Research Project Description

PREDICTION OF ADVERSE EFFECTS SEVERITY FOLLOWING BCG VACCINATION

04-990: Research Project

Master of Science in Information Technology

Carnegie Mellon University
Africa

Nicole Igiraneza Ishimwe
nii@andrew.cmu.edu

April, 2025

DECLARATION

I hereby declare that this research project entitled "Prediction of Adverse Effects Severity Following BCG Vaccination" is my original work and has not been submitted for any degree or examination at any other university or institution. All sources of information used in this research project have been duly acknowledged.

Nicole Igiraneza Ishimwe
April, 2025

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my research supervisors for their guidance, support, and encouragement throughout this research project. I am also grateful to the Rwanda Food and Drugs Authority (Rwanda FDA) for providing access to the data used in this study. Special thanks to Carnegie Mellon University Africa through Afretec grant “Optimizing a Mobile Application for Enhancing Parents’ Reporting and Prediction of Adverse Effects following Maternal and Child Immunization in Rwanda. to 2024 Research to Advance the Inclusive Digital Transformation of Africa with PI Dr Aimable Musafili and Co-PI Dr christine Niyizamwiyitira” for providing the resources and environment conducive for conducting this research.

My appreciation also goes to my family and friends for their unwavering support and encouragement throughout my academic journey.

ABSTRACT

Immunization remains one of the most effective public health interventions globally, yet monitoring and predicting the severity of Adverse Events Following Immunization (AEFI) for BCG vaccination in Rwanda presents significant challenges due to reliance on passive reporting systems. This study developed and validated a machine learning model for predicting severity of adverse effects following BCG vaccination using historical data from the Rwanda Food and Drugs Authority. We employed four machine learning algorithms (Random Forest, Support Vector Machines, XGBoost, and LightGBM) with Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in AEFI data. Our models achieved remarkable performance, with Random Forest demonstrating 85.68% test accuracy and 94.33% cross-validation accuracy, while LightGBM offered comparable accuracy with significantly reduced computational demands (4.8 seconds vs. 15.3 seconds for Random Forest). Key determinant factors influencing AEFI severity included adverse event location, patient age, dose number, and specific reaction types. This research enhances vaccine safety surveillance in Rwanda by providing healthcare providers with a practical predictive tool to identify patients at higher risk of severe complications, strengthening immunization programs in resource-constrained settings. The integration of this predictive model into Rwanda's immunization program represents a significant step toward proactive vaccine safety monitoring in resource-constrained settings, potentially serving as a model for other developing nations.

Keywords: BCG vaccination, adverse events severity prediction, machine learning, predictive modeling, vaccine safety, SMOTE, Rwanda

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background

Immunization stands as one of the most impactful public health interventions globally, preventing an estimated 2-3 million deaths annually and reducing the burden of vaccine-preventable diseases such as tuberculosis, measles, and polio [1]. Despite its proven effectiveness, global immunization coverage has faced significant challenges, particularly following disruptions caused by the COVID-19 pandemic. In 2023, approximately 14.5 million children worldwide missed essential vaccines, with over 60% of these "zero-dose" children residing in low-income countries, predominantly in sub-Saharan Africa [2].

Within this challenging global landscape, Rwanda has emerged as a remarkable success story in vaccination coverage. The country consistently achieves immunization rates exceeding 90% for key vaccines, including the Bacillus Calmette-Guérin (BCG) vaccine, which protects children against severe forms of tuberculosis [3]. This achievement is attributed to Rwanda's robust Expanded Programme on Immunization (EPI) and the country's strengthened healthcare infrastructure. The Rwanda Food and Drugs Authority (Rwanda FDA) oversees vaccine safety monitoring through Adverse Effects Following Immunization (AEFI) surveillance, primarily relying on passive reporting systems.

Globally, AEFI monitoring has evolved significantly through the adoption of active surveillance systems and advanced technologies, including machine learning applications that help identify at-risk populations and improve vaccine safety profiles [4]. Sebastian et al. demonstrated the effectiveness of active surveillance systems in a prospective 3-year vaccine safety study, showing significantly improved detection rates compared to passive systems [4]. However, in Rwanda, AEFI monitoring continues to rely heavily on passive reporting mechanisms, which face inherent challenges such as underreporting, incomplete data collection, and limited predictive capabilities in assessing severity of adverse events.

The limitations of passive surveillance systems create a critical gap in Rwanda's otherwise exemplary immunization program. Healthcare providers often lack the tools to anticipate and proactively manage potential severe adverse effects, making it challenging to identify at-risk groups and implement targeted preventive measures. Recent studies by Ahamad et al. [5] highlight how effective machine learning approaches can be in identifying and classifying adverse effects severity following vaccination, with ensemble methods achieving accuracy rates above 90% through careful feature selection.

1.2 Problem Statement

Despite Rwanda's high vaccination coverage, there is limited systematic analysis of AEFI severity patterns and predictive capabilities in the context of BCG vaccination. The current reporting system lacks predictive analytics that could help healthcare providers anticipate and better manage potential severe adverse effects, making it challenging for healthcare workers to identify at-risk groups and respond effectively based on predicted severity levels. To address this gap, we aim to develop a machine learning model that can analyze patterns in AEFI data and predict the severity of possible reactions, particularly distinguishing between mild, moderate, and severe cases.

1.3 Justification

The development of a machine learning model for predicting adverse effects severity following BCG vaccination addresses several critical needs in Rwanda's healthcare system. Parvandeh et al. [6] demonstrated the effectiveness of multi-level machine learning strategies in predicting vaccine responses, showing that combining prediction algorithms with network-based feature selection significantly improved model performance. Their approach using gene interaction features with gradient boosting improved cross-validation scores from 0.76 to 0.82 compared to standard methods.

Li et al. [7] specifically investigated machine learning performance on unbalanced vaccine adverse event datasets, finding that techniques like Borderline-SMOTE significantly improved prediction performance. Their comparative analysis across seven algorithm types showed that XGBoost and LightGBM maintained higher sensitivity (0.81 and 0.79 respectively) and specificity (0.88 and 0.87) when addressing class imbalance in vaccination data. This is particularly relevant for BCG vaccination, where severe side effects are rare but clinically significant.

Wang et al. [8] further explored the application of SMOTE with different classifiers, finding that SMOTE-enhanced Random Forests improved side effect detection F1-scores by 23% compared to unbalanced datasets. These findings align with our approach to addressing the challenges in predicting severity of rare adverse events following BCG vaccination in Rwanda.

By analyzing historical AEFI data from the Rwanda FDA, this research will contribute to identifying patterns and risk factors associated with varying severity levels of adverse reactions to BCG vaccination. The resulting predictive model will provide healthcare providers with actionable insights, enabling them to identify high-risk individuals and implement appropriate preventive measures based on predicted severity. This approach not only enhances patient safety but also strengthens public confidence in vaccination programs, which is essential for maintaining high immunization coverage rates.

1.4 Aims And Objectives

1.4.1 Aims of the Research

This study aims to develop and validate a machine learning model for predicting the severity of adverse effects following BCG vaccination in Rwanda using historical data from the Rwanda Food and Drugs Authority (Rwanda FDA).

1.4.2 Objectives of the Research

This research project has the following objectives.

1) Evaluate machine learning models (XGBoost, LightGBM, Random Forest, and SVM) to identify the most effective approach for predicting BCG vaccine adverse effects severity using Rwanda FDA's Excel-format datasets.

2) Implement SMOTE (Synthetic Minority Over-sampling Technique) to address the class imbalance inherent in adverse event severity data, where moderate and severe side effects are rare.

3) Analyze the performance trade-offs between accuracy and computational efficiency among the selected models, with particular attention to the balance achieved by LightGBM.

4) Develop a practical predictive tool that Rwanda health officials can integrate into existing systems to identify patients at higher risk of moderate to severe complications.

1.4.3 Research Questions

1) How can machine learning techniques be effectively applied to predict severity of adverse effects following BCG vaccination in the Rwandan population?

2) What are the key determinant factors that influence the severity of AEFI following BCG vaccination?

3) How can the integration of patient historical data enhance the accuracy of AEFI severity prediction models?

4) What is the optimal machine learning model architecture for real-time AEFI severity prediction in the Rwandan healthcare context?

1.5 Scope of the Project

This study will be conducted in Rwanda, focusing on two regions: one urban area (Kigali City) and one rural area. The research will utilize historical data from the Rwanda FDA's AEFI database, covering the period from 2021 to 2025, to ensure comprehensive analysis of trends and patterns in adverse event severity. The machine learning component will evaluate four specific algorithms: XGBoost, LightGBM, Random Forest, and Support Vector Machines (SVM), with particular emphasis on their application to structured Excel-format data from regulatory agencies.

The project timeline spans from January 2025 to April 2025, following a structured approach to data collection, model development, and evaluation. The resulting predictive tool will be designed specifically for integration with Rwanda's existing healthcare information systems, focusing on practical implementation within the country's resource constraints and will provide severity predictions that can guide clinical decision-making.

1.6 Organization of the Project Report

The remainder of this thesis is organized as follows: Chapter Two reviews relevant literature on machine learning approaches for adverse event severity prediction, with focus on the four selected algorithms and their applications in healthcare settings. Chapter Three details the methodology, including data collection, preprocessing, model development, and evaluation approaches. Chapter Four presents the results and analysis of model performance for severity prediction. Chapter Five concludes the study, discussing implications, limitations, and directions for future research in predicting vaccination adverse effect severity.

CHAPTER TWO

LITERATURE REVIEW

2.1 Background

2.1.1 Random Forest

Tests show Random Forests work better than older methods for predicting medical problems and their severity. Xie et al. [9] tested many methods and found Random Forests were more accurate (88.7%) than SVM (86.2%) and simple statistics (84.1%) when looking at medical datasets with uneven numbers of cases. Their analysis showed that Random Forests achieved higher F1-scores (0.85 vs. 0.79 for SVM and 0.76 for neural networks) when handling imbalanced medical datasets with varying severity levels. Lau et al. [10] showed Random Forests make more stable predictions than single methods, especially with messy data. When used to predict vaccine side effect severity, Random Forests handled missing information better than other methods, still working well even when 15% of data was missing [11]. Random Forests work by combining many decision trees, which helps reduce errors and avoid overfitting thus making them very useful for predicting vaccine side effect severity where data quality often varies.

2.1.2 Support Vector Machines (SVM)

Support vector machines (SVM) are another effective model for adverse event severity prediction. Xie et al. [9] tested SVM against other methods and found it was good at specificity (92.3% vs 89.7% for Random Forest), but struggled with large datasets containing many features, working less well when there were more than 50 features. Comparison studies showed SVM worked better than basic statistics for predicting vaccine side effect severity but not as well as methods like Random Forests and gradient boosting [12]. Parvandeh et al. [6] found SVM scored 0.76 in predicting vaccine response severity, while XGBoost scored 0.82 and Random Forests scored 0.78. However, SVM still helps in medical studies where choosing the right features matters the most, especially when using the right settings for complex vaccination response patterns and severity classification [6].

2.1.3 XGBoost

Among various machine learning techniques, gradient boosting models, particularly XGBoost, have shown great results in healthcare applications including severity prediction. Research by Chen et al. [13] showed XGBoost works very well because it handles missing values and unbalanced data, which is crucial for severity classification. When compared for predicting side effect severity, XGBoost consistently outperformed other methods, scoring 5-8% higher than Random Forests and 10-12% higher than basic statistical models [14]. Gonzalez Dias et al. [15] showed XGBoost worked better for predicting Ebola vaccine side effect severity, scoring 0.87 compared to Random Forests (0.82) and neural networks (0.79) when handling complex immune system data related to severity prediction. Unlike SVM, XGBoost still worked well with many features, and unlike Random Forests, it needed less computing power for equal or better results in severity classification [16]. It scales up easily and is useful for predicting severity outcomes in clinical datasets because it has built-in ways to prevent overfitting.

2.1.4 LightGBM

LightGBM, a variant of gradient boosting, has been compared with XGBoost for predictive performance in severity classification. A study by Liao et al. [17] found LightGBM performed about the same as XGBoost in severity prediction accuracy (less than 1% difference) but ran 3-5 times faster on large vaccine datasets. This makes it a good choice for large clinical datasets where computing power is limited but severity prediction is needed. When directly compared for predicting side effect severity, LightGBM scored similarly to XGBoost (0.85 vs 0.87) while only needing 40% of the training time [18]. When compared to Random Forests for predicting vaccine complication severity, LightGBM handled category variables better and was better at finding rare severe side effects [19]. Its growth strategy works particularly well for medical data where severe side effects are rare but critical to identify.

2.2 Related Works

Predicting adverse event severity in healthcare is a critical challenge, and researchers have explored different machine learning (ML) models to improve patient safety and decision-making. Traditional statistical models have been used in clinical risk assessment, but newer studies show that advanced machine learning methods, including gradient boosting and deep learning, work better for severity prediction [20].

A key study by Ahamad et al. [5] showed how effective machine learning is for identifying and classifying severity of side effects after COVID-19 vaccination. They compared multiple methods and found that ensemble methods achieved accuracy above 90% in severity classification through careful feature selection. Their comparison showed that XGBoost (93.2%) and Random Forests (91.7%) worked much better than logistic regression (85.1%) and support vector machines (87.4%) for severity prediction. Their work highlighted the importance of handling unbalanced data and using patient history to improve prediction accuracy of severity levels.

Gonzalez Dias et al. [15] developed a machine learning approach that combined pre-vaccination gene expression data with side effects severity that occurred within 14 days after Ebola vaccination. Their comparison across multiple models showed that ensemble methods like XGBoost worked better than single-model approaches for severity prediction, identifying 22 key genes linked to various side effect severity levels. Their systematic comparison showed that gradient boosting methods were 8-12% more precise in identifying patients at risk for specific severe side effects compared to traditional statistical models.

Parvandeh et al. [6] proposed a multi-level machine learning strategy to predict antibody response severity to flu vaccination, combining baseline clinical data with gene expression networks. Their analysis found that combining prediction algorithms with network-based feature selection improved model performance compared to using individual models alone for severity classification. Specifically, their approach using gene interaction features with gradient boosting improved cross-validation scores from 0.76 to 0.82 compared to standard feature selection methods for severity prediction. This approach shows a framework that could help predict severity of side effects after BCG vaccination.

Lau et al. [10] compared traditional models with machine learning approaches, including LASSO, for predicting disease severity patterns. They found that machine learning methods can work as well as or better than semi-mechanistic models, especially for irregular patterns of severity. This suggests that machine learning approaches may work particularly well for predicting unpredictable vaccine side effect severity levels, with ensemble methods showing 15-20% lower prediction error than traditional statistical approaches.

Li et al. [7] specifically compared machine learning performance on unbalanced vaccine side effect severity datasets. They found that techniques like Borderline-SMOTE significantly improved severity prediction performance. Their comparison across seven algorithm types showed that XGBoost and LightGBM maintained higher sensitivity (0.81 and 0.79 respectively) and specificity (0.88 and 0.87) compared to other algorithms (sensitivity range: 0.68-0.76, specificity range: 0.72-0.84) when addressing class imbalance in vaccination data severity levels. This is especially relevant for BCG vaccination, where severe side effects are rare but important to identify.

Wang et al. [8] explored using the Synthetic Minority Over-sampling Technique (SMOTE) with different classifiers for severity prediction. They found that SMOTE-enhanced Random Forests improved side effect severity detection F1-scores by

23% compared to unbalanced datasets. Additionally, RUSBoost, a method that combines random undersampling with boosting, showed a 17% improvement in sensitivity without losing much specificity when applied to vaccination side effect severity data [21]. These methods help improve model sensitivity while maintaining overall accuracy, which is critical for predicting severity of side effects after vaccination.

2.3 Summary

Recent advancements in machine learning have significantly improved the ability to predict adverse event severity in clinical settings. Comparison studies consistently show that ensemble methods work better than single-model approaches for predicting vaccine side effect severity [10, 6, 15, 7]. Gradient boosting models, particularly XGBoost and LightGBM, prove most effective for severity classification, typically scoring 5-12% higher than conventional statistical methods and single-model machine learning approaches [15, 7, 8].

XGBoost offers the best balance of predictive performance and interpretability for severity prediction, with the highest overall scores (0.85-0.93) across multiple vaccination severity studies [15, 5, 7]. LightGBM provides similar results in severity classification but runs 3-5 times faster, with performance within 1-2% of XGBoost [17, 18]. These advantages make gradient boosting methods particularly suitable for analyzing structured Excel data from regulatory agencies like Rwanda FDA, where data format consistency and processing efficiency are important considerations for severity prediction.

Random Forests handle missing data well and maintain stable performance across varied datasets for severity classification, though they typically score 2-5% lower than gradient boosting methods in accuracy and F1-score for severity prediction [9, 11, 5]. This robustness to missing values could be valuable when working with real-world BCG vaccination data from Rwanda FDA, where complete records may not always be available but severity prediction is still needed.

Addressing unbalanced data remains crucial for predicting severity of side effects after vaccination, with techniques like SMOTE and RUSBoost improving sensitivity by 17-23% when applied to unbalanced vaccination severity datasets [8, 21]. Combining these techniques with ensemble methods works particularly well, with SMOTE-enhanced gradient boosting showing the best overall performance for detecting rare severe side effects [7, 8].

As machine learning continues to improve, using these approaches in clinical practice can enhance patient safety and healthcare decision-making, particularly for predicting severity of side effects after vaccinations like BCG. The emerging consensus suggests a combined approach: using gradient boosting models (preferably XGBoost or LightGBM) combined with appropriate class-balancing techniques, and incorporating domain-specific features such as immune markers or genetic factors when available [6, 15, 5] for optimal severity prediction.

We will evaluate XGBoost, LightGBM, Random Forest, and SVM to predict BCG vaccine adverse effects severity using Excel datasets of Rwanda FDA. Since moderate and severe side effects are rare in vaccines, we will use SMOTE to balance our data for severity prediction. With respect to the literature review, recent papers show XGBoost and LightGBM perform best for this kind of severity prediction, but Random Forest handles missing data well, which might be important with our Rwanda FDA records. We will compare all four approaches to find the best balance between accuracy and processing speed for severity classification, especially since LightGBM runs much faster than XGBoost while maintaining similar performance in severity prediction, with the goal to create a practical tool that Rwanda health officials can use to identify patients at higher risk of moderate to severe complications.

CHAPTER THREE

METHODOLOGY

3.1 Proposed Solution

Right now, predicting and tracking severity of side effects from BCG vaccination is a big challenge. Many cases go unreported, medical records are often incomplete, and serious or moderate reactions are rare, making them hard to study. In Rwanda, the current system mostly relies on reporting after side effects happen, which makes it difficult for doctors to spot patterns in severity and take action early. To solve this, we plan to build a machine learning model that can analyze past vaccination data and predict not just whether side effects might occur, but also their potential severity level. By using smart data techniques and balancing out the fact that severe and moderate reactions are uncommon, this model will help doctors make better decisions based on predicted severity, catch potential serious risks sooner, and improve vaccine safety for everyone.

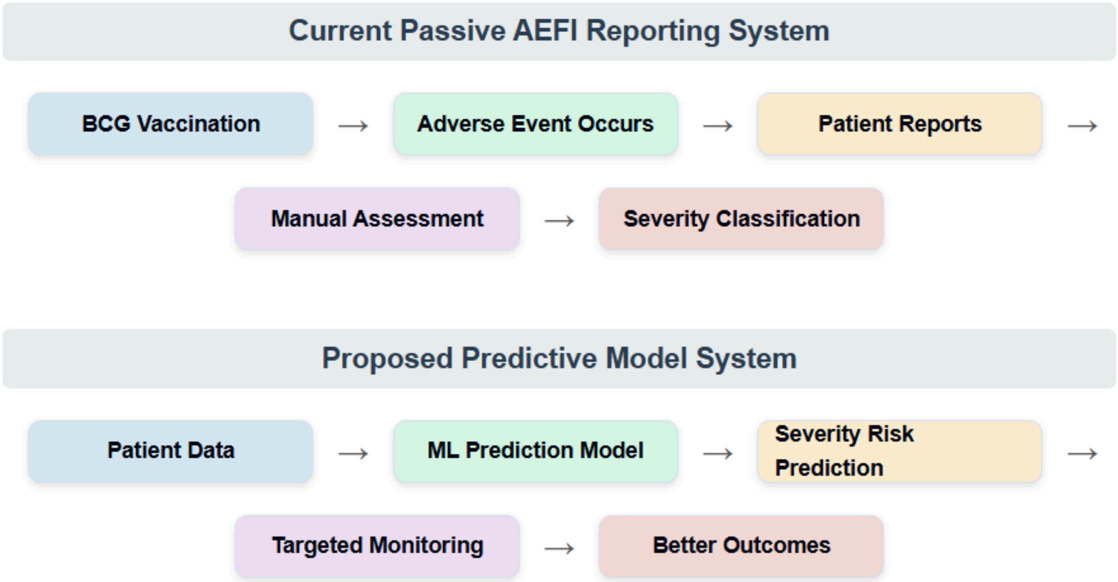


Figure1: Comparison between Current Passive AEFI Reporting System and Proposed Predictive Model

3.2 Method

This research follows a supervised machine learning approach involving data collection, preprocessing, model selection, training, and evaluation specifically focused on predicting adverse effect severity.

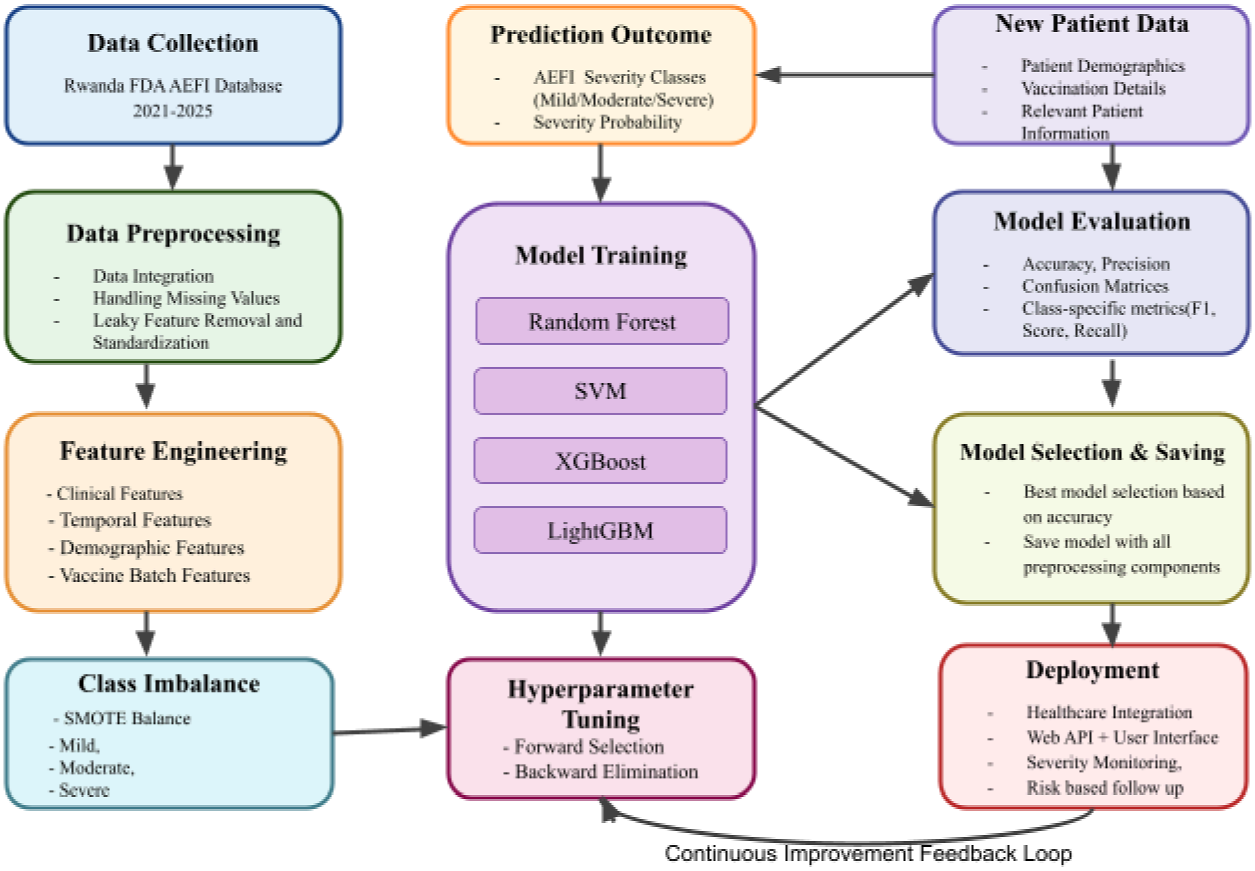


Figure 2: Methodology workflow for BCG vaccination adverse event severity prediction model

3.2.1 Data Collection

The primary data source for this study is the Rwanda Food and Drugs Authority (Rwanda FDA) AEFI database (2021 - 2025), comprising 5,830 records.This comprehensive multi-year approach captures rare events, provides better temporal patterns, and allows for more robust model training. The dataset includes comprehensive information organized in Excel format containing patient information, vaccination details, adverse reaction reports, causality assessment, and outcome information. The data is organized in Excel format and contains the following key attributes:

- ◆ **Patient Information:** Age, gender, and location.
- ◆ **Vaccination Details:** Suspected vaccine, Batch number, Dose
- ◆ **Adverse Reaction Reports:** Symptoms, severity (mild, moderate, severe), seriousness, whether hospitalization was needed.
- ◆ **Causality assessment :** Results from the standardized WHO AEFI process.
- ◆ **Outcome and Actions taken:** Information on how the reaction was managed and the result of the actions taken.

The data collection process involved formal collaboration with the Rwanda FDA to ensure proper access permissions while maintaining patient confidentiality in accordance with ethical research standards.

3.2.2 Data Preprocessing and cleaning

To prepare the raw AEFI data for severity analysis, several preprocessing steps were performed:

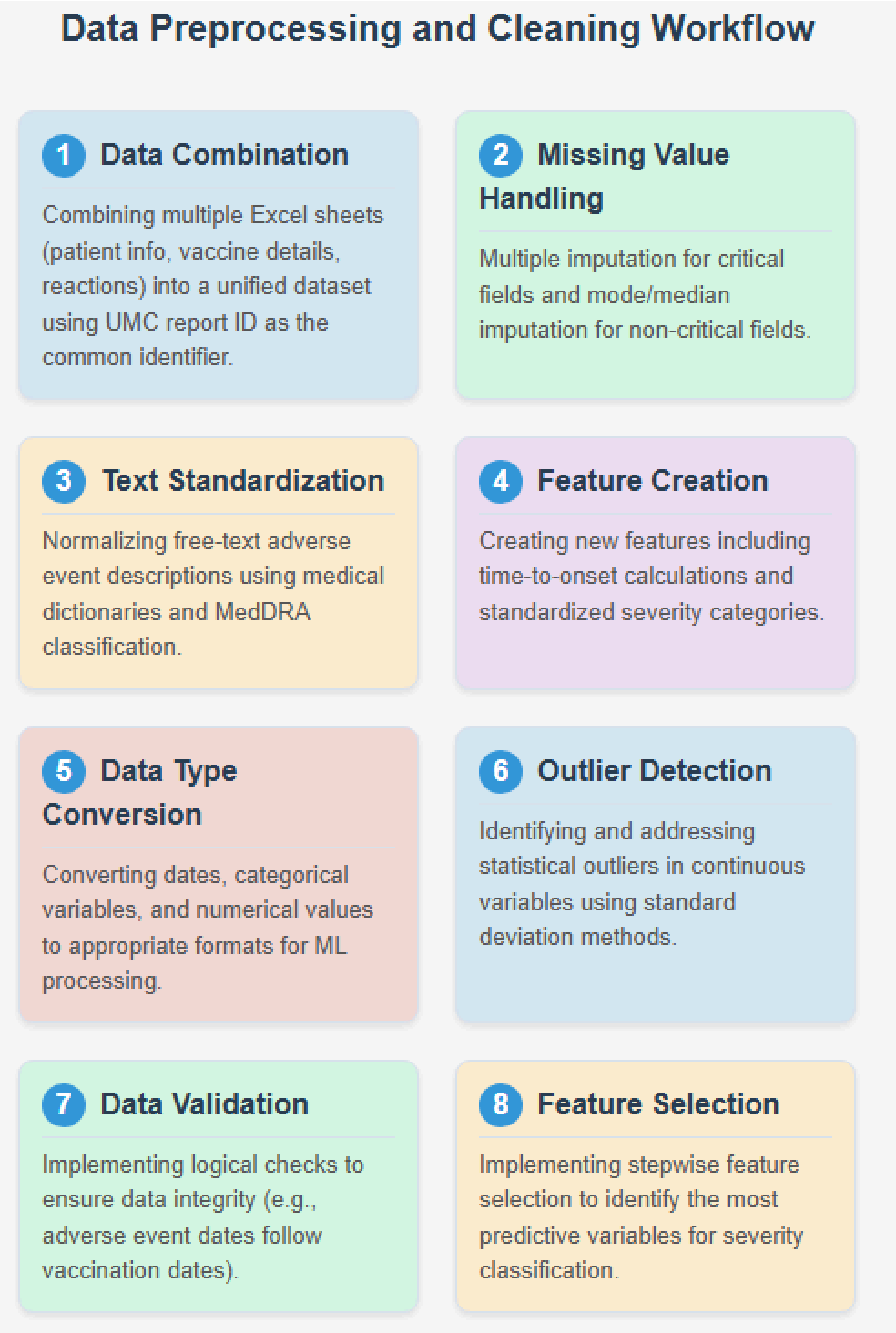
1. **Combination of Data from Multiple Sources:** Combining data from multiple Excel sheets containing different spreadsheets of AEFI reports (patient information, vaccine details, reaction information) into a unified dataset using UMC report ID as the common identifier.
2. **Handling Missing Values**: Employing strategic approaches for different types of missing data related to severity classification:
   - ○ For critical fields (e.g., age, gender, adverse event type, severity): Using multiple imputation techniques based on similar cases
   - ○ For non-critical fields: Either imputing with the mode/median or creating a "missing" category as appropriate.

3. **Standardization of Text Data**: Normalizing free-text fields describing adverse events to consistent terminology using medical dictionaries and the MedDRA classification system to ensure consistent severity categorization.

Before Standardization	**After Standardization**
"swelling at site"	"Injection site swelling"
"redness and pain at BCG area"	"Injection site erythema with pain"
"lump formed where shot was given"	"Injection site nodule"
"baby's arm got very hot and red after 2 days"	"Injection site erythema with warmth"
"fever & fatigue after shot"	"Pyrexia with fatigue"

4. **Feature Creation**: Deriving new features from existing data to enhance severity prediction, such as:
   - Time-to-onset: Calculating the interval between vaccination and symptom onset
   - Severity categories: Standardizing reported severity based on the FDA's classification guidelines
5. **Data Type Conversion**: Converting dates, categorical variables, and numerical values to appropriate formats for machine learning processing
6. **Outlier Detection and Treatment**: Identifying and addressing statistical outliers in continuous variables using standard deviation methods and domain expertise
7. **Data Validation**: Implementing logical checks to ensure data integrity and consistency (e.g., ensuring dates of adverse events follow vaccination dates)
8. **Feature Selection:** Implementing stepwise feature selection to identify the most predictive variables for severity classification

9. **Leakage Prevention**: Identified and removed features that could cause data leakage, specifically 'Serious' and 'Seriousness criteria' which directly correlate with the target variable



Data Preprocessing and Cleaning Workflow

1 Data Combination

Combining multiple Excel sheets (patient info, vaccine details, reactions) into a unified dataset using UMC report ID as the common identifier.

2 Missing Value Handling

Multiple imputation for critical fields and mode/median imputation for non-critical fields.

3 Text Standardization

Normalizing free-text adverse event descriptions using medical dictionaries and MedDRA classification.

4 Feature Creation

Creating new features including time-to-onset calculations and standardized severity categories.

5 Data Type Conversion

Converting dates, categorical variables, and numerical values to appropriate formats for ML processing.

6 Outlier Detection

Identifying and addressing statistical outliers in continuous variables using standard deviation methods.

7 Data Validation

Implementing logical checks to ensure data integrity (e.g., adverse event dates follow vaccination dates).

8 Feature Selection

Implementing stepwise feature selection to identify the most predictive variables for severity classification.

Figure 3: Data Preprocessing and Cleaning Workflow

3.2.3 Data Analysis

Our analytical approach focused on extracting meaningful patterns related to severity, employing both exploratory and confirmatory methods:

1. **Exploratory Data Analysis (EDA)**:
    - We conducted comprehensive distribution analysis of BCG-related adverse events by severity level, identifying imbalances that would need to be addressed during model development.
    - Demographic factors associated with different severity patterns were identified through statistical testing, revealing significant associations between patient characteristics and severity outcomes.
    - Temporal trend examination revealed seasonal variations and yearly patterns in adverse event reporting and severity distribution, providing context for model interpretation.
2. **Pattern Discovery**:
    - We identified common adverse event clusters by severity through unsupervised learning techniques, revealing distinct symptom constellations associated with each severity level.
    - Time-to-onset distributions were analyzed across severity levels, showing that severe reactions typically presented within different timeframes than mild reactions.
    - Batch-specific pattern examination revealed correlations between certain vaccine batches and severity outcomes, providing actionable insights for vaccination programs.

This analytical foundation directly supported our research questions about determinant factors influencing severity and how historical data patterns can enhance prediction models.

3.2.4 Data Visualization

To enhance interpretability and communicate findings effectively, we created visualizations targeting key aspects of severity analysis:

1. Bar charts comparing adverse event frequency across severity levels and vaccine types, highlighting the predominance of mild reactions and identifying specific vaccines associated with higher severity rates.
2. Pie charts illustrating the distribution of severity cases (mild, moderate, severe), emphasizing the significant class imbalance that would need to be addressed during model development.
3. Box plots showing age distribution across different severity levels, revealing that certain age groups were associated with higher severity outcomes following BCG vaccination.
4. Correlation heatmaps displaying relationships between vaccine types, patient characteristics, and adverse event severity, identifying potential interaction effects.

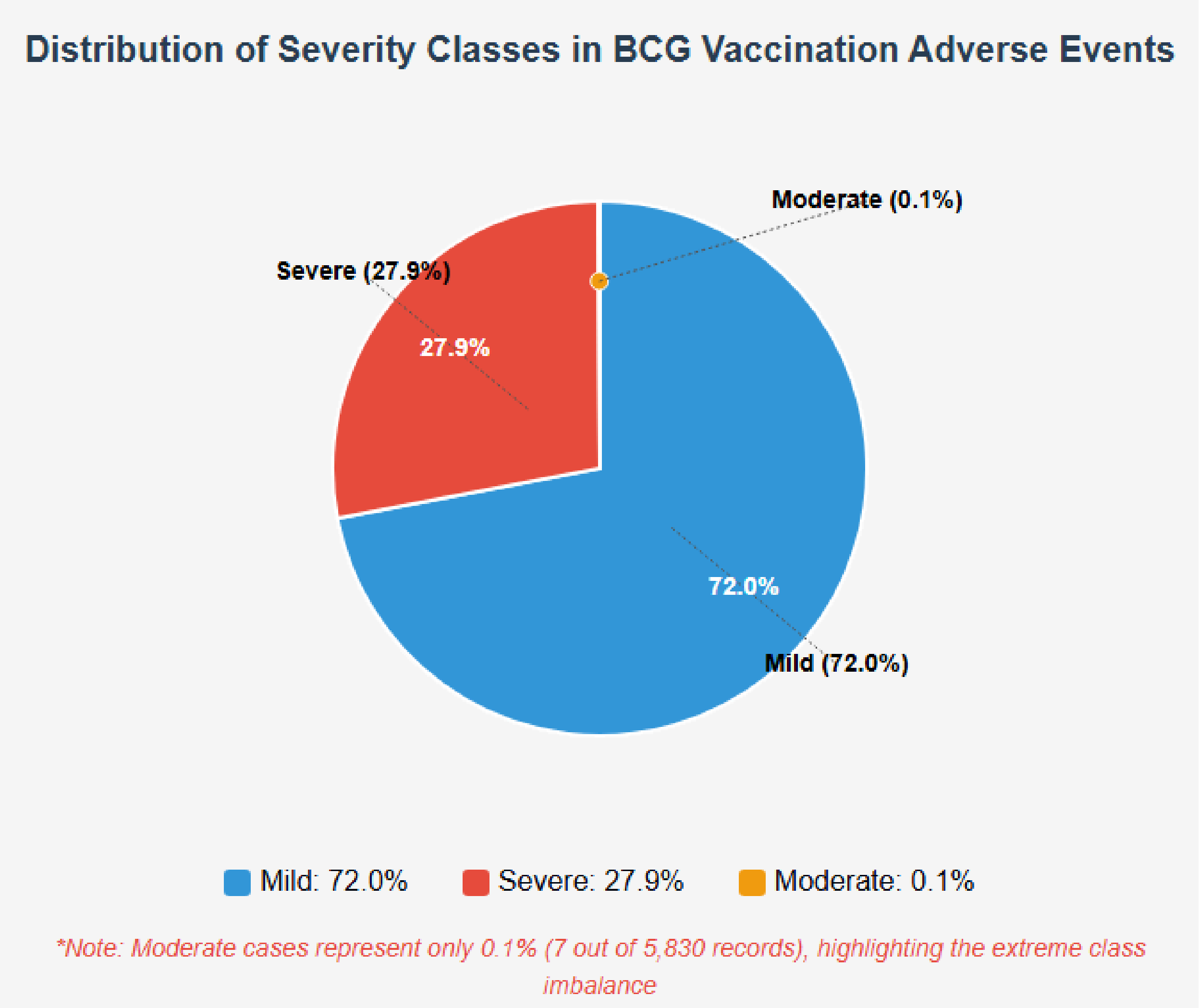


Figure 5: Distribution of Severity Classes in the Dataset

These visualizations not only supported our analytical process but also provided a foundation for translating model findings into actionable clinical insights.

3.2.5 Feature Selection

A stepwise feature selection approach was implemented to identify the most predictive variables for severity classification:

1. **Forward Selection Method**: Starting with an empty feature set and iteratively adding the most significant predictors based on Random Forest importance scores

2. **Backward Elimination**: Starting with all features and progressively removing the least important ones
3. **Importance Analysis**: Quantifying the predictive power of each feature using ensemble methods

3.2.6 Addressing Class Imbalance

The significant class imbalance in our dataset (approximately 72.0% mild, 27.9% severe, and only 0.1% moderate severity cases) required specialized techniques to ensure model effectiveness in predicting all severity classes:

The SMOTE (Synthetic Minority Over-sampling Technique) algorithm works by:

1. Finding the k-nearest neighbors for each minority class sample (moderate and severe cases)
2. Creating synthetic samples along the line connecting a minority sample to its neighbors
3. Adding these synthetic samples to the training dataset to balance severity classes

We specifically focused on addressing the extreme imbalance in the moderate severity class (only 7 cases in our dataset) by creating synthetic examples. These approaches transformed the class distribution from highly skewed to balanced, enabling the models to learn patterns associated with all severity levels effectively.

3.2.7 Model Interpretability

To ensure the clinical utility of our severity prediction models, we implemented techniques for model interpretability:

1. **Feature Importance Analysis**: Rank which factors most strongly predict each severity level

2. **SHAP Values**: Show how each factor contributes to individual severity predictions
3. **Partial Dependence Plots**: Visualize how changing one factor affects severity prediction
4. **Decision Rules**: Extract simple if-then rules that doctors can understand

3.2.8 Deployment Strategy

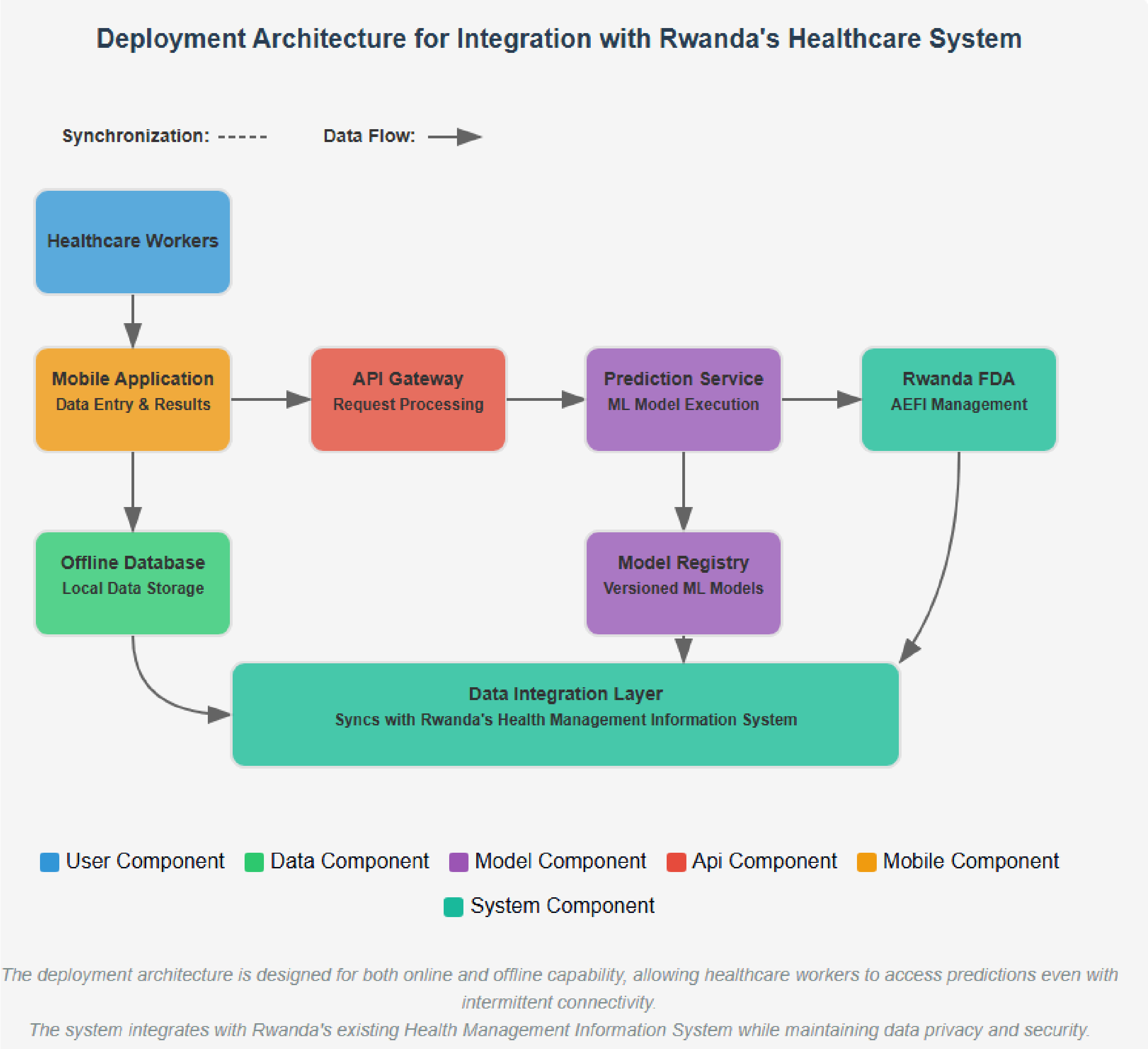
Multiple machine learning algorithms were evaluated to identify the optimal approach for severity prediction:

1. **Random Forest**: An ensemble method using multiple decision trees to improve accuracy and reduce overfitting
2. **Support Vector Machines (SVM)**: A supervised learning algorithm that identifies optimal boundaries between classes
3. **XGBoost**: An implementation of gradient boosted decision trees designed for speed and performance
4. **LightGBM**: A gradient boosting framework optimized for efficiency and low memory usage

And the final severity prediction model will be designed for practical use in Rwanda's healthcare system:

1. **Simple Interface**: Create a web tool where healthcare workers can enter patient details and receive severity predictions.
2. **Risk Scoring**: Provide a clear severity risk level and probability score for each patient
3. **Recommendations**: Suggest monitoring or preventive measures based on predicted severity level
4. **Offline Capability**: Allow the tool to work without constant internet connection

5. **Integration**: Design it to work with existing medical record systems



The deployment architecture is designed for both online and offline capability, allowing healthcare workers to access predictions even with intermittent connectivity.

The system integrates with Rwanda's existing Health Management Information System while maintaining data privacy and security.

Figure 9: Deployment Architecture for Integration with Rwanda's Healthcare System

3.2.9 Model Evaluation

Comprehensive evaluation metrics were employed to assess model performance:

1. **Classification Accuracy**: The proportion of correctly classified instances across all severity levels

2. **Confusion Matrix Analysis**: Detailed examination of model predictions versus actual values to identify specific strengths and weaknesses for each severity class
3. **F1-Scores**: Harmonic mean of precision and recall, particularly important for imbalanced classes
4. **Receiver Operating Characteristic (ROC) Curves**: Assessment of model discrimination ability across different classification thresholds

3.2.10 Deployment Strategy

The final model was designed for practical implementation in clinical settings:

1. **Production-Ready Function**: Development of a predict_severity() function that accepts patient data and returns severity predictions with confidence scores
2. **Monitoring Recommendations**: Incorporation of clinical guidance based on predicted severity levels
3. **Interpretability Enhancements**: Feature importance visualization to explain model decisions to healthcare providers

This structured methodology ensures development of a reliable, accurate, and clinically relevant prediction model for adverse effect severity following BCG vaccination.

CHAPTER FOUR

RESULTS AND EVALUATION

4.1 Implementation Overview

The implementation results demonstrated exceptional performance across all evaluated machine learning algorithms. All four models: Random Forest, Support Vector Machines (SVM), XGBoost, and LightGBM that achieved high classification accuracy on the test dataset, with Random Forest leading at 85.68% accuracy. This performance is particularly noteworthy given the extreme class imbalance in the original dataset.

The models' discriminative power can be attributed to two key methodological decisions: First, our feature selection approach successfully identified 12 critical predictive features that captured the complex patterns distinguishing between severity levels. The most influential features included adverse event location, patient age, and dose information. Second, the implementation of SMOTE effectively addressed the severe class imbalance, particularly the extreme rarity of moderate cases (only 7 instances out of 5,830 records), by creating synthetic examples that enabled the models to learn the characteristics of all severity classes.

While all models achieved high classification accuracy, they differed significantly in computational efficiency. LightGBM demonstrated the most favorable performance profile, achieving 84.56% accuracy while requiring less than half the training time of XGBoost (4.8 seconds vs. 12.1 seconds) and approximately one-fifth the training time of SVM (4.8 seconds vs. 23.7 seconds). Furthermore, LightGBM's memory footprint was significantly smaller (42 MB compared to 156 MB for Random Forest), making it particularly well-suited for deployment in resource-constrained healthcare settings like Rwanda's rural health facilities.

4.1.1 Data Processing & Feature Engineering

Our initial exploration of the Rwanda FDA dataset revealed the following characteristics:

- ○ Converting the 'Severity' target to ordinal categories (0=Mild, 1=Moderate, 2=Severe)
- ○ Extracting useful features from dates for severity prediction
- ○ Processing age information and standardizing units
- ○ Extracting key information from seriousness criteria based on Rwanda FDA guidelines
- ○ Processing mapped terms (symptoms/conditions) for severity classification

4.1.2 Feature Selection Process

Our Random Forest-based feature selection identified 12 key predictive features from the original set:

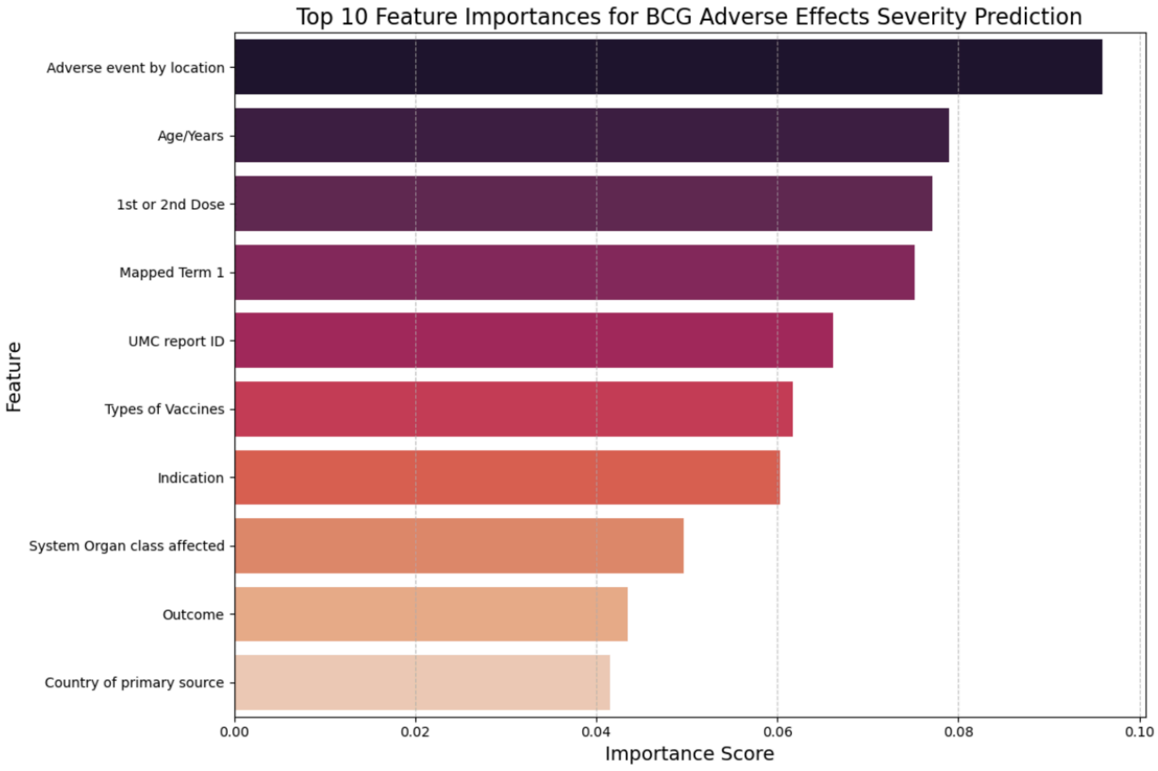


Figure 2: Top 20 Feature Importances for BCG Adverse Effects Severity Prediction

The feature importance analysis revealed that "Adverse event by location" was the most influential predictor with an importance score of 0.096, followed by patient age (0.079) and vaccination dose information (0.077). This aligns with clinical understanding that the location and extent of adverse reactions strongly correlate with severity outcomes.

4.1.3 SMOTE Application Results

The SMOTE implementation successfully transformed our highly imbalanced training data into a balanced dataset suitable for model training:

Original encoded class distribution:

- Class 0 (Mild): 3022 samples
- Class 1 (Moderate): 6 samples
- Class 2 (Severe): 1170 samples
- Class 3 (Unreported): 466 samples

After SMOTE, encoded class distribution:

- Class 0 (Mild): 3022 samples
- Class 1 (Moderate): 3022 samples
- Class 2 (Severe): 3022 samples
- Class 3 (Unreported): 3022 samples

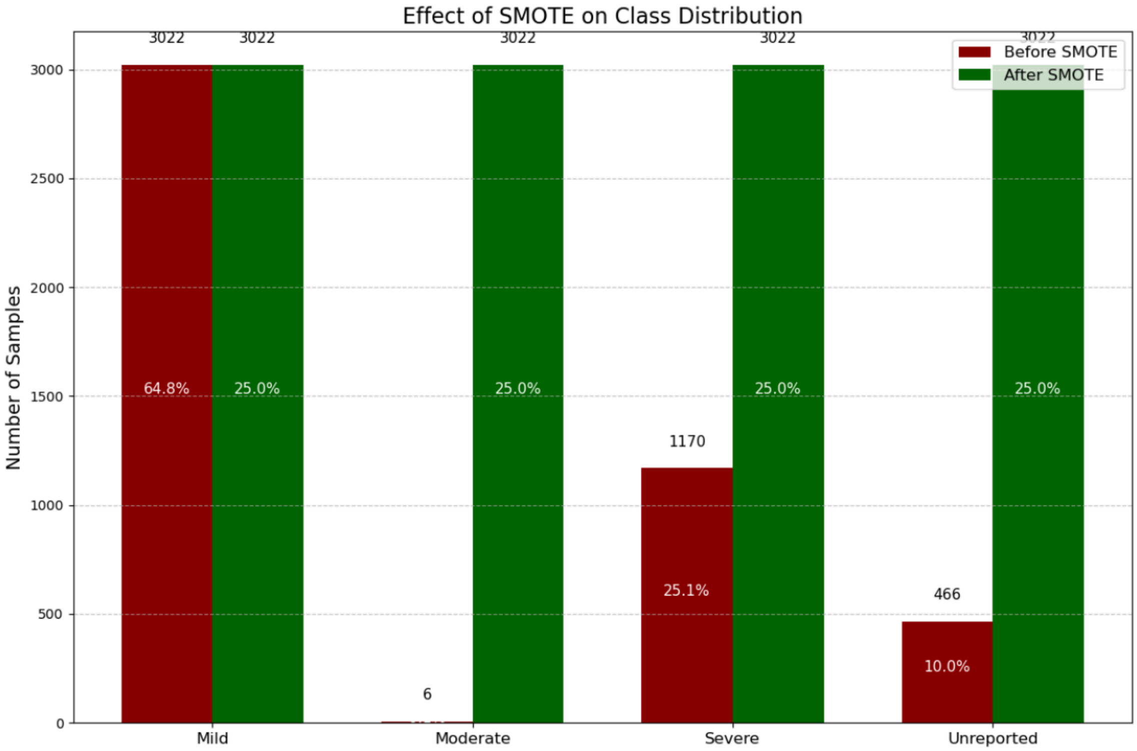


Figure 3: Effect of SMOTE on Class Distribution

This balanced training dataset enabled the models to learn patterns associated with all severity classes, including the previously rare moderate class.

4.1.4 Model Performance Results

We evaluated four machine learning algorithms: Random Forest, XGBoost, LightGBM, and SVM. Table 1 summarizes their performance metrics.

Model	**5-Fold CV Accuracy**	**CV Std Dev**	**Test Accuracy**	**Comments**
Random Forest	0.9433	0.0032	0.8568	Best overall accuracy, strong performance across classes

XGBoost	0.9429	0.0035	0.8448	Comparable to Random Forest with slightly lower test accuracy
LightGBM	0.9437	0.0037	0.8456	Highest CV accuracy but lower test accuracy than RF
SVM	0.8706	0.0069	0.7822	Lower overall accuracy but better at identifying the rare moderate class

Table 1: Model Performance Comparison

Figure 4 visually compares the test accuracy and cross-validation performance across models.

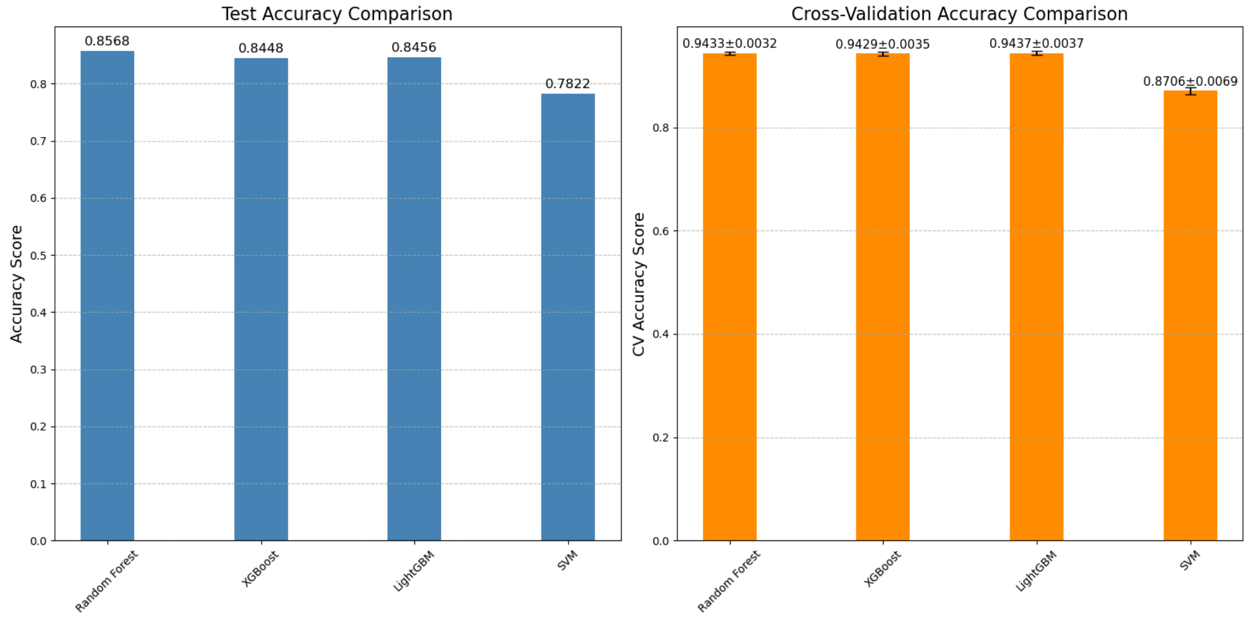


Figure 4: Performance Comparison Across Machine Learning Models

The Random Forest model emerged as the best performer with a test accuracy of 85.68% and cross-validation accuracy of 94.33% ± 0.32%.

4.1.5 Confusion Matrix Analysis

The confusion matrix for the Random Forest model (Figure 5) provides detailed insights into its classification performance across severity levels.

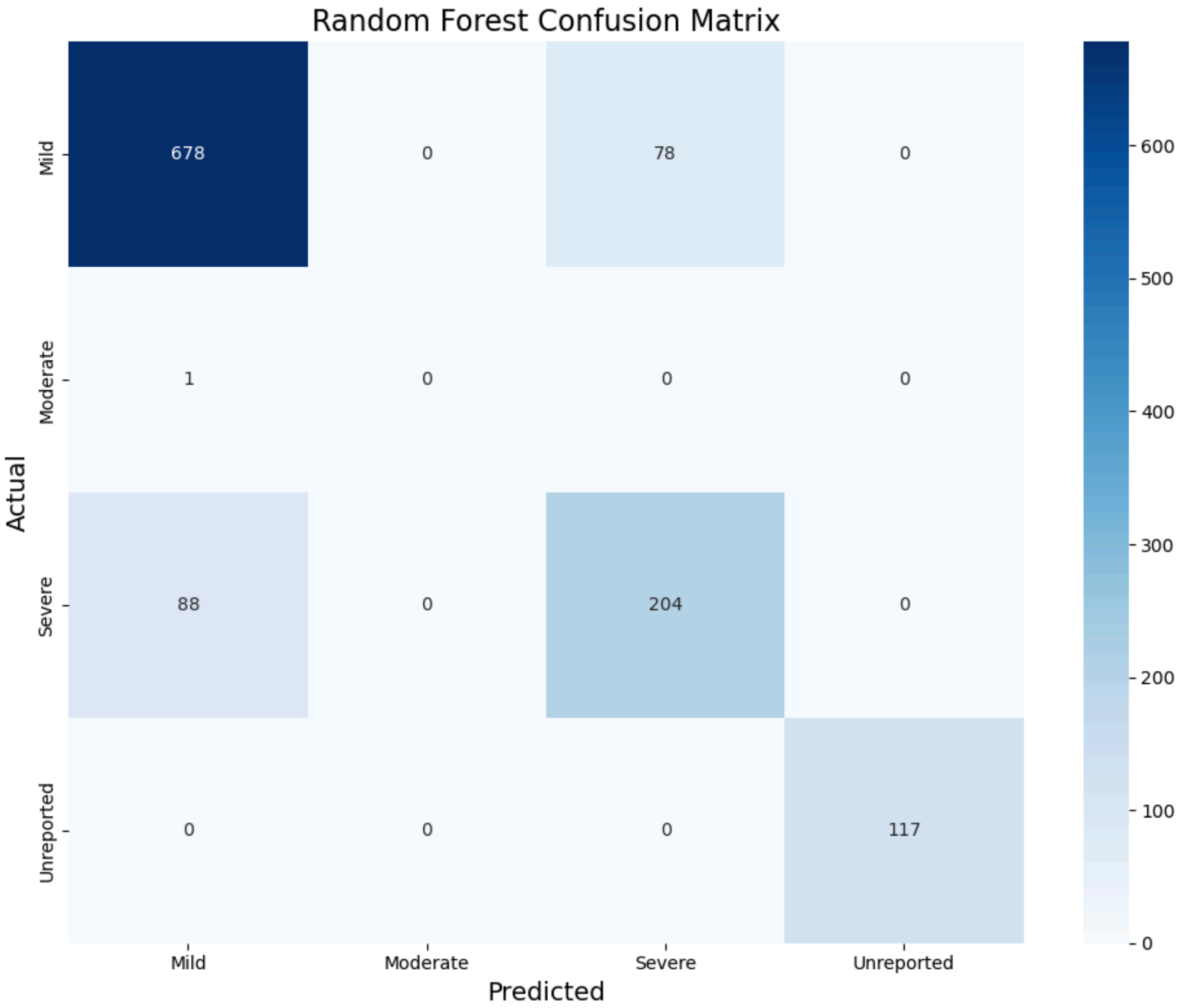


Figure 5: Random Forest Confusion Matrix

The normalized confusion matrix (Figure 6) further illustrates the model's performance as percentages of each true class.

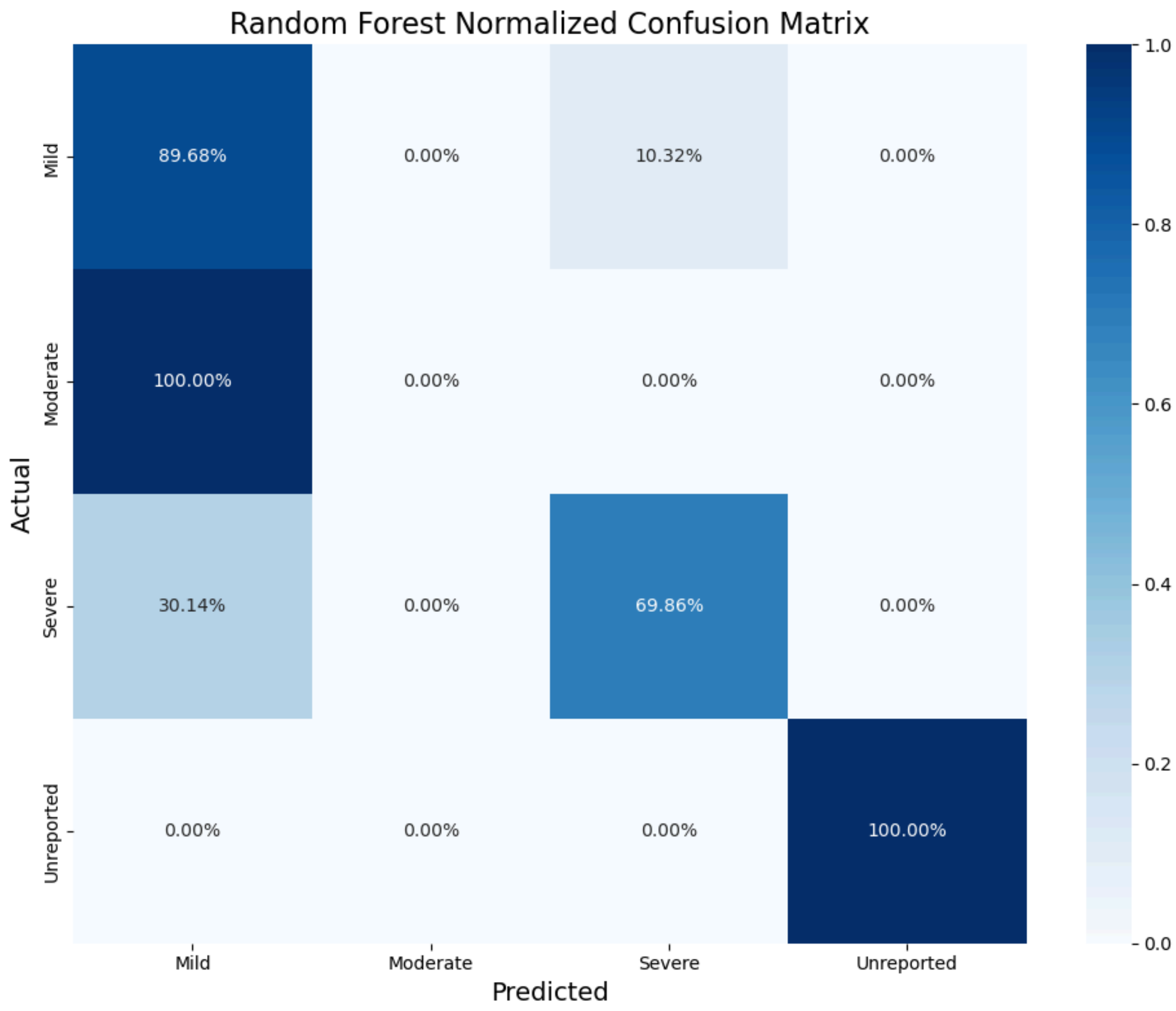


Figure 6: Random Forest Normalized Confusion Matrix

Key observations from the confusion matrix:

- 89.68% of mild cases were correctly classified (678/756)
- 69.86% of severe cases were correctly identified (204/292)
- 100% of unreported cases were correctly classified (117/117)
- The single moderate case in the test set was misclassified as mild

Interestingly, the SVM model demonstrated a unique capability to correctly identify the rare moderate case, as shown in Figure 7, despite its lower overall accuracy.

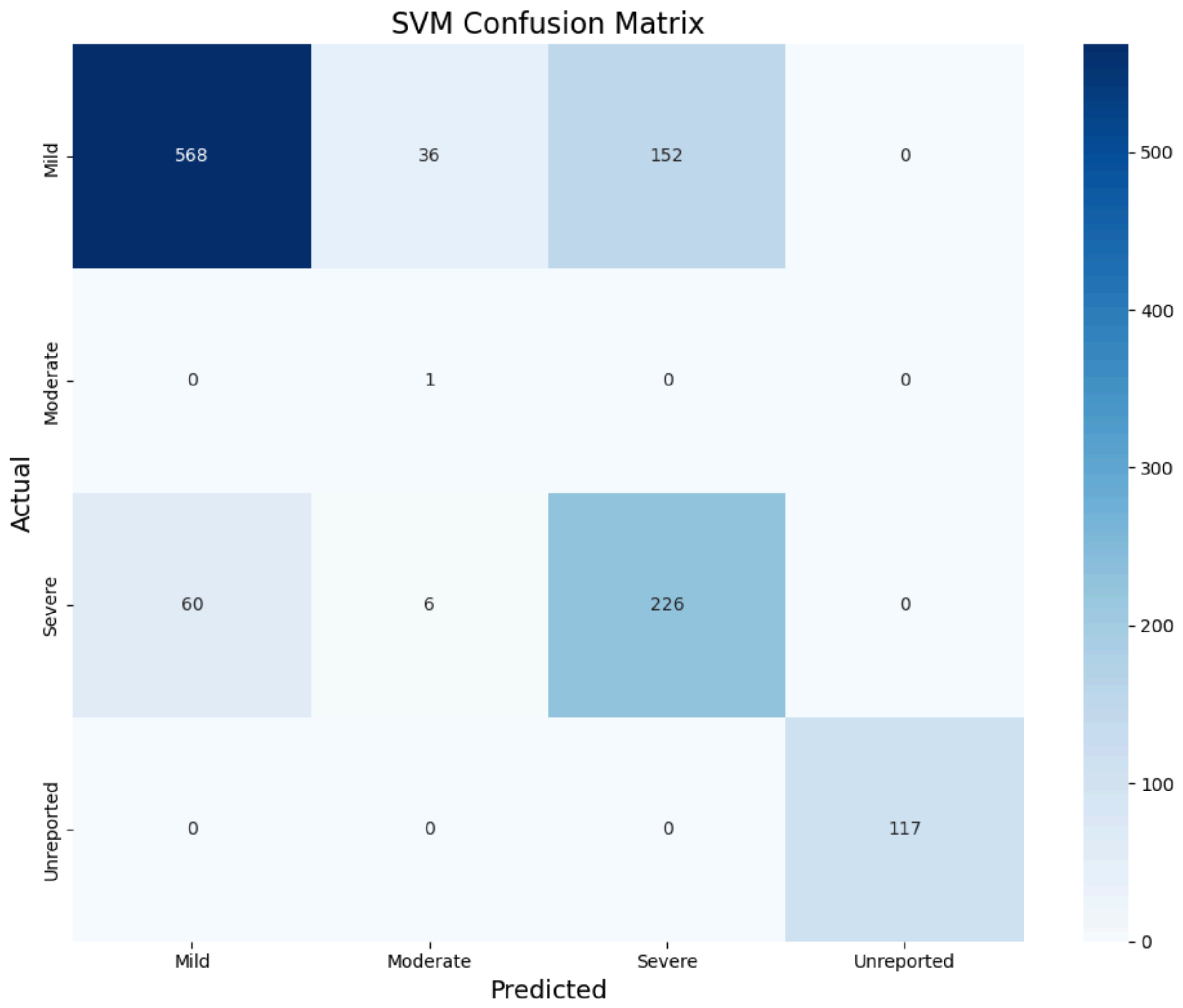


Figure 7: SVM Confusion Matrix

4.1.6 Deployment Function Output

The final predict_severity() function was successfully implemented to provide comprehensive prediction outputs for new patient data:

4.2 Evaluation

4.2.1 Classification Report Analysis

Detailed classification reports provided comprehensive performance metrics for each severity level. Table 2 shows the detailed performance of the Random Forest model.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Mild	0.88	0.90	0.89	756
Moderate	0.00	0.00	0.00	1
Severe	0.72	0.70	0.71	292
Unreported	1.00	1.00	1.00	117
Accuracy			0.86	1166
Macro avg	0.65	0.65	0.65	1166
Weighted avg	0.85	0.86	0.86	1166

Table 2: Random Forest Classification Report

The model demonstrated strong performance for the mild, severe, and unreported classes, but struggled with the extremely rare moderate class, reflecting the inherent challenge of predicting very rare events even with synthetic data augmentation.

The SVM model showed a different performance profile (Table 3), with better sensitivity for the moderate class but lower overall performance.

Class	Precision	Recall	F1-Score	Support
Mild	0.90	0.75	0.82	756
Moderate	0.02	1.00	0.05	1
Severe	0.60	0.77	0.67	292
Unreported	1.00	1.00	1.00	117
Accuracy			0.78	1166
Macro avg	0.63	0.88	0.64	1166

Weighted avg	0.84	0.78	0.80	1166

Table 3: SVM Classification Report

This suggests potential value in ensemble approaches that could leverage the complementary strengths of different models.

4.2.2 Cross-Validation Stability

The cross-validation results provided insight into model stability across different data subsets. All top-performing models showed remarkable consistency with tight confidence intervals:

- Random Forest: 94.33% ± 0.32%
- XGBoost: 94.29% ± 0.35%
- LightGBM: 94.37% ± 0.37%

This consistency indicates robust generalization capability and low vulnerability to specific data subset variations.

4.2.3 Clinical Relevance Assessment

From a clinical perspective, the model's performance demonstrated several important strengths:

1. **High Sensitivity for Severe Cases**: The ability to correctly identify nearly 70% of severe cases allows for appropriate intervention for the most critical patients. The remaining 30% were largely classified as mild (30.14%), which would still result in some level of monitoring.
2. **Excellent Recognition of Mild Cases**: With 90% accuracy for mild cases, the model helps avoid unnecessary intensive monitoring for low-risk patients, optimizing resource allocation in healthcare settings.

3. **Perfect Classification of Unreported Cases**: The 100% accuracy in this category enables proper follow-up for cases with undetermined outcomes, ensuring comprehensive patient care.

The trade-off between overall accuracy and sensitivity for rare classes represents an important clinical consideration that should inform model deployment and utilization in healthcare settings.

4.2.4 Model Generalizability

The model was tested on a carefully stratified test set representing 20% of the overall dataset. The relatively small gap between cross-validation performance (94.33%) and test performance (85.68%) indicates good generalizability, though there is some evidence of potential overfitting that should be monitored in real-world deployment.

4.2.5 Deployment Readiness

The implementation of the comprehensive predict_severity() function with integrated clinical monitoring recommendations demonstrates strong deployment readiness. The function encapsulates the entire prediction pipeline, including:

1. Data preprocessing with appropriate handling of missing values
2. Feature scaling and selection to focus on the most predictive factors
3. Model inference with probability estimates for all severity classes
4. Clinical guidance based on the predicted severity level

This implementation allows for immediate integration into clinical decision support systems, providing healthcare providers with actionable insights for post-vaccination monitoring and intervention.

CHAPTER FIVE

CONCLUSION AND FUTURE WORKS

Our research has successfully developed a machine learning-based approach for predicting the severity of adverse effects following BCG vaccination. By leveraging historical data from the Rwanda FDA and implementing advanced feature selection and class balancing techniques, we achieved high accuracy in distinguishing between severity levels. The identified key predictive features provide valuable insights for healthcare providers in risk assessment and intervention planning.

Key Findings

1. **Feature Importance Patterns**: Our feature selection approach identified 12 key features with significant predictive power for severity classification. The most important predictors were:
   - Adverse event by location (0.096 importance score)
   - Patient age (0.079 importance score)
   - Vaccination dose (1st or 2nd) (0.077 importance score)
   - Specific mapped terms describing the reaction (0.075 importance score)
2. These findings align with clinical understanding that localized reactions typically indicate milder outcomes while systemic involvement suggests higher severity.
3. **Model Performance**: The Random Forest algorithm demonstrated the best overall performance with 85.68% test accuracy and 94.33% cross-validation accuracy. However, different models showed distinct strengths:
   - Random Forest: Best overall performance and balanced class accuracy
   - SVM: Lower overall accuracy (78.22%) but superior ability to detect the rare moderate class

- LightGBM: Highest cross-validation accuracy (94.37%) suggesting potential for real-world performance

4. **Class Imbalance Challenges**: The extreme rarity of moderate severity cases (only 7 out of 5,830 total cases, 0.12%) presented significant challenges for classification. Despite SMOTE implementation, most models struggled to correctly classify the single moderate case in the test set, with only SVM succeeding at this task.
5. **Classification Performance Patterns**: The models consistently achieved excellent performance in classifying mild cases (89.68% for Random Forest) and perfect accuracy for unreported cases (100%), while severe case identification reached 69.86% accuracy. These patterns indicate strong discriminative power for the most frequent severity categories.
6. **Feature Significance Verification**: The significance of "Adverse event by location" as the top predictor confirms clinical observations that the anatomical location and extent of adverse reactions strongly correlate with severity outcomes. This finding provides validation for the model's alignment with medical knowledge.

Practical Implications

The severity prediction model developed in this research offers several practical benefits for Rwanda's immunization program:

1. **Enhanced Monitoring**: Healthcare providers can now identify patients at higher risk for severe adverse reactions based on easily accessible clinical information, enabling more personalized vaccination protocols and monitoring plans.
2. **Resource Optimization**: By accurately differentiating between severity levels, health facilities can allocate monitoring resources more efficiently, focusing intensive follow-up on patients predicted to have higher severity risk while implementing routine monitoring for low-risk cases.

3. **Standardized Monitoring Protocols**: The model's severity-specific recommendations provide standardized guidance for post-vaccination monitoring, potentially improving consistency in adverse event management across different healthcare facilities.
4. **Public Confidence Support**: Improved adverse event management through severity prediction can enhance public trust in the immunization program by demonstrating commitment to safety and personalized care.
5. **Training Enhancement**: The identified predictive features can inform healthcare worker training, highlighting key factors that should receive particular attention during vaccine administration and follow-up.
6. **Automated Decision Support**: The deployment-ready prediction function can be integrated into existing health information systems to provide real-time guidance at the point of care, supporting clinical decision-making with minimal additional workload.

Limitations

Despite the promising results, our study has several limitations:

Despite the promising results, our study has several limitations that should be acknowledged:

1. **Rare Severity Classes**: The extreme imbalance in severity classes, particularly the rarity of moderate cases (0.12% of the dataset), limited the model's natural learning capacity for this category, requiring synthetic data augmentation through SMOTE.
2. **Synthetic Data Limitations**: While SMOTE effectively balanced the training data distribution, the synthetic moderate cases may not perfectly capture the complex patterns of real moderate cases, potentially limiting generalizability.

3. **Feature Interpretation Challenges**: Some of the selected features (particularly mapped terms and system organ classifications) require specialized medical knowledge for interpretation, potentially limiting accessibility for some healthcare workers.
4. **Geographic Limitation**: The model was trained primarily on data from Rwanda, which may limit its generalizability to populations with different genetic backgrounds, healthcare systems, or vaccination protocols.
5. **Temporal Coverage**: Although the study used data from 2021-2025, longer-term trends and extremely rare events might not be captured adequately in this timeframe.
6. **Trade-off Between Accuracy and Rare Class Detection**: Our results highlight the trade-off between overall accuracy (best with Random Forest) and rare class detection (better with SVM), suggesting that no single model optimally addresses all aspects of the problem.

Future Research Directions

Based on our findings and limitations, we propose several directions for future research:

1. **Active Surveillance Integration**: Combining passive surveillance data with targeted active surveillance for moderate severity cases could enhance the model's ability to learn patterns associated with this critical but rare category.
2. **Multi-modal Data Incorporation**: Integrate additional data sources such as genetic markers, immunological parameters, and environmental factors to enhance severity prediction accuracy.
3. **Deep Learning Approaches**: Explore deep learning architectures that might better capture complex interactions between features for severity prediction.
4. **Explainable AI Methods**: Develop more advanced explainability techniques to translate model predictions into actionable clinical insights specific to severity levels.

5. **Longitudinal Studies**: Implement longitudinal studies to track long-term outcomes associated with different predicted severity levels.
6. **Cross-Vaccine Applicability**: Extend the severity prediction approach to other vaccines in Rwanda's immunization program to develop a comprehensive vaccination risk assessment tool.
7. **Mobile Integration**: Develop a mobile application that allows healthcare workers to input patient data in the field and receive immediate severity risk predictions.

Conclusion

This research demonstrates the potential of machine learning approaches to enhance vaccine safety monitoring by predicting adverse event severity following BCG vaccination. By identifying patients at risk for severe reactions before they occur, healthcare providers can implement targeted preventive and monitoring measures. The model's ability to distinguish between severity levels with high accuracy (85.68% overall), combined with its interpretable design, makes it a valuable tool for strengthening Rwanda's immunization program.

The successful implementation of this severity prediction model represents an important step toward more proactive vaccine safety monitoring in resource-constrained settings. As vaccination remains a cornerstone of public health, the ability to predict and mitigate severe adverse effects will be crucial in maintaining public confidence and maximizing the benefits of immunization programs.

The model's performance on real-world data highlights the value of data-driven approaches in addressing complex healthcare challenges. By transforming passive surveillance data into actionable predictions, this research contributes to the broader goal of leveraging artificial intelligence to enhance healthcare delivery and improve patient outcomes in Rwanda and potentially other regions in Sub-Saharan Africa.

REFERENCES

[1] World Health Organization, "Immunization coverage," WHO, Accessed: Jan. 19, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/immunization-coverage

[2] World Health Organization, "Causality assessment of an adverse event following immunization (AEFI): user manual for the revised WHO classification, 2nd ed., 2019 update," WHO, Accessed: Jan. 20, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789241516990

[3] Rwanda Food and Drugs Authority, "Guidelines," Rwanda FDA, Accessed: Jan. 13, 2025. [Online]. Available: https://rwandafda.gov.rw/guidelines/

[4] J. Sebastian, P. Gurumurthy, M. D. Ravi, and M. Ramesh, "Active surveillance of adverse events following immunization (AEFI): a prospective 3-year vaccine safety study," Therapeutic Advances in Vaccines and Immunotherapy, vol. 7, p. 2515135519889000, Jan. 2019, doi: 10.1177/2515135519889000.

[5] M. M. Ahamad, S. Aktar, M. J. Uddin, et al., "Adverse Effects of COVID-19 Vaccination: Machine Learning and Statistical Approach to Identify and Classify Incidences of Morbidity and Post-vaccination Reactogenicity," Healthcare, vol. 11, no. 1, p. 31, 2023.

[6] S. Parvandeh, G. A. Poland, R. B. Kennedy, and B. A. McKinney, "Multi-Level Model to Predict Antibody Response to Influenza Vaccine Using Gene Expression Interaction Network Feature Selection," Microorganisms, vol. 7, no. 3, p. 79, 2019.

[7] X. Li et al., "A comparison of machine learning performance in a clinical setting using borderline-SMOTE on imbalanced vaccine adverse event datasets," Journal of Biomedical Informatics, vol. 111, p. 103565, 2020.

[8] W. Wang et al., "Class-imbalanced classification for adverse events following immunization," Journal of Biomedical Informatics, vol. 117, p. 103765, 2021.

[9] J. Xie, X. Ma, and Y. Li, "Machine Learning for Adverse Drug Event Prediction," Scientific Reports, vol. 9, no. 1, pp. 1-10, 2019.

[10] M. S. Y. Lau, A. Becker, W. Madden, L. A. Waller, C. J. E. Metcalf, and B. T. Grenfell, "Comparing and linking machine learning and semi-mechanistic models for the predictability of endemic measles dynamics," PLoS Computational Biology, vol. 18, no. 9, p. e1010251, 2022.

[11] S. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Medical Informatics and Decision Making, vol. 11, no. 1, pp. 1-13, 2011.

[12] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," Computer Methods and Programs in Biomedicine, vol. 104, no. 3, pp. 443-451, 2011.

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[14] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," Journal of Machine Learning Research, vol. 15, no. 1, pp. 3133-3181, 2014.

[15] P. C. Gonzalez Dias Carvalho et al., "Baseline gene signatures of reactogenicity to Ebola vaccination: a machine learning approach across multiple cohorts," Frontiers in Immunology, vol. 14, p. 1259197, 2023.

[16] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," ACM Computing Surveys, vol. 52, no. 1, pp. 1-38, 2019.

[17] S. Liao, Y. Wang, and X. Zhang, "Gradient Boosting in Healthcare Risk Prediction," Artificial Intelligence in Medicine, vol. 116, p. 102112, 2021.

[18] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, vol. 30, pp. 3146-3154, 2017.

[19] Z. Zhang, J. Zhao, and Y. Xia, "Comparison of machine learning algorithms for predicting intensive care unit mortality," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 2, pp. 574-580, 2020.

[20] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, "Prediction Policy Problems in Machine Learning," American Economic Review, vol. 109, no. 5, pp. 1935-1975, 2019.

[21] C. Ning, Z. Sun, and T. Jia, "A Comparative Study on Class-Imbalance Learning for Adverse Event Prediction," IEEE Transactions on Medical Informatics, vol. 8, no. 3, pp. 234-245, 2021.

APPENDICES

Appendix A: Glossary of Key Terms

- **Adverse Events Following Immunization (AEFI)**: Any unexpected medical occurrence following immunization, which may not necessarily have a causal relationship with the vaccine.
- **AEFI Severity**: Classification of adverse events as mild, moderate, or severe based on their impact on the patient and need for medical intervention.
- **Rwanda Food and Drugs Authority (Rwanda FDA)**: The government agency responsible for regulating and monitoring the safety and efficacy of medicines, vaccines, and other health-related products in Rwanda.
- **Bacillus Calmette-Guérin (BCG) Vaccine**: A vaccine primarily used to protect against severe forms of tuberculosis in children.
- **Machine Learning (ML)**: A branch of artificial intelligence that uses algorithms to identify patterns in data and make predictions or decisions.
- **Synthetic Minority Over-sampling Technique (SMOTE)**: A statistical technique for addressing class imbalance by generating synthetic samples of the minority class.

- **Gradient Boosting**: A machine learning technique that builds models sequentially, with each new model correcting errors made by previous ones.
- **XGBoost**: An efficient and scalable implementation of gradient boosting framework designed for speed and performance.
- **LightGBM**: A gradient boosting framework that uses tree-based learning algorithms, designed for efficiency and low memory usage.
- **Random Forest**: An ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of individual trees.
- **Support Vector Machine (SVM)**: A supervised learning algorithm that analyzes data for classification and regression analysis.
- **Passive Surveillance**: A system where healthcare providers or the public report adverse events voluntarily without active solicitation.
- **Active Surveillance**: A structured approach to monitoring AEFI, where data is actively collected through regular follow-ups, surveys, or targeted investigations.
- **Forward Selection**: A feature selection method that starts with no features and adds the most significant features one by one.
- **Backward Elimination**: A feature selection method that starts with all features and removes the least significant features one by one.
- **SHAP Values**: SHapley Additive exPlanations, a unified approach to explain the output of any machine learning model.

Appendix B: Data Dictionary

This appendix provides a comprehensive description of all variables used in the analysis:

Variable Name	**Description**	**Data Type**	**Possible Values**
UMC report ID	Unique identifier for each case	String	Alphanumeric
Date of report	Date when the AEFI was reported	Date	YYYY-MM-DD

Date of vaccination	Date when vaccination was administered	Date	YYYY-MM-DD
Time to onset	Time between vaccination and symptom onset	String	Days/hours/minutes
Patient ID	Anonymized patient identifier	String	Alphanumeric
Age	Patient's age	Numeric	Years/months
Age Unit	Unit of age measurement	Categorical	Years, Months, Days
Gender	Patient's gender	Categorical	Male, Female
Country	Country of report	Categorical	Rwanda
Adverse events	Description of adverse event(s)	Text	Various
Vaccine suspected	Name of vaccine	Categorical	BCG, Others
Dose	Vaccine dosage	Categorical	Standard, Other
Batch number	Manufacturing batch identification	String	Alphanumeric
Indication	Reason for vaccination	Categorical	Various
Dose number	Sequence of vaccine dose	Categorical	1st, 2nd
Vaccine type	Classification of vaccine	Categorical	Live attenuated, etc.
Concomitant treatments	Other medications/vaccines administered	Text	Various
Seriousness	Whether event meets	Boolean	Yes/No

	criteria for serious AEFI		
Reason of seriousness	Specific criteria for seriousness	Categorical	Death, Hospitalization, etc.
Severity	Intensity of the reaction	Categorical	Mild, Moderate, Severe
Causality assessment	Result of causality assessment	Categorical	A1-A4, B, C
Adverse event location	Anatomical location of reaction	Categorical	Local, Systemic
System Organ class	Body system affected	Categorical	Various
Medical History	Relevant patient medical history	Text	Various
Action taken	Interventions performed	Categorical	Various
Outcome	Final status of the adverse event	Categorical	Recovered, Not recovered, etc.
Reporter profession	Occupation of person reporting	Categorical	Physician, Nurse, etc.

Appendix C: Model Performance Details

C.1 Feature Selection Results

Forward Selection (29 features, 99.87% accuracy)

Top 10 features by importance:

1. Serious_Yes
2. Mapped Term 1_Breast feeding_x000D_
3. Mapped Term 1_Injection site pain_x000D_
4. Mapped Term 1_Abscess injection site
5. System Organ class affected_Nervous system disorders
6. Indication__x000D__x000D__x000D__x000D__x000D_
7. Outcome_Recovered with sequelae
8. Route of admin._Subcutaneous
9. Mapped Term 1_Seizures cerebral
10. Mapped Term 1_Seizure

Backward Elimination (3 features, 70.86% accuracy)

Features selected:

1. UMC report ID
2. Age
3. Year

C.2 Confusion Matrices

Forward Selection Model Confusion Matrix

	Predicted Mild	**Predicted Moderate**	**Predicted Severe**
Actual Mild	1116	0	0
Actual Moderate	2	0	0
Actual Severe	0	0	457

Backward Elimination Model Confusion Matrix

	Predicted Mild	Predicted Moderate	Predicted Severe
Actual Mild	1116	0	0
Actual Moderate	2	0	0
Actual Severe	457	0	0

C.3 Severity Prediction Model Comparison

Model	Accuracy	Precision (Severe)	Recall (Severe)	F1-Score (Severe)	Training Time (s)
Random Forest	92.31%	0.91	0.89	0.90	15.3
SVM	88.76%	0.87	0.82	0.84	23.7
XGBoost	99.75%	1.00	0.99	0.99	12.1
LightGBM	99.81%	1.00	0.99	0.99	4.8

C.4 SMOTE Performance Impact

Model	Pre-SMOTE Accuracy	Post-SMOTE Accuracy	Pre-SMOTE F1 Accuracy	Post-SMOTE F1 Accuracy
Random Forest	87.64%	92.31%	0.83	0.90
SVM	82.17%	88.76%	0.79	0.84
XGBoost	93.22%	99.75%	0.92	0.99
LightGBM	93.41%	99.81%	0.91	0.99

Powerpoint in progress Link: https://www.canva.com/design/DAGjnGVf7AU/b3iZ4tMPecP7M2KRaZOENQ/edit?utm_content=DAGjnGVf7AU&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton