	289
1	Main low-missingness model	Logistic Regression	Optimised F1	0.723597	0.922212	0.17636	0.215103	0.954347	0.193814	0.732214	0.145444	9177	439	343	94

=====

Main low-missingness model - XGBoost

=====

	model_set	model	threshold_type	threshold	accuracy	precision	recall_sensitivity	specificity	f1_score	roc_auc	pr_auc	true_negative	false_positive	false_negative	true_positive
0	Main low-missingness model	XGBoost	Default 0.50	0.500000	0.725853	0.097536	0.643021	0.729617	0.169379	0.730159	0.149943	7016	2600	156	281
1	Main low-missingness model	XGBoost	Optimised F1	0.719994	0.922312	0.179104	0.219680	0.954243	0.197328	0.730159	0.149943	9176	440	341	96

=====

Main low-missingness model - Random Forest

=====

	model_set	model	threshold_type	threshold	accuracy	precision	recall_sensitivity	specificity	f1_score	roc_auc	pr_auc	true_negative	false_positive	false_negative	true_positive
0	Main low-missingness model	Random Forest	Default 0.50	0.500000	0.766537	0.100084	0.546911	0.776518	0.169204	0.727281	0.146132	7467	2149	198	239
1	Main low-missingness model	Random Forest	Optimised F1	0.678108	0.924600	0.182178	0.210526	0.957051	0.195329	0.727281	0.146132	9203	413	345	92

Fig 7. Model Performance Evaluation

Figure 7 illustrates the performance of Logistic Regression, Random Forest, and XGBoost on the low-missingness infant mortality dataset. The models showed similar discriminative ability, with ROC-AUC values ranging from 0.727 to 0.732. Logistic Regression achieved the highest ROC-AUC (0.732), followed by XGBoost (0.730) and Random Forest (0.727).

At the default threshold (0.50), Logistic Regression had the highest sensitivity (0.661), followed by XGBoost (0.643), while Random Forest performed lowest (0.546). Conversely, Random Forest achieved higher specificity (0.777) and accuracy (0.767) than the other models. With F1-optimized thresholds, accuracy and specificity increased markedly (≈ 0.92 – 0.95), but sensitivity dropped sharply (≈ 0.21 – 0.22), indicating reduced detection of infant deaths. Under this setting, XGBoost achieved the highest F1-score (0.197) and precision (0.179).

Overall, XGBoost provided the best balance between precision and recall, while Logistic Regression performed best in ROC-AUC and sensitivity, and Random Forest favored specificity.

4.11 Performance of the Extended Sensitivity Models

The extended sensitivity model included all selected predictors, including those with at least 90% incomplete data such as antenatal care, skilled prenatal care, place of delivery, skilled delivery and small birth size. These variables were included to assess whether adding health service and birth-related variables improved predictive performance. The extended model was interpreted cautiously due to high degree of missing data.

EXTENDED MODEL PERFORMANCE COMPARISON															
	model_set	model	threshold_type	threshold	accuracy	precision	recall_sensitivity	specificity	f1_score	roc_auc	pr_auc	true_negative	false_positive	false_negative	true_positive
0	Extended sensitivity model	Logistic Regression	Default 0.50	0.500000	0.720979	0.099188	0.670481	0.723274	0.172810	0.738293	0.163490	6955	2661	144	293
4	Extended sensitivity model	XGBoost	Default 0.50	0.500000	0.714811	0.096346	0.663616	0.717138	0.168262	0.738135	0.155957	6896	2720	147	290
2	Extended sensitivity model	Random Forest	Default 0.50	0.500000	0.753009	0.098824	0.576659	0.761023	0.168731	0.727840	0.146260	7318	2298	185	252
1	Extended sensitivity model	Logistic Regression	Optimised F1	0.754232	0.929772	0.204396	0.212815	0.962354	0.208520	0.738293	0.163490	9254	362	344	93
5	Extended sensitivity model	XGBoost	Optimised F1	0.675546	0.909878	0.167376	0.270023	0.938956	0.206655	0.738135	0.155957	9029	587	319	118
3	Extended sensitivity model	Random Forest	Optimised F1	0.663557	0.924003	0.181287	0.212815	0.956323	0.195789	0.727840	0.146260	9196	420	344	93

Figure 8: Model Performance Evaluation

Figure 8 indicates the performance of Logistic Regression, Random Forest, and XGBoost on the extended infant mortality dataset. The models demonstrated comparable moderate predictive performance, with ROC-AUC values of approximately 0.738 for Logistic Regression and XGBoost, and 0.727 for Random Forest.

At the default threshold (0.50), Logistic Regression (0.670) and XGBoost (0.664) achieved higher sensitivity than Random Forest (0.577), indicating better identification of infant mortality cases. In contrast, Random Forest showed slightly higher accuracy (0.753) and specificity (0.761), reflecting improved classification of survivors. Under the optimized F1 threshold, accuracy increased to approximately 0.92–0.93 and specificity to 0.95–0.96, while sensitivity decreased to approximately 0.20–0.21, indicating a trade-off between reduced false positives and increased missed deaths.

Overall, model performance remained closely aligned across algorithms, with Logistic Regression showing marginally better overall performance, closely followed by XGBoost.

4.12 Improved Main Model With Class Weighting and SMOTE Sensitivity Analysis

To address class imbalance, additional improved models were developed using class weighting and SMOTE sensitivity analysis.

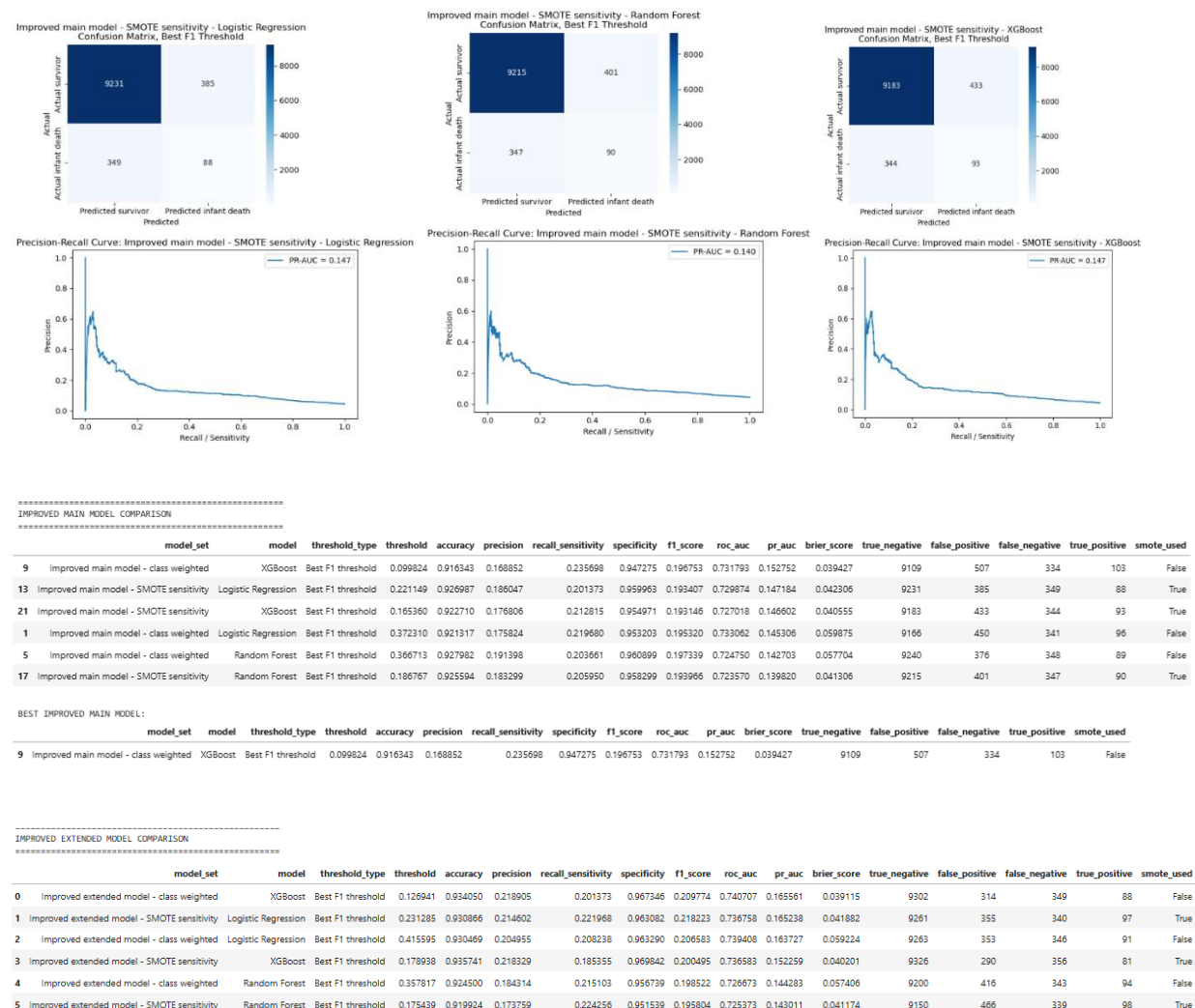


Figure 9: Evaluation of Improved Models

highest F1-score (0.218), while SMOTE Random Forest achieved the highest recall (0.224), indicating that the best-performing model varied depending on the evaluation metric.

Overall, the extended models provided only marginal improvement over the main models, suggesting limited additional predictive value from the extra variables. Given data completeness and practical applicability, the class-weighted XGBoost main model was selected as the final model due to its best overall balance of performance, calibration, and robustness.

4.13 Comparison of the Best Non-Improved and Improved Models

MODEL COMPARISON TABLE													
Model category		Model	Accuracy	Precision	Sensitivity/Recall	Specificity	F1-score	ROC-AUC	PR-AUC	True positives	False negatives	False positives	True negatives
0	Best non-improved model	Extended Logistic Regression	0.9298	0.2044	0.2128	0.9624	0.2085	0.7383	0.1635	93	344	362	9254
1	Best improved model	Improved extended class-weighted XGBoost	0.9341	0.2189	0.2014	0.9673	0.2098	0.7407	0.1656	88	349	314	9302

Figure 10: Comparison of Best Improved and Non-Improved Models

Figure 10 compares the best-performing non-improved and improved models. Overall, both models performed similarly, although the improved extended class-weighted XGBoost model showed a slight overall advantage across most metrics.

The improved extended XGBoost model achieved marginally higher accuracy, precision, specificity, F1-score, ROC-AUC, and PR-AUC, and produced fewer false positives, indicating better classification of infants who survived. In contrast, the extended Logistic Regression model demonstrated slightly higher sensitivity, indicating better detection of infant mortality cases.

Overall, the extended class-weighted XGBoost model provided the best overall performance; however, due to reliance on high-missingness predictors, it was retained only as a sensitivity model. The improved main class-weighted XGBoost model was selected as the primary model due to its strong performance on more complete and reliable data.

4.14 Interpretation of predictor importance in the improved Best Model.

The predictor importance results show that birth interval and short gestation were by far the most influential variables in the improved main XGBoost model, with consensus importance scores of 0.828 and 0.827, respectively. These were followed by parity (0.492), multiple birth (0.457), and maternal age at birth (0.378). Variables such as wealth (0.288) and child sex (0.131) had smaller but still noticeable contributions, while improved water (0.056), education (0.056), improved toilet (0.044), working status (0.017), and union status (0.007) had only limited influence on the model.

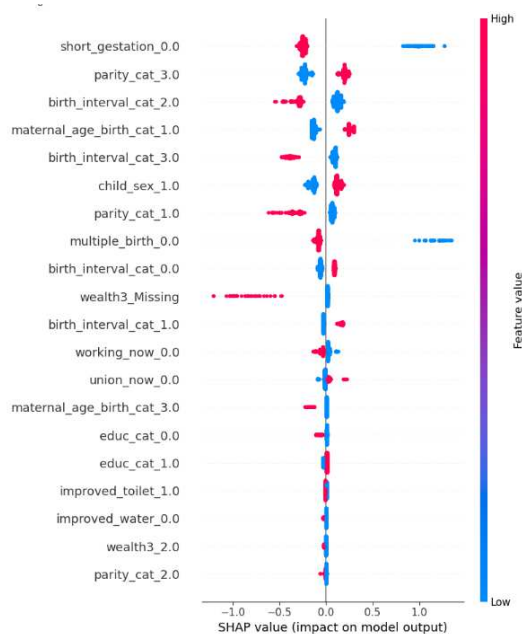
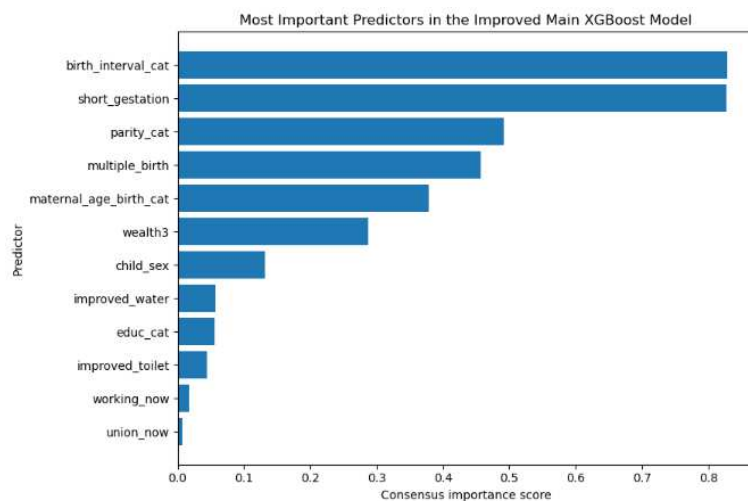


Figure 11: Predictor Importance Results

Figure 11 presents the SHAP-based feature importance results for the final model. The analysis indicates that short gestation, higher parity, shorter birth intervals, multiple birth status, younger maternal age, and male sex are associated with an increased predicted risk of infant mortality. In contrast, longer birth intervals, lower parity, and singleton births are associated with reduced risk.

The widest distribution of SHAP values was observed for short gestation and birth interval, indicating that these variables had the greatest influence on model predictions and were the most important predictors of infant mortality.

Overall, the improved main XGBoost model was primarily driven by reproductive and birth-related factors rather than broader household or socioeconomic characteristics. This suggests that infant mortality risk in this study was largely determined by pregnancy timing, birth history, and maternal reproductive profile, with comparatively limited influence from household living conditions.

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

This chapter discusses the findings of the study in relation to the research aim, objectives, literature reviewed in Chapter Two, and the methodology described in Chapter Three. The study aimed to develop and evaluate machine learning models for predicting infant mortality in Malawi using the 2024 Malawi Demographic and Health Survey data. Specifically, the study sought to identify key predictors of infant mortality, develop prediction models using Logistic Regression, Random Forest and XGBoost, compare their predictive performance, and determine which modelling approach offered the best balance between predictive performance and interpretability.

The chapter begins by summarizing the main findings, followed by a discussion of the results according to the study objectives. It then presents the implications of the findings for public health policy and practice, discusses the methodological contribution of the study, outlines the limitations, and provides recommendations for policy, practice and future research.

5.1 Summary of the Main Findings

The analysis included 50,261 child records with a valid infant survival outcome. Infant death was a relatively rare outcome in the dataset, affecting 2,187 children, which represented 4.35% of the unweighted sample and 4.72% of the weighted sample. This confirmed that the dataset had strong class imbalance, with most children surviving infancy.

The descriptive and bivariate analyses showed that infant death was mainly associated with reproductive and biological factors. Short gestation, multiple birth status, birth interval, maternal age at birth, parity and birth order showed the strongest crude associations with infant death. However, the Cramer's V results showed that the strength of individual associations was generally weak, suggesting that no single

predictor explained infant death on its own. This supported the use of multivariable and machine learning models.

The adjusted Logistic Regression results showed that multiple birth status and short gestation were among the strongest independent predictors of infant death. Higher parity also increased the odds of infant death, while some birth interval categories were associated with lower odds. These findings were supported by model interpretation results, which identified short gestation, birth interval, parity, maternal age at birth and multiple birth status as the most influential predictors.

The model development results showed that XGBoost performed best among the main low-missingness models. After class imbalance handling, the improved main class-weighted XGBoost model was selected as the primary predictive model because it achieved the best balance between predictive performance and data completeness. The improved extended class-weighted XGBoost model achieved the highest overall PR-AUC and better calibration, but because it included variables with very high missingness, it was treated as the best sensitivity model rather than the primary model.

Overall, the study found that machine learning provided some improvement over traditional modelling, but the improvement was modest. Logistic Regression remained important because it produced interpretable odds ratios, while XGBoost offered the strongest predictive performance. These findings support a balanced conclusion: machine learning can add value, but interpretability and data quality remain essential for public health use.

5.2.3 Health service variables and missing data

Health service variables such as antenatal care, place of delivery, skilled prenatal care, skilled delivery and small birth size had very high missingness, with approximately 90% missing for some variables. This affected their use in model development. In line with the methodology, these variables were excluded from the primary low-missingness model and retained only in the extended sensitivity model.

The extended sensitivity model showed slightly better predictive performance than the main model. For example, the improved extended class-weighted XGBoost model

achieved a PR-AUC of 0.166 compared with 0.153 for the improved main class-weighted XGBoost model. It also showed slightly better precision and specificity. However, the extended model relied on predictors with very high missingness and, in some cases, appeared to learn from patterns of missingness rather than from the underlying health service information itself.

This means that the extended model was useful for testing whether additional variables could improve predictive performance, but it was not suitable as the primary model. The improved main class-weighted XGBoost model was therefore retained as the primary selected model because it relied on more complete data and provided a stronger basis for practical interpretation.

5.2.4 Predictive performance of Logistic Regression, Random Forest and XGBoost

The study compared Logistic Regression, Random Forest and XGBoost. Logistic Regression was included as the baseline model because it is widely used in public health and provides interpretable odds ratios. Random Forest and XGBoost were included because they can capture non-linear relationships and interactions more flexibly.

The results showed that XGBoost performed best among the main low-missingness models. The improved main class-weighted XGBoost model achieved the best PR-AUC among the main models and was therefore selected as the primary predictive model. Although the improvement over Logistic Regression was modest, it offered the best balance between predictive performance and data completeness.

The improved extended class-weighted XGBoost model achieved the strongest overall predictive performance, with slightly higher PR-AUC, precision, specificity, F1-score and ROC-AUC than the improved main model. However, because it included predictors with very high missingness, it was retained as the best sensitivity model rather than the primary model. This distinction is important because model performance needs to be interpreted alongside data quality.

Random Forest did not outperform Logistic Regression or XGBoost. This may be because the predictors were mostly categorical and may not have contained enough complex non-linear signal for Random Forest to provide a clear advantage. Logistic Regression, however, remained highly competitive and performed strongly relative to the machine learning models. This suggests that, for this dataset, the additional predictive gain from machine learning was present but small. XGBoost offered the strongest overall prediction, but Logistic Regression remained valuable because it was easier to interpret and produced meaningful odds ratios.

5.2.5 Model improvement and class imbalance handling

The study tested model improvement using class weighting and SMOTE sensitivity analysis. Class weighting improved the XGBoost model and was more useful than SMOTE for the main analysis. The improved main class-weighted XGBoost model achieved a PR-AUC of 0.153 and sensitivity of 23.57%, which was slightly better than the earlier main XGBoost model.

The improved extended class-weighted XGBoost model achieved the highest overall PR-AUC of 0.166 and the lowest Brier score, suggesting better ranking of the minority class and better probability calibration. However, because it was based on variables with very high missingness, it was interpreted as the best sensitivity model rather than the primary model.

The comparison between the best non-improved model and the best improved model showed that the gain from model improvement was modest. The best non-improved model was the extended Logistic Regression model, while the best improved model was the improved extended class-weighted XGBoost model. The improved XGBoost model had higher accuracy, precision, specificity, F1-score, ROC-AUC and PR-AUC, but the Logistic Regression model had slightly higher sensitivity and identified more infant deaths. This shows that the improvement was useful, but small.

Overall, the findings support the use of class weighting as the main imbalance-handling approach and SMOTE as a sensitivity analysis. They also reinforce the importance of evaluating models using multiple metrics rather than relying on a single performance measure.

5.2.6 Interpretability and public health usefulness.

A major objective of the study was to identify a modelling approach that balances prediction and interpretability. The findings show that XGBoost was the strongest predictive approach, while Logistic Regression remained essential for interpretation. Logistic Regression explained the direction and magnitude of relationships through odds ratios, while XGBoost interpretation methods identified the predictors that contributed most strongly to prediction.

The interpretation results showed that short gestation, birth interval, parity, maternal age at birth and multiple birth status were the most meaningful predictors of infant death. These predictors are understandable and relevant to public health. They point to areas such as reproductive health, birth spacing, maternal risk identification, and improved care for high-risk pregnancies and newborns.

However, the modest sensitivity and PR-AUC values show that the model should not be used as a stand-alone decision-making tool without further validation. Instead, the findings are more useful for understanding risk patterns and guiding further work on risk stratification. Any operational use of the model would require external validation, calibration assessment, and careful consideration of how the model would be used in practice.

5.3 Discussion in Relation to Study Objectives

Objective 1: To identify key predictors of infant mortality using Malawi Demographic and Health Survey data

This objective was achieved. The study identified short gestation, birth interval, parity, maternal age at birth and multiple birth status as the most important predictors of infant death. These predictors appeared consistently across the statistical and machine learning interpretation methods. The findings show that infant mortality in the dataset was mainly driven by reproductive and biological factors, although household and health service factors also contributed to some extent.

Objective 2: To develop prediction models for infant mortality using Logistic Regression, Random Forest and XGBoost

This objective was achieved. The study developed and evaluated Logistic Regression, Random Forest and XGBoost models using both main low-missingness and extended sensitivity datasets. Logistic Regression served as the interpretable baseline model, while Random Forest and XGBoost were used to assess whether machine learning models could improve predictive performance.

Objective 3: To compare the predictive performance of the developed models

This objective was achieved. The models were compared using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC and PR-AUC. Because infant death was a minority class, PR-AUC, sensitivity and F1-score were given greater attention than accuracy alone. The results showed that XGBoost performed best among the main low-missingness models, while the improved extended class-weighted XGBoost model performed best overall in the sensitivity analysis. However, the overall improvements were modest.

Objective 4: To determine which modelling approach offers the best balance between predictive performance and interpretability

This objective was achieved. The improved main class-weighted XGBoost model was selected as the primary predictive model because it performed best among the low-missingness models and did not rely heavily on incomplete variables. The improved extended class-weighted XGBoost model was identified as the best sensitivity model because it achieved the highest overall PR-AUC, but it was interpreted cautiously because of very high missingness in the added predictors. Logistic Regression remained important as the interpretable baseline model because it provided odds ratios and helped explain the direction of predictor effects.

5.5 Methodological Contribution of the Study

This study contributes methodologically by applying and comparing traditional and machine learning approaches to infant mortality prediction using recent Malawi DHS data. It demonstrates that machine learning can provide some improvement in predictive performance, but that the gain may be modest when using survey-based variables alone.

The study also shows the value of combining predictive performance with interpretability. XGBoost provided the strongest prediction, while Logistic Regression provided clearer explanation through odds ratios. At the same time, the study demonstrates the importance of distinguishing between a primary model, selected on the basis of both performance and data completeness, and a sensitivity model, which may perform slightly better but rely on variables with substantial missingness.

In addition, the study highlights the importance of handling class imbalance carefully. The findings show that accuracy alone is not enough when the outcome is rare. PR-AUC, sensitivity, F1-score and confusion matrix results provided a more realistic assessment of model usefulness.

Finally, the study demonstrates a transparent modelling workflow, including missing data assessment, multicollinearity checks, main and sensitivity modelling, class imbalance handling, model comparison and interpretation. This strengthens reproducibility and provides a useful approach for similar studies using DHS data.

5.6 Limitations of the Study

The study had several limitations.

First, the analysis used secondary cross-sectional survey data. Although infant survival status was derived from retrospective birth history, the data do not allow strong causal conclusions. The models identify predictors and risk patterns, but they do not prove that these factors directly caused infant death.

Second, some variables relied on maternal recall, including age at death, birth history, size at birth and antenatal care. Recall error may have affected the accuracy of some variables, especially for births that occurred several years before the survey.

Third, several important clinical variables were unavailable or incomplete. These include birth weight, newborn complications, congenital conditions, causes of death, quality of care, timing of care seeking, neonatal infection and detailed clinical history. The absence of these variables may have limited model performance.

Fourth, some health service variables had very high missingness. Although they were included in the extended sensitivity model, their interpretation was limited because the model may have learned missing-data patterns rather than true service-use effects.

Fifth, infant death was a rare outcome, leading to class imbalance. Although stratified splitting, class weighting and SMOTE sensitivity analysis were used, the models still had modest sensitivity and PR-AUC.

Sixth, the model was developed using Malawi DHS data only. External validation using another dataset or future survey round was not conducted. Therefore, the model's generalisability to other populations or future time periods remains uncertain.

5.9 Conclusion

This study developed and evaluated Logistic Regression, Random Forest and XGBoost models for predicting infant mortality in Malawi using 2024 Malawi Demographic and Health Survey data. The findings showed that infant death was a rare outcome, which made prediction challenging and required careful handling of class imbalance.

The most important predictors of infant death were short gestation, birth interval, parity, maternal age at birth and multiple birth status. These findings are consistent with the literature showing that infant mortality is strongly influenced by biological and reproductive factors.

Among the models, the improved main class-weighted XGBoost model was selected as the primary predictive model because it performed best among the low-missingness models and avoided heavy reliance on incomplete variables. The improved extended class-weighted XGBoost model achieved the best overall PR-AUC, but it was treated as the best sensitivity model because it included predictors with very high missingness. Logistic Regression remained important because it provided interpretable odds ratios and helped explain the direction and magnitude of predictor effects.

Overall, the study found that machine learning can add value to infant mortality prediction, but the improvement over Logistic Regression was modest. The findings

therefore support a balanced approach: machine learning may improve prediction, but interpretability and data quality remain essential if the results are to be meaningful for public health decision.

Declarations

Ethics approval and consent to participate

This study used secondary data from the 2024 Malawi Demographic and Health Survey (MDHS), which is publicly available upon request from the DHS Program. The MDHS protocols were reviewed and approved by the relevant national ethics committees, and informed consent was obtained from all participants prior to data collection. No additional ethical approval was required for this secondary analysis of anonymized data.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from the DHS Program upon reasonable request through their official website (<https://dhsprogram.com>). The dataset used is the 2024 Malawi Demographic and Health Survey.

Competing interests

The author declares that there are no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

A. Ngosi conceived the study, designed the methodology, conducted the data analysis, interpreted the results, and drafted and revised the manuscript. The author read and approved the final manuscript.

Acknowledgements

I would like to express sincere gratitude to Dr Ramazan for his guidance and throughout this research. I also acknowledge the University of Portsmouth for providing a supportive learning environment and access to academic resources.

I acknowledge The DHS Program, the Malawi National Statistical Office, and ICF for granting access to the 2024 Malawi Demographic and Health Survey data.

I am especially grateful to my wife, Dr Mercy Ngosi and my lovely daughters Watanja, Watipa and Wantwa, for their continuous support and encouragement throughout this research.

References

1. Adegbosin, A. E., Stantic, B., & Sun, J. (2020). Efficacy of deep learning methods for predicting under-five mortality in 34 low-income and middle-income countries. *BMJ Open*, 10(8), e034524. <https://doi.org/10.1136/bmjopen-2019-034524>
2. Baraki, A. G., Akalu, T. Y., Wolde, H. F., Lakew, A. M., & Gonete, K. A. (2020). Factors affecting infant mortality in the general population: Evidence from the 2016 Ethiopian Demographic and Health Survey (EDHS); a multilevel analysis. *BMC Pregnancy and Childbirth*, 20, Article 299. <https://doi.org/10.1186/s12884-020-03002-x>
3. Bizzego, A., Bornstein, M. H., Gabrieli, G., Deater-Deckard, K., Lansford, J. E., Bradley, R. H., Costa, M., & Esposito, G. (2021). Predictors of contemporary under-5 child mortality in low- and middle-income countries: A machine learning approach. *International Journal of Environmental Research and Public Health*, 18(3), Article 1315. <https://doi.org/10.3390/ijerph18031315>
4. Black, J. E., Kueper, J. K., & Williamson, T. S. (2023). An introduction to machine learning for classification and prediction. *Family Practice*, 40(1), 200–204. <https://doi.org/10.1093/fampra/cmac104>
5. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
8. Croft, T. N., Allen, C. K., & Zachary, B. W. (2023). Guide to DHS statistics: DHS-8. ICF. <https://www.dhsprogram.com/publications/publication-dhsg1-dhs-questionnaires-and-manuals.cfm>
9. Dessie, M. B., & Adisu, M. A. (2025). An artificial neural network approach for predicting infant mortality status in Ethiopia. *BMC Public Health*, 25, Article 3515. <https://doi.org/10.1186/s12889-025-24884-6>

10. Manda, S. O. M. (1999). Birth intervals, breastfeeding and determinants of childhood mortality in Malawi. *Social Science & Medicine*, 48(3), 301–312. [https://doi.org/10.1016/S0277-9536\(98\)00359-1](https://doi.org/10.1016/S0277-9536(98)00359-1)
11. Mfateneza, E., Rutayisire, P. C., Biracyaza, E., Musafiri, S., & Mpabuka, W. G. (2022). Application of machine learning methods for predicting infant mortality in Rwanda: Analysis of Rwanda Demographic Health Survey 2014–15 dataset. *BMC Pregnancy and Childbirth*, 22, Article 388. <https://doi.org/10.1186/s12884-022-04699-8>
12. National Statistical Office (NSO) [Malawi] & ICF. (2025). 2024 Malawi Demographic and Health Survey. NSO and ICF. <https://www.dhsprogram.com/pubs/pdf/FR398/FR398.pdf>
13. Oduse, S., Zewotir, T., & North, D. (2021). The impact of antenatal care on under-five mortality in Ethiopia: A difference-in-differences analysis. *BMC Pregnancy and Childbirth*, 21, Article 44. <https://doi.org/10.1186/s12884-020-03531-5>
14. Samuel, O., Zewotir, T., & North, D. (2024). Application of machine learning methods for predicting under-five mortality: Analysis of Nigerian Demographic Health Survey 2018 dataset. *BMC Medical Informatics and Decision Making*, 24, Article 86. <https://doi.org/10.1186/s12911-024-02476-5>
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>
16. UNICEF. (2024). Delivery care. UNICEF Data. <https://data.unicef.org/topic/maternal-health/delivery-care/>
17. United Nations Inter-agency Group for Child Mortality Estimation. (2026). Levels & trends in child mortality: Report 2025. United Nations Children's Fund. <https://data.unicef.org/resources/levels-and-trends-in-child-mortality-2025/>
18. World Health Organization. (2024). Newborn mortality. <https://www.who.int/news-room/fact-sheets/detail/newborn-mortality>
19. World Health Organization, UNICEF, United Nations Population Fund, World Bank Group, & United Nations Population Division. (2023). Improving maternal and newborn health and survival and reducing stillbirth: Progress report 2023. World Health Organization. <https://www.who.int/publications/i/item/9789240073678>