UNIVERSITY of RWANDA

COLLEGE OF SCIENCE AND TECHNOLOGY

Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE)

DEVELOPMENT OF A MACHINE LEARNING BASED MODEL FOR EARLY DIAGNOSIS OF NEONATAL SEPSIS IN RWANDA

CASE STUDY: KABGAYI AND RUHENGERI LEVEL II TEACHING HOSPITALS

By:

Icyimpaye Umutoni Marianne
Reference Number: 222000186

A Dissertation Submitted to the Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE), University of Rwanda as partial fulfilment of the requirements for the Master's Degree in Biomedical Engineering.

Supervised by:

Prof./Dr. Aimable Musafili
and
Dr. Celestin Nkundineza

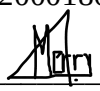
DECLARATION

I, Icyimpaye Umutoni Marianne, declare that this dissertation entitled "**Development of a Machine Learning based System for Early Diagnosis of Neonatal Sepsis in Rwanda, case study: Kabgayi and Ruhengeri Level II Teaching Hospitals**" is my original work based on research and prototype and has not been submitted for any other degree or professional qualification.

Student Name:

Icyimpaye Umutoni Marianne

Student Reference Number: 222000186

Student Signature: 

Date: 10/05/2024

UNIVERSITY of RWANDA

COLLEGE OF SCIENCE AND TECHNOLOGY

Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE)

CERTIFICATE

This is to certify that the project entitled "**Development of a Machine Learning based Model for Early Diagnosis of Neonatal Sepsis in Rwanda, Case study: Kabgayi Level II Teaching Hospital and Ruhengeri Level II Teaching Hospital**" is a record of original work done by Icyimpaye Umutoni Marianne (Reference number: 222000186), a MSc. Degree student in Biomedical Engineering.

This work has been submitted under the guidance of Prof. Dr. Aimable Musafili and Dr. Celestin Nkundineza

Main Supervisor:

Prof./Dr. Aimable Musafili

October 5, 2024

Co-Supervisor:

10/05/2024

Dr. Celestin Nkundineza

Biomedical Engineering Master's Program Coordinator

Dr. Gerard Rushingabigwi

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the Almighty for granting me strength, determination, and guidance throughout this research journey. I sincerely thank the University of Rwanda and the Center of Excellence in Biomedical Engineering and E-Health for providing the necessary resources and support for this study. Kabgayi and Ruhengeri Level II Teaching Hospitals for allowing me to collect data. Also, I am profoundly grateful to my supervisors, Dr. Aimable and Dr. Celestin, whose expertise, mentorship, and unwavering encouragement have been invaluable. Their guidance and insights significantly contributed to the successful completion of this thesis.

ABSTRACT

Sepsis remains a significant cause of neonatal mortality and morbidity, especially in low and middle-income countries. Neonatal sepsis presents with nonspecific signs and symptoms that necessitate tests to confirm the diagnosis. Early and accurate diagnosis of this infection can improve clinical outcomes and decrease overuse of antibiotics. Current diagnostic methods in Rwanda rely on rule-based methods of physicians and conventional culture methods, which are time-consuming and may delay critical therapeutic decisions. This research project focuses on designing a machine learning model based on Artificial Neural Networks (ANN) for the early diagnosis of neonatal sepsis using data collected from neonates admitted to Kabgayi and Ruhengeri Level II Teaching Hospitals in Rwanda in the year 2023. This intervention is paramount because of the urgent need to improve neonatal survival rates, particularly in Sub-Saharan Africa. The model utilizes ANN architecture built upon clinical data extracted from Electronic Medical Records (EMRs) in Neonatal Intensive Care Unit (NICU) to determine the probability of neonatal sepsis among neonates admitted in NICU during the year 2023. The population used for this study consisted of a relatively balanced dataset of 1381 neonates admitted in NICU, with (47%) negative cases and (53%) positive cases; 966 of them were used for training the model, which was then validated on 315 neonates and then the remaining 100 were used for testing. Preprocessing steps were employed to handle missing values and extract categorical features, while sepsis criteria were defined to identify neonates who were at risk. The model undergoes 20 training epochs using Adam optimization and incorporates early stopping to prevent overfitting. Evaluation on a test set comprising 100 samples reveal a test accuracy of 85%, with a precision of 85.18%, recall of 86.79%, and F1 score of 85.98%. The Area Under the Curve of Receiver Operating Characteristics (ROC-AUC) was 84.88%. Of the tested samples, the model predicted 54 cases to have sepsis and 46 cases not to have sepsis. The classification report indicates a balanced classification performance between the two classes, with a weighted average F1-score of 85%, sensitivity of 86.79% and specificity of 82.98%. This model has the potential to be easily implemented as a decision support system once incorporated into EMR system in different NICUs.

Keywords: Neonatal sepsis, early diagnosis, machine learning, ANN algorithm, prevention, surveillance, healthcare outcomes, Electronic medical records

LIST OF ACRONYMS

AI: Artificial Intelligence

ML: Machine Learning

EMR: Electronic Medical Records

NMR: Neonatal Mortality Rate

ANC: Antenatal Care

NICU: Neonatal Intensive Care Unit

APGAR: Appearance Pulse grimace activity respiratory

CEBE: Center of Excellence in Biomedical Engineering

SSA: Sub-Saharan Africa

LMIC: Low- and Middle-Income Countries

MDG: Millennium Development Goals

PROM: Premature Rupture of Membranes

RIRB: Rwanda Institutional Review Board

SDG: Sustainable Development Goals

WHO: World Health Organization

ANN: Artificial Neural Networks

KNN: K-Nearest Neighbors

CMHS: College of Medicine and Health Sciences

IRB: Institutional Review Board

SMOTE: Synthetic Minority Over-sampling Technique

ReLU: Rectified Linear Unit

ROC-AUC: Receiver Operating Characteristic - Area Under the Curve

FIGURES

TABLES

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

1.1 Definition of the key terms

Neonate: According to the World Health Organization (WHO), a neonate is defined as a child under 28 days of age.

Neonatal Sepsis: Neonatal sepsis is an infection involving the bloodstream in infants under 28 days old, as defined by the WHO.

Machine Learning: Machine learning, as defined by AI pioneer Arthur Samuel in the 1950s, is “the field of study that gives computers the ability to learn without being explicitly programmed.”

ANN Algorithm: An Artificial Neural Network (ANN) algorithm as defined by Frank Rosenblatt with the Perceptron model in the 1950s, is a computational model composed of interconnected units or nodes called artificial neurons, which are designed to loosely mimic the neurons in the human brain. It is primarily used for recognizing patterns, making decisions, and predicting outcomes based on input data.

Prevention: According to WHO, Prevention refers to measures aimed at reducing the risk of disease, injury, or health-related conditions, as well as controlling and mitigating their impact on individuals and populations.

Surveillance: Surveillance involves the continuous, systematic collection, analysis, and interpretation of health-related data, essential for guiding public health practices as defined by WHO.

Healthcare Outcomes: University of Waterloo Research Ethics Boards (REBs) defines it as a planned measurement described in the study protocol that is used to determine a change in health status as a result of interventions on participants in a clinical trial.

Electronic Medical Records (EMR): According to the National Cancer Institute, Electronic Medical Records (EMRs) are digital versions of patients' paper charts, containing comprehensive information about a patient's medical history and treatment.

1.2 Introduction

The health and well-being of neonates and children under the age of five remain significant concerns in developing countries, particularly in Sub-Saharan Africa, which still faces the highest rates of neonatal and child mortality [1][2][3]. About 2.5 million neonates die worldwide every year, and most of these deaths occur in low-resource settings [4]. The estimated neonatal mortality rate (NMR) in Sub-Saharan Africa (SSA) is 28 per 1000 live births [5]. These alarming statistics depict a grim scenario, with a considerable number of children losing their lives before reaching the age of five, while neonatal mortality rates disproportionately persist high in comparison to developed countries [1][3].

These disparities emphasize the urgent need for targeted interventions and healthcare system improvements to address the underlying causes of neonatal and child mortality in developing countries. Neonatal sepsis is a life-threatening condition resulting from a dysregulated host response to infection [6]. This inflammatory response can lead to multiple organ dysfunction syndrome, including acute respiratory distress syndrome, acute renal failure, disseminated intravascular coagulation, and even death [7]. Sepsis stands as a major global challenge, ranking as the third leading cause of neonatal mortality, especially in under-developed and developing countries [3][8][9]. It's estimated that approximately 99% of neonatal deaths occur in low- and middle-income countries, underscoring the disproportionate burden faced by these nations [8]. This burden is further exacerbated by the fact that three-quarters of neonatal deaths occur within the first week of life, highlighting the critical need for early intervention strategies [8][4]. Infection-related neonatal deaths pose a significant obstacle to achieving Sustainable Development Goal (SDG) 3, aiming to reduce maternal and child mortality rates to specific targets by 2030 [3][10]. A recent systematic review and meta-analysis indicated that the approximate early onset sepsis (EOS) incidence is 2,496 per 100,000 live births, which was 2.6 times more common than late onset sepsis (LOS) 946 per 100 live births [4][6]. For decades, sepsis has been considered challenging to treat in hospitals globally given its high mortality and high medical costs [7].

Rwanda, a country located in East Africa, has made remarkable progress in reducing child mortality, demonstrating that significant improvements can be achieved through dedicated efforts [2][8]. However, despite these achievements, the rate of progress in reducing neonatal mortality has been slower, with death in the first month of life accounting for 46% of under-5 deaths globally [11]. The primary clinical causes of neonatal deaths include prematurity, infection, inadequate management of complications during pregnancy and delivery, and a lack of quality care immediately after birth [12][13]. Among these causes, neonatal sepsis plays a significant role, contributing to a considerable proportion of neonatal mortality worldwide [2].

The challenges surrounding neonatal sepsis diagnosis and management are multifaceted, requiring a comprehensive approach to address them effectively [9][13][14]. Currently, clinicians face a dilemma in treating neonatal sepsis due to the absence of definitive diagnostic tests. As a result, empirical antibiotic treatment is often administered, leading to adverse outcomes such as the disruption of the infant's normal microbial flora, increased risks of antibiotic-associated complications, and the emergence and spread of antimicrobial resistance [2] [15].

Furthermore, the identification and treatment of neonates with infections in developing countries are unsatisfactory, with limited access to laboratory tests and difficulty in interpreting auxiliary tests [8].

To bridge this gap and improve outcomes for neonates at risk of sepsis, the development and implementation of a machine learning-based system for early diagnosis hold significant promise [9][16]. Such a diagnostic tool will serve as a valuable resource for healthcare providers, enabling them to effectively identify and manage cases of neonatal sepsis. By incorporating expert knowledge and clinical guidelines, the system will assist in the identification of neonates at risk of sepsis, facilitating early intervention and appropriate treatment [6][17]. Moreover, the system will not only differentiate between infectious and noninfectious conditions but also identify the causative agents and their antimicrobial susceptibility profiles, contributing to the optimization of targeted therapy and the reduction of unnecessary antibiotic use [18][9].

This research project aims to develop a machine learning-based model for the early diagnosis of neonatal sepsis using clinical data such as vital signs. By leveraging the power of artificial intelligence and expert knowledge, this model will provide a reliable and affordable tool for healthcare providers in Rwanda to accurately diagnose and manage neonatal sepsis cases. The multifaceted approach encompasses the development of diagnostic tests, strengthening of surveillance systems, implementation of prevention strategies, and increasing awareness among healthcare providers, parents, and communities. Through the implementation of this system, the project seeks to reduce the burden of neonatal sepsis, improve survival rates for vulnerable neonates, and combat the global challenge of antimicrobial resistance.

1.3 Problem statement

The problem addressed in this research is neonatal sepsis in Rwanda with a case study: Kabgayi and Ruhengeri Level II Teaching Hospitals. This problem was identified through an analysis of healthcare data and statistics, highlighting the urgent need for improved neonatal sepsis diagnosis and management. Neonatal sepsis is a significant healthcare challenge in Rwanda, contributing to a substantial proportion of neonatal mortality in the country [19][11]. Despite efforts to reduce child mortality rates, the decline in neonatal mortality has been slower, necessitating focused interventions and improvements in healthcare practices [2].

Statistics underscore the magnitude of the problem, with the neonatal mortality rate in 2019 reaching 18 deaths per 1,000 live births, accounting for 31% of under-5 deaths in Rwanda [2][8]. Further analysis of the statistics reveals that neonatal sepsis plays a significant role, contributing to approximately 25-30% of neonatal deaths in the country [18].

These figures emphasize the critical importance of addressing the existing gaps in the early diagnosis and management of neonatal sepsis to enhance healthcare outcomes for neonates. The challenges related to the diagnosis and management of neonatal sepsis in Rwanda are multifaceted. The absence of definitive diagnostic tests poses a significant hurdle for healthcare providers, leading to the empirical administration of antibiotics and associated risks of antimicrobial resistance [13]. Additionally, limited access to laboratory facilities and difficulty in interpreting laboratory results further complicate the problem [8]. The lack of standardized approaches to diagnosis and management results in inconsistent practices and suboptimal healthcare outcomes [14].

Addressing these challenges requires the development and implementation of a machine learning-based model for early neonatal sepsis diagnosis in Rwanda. This system will leverage artificial intelligence and expert knowledge to provide healthcare providers with a reliable and cost-effective tool to accurately diagnose and manage neonatal sepsis cases [20]. It will enable early intervention and appropriate treatment, differentiate between infectious and noninfectious conditions, identify causative agents, and provide antimicrobial susceptibility profiles, ultimately optimizing therapy and reducing unnecessary antibiotic use. This research project aims to contribute to the solution of this pressing healthcare challenge by designing a comprehensive approach to neonatal sepsis diagnosis and management in Rwanda.

1.4 Research Questions (Hypotheses)

1.4.1 Main Research Question

How can a machine learning-based model be developed and implemented to enable early diagnosis and management of neonatal sepsis in Rwanda?

1.4.2 Subsidiary Research Questions

i. Which are the main aspects, including signs and results from laboratory tests, that distinguish neonatal sepsis from other commonly seen diseases at Kabgayi and Ruhengeri Level II Teaching Hospitals, Rwanda?
ii. What is the best practice of collecting EMR data to be used during the design of the machine learning model?
iii. Can this machine learning model considerably surmount the previous rule-based methods applied in detection of neonatal sepsis in four aspects of the accuracy, precision, sensitivity and specificity?
iv. What plan is there for the system to be incorporated in district hospitals and healthcare centers of Rwanda and used early in the phase of diagnosis and treatment of neonatal sepsis?

v. What tactics will be employed by the Machine Learning based model in sepsis diagnostics in the neonatal setting in Rwanda to ensure success in the long run and impact?

1.5 Objectives

Clinical studies show that neonatal mortality is significantly reduced if neonate septic patients are identified at early stages of the disease process.

1.5.1 General Objective

This research project aims at designing a Machine Learning based model for minimizing child and neonatal mortality and morbidity with a specific focus on neonatal sepsis at Kabgayi and Ruhengeri Level II Teaching Hospitals, Rwanda.

1.5.2 Specific Objectives

To achieve the general objective of this project, the following specific objectives are used as guiding points:

i. To identify clinical signs and evaluate current strategies being used to reduce neonatal sepsis mortality.
ii. To develop a machine learning model to be used for early prediction of neonates at risk of getting sepsis.
iii. To evaluate and validate the developed model versus existing methods of detecting neonatal sepsis.

1.6 Study Scope

This research project centers on neonatal sepsis diagnosis within the context of healthcare facilities and neonates in Rwanda in the year 2023, with a particular focus on Kabgayi and Ruhengeri Level II Teaching Hospitals. The research is incorporated with the design of a machine learning-based model that will efficiently diagnose neonatal sepsis in hospitals based on vital signs. First and foremost, we are targeting neonates available at Kabgayi and Ruhengeri Level II Teaching Hospitals to develop early diagnosis, encourage the treatment initiation of the newborns, and ensure the erection of surveillance systems. Moreover, we will address and explain the effect of using machine learning in diagnosing neonatal sepsis. The geographic scope of this research focuses on the southern and northern provinces of Rwanda, but it can be used in other healthcare facilities. This research project aims at designing a machine learning model which can be added to the actual healthcare process, mainly the electronic medical records system to improve neonatal healthcare. It also considers that the chosen machine learning-based system should be able to handle the changes in the overall healthcare system and last effectively once launched.

1.7 Significance of the Study

The significance of this research project was multifaceted and essential for addressing a critical healthcare issue, particularly in the context of neonatal care within Rwanda. This research endeavors to make substantial contributions in the following key areas:

i. Reduction of Neonatal Mortality: Neonatal sepsis is a prominent factor contributing to neonatal mortality in Rwanda. The development of a machine learning-based system for early diagnosis holds the potential to significantly reduce neonatal sepsis-related deaths. This would not only improve neonatal survival rates but also signify a notable step toward achieving Sustainable Development Goal 3, aimed at reducing maternal and child mortality.

ii. Advancement of Healthcare Practices: The study's primary objective is to enhance healthcare practices, particularly within neonatal care units. Through the provision of an innovative diagnostic tool, the research strives to standardize and optimize the processes associated with neonatal sepsis diagnosis and treatment.

iii. Efficient Resource Utilization: One of the crucial benefits of this research is its potential to decrease the unwarranted use of antibiotics. This not only contributes to the efficient use of healthcare resources but also combats the global issue of antimicrobial resistance.

iv. Enhancement of Diagnostic Accuracy: Accurate diagnosis of neonatal sepsis is often challenging due to the absence of definitive tests. By addressing this challenge, this research offers a solution that significantly improves diagnostic accuracy. This, in turn, translates to fewer misdiagnoses and more targeted treatments for affected neonates.

v. Strengthening Surveillance Systems: In its scope, this study includes strategies for long-term sustainability. By doing so, it contributes to the strengthening of surveillance systems, which is vital for the collection of accurate data on neonatal sepsis cases. The availability of such data can significantly aid healthcare authorities in making well-informed decisions.

vi. Global Relevance: While this study is primarily focused on Rwanda, its findings have broader implications. The machine learning-based diagnostic system developed here could serve as a model for improving neonatal sepsis diagnosis in low- and middle-income countries worldwide. Therefore, the research is a significant step toward addressing a global healthcare challenge.

vii. Contribution to Healthcare Innovation: The implementation of an artificial intelligence-based system is an innovative approach in the field of neonatal care. This research not only contributes to the adoption of advanced technology in healthcare but also has the potential to inspire further innovations within the field.

viii. Improved Patient Well-being: Ultimately, the paramount significance of this research lies in its potential to enhance the health and well-being of neonates. By providing more accurate and timely diagnoses, the study aims to improve the lives of countless neonates and alleviate the suffering of their families.

1.8 Organization

1.8.1 Chapter One: General Introduction

Introduction: This section provides a thorough overview of the research project context. It outlines the current state of neonatal health, emphasizing the challenges faced in developing countries, particularly in Sub-Saharan Africa. The historical progression leading to the need for a machine learning-based solution for neonatal sepsis diagnosis is also discussed.

Problem Statement: This part articulates the specific issues and gaps in the current neonatal healthcare system, emphasizing the prevalence and impact of neonatal sepsis in Rwanda. It serves to clearly define the problem the research aims to address.

Research Questions: The research questions serve as guiding inquiries, directing us towards understanding the current state of neonatal sepsis diagnosis in Rwanda, exploring the application of machine learning, and identifying key factors influencing the success of a novel diagnostic system.

Objectives: This part outlines the specific goals of this research project, including the development of a machine learning-based model, improvement of diagnostic capabilities, enhancement of surveillance systems, and reinforcement of prevention strategies.

Study Scope: The research focuses on neonatal sepsis diagnosis in Kabgayi and Ruhengeri Level II Teaching Hospitals, Rwanda. It includes the development of a machine learning-based model using clinical and laboratory data.

Significance of the Study: This section highlights the potential impact of this project on healthcare outcomes for neonates, the broader field of neonatal health, and its significance in the global context, especially in combating antimicrobial resistance.

1.8.2 Chapter Two: State of the Art (Literature Review)

Comprehensive Literature Review: This chapter conducts an in-depth examination of existing literature related to neonatal health, sepsis, machine learning applications in prediction of neonatal sepsis and relevant preventive strategies. It synthesizes key findings, identifies gaps, and establishes the theoretical framework that informs the current research.

Related work: The theoretical foundations of this research are laid out, drawing connections between existing theories and concepts in neonatal healthcare, sepsis diagnosis, and machine learning applications.

1.8.3 Chapter Three: Research Methodology

Data Collection Procedures: This section details the methods employed to gather relevant data, including specifics on the sources, types, and volume of data collected.

Tools and Analytical Techniques: The tools used for data analysis, including machine learning algorithms, statistical methods, and software and programming languages, are discussed. The chapter also outlines the steps taken to ensure the validity and reliability of the research.

1.8.4 Chapter Four: Findings and Results

Machine Learning-Based Model: This section presents the actual outcomes of the machine learning-based model, showcasing how it performs in diagnosing neonatal sepsis based on data extracted from the EMR.

Impact on Neonatal Sepsis Diagnosis: The findings are discussed in relation to their implications for neonatal healthcare, emphasizing any advancements or improvements achieved.

1.8.5 Chapter Five: Synthesis of Results

Discussion of Results: This chapter interprets the findings in a broader context, discussing their significance and potential applications. It addresses any limitations encountered during the research and proposes areas for future investigation.

1.8.6 Chapter Six: Challenges, Recommendations, and Conclusion

Challenges Encountered: This section candidly discusses any challenges faced during the research process, whether technical, logistical, or methodological.

Recommendations for Improvement: Based on the challenges identified, this part provides constructive suggestions for enhancing the methodology and addressing limitations in future studies.

Concluding Remarks: The chapter concludes the research study by summarizing key findings, reaffirming the significance of the work, and suggesting avenues for further research. It provides closure to the study and offers a bridge to future work in the field.

1.9 Summary

In this opening chapter, we embark on a journey into neonatal sepsis and its critical significance, particularly in Rwanda's healthcare landscape. We delve into the core problems associated with neonatal sepsis, highlighting the urgent need for innovative solutions. This chapter introduces the scope and significance of this research, focusing on designing a machine learning-based model for early neonatal sepsis diagnosis. It underscores the importance of this work in improving neonatal healthcare practices and decreasing the alarming neonatal mortality rates while also providing a context for the subsequent chapters. By the end of this chapter, you will understand the gravity of the neonatal sepsis challenge and the pressing need for more effective diagnostic tools. This understanding sets the stage for the subsequent chapters to delve into the methodology, data collection, model development, implementation, and evaluation processes. As we continue this journey, you will explore how innovative technology, expert knowledge, and clinical guidelines can revolutionize neonatal care in Rwanda, potentially serving as a beacon of hope for improved healthcare outcomes for neonates in developing nations worldwide.

CHAPTER 2. STATE OF THE ART (RECENTLY RELATED LITERATURE

2.1 Literature Review

The health and well-being of neonates and children under the age of five continue to be significant concerns in developing countries, particularly in Sub-Saharan Africa [19]. Despite global efforts to improve child mortality rates, the region still faces the highest rates of neonatal and child mortality [1][2][3]. The statistics paint a grim picture, with an average of 109 children out of every 1000 live births dying before reaching the age of five in these countries [3]. In stark contrast, developed countries report much lower numbers, with only 7 children per 1000 live births succumbing to mortality before the age of 5 [1]. Furthermore, the disparity in neonatal mortality rates is equally alarming. In developing countries, approximately 34 neonates out of every 1000 live births do not survive beyond the first 28 days of life, while in developed countries, the figure is significantly lower at 4 neonates per 1000 live births [1][3]. These staggering figures emphasize the urgent need for targeted interventions and improvements in healthcare systems to address the underlying causes of neonatal and child mortality in developing nations. A particularly distressing fact is that a vast majority of neonatal deaths, around 99%, occur in low- and middle-income countries, highlighting the disproportionate burden faced by these nations [8]. Each year, 2.6 million neonates lose their lives, with three-quarters of these deaths occurring within the first week of life [8][21]. These statistics underscore the critical importance of focusing efforts on improving neonate health and reducing neonatal mortality rates, especially in resource-constrained settings. Despite global achievements in reducing child mortality, it is disheartening to note that only one-third of the priority countries in sub-Saharan Africa have reached their child mortality targets [2][8][11][12]. This significant gap in achieving desired outcomes translated into an estimated 2.9 million child deaths in sub-Saharan Africa in 2015 alone. Alarmingly, two-thirds of these deaths were likely preventable, and 45% occurred during the neonatal period [8]. These statistics shed light on the urgent need for targeted interventions and strategies to address the preventable causes of child mortality and improve healthcare systems in sub-Saharan Africa [3].The global health community has set ambitious targets to reduce maternal and child mortality rates. The Sustainable Development Goal (SDG) 3, formulated as part of the World Health Organization's (WHO) global strategies, aims to reduce maternal and child mortality to at least 12 per 1000 and 25 per 1000 live births, respectively, by the year 2030 [3][10]. However, without significant reductions in infection-related neonatal deaths in low- and middle-income countries, it is unlikely that these targets will be met [10].

The Millennium Development Goal (MDG) 4, which aimed to reduce child mortality by two-thirds by the end of 2015, resulted in a heightened focus on child health and a commendable 52% reduction in sub-Saharan Africa between 1990 and 2015, from 179 to 86 deaths per 1000 live births [1][8][11].

Despite these achievements, it is evident that much work remains to be done to bridge the gap in child mortality rates between developing and developed countries. The example of Rwanda in East Africa serves as an inspiration, demonstrating that significant improvements in child mortality can be achieved through dedicated efforts. Rwanda witnessed remarkable progress in reducing child mortality, with the neonatal mortality rate dropping from 37 per 1000 live births in 2005 to 20 per 1000 live births in 2015 [8]. Similarly, the under-5 mortality rate decreased from 152 per 1000 live births to 50 per 1000 live births during the same period [2][8].

The health of neonates and children under five remains a significant concern in developing countries, and the highest rates of neonatal and child mortality are still in Sub-Saharan Africa [1][2][3]. On average 109 children per 1000 live births die before the age of five whereas in developed countries, only 7 children per 1000 live births die before the age of 5 and where 34 neonates per 1000 live births die before 28 days of life, whereas only 4 neonates per 1000 live births die before 28 days of life in developed countries [1]. Every year 2.6 million neonates die; three fourths of these deaths occur in the first week of life, and almost all (99%) in low- and middle-income countries [8]. Despite global achievements, only one-third of the priority countries in sub-Saharan Africa reached their child mortality target [2][8][11][12]. This achievement gap translated into an estimated 2.9 million child deaths in 2015 in sub-Saharan Africa, of which two-thirds were likely from preventable causes and 45% occurred in the neonatal period [8].

Without a significant reduction of infection-related neonatal deaths in LMICs it is unlikely that Sustainable Development Goal (SDG) 3 as one of the Global strategies of WHO [2], aim to reduce maternal and child mortality to at least 12 per 1000 and 25 per 1000 livebirths respectively by 2030 will be met [3][10]. Additionally, the Millennium Development Goal (MDG) 4 to reduce child mortality by two thirds by the end of 2015 resulted in an unprecedented focus on child health and a 52% reduction in sub-Saharan Africa between 1990 to 2015, from 179 to 86 deaths per 1000 live births [1][8][11].

In East Africa, Rwanda has made dramatic improvements in child mortality, moving from 37 neonatal and 152 under-5 deaths per 1000 live births in 2005 to 20 and 50 per 1000 live births, respectively, in 2015 [11]. However, the rate of progress in the neonatal period has been slower, with death in the first month of life accounting for 46% of under-5 deaths globally [11].

The main clinical causes of death include prematurity, infection, inadequate management of complications of pregnancy and delivery, and lack of quality care immediately after birth [12][13]. Referring to low-income and middle-income countries (LMICs), Neonatal sepsis is the third leading cause of neonatal mortality, only behind prematurity and intrapartum-related complications (or birth asphyxia), with LMICs bearing the burden of 99% of global neonatal mortality, highlighting global disparity [3][8][12]. Neonates, especially preterm, are more susceptible to infections than children at any other age period [8].

Neonatal sepsis represents a significant global challenge, ranking as the third leading cause of neonatal mortality, particularly in developing countries [20]. Despite notable advancements in medical practices, the diagnosis and management of neonatal infections continue to present formidable difficulties [22][8]. One of the major complexities arises from the presence of noninfectious conditions that closely resemble sepsis, a concern particularly pronounced in preterm infants [3][10]. Compounding this issue is the absence of reliable diagnostic tests, making it challenging to accurately differentiate between infectious and noninfectious conditions. The impact of neonatal sepsis on mortality rates is profound, accounting for 13% of all neonatal deaths and a staggering 42% of deaths occurring within the first week of life [2]. Indeed, strategies that can prevent and treat neonate with sepsis are essential to accelerate the progress of neonate survival. In many developing country settings, however, the identification and treatment of neonates with infection is unsatisfactory [23].

Identification of risk factors and early institution of therapy thereby can improve neonatal mortality and morbidity [2][11]. Unfortunately, developing countries often lack comprehensive surveillance systems, leading to significant underreporting of neonatal sepsis cases, particularly those that occur outside healthcare facilities. Consequently, the true magnitude of its impact on mortality may be even higher than currently estimated [3]. Neonates, especially preterm infants, exhibit heightened vulnerability to infections compared to children in other age groups.

This susceptibility stems from various factors, including impaired cytokine production, decreased expression of adhesion molecules in neutrophils, and a reduced response to chemotactic factors, all of which compromise the effectiveness of the innate immune system [8]. Given the high-risk nature of neonatal sepsis, clinicians face a challenging dilemma when it comes to treatment. In the absence of definitive diagnostic tests, they are compelled to administer empirical antibiotics to infants displaying risk factors or suspected signs of sepsis [7].

However, this approach comes with several drawbacks. Firstly, the use of broad-spectrum antibiotics can contribute to adverse outcomes, including the disruption of the infant's normal microbial flora, increased risks of antibiotic-associated complications, and the potential for drug-related adverse effects [24].

Secondly, prolonged empirical antibiotic treatment contributes to the global concern of antimicrobial resistance, as it exposes both the individual infant and the broader population to the selective pressure that drives the emergence and spread of resistant pathogens [2][8].

Innate immunity is affected by impaired cytokine production, decreased expression of adhesion molecules in neutrophils and a reduced response to chemotactic factors [8]. Also, transplacental passage of antibodies starts during the second trimester and achieves its maximal speed during the third trimester [8]. As a result, most preterm neonates have significantly reduced humoral responses [25][8]. In developing countries, clinically diagnosed sepsis is present in 49–170 per 1000 live births, culture-proven sepsis in 16 per 1000 live births and neonatal meningitis in 0.8–6.1 per 1000 live births [8][26]. In developing countries, most pathogens isolated in the hospital setting before 72 h of life are similar to those isolated afterward; it is likely that highly unclean delivery practices lead to infections with nosocomial agents very early in life [27]. In addition, most neonate are born at the household and might get infected with community acquired pathogens even after 72 h [8]. As a result, several authors have classified neonatal sepsis in developing countries as community- and hospital-acquired instead of early and late-onset.

Sepsis is defined as a dysregulated host response to infection that results in life-threatening organ dysfunction. Neonatal sepsis is divided into early-onset (if clinical signs start before 7 days of life) and late-onset (if clinical signs start afterward) [2][8][11] and it is caused by factors related to both maternal and neonatal such as pathogens, prolonged rupture of membrane (PROM), urinary tract infection, intrapartum fever, instrumental delivery, prematurity, chorioamnionitis, frequent vaginal examination, never attend antenatal care (ANC), home delivery, meconium-stained amniotic fluid, contaminated foods intake, low birth weight, complicated or instrument-assisted delivery, low appearance pulse grimace activity respiration (APGAR) scores and invasive procedures during hospital admission [8][11][10][28].

Despite the high burden of neonatal sepsis, high-quality evidence such as, lack of randomized controlled trials, research studies, and lack of consensus among experts regarding best practices in early diagnosis and treatment is still lacking [8]. One of the major difficulties in the management of neonatal sepsis is getting an accurate diagnosis. Unlike older patients, neonates have very subtle presentations, and multiple conditions resemble neonatal sepsis [8].

Auxiliary tests have limited value, access to laboratory tests in some developing countries is limited, and are difficult to interpret due to low sensitivity and changing normal ranges during the neonatal period [2][8]. Also, blood cultures also lack sensitivity due to specific characteristics of the neonatal population and blood cell count is difficult to interpret in the neonatal period because it varies significantly with day of life and gestational age [29]. Sepsis shares a similar clinical presentation to other common conditions in the neonatal period such as temperature instability, heart rate =>180 beats/min or <=100 beats/min, and even respiratory rate >60 breaths/min plus grunting or desaturations among others [8][30].

Addressing the challenges surrounding neonatal sepsis requires a multifaceted approach. Enhancing diagnostic capabilities through the development of reliable, rapid, and affordable diagnostic tests is essential [31]. These tests should not only differentiate between infectious and noninfectious conditions but also identify the causative agents and their antimicrobial susceptibility profiles [18]. Additionally, strengthening surveillance systems in developing countries is crucial to capture accurate data on the burden and outcomes of neonatal sepsis. Prevention strategies should also be prioritized, including interventions to improve maternal and neonatal health, such as antenatal care, clean delivery practices, and appropriate umbilical cord care [13][14]. Promoting breastfeeding, immunization programs, and infection control measures in healthcare settings are also vital components of a comprehensive approach. Furthermore, efforts to raise awareness among healthcare providers, parents, and communities about the symptoms and clinical signs of neonatal sepsis can contribute to early detection and timely intervention [32]. Education and training programs should focus on proper antibiotic use, emphasizing the importance of targeted therapy based on accurate diagnosis whenever possible.

In conclusion, neonatal sepsis represents a significant global health challenge, particularly in developing countries. The difficulties in diagnosing and managing neonatal infections, compounded by the absence of reliable diagnostic tests, result in clinicians resorting to empirical antibiotic treatment [8][13][14]. However, the use of broad-spectrum antibiotics and prolonged empirical therapy carry adverse consequences and contribute to the rise of antimicrobial resistance [2]. Addressing these challenges requires a comprehensive approach encompassing improved diagnostics, strengthened surveillance systems, prevention strategies, and increased awareness. By addressing these critical areas, we can work towards reducing the burden of neonatal sepsis, improving outcomes for vulnerable neonates, and combating the global challenge of antimicrobial resistance [28][30]. Also, given the high incidence and mortality of sepsis in preterm infants and its long-term consequences on growth and development, efforts to reduce the rates of infection in this vulnerable population are one of the most important interventions in neonatal care [18].

Early diagnosis, management, and treatment can considerably decrease the risk of neonatal sepsis, and improve the outcome by providing an early and correct diagnosis of neonatal sepsis [28]. Identification of risk factors and timely initiation of treatments can significantly decrease neonatal mortality and morbidity [8].

This is why, there is a need for strategies that can prevent, detect or treat neonates with sepsis are essential to accelerate the progress of neonate survival. So, this project proposal proposes to develop a machine learning-based model for the early diagnosis of neonatal sepsis using clinical and laboratory data. The system will incorporate expert knowledge and clinical guidelines to identify neonates at risk of sepsis and facilitate early intervention [5]. The project proposal has the potential to improve outcomes for neonates with sepsis and contribute to the development of innovative solutions for this pressing healthcare challenge.

The development and implementation of a machine learning based model for neonatal sepsis diagnosis hold significant promise in improving healthcare outcomes for neonates at risk of sepsis [31]. This diagnostic tool would serve as a valuable resource for healthcare providers, enabling them to effectively identify and manage cases of neonatal sepsis. By utilizing clinical data as input, the trained model used the information to generate a diagnosis of either neonatal sepsis or no sepsis [33]. One of the key advantages of this machine learning based model is its potential to enhance neonatal outcomes. By providing healthcare providers with a standardized and reliable diagnostic tool, the model would enable early detection and intervention, leading to prompt initiation of appropriate treatment. This timely response can significantly reduce morbidity and mortality rates associated with neonatal sepsis, ensuring better health outcomes for affected infants [33][34][35]. In addition to improving clinical outcomes, the implementation of a machine learning based model for neonatal sepsis diagnosis can also yield economic benefits. By streamlining the diagnostic process, healthcare costs can be reduced. The model would minimize unnecessary diagnostic tests and interventions, optimizing resource utilization and minimizing healthcare expenses associated with managing suspected cases of neonatal sepsis [35]. Moreover, the model's efficiency in providing accurate diagnoses would help prevent prolonged hospital stays and unnecessary treatments, further reducing the financial burden on healthcare systems.

Integrating the machine learning-based model into existing EMRs is another advantage that enhances its accessibility and usability. This integration will allow healthcare providers to incorporate the diagnostic tool into their routine workflow seamlessly. Regardless of whether all healthcare facilities or systems adopt the model, it can still be effectively integrated into EMRs utilized by healthcare providers who leverage its benefits.

This interoperability ensures widespread availability and accessibility of the model, making it a valuable resource for healthcare professionals in various settings. Furthermore, the machine learning-based model can contribute to increased efficiency in diagnosing and treating sepsis. By automating the diagnostic process, the model reduces the dependence on subjective judgment and variability in diagnostic decision-making. It provides healthcare providers with a standardized approach, ensuring consistency and accuracy in diagnosing neonatal sepsis. This increased efficiency translates into the timely initiation of appropriate treatments, optimizing patient care, and improving overall healthcare delivery.

To sum up, designing and implementing this machine learning based model for early neonatal sepsis diagnosis offers numerous benefits for healthcare providers and neonates at risk of sepsis. This diagnostic tool will facilitate early detection and intervention, improving neonatal outcomes. The model's streamlined approach can also reduce healthcare costs and time associated with managing neonatal sepsis. Its integration into existing EMRs ensures widespread accessibility and usability. By enhancing efficiency in diagnosis and treatment, the machine learning-based model holds great potential to positively impact the management of neonatal sepsis and improve the overall quality of care provided to vulnerable neonates.

2.2 Related Work

This section underpinning literature review draws from various domains to construct a comprehensive understanding of the health challenges faced by neonates and children under the age of five in developing countries, particularly in Sub-Saharan Africa. The conceptual framework integrates key elements from neonatal healthcare, global health targets, and the specific context of neonatal sepsis, incorporating both epidemiological and clinical perspectives. Several papers were reviewed that used artificial neural networks to predict sepsis neonatal. In every article, we analyzed the model building process, variable selection, ground truth, training and test datasets, overfitting avoidance, error estimate, and AUC ROC (Area Under the Curve) information.

Cristhyne Le´on1 [6]. Presents Recurrent Neural Networks for Early Detection of Late Onset Sepsis in Premature Infants Using Heart Rate Variability. The dataset used consisted of 259 premature infants; 193 of them were used for training the model, which was then tested in the remaining 66 infants. Thus, they obtained an area under the receiver operating characteristics curve (AUROC) of more than 80% for the 24 hours before the onset of the infection, and reaching 90.4% (95% CI [88.1%, 92.6%]) six hours before the time of the infection.

Fernando López-Martínez [36]. The purpose of this study is to develop a non-invasive neural network classification model for early neonatal sepsis detection.

The data used in this study is from Crecer's Hospital Center in Cartagena-Colombia. An imbalanced dataset of 555 neonates with (66%) of negative cases and (34%) of positive cases was used for this study. The study results show a sensitivity of 80.32%, a specificity of 90.4%, precision on the positive predicted value of 83.1% in the test sample and a calculated area under the curve of 92.5% (95% Confidence Interval [91.4 - 93.06]).

Dž. Gojak* [9]. A database of 1,000 data was used to develop the artificial neural network, of which 200 were healthy and 800 were sick. The obtained training effect was 98.93%, sensitivity was 98.75%, specificity was 95.50% and accuracy was 98.33%.

Addy Cecilia Helguera-Repetto [20]. A predictive model was obtained by training and validating an artificial Neural Networks (ANN) algorithm with a balanced dataset consisting of preterm and term non-septic or septic neonates (early- and late-onset), with negative and positive culture results, respectively, using 25 maternal and neonatal features. The outcome of the model was sepsis or not. The performance measures of the model, evaluated with an independent dataset, outperformed physician's diagnosis using the same features based on traditional scoring systems, with a 93.3% sensitivity, an 80.0% specificity, a 94.4% AUROC, and a regression coefficient of 0.974 between actual and simulated results.

2.2.1 Global Health Disparities

The literature review begins by establishing a theoretical foundation in global health disparities, emphasizing the persistent challenges faced by developing countries, especially in Sub-Saharan Africa. The stark contrast in neonatal and child mortality rates between developed and developing regions serves as a critical backdrop. This draws on established theories within global health literature that highlight socioeconomic, infrastructural, and healthcare access factors contributing to health disparities.

2.2.2 Global Health Goals

The Sustainable Development Goals (SDG) framework, particularly SDG 3 focused on maternal and child health, serves as a theoretical anchor. The inclusion of specific targets related to neonatal and child mortality rates sets the stage for examining the progress, gaps, and challenges in achieving these global health objectives. The Millennium Development Goal (MDG) 4, with its historical context, provides an additional layer to the understanding of global efforts to reduce child mortality.

2.2.3 Specific Focus on Neonatal Sepsis

Within the broader context, the theoretical framework narrows its focus to neonatal sepsis as a critical determinant of neonatal mortality. The inclusion of theoretical concepts related to infectious diseases, immunology, and epidemiology elucidates the intricate dynamics of neonatal sepsis. This section builds on existing theories that underscore the vulnerability of neonates to blood infections and the unique challenges in diagnosing and managing neonatal sepsis.

2.2.4 Impact of Health Systems and Interventions

Theoretical considerations extend to the impact of health systems and interventions on neonatal health. The framework weaves theoretical concepts related to healthcare infrastructure, surveillance systems, and the effectiveness of interventions. This involves drawing on health system strengthening theories, the role of preventive strategies, and the socio-cultural factors influencing healthcare practices.

2.2.5 Machine Learning and Healthcare

This section introduces concepts from the fields of machine learning and artificial intelligence. Drawing from literature in medical informatics, this section explores how machine learning applications can contribute to healthcare improvement. Theoretical underpinnings related to the integration of technology, diagnostic accuracy, and the potential economic benefits are integrated.

2.2.6 Challenges and Ethical Considerations

The theoretical considerations extend to the challenges associated with neonatal sepsis diagnosis and the ethical dimensions of implementing technology-driven solutions. This involves drawing on theories related to medical ethics, equity in healthcare access, and the socio-cultural implications of deploying advanced technologies in resource-constrained settings.

2.3 Summary

In exploring existing literature, profound insights have been gained regarding the critical challenges surrounding neonatal and child health in developing nations, specifically Sub-Saharan Africa. The global health disparities are starkly evident, with alarmingly high rates of neonatal and child mortality persisting despite global initiatives. Theoretical frameworks have illuminated the intricate dynamics of neonatal sepsis, emphasizing its status as a formidable contributor to mortality, particularly in resource-constrained settings. The extensive literature review highlights the urgent need for targeted interventions. The global health goals, including the Sustainable Development Goals (SDG) 3 and the historical context of the Millennium Development Goal (MDG) 4, provide a framework for understanding the ambitious targets and the significant gaps in achieving them. The progress witnessed in countries like Rwanda serves as a beacon of hope, demonstrating the transformative potential of dedicated efforts.

The theoretical foundation addresses the challenges in diagnosing neonatal sepsis, emphasizing the lack of evidence-based practices and the difficulty of differentiating between infectious and noninfectious conditions. This highlights the critical role of theoretical frameworks and practical insights, which are essential for comprehending the issue. Machine learning presents a promising opportunity to improve diagnostic accuracy and neonatal outcomes. Moving forward, the goal is to transform these insights into actionable strategies.

By integrating theoretical frameworks with practical insights from the literature, we aim to develop a machine learning-based model for early neonatal sepsis diagnosis. Many existing models in the literature lack some of the important evaluation metrics, such as validation accuracy, F1 score, and AUC ROC, which are crucial for assessing overfitting and biases.

Without these metrics, it is difficult to evaluate the true performance of these models, leading to potential overfitting and biased results. Therefore, an approach that includes these evaluations is necessary to provide a more robust and reliable assessment of model performance. Our objective is to contribute to academic discourse and make significant progress in addressing urgent healthcare issues faced by neonates in resource-limited settings. The insights gained in this chapter inspired us with a strong sense of purpose and a commitment to translating knowledge into impactful solutions in the subsequent chapters.

CHAPTER 3. RESEARCH METHODOLOGY

This chapter serves as the compass guiding the study's journey. It unveils the methodology, exploring the system design intricacies, tools for data collection, and the art of designing a machine learning model. The choice of design methodologies is justified, providing a theoretical foundation. Data collection methods and tools are dissected, ensuring precision aligns with neonatal sepsis research goals. The chapter culminates in the practical realm with insights into prototyping, bringing the proposed solution to life.

3.1 Research Process

This sub-chapter delves into a thorough review of existing research works closely aligned with our pursuit of designing a machine learning-based model for neonatal sepsis diagnosis. By critically engaging with prior studies, we aim to build upon established knowledge, identify gaps, and refine our research questions. This process not only validates the significance of our work but also positions our study within the broader context of advancements and challenges in neonatal healthcare and machine learning applications. As we navigate the troves of existing research, we leverage the wisdom of those who have treaded similar paths, enriched our methodology and ensured a robust foundation for innovation.

3.1.1 The Research Plan

The development of a machine learning based model for early diagnosis of neonatal sepsis involved several procedures and techniques to identify, collect, and analyze data.

3.1.1.1 Study Design and Ethical Considerations

The first step was to design a study plan that outlined the objectives, scope, and ethical considerations of the research. This involved obtaining an ethical clearance from the College of Medicine and Health Sciences (CMHS) institutional review board (IRB) that ensured compliance with data protection regulations.

3.1.1.2 Data Identification and Collection

The data used in this study were obtained from two different databases, which consisted of Neonatal Intensive Care Unit (NICU) records from Kabgayi and Ruhengeri Level II Teaching Hospitals, including a total of 1381 neonate records, where 659 neonates' data were collected from Kabgayi Level II Teaching Hospital and 722 from Ruhengeri Level II Teaching Hospital, collected in the year 2023 with both hospitals approval, their Research Committee, and the CMHS Institutional Review Board. Each record consisted of summaries of hourly clinical signs, some laboratory values, and demographic variables. Specifically, the data contained 25 clinical variables: 8 clinical signs variables, 14 laboratory variables, and 6 demographic variables.

3.1.1.3 Data Preprocessing and Cleaning

Python version 3.10.11, operating on a 64-bit system within Visual Studio Code version 1.88.1, was used to develop this machine-learning model aimed at early predicting neonatal sepsis. Then, data preprocessing and cleaning were used to prepare the dataset containing neonates' medical records for further analysis and modelling. The essential data manipulation, visualization, and machine learning libraries were imported. The NumPy library was utilized for numerical computations, while Pandas was employed for data manipulation tasks, such as loading the dataset from an Excel file and handling data frames. Matplotlib was used for data visualization, enabling the creation of various plots and charts. Additionally, sci-kit-learn libraries, including train_test_split, MinMaxScaler, OneHotEncoder, Pipeline, Column Transformer, and SMOTE, are imported to facilitate data preprocessing tasks such as feature scaling, encoding categorical variables, handling missing values, and addressing class imbalance.

Data preprocessing followed, where the neonatal clinical data dataset was loaded from an Excel file. The subsequent step involves handling missing values in the dataset. Specifically, columns and rows with missing data were identified and removed from the dataset using the drop function. Additionally, datetime columns were dropped as they may contribute little to the predictive model. Then, categorical variables were identified, and if any non-numeric values were present, they were encoded using one-hot encoding to convert them into a numerical format suitable for machine learning algorithms. This process ensures that the dataset is free from missing values and that all features are in a suitable format for model training.

3.1.1.4 Feature Selection and Engineering

After completing the data preprocessing and cleaning steps, we began the process of feature selection and engineering. This phase was crucial for identifying relevant features that contribute to the diagnosis of neonatal sepsis and enhancing the model's predictive power through feature engineering techniques.

First, numerical and categorical columns were defined based on domain knowledge and the characteristics of the datasets used in model development. **Numerical columns included**:

- **Age of the neonates**: Represented by the number of days since birth.
- **Blood oxygen saturation**: Indicates the percentage of oxygen saturation in the blood, a critical measure of respiratory function, with values above 95% considered normal for neonates.
- **Pulse**: Refers to the heart rate measured in beats per minute, an indicator of cardiovascular health, with normal values ranging between 80 and 180 bpm.
- **Respiratory rate**: Measures the number of breaths taken per minute, with normal values ranging between 30 and 60 breaths per minute.
- **Temperature**: Body temperature measured in degrees Celsius, with normal values ranging between 36.5°C and 38°C.
- **Weight**: Categorized as below 1.5 kg, between 1.5 kg and 2.5 kg, and above 2.5 kg.

These numerical columns represent clinical signs and measurements. Categorical columns included **clinical symptoms**, which indicate specific medical conditions and signs such as fever, jaundice, seizure, vomiting, respiratory distress, nursing strike, rhinorrhea, diarrhea, and poor feeding.

Following the identification of features, we defined a preprocessing pipeline using Scikit-learn's Pipeline and Column Transformer. This pipeline encapsulated the preprocessing steps for numerical and categorical features separately. For numerical features, a MinMaxScaler was applied to scale the values to a range between 0 and 1, ensuring all features contribute equally to the model. For categorical features, a OneHotEncoder with the parameter drop='first' was used to encode categorical variables into binary vectors, enabling the model to interpret them properly.

3.1.1.5 System Development and Training

In this phase, the implementation of a machine learning model using the keras library in Python, which provides a user-friendly interface for building and training neural networks was developed. The cleaned and combined dataset with a total number of 1381 neonates was split into training, validation, and testing sets using the train test split function from sci-kit-learn to facilitate the model development and evaluation process, ensuring that the model's performance can be assessed on unseen data. A train size was 966, validation size 315 and lastly, a test size of 24% which is equal to 100 neonates.

During model training, the Adam optimizer with a learning rate of 0.01 was utilized. The training process was executed over 20 epochs with a batch size of 19. Additionally, the model utilized early stopping with a patience of 5 epochs to prevent overfitting. After that, a neural network model architecture was defined using the Keras functional API. The model architecture includes 5 layers of neurons, each with specified activation functions such as ReLU (Rectified Linear Unit) for intermediate layers with 512, 256, and 128, number of neurons and a single neuron and sigmoid for the output layer, which is suitable for binary classification tasks. Additionally, batch normalization layers are incorporated to improve the model's generalization and prevent overfitting by reducing internal covariate shifts and randomly dropping neurons during training, respectively.

Once the model architecture was defined, the model is compiled using the Adam optimizer with a specified learning rate and binary cross-entropy loss function, which is suitable for binary classification tasks. Additionally, the model was configured to optimize for accuracy during training and evaluation. The training process commenced by calling the fit method on the model object, passing the resampled training data (X_train_resampled and y_train_resampled) as input. During training, the model iteratively adjusts its parameters using gradient descent algorithms to minimize the loss between predicted and actual outcomes on the training data. The training process is monitored using an early stopping callback, which stops training if the validation loss does not improve for 5 number of epochs, helping to prevent overfitting.

3.1.1.6 System Implementation and Performance Evaluation

Following this training, the model is further evaluated using a separate testing dataset to judge the performance of the model based on a suite of evaluation metrics, as indicated below. These were specificity, sensitivity, accuracy, precision, recall, F1 score, and ROC-AUC score, calculated as a measure to evaluate the model performance in predictive analytics. The accuracy measured the proportion of the number of correctly classified instances to the instances considered while Precision measured the proportion of true positive predictions among all the instances classified as positive, which actually revealed the ability to avoid false alarms. Recall is a definition rate of the actual positive instances against the positive predictions; it means how much the model predicts those positive instances that have really occurred in the data. The F1 score is the harmonic mean of precision and recall. In a sense, it balances the tradeoff between precision and recall, giving a single metric with overall performance. The ROC-AUC score represented the area under the curve of the receiver operating characteristic, summarizing the model's ability to discriminate between the positive and negative instances at a different threshold level. All this was done on both training and testing data.

3.1.1.7 Documentation and Reporting

A thoroughly documented report was diligently compiled, encompassing a comprehensive narrative of the project's progression. This narrative includes the exploration of preliminary discoveries, the iterative refinement of the development process, meticulous testing stages, and, finally, the definitive outcomes stemming from the evaluation of the model's performance.

3.2 Research Design Method

This is a retrospective cohort study that looked into clinical data for all neonates born and hospitalized in the year 2023 at Kabgayi and Ruhengeri Level II Teaching Hospitals, located in the Southern Province and Northern Province of Rwanda. Its main objective was to come up with a machine learning model for the early prediction of neonatal sepsis.

The model considered fine-tuning its accuracy and precision by including expert advice and adherence to clinical guidelines on the identification of risk to the neonates and others among the metrics, which included clinical signs of neonates as predictors and gestational age and birth weight as two of the confounders. In an effort to minimize bias, data-cleaning processes, gauged using considerations for statistical power, and effect size in the calculations were applied. Processing and grouping data for the analysis stage have been done appropriately. Data summarization in this case was descriptive, using the statistical data summaries. That is, the model performance was carried out through a rigorous test and benchmarking with the currently existing diagnostic methodologies. Most importantly, the whole process follows the ethical approval protocols.

3.2.1 System Design Flowchart

This flowchart (Figure 1) provides a high-level overview of the entire process used to develop the machine learning model, from data loading and preprocessing to model training, evaluation, testing, and prediction. Each step is represented with a specific operation where some are combined, and arrows indicate the flow of the process.

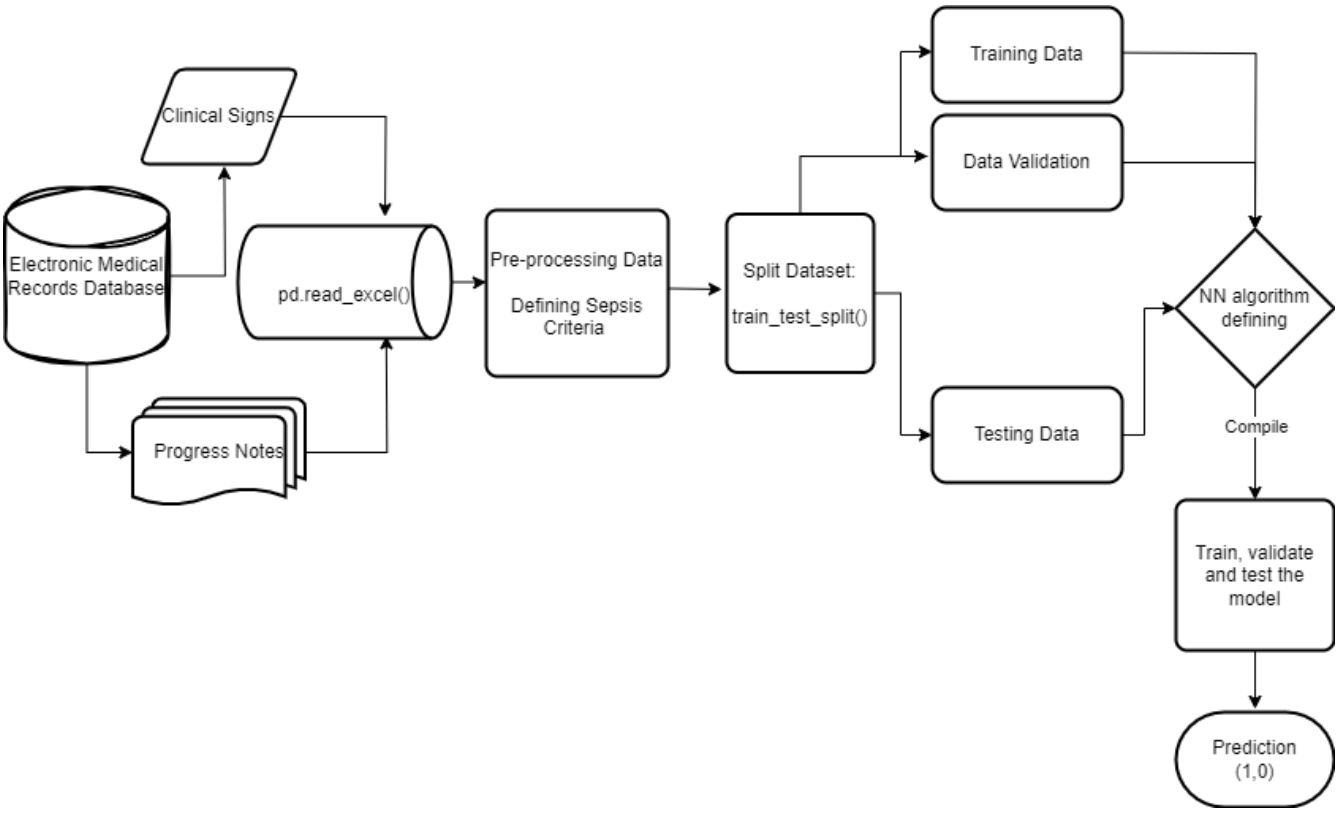


Figure 1: system design flowchart

3.2.2 Dataset

After obtaining authorization from Kabgayi and Ruhengeri Level II Teaching Hospitals, the next step was to engage with both hospitals' research committees to discuss the data usage and integrity protocols. Following this, they liaised with the hospital's IT department to acquire the dataset containing neonatal patient records for the year 2023 as it is shown in table 1 below. Certain variables were redacted from the dataset to maintain confidentiality, including neonates' names, demographic details, and healthcare personnel names. The datasets were provided in two parts: one containing clinical indicators such as body temperature, weight, pulse, respiratory rate, and blood oxygen saturation, while the other included complaints and symptoms like jaundice, vomiting, and excessive crying.

After merging the two datasets to consolidate records for individual neonates, manual cleaning was performed to rectify writing errors and missing information, and we discarded rows and columns with missing data. Consequently, the resulting dataset comprised 1381 neonatal records, retaining 19 out of the initial 25 variables. Among those variables, the ones that were considered in the model designing were system patient ID, blood oxygen saturation, clinical symptoms, height, pulse, respiratory rate, temperature, and weight which make a total of 15 variables.

Table 1: Datasets used for the model training, validation and testing

N^0	Hospital	Period (2023)	Dataset Collected	Dataset Used	Variables (Features)
1	Kabgayi Level II Teaching H.	Jan-December	22,	659	15
2	Ruhengeri Level II Teaching H.	Jan-June	35,634	722	15

3.2.3 ANN development

Development of the machine learning model was developed using the Artificial Neural Network algorithm, which includes the importation of all required libraries, i.e., NumPy, pandas, scikit-learn, Matplotlib, imbalanced-learn (imblearn), and TensorFlow, in that these libraries include the manipulation of data, preprocessing, the building of the model, and evaluation. The clinical records of this neonatal dataset are loaded using Pandas from an Excel file. Further, the first phase of data pre-processing is done, where the missing values are handled, and the dataset is prepared for modelling. This includes the function to drop the rows or columns having missing data, encoding the categorical variables using OneHotEncoder, and defining columns of numeric and categorical type. After preprocessing, data will be split into features (X) and target variables (y) to be used for training and testing the model. Handling an imbalance in the class distribution will be done through the Synthetic Minority Over-sampling Technique (SMOTE) to the training data. We then use the Keras API of TensorFlow to define the architecture for the neural network model.

The model was made of a total of 5 dense layers, and the activation functions are rectified linear unit (ReLU), with batch normalization layers following each dense layer to address the dropouts in the model so as not to result in overfitting. At the output layer, a sigmoid activation function was used in producing the binary classification predictions. After defining the model architecture, it's then compiled with an Adam optimizer and binary cross-entropy loss function. In this case, a stopping factor has been introduced to check on or prevent overfitting at training. The model training was done on the resampled training data, and validation was done as well on validation dataset.

Consequently, after training, it undertakes the model performance assessment with the following measures of accuracy, precision, recall, F1 score, and ROC-AUC score.

Such metrics gave a closer view of the predictive power and overall performance of the model regarding different aspects of classification accuracy. Next, model evaluation metrics were printed to the console, and bar plots and histograms for the analyzed model performance are plotted. The testing dataset was then predicted, while further analysis based on some criteria, like identifying the neonates who were predicted to have sepsis and plotting their status, was done. Overall, the ANN development involved model building, data preprocessing, training, evaluation, and analysis, and therefore developed an effective predictive model for neonatal sepsis. The final NN architecture as shows in figure 2 consisted of an input layer with 15 neurons equals to the total features used, followed by three hidden layers, and an output layer. The first hidden layer has 512 neurons, the second has 256 neurons, and the third has 128 neurons. Each hidden layer uses the ReLU (Rectified Linear Unit) activation function, which introduced non-linearity by setting negative values to zero and passing positive values unchanged, thereby helping the network learn complex patterns more effectively. Batch Normalization was applied to each hidden layer to standardize inputs, which stabilized and accelerated training by reducing internal covariate shift. The output layer comprised of 2 neurons with a Sigmoid activation function, suitable for binary classification tasks as it maps outputs to a probability range between 0 and 1.

This architecture, with its combination of ReLU, Sigmoid, and Batch Normalization, and learning rate of 0.01 ensured effective learning and accurate predictions.

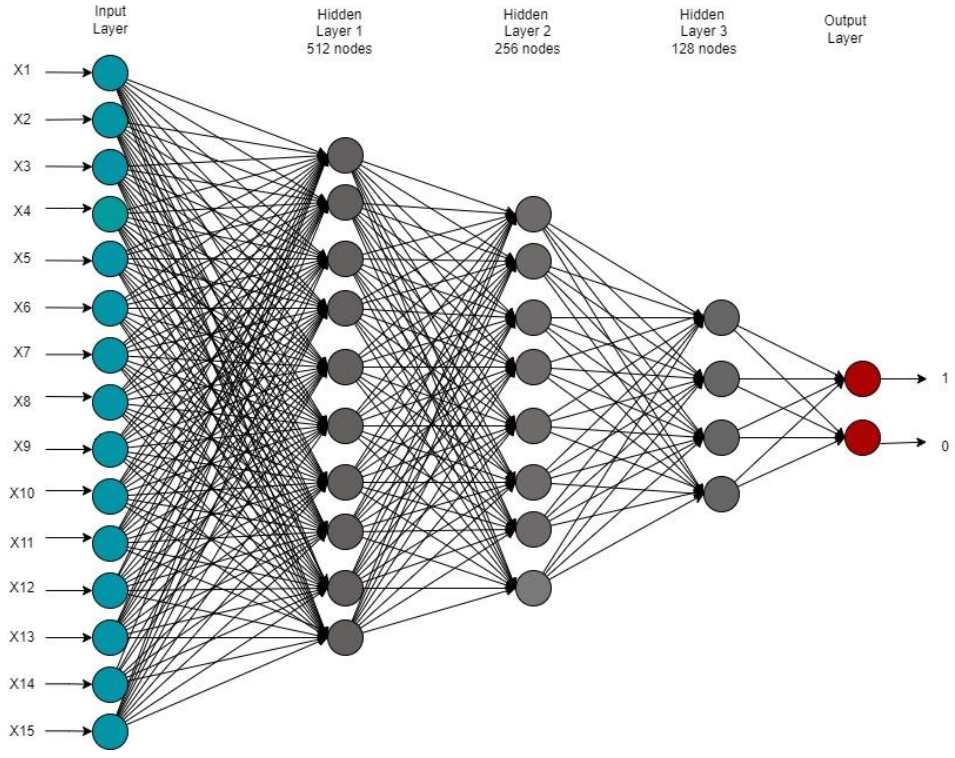


Figure 2: ANN model for predicting neonatal sepsis.

3.3 Summary

In selecting a method for this research, a comprehensive review of various approaches related to the study on early diagnosis of neonatal sepsis was conducted. After careful consideration, the specific research method identified was machine learning with a supervised learning task, particularly utilizing artificial neural networks (ANN). This involved collecting data from Kabgayi and Ruhengeri Level II Teaching Hospitals, to be used in training a neural network algorithm. Subsequently, the algorithm undergoes testing, and generates an output predicting whether a neonate tends to have sepsis or not using the system patient ID.

The choice of Kabgayi and Ruhengeri Level II Teaching Hospitals as the primary research sites was made following a thorough examination of neonatal sepsis cases in Rwandan hospitals, considering factors such as the quality of data recording practices and number of patients admitted there. Kabgayi and Ruhengeri Level II Teaching Hospitals emerged as the preferred locations due to their robust data recording mechanisms and number of neonates admitted in these hospitals such as Gisenyi, and Rwinkwavu (These hospitals were in the chosen list of hospitals, we were considering for data collection). This decision underscores the importance of pragmatic considerations in research planning and execution, ensuring a balance between the research objectives and the practical constraints that may arise during the study. The 15 months' timeline and the allocated budget of 580,000 Rwf were designed to accommodate the specific needs of this research, taking into account the unique circumstances surrounding neonatal sepsis data collection and analysis at Kabgayi and Ruhengeri Level II Teaching Hospital.

CHAPTER 4. THE PROJECT RESULTS FROM SIMULATION AND/OR IMPLEMENTATION

4.1 The output from the testing set of the machine learning model

The updated and purposely collected data adequately trained, tested, and validated the artificial neural network (ANN) model in most instances for the detection of early neonatal sepsis, focusing more on different aspects of its performance and the power of prediction it can make. The training process comprised of 20 epochs, with each epoch taking approximately 1 second apart from the first one which took around 8 seconds to complete. During testing as shows in figure 3, the model gave a test accuracy of about 85.00%, where precision, recall, and F1-score had different performance inclinations in the classification of the sepsis cases on evaluation using the testing data set. On the other hand, the precision from the model for sepsis prediction was about 85.18%, which implies the model has a moderate false-positive rate. The recall was about 86.79%, meaning it could capture a big portion of the true positive cases. The F1 score, a harmonic mean of precision and recall, was kinder good at approximately 85.98%, suggesting a good performance of the model. The ROC-AUC score, which is a measure of the ability to discriminate, had an approximate score of 84.88%. Apart from this, during the evaluation process, it also produced a classification report in a clear and distributed way for the performance of the model into positive and negative classes.

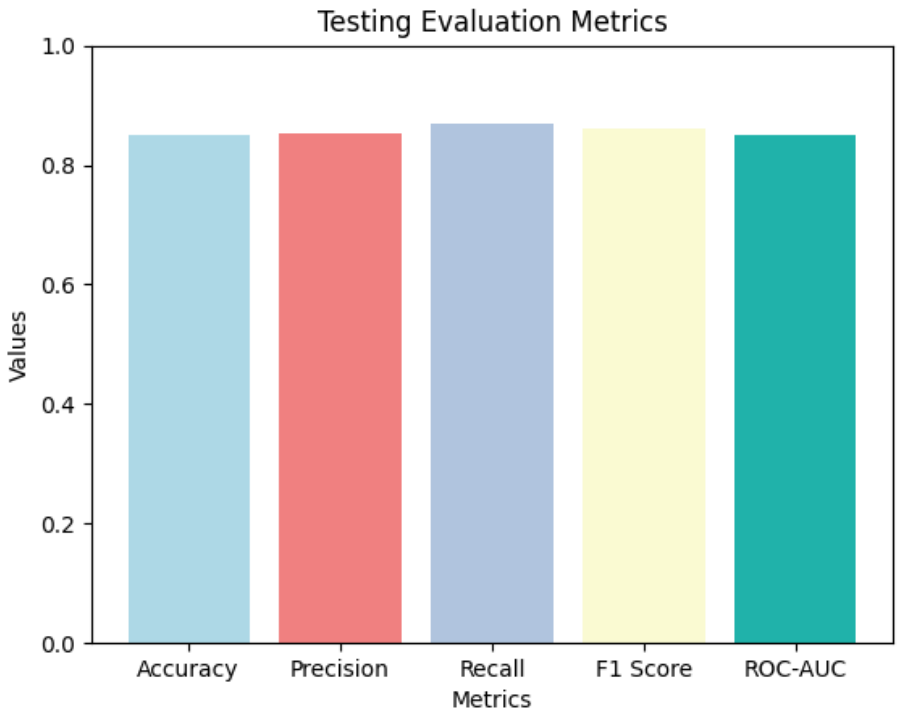


Figure 3: Evaluated Metrics results

Figure 4 represents a SHAP (SHapley Additive exPlanations) visualizations, the colors and dots represent different aspects of feature contributions to model predictions. Red and Purple dots represent higher feature values with the most impact on the model prediction while blue dots represent lower feature values. These colors indicate lower values of the feature. The blue dots show how the model's prediction changes when the feature value is low.

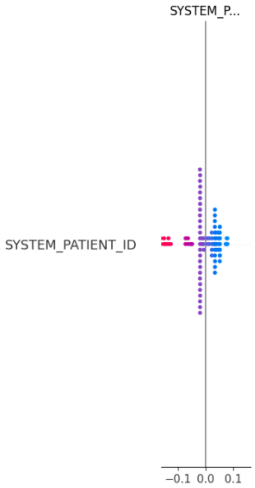


Figure 4: SHAP (SHapley Additive exPlanations) visualizations, X-Axis: Represents the feature value while Y-Axis: Represents the SHAP value.

Also, the outputs showed the model predictions and corresponding System Patient IDs to be used in further analysis of the predicted sepsis cases. The training and overall evaluation process took quite a while since training epoch times varied from 12 to 1 second for each; while testing, approximately 3ms was taken per step. Overall, the model presented a fair performance in the detection of neonatal sepsis, but it needs necessary improvement, especially precision and F1 score metrics. Further refinement and optimization of the model may have further polished clinical utility and predictive accuracy. Figure 5 shows neonates predicted to have sepsis and those who are not likely to have neonatal sepsis.

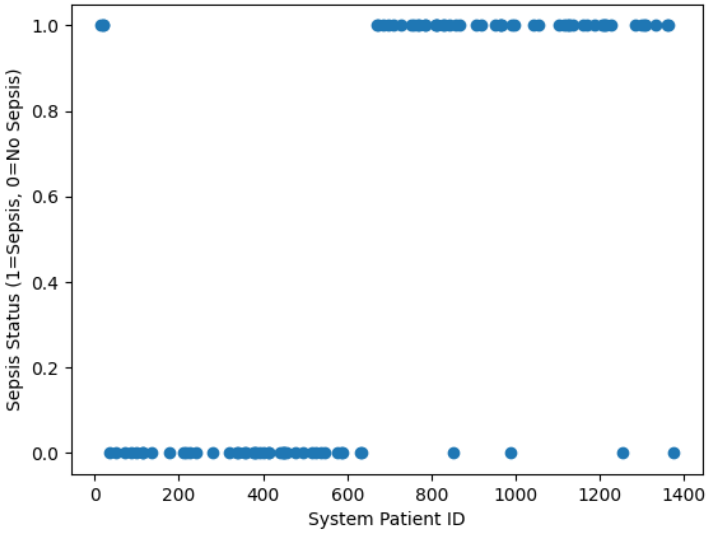


Figure 5: Histogram of System Patient ID VS their Status

Out of 100 instances in the testing dataset, 54 neonates were predicted to have sepsis (class 1), while the minority, 46 neonates, did not have sepsis (class 0). A total number of 966 neonates were used in the training of the model while 314 neonates' data were used for validation. The following are the patients' IDs and their corresponding numbers which are predicted to have neonatal sepsis.

Also figure 6 represents sensitivity and specificity which provides a visual representation of a classification model's performance across different threshold settings. Sensitivity, also known as recall or true positive rate, measures the model's ability to correctly identify positive cases while specificity, on the other hand, measures the model's ability to correctly identify negative cases. While testing, the model showed Sensitivity of around 86.79% and Specificity of 82.98%, suggesting a good performance of the model in differentiation true positive from true negative. These two parameters were calculated according to the following equations:

$$\boldsymbol{Sensitivity\ (Recall)} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP+FN})}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{(\boldsymbol{TN+FP})}$$

Where: TP – true positive, TN – true negative

FP – false positive, FN – false negative

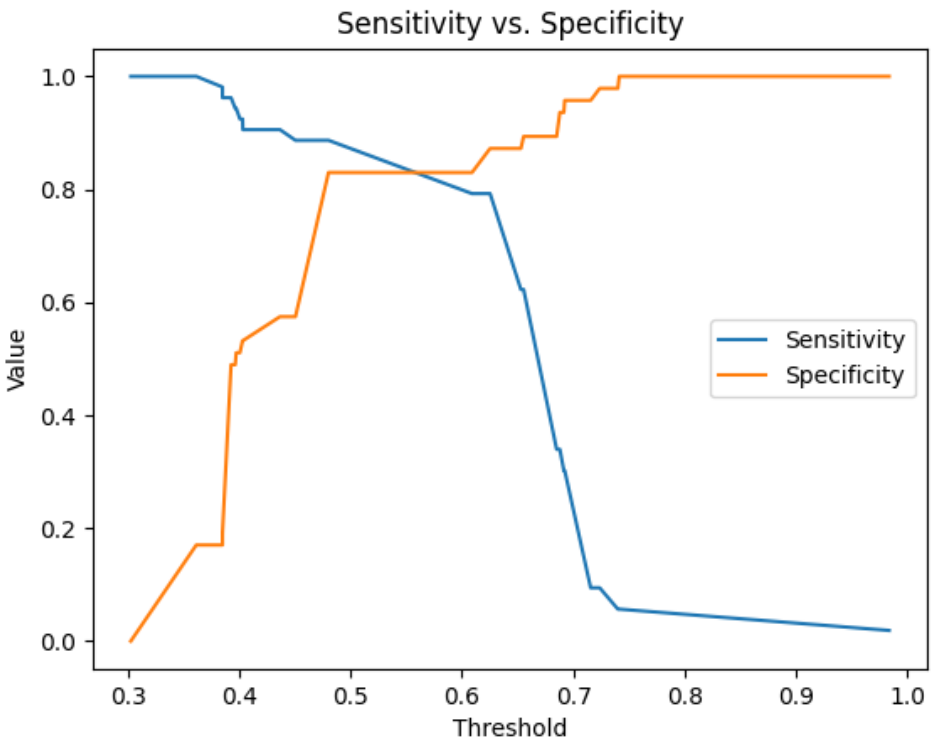


Figure 6: A Plot generated from the output showing the sensitivity and specificity of the model during testing

Table 2 shows how the accuracy varied with the loss over the total number of 20 epochs used during training, the early stopping used during the training since there is the `**Early Stopping**` callback caused the testing set to stop at 16 epochs. This callback monitors the validation loss metric, and because it did not show any improvement after 5 consecutive epochs the training was stopped, which made the total epochs to be 16 instead of 20 to meet the criteria set. This technique was used to prevent overfitting and unnecessary computation. This helped in visualizing how the model was performing during the training and validation phase in the first 10 epochs out of 16 epochs performed.

Table 2: The model's performance over time during training and validation set

N	Epochs	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1.	Epoch	0.5798	0.8787	0.5365	3.1058
2.	Epoch	0.6854	0.6507	0.5270	2.7918
3.	Epoch	0.6839	0.6385	0.5270	1.5527
4.	Epoch	0.6576	0.6637	0.6254	1.0694
5.	Epoch	0.6843	0.6779	0.6508	0.8338
6.	Epoch	0.6705	0.6313	0.6254	0.8593
7.	Epoch	0.6855	0.6461	0.6508	0.7104
8.	Epoch	0.6940	0.6210	0.6286	0.7987
9.	Epoch	0.6882	0.6336	0.6444	0.7707
10.	Epoch	0.6933	0.6141	0.6190	0.8000

The model's performance over 16 epochs demonstrated significant learning and generalization improvements, as seen by the gradual increase in training accuracy from 57.98% to approximately 73.09% and a consistent decrease in training loss from 87.87% to 61.41%. Initially, the validation accuracy was low of about 53.65%, but it improved significantly, stabilizing around 70% in later epochs as well, indicating a better generalization to unseen data. Additionally, the validation loss exhibited a good progress towards the end, which suggested little change of overfitting.

Figure 7 below shows a graph on how the model loss and accuracy kept on varying over the 16 epochs performed.

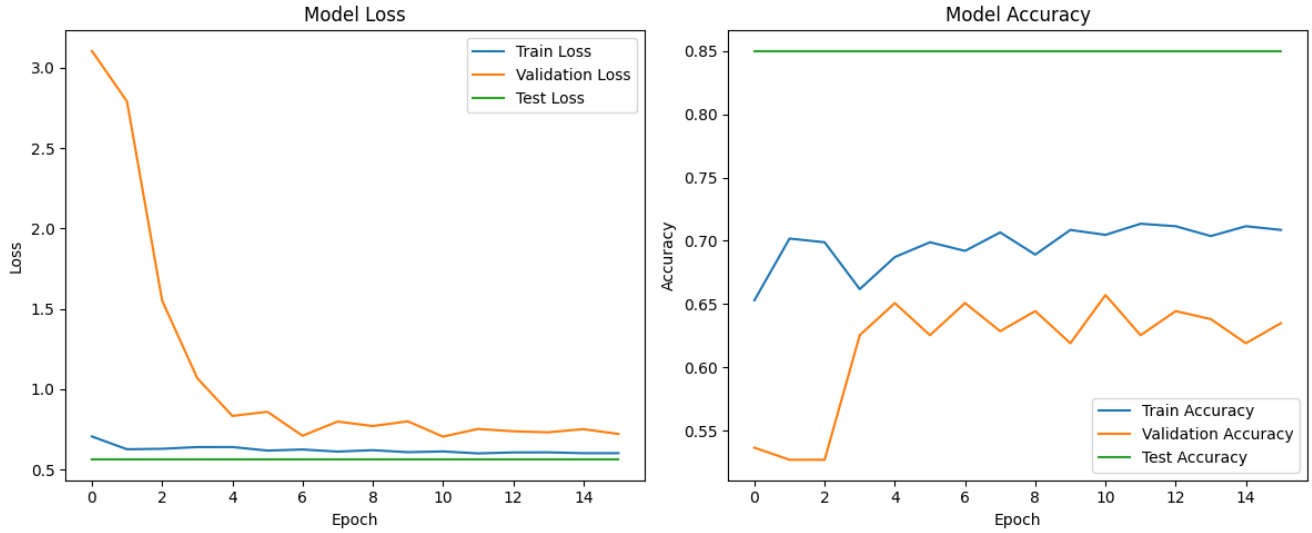


Figure 7:Model loss vs Model accuracy during training, validation, and testing phases

The ROC-AUC score, which is a measure of the ability to discriminate, had a score of 73.09% during training while during testing it scored 84.88%. Figure 8 below shows the curve showing the performance of the model compared to how it differentiates true positive from false positive during testing.

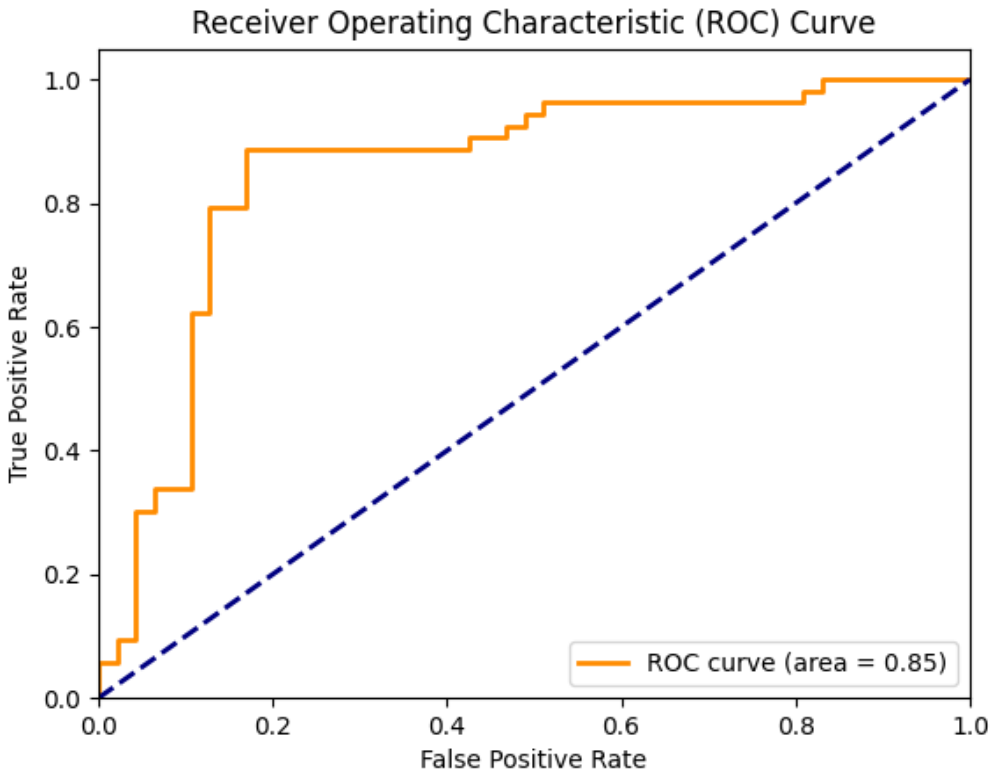


Figure 8: Receiver Operating Characteristic (ROC) Curve' of how the model differentiates true positive from false positive

Figure 9 shows the confusion matrix which is a key evaluation metric for classification problems. It provided a detailed breakdown of how well the model's predictions match the actual labels.

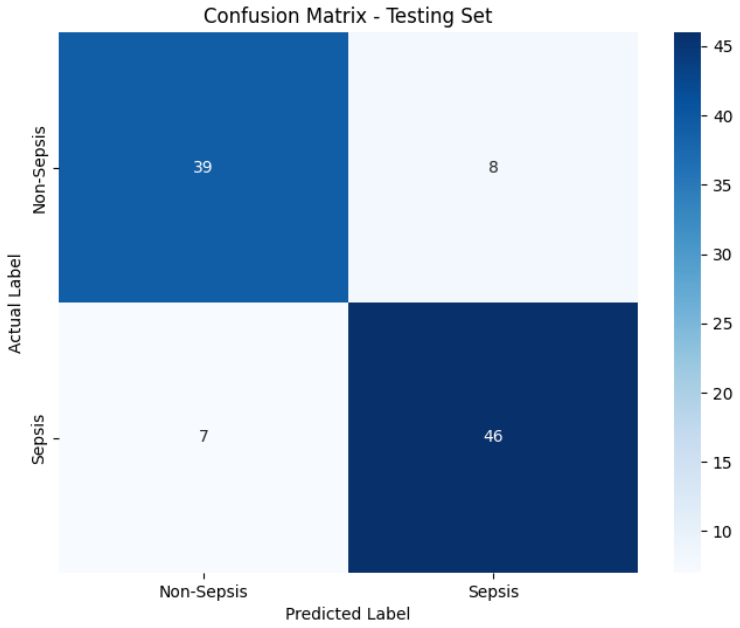


Figure 9: Confusion matrix to evaluate the classification performance

4.2 Summary

This model was cross-validated with the testing dataset, showing approximately 85.00% accuracy for the overall model's performance. The precision of the model in detecting sepsis was at 85.18%, meaning that among the detected cases, the ones corresponding to true positives are more compared to those that correspond to other cases. This, therefore, translated to 86.79 % recall of sepsis, a figure that generally indicates that the model can pick approximately 85.00% of the actual sepsis cases. This result was able to highlight a precision value of around 85.98 % with an F1 score, thus indicating that indeed the model had capability in discriminating true sepsis from non-sepsis cases. Generally, the ability to realize accurate predictions concerning the occurrence of cases of sepsis is somehow actualized by this model, though it has attained a score of ROC-AUC of about 84.88%, which is an indication of capability of the model to discrimination neonatal sepsis cases.

CHAPTER 5. CONCLUSION AND RECOMMENDATION

5.1 Limitations and Strengths

This research project faced data collection challenges that slowed whole process down. Some crucial details, like hourly information, mothers' information throughout the pregnancy stage and during labor, and lab tests, were missing from the datasets. Additionally, the data included entries in multiple languages (e.g., French) and used different terminologies, making data cleaning difficult. Also, the fact that in the datasets collected from both hospitals, neonates with vital signs which are likely to fit those predicted to have neonatal sepsis were moderate compared to those not included. These limitations prevented the model from being on the same scale as the existing ones and it cannot integrate with the existing electronic medical record (EMR) system at the current stage. Despite these challenges, the research demonstrated significant potential. The model's innovative approach and adaptability show promise for future advancements. With comprehensive optimization and the incorporation of a more extensive and diverse dataset, the model is expected to overcome current limitations and provide valuable contributions to neonatal sepsis diagnosis. The resilience shown in addressing these data-related obstacles underscores the model's capacity for growth and improvement, setting a strong foundation for impactful future developments in this critical area.

5.2 Conclusion

This study presented the development of a machine learning-based model for early prediction of neonatal sepsis using datasets from Kabgayi and Ruhengeri Level II Teaching Hospitals, located in Rwanda. We used a neural network algorithm. The model showed a comprehensive approach leading to valuable insights, from data collection, algorithm development, model training, and rigorous evaluation to continuous improvement. On the other hand, the evaluation's output with our model indicates complexities and challenges in developing an accurate tool for neonatal sepsis diagnostics. Despite significant efforts toward its improvement, the model's performance on the testing dataset yielded moderate precision and recall for sepsis detection. Improvement can be predicted through this project for neonatal healthcare outcomes, but implementation is still fine-tuning. It provided a good space for this project to be a real pace-setter in developing countries for replication to fight against neonatal mortality and antimicrobial resistance and a good starting point for researchers whose focus is in improving healthcare through non-invasive diagnosis of diseases.

5.3 Recommendations

It will only demand more from the continuous improvement of machine learning techniques and further reinforcement of interdisciplinary collaborations that will be critically needed for the research community to grapple with the new health challenge, neonatal sepsis. Using further novel data sources, algorithm optimization, and model interpretability could improve the predictive system's fit within resource-constrained settings, such as Kabgayi and Ruhengeri Level II Teaching Hospitals in Rwanda. Open sharing of the datasets and methodologies will further help for the broader validation and adoption of such innovations in varied healthcare settings. To the end-users, it prioritizes investments in technology-enabled solutions for early disease detection and prevention. Further integration of such predictive models within the clinical routine workflows and further evidence-based practices shall greatly help achieve better neonatal health outcomes within the region, not forgetting the reduction of the current high morbidity and mortality rates. In addition, for the system to remain effective and sustainable within the clinical context of the hospital, there must be continuous determination of the effectiveness and necessary adaptations according to the feedback obtained from the actual application.

5.4 Future plan

We plan to focus on optimizing the current neonatal sepsis prediction model to ensure its seamless integration into the EMRs systems used in hospitals across the country and also beyond in the whole region. This will involve extensive hyperparameter tuning and model refinement to enhance accuracy and robustness, coupled with rigorous validation through cross-validation techniques. To ensure practical applicability, we opt to collaborate with healthcare professionals to tailor the model to clinical workflows and conduct pilot implementations in different hospitals. Additionally, we aim to publish these findings in peer-reviewed journals and at different medical and engineering conferences in order to reach a broad audience of researchers, engineers, and healthcare providers among others. By engaging in interdisciplinary collaboration and actively seeking feedback from end-users, we shall work towards refining the model and promoting its adoption, ultimately improving patient outcomes and advancing the field of neonatal care.

REFERENCES

[1] J. June, "MINISTRY OF HEALTH," no. July 2013, 2018.

[2] A. M. Kasie, M. Kassaw, and M. A. Getu, "Neonatal sepsis and its associated factors in East Africa : a systematic review and meta-analysis , 2019," pp. 1–42, 2019.

[3] J. T. Id *et al.*, "Neonatal sepsis at Mulago national referral hospital in Uganda : Etiology , antimicrobial resistance , associated factors and case fatality risk," vol. 46, pp. 1–14, 2020, doi: 10.1371/journal.pone.0237085.

[4] H. Attia Hussein Mahmoud *et al.*, "Insight Into Neonatal Sepsis: An Overview," *Cureus*, vol. 15, no. 9, pp. 1–12, 2023, doi: 10.7759/cureus.45530.

[5] D. P. Ezeobi, "A Neonatal Sepsis Prediction Algorithm Using Electronic Medical Record (EMR) Data from Mbarara Regional Referral Hospital (MRRH)," pp. 1–30, 2022.

[6] C. Leon, P. Pladys, A. Beuchee, and G. Carrault, "Recurrent Neural Networks for Early Detection of Late Onset Sepsis in Premature Infants Using Heart Rate Variability," *Comput. Cardiol. (2010).*, vol. 2021-Septe, pp. 1–4, 2021, doi: 10.23919/CinC53138.2021.9662715.

[7] Y. Y. Kuo, S. T. Huang, and H. W. Chiu, "Applying artificial neural network for early detection of sepsis with intentionally preserved highly missing real-world data for simulating clinical situation," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12911-021-01653-0.

[8] A. Zea-vera and T. J. Ochoa, "Challenges in the diagnosis and management of neonatal sepsis," pp. 1–13, 2015, doi: 10.1093/tropej/fmu079.

[9] D. Gojak *et al.*, "The use of artificial intelligence in the diagnosis of neonatal sepsis," *IFAC-PapersOnLine*, vol. 55, no. 4, pp. 62–67, 2022, doi: 10.1016/j.ifacol.2022.06.010.

[10] R. Milton *et al.*, "Articles Neonatal sepsis and mortality in low-income and middle- income countries from a facility-based birth cohort : an international multisite prospective observational study," pp. 661–672, 2022, doi: 10.1016/S2214-109X(22)00043-2.

[11] N. Gupta *et al.*, "Causes of death and predictors of childhood mortality in Rwanda : a matched case-control study using verbal social autopsy," pp. 1–9, 2018.

[12] M. Nyishime *et al.*, "A retrospective study of neonatal case management and outcomes in rural Rwanda post implementation of a national neonatal care package for sick and small infants," pp. 1–11, 2018.

[13] A. R. Ndayizeye, E. Sibomana, I. Nyaziyose, C. J. Conard, and P. Cartledge, "Open access Neonatal antibiotic use at a district and teaching hospital in Rwanda – a retrospective descriptive study," vol. 76, no. 2, pp. 1–6, 2019.

[14] O. Urayeneza *et al.*, "Increasing evidence - based interventions in patients with acute infections in a resource - limited setting : a before - and - after feasibility trial in Gitwe , Rwanda," *Intensive Care Med.*, vol. 44, no. 9, pp. 1436–1446, 2018, doi: 10.1007/s00134-018-5266-x.

[15] "Acta Paediatrica - 2023 - Honor - Vital sign-based detection of sepsis in neonates using machine learning.pdf."

[16] M. A. Adewole K. S., Hambali and M. K. Jimoh, "Rule-Based Expert System for Disease Diagnosis," *Proc. Int. Conf. Sci. Technol. Educ. Arts, Manag. Soc. Sci. (iSTEAMS Res. Nexus).*, no. July, pp. 183–190, 2014, [Online]. Available: isteams.org

[17] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–10, 2020, doi: 10.1038/s41746-020-0221-y.

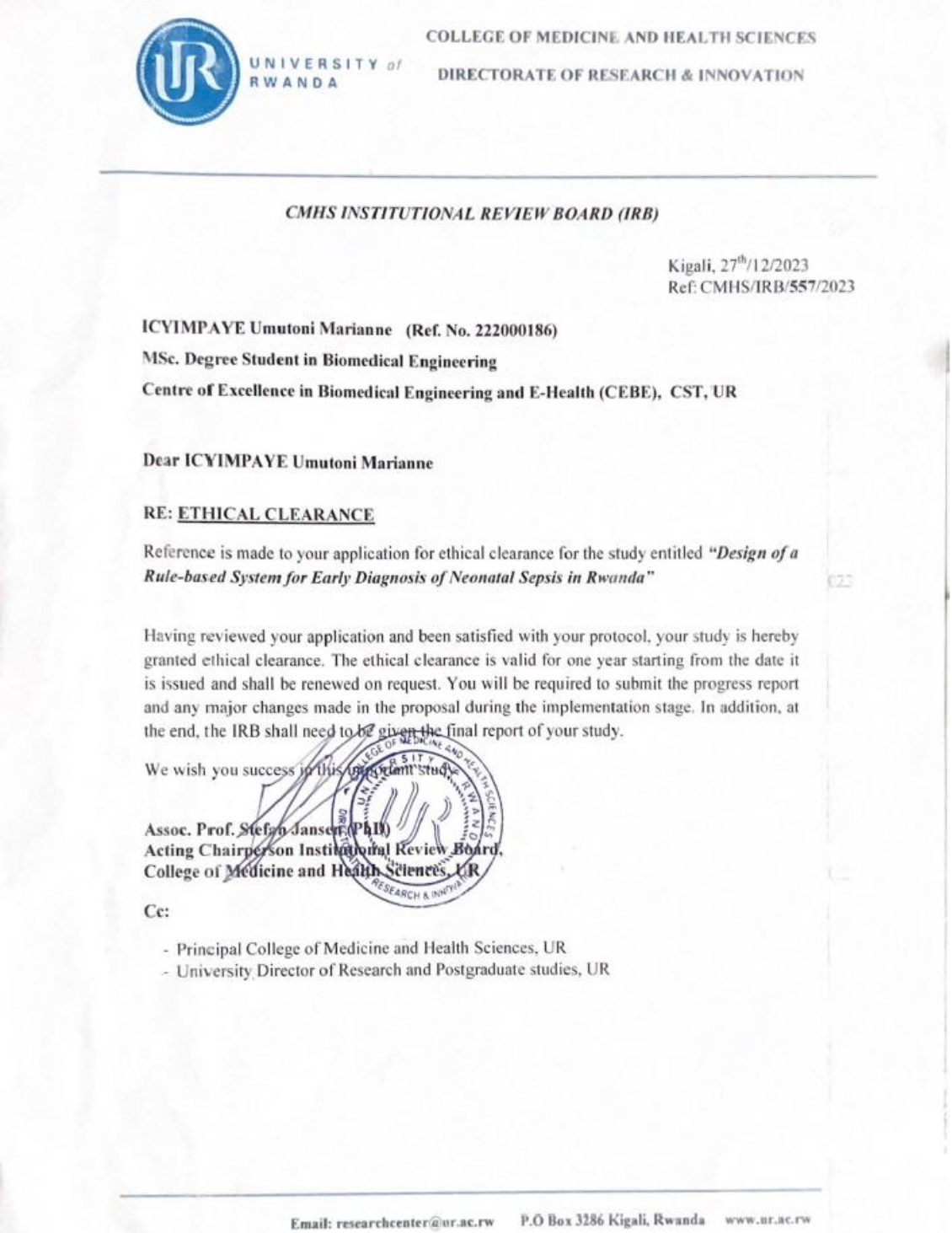
[18] D. A. H. Id *et al.*, "Sepsis in two hospitals in Rwanda : A retrospective cohort study of presentation , management , outcomes , and predictors of mortality," pp. 1–15, 2021, doi: 10.1371/journal.pone.0251321.

[19] E. Mfateneza, P. C. Rutayisire, E. Biracyaza, S. Musafiri, and W. G. Mpabuka, "Application of machine learning methods for predicting infant mortality in Rwanda: analysis of Rwanda

demographic health survey 2014–15 dataset," *BMC Pregnancy Childbirth*, vol. 22, no. 1, pp. 1–13, 2022, doi: 10.1186/s12884-022-04699-8.
[20] A. C. Helguera-Repetto *et al.*, "Neonatal Sepsis Diagnosis Decision-Making Based on Artificial Neural Networks," *Front. Pediatr.*, vol. 8, no. September, 2020, doi: 10.3389/fped.2020.00525.
[21] A. J. Hamilton *et al.*, "Machine learning and artificial intelligence: Applications in healthcare epidemiology," *Antimicrob. Steward. Healthc. Epidemiol.*, vol. 1, no. 1, pp. 1–6, 2021, doi: 10.1017/ash.2021.192.
[22] M. S. Edwards, "Clinical features and diagnosis of sepsis in term and late preterm infants," *InUptodate,Kim,MS(Ed),UptoDate,Waltham, MA*, pp. 1–21, 2024.
[23] H. S. F. Fraser *et al.*, "User Perceptions and Use of an Enhanced Electronic Health Record in Rwanda With and Without Clinical Alerts: Cross-sectional Survey," *JMIR Med. Informatics*, vol. 10, no. 5, pp. 1–22, 2022, doi: 10.2196/32305.
[24] P. K. Aksoy, "A Decision Support System on Artificial Intelligence Based Early Diagnosis of Sepsis," 2022.
[25] P. T. Cartledge, F. S. Ruzibuka, F. Rutagarama, S. Rutare, T. Rogo, and P. T. Cartledge, "Antibiotic prescribing practices in three neonatology units in Kigali , Rwanda . – an observational study," vol. 20, no. 4, pp. 1646–1654, 2020.
[26] I. M. Stefanovic, "Neonatal sepsis," *Biochem. Medica*, vol. 21, no. 3, pp. 276–281, 2011, doi: 10.36106/ijsr/2409139.
[27] A. C. Helguera-Repetto *et al.*, "Neonatal Sepsis Diagnosis Decision-Making Based on Artificial Neural Networks," *Front. Pediatr.*, vol. 8, pp. 1–19, 2020, doi: 10.3389/fped.2020.00525.
[28] R. R. Hospital, E. Nimukuze, B. Mukarwego, and B. Ge, "Trends and Factors Associated with Neonatal Sepsis at," vol. 8, no. 3, 2021.
[29] Y. G. Robi and T. M. Sitote, "Neonatal Disease Prediction Using Machine Learning Techniques," *J. Healthc. Eng.*, vol. 2023, 2023, doi: 10.1155/2023/3567194.
[30] H. Magge *et al.*, "The All Babies Count Initiative : Impact of a Health System Improvement Approach on Neonatal Care and Outcomes in Rwanda," pp. 505–517, 2020.
[31] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," *Healthc.*, vol. 10, no. 3, pp. 1–19, 2022, doi: 10.3390/healthcare10030541.
[32] C. M. Bech *et al.*, "Risk factors for neonatal sepsis in Sub-Saharan Africa: A systematic review with meta-analysis," *BMJ Open*, vol. 12, no. 9, 2022, doi: 10.1136/bmjopen-2021-054491.
[33] D. Wendimu and K. Biredagn, "Developing a knowledge-based system for diagnosis and treatment recommendation of neonatal diseases Developing a knowledge-based system for diagnosis and treatment recommendation of," *Cogent Eng.*, vol. 10, no. 1, 2023, doi: 10.1080/23311916.2022.2153567.
[34] W. Song, S. Y. Jung, H. Baek, and C. W. Choi, "A Predictive Model Based on Machine Learning for the Early Detection of Late-Onset Neonatal Sepsis : Development and Observational Study Corresponding Author :," vol. 8, pp. 1–16, doi: 10.2196/15965.
[35] S. F. Feleke, B. Mulu, M. Azmeraw, D. Temesgen, and M. Dagne, "Clinical Prediction Model Development and Validation for the Detection of Newborn Sepsis , Diagnostic Research Protocol," no. November, pp. 8025–8031, 2022.
[36] F. López-Martínez, E. R. Núñez-Valdez, J. Lorduy Gomez, and V. García-Díaz, "A neural network approach to predict early neonatal sepsis," *Comput. Electr. Eng.*, vol. 76, pp. 379–388, 2019, doi: 10.1016/j.compeleceng.2019.04.015.

APPENDICES

Appendix 1: Mode of Data Collection (1. Ethical clearance, 2. Approval letter, 3. Amendment letter, and Confidentiality)



UNIVERSITY of RWANDA

COLLEGE OF MEDICINE AND HEALTH SCIENCES

DIRECTORATE OF RESEARCH & INNOVATION

CMHS INSTITUTIONAL REVIEW BOARD (IRB)

Kigali, 27th/12/2023
Ref: CMHS/IRB/557/2023

ICYIMPAYE Umutoni Marianne (Ref. No. 222000186)
MSc. Degree Student in Biomedical Engineering
Centre of Excellence in Biomedical Engineering and E-Health (CEBE), CST, UR

Dear ICYIMPAYE Umutoni Marianne

RE: ETHICAL CLEARANCE

Reference is made to your application for ethical clearance for the study entitled ***"Design of a Rule-based System for Early Diagnosis of Neonatal Sepsis in Rwanda"***

Having reviewed your application and been satisfied with your protocol, your study is hereby granted ethical clearance. The ethical clearance is valid for one year starting from the date it is issued and shall be renewed on request. You will be required to submit the progress report and any major changes made in the proposal during the implementation stage. In addition, at the end, the IRB shall need to be given the final report of your study.

We wish you success in this research study.

Assoc. Prof. Stefan Jansen (PhD)
Acting Chairperson Institutional Review Board,
College of Medicine and Health Sciences, UR

Cc:

- Principal College of Medicine and Health Sciences, UR
- University Director of Research and Postgraduate studies, UR

Email: researchcenter@ur.ac.rw P.O Box 3286 Kigali, Rwanda www.ur.ac.rw

KABGAYI DIOCESE

MINISTRY OF HEALTH
KABGAYI LEVEL 2 TEACHING HOSPITAL
B.P: 66 GITARAMA-RWANDA
E-mail: info@kabgayihospital.rw
: kabgayi.hospital@moh.gov.rw
: kabgayihospital1@gmail.com

Kabgayi, March 6th, 2024.

Ref. : N° 498/HOP/MJB/bcm

To ICYIMPAYE UMUTONI Marianne

Re : Acceptance letter.

Dear Marianne,

I am writing this letter in order to let you know that we have received your letter in which you requested the permission to collect data at Kabgayi Hospital for research purpose "**Design of a machine learning based sytem for early diagnosis of neonatal sepsis at Kabgayi District Hospital**".

In fact, Kabgayi Hospital accepts your request and allows you to collect data. So, you are welcome to conduct your research but you have to meet our Research Committee for more information.

Best regards,

HOPITAL KABGAYI CESAR

Dr. MUVUNYI Jean Baptiste
Director General of Kabgayi Hospital

CC :
- Research Committee

KABGAYI DIOCESE

MINISTRY OF HEALTH
KABGAYI LEVEL 2 TEACHING HOSPITAL
B.P: 66 GITARAMA-RWANDA
E-mail: info@kabgayihospital.rw
: kabgayi.hospital@moh.gov.rw
: kabgayihospital1@gmail.com

Health Management Information System Unit

USERNAME AND PASSWORD AGREEMENT

This agreement sets out the terms and conditions on which username and passwords are provided to OpenMRS/EMR Users. By requesting a EMR user name and accessing information, data or materials contained in EMR databases, you hereby agree to accept the following terms and conditions:

Terms and Conditions:

1. Usernames and passwords must not be shared with anyone. Sharing of IDs is not acceptable as individual accounts can be provided.

2. It is the responsibility of the user who requested the username for the EMR to:
    a. **Inform hospital management for any changes**. Users of EMR agree to discontinue use of their account immediately upon termination from their employer and to notify Hospital IT to close the account at such time.
    b. **Identify and report data quality issues**: While all users and program staff seeks to ensure that data are of the best quality possible, in the course of detailed secondary analysis it is likely that users will note anomalies due to missing data or errors of data transcription and data entry. Data users are required to report such issues to the department responsible for the specific data source so that appropriate action can be taken to correct errors before publishing analyses that might be influenced by these issues.
    c. **Ensure data security and integrity:** Authorized users become the custodians of the data that they are given to use, they must ensure that the data are not shared with others who are not included in the data access agreement. They must not intentionally falsify data and ensure that the integrity of the data is maintained while it is in their possession.

3. **Security of passwords is the responsibility of the user**. Passwords should be changed regularly and chosen in a suitable manner as to prevent unauthorized access to the Account. Clients are urged to change temporary passwords immediately on receipt. Passwords should contain numbers, letters and non-alphanumeric characters. Passwords should not contain personal information which is easily guessed by others such as birth dates, anniversary dates and phone numbers.
4. The user understands that:

UNIVERSITY of RWANDA

COLLEGE OF MEDICINE AND HEALTH SCIENCES

DIRECTORATE OF RESEARCH & INNOVATION

CMHS INSTITUTIONAL REVIEW BOARD (IRB)

Kigali, 12th /June/2024
No 487/CMHS IRB/2024

Icyimpaye Umutoni Marianne
Msc Biomedical Engineering
Centre of Excellence in Biomedical Engineering and E- Health (CEBE), CST, UR

Re: Amendment of ethical clearance

Dear

We thank you for submitting your request for research project amendments in the project titled ***"Design of a Machine Learning based Model for Early Diagnosis of Neonatal Sepsis in Rwanda***

After reviewing your request, the amendments have been approved as follows:

- The data collection sites have been extended to include Ruhengeri Referral Hospital in addition to Kabgayi Level II Teaching Hospital.
- The sample size has been increased from 659 to 5000.
- The sampling strategy will involve collecting data from the electronic medical records of neonates born in 2023.from Both hospitals

We wish you success in this important study.

Assoc.Prof. Stefan LANSEN
Ag Chairperson Institutional Review Board,
College of Medicine and Health Sciences, UR
Cc:

- Principal College of Medicine and Health Sciences, UR
- University Director of Research and Innovations, UR

REPUBLIC OF RWANDA MINISTRY OF HEALTH	RUHENGERI LEVEL TWO TEACHING HOSPITAL Po.Box: 57, MUSANZE Website : rrh.gov.rw ruhengeli.hospital@moh.gov.rw	Client centered Service Integrity Teamwork Innovation

Musanze, 08 JUL 2024

Ref. 755 /RL2TH /DG/2024

ICYIMPAYE UMUTONI Marianne
Tel: 0788721793

Re: **Your request for data Collection**

Dear ICYIMPAYE;

Reference is made to your letter dated on 24th June , 2024 applying permission for data collection of the research project entitled: ***"Design of a machine learning-based model for early diagnosis of neonatal sepsis in Rwanda, Case study: Ruhengeri Level Two Teaching Hospital".***

We have the pleasure to inform you that you are allowed to conduct the above mentioned research project .However you're obliged to have all the required equipments for use and the final project report will be shared with Ruhengeri Level II Teaching Hospital.

Best regards.

Dr MUHIRE Philbert
Director General of Ruhengeri Level Two Teaching Hospital
Cc:

-Chair of ethic committee

Appendix 2: The Utilized Codes

```
# Importing all the necessary libraries
import os
import random
import numpy as np
import pandas as pd
import shap
import tensorflow as tf
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, OneHotEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
roc_auc_score, roc_curve, classification_report, confusion_matrix
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from imblearn.over_sampling import SMOTE
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, Dense, Dropout, BatchNormalization
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping, ModelCheckpoint

# Set seeds for reproducibility
def set_seeds(seed=22):
    os.environ['PYTHONHASHSEED'] = str(seed)
    random.seed(seed)
    np.random.seed(seed)
    tf.random.set_seed(seed)

# Ensure deterministic behavior
def set_deterministic():
    os.environ['TF_DETERMINISTIC_OPS'] = '1'
    os.environ['TF_CUDNN_DETERMINISTIC'] = '1'

set_seeds()
set_deterministic()

# Loading the merged Neonatal dataset
try:
    dataset = pd.read_excel(r'E:\Local Disk D\Neonatal sepsis dataset\kabgayi and
ruhengeri_merged_neonates_data.xlsx')
except FileNotFoundError:
    print("Error: File not found. Please check the file path.")
    raise
except Exception as e:
    print("An unexpected error occurred:", str(e))
    raise

# Performing initial data preprocessing by dropping columns and rows with missing data
```

```
missing_columns = dataset.columns[dataset.isnull().any()].tolist()
dataset.drop(missing_columns, axis=1, inplace=True)
datetime_columns = ['ENCOUNTER_DATETIME']
dataset.drop(columns=datetime_columns, inplace=True, errors='ignore')
dataset.dropna(inplace=True)

# Reset the index
dataset.reset_index(drop=True, inplace=True)

# Identify columns with non-numeric values
non_numeric_columns = dataset.select_dtypes(exclude=['float64', 'int64']).columns

# Define non-numeric categorical columns excluding the problematic column
categorical_columns = [col for col in non_numeric_columns if col !=
'ASPERGILLUS_BRONCHITIS___1F20.14']

# Add 'CLINICAL_SYMPTOMS' to categorical_columns if it's not already present
if 'CLINICAL_SYMPTOMS' not in categorical_columns:
    categorical_columns.append('CLINICAL_SYMPTOMS')

# Extract sub-columns containing symptoms
clinical_symptom_columns = [col for col in dataset.columns if 'CLINICAL_SYMPTOMS' in
col]

# Create dummy variables for sub-columns
for column in clinical_symptom_columns:
    dataset = pd.get_dummies(dataset, columns=[column])

# Identify remaining non-numeric categorical columns
remaining_categorical_columns = [col for col in categorical_columns if col not in
clinical_symptom_columns]

# Create dummy variables for remaining categorical columns
for column in remaining_categorical_columns:
    dataset = pd.get_dummies(dataset, columns=[column])

# Defining numerical columns
numerical_columns = ['DAYS', 'BLOOD_OXYGEN_SATURATION', 'PULSE',
'RESPIRATORY_RATE', 'TEMPERATURE_(C)_1', 'WEIGHT_(KG)_1']

# Defining preprocessing pipeline
numeric_transformer = Pipeline(steps=[
    ('scaler', MinMaxScaler())
])

categorical_transformer = Pipeline(steps=[
    ('encoder', OneHotEncoder(drop='first'))
])

preprocessor = ColumnTransformer(transformers=[
```

```
    ('num', numeric_transformer, numerical_columns),
    ('cat', categorical_transformer, remaining_categorical_columns)
])

# Defining sepsis criteria
def calculate_sepsis_criteria(data):
    # Define signs that may indicate neonatal sepsis
    sepsis_signs = ['Fever', 'Jaundice', 'Seizure', 'Vomiting', 'Respiratory distress', 'Nursing strike',
'Rhinorrhea', 'Diarrhea', 'Poor feeding']

    # Drop the problematic column before encoding categorical columns
    data = data.drop(columns=['ASPERGILLUS_BRONCHITIS___1F20.14'], errors='ignore')

    # Encode categorical columns
    data_encoded = pd.get_dummies(data)

    # Filter columns related to clinical symptoms
    clinical_symptom_columns = [col for col in data_encoded.columns if
'CLINICAL_SYMPTOMS' in col]

    # Conditions for suspected neonatal sepsis based on symptoms and other vital signs
    return (
        (~data_encoded['TEMPERATURE_(C)_1'].between(36.5, 38)) |
        ((~data_encoded['RESPIRATORY_RATE'].between(30, 60)) &
(~data_encoded['PULSE'].between(80, 180)) & (data_encoded['WEIGHT_(KG)_1'] < 1.5)) |
        ((~data_encoded['RESPIRATORY_RATE'].between(30, 60)) &
(data_encoded['BLOOD_OXYGEN_SATURATION'] < 95) &
(data_encoded['WEIGHT_(KG)_1'] < 1.5)) |
        ((~data_encoded['PULSE'].between(80, 180)) &
(data_encoded['BLOOD_OXYGEN_SATURATION'] < 95) &
(data_encoded['WEIGHT_(KG)_1'] < 1.5)) |
        ((~data_encoded['RESPIRATORY_RATE'].between(30, 60)) &
(~data_encoded['PULSE'].between(80, 180)) &
(~data_encoded['WEIGHT_(KG)_1'].between(1.5, 2.5))) |
        ((~data_encoded['RESPIRATORY_RATE'].between(30, 60)) &
(data_encoded['BLOOD_OXYGEN_SATURATION'] < 95) &
(~data_encoded['WEIGHT_(KG)_1'].between(1.5, 2.5))) |
        ((~data_encoded['PULSE'].between(80, 180)) &
(data_encoded['BLOOD_OXYGEN_SATURATION'] < 95) &
(~data_encoded['WEIGHT_(KG)_1'].between(1.5, 2.5))) |
        ((~data_encoded['RESPIRATORY_RATE'].between(30, 60)) &
(~data_encoded['PULSE'].between(80, 180)) & (data_encoded['WEIGHT_(KG)_1'] > 2.5 )) |
        ((~data_encoded['RESPIRATORY_RATE'].between(30, 60)) &
(data_encoded['BLOOD_OXYGEN_SATURATION'] < 95) &
(data_encoded['WEIGHT_(KG)_1'] > 2.5)) |
        ((~data_encoded['PULSE'].between(80, 180)) &
(data_encoded['BLOOD_OXYGEN_SATURATION'] < 95) &
(data_encoded['WEIGHT_(KG)_1'] > 2.5)) |
        (data_encoded[clinical_symptom_columns].apply(lambda row: any(symptom in row.values
for symptom in sepsis_signs), axis=1)))
```