Neonatal Disease Prediction using Machine Learning Techniques

Yohanes Gutema Robi

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

September 2022

Adama, Ethiopia

Neonatal Disease Prediction Using Machine Learning Techniques

Name: Yohanes Gutema Robi

Advisor: Dr. Tilahun Melak

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

September 2022

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled '**Neonatal diseases Prediction using machine learning techniques**" is my original work. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation

Yohanes Gutema _____ _____

Name of student Signature Date

RECOMMENDATION

I/we, the advisor(s) of this thesis, hereby certify that I/we have read the revised version of the thesis entitled "**Neonatal Diseases Prediction using Machine Learning Techniques**" prepared under my/our guidance by **Yohanes Gutema** submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering.

Therefore, I/we recommend the submission of the revised version of the thesis to the department following the applicable procedures.

_____	_____	_____
Major Advisor	Signature	Date
_____	_____	_____
Co-advisor	Signature	Date

APPROVAL SHEET

I/we, the advisors of the thesis entitled "**Neonatal diseases Prediction using machine learning techniques**" and developed by **Yohanes Gutema** hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Major Advisor/Supervisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis by Yohanes Gutema have read and evaluated the thesis entitled **Neonatal Diseases Prediction using Machine Learning Techniques**" and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering

Chairperson	Signature	Date
Reviewer 1	Signature	Date
Reviewer 2	Signature	Date

Finally, approval and acceptance of the thesis proposal are contingent upon the submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	
School Dean	Signature	
Office of Postgraduate Studies, Dean	Signature	

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to Almighty God for helping me from the early stage of this study until its successful accomplishment in such a beautiful manner.

Then I would like to thank my advisor, Dr. Tilahun Melak for his valuable guidance and advice. He inspired me greatly to work on this research. His willingness to motivate me contributed tremendously to this study. Without his genuine and professional guidance, this study would not have been completed. I cannot have enough words to thank you for your dedicated and highly qualified support, and understanding.

I am also indebted to give special thanks to the staff members of Asella Compressive Hospital for their cooperation during the data collection process. Besides, I am so eager to express my warm gratitude to Dr. Gemechu Edae (pediatrician), for their interest, guidance, and commitment to sharing knowledge. It is my pleasure also to express my thanks to all my family members for their encouragement, and support in accomplishing this thesis work.

Finally, I would like to thank sincerely all my friends who helped me with their valuable support during the entire process of my study.

TABLE OF CONTENTS

APPROVAL SHEET iii
ACKNOWLEDGMENTS iv
TABLE OF CONTENTS v
LIST OF TABLES viii
LIST OF FIGURES ix
LIST OF EQUATIONS xi
LIST OF ABBREVIATIONS xii
ABSTRACT xv
CHAPTER ONE 1
1. INTRODUCTION 1
1.1. Background of the Study 1
1.2. Motivation of the study 3
1.3. Statement of the Problems 4
1.4. Research Questions 5
1.5.1. General Objective 5
1.5.2. Specific Objectives 5
1.5. Significance of the Study 5
1.6. Scope and Limitation of the study 6
Scope of the Study 6
Limitation of the study 6
1.7. Organization of the Thesis 6
CHAPTER TWO 8
2. LITERATURE REVIEW 8
2.1 Neonatal Disease and its Global Burden 8
2.2. Common Neonatal Diseases 9
2.2.1. Respiratory Distress Syndrome (RDS) 9
2.2.2. Necrotizing Enter Colitis in the Newborn 10
2.2.3. Sepsis 11
2.2.4. Perinatal Asphyxia 12
2.3. Overview of Machine Learning 13
2.3.1. Supervised Machine Learning 13
2.3.2. Unsupervised Learning 14
2.3.3. Semi-Supervised machine learning 14

2.3.4. Reinforcement Learning 14
2.3.5. Ensemble Learning 14
2.4. Machine Learning in Healthcare Sector 15
2.4.1. Machine Learning Approach in Disease Prediction 16
2.5. Feature Selection 19
2.6. Related Works with Neonatal Disease prediction 21
CHAPTER THREE 27
3. RESEARCH METHODOLOGY 27
3.1. Research Design 27
3.1.1. Problem Identification 28
3.1.2. Data Collection 28
3.2. Neonatal Disease Classification Modeling 29
3.2.1. Dataset Building 29
3.2.2. Data Preprocessing 29
3.2.3. Feature Selection 31
3.2.4. Machine learning Models 32
3.3. Predictive Model Evaluations 32
3.3.2. Performance Evaluation Metrics for Predictive Modeling 34
CHAPTER FOUR 37
4. PROPOSED SOLUTION FOR NEONATAL DISEASES PREDICTION 37
4.1. Proposed Neonatal Disease prediction Model Architecture 37
4.2. Understanding of the Data 38
4.3. Data preprocessing 42
4.3.1. Data Cleaning and Handling Missing Values 42
4.3.2. Handling Categorical Data 43
4.4. Feature Selection 44
Feature Elimination with cross-validation (RFECV) 45
4.5. Machine Learning Model Building 45
4.5.1 Hyperparameter Tuning 47
4.6. Model Evaluation and Prototyping 48
CHAPTER 5 50
5. IMPLEMENTATION AND EXPERIMENTATION 50
5.1. Development Tools 50
5.2. Experimentation Environment 51
5.3. Dataset Description 52

5.4. Preprocessing Implementations 54
5.4.1. Handling missing Values Implementations 54
5.4.2. Implementation of Handling categorical values 55
5.5. Implementation of Feature Selection 55
5.6. Machine Learning Models Implementation 56
5.6.1. Hyperparameter Tuning Implementation 56
5.7. Model Testing and Evaluation Implementation 58
CHAPTER SIX 59
6. RESULTS, EVALUATION, AND DISCUSSIONS 59
6.1. Dataset Class Distributions Results 59
6.2. Feature importance and Selection Result 60
6.3. Model Building 61
6.4. Models Performance and Evaluation Result 62
6.4.1. Performance and Evaluation Result of models with original Features 62
6.4.2. Performance and Evaluation Result of the models with Feature Selection 65
6.5. Hyperparameter Tuning Result 70
6.6. Discussion 71
CONCLUSION, AND FUTURE WORK 74
Conclusion 74
Recommendation and Future Works 75
APPENDIX i
Appendix A1: Samples dataset ii
Appendix B: sample code iii
B1: Stacking ensemble sample code iii
B2: Sample Code for Integrating ML Model with Flask Server iv
Appendix C Classification Reports v

LIST OF TABLES

Table 2-1 Common Neonatal Diseases 13
Table 2-2 Pros and Cons of Machine learning algorithms used frequently 18
Table 3-1 Confusion matrix for four-class classification 36
Table 4-1 Missing Value Imputation 43
Table 4-2 Categorical data 44
Table 5-1 Description of the Tools and Python Packages Used During the Implementation. 50
Table 6-1 Feature Selection Result for each selected Algorithm 61
Table 6-2 SVM, RF, XGB, and stacking models accuracy result without FS 64
Table 6-3 Performance metrics for SVM, RF, and XGB models with original features 65
Table 6-4 Table SVM, RF, XGB, and Stacking and models accuracy result with FS 68
Table 6-5 Performance Metrics for RF and XGB with RFECV models 69
Table 6-6 Summary of Model Evaluation with and without FS techniques 69
Table 6-7 Parameters used in RF as returned by randomized search CV 70
Table 6-8 Parameters used in XGB as suggested by randomized search CV 71
Table 6-9 Performance metric result for SVM model with and without Normalization 71

LIST OF FIGURES

Figure 2:1 Leading causes of death in developing countries 9
Figure 2:2 stacking ensemble flow diagram 15
Figure 2:3 Boosting ensemble flow diagram 15
Figure 2:4 Feature Selection 19
Figure 2:5 Filter Method Flowchart 19
Figure 2:6 Wrapper Method flowchart 20
Figure 2:7 Embedded Method Flowchart 20
Figure 3:1 Flowchart of methodology for neonatal disease classification 27
Figure 3:2 stacking ensemble architecture 32
Figure 3:3 Algorithm for the 10-Fold Cross-Validation 34
Figure 4:1 Proposed Neonatal Disease prediction Model Architecture 38
Figure 4:2 Thermometer shows the body temperature 39
Figure 4:3 Dataset Preprocessing step 42
Figure 4:4 Feature Selection workflow diagram 44
Figure 4:5 Model building flow diagram 45
Figure 4:6 Decision function for linearly separable problem 46
Figure 4:7 Deployment Architecture for Neonatal disease prediction 49
Figure 5:1 Implementation of loading dataset using panda's library 54
Figure 5:2 Implementation of Handling missing values 54
Figure 5:3 Handling Categorical Data 55
Figure 5:4 Splitting Dataset into Dependent and Independent features 55
Figure 5:5 Applying Stratified 10-fold CV on Models 58
Figure 6:1 Class distribution of four-class 59
Figure 6:2 distribution of features in dataset 60
Figure 6:3 Feature Importance Using Random Forest 60
Figure 6:4 Confusion Matrix for SVM without (left) and with Normalization (right) 62
Figure 6:5 Confusion Matrix for RF without (left) and with Normalization(right) 63
Figure 6:6 Confusion Matrix for XGB without (left) and with Normalization(right) 64
Figure 6:7 Confusion Matrix for SVM with RFECV without and with Normalization 66
Figure 6:8 Confusion Matrix for RF with RFECV without and with Normalization 66
Figure 6:9 Confusion Matrix for XGB with RFECV without and with Normalization 67
Figure 6:10 Confusion matrix for Stacking ensemble with FS 67

Figure 6:11 Learning curve representing training and validation scores *68*
Figure 6:12 Model Evaluation Results *70*
Figure 6:13 GUI for Neonatal Diseases Diagnosis prototype *73*

LIST OF EQUATIONS

Equation 3:1 Feature normalization Formula.....*31*
Equation 3:2. Accuracy Formula.....*35*
Equation 3:3 Precision Formula.....*35*
Equation 3:4 Recall Formula.....*35*
Equation 3:5 F1-score Formula *35*

LIST OF ABBREVIATIONS

A

AI: Artificial Intelligence

ANN: Artificial Neural Network

APGAR: Activity, Pulse, Grimace, Appearance, Respiration

ACH: Asella Compressive Hospital

C

CRP. C-Reactive Protein

CSV · Common Separated Value

CV · Cross Validation

D

DT · Decision Tree

E

EDHS · Ethiopian Demographic and Health Survey

F

FMOH · Federal Minster of Health

FS · Feature Selection

G

GA. Gestational Age

GUI: Graphical User Interface

I

ICD - International Classification of Diseases

ICSCR**:** Intercostal Sub Costal Retractions

IEEE · International Electrical and Electronics Engineering

U

UFS · Univariate Feature Selection

XGB: Extreme Gradient Boosting

L

LDA · Linear Discriminant Analysis

LLVW: Low Lung Volumes and Whiteout

LR · Linear Regression

M

MDG. Millennium Development Goal

ML · Machine Learning

N

NB Naïve Bayes

NEC. Necrotizing Enter colitis

NICU. Neonatal Intensive Care Unit

P

PA · Parental Asphyxia

R

RDS. Respiratory Distress Syndrome

RF · Random Forest

RFE · Recursive Feature Elimination

RFECV · Recursive Feature Elimination with cross-validation

RL · Reinforcement Learning

RMSE · Root Mean Square Error

RR: Respiratory Rate

S

SDG. Sustainable Development Goal

SFS · Sequential Forward Selection

SMOTE: Synthetic Minority Over-Sampling Technique

SVM · Support Vector Machine

SpO_2: oxygen Saturation

W

WBC. White Blood Cells

WHO · World Health Organization

ABSTRACT

The newborn baby is defined as an infant in the first 28 days of life after birth. In the world, neonatal mortality is a big problem that accounts for the lion's share of under-five mortality. Neonatal diseases are one of the deadliest diseases for children under five in the world. Countries and health organizations make a great effort to eradicate this global burden of disease. However, some obstacles have a hindering effect on the fight against the disease. This lower number of neonatal health professionals, especially in sub-Saharan countries such as Ethiopia, is the main obstacle to increasing the accessibility of diagnosis for all neonatal patients. Most neonatal health professionals decide the type of illness only on clinical diagnosis (oral questions and answers). This method often led to misdiagnosis because the health professional may not have a complete picture of all the variables that have a contributing effect on neonatal disease as well as, and some health professionals may curiously not diagnose all newborns due to workload. In addition, neonatal diseases often have similar symptoms, which also led to misdiagnosis and high attention. These problems can be minimized with the implementation of a predictive system that can take all the necessary data from the patient and classify the most likely neonatal disease. The availability of relevant historical data and emerging technologies such as machine learning and data science can address this type of problem. Therefore, this study proposed a predictive model of specific neonatal diseases based on data from neonatal patients collected at the Asella Compression Hospital. The newly developed model stacking suite was compared to a machine learning algorithm such as XGBoost (XGB), Random Forest (RF), and support vector machine (SVM) with and without recursive feature removal with cross-validation (RFECV). under cross-stratified k-fold validation. The results of the performance evaluation showed that the newly developed model stacking ensemble with the RFECV method performs better than other models, which obtained an accuracy of 97.04. The proposed neonatal predictive model classifies a common neonatal disease based on a multi-class prediction approach to help domain experts.

Keywords*: Neonatal Diseases, Machine Learning, Global Burden, Stacking*

CHAPTER ONE

1. INTRODUCTION

This chapter explains the introductory part of this study including the background of the study, a statement of the problems, research questions answered, motivation of the study, objective of the study, scope, limitations, and organization of the study.

1.1. Background of the Study

The neonatal period is the critical time of human life because the newborn baby is vulnerable to various problems due to the baby should adapt to the new environment and should complete several physiological adjustments essential for life external of the uterus(WHO, 2021) (Mitiku, 2021). Children face the highest risk of dying in their first month of life. Globally more than 2.4 million children died in the first month of life according to the 2019 report, approximately around 6,700 neonates died every day(WHO, 2020).

The neonatal mortality rate shows disparities across the region and countries. The mortality rate of neonatal was the highest in sub-Saharan Africa and South Asia, with the mortality rate approximated at 34 and 29 deaths respectively in 2019. Across the countries, the risk of neonatal death was about 55 times higher in the highest mortality country than in the lowest mortality country according to a study by the world health organization (UNICEF's Data & Analytics, 2020)[1]. One-third of the world's mortality of neonatal burden is contributed by Africa specifically in sub-Saharan Africa; It is the highest neonatal mortality region in the world due to different diseases and is still far off from reaching Sustainable Development Goal (SDG-3)(Forum, 2021). Neonatal morbidity and mortality are significant contributors to under-five morbidity and mortality in a low-income country like Ethiopia(WHO, 2021). According to, the 2019 Ethiopia Mini Demographic and Health Survey (EMDHS) report the neonatal mortality rate accounts for about 30 per 1000 live births(Ethiopian Public Health Institute Addis Ababa, 2019).

The major illness that caused the highest percentage of neonatal mortality is sepsis, pneumonia, tetanus and diarrhea, intrapartum-related complication, and asphyxia. In Ethiopia, the most

[1] https://www.unicef.org/topics/sustainable-development-goals

common illness leading to neonatal mortality are sepsis, birth asphyxia, necrotizing enter colitis(NEC), and respiratory distress syndrome (RDS) according to a 2019 survey(Ethiopian Public Health Institute Addis Ababa, 2019). Neonatal conditions are the leading cause of death in low-income countries in 2019 according to the WHO global health estimate(WHO, 2020). The most common neonatal diseases such as sepsis, respiratory distress syndrome, birth asphyxia, and necrotizing enter colitis are the leading diseases of neonates accounting for 26%,23%,19%,,3-7% respectively (Mekonnen et al., 2021)(Mikhno & Ennett, 2016)(Delele et al., 2021) which also takes the highest percentage in Ethiopia(WHO, 2021).

Neonates were dying and suffered daily due to, a lack of diagnostic tools, diagnostic delay, and lack of quality care, and treatments for neonatal conditions are challenging(Demisse et al., 2017). Moreover, the lack of early identification of neonatal diseases often results in inappropriate use of antibiotics and results in increasing the risk of the development of antimicrobial resistance. Some neonatal diseases have often similar symptoms, for instance, neonatal sepsis is very similar to other-threatening diseases such as perinatal asphyxia and necrotizing enter colitis which makes it difficult to accurately diagnose and treat. In developing countries like Ethiopia, there are few diagnostic tools exist, in that addition, the results of diagnostic tests take up to 48 hours, which is often too long to wait as the condition of a neonate with neonatal diseases can deteriorate rapidly(McDougall & Hunt, 2018). Therefore, early identification of neonatal disease is very crucial to reducing the socioeconomic burden on society and neonatal mortality and morbidity rate.

The neonatal period is unique to the first 28 days of life after birth and represents around 40% of all mortality in children under five. Among many neonatal diseases, the main contributors to the global burden of morbidity are sepsis, RDS, and birth asphyxia ((IHME), 2019). Neonatal diseases exert a heavy burden on families, society, and the health system. Currently, preventive methods, diagnostic tools, and treatments for neonatal diseases remain limited due to the complex cause of these diseases. Many of these preventive approaches focus on maternal health before birth, such as maternal immunization and efforts to guarantee a healthy pregnancy(Victorio, 2021)(Yuan et al., 2022). To minimize the stated challenges, early detection of neonatal diseases with machine learning techniques is applied.

Machine learning is a sub-branch of Artificial intelligence (AI) that use data and algorithm to compute the performance and innovative opportunities to understand data processing in the health domain in general and neonatal disease classification in particular. Machine learning is

applied in the health systems to improve health sectors in several areas such as disease diagnosis, drug recommendation, personalized medicine, medical treatment, clinical trial improvement and epidemic control, and so on.

Based on the above problems and challenges on one hand and the application of machine learning on another hand, this study aimed to classify neonatal disease using developed stacking machine-learning techniques. Therefore, neonatal data were collected from Asella Compressive Hospital neonatal patient history and NICU and classify four common neonatal diseases such as sepsis, respiratory distress syndrome (RDS), Parental asphyxia (PA), and necrotizing enter colitis (NEC) which is difficult to diagnosis and need early identification are classified.

1.2. Motivation of the study

The WHO survey and different statistical investigations indicate that neonatal mortality and morbidity is the main problem of the global burden of disease. Specifically, in the sub-Saharan region, including Ethiopia, the neonatal mortality rate became the main contributor to the death of children under five years of age(WHO, 2020).

Neonatal conditions such as prematurity, jaundice, apnea, and disease transfer from mother to child are easily identified by clinical diagnosis(Forum, 2021). But due to of similarity of disease symptoms, the lack of neonatologists and pediatricians as well as the lack of computer-aided diagnosis to make early identification of this common neonatal disease namely Sepsis, RDS, PA, and NEC (Desalew et al., 2020). Therefore, this disease needs early identification to reduce the neonatal mortality rate by classifying the mentioned diseases. Globally different statistical analysis research was conducted on neonatal mortality and factor associated with neonatal diseases. In Ethiopia, a similar investigation was conducted on newborn mortality and its impacts on the socioeconomic burden in society.

Based on the identified problems the researcher was motivated to work on common neonatal diseases' classification using machine-learning techniques that can analyze medical data and classify the disease properly based on the collected dataset. Therefore, the study is interested to propose a classification model for neonatal disease classification early using machine learning techniques to overcome the above problems. For domain expert, it helps to identify neonatal diseases early, make fast decision-making, and help to give proper treatment to the newborn, which reduce the mortality rate.

1.3. Statement of the Problems

The period from birth to the first 28 days of life after birth is the neonatal period and its mortality holds more than 40% of all mortality in children under five ages. Newborn health is very important for human life survival because life starts at this age. Neonatal are vulnerable to different environments and easily affected by the disease. Neonatal disease is a heavy burden on families, society, and the health system as well as the government. Among neonatal condition, many of them are controlled and easily identified by the clinical diagnosis but sepsis, RDS, PA, and NEC are difficult to identify and holds the highest percentage of neonatal mortality(WHO, 2021)(Gupta & Choudhury, 2017).

Currently, preventive methods, diagnostic tools, and treatments for neonatal diseases remain limited due to the complex cause of these diseases. Many of these preventive approaches focus on maternal health before birth, such as maternal immunization and efforts to guarantee a healthy pregnancy(Victorio, 2021)(Yuan et al., 2022). Recently, the world is facing a shortage of 4.3 million healthcare institution workers(Oleribe et al., 2019). According to Sousan research in 2016, Ethiopia is below the WHO standard for health workers to population ratio the standard is health workers 2.18 per 1000 population(Valizadeh et al., 2016). There are various risk factors related to newborn mortality have been identified in studies undertaken in Ethiopia(Sharew et al., 2018)(Mitiku, 2021). Among these factors lack of diagnostic tests results in the inappropriate use of antibiotics, which increases the risk of the development of antimicrobial resistance, the similarities of neonatal disease symptoms are the main challenge to easily identify and making difficult to accurately diagnose and treat (McDougall & Hunt, 2019). In addition, the diagnostic tests may take up to 48 hours, which is too long to wait as the condition of neonates with a neonatal disease can deteriorate rapidly. Furthermore, low health service, disparities in hospital setups or accessible equipment, lack of antenatal care, diagnostic delay, and medical practitioners cannot easily predict the disease, which demands expertise and higher knowledge to diagnose (Desalew et al., 2020)(UNICEF's Data & Analytics, 2020). Therefore, early detection enables the expert to diagnose the newborn to give treatment and reduce the mortality and morbidity of the neonates. This study proposed a machine learning approach to solve the above problem in healthcare institutions by predicting neonatal disease based on the collected dataset. Furthermore, the study helps to achieve early diagnosis of disease, suggest experts to recommend appropriate treatments, and helps the government policies toward SDG-3.

1.4. Research Questions

This study attempt to answer the following questions**.**

RQ1. What are the necessary features to prepare an effective neonatal disease dataset in applying machine-learning models?

RQ2. How the performance of the prediction model is improved using ensemble techniques?

RQ3: To what extent the developed prototype system works for the diagnosis and treatment recommendations of neonatal diseases? Objectives of the Study

1.5.1. General Objective

The general objective of this study is to design a neonatal disease classification model using machine-learning techniques, which help experts to diagnose and treat patients.

1.5.2. Specific Objectives

To achieve the general objectives of this study the research has the following specific objectives

- ✓ Review literature concerning machine learning concepts to predict neonatal diseases
- ✓ Collect data and prepare a dataset for model training
- ✓ Identify the relevant set of features from the original feature set for neonatal diseases prediction
- ✓ Building ensemble machine learning to optimize model performance
- ✓ Evaluate the neonatal diseases prediction model's performance
- ✓ Design prototype for neonatal diseases prediction

1.5. Significance of the Study

The significance of this study is described from different perspectives. The study contributed to providing a classification model of neonatal diseases for health workers in health sectors easily. The proposed model plays a significant role to update or improve the knowledge and experience of the health professional to diagnose and treat newborn diseases. The research benefits health sector researchers or professionals to maximize awareness of the issue of newborn health amid decision-making. The study support government policy and strategies toward newborn and strength the existing policies like Sustainable Development Goals (SDG). The prototype is important to give training in the areas where shortages of experienced experts

are available. Hence, it is also advantageous for a rural area that has a computer system and a lack of medical professionals and medical facilities.

1.6. Scope and Limitation of the study

Scope of the Study

This study proposed to ensemble stacking neonatal disease classification algorithm to predict common neonatal diseases such as sepsis, PA, RDS, and NEC model using machine-learning algorithms that enables the experts to give appropriate treatment and recommendations based on data collected from Asella Comprhenessive Hospital Neonatal Intensive Care Unit (NICU) dataset, which is 2018-2021. The study focused on Asella Comprhenessive Hospital dataset because the hospital has its pediatrician department, works in a child care center and its serve for many rural areas of the population where the newborn will be affected by different disease due to poor living condition, educational status, and lack of follow up during antenatal care. Only three machine learning techniques such as RF, xgboost, and SVM, and form stacking linear regression model with RFECV feature selection methods.

Limitation of the study

Though there are many hospitals and health centers, in Ethiopia, it is very difficult to find organized data in electronic format and its time consuming to change the data from manually recorded to electronic data. Furthermore, it is difficult even to collect data because different hospitals are not interested to give the data for privacy purposes and other issues related to willingness to support. Data were only collected from the Asella Compressive Hospital and there is no standard rule to select relevant features to predict neonatal diseases. Due to a lack of data problems, this study considers only four common neonatal diseases namely Sepsis, NEC, RDS, and PA not considering the severity of the diseases.

1.7. Organization of the Thesis

This section, described the organization of the whole study as follows.

Chapter 1 Introduction: the introduction part of the study includes the background, motivation, statement of the problem, research questions, significance of the study, scope, and limitation of the study described.

Chapter 2 Literature Review: states the scientific literature related to neonatal and machine learning concepts and related works

Chapter 3 Methodology: this chapter defined the research methodology, the source of the data, how to prepare the dataset, feature selection, and model evaluation.

Chapter 4 proposed solution: this chapter states the proposed solution for neonatal disease classification, proposed feature selection, data understanding, the model selected, and evaluation methods used.

Chapter 5 Implementation: describe the implementation and experimentation of the proposed solution**,** implementation of features selection, data preprocessing, working environment data set description model implementation, and evaluation.

Chapter 6 Result and Discussion: discusses the result of the dataset class distribution, feature selection, and performance evaluation with and without feature selection

Conclusion and future work: it describes the conclusion for the study and future work for further investigation of this research work

CHAPTER TWO

2. LITERATURE REVIEW

This chapter presents a review of surveys of academic articles, journals, conference papers, and other sources related to neonatal, and machine learning to provide a brief understanding of what done so far, and what will be done in this thesis. The review presented in this chapter deals with neonatal diseases, machine-learning techniques, data pre-processing, feature selection methods, prediction model, and model evaluation approach.

2.1 Neonatal Disease and its Global Burden

Life starts in normal health is very important for all newborns. In the first 28 days, the neonatal period is very especially serious. The period represents the most serious time for a newborn's survival due to the neonates are vulnerable to various problems and should adapt to the new environment as well as the newborn should complete several physiological adjustments essential for life external of the uterus(Delele et al., 2021). Neonatal diseases disturb the normal state of body organs and make newborns abnormal function and proceed to newborn mortality if it doesn't get an early diagnosis, treatment, and care (Gupta & Choudhury, 2001). Neonatal mortality contributes to a higher percentage of the mortality rate of young children, especially in underdeveloped countries, including Ethiopia, where the living condition is low(Gizaw et al., 2014). In Ethiopia, scientific confirmation on the level of a community practice of essential newborn care is scarce and questionable. Neonatal mortality has a great impact on socioeconomic development and quality of life, which also takes a significant part of mortality in children under five years of age. In 2019, global health data exchange (GHDX) report more than 2.4 million neonatal death every day, and sub-Saharan Africa holds 60%. Sepsis, birth asphyxia, Necrotizing Enter colitis(NEC), and respiratory distress syndrome of the newborn are common neonatal diseases((IHME), 2019) (Najafian & Khosravi, 2020)(Abdo et al., 2019)(Getabelew et al., 2018)(Alsaied et al., 2020).

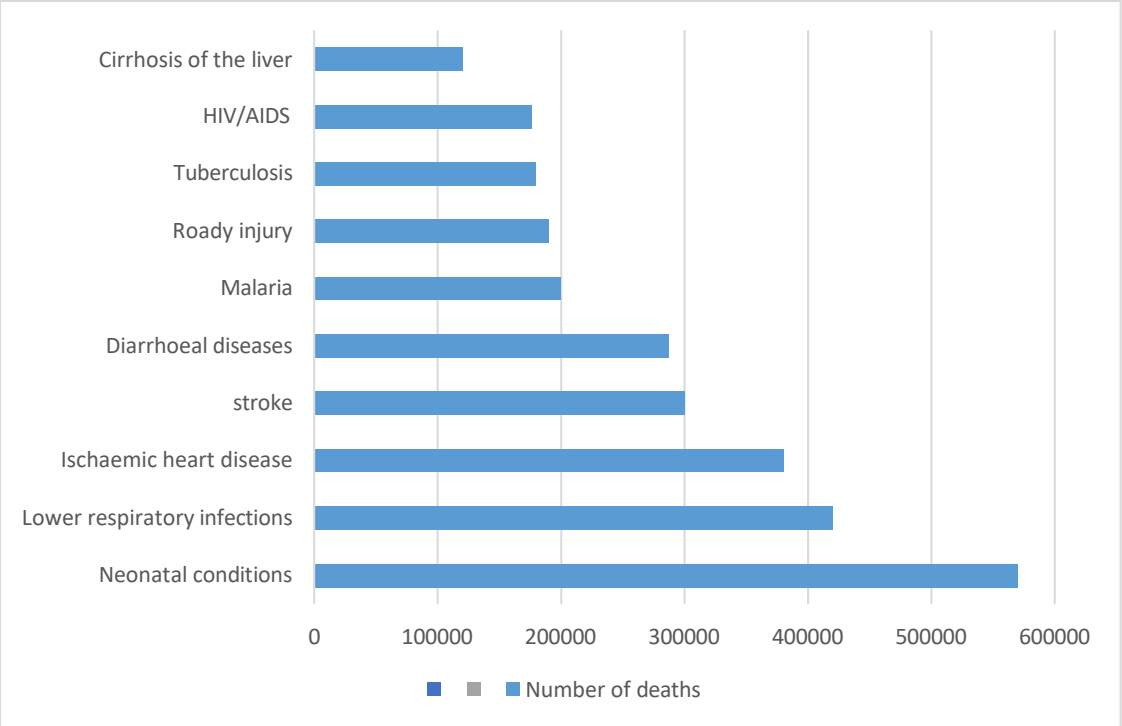

Figure 2:1 Leading causes of death in developing countries[2]

2.2. Common Neonatal Diseases

The most common neonatal communicable diseases which hold around 75 % of newborn deaths are due to four main diseases sepsis, birth asphyxia, Necrotizing Enter colitis, and respiratory distress syndrome(WHO, 2021)(Delele et al., 2021).

2.2.1. Respiratory Distress Syndrome (RDS)

Neonatal respiratory distress syndrome is a common reason, which increases the mortality and morbidity of neonates. RDS is a leading clinical dilemma faced by preterm infants and is directly related to structurally undeveloped and surfactant deficient lungs. RDS is caused by pulmonary surfactant deficiency in a newborn's lungs, which is more common in babies born before 37 weeks. Surfactant is a foamy substance that saves the lungs from fully expanding so that neonates can breathe in air once they are born. Without enough surfactant, the lungs collapse and the neonate have to work hard to respire. The baby might not be able to respire enough oxygen to support the body part(Liu & Sorantin, 2019).

[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

2.2.1.1 Diagnosis of RDS

Neonatal RDS occurs in infants whose lungs have not yet fully developed. The more premature the baby, the higher the risk of RDS after birth. RDS is diagnosed by a combination of clinical signs and symptoms, laboratory tests, and chest X-rays. Symptoms of RDS include tachypnea, lack of oxygen at birth or in the first few hours after birth, grunting, intercostal subcostal retraction, decreased air entry, problem with reflexes, and hypothermia or temperature instability (Liu & Sorantin, 2019) (Najafian & Khosravi, 2020).

A blood oxygen level test is used to check that the baby's lungs are working properly and measure the acid-base balance in the blood. Oxygen saturation (SpO2) measures the amount of hemoglobin in the blood. Hemoglobin is a protein found in red blood cells and responsible for transporting oxygen from the lungs to the rest of the body. An x-ray or chest x-ray is a quick and painless imaging test that uses certain electromagnetic waves to take pictures of the drawings in and around the chest. To confirm detection, the clinical team takes x-rays of the newborn's chest. An X-ray of a preterm infant with RDS will show lung and air volumes in the lung airways that are black in contrast to the encompassing white regions that do not contain air (Najafian & Khosravi, 2020).

2.2.2. Necrotizing Enter Colitis in the Newborn

Necrotizing enter colitis (NEC) disease is a major cause of morbidity and mortality in the neonatal, and its burden became more prominent. NEC has one of the most destructive diseases occurring in a neonatal intensive care unit (NICU) and this can damage the tender tissues and lead to neonatal mortality as well can cause other diseases in the baby but most babies completely recover once they receive treatment in time. (Alsaied et al., 2020)(Neu, 2020).

2.2.2.1 Diagnosis of NEC

NEC is diagnosed by running various laboratory tests and doing a physical examination. During the examination, the pediatrician or neonatologist gently touches the baby’s abdomen to check for pain, swelling, and vomiting. Then different laboratory tests like WBC, temperature, and blood cultures C - reactive protein were taken. Also, stool guaiac test and doctor may order certain blood tests to measure the baby's platelet levels and white blood cell count. Platelets make it possible for blood to clot. White blood cells help fight infection. Low platelet levels or a high white blood cell count can be a sign of NEC. The neonatologist may need to insert a needle into your baby's abdominal cavity to check for fluid in the intestine. The

presence of intestinal fluid usually means that there is a hole in the intestine(Holm, 2021),(Neu, 2020)

2.2.3. Sepsis

Newborn babies are at a higher risk of infection because of their weak immune systems related to their age. Bacteria and viruses cause most infections in newborn babies. Globally, neonatal sepsis is a leading reason for inflated death and illness of neonates; it is responsible for approximately 30%-50% of the total neonatal deaths in developing countries. Moreover, it is also a common reason for neonatal death in Ethiopia(Getabelew et al., 2018). Neonatal sepsis is a kind of infection and specifically refers to the presence in a very infant of a bacterial bloodstream. A newborn who has an infection and develops sepsis can have inflammation all through the body, leading to organ dysfunction. Sepsis is a disorder in which the body's immune system overreacts to an infection, causing white blood cells to attack the body's cells and tissues(Getabelew et al., 2018)(Ershad et al., 2019). Sepsis in newborn babies can progress fast and early diagnosis and treatment are vital for improved outcomes.

2.2.3.1. Diagnosis of Neonatal Sepsis

Neonatal sepsis is described as a clinical syndrome of bacteremia with general signs and symptoms of infection in the first month of life. Signs and symptoms are nonspecific in neonatal sepsis. Therefore, the differential diagnosis is important. Neonatal sepsis can present with abnormal breathing, tachypnea, vomiting, abnormal activity (seizures), temperature instability, and absent or abnormal neonatal reflexes(Lashkari, 2021). Laboratory tests are a medical procedure that is intended for testing a sample of blood, urine, or other tissues or substances taken from the body to help diagnose a disease (Fattah et al., 2019). Sample of blood taken from the body and tested in a lab. Blood tests are usually done to check how the body copes with illness, inflammation, and infection. A common blood test measures how many leukocytes are circulating in baby blood, among other things. WBC (also called leukocytes) fight bacteria, viruses, and other organisms the baby's body identifies as a danger. Blood cultures are medical laboratory tests used to distinguish an infection in the blood and identify the cause(Cabrera-Quiros et al., 2021). Blood culture for the significant diagnosis of bacterial and fungal infections. A C-reactive protein (CRP) test measures the amount of C-reactive protein in the blood. CRP is a well-established biomarker of infection and inflammation. Because the amount of CRP raised significantly during acute inflammation than the levels of

the other acute phase reactants, the CRP test has been used to identify infection or sepsis (Song et al., 2020)(Lashkari, 2021).

2.2.4. Perinatal Asphyxia

Perinatal asphyxia (PA) is one of the major contributors to neonatal morbidity and mortality across the globe. In Africa including Ethiopia PA remains a severe disease that places economic and social-emotional burdens on families and society(Abdo et al., 2019). Perinatal asphyxia is defined as the failure to establish breathing at birth and sustain enough respiration after delivery and is characterized by a marked impairment of gas exchange (Abdo et al., 2019). Perinatal asphyxia occurs when there is no blood flow and gas exchange to or from the embryo in the hours leading up to, during, or after birth. When before birth or immediately after the birth blood circulation is halted altogether, and there is a halfway or complete lack of oxygen to the vital tissues (Abdo et al., 2019).

2.2.4.1. Diagnosis of Perinatal Asphyxia

When a baby's brain and other organs do not get enough oxygen and nutrients before, during, or right after birth, called perinatal asphyxia. Each baby may experience symptoms of perinatal asphyxia differently. However, the following are the most common symptoms; slow heart rate, bluish or pale skin color, abnormal activity (seizures), poor muscle tone and reflexes, resuscitation, tachypnea (rapid respirations), temperature instability, and failure of air to enter the lungs. A neonatal clinical assessment can be performed using APGAR (Activity, Pulse, Grimace, Appearance, Respiration) score test to assess the health of the infant's physical condition immediately after delivery and to decide the urgent need for additional medical or emergency care (Shirwaikar et al., 2017),(Shirwaikar et al., 2017). The subsequent test is used to diagnose Perinatal asphyxia: Severe acid levels pH less than 7.00 in the arterial blood of the umbilical cord, APGAR test score zero to five for longer than five minutes to evaluate the baby’s condition (Shirwaikar et al., 2017). The score is based on measures of heart rate, breathing effort, the color of skin, muscle tone, and response to stimulation. Therefore, the scores obtainable are between zero and ten, with ten being the highest possible score. A total score of six to ten is considered normal and an abnormal APGAR score indicates depressed vitality. However, several possible causes of abnormal APGAR scores exist, such as perinatal asphyxia (Shirwaikar et al., 2017).

Table 2-1 Common Neonatal Diseases

Diseases	Description	Treatment
RDS	a lack of a slippery substance in the lungs called surfactant	CPAP+incubators+ Antibiotics (ampicillin)
Sepsis	a blood infection that occurs in an infant younger than 90 days old.	Antibiotics: - ampicillin and gentamicin
PA	failure to establish breathing at birth	Resuscitation +oxygen +Therapy + Antibiotics (ampicillin)
NEC	a serious gastrointestinal problem that mostly affects premature babies	Antibiotics(gentamicin), metronidazole (AGM)

2.3. Overview of Machine Learning

The advent of the Internet, computing, and other technologies gave rise to numerous advances in many fields, such as medicine, agriculture, e-commerce, and law (Rabbi, 2019). In this era, many organizations use computer systems and artificial intelligence all over the world to satisfy their customers, save time and other resources, and gain more profit. Machine learning is a branch of computer science and AI, which work with the use of a huge amount of different organizational data. It has become an interesting and important research area for solving very complex problems with a huge amount of data and is used for the detection of target patterns from the data. Machine learning uses a huge amount of data and algorithms to mimic the way a human being can learn and improve its performance through experience. It is about training the machine on what to do and having it do it automatically based on experience without explicit programming. According to different scholars, machine-learning approaches are classified into supervised, unsupervised, semi-supervised, and reinforcement learning approaches.

2.3.1. Supervised Machine Learning

Supervised learning is a type of machine learning algorithm, that develops a machine-learning model based on a known label. This algorithm consists of a target or dependent attributes, which are predicted from a given set of predictor or independent attributes. The data used in the supervised learning algorithm is labeled or categorized data. The learned rule is then used to categorize unseen data with unknown outputs. Classification and regression are the two common tasks of supervised machine learning. Some examples of supervised learning are

Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, K-Nearest Neighbor, Logistic Regression, etc.

2.3.2. Unsupervised Learning

Unsupervised learning is the algorithm used to work with unlabeled data. This algorithm is a very powerful tool for analyzing data and identifying patterns and trends without the unknown label, that instance belongs. It is the learning algorithm applicable to grouping a given data set into different groups. K-means is the most common example of unsupervised learning.

2.3.3. Semi-Supervised machine learning

Semi-supervised Machine Learning some sorts of machine learning techniques are neither fully supervised nor fully unsupervised. It falls between the two and is called semi-supervised machine learning. Semi-supervised algorithms mostly use small-supervised learning parts that mean a small amount of pre-labeled annotated datasets, and a large unsupervised learning component that means many unlabeled datasets for training

2.3.4. Reinforcement Learning

Reinforcement Machine Learning Reinforcement machine learning is a machine learning technique in which the agent finds a suitable action to take in a given situation to maximize reward, and it also discovers optimal results by a process of trial and error (Perner, 2015).

2.3.5. Ensemble Learning

Ensemble methods are machine learning techniques that combine several single models to produce one optimal predictive model[3].there are different types of ensemble learning models such as bagging, random forest, stacking, and boosting. Ensemble methods are general meta techniques that seek better performance in the prediction. Used for both classification and regression tasks.

Stacking ensemble learning[4]

Stacking is an ensemble method that seeks a diverse group of base classifiers by varying the model types fit on training data and combining the base model for prediction. Stacking is a general method in which the base learner is trained individually and combined. The individual

[3] *https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f*

[4] *https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/*

learners are called level-1 learners and the models that are used to combine individual learners are the second levels learners.

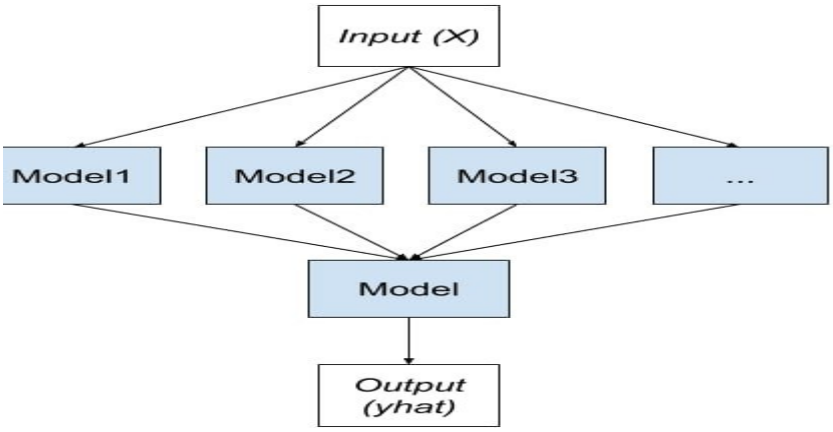

Figure 2:2 stacking ensemble flow diagram

Boosting ensemble learning[5]

It combines many weeks of learners and converts them into strong learners. The popular algorithm is Adaboost, gradient boosting, and stochastic gradient boosting(XGBoost and similar

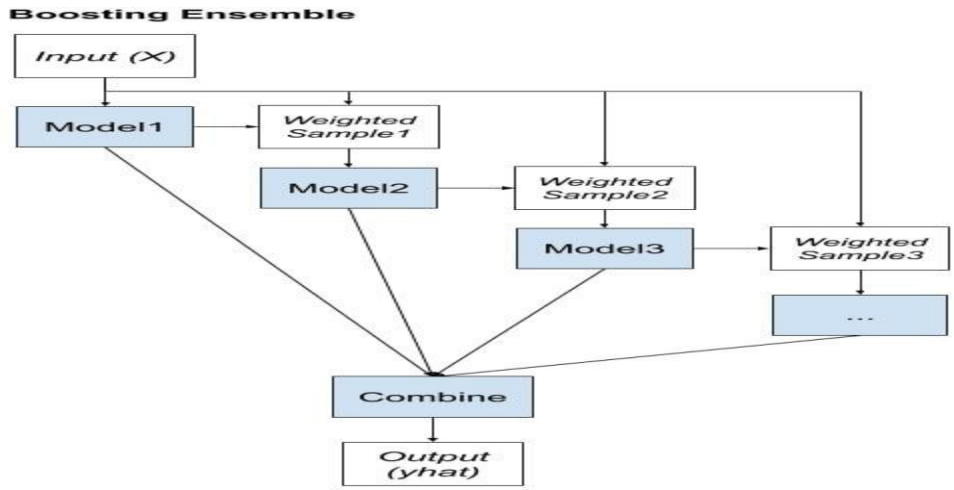

Figure 2:3 Boosting ensemble flow diagram

2.4. Machine Learning in Healthcare Sector

The health sector is one of the important sectors in one country, which determine the development, and life of the society. Good quality healthcare helps to prevent disease and improve the quality of life in society. How every in a developing country like Ethiopia,

healthcare sectors face multiple challenges such as increasing burden of disease, morbidity, mortality, disability, greater demand for health services, delayed diagnosis, higher social expectation, lack of manpower, challenges related to inefficiency, and low productivity of health workers(Atun, 2015). A fundamental transformation of the healthcare system is the key factor to overcoming these challenges and achieving Sustainable development goals (SDG) by 2030. Therefore, machine learning is the most feasible indicator and an interesting area in digital technology which is used to overcome the challenges of the healthcare system with less, and could be the catalyst for health system transformation(Badawi et al., 2014).

Machine Learning plays a great role in preventive healthcare issues and is an early prediction of the disease. This early prediction makes the diagnosis of the disease simple and enables the practitioners to treat the patients early. After screening the disease, healthcare providers or experts could then use this predictor to identify high-risk individuals and recommend tests or other interventions for the disease (Kohli & Arora, 2018). In healthcare, machine learning can be applied in different tasks like identifying diseases and diagnosis, drug discovery and manufacturing, medical imaging diagnosis, personalized medicine, outbreak prediction, etc.
Hence, to solve the diagnosis and treatment problems machine learning techniques play a crucial role in the health sector to identify correctly the presence or absence of the disease based on the patient history data. This study aims to work on labeled data to predict Neonatal Disease, by using supervised machine-learning techniques.

2.4.1. Machine Learning Approach in Disease Prediction

Today, people face various diseases due to lifestyle, environmental conditions, and other factors. So earlier disease prediction becomes an important task to save the life of the population. But accurate disease prediction easily based on symptoms becomes too difficult for medical experts (Dahiwade et al., 2019)(Jayachandiran, 2018). Using the latest technology in the health system will significantly make these difficulties easy to handle(Jayachandiran, 2018). Currently, machine learning uses successfully in the healthcare system in predicting and diagnosing many serious diseases based on patient history data. In health care systems machine learning allows the health expert to process complex and huge amounts of datasets and then analyze the dataset in understandable clinical knowledge(Dahiwade et al., 2019). Therefore, machine learning applied to the healthcare system has a significant impact on disease prediction to deliver treatment and increase the accuracy of the disease diagnosis. Different researchers used different machine learning algorithms, as discussed below.

Random Forest (RF): It is a supervised machine learning algorithm called random decision forest, which is an ensemble machine-learning method for classification, and regression. It operates by generating multiple trees of randomly sub-sampled features. Many researchers selected the random forest machine learning algorithm for disease prediction due to its comparatively good accuracy, robustness, and ease to use model(Daunhawer et al., 2019).

Support Vector Machine (SVM): it is a supervised machine learning algorithm used for prediction and regression problems, therefore, for neonatal disease prediction. SVM works by taking the data points and making a hyperplane that separates the best tags and the hyperplane is a line, which is known as a decision boundary for labels. For this reason, most of the studies done in machine learning disease prediction used SVM due to its accuracy, supports multiple classes, and don't suffer the condition of overfitting(Shirwaikar et al., 2016)(Dahiwade et al., 2019).

Decision Tree (DT): These are supervised machine learning algorithms that are used for both classification and regression problems. It solves the problems by transforming the data into a tree structure and sorting it by its feature values. Each node in a decision tree denotes features in an instance to be classified, and each leaf node represents a class label to which the instances belong. This model uses a tree structure to partition the dataset based on the condition as a predictive model that maps observations on an item to decide on the target value of instances (et al., 2017).

Artificial neural networks (ANN): - is the fundamental types of neural networks that are inspired by the biological brain and include different layers i.e., input layer, hidden layer, and output layer nodes(Roy Chowdhury et al., n.d.). Each connection input layer signal can transmit it to hidden layers and the hidden layer process it and transform it into output concerning the activation function.

Gradient Boosting: this is a type of machine learning used for both classification and regression problems. Gradient boost is an ensemble-learning method, that groups weak predictive models specifically a decision tree, and the resulting algorithm is known as gradient boosted trees, which outperforms random forest. It builds the predictive machine-learning model in a step-wise manner like others boosting techniques do, and it advances the performance of the model.

Table 2-2 Pros and Cons of Machine learning algorithms used frequently

ML Algorithms[6]	Pros	Cons
RF	used for both classification and regression tasks, no scaling or transforming of variables is worrying, not affected by outliers to a fair degree, has good performance even on imbalance dataset	Is like a black box algorithm, with little control over what models do, are not easily interpretable
DT	It's like white both models, require fewer data processing, and irrelevant features won't affect the decision tree	It tends to overfit and requires high computational time
NB	Perform well in a small dataset, fast prediction	Assumption of independence between features
KNN	doesn't make assumptions regarding the dataset, simple forms of machine learning	Works on the small dimensional dataset, can't handle data with missing values
XGB	High predictive accuracy, flexible and can handle missing values	Computational expensive and less interpretable
SVM	Thrives in high dimension, fast prediction used for both regression and classification	are not the most efficient algorithms, Overfitting is another potential side effect of Support Vector Machines

[6] *https://medium.com/analytics-vidhya/ml-algorithms-pros-cons-and-suitable-usages-b377c3c09f1b*

2.5. Feature Selection

Feature selection techniques are used to reduce the number of features using appropriate methods and select relevant features to develop the model. The aim of using feature selection is to avoid inconsistent data, and noisy data, and to reduce the features of the dataset for better performance. Machine learning models can choose the best features automatically based on the types of problems being solved. This helps to reduce the computational cost, better model interpretability, and leads to better performance(Dhal & Azad, 2021).

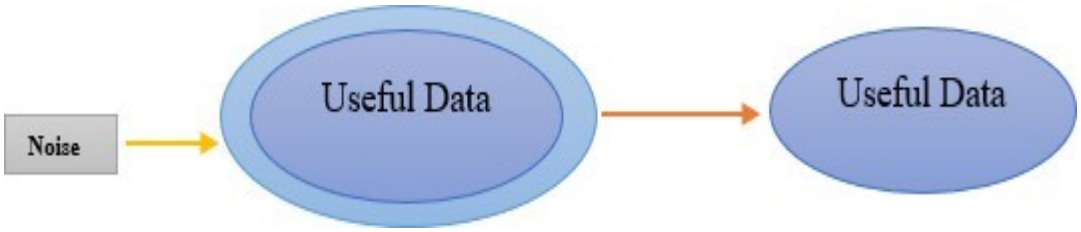

Figure 2:4 Feature Selection Adopted from[7]

Feature Extraction: extracting and transforming relevant features from original dataset features and reducing them to more manageable datasets without losing relevant data and important features. Principal component analysis (PCA), independent component analysis, and linear discriminant analysis algorithms are some of the commonly used feature extractions used in the health dataset.

Feature Selection methods: is approaches are classified into four categories, such as filter approach, wrapper approach, and embedded approach(Kaushik, 2016).

Filter Methods: This method is frequently used as preprocessing step, and to assess the relevance of features by using a ranking procedure that consequently removes the least-scoring features. The methods are found to be fast, scalable, computationally easy, and independent of any machine learning classifier(Kaushik, 2016).

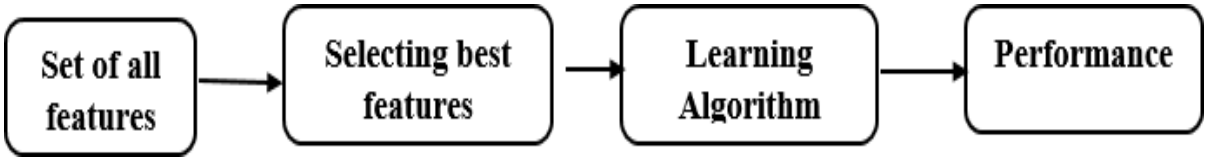

Figure 2:5 Filter Method Flowchart adopted from[8]

[7] *https://scikit-learn.org/stable/modules/feature_selection.html*
[8] *https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/#h2_1*

Wrapper Methods: It is similar to the filter method in identifying a relevant set of features except that they make use of a predefined classifier algorithm instead of an independent measure of the subset evaluation. This method gives better results related to filter methods, but they tend to be more computationally expensive when the number of features becomes very large(Kaushik, 2016).

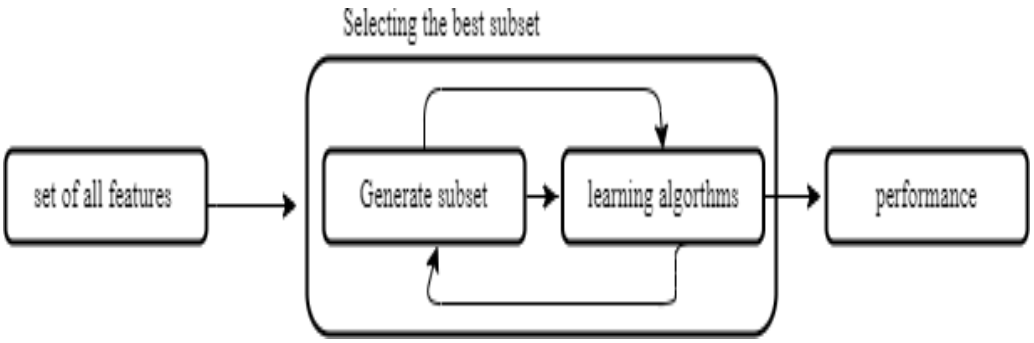

Figure 2:6 Wrapper Method flowchart adopted from[9]

Embedded Methods: It combines the qualities of both filter, and wrapper methods to generate the best feature subset. The algorithms that have their built-in feature selection methods employ it. LASSO and RIDGE regression are some of the most popular examples of this method which have an inbuilt penalization function to reduce overfitting(Kaushik, 2016).

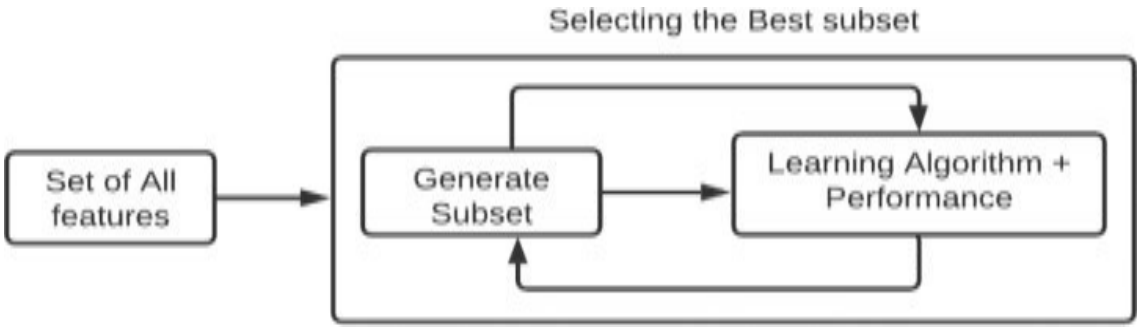

Figure 2:7 Embedded Method Flowchart adopted from[10]

[9] *https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/*

[10] *https://ml-concepts.com/2021/10/07/3-embedded-methods/*

Table 2.3 Feature Selection methods(Ibrahim et al., 2018),(Kaushik, 2016)

FS Methods[11]		Pros	Cons	Examples
Filter	Univariate	Independent of the classifiers, faster, scalable &computationally simple	May lead to poor feature subset, no relation with features	Information gain, ANOVA
	Multivariate	Better computational complexity, best for huge features	Slower, require high levels of mathematical calculations	CFS, Markov blanket filter
Wrapper	Randomized	Less prone to local optima, classifier dependent,	Classifier-dependent selection, Higher risk of overfitting	Randomized hillclimbing , GA
	Deterministic	Best with a huge number of features	Overfitting problems	SFS, SBE, RFCV
Embedded		Classifier dependent, Better computational complexity	Classifier dependent Selection.	DT, Weighted NB, Feature Importance

Note: GA: Genetic Algorithm, ANOVA: Analysis of Variance, CFS: Correlation-based Feature Selection, SBE: Sequential Backward Elimination

2.6. Related Works with Neonatal Disease prediction

There are several studies by different researchers, which have applied machine learning in the prediction of different diseases. Machine learning trained on data can accurately predict neonatal disease. Nevertheless, the major number of prediction failures decreases its inappropriate for individual treatment decisions.

In 2011, Chowdhury stated that the demand for ANN application in the field of medical decisions was increasing for better performance in predicting disease diagnosis. The researcher represents the use of ANN for neonatal disease diagnosis prediction by training a multilayer perception with a BP learning algorithm to identify a design pattern for the prediction of neonatal diseases. They compare different training algorithms of MLP such as Conjugate

[11] *https://sebastianraschka.com/faq/docs/feature_sele_categories.html*

Gradient Descent, and Quick Propagation, which are used for the prediction of neonatal diseases. The proposed model uses 94 cases of different symptoms and signs as a parameter to test the model and improved 75 % of accuracy. However, the method and the size of the data they used are very small so, it needs further research to get high accuracy with different methodologies and models(Roy Chowdhury et al., n.d.).

In Iran, Reza Safdari(Safdari et al., 2016a), developed a fuzzy expert system that predicts neonatal death risk. In this study, the researchers prepared questionnaires and the questionnaire was given to neonatologists to acquire knowledge and form the neonatologist's answers they identified important factors. The researchers combined computational and fuzzy models based on an inference system for the prediction of neonatal death risk. They used MATLAB and C# to build the model and GUI respectively. The results showed 90%, 83%, and 97% accuracy, sensitivity, and specificity of the model respectively.

A study conducted by Rudresh and Shiwwaiker, focused on the prediction of apnea episodes during the first week of a child's birth using machine-learning algorithms. The data consists of 229 examples of neonates admitted to the NICU. The researcher developed a model using different machine learning algorithms such as DT, SVM, and RF for the prediction of apnea episodes but provided that the medically expected accuracy improved. Among the results obtained, an accuracy of 88% using the random forest algorithm. Finally, the authors recommended further study to apply the machine-learning model to predict other neonatal diseases such as jaundice, respiratory distress syndrome, and sepsis,(Shirwaikar et al., 2016).

The study conducted by Bitew(Bitew et al., 2020), shows the risk of under-five mortality in Ethiopia using machine-learning models to predict its important sociodemographic determinates in Ethiopia based on 2016 EDHS data. The researcher used different machine learning models such as RF, LR, and KNN for the prediction of under-five mortality factors in Ethiopia. The results of the research show and consider the disparities in the mortality rate of those under five across the region of Ethiopia with the highest accuracy of 67.2 for the random forest. The limitation of the paper was focused only on surviving women this means they did not consider neonatal and maternal mortalities and this led to a decrease under five mortalities rate.

(Sheikhtaheri et al., 2021) were introduced to a Prediction of neonatal deaths in NICUs development and validation of machine learning models. The researchers applied machine-

learning techniques to improve the neonatologist or pediatrician's ability to predict the mortality of neonatal and its risk. The dataset was collected from Tehran, Iran in two phases first the neonatal death factors including diseases were identified and then they train, tested, and measures the performance of different algorithms like ANN, RF, CHART, SVM, ensembles were developed and modeled and SVM had the best accuracy having 94%. However, the researcher used imbalanced data, and neonatal diseases are considered as one feature and focused on only the risk and death of neonatal mortality.

(Shirwaikar et al., 2017) were explored selected supervised machine learning techniques for analysis of neonatal data used to build predictive models for diagnosis of neonatal diseases. The authors reviewed literature about use in neonatal disease diagnosis and monitoring of health with the methodological details and used accuracy, specificity, sensitivity, and ROC curves for model performance evaluation. According to this study, they concluded that neural networks, support vector machines, and decision trees have been considered efficient in neonatal data and ensemble techniques showed superior predictive capacity.

Table 2-4 Summary of related works

Authors and Year	Title	Methods and Findings	Gaps
(Roy Chowdhury et al., 2016)	Artificial Neural Network model for Neonatal Disease Diagnosis	ANN with an accuracy of 75%	Datasets used are too small and contain 94 rows, selected feature selection methods were time-consuming, as well as model evaluation techniques, are not clear. In addition, the clinical features dataset, which is very essential to predict neonatal disease, was not addressed.
(Shirwaikar et al., 2017)	Supervised Learning techniques for analysis of Neonatal Data	ANN, NB, SVM, LR with the highest accuracy of 89 SVM	Small dataset only demographic features were considered, and the class were highly unbalanced which decrease the result of the model
(Shirwaikar et al., 2016)	Machine Learning Techniques for Neonatal Apnea Prediction	DT, SVM, and RF with the highest accuracy 88% RF	It is binary classes having neonatal apnea or not for apnea & didn't address dataset outside NICU with a small dataset focusing on demographic data
(Safdari et al., 2016)	fuzzy expert system to predict the risk of neonatal death	Fuzzy-model inference system (accurac90%)	The researchers focus on the burden of neonatal mortality rather than identifying the cause of death

(Bitew et al., 2020)	Machine-learning approaches for predicting under five mortality determinants in Ethiopia: evidence from 2016 EDHS	RF, LR, and KNN with the highest accuracy of 67.2% with RF	Even though this research was done in Ethiopia, they only focus on children under five age and they didn't include neonatal data. Hence neonatal mortality holds the highest percentage of under-five ages. so, this is the main limitation of the study
(Sheikhtaheri et al., 2021)	Prediction of neonatal deaths in NICUs: development and validation of machine learning models	ANN, RF, CHART, SVM, and ensembles with the highest accuracy of 94 % SVM	Focus on neonatal death and more focus on machine learning application in neonatal domain than a prediction of the diseases and used imbalanced dataset

By considering, the gap listed under the related work in the table above along with neonatal mortality and morbidity challenges in Africa particularly Ethiopia, this study comes up with a solution for already identified problems, to overcome neonates, and health expert's challenges by automating neonatal disease prediction using ensemble machine-learning techniques. Most of the existing neonatal disease was done to identify the presence or absence of only one disease, a small dataset as well as statistical research based on only the NICU dataset. Therefore, this research will focus on building machine learning models to solve the problems phased by WHO, FMOH, and health experts in determining and predicting neonatal diseases. Thus, the study outcomes can also reduce mortality and morbidity, save time, and effort, and generally improves accuracy for neonatologist or patrician in neonatal disease prediction with the real-world dataset.

CHAPTER THREE

3. RESEARCH METHODOLOGY

In this chapter overall methodology and the procedure of developing a neonatal disease prediction model based on determinant attributes using machine-learning techniques are discussed. Machine learning is applied in the health sector nowadays to manage a huge amount of data, which is complex for humans to analyze easily. To build these prediction machine-learning models different methods, techniques and procedures have to be used in each phase. Moreover, in this chapter data collection methods, preprocessing techniques, and machine learning models used in designing a neonatal disease prediction model are explained. Finally, the research performance evaluation metrics are described.

3.1. Research Design

Research design is a procedure that the researcher follows to collect data, and preprocessing to answer the research questions. Research design may be quantitative, qualitative, or hybrid research designs. This study implemented an experimental research design approach because the quantitative research design approach finds the relationship between two variables i.e., independent variable and dependent variables within the data samples, and involves identifying research objectives and building machine learning models to prove the concept(Kamiri & Mariga, 2021).

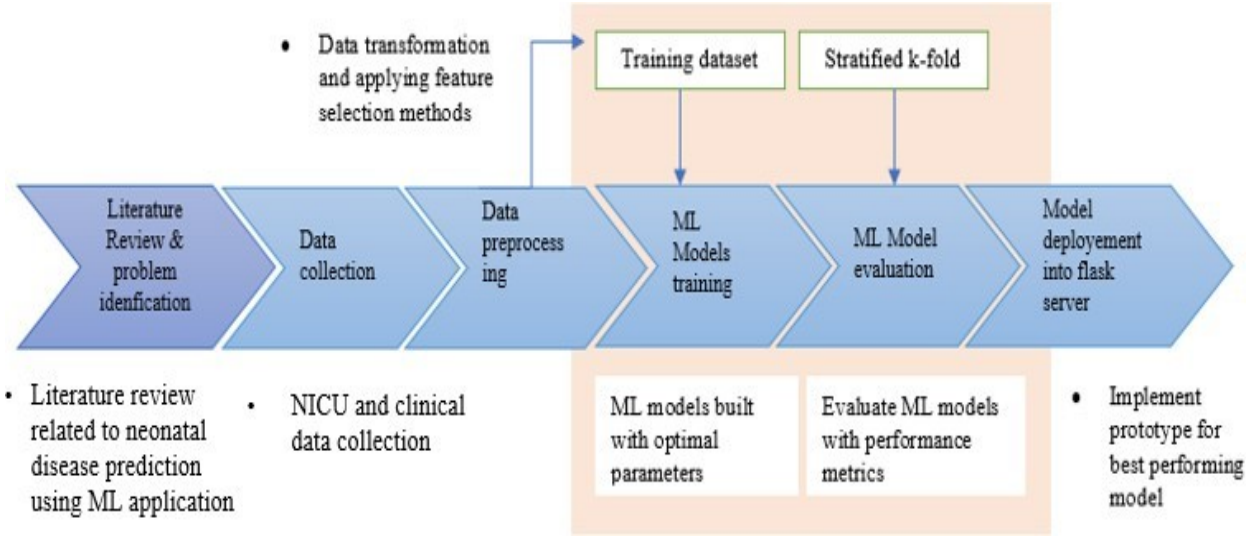

Figure 3:1 Flowchart of methodology for neonatal disease classification

3.1.1. Problem Identification

It is an important step of this study, which can support the understanding of the research area or problem domain. To start the actual machine-learning task, we need to be able to define the problem in a well-defined way and have a good understanding of the dataset used in the machine-learning model. Because the purpose of machine learning techniques is determined by the type of problem that needs to be solved and the data (Singh & Chauhan, 2020). In this study, the review of different publications that focus on machine learning applications in health care was used as a source of support. In addition, the discussion with domain experts was carried out; relevant documents were analyzed in depth to raise awareness about neonatal diseases. However, the investigator is unable to find published research on the prediction of neonatal diseases, namely NEC, sepsis, respiratory distress syndrome, and perinatal asphyxia in the Ethiopian context. Consequently, the researcher was motivated to develop an ensemble machine-learning model from SVM, RF, and XGboost.

3.1.2. Data Collection

Data used for this research were taken from newborn patients' cards admitted to the NICU that attends their treatments or died during the period 2018 to 2021 from Asella Compressive Hospital, Asella, and Oromia, Ethiopia. In addition, different features to diagnose neonatal disease are identified by asking the neonatologist and pediatrician, assessing different local and global references on neonatal disease, and considering online available features of the neonatal disease dataset. The data were collected by reviewing the patient history in the NICU by two nurses working in the NICU. The collected data have 20 features 19 independent and 1 dependent feature with a 4-target class. Asella Compressive Hospital is the largest public and teaching hospital in Ethiopia, which is administered by Arsi University and responsible to the Federal Minster of Health of Ethiopia. According to a different source of information, Asella Compressive Hospital is serving many neonatal patients including patient from Bale and Arsi zones, and have its own pediatric and childcare center with pediatrician expert. In addition, the hospital opens up for the service and has a good opportunity for researchers and policymakers for program monitoring and evaluation.

3.2. Neonatal Disease Classification Modeling

3.2.1. Dataset Building

The main objective of this study is to classify neonatal diseases using machine-learning techniques that classify neonatal diseases for common diseases. To achieve the aim of this study the data was taken from Asella Compressive Hospital and should be preprocessed because it was originally raw data, and not suitable for machine learning model training and testing. To build a dataset for neonatal diseases model development raw dataset, then preprocessing data and dataset labeling for neonatal diseases prediction is a key step.

3.2.2. Data Preprocessing

Preprocessing is a method of making the noisy, unclean, filling missing values, disorganization, and un-categorical data into a suitable form for machine learning models. Preprocessing is a very important step in machine learning for model training because quality datasets produce a good performance in training machine learning. Therefore, the dataset should have been preprocessed before it feeds into machine learning for model training. The data taken from Asella Compressive Hospital was firstly raw data, which contain incomplete, noisy, inconsistent, inaccurate, and irrelevant data points. The data preprocessing method includes cleaning noise data, handling missing values, filling missed data by its mean, mode, variance, or standard deviation, handling categorical data, and generally, it helps us to prepare the clean dataset for model training.

Cleaning Noise Data: This is a process of detecting inaccurate, incomplete, noisy, outliers and correcting errors in the datasets, which negatively affect the performance of the model, then it helps machine-learning models to get a quality dataset, that improves the accuracy of the machine learning model.

Handling Missing Values: Due to some mistakes such as unfilled patient history data by an expert or while collecting data some values are missing, which leads to missing diagnosis results and a great impact on model accuracy. Missing values in the dataset can be handled in several ways including dropping missing values if it has the least impact on individual instances, filling missing values, replaced by zero values which will bring no change to the model. Furthermore, missing datasets can be handled by data imputation, which replaces the missed values by their mean, mode, or standard deviation, predicting missed values from an available dataset.

Handling Categorical Data: in this step, collected data is transformed into the categorically required data format because the collected data consists of real data, decimals, and nominal values data. Therefore, this nominal data should be converted into numerical data of the form 0 and 1. For instance, the "Vomiting" attribute has nominal values that can be converted as 0- for 'No' and 1- for 'Yes'. In handling categorical data and after preprocessing the data should be in CSV file forms which hold the entire integer and float values data.

Handling Imbalanced Data: it is common in multi classes and the dataset collected for this study contains a slightly class imbalance that requires to be handled. In this study, the class imbalance problem is solved by setting the class weight to balance in hyperactive parameter setting for random forest algorithm, choosing ensemble algorithms like the RF method, which have a built-in approach to combat class imbalance, and SVM which addresses this by assigning different weights to positive and negative instances for imbalanced data. Using an under-sampling method, which may cause loss of useful information, and Over-sampling like SMOTE (Synthetic Minority Over-Sampling Technique) which adds noise to the dataset that increases the chance of model overfitting (Tahir *et al*., 2019). Therefore, without over-sampling, or under-sampling data random forest classifier and XGBoost can be aware of imbalanced data by assigning class weight to balanced so that it can penalize the majority class to have equal weight with the minority class (Zichen Wang, 2018), (Zhu *et al*., 2018). Handling imbalanced data: it is common in multi classes and the dataset collected for this study contains a slightly class imbalance that requires to be handled.

Feature Scaling: Feature scaling is an important task in many machine learnings because the range of feature values in the raw data varies (Bhandari, 2020). Feature scaling can help us to have equally scaled features that contribute equally to the result. However, some machine learning algorithms that exploit distance or similarity in the form of scaler products between data points, such as KNN and SVM, are sensitive to feature scaling. On the other hand, machine learning algorithms such as XGBoost (XGB), random forest (RF), and Naïve Bayes are insensitive to feature scaling (Brownlee, 2020).

In this study from different feature scaling methods, the StandardizeScaller is used in which how every XGB and RF do not affect by feature scaling but support vector machine affected by feature scaling. Feature standardization is removing the mean and scaling to unit variance.

StandardizeScaller is done as follows:

$$\mathbf{X' = \frac{X-u}{s}} \qquad \text{Equation 3:1}$$

Where:

X= score of a sample

u and s = are the mean of the training samples and s is the standard deviation respectively

3.2.3. Feature Selection

It is very important to identify the most relevant features to generate the best results from data using the machine-learning algorithm. Features selection is the process of selecting the most effective features from a large number of input variables or original datasets because, in reality, all the features from the original dataset may have less important for machine learning predictive modeling. Therefore, the goal of feature selection is to select a relevant and essential set of features to reduce the computational cost of the machine learning predictive modeling to enhance the effective performance of the model. In addition, features selection is related to reducing the dimension of the dataset or dimension reduction, which seeks fewer for modeling However, the dimensionality reduction method is reducing the number of features by creating a new combination of attributes, whereas feature selection methods are a way of including, or excluding features present in the dataset without changing them. Filter, wrapper, and embedded feature selection techniques are the three core classes of feature selection methods used in different clinical datasets(Sun et al., 2019),(Phyu & Oo, 2016),(Awada et al., 2012)(Remeseiro & Bolon-Canedo, 2019).

Filter methods are one of the best classes of feature selection techniques that pick the most appropriate columns of features independent of the machine-learning algorithm, and this method measures the significance of features by their correlation with the target variable.

Wrapper Methods are another commonly used feature selection methods to select a set of most relevant features, or columns as a search problem, where different combinations are already prepared, estimated, and compared to other combinations. A model-based machine-learning algorithm is used to assess the group of features and assign a score based on model accuracy. This feature selection technique examines the importance of each subset of a feature or column by actually training a model on it.

Feature selection methods such as univariate from filter method, Recursive Feature Elimination, and sequential forward selection from wrapper method are best and easy to use (Shaikh, 2018). Recursive Feature Elimination (RFE) is a prevalent and very common feature selection algorithm. Because it is easy to configure, and use, very effective at selecting those features in a training dataset that are most relevant in predicting the target variable (Abidin *et al*., 2020). It can be used with cross-validation. In this study before applying these feature selection techniques, categorical data should be transformed into numeric using label encoding since chosen feature selections are used only for the numerical dataset.

3.2.4. Machine learning Models

This study aims to propose a neonatal disease prediction model using base learning models and combine them to form better-performed predictive models. To propose the ensemble learning models SVM, RF, and XGB are combined with and without features selection methods.

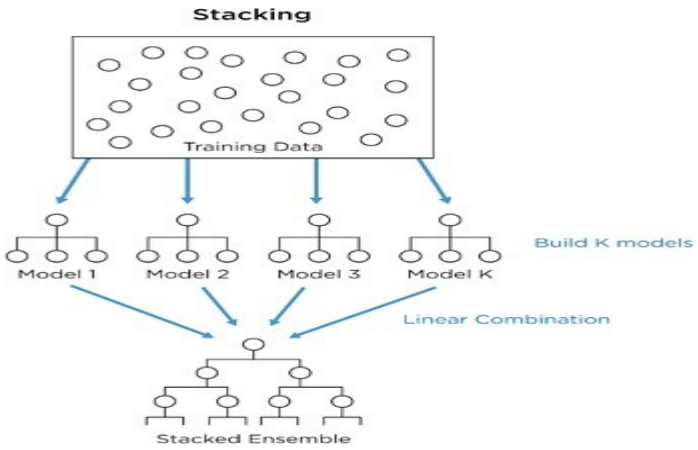

Figure 3:2 stacking ensemble architecture adopted from[12]

3.3. Predictive Model Evaluations

It is an effective step, very crucial, and an integral part of the machine learning model development process to estimate the performance of predictive modeling and used to measure how well the chosen model will perform on unseen data. It is intended to evaluate the generalization accuracy of the model on unseen data. The performance evaluation method may be holdout and cross-validation. Holdout evaluation aims to test a model on different data that it was trained on which provides an unbiased estimate of learning performance. In holdout,

[12] *https://medium.com/ml-research-lab/stacking-ensemble-meta-algorithms-for-improve-predictions-f4b4cf3b9237*

method data is randomly divided into a training set, validation set, and test set. Cross-validation is the technique used for evaluating model performance on a test dataset that is not seen by the model, which helps to evade overfitting in predictive models especially in the case when the amount of data may be very limited.

3.3.1. Cross-Validation

Holdout: This is a basic approach in which a large dataset is randomly separated into two subsets such as training and test sets. The training dataset is used to build machine learning models, whereas the test dataset is unseen instances among the subset of the dataset to access the future performance of the models (Dantas, 2020). The holdout method is better when there is a large dataset.

K-fold Cross-Validation: A technique that assists us to estimate the effective performance or accuracy of machine learning models. This technique is also used to compare several machine-learning models so that it helps us to pick the most appropriate machine-learning model that can give the best accuracy score. Therefore, machine learning models can be built with the dataset, and it can be checked how much the model could perform using this method on unseen data. It is a cross-validation method with the following steps. As such first, split the entire dataset into k-number of folds. For each fold of the dataset, build the model on k-1 folds of the training dataset. Then, test the model to assess the performance of the kth fold. This step is repeated until each of the k-folds has been used as the test set. The average k-scored accuracy is called cv accuracy and served as a performance metric of the model (Dantas, 2020).

Stratified K-Fold Cross Validation: Since k-fold CV randomly shuffles the dataset, and divided it into folds, the probability to get highly imbalanced folds might be high that causing model training to be biased. It might work well for data with a balanced class distribution. However, when class distribution is imbalanced, even if it is too small, one or more folds will have few or no samples from the minority class. This causes model evaluation to be misleading, as the model needs only to predict the majority class correctly. The stratified form of k-fold CV that conserves the imbalanced class distribution in each fold is called stratified k-fold CV, and it will enforce the class distribution in each split of the data to match the distribution in the complete training dataset (Dantas, 2020). Since the data collected, contain imbalanced class distribution, the stratified k-fold cross-validation technique is used for this study.

Figure 3:3 Algorithm for the 10-Fold Cross-Validation

3.3.2. Performance Evaluation Metrics for Predictive Modeling

The idea of building a predictive model using a machine-learning algorithm works on the principle of constructive feedback. Build the model using a machine-learning algorithm, get feedback on evaluation metrics, make improvements, and continue until you get the desired accuracy. Performance evaluation metrics can evaluate the performance of the model. An important component of evaluation metrics is their ability to distinguish between the results of different models. There are several types of predictive model performance evaluation metrics used in this study to evaluate the performance of the selected predictive modeling, such as accuracy, precision, recall, f1-score, and confusion matrix. When classification is performed, four types of results can emerge which are listed below.

- **True positives** (TP) have occurred when predicting an instance belonging to a class and it is classified correctly, i.e., when the predicted value is equal to the actual value. The actual value was positive and the model predicted a positive value.
- **True negatives** (TN) have occurred when the model predicts an example that does not belong to a class and it does not belong to that class, i.e., the predicted value is the same as the actual value. The model predicts the value as negative, and the actual value was negative.
- **False positives** (FP) have occurred when the predicted instances belong to a class when in reality it does not, which means the predicted value was wrongly predicted. The true value was negative, but the model incorrectly predicted it as positive.

- **False negatives** (FN) have happened when the model predicts examples do not belong to a class when in fact it does, i.e., the predicted value was mistakenly predicted, the actual value was positive, but the model predicted as negative.

Accuracy (ACC) is the most common and popular performance evaluation metric for classification tasks and is defined as the percentage of correct predictions for the test dataset. It can be calculated by dividing the number of correctly classified by the number of total predictions.

$$\text{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \qquad \text{Equation 3:2}$$

Precision is the ratio of the entire number of true positive results to the number of positive results predicted by the classifier model, and it is how close a measured value is to each other.

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \qquad \text{Equation 3:3}$$

Recall is the total number of true positives divided by the number of all relevant samples, or it is the ratio of true positives to true positives plus false negatives.

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \qquad \text{Equation 3:4}$$

F1-score is the harmonic mean of the precision, and recall, whereas the F1-score reaches its best values at 1 i.e., perfect precision and recall, and worst at zero.

$$\text{F1 Score} = 2 * ((\frac{Precision*Recall}{Precision+Recall})) \qquad \text{Equation 3:5}$$

Confusion Matrix is a method that summarizes the classification performance of the models concerning some test datasets, and the outcome can be two or more classes. It is used to classify the data in a matrix format which is indexed in one dimension by the true class of an object, and the class that classifier assigns in the other dimension. The confusion matrix was used with four class classifications in this study in which neonatal diseases were classified as, RDS, sepsis, NEC, and PA.

Table 3-1 Confusion matrix for four-class classification

		Predicted values			
Actual Values		Sepsis	RDS	PA	NEC
	Sepsis	True Sepsis	False RDS	False PA	False NEC
	RDS	False Sepsis	True RDS	False PA	False NEC
	PA	False Sepsis	False RDS	True PA	False NEC
	NEC	False Sepsis	False RDS	False PA	True NEC

CHAPTER FOUR

4. PROPOSED SOLUTION FOR NEONATAL DISEASES PREDICTION

In this chapter, the proposed neonatal prediction model using ensemble a machine learning approach to classify neonatal diseases was discussed in detail in consideration of solving the issues discussed in chapter one and finding the answer to the stated research questions following to fill the gap discussed in the literature review. The high-level architecture of the proposed solution is illustrated below to show the details of each component that exists within the proposed framework.

4.1. Proposed Neonatal Disease prediction Model Architecture

The idea behind this research study is to design an ensemble of a machine-learning model for neonatal disease prediction that leads to a generous improvement in classification of neonatal diseases. Which classify the neonatal disease as necrotize enter colitis(NEC), sepsis, respiratory distress syndrome(RDS), and parental asphyxia(PA) based on the neonatal dataset collected from Asella Compressive Hospital.

The methodology attempts towards exploring the power of machine learning algorithms in predicting neonatal disease. Preprocessing platform, algorithm platform, and training platform, as shown in the following figure. The proposed architecture of this study designates that, the collected data was initially raw data, and undergoes preprocessing stages such as data cleaning, handling missing values, transforming data, and handling categorical values to make the dataset suitable for the machine-learning algorithm. After the dataset is preprocessed, the next step is selecting appropriate feature selection techniques to identify the capability of the relationship between targets and features and in this, recursive feature elimination with cross-validation is used. Then preprocessed data is fed into the selected machine learning models i.e SVM, RF, and XGB with and without the feature selection method, and the result of the three selected models is combined to form a stacking ensemble which is the new proposed model for neonatal diseases prediction. Finally, the performance of the models is measured with and without feature selection methods under stratified k-fold cross-validation with k=10 with other performance metrics to figure out the most performed model (Novaes *et al*., 2021).

Different performance evaluation metrics are used to get the most performed model, which can give the best future prediction on the test dataset. After all of these stages are completed, model deployment is the next, and final step. Deployment is a method that is used to integrate a machine learning predictive model into the existing production environment to give a real-world decision based on data to help domain experts, pediatricians, and patients. Hence early disease prediction can reduce the mortality and the risk of morbidity of neonates.

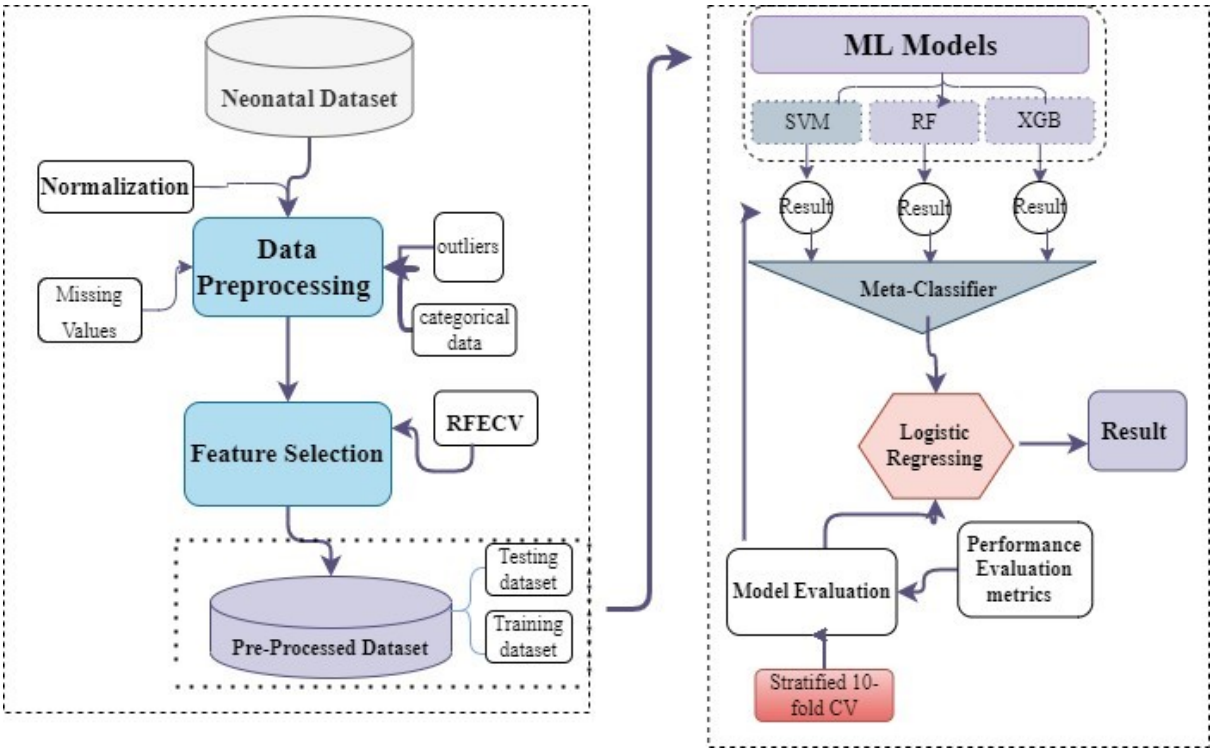

Figure 4:1 Proposed Neonatal Disease prediction Model Architecture

4.2. Understanding of the Data

After defining the problem, the next step is data understanding to recognize the accessible dataset. The quality and quantity of available datasets have a significant impact on the performance of machine learning and knowledge discovery. This phase mainly focuses on collecting the target dataset and deciding the type, format, and number of the datasets. Datasets are checked for completeness, redundancy, missing values, plausibility of feature values, etc. Finally, the step includes the effectiveness of the dataset concerns machine learning techniques(Eaton, 2018) There is no database in Asella Compressive Hospital. As the hospital keeps the record of each patient in a manual format and it was needed to register each instance of data into Microsoft Excel 2016 for further analysis activity and getting ready for input to the

Anaconda Jupiter notebook tool. Three years of datasets were collected from Asella Compressive Hospital. The total size of the initial datasets in excel format is 233KB

The datasets used for the prediction of neonatal diseases were collected from Asella Compressive Hospital, Ethiopia, from 2018 up to 2021 with a total number of 2298 instances and 20 features, whereas 19 were independent features, and 1 dependent feature (including four class labels). The exam results of one patient are represented by a piece instance in the dataset, which is collected, from the discharge summary and physical examination of neonatal disease. The registered datasets include admission information, delivery information, symptoms, and laboratory and x-ray results of the patients as listed below.

Temperature: - the temperature is the specific degree of hotness or coldness of the body. Body temperature is measured by a clinical thermometer and represents a balance between the heat produced by the body and the heat it loses. In the collected dataset, the value of body temperature was scored in degrees Celsius (°C).

Figure 4:2 Thermometer shows the body temperature adopted from [13]

Respiratory Rate (RR): - it is the number of breaths a baby takes per minute or, more formally,
the number of movements indicative of inspiration and expiration per unit time. The respiratory rate is usually determined by counting the number of breaths for one minute by counting how many times the chest rises. The aim of measuring respiratory rate is to determine whether the respirations are normal, abnormally high, or low. The normal respiration rate for a newborn at rest is 30 to 60 breaths per minute (bpm) (Eichenwald et al., 2021a)

[13] *https://flo.health/being-a-mom/your-baby/baby-health-and-safety/newborns-temperature*

Vomiting: - It is the forceful throwing up of stomach contents through the mouth. **Grunting: -** it is an expiratory sound caused by the sudden closure of the glottis during expiration
in an attempt to increase airway pressure and lung volume and prevent alveolar atelectasis (Eichenwald et al., 2021b).

Reflexes: - neonatal reflexes or primitive reflexes are the inborn behavioral patterns that develop
during uterine life. These reflexes are important for a newborn's survival immediately after birth, including sucking, rasping, blinking, Moro, rooting, and asymmetric tonic neck reflex(Eichenwald et al., 2021b).

Blood Cultures: - It is the gold standard to detect the presence of bacteria or fungi in the blood, to identify the type present, and to guide treatment. Testing is used to identify a blood infection (septicemia) that can lead to sepsis, a serious and life-threatening complication(Ershad et al., 2019).

Crying: - crying is a newborn's main way of telling you what they need. Sound can spur you into action, even when you are asleep. Infants normally cry for about one to three hours a day. It is perfectly normal for an infant to cry when hungry, thirsty, tired, lonely, or in pain(Eichenwald et al., 2021).

APGAR score: - is a rapid test that is done on a baby between 1 and 5 minutes after birth. With this score, the system provides an accepted and convenient method of reporting the status of the newborn immediately after birth. The APGAR score consists of five components: Appearance, Pulse, Grimace, Activity, and Breathing. A score of six to ten is normal and is a sign that the newborn is in good condition health. Any score below six is a sign that the baby needs medical attention. (Antonucci et al., 2014).

Gestational Age (GA): - Gestation is the period between conception and birth. During this time,
the baby grows and develops inside the mother's womb. It is measured in weeks, from the first day
of the woman's last menstrual cycle to the current date. Neonates can be classified based on gestational age such as Pre-term: less than 259 days (37 weeks), the term: 259–293 days (37–41 weeks). Post-term: 294 days (42 weeks) or more (Arcangela Lattari Balest, 2021).

Seizures: - a seizure is caused by sudden, abnormal, and excessive neural activity in the brain. By definition, neonatal seizures occur during a full-term infant's neonatal period, the first 28 days of life. Many of the visible signs of neonatal seizures, such as chewing movements, unusual cycling, random or errant eye movements, and beatings (Victorio, 2021).

Resuscitate: - the newborn is resuscitated when different indications such as the heart have occurred like a heart rate that remains below 100 beats per minute (bpm) for thirty seconds, poor respiratory failure activity manifested by panting, bluish skin, and poor muscle tone (Antonucci et al., 2014).

Oxygen Saturation (SpO2): -measures the amount of oxygen-carrying hemoglobin in the blood relative to the amount of hemoglobin that does not carry oxygen. The human body requires and regulates a very precise and specific balance of oxygen in the blood. Normal arterial blood oxygen saturation levels in newborns are 90% to 100%. SpO2 < 90percentage can result in very serious problems in neonates (Arcangela Lattari Ballast, 2021. Pulse oximetry is a test used to measure the level of oxygen (oxygen saturation) in the blood. It is an easy and painless measure of how well oxygen is delivered to the parts of the baby's body farthest from the baby's heart, such as the arms and legs.

C-reactive protein (CRP): - a C-reactive protein test measures the level of C-reactive protein in the baby's blood. CRP is a protein made by the baby's liver. It is sent into the baby's bloodstream in response to inflammation. Inflammation is the baby's body's way of protecting tissues if the baby has been injured or has an infection. CRP levels < 7 mg/L are considered normal in newborns while 7 to 20 mg/L as cut-off levels indicate the presence of sepsis or another infection (Bunduki & Adu-Sarkodie, 2020).

White Blood Cells (WBC): -a white blood cell count is a blood test to measure the number of white blood cells in the blood. White blood cells are also called leukocytes. They help fight infections. On the day of birth, a newborn has a high white blood cell count, with normal values for total WBC ranging from 8,000/μL to 20,000/μL or per microliter. A low white blood cell count (< 8,000/μL) is associated with an increased odds ratio for sepsis than an elevated white blood cell count (>20,000/μL) (Bunduki & Adu-Sarkodie, 2020b).

Low Lung Volumes and Whitening (LLVW): - a premature newborn born at less than 37 weeks gestation has the added complication of achieving these changes with relatively immature lungs. Lung bleaching occurs when the black in a lung field is replaced by white,

indicating that there is less air entering the alveoli, perhaps due to mechanical obstruction, increased fluid, consolidation, or other causes (SABEL & WADSWORTH, 2016)

Intercostal Subcostal Retractions (ICSCR): - Occurs when lung compliance is poor or the airways resistance is high, as a result of the negative intrapleural pressure generated by the contraction of the diaphragm and accessory muscles of the chest wall. Retractions, the inward movement of the skin of the chest wall or inward movement of the breastbone (sternum) during inspiration, is an abnormal breathing pattern. Intercostal retractions are an inward movement of the skin between the ribs. While subcostal retractions are an inward movement of the abdomen just below the rib cage (Van Haeringen et al., 2019).

4.3. Data preprocessing

It is a data preparation technique, such as data cleaning, filling in missing values, organizing raw data, handling categorical values, and data transformation to make it a suitable form for machine learning models. Here, the already prepared dataset is converted into a common separated value (CSV) format. CSV extension since the python CSV module gives the python programmer the ability to parse the CSV formatted file for further preprocessing using python code in machine learning. The data collected for this study contains some inconsistent data that need to be cleaned, and few missing values that need to be filled, data to be transformed for ensuring better data quality and categorical values to be handled. The data preprocessing flow diagram is shown below.

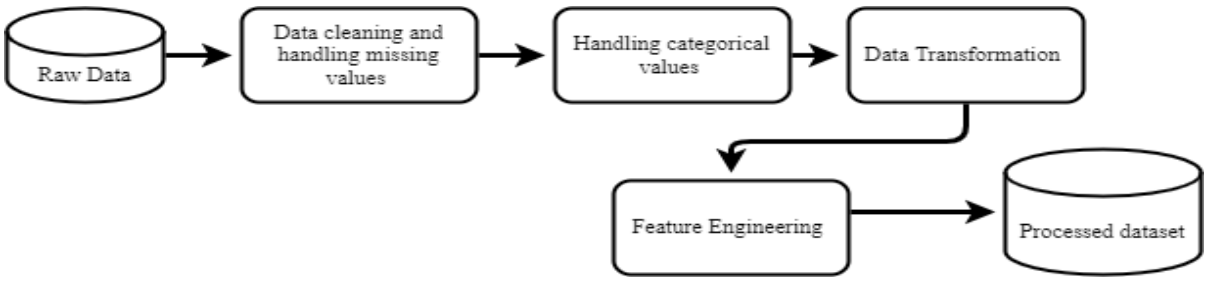

Figure 4:3 Dataset Preprocessing step

4.3.1. Data Cleaning and Handling Missing Values

Data cleaning is the task included in the data preprocessing that is used to fix, or eradicate incorrectly formatted or incomplete data found in a dataset. Hence, the collected data for this study contains too few incorrectly formatted and inappropriate data that need to be removed. In addition to incorrectly formatted data present within the dataset, it also contains missing

values. Missing data are a common problem that can affect the data analysis process. Missed values in the dataset will minimize the accuracy of a classification and affect the models to be produced from the classification algorithm. The dataset collected from ACH has total records of 2298 and 20 features including the class label. Table 4.1 shows, the name of the features, the number of missing values, and appropriate imputation techniques. Based on this, the mean values of numeric attribute that contains missing data were used to fill in the missing values. In the case of a categorical attribute, the mode, which is the most frequent value, used to fill missed values

Table 4-1 Missing Value Imputation

Attribute Name	**Data Type**	**Number of Missing Values**	**Missing values in %**
AF	Nominal	14	0.61%
Grunting	Nominal	7	0.3%
WBC	Numeric	12	0.52%
Vomiting	Nominal	23	1%
LLVW	Nominal	3	0.13%
Respiratory Rate (RR)	Numeric	120	5.22%
Blood cultures	Nominal	53	2.31%
C-Reactive Protein (CRP)	Nominal	25	1.1%
Grunting	Nominal	49	2.13%
Temperature	Numeric	22	0.96%
Weight	Numeric	182	7.92%
Heart Rate	Numeric	295	12.84%

4.3.2. Handling Categorical Data

In this study, categorical data is converted to a numeric datatype to make it suitable to build machine-learning models. Therefore, in this study, the label encoder encodes labels with a value between 0 and n_classes-1 where n is the number of distinct labels, in this if the label repeats it allows the same value as assigned earlier. The following are categorical data

Table 4-2 Categorical data

Attribute Name	Data Type
AF	Nominal
Grunting	Nominal
Vomiting	Nominal
LLVW	Nominal
Blood cultures	Nominal
C-Reactive Protein (CRP)	Nominal
Grunting	Nominal
Seizures	Nominal
Reflexes	Nominal
GA	Nominal
SpO2	Nominal
ICSCR	Nominal
APGAR score	Nominal
Crying	Nominal

4.4. Feature Selection

After data cleaning, handling missing values and handling categorical values the next step is to analyze the most relevant features that contribute to boosting the performance of models and remove the features that increase the noise in the models. Feature selection is aimed to reduce the number of features to those that are believed to be the most convenient to enhance the predictive power of machine learning models, and reduce the overfitting, and training time of the models. Hence, recursive feature elimination with cross-validation(RFECV) is used for this study.

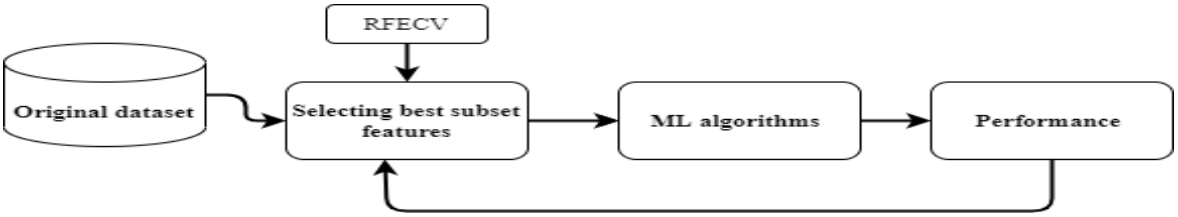

Figure 4:4 Feature Selection workflow diagram

Feature Elimination with cross-validation (RFECV)

It is the backward feature selection method of the predictors that builds a model's overall set of features and calculates an importance score for each feature. The weakest predictor with the smallest important scores is removed, the model is re-trained, and importance scores are computed again. Then, the number of feature subsets to estimate each subset's sizes identified[14]. Therefore, the subset size is used as a parameter tuning for RFE. The feature's subset size improves the criteria used to choose the best feature(s) based on the feature's importance score rankings. The best subgroup is then used to train the final machine learning predictive model. To get the best set of features from the original features set, cross-validated RFE was used to score different feature subsets and select the best scoring set of features

4.5. Machine Learning Model Building

The objective of this study is to build automated neonatal disease prediction to identify whether the disease is NEC, sepsis, RDS, or perinatal asphyxia based on the given features both clinical and instrumental diagnosis parameters using machine learning algorithms. Therefore in this study stack ensemble, machine learning models are developed by combining XGB, RF, and SVM. The model is evaluated using stratified 10-fold CV, and other performance metrics to identify the best performing model for final neonatal diseases prediction and deployment.

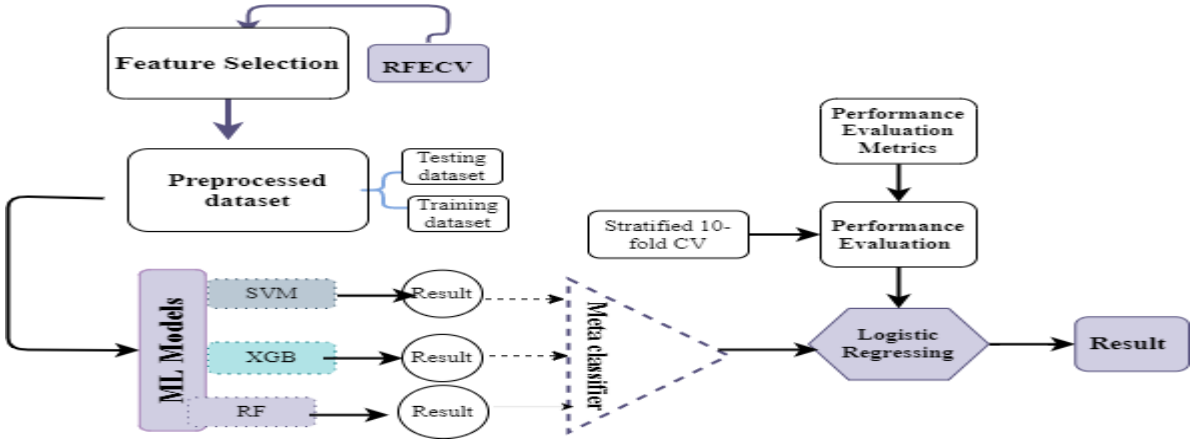

Figure 4:5 Model building flow diagram

Support Vector Machine (SVM) Classifier Algorithms

Support vector machines also called supporting vector networks, are the collection of similar supervised learning methods used for classification and regression. The SVM training

[16] *https://machinelearningmastery.com/rfe-feature-selection-in-python/*

algorithm is the non-probabilistic, multiple, linear classifier. The support vector machine aims to find an optimal boundary between possible outputs. This study proposed the One-Vs-Rest (OVR) strategy of the Support Vector Machine (SVM) machine learning classifier because a simple form of the SVM algorithm is a binary classifier. To classify multiple groups of classes, we applied an OVR method along with the SVM algorithm, which is used, for multiple class classifications. This method separates each class from the rest of the classes in the dataset. In this study, one-vs-rest classification is usually preferred, since the results are mostly similar, but the runtime is significantly less[15].

Support vector machine constructs a hyper-plane or set of hyper-planes in a high dimensional space, which can be used for classification tasks. If the hyperplane has the largest distance to the nearest training data points it has good separation for classification i.e. the larger margin the lower the error for the classifier.

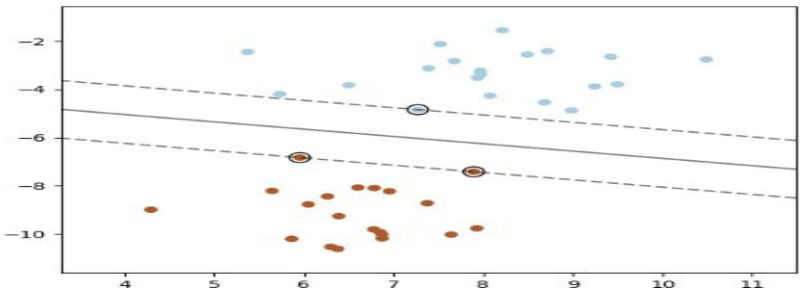

Figure 4:6 Decision function for linearly separable problem

Random Forest Classifier Algorithm (RF)

Is an ensemble of classifiers, usually composed of a decision tree which is used for solving both classification and regression problems, and is used to build neonatal diseases, classification model[16]. As the name indicates this algorithm randomly creates a forest with several decision trees. The higher number of trees in the forest, the greater the accuracy result. The working principle of RF is, first it selects random K data points from the training dataset. Second, it builds decision trees associated with the designated data points. Third, it chooses the number N for the decision trees that wanted to build, then repeats steps 1 & 2. In addition, it finds the predictions of each decision tree, and assign the new data instances to the category that wins the majority votes11 and the detail of the random forest algorithm, major difference

[15] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[16] *https://machinelearningmastery.com/rfe-feature-selection-in-python/*

between Bagging and Boosting ensemble methods is explained in (Zhang & Haghani, 2015). In that, Random Forest is a bagging method, whereas XXGBoost is an ensemble boosting method.

XGBoost (XGB) Classifiers

XGBoost is an extended version of gradient boosting decision trees designed for the speed and performance of machine learning. Xgboost is used for both classification and regression tasks in machine learning techniques[17]. Some necessary features of XGboost are the following

- ✓ Parallelization:- implemented on multiple CPU cores to train
- ✓ Regularization: Xgboost uses different regularization to avoid overfitting
- ✓ Non-linearity: can learn to form non-linear data
- ✓ Cross-validation:- built-in
- ✓ Scalability

Stacking ensemble

Stacking is an extended form of machine learning technique in which all base models equally participate as per their performance weights and build a new model with better predictions[18].

To implement stack ensemble learning we follow the following steps

- splitting data into training and testing
- Predict the individual model performance
- adds individual model prediction to meta-algorithm
- Repeat steps 2 and 3. Now train the Meta model on level 1 prediction, where these predictions will be used as features for the model.
- Finally, Meta learners can now be used to predict test data in the stacking mode

4.5.1 Hyperparameter Tuning

Method of choosing a set of hyperparameters for a machine learning algorithm that allows the model to optimize its performance. Setting hyperparameters in the right way is the only method to extract the most performance from your models. This can be done with manual or automatic hyperparameter configuration. In the first method, different groups of hyperparameters are set and experimented with manually. This is tedious and may not be practical in cases where there

[17] *https://neptune.ai/blog/xgboost-everything-you-need-to-know*
[18] *https://www.javatpoint.com/stacking-in-machine-learning*

are many hyperparameters to try (Liashchynskyi, 2019). And in this last method, the best hyperparameters are found using an algorithm that automates and optimizes the process. The second method is used in this study. Grid search and random search are the two most commonly used algorithms for hyperparameter optimization, as grid search is a traditional method of hyperparameter optimization that simply performs a complete search on a given subset of the algorithm's hyperparameter space. Training and exhaustively generates candidates from a specific grid of parameter values and this method suffers from high dimensional spaces, but can often be easily parallelized as the hyperparameter values the algorithm works with are usually independent of each other (Liashchynskyi, 2019). While random search overrides the entire selection of all combinations by their random selection. It outperforms a grid search, particularly if only a few hyperparameters affect the performance of the machine learning algorithm (Liashchynskyi, 2019), and this technique is used for this study.

4.6. Model Evaluation and Prototyping

Model evaluation is used to estimate the generalization accuracy of the predictive model on the out-of-sample dataset. Therefore, stratified k-fold CV with k=10 is used to identify the best-performing model for neonatal disease prediction. Therefore, the model is trained with the entire dataset and its performance is assessed using stratified 10-fold cross-validation with other performance metrics which are appropriate to the model such as confusion matrix, accuracy, precision, recall, and f1-score as discussed in chapter three in section 3.4.2

4.6.1. Prototype Development

After building a machine-learning model to automate neonatal disease prediction, the next step is to understand how the model works in the real world to predict unseen datasets. Once a model is trained and evaluated, the best performing classifier is deployed to the flask server using HTML Webpage. So, it can be used by health experts, pediatricians, or neonatologists to make accurate, and automatic predictions once feature data are given to it. It is fast, and accurate compared to the conventional neonatal disease classification which requires effort, and is time-consuming

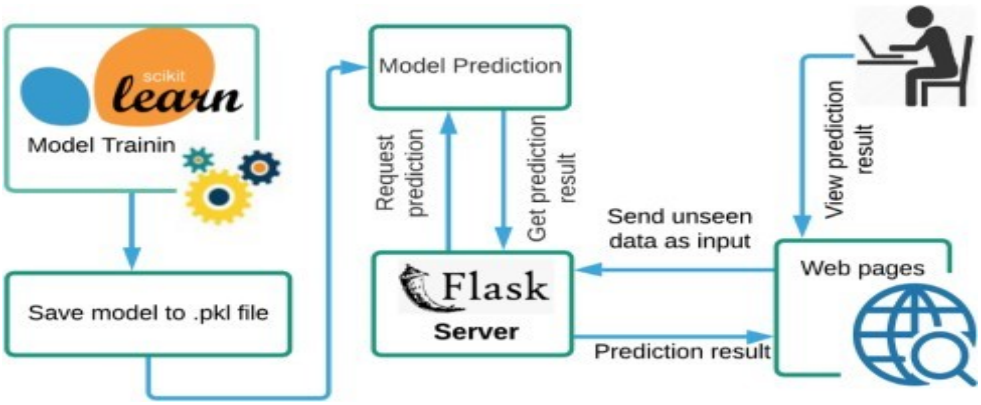

Figure 4:7 Deployment Architecture for Neonatal disease prediction

CHAPTER 5

5. IMPLEMENTATION AND EXPERIMENTATION

In this chapter, how the models to be evaluated are implemented, and the tools needed for experimentation to achieve the proposed framework for neonatal disease classification is discussed. The study used is to help medical experts, health sectors, and patients in diagnosing the class of neonatal diseases. Before the actual implementation of the models, and deploying the model into the flask server, different software packages, and tools used are discussed in the following sections.

5.1. Development Tools

In this study, several tools and packages are used to implement the proposed solution. Python programming language has used this study for implementing and experimenting with each proposed solution from the data preprocessing to the final phase, which is the model-building phase and prototype development. The reason to choose python is that it is popular and it is the language of choice for developers, researchers, and data scientists that make the machine learning models development simple and clear. Besides, Python is a general-purpose programming language that can be used for processing text, numbers, images, and scientific data easily. It is also simple and easy to understand.

Table 5-1 Description of the Tools and Python Packages Used During the Implementation

Tools and packages	Version	Description
Anaconda navigator[19]	2.1.0	Help us to launch development applications and easily manage anaconda packages, environments, and channels without the need to use command-line commands.
Python[20]	3.8.8	Python is a popular platform easy to learn, powerful programming language to develop a machine learning application
Jupiter Notebook[21]	6.3.0	It is a free web application that enables us to create and share documents. It is useful for data cleaning and

[19] Anaconda Navigator — Anaconda documentation
[20] https://www.python.org/doc/essays/blurb/
[21] https://jupyter.org/

		transformation, modeling, data visualization, and machine learning
Scikit-learn[22]	0.24.1	Is a python module used in machine learning for handling basic ML tasks clustering, regression, and classification
Pandas	1.2.4	Pandas is an important package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data. It enables integrating and filtering of data, as well as loading it from other external sources like Excel and handling the dataset.
NumPy	1.20.1	Array processing for numbers, strings, and objects. This study uses it for handling converting the text to numeric data for features, training, and testing the model.
Matplotlib	3.3.4	Publication quality figures in python. This study uses it for data and results visualization.
Seaborn	0.11.1	Statistical data visualization
Flask server[23]	1.1.1	Micro framework for making web services in python. This study uses it for implementing the prototype for selected models
Microsoft Word	2016	For documentation purpose
Microsoft Excel	2016	or dataset preparation

5.2. Experimentation Environment

After completing the aforementioned tasks, the next procedure is initiating the experiment with the selected algorithm. In the progress of this study, Python version 3.8.8 is utilized for algorithm training and designing a model. Python is a programming language that is easy to learn. And it is unique by being efficient in high-level data structures. Nowadays it is an ideal solution for different programmers, data analysts, engineers, etc. All the experiment is executed on a personal computer with a configuration of Intel® core ™ i7 CPU @ 1.88 GHz 1.99 GHz, 8.00 GB of installed memory (RAM), and 64-bit Microsoft Windows 10 Pro operating system are used in this study.

[2222] https://scikit-learn.org/stable/
[23] https://realpython.com/tutorials/flask

5.3. Dataset Description

The dataset used in this study is collected from Asella Compressive Hospital, Oromia, Ethiopia patient history data from a manually recorded document. First, the researcher assessed different information about the neonatal disease in Ethiopia and reviewed different material about the hospital, which works on neonatal disease treatment. Then Asella Compressive Hospital was selected to collect the data to prepare the dataset used for this study. Then the patient history data from the year 2018 to 2021 was collected by considering the identified attributes for the dataset preparation purpose. Due to a lack of full information and time, all-patient history data was not collected. This process results in a total number of 2298 patient data with NEC, sepsis, RDS, and PA. The dataset had missing values. By applying the data preprocessing method to fill the missing values, the final four-class labeled dataset contains a total of 2298 rows with 20 columns including one target class of data labeled as NEC, sepsis, RDS, and PA. The detailed method for building the dataset was discussed in chapters three and chapter four. In addition, the result labeling and size of each class in the dataset are presented in chapter six.

Table 5:2 Description of the neonatal disease dataset

No	Symbol	Feature full name	Type	Description
1	GA	Gestational Age	Nominal	Is the period between conception and birth and has values as s Pre-term, term and post-term
2	Temp	Temperature(^{0}C)	Numeric	It is a specific degree of hotness or coldness of the body of a newborn baby measured in Celsius(^{0}C)
3	RR	Respiratory Rate(bpm)	Numeric	It is the number of breaths a baby takes per minute
4	Reflexes	Reflexes	Nominal	Reflexes or primitive reflexes are the inborn behavioral patterns that develop during uterine life
5	APGAR score	Appearance, Pulse, Grimace, Activity, and Respiration score	Numeric	The APGAR score comprises 5 components: Appearance, Pulse, Grimace, Activity, and Respiration and score of 6-10 is normal, if the

				score is <6 baby needs medical attention
6	Resuscitate	Resuscitate	Nominal	Emergency procedures performed by a doctor to support newborns who are at risk
7	Seizures	Seizures	Nominal	Is caused by sudden, abnormal, and excessive neuronal activity in the brain
8	HR	Heart Rate(bpm)	Numeric	Is the number of times each minute that newborn heart beats which is normally between 70 to 190 bpm
9	SpO2	Oxygen Saturation (Pulse oximetry)	Numeric	It measures the amount of oxygen-carrying hemoglobin in the blood relative to the amount of hemoglobin not carrying oxygen and is measured in Pulse oximetry
10	Crying	Crying	Nominal	It's a sound that can spur you into action, even when you are asleep
11	CRP	C-Reactive Protein(mg/L)	Numeric	c-reactive protein test measures the level of c-reactive protein in baby blood and measured in mercury per liter
12	WBC	White Blood Cells(/µL)	Numeric	count is a blood test to measure the number of white blood cells in the blood measured in a microliter**(µL)**
13	LLVW	Low Lung Volumes and Whiteout	Nominal	complication happens due to changes in immature lungs. Lung whiteout occurs when the black in a lung field is replaced by white
14	Weight	Weight	Numeric	
15	ICSCR	Intercostal Sub Costal Retractions	Nominal	It occurs when lung compliance is poor or airway resistance is high
16	Grunting	Grunting	Nominal	It is an expiratory sound caused by the sudden closure of the glottis

17	Vomiting	Vomiting	Nominal	It is the forceful throwing up of stomach contents through the mouth
18	B C	Blood Cultures	Nominal	It is the gold standard to detect the presence of bacteria or fungi in the blood
19	AF	Antenatal Follow-up	Nominal	Supervision of humans during pregnancy to monitor the progress of fetal growth
20	Diseases	Class	Nominal	Target classes containing NEC, Sepsis, RDS and Perinatal Asphyxia According to WHO classification

5.4. Preprocessing Implementations

Preprocessing implementation includes handling missing values, handling categorical values, and feature selection. All of them are implemented using a python programming language as discussed in chapters three and four to make it suitable for the machine learning algorithm to build predictive machine learning models.

```
#importing necessary library & loading Unprocessed dataset
import pandas as pd
ND_Dataset = pd.read_csv ('ND_Dataset.csv')
```

Figure 5:1 Implementation of loading dataset using panda's library

5.4.1. Handling missing Values Implementations

To fill the missing values in the dataset, we have used the fillna() method that replaces the missing values by calculating the mean and mode value. The python Simple Imputer module is used to fill missing values using the mean strategy.

```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(missing_values=np.nan,strategy='mean')
imputer.fit(X)
X=imputer.transform(X)
```

Figure 5:2 Implementation of Handling missing values

5.4.2. Implementation of Handling categorical values

To build a correct and appropriate dataset and to have an accurate model none numeric data should be transformed into numerical data. Therefore, these categorical values are handled and implemented as discussed in chapter four and implemented as the following

```
from sklearn.preprocessing import LabelEncoder
lb_encoder=LabelEncoder()
categorical_column=ND_Dataset.columns.tolist()
for column in categorical_column:
    if ND_Dataset[column].dtype=='object':
        ND_Dataset[column]=lb_encoder.fit_transform(ND_Dataset[column])
```

Figure 5:3 Handling Categorical Data

After cleaning the dataset, the features split into independent and dependent features. Then the data without feature selection is used as X training and the labels, which contain the class or

labels dataset belong set as Y. Then, the classifiers train by using the X and Y of the entire dataset.

To train using the fit() method with the appropriate set parameter for each classifier instantiate the class.

```
 # Loading dataset and spliting it into independet and dependet features
import pandas as pd
ND_dataset = pd.read_csv (r'C:\Users\jo\Desktop\NeonatalD.csv')
#Independet and dependet variable X and y respectively
X=ND_dataset.drop(['Neonatal_DP'],1) # attributes to determine dependet variable /Labal/Class
y=ND_dataset['Neonatal_DP']#dependet variable
```

Figure 5:4 Splitting Dataset into Dependent and Independent features

5.5. Implementation of Feature Selection

Feature selection is a technique that selects the best features set among the original features to improve the predictive power of machine learning algorithms feature selection is implemented as discussed in the feature selection section of chapter four.

In this study, Recursive Feature Elimination with Cross-Validation is used to select the most important features.

Recursive Feature Elimination with Cross Validation: this method allows to removes of the weakest features iteratively until the best features set to maximize the predictive model are

achieved. It is a wrapper method of feature selection used with cross-validation to score different features subsets and select the best scoring from the collection of features and implemented as follows

```
rf_rfecv=RandomForestClassifier(n_estimators=1200, max_features='auto', class_weight='balanced',
                                max_depth=15,criterion='entropy',min_samples_split=2,random_state=42, n_jobs=-1)
cv=StratifiedKFold(n_splits=10, random_state=42, shuffle=True)
rfecv=RFECV(estimator=rf_rfecv,step=1,cv=cv,scoring='accuracy')
rfecv=rfecv.fit(X,y)
print('optimal number of features:',rfecv.n_X)
print('Best features:', X.columns[rfecv.support_])
```

Figure 5:5 Recursive Feature Elimination with cross-validation implementation

5.6. Machine Learning Models Implementation

The new proposed machine learning model stacking ensemble is compared with other models such as Random Forest (RF), XGBoost, and Support vector machine(SVM) and the new model is implemented to automate neonatal diseases classification, and the models are evaluated using stratified 10-fold cross-validation along with performance metrics. To do this, all required python libraries are imported as follows.

```
# Importing necessary libriries and packes used for modeling anf evaluation ND_Classification model
import pandas as pd
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
import numpy as np
import sklearn.metrics
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
from matplotlib import cm
import pickle
import mlxtend
from sklearn import metrics
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from collections import Counter
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier,GradientBoostingClassifier
from sklearn.metrics import f1_score,confusion_matrix
from sklearn.metrics import classification_report
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
from sklearn.feature_selection import RFECV
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from mlxtend.feature_selection import SequentialFeatureSelector
from mlxtend.plotting import plot_sequential_feature_selection as plot_Selector
from sklearn.exceptions import ConvergenceWarning
```

Figure 5:6 Importing the libraries necessary for the Modeling

5.6.1. Hyperparameter Tuning Implementation

Hyperparameter tuning discussed in chapter four under section 4.5.1 is implemented on our land suitability dataset in this section. Randomized search cv is used for this study to select the best hyperparameter automatically that optimizes the performance of the models used in this

study such as RF, SVM, and XGB. The sample code used to implement hyperparameter tuning for RF, XGB, and SVM is depicted in Appendix C:2. Random forest is a supervised machine learning algorithm that is flexible and easy to use that fits several decision tree classifiers on several sub-samples of the dataset and it uses averaging to enhance the predictive accuracy and resist over-fitting. It is implemented using the RandomForestClassifier () method from sklearn of ensemble package that is used to fit and train the classifier with the best parameters suggested by randomized search cv and is implemented as follows.

```
#instatiate Random Forest Classifier
RFM=RandomForestClassifier(n_estimators=1200,random_state=0,max_features='auto',class_weight=None,
                           criterion='entropy',min_samples_split=2,n_jobs=-1)
RFM.fit(X,y)
```

Figure 5:7 Python code to create and fit RF model

Another classifier proposed in this study is the Support Vector Machine. The study used the LinearSVC () classifier from the sklearn package to build the SVM model. This classifier is creating instances with a basic parameter like OVR to classify the data set of various classes.

```
#Instantiate Support Vector Machine
SVM=LinearSVC(C=1,random_state=0,multi_class='ovr',class_weight='None')
SVM.fit(X,y)
```

Figure 5:8 Python code to create and fit the SVM model

The third selected algorithm is the XGBoost classifiers method, which is a predictive method machine learning model, and is implemented by the XGBClassifier () method of the sklearn of ensemble package which is used to fit and train the model with the best parameters.

```
# xgboosting
from xgboost import XGBClassifier
xgb=XGBClassifier(learning_rate = 0.05, n_estimators=300, max_depth=5)
xgb.fit(x, y)
```

Figure 5:9 Python code to create and fit the XGB model

The proposed model is a stacking classifier by combining SVM, XGB, and RF

```
from mlxtend.classifier import StackingCVClassifier
LR = LogisticRegression()
sclf = StackingCVClassifier(classifiers=[SVM, rf,XGB],
                            meta_classifier=LR,
                            random_state=RANDOM_SEED)
```

Figure 5:10 python code to create stacking model

5.7. Model Testing and Evaluation Implementation

After data preprocessing, feature selection, and model building, the next step is to evaluate the machine learning model performance using stratified k-fold cross-validation with k = 10 with other performance metrics such as accuracy, confusion matrix, precision, recall, and f1 score. Stratified K-fold cv is implemented using cross_val_score () and cross_val_predict () method. cross_val_score () is a Sklearn Python model selection package that takes estimator, features, target, number of folds, score, and n_jobs parameters. In this study, the entire dataset is passed to a 10-fold stratified cv, and n_iobs indicates the number of CPUs used to do the calculation. n_iobs equal to -1 means 'all CPUs' used for computation, and equal to 1 is the other way around, and cross_val_predict () returns the predicted target values for the fold used as evidence. Python modules in the sklearn method like Classification report () that is used to build a text report showing the main ranking metrics, precision score () which returns the accuracy score of the model, and confusion matrix () which is used to evaluate accuracy classifications are used in this study.

```
from sklearn.model_selection import cross_val_score
skfold=StratifiedKFold(n_splits=10,random_state=42,shuffle=True)
Accu=cross_val_score(model,X,y,scoring='accuracy',cv=skfold,n_jobs=-1)
# To display the accuracy for each fold
print("Accuracy Scores of each Fold per iteration")
print("================================================")
fold=1
for i in acc:
    print("Accuracy of Fold",fold,"is:",round(i*100,2))
    print("------------------------------------------------")
    fold=fold+1
#To display mean accuracy of all folds
print("Mean Accuracy of all Folds",round(acc.mean()*100,2))
```

Figure 5:5 Applying Stratified 10-fold CV on Models

Finally, the prototype for predicting neonatal disease is implemented based on the new proposed ensemble machine learning algorithm and then deployed using flask web services for it to be able to accept a new input from the user and then return the prediction results.

CHAPTER SIX

6. RESULTS, EVALUATION, AND DISCUSSIONS

The main aim of this chapter is to present the details of the experimental results obtained from this study. The overview of the dataset used and its class distribution, model performance, evaluation results, and the effect of feature selection on performance are discussed in this chapter. Furthermore, the results of selected models and newly developed stack ensemble compared with selected three models are discussed, and finally, the best performing model is deployed using the flask server. A detailed comparison of this research with previous studies is also provided to ensure that this study fills the gaps identified in the literature review, answers the research question, and implements the specific research objectives mentioned.

6.1. Dataset Class Distributions Results

The total dataset size used for this study is 2298 rows with 20 columns including one target class after preprocessing the row data collected from ACH. The study classified neonatal disease into four common classes with 711 patients are sepsis disease,648 newborns with Respiratory Distress Syndrome(RDS) disease,527 with Parental asphyxia(PA), and 412 neonates with Necrotizing enterocolitis (NEC). Class distribution in four class dataset are shown in detail in figure 6.1 below

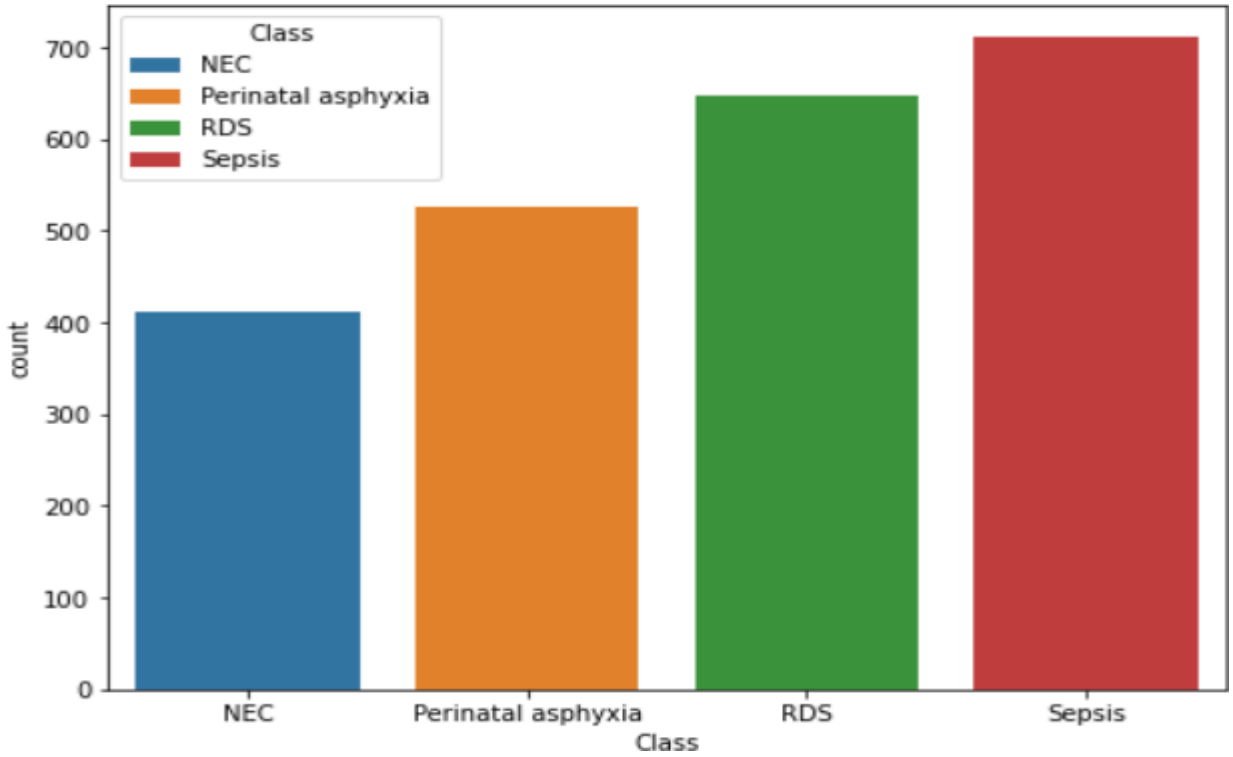

Figure 6:1 Class distribution of four-class

As illustrated in Fig. 6.1 above, the distribution of classes in the collected ACR dataset after pre-processing with four classes, there is a slight class imbalance in the dataset in that the instances are not equally distributed among the target classes. This is because more infants in the hospital (where the dataset is collected) were more affected by those with a high number of counts.

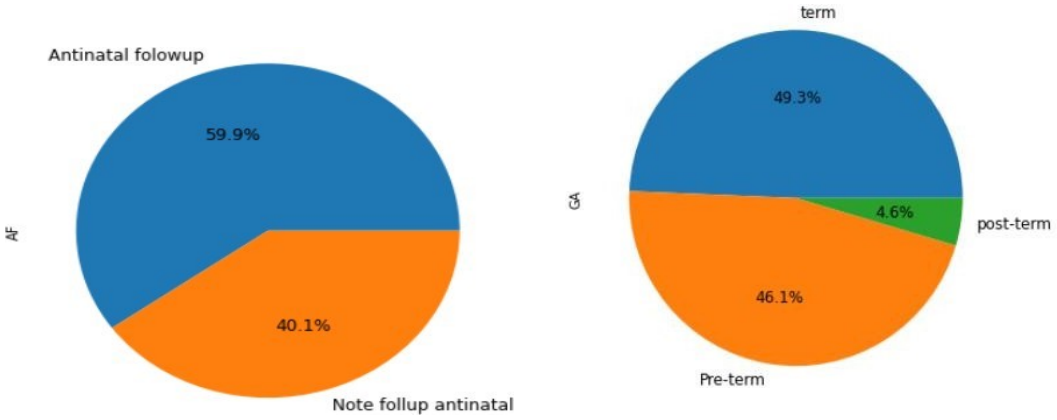

Figure 6:2 distribution of features in a dataset

As shown in the above table the result of the study shows that from the left to right only 59.9 % of women followup anti-natal care and 40.1% didn't follow up during their pregnancy. The right side figure shows that 49.3%,4.6%, and 46.1% of neonatal were born as a term, pre-term, and post-term respectively. This indicates that there is a small number of newborn babies born post-term.

6.2. Feature importance and Selection Result

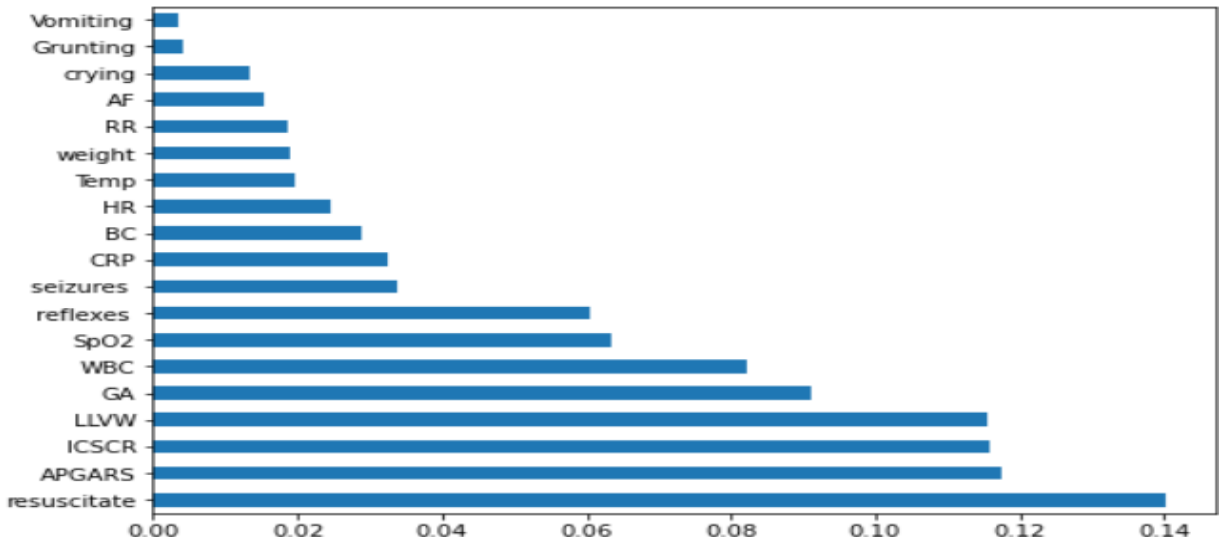

Figure 6:3 Feature Importance Using Random Forest

Figure 6.3 indicates the ranking of features based on their relevance score based on the random forest classifier algorithm. The importance of features indicates how important the features are in deciding the target accuracy of the model. The most important functions have a higher performance score compared to the function with the lowest score. The comparison of features is based on comparing the importance scores of other features in the data set.

In this study feature selection methods are applied to select the most appropriate feature set to improve the predictive power of machine learning algorithm classifiers. Recursive feature elimination with cross-validation (RFECV), where used in the training of the SVM, RF, XGB, and stacking ensemble models used in the training dataset. The below table shows the feature selection result for each selected algorithm.

Table 6-1 Feature Selection Result for each selected Algorithm

Feature Selection Methods	Selected ML Models	Selected features	Original features
RFECV	SVM	12	19
	RF	12	19
	XGB	12	19
	Stacking	12	19

Table 6.1 shown above indicates the number of features selected RFECV methods with RF models different from the original features. The performance of the models with and without the feature selection method on the selected feature set is discussed in the following section.

6.3. Model Building

The proposed stacking ensemble model was compared with different machine learning models such as Support Vector Machine, Random Forest, and XGBoost. To build the model first, the dataset features are split into dependent and independent features. Independent feature datasets are features with values, which do not determine by other whereas dependent values are the target or class disease, which depends on the independent features. Then the models are built on multiclass datasets with and without using feature selection techniques.

6.4. Models Performance and Evaluation Result

The Experiment built a stacking ensemble classifier algorithm that combines three models such as SVM, RF, and XGB with RFECV based on a random forest classifier and develops a stacking classifier. Stratified 10-fold cross-validation is used in testing and training models' performance with other performance evaluation metrics as discussed in previous chapters three and four. In stratified 10-folds, cross-validation techniques select datasets in the same proportion. The models are trained using 10-1 folds and tested using one remaining fold; the process is iterative for each fold. In this study, the performance of selected models with and without feature selection is discussed to identify the most performing model based on the selected features. Finally, the newly developed stack ensemble classifier model will be deployed on the web server using the flask server for neonatal disease classification.

6.4.1. Performance and Evaluation Result of models with original Features

The performance evaluation result of Stacking, SVM, RF, and XGB models based on original features of the neonatal disease dataset without applying any feature selection methods are discussed.

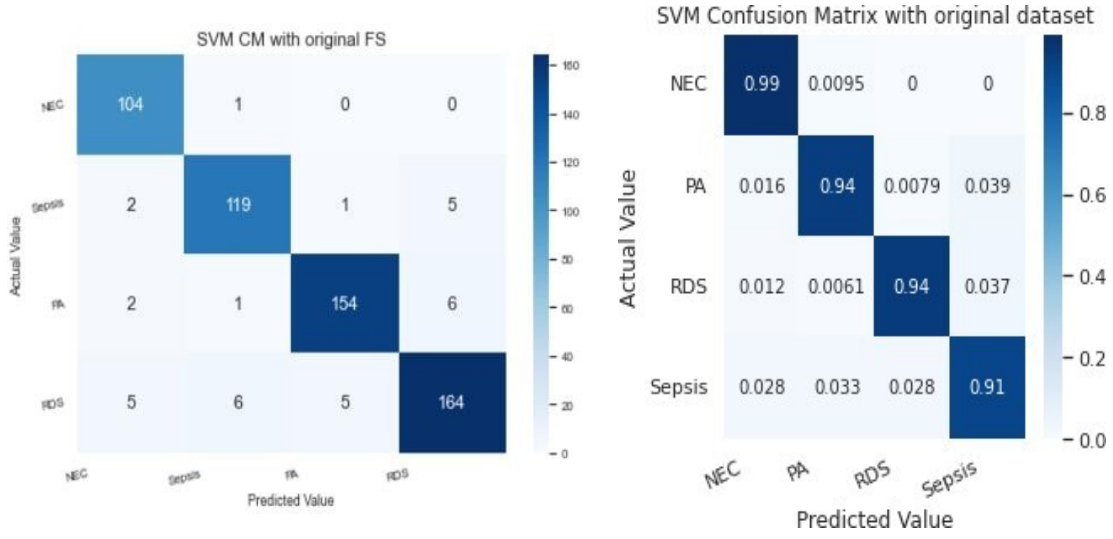

Figure 6:4 Confusion Matrix for SVM without (left) and with Normalization (right)

As can be seen from Figure 6.4 above, the performance of the SVM algorithm is examined on neonatal dataset using stratified confusion matrix, as such 104 out of 105, 119 out of 127, 154 out of 163, and 164 out of 180 dataset instances are correctly classified as NEC, PA, RDS, and Sepsis respectively. Where only the algorithm misclassifies 1 instance of NEC as PA, 2,1,5 instances of PA as NEC, RDS, and sepsis respectively, 2,1, and 5 instances of RDS as NEC,

PA, and sepsis respectively, and 5,6 and 5 instances of Sepsis as NEC, PA, and RDS respectively. The normalized confusion matrix for SVM is also illustrated on the right side in figure 6.4 above which is the same as the left, except it indicates correctly classified instances in the form of a percentage.

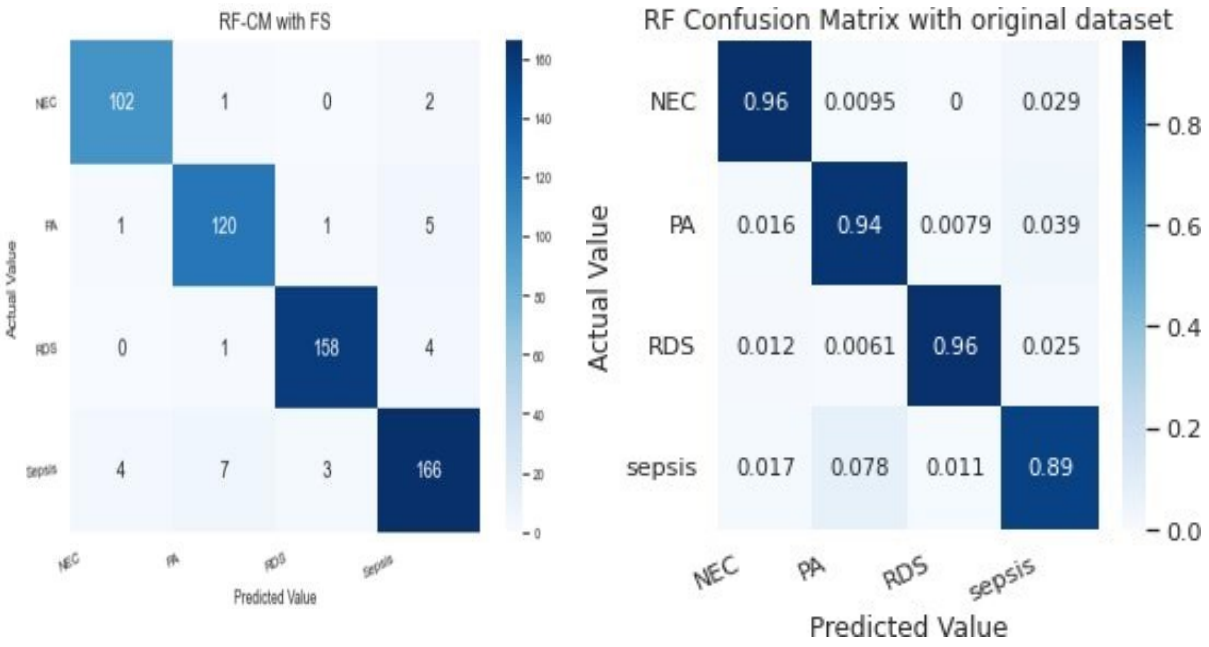

Figure 6:5 Confusion Matrix for RF without (left) and with Normalization(right)

Confusion matrix to assess the performance of Random forest algorithm on the test dataset also illustrated in figure 6.5 above, in which the algorithm correctly classifies 102 out of 105 as NEC, 120 instances of PA out of 127 as PA, 158 instances out of 163 as RDS, and 166 instances of sepsis out of 180 as Sepsis respectively. RF algorithm only misclassifies 1, 2 instances of NEC as PA and sepsis respectively, 1,1,5 instance of PA as NEC, RDS, and Sepsis respectively, and 1, 4 instances of RDS as PA, and sepsis, respectively, and 4,7,3 instances of sepsis as NEC, PA, RDS respectively. As shown in figure 6.5 above, the normalized confusion matrix for RF also shown to indicate correctly classified instances in percentage form.

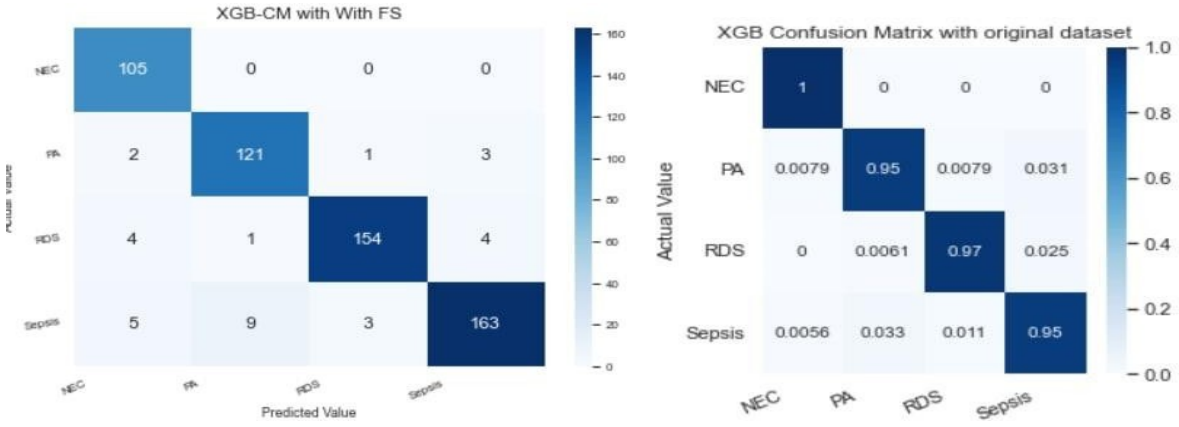

Figure 6:6 Confusion Matrix for XGB without (left) and with Normalization(right)

Similarly, another algorithm used for this study is XGB and its performance is measured using a confusion matrix as shown in figure 6.6 above. Accordingly, the algorithm classifies instances of 105 out of 105, 121 out of 127, 154 out of 163, and 163 out of 180, as Sepsis, NEC, RDS, and PA respectively. However, the algorithm misclassifies 2,1,3, instances of PA as NEC, RDS, and sepsis respectively,4, 1,4 instances of RDS as NEC, PA, and Sepsis respectively, and 5,9,3 instances of sepsis as NEC, PA, and RDS respectively. The confusion matrix on the right side of figure 6.6 is also the same as the left, except it is in the normalized form.

Table 6-2 SVM, RF, XGB, and stacking models accuracy result without FS

Models	**Stratified 10-fold Accuracy result(%) for SVM,RF,XGB**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc(%)
SVM	96.52	96.96	92.61	96.09	93.04	95.65	95.65	95.22	94.76	92.32	94.86
RF	97.39	96.96	94.35	96.09	93.91	96.52	97.83	95.65	96.07	93.07	95.78
XGB	97.39	95.22	94.78	97.39	93.94	96.96	97.39	95.65	96.07	94.32	**95.91**
Stacking	97.78	96.50	95.63	94.97	95.49	96.96	97.47	97.03	95.91	96.83	**96.69**

Table 6.2 shown above indicates the stratified 10-fold cross-validation accuracy of each fold for RF, XGB, SVM, and **stacking** models in which **stacking** scores the highest accuracy of **96.69** than RF, XGB, and SVM with an accuracy of 95.78, 95.91, and 94.86 respectively. Moreover, the performance metrics of the three algorithms are shown in the table below.

Table 6-3 Performance metrics for SVM, RF, and XGB models with originals features

Models	**Stratified 10-fold CV score of performance Evaluation metrics**		
	Precision (%)	Recall (%)	F1-score (%)
SVM	95.14	95.68	95.11
RF	95.8	95.70	95.70
XGB	96.01	96.13	95.79
Stacking	**96.70**	**97.0**	**96.90**

Table 6.3 illustrates the scores of performance metrics used in this study for each model. Stacking models outperform the other models in all selected matrices in this study.

6.4.2. Performance and Evaluation Result of the models with Feature Selection

Based on the RF model RFECV with cross-validation was used to select the best feature subset to enhance the predictive power machine, learning models. Therefore, the performance of RF, XGB, and SVM and Stacking ensemble models in terms of confusion matrix, and other performance metrics under stratified 10-fold cross-validation with feature selections method mentioned below and performance of the model compared with the original features.

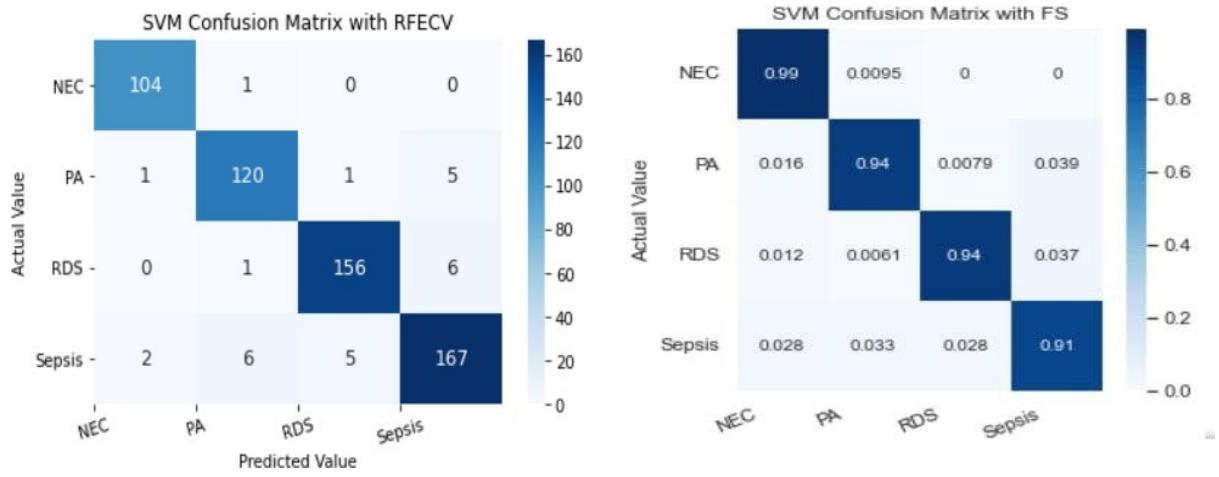

Figure 6:7 Confusion Matrix for SVM with RFECV without and with Normalization

Figure 6.7 shown above, that the confusion matrix of the SVM model with recursive feature elimination with cross-validation is shown with RF which correctly classified 104 instances out of 105,120 out of 127,156 out of 163,167 out of 180 classified as NEC, PA, RDS, and sepsis respectively. Even though the algorithm misclassifies 1 instance of NEC as PA, 1,1, 5 instances of PA as NEC, RDS and sepsis,1,6 instances of RDS as PA and sepsis, and 2,6,5 instances of Sepsis as, NEC, PA, and RDS respectively. The right side in figure 6.7 above shows a normalized confusion matrix for RF with RFECV, which shows correctly classified in a percentage format.

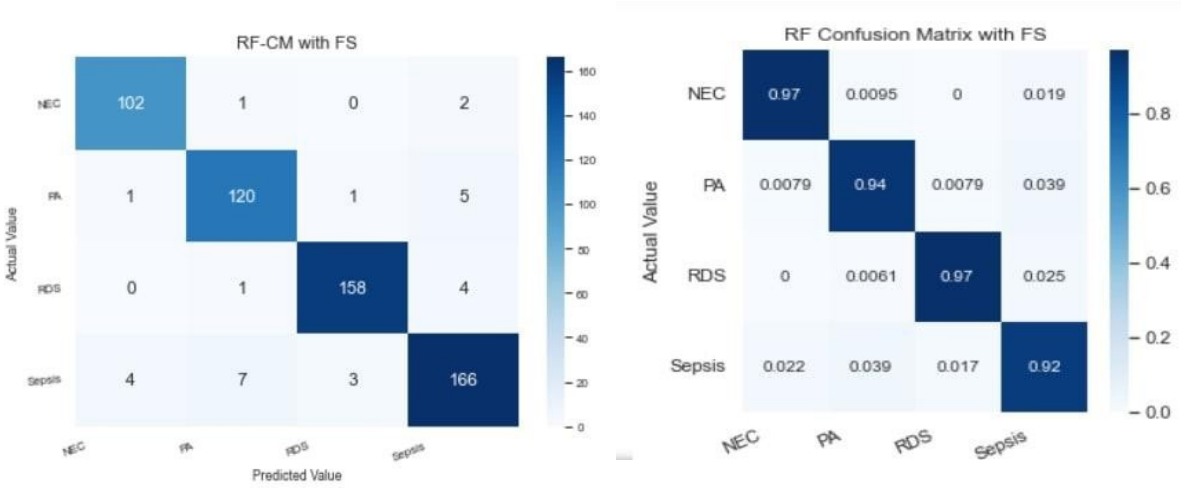

Figure 6:8 Confusion Matrix for RF with RFECV without and with Normalization

Figure 6.8 shown above, that the confusion matrix of the Random forest model with recursive feature elimination with cross-validation is shown with RF which correctly classified 102 instances out of 105,120 out of 127,158 out of 163,166 out of 180 classified as NEC, PA, RDS, and sepsis respectively. Even though the algorithm misclassifies 3 instances of NEC as sepsis, 2, 5 instances of PA as RDS and sepsis,1,4 instances of RDS as PA and sepsis, and 1,9,2

instance of NEC, P, and RDS respectively. The right side in figure 6.8 above shows a normalized confusion matrix for RF with RFECV, which shows correctly classified in a percentage format.

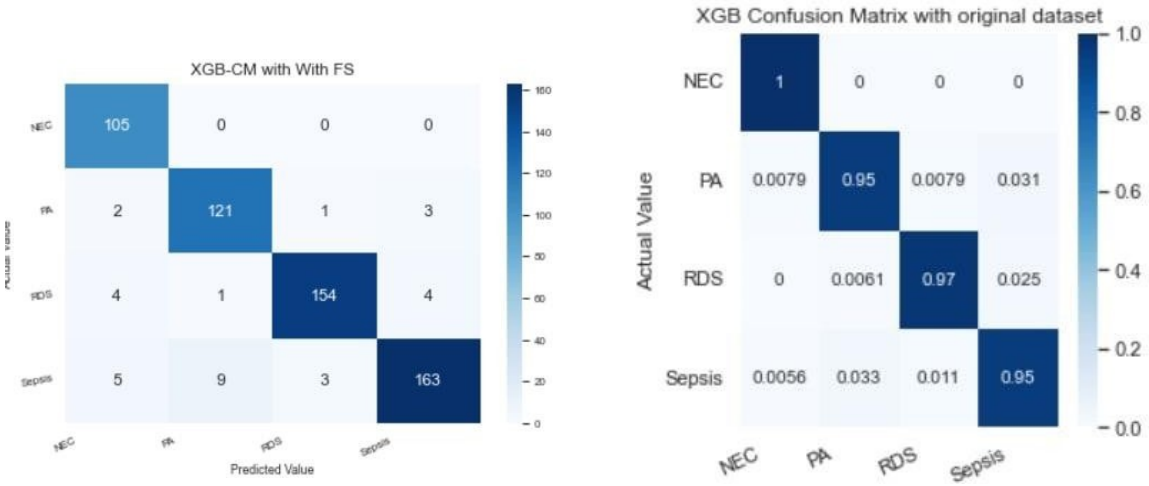

Figure 6:9 Confusion Matrix for XGB with RFECV without and with Normalization

Figure 6.9 shown above, that the confusion matrix of the XGBoost model with recursive feature elimination with cross-validation is shown with RF which correctly classified 105 instances out of 105,121 out of 127,154 out of 163,163 out of 180 classified as NEC, PA, RDS, and sepsis respectively. Even though the algorithm misclassifies 2, 1,3 instances of PA as NEC, RDS, and sepsis,4,1,4 instances of RDS as NEC, PA, and sepsis, and 5,9,3 instances of sepsis, as NEC, PA, and RDS respectively. The right side in figure 6.9 above shows a normalized confusion matrix for RF with RFECV, which shows correctly classified in a percentage format.

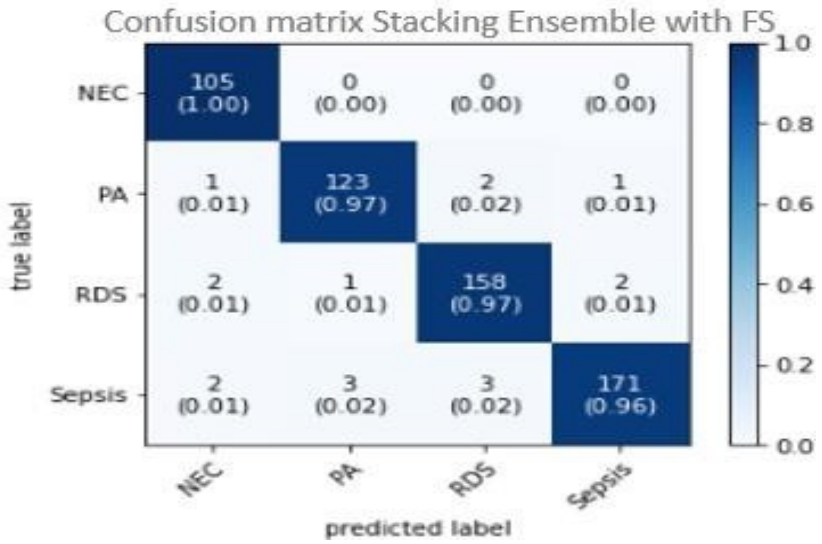

Figure 6:10 Confusion matrix for Stacking ensemble with FS

Figure 6.10 shown above, that the confusion matrix of the Stacking ensemble model with recursive feature elimination with cross-validation is shown with RF which correctly classified 105 instances out of 105,123 out of 127,158 out of 163,171 out of 180 classified as NEC, PA, RDS, and sepsis respectively. Even though the algorithm misclassifies 1, 2,1 instances of PA as NEC, RDS, and sepsis,2,1,2 instances of RDS as NEC, PA, and sepsis, and 2,3,3 instances of sepsis, as NEC, PA, and RDS respectively.

Performance evaluation metrics results of SVM, RF, XGB, and stacking ensemble models with Recursive Feature Elimination with cross-validation feature selection method, which is evaluated using stratified 10-fold cross-validation, are illustrated in the table shown below

Table 6-4 Table SVM, RF, XGB, and Stacking and models accuracy result with FS

	Stratified 10-fold Accuracy result(%) for SVM,RF,XGB &stacking										
Models	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc(%)
SVM	96.96	97.39	94.35	96.52	93.48	94.78	95.65	94.78	94.76	94.32	95.3
RF	97.83	97.39	93.48	96.96	96.09	96.09	96.96	95.65	97.82	96.07	96.43
XGB	97.39	97.83	94.78	97.83	93.91	96.96	97.39	95.22	98.25	96.94	96.65
Stacking	97.78	96.50	95.63	94.97	95.49	96.96	97.47	97.03	95.91	96.83	**97.04**

Figure 6.11 shows the difference between the training and validation accuracy of the models.

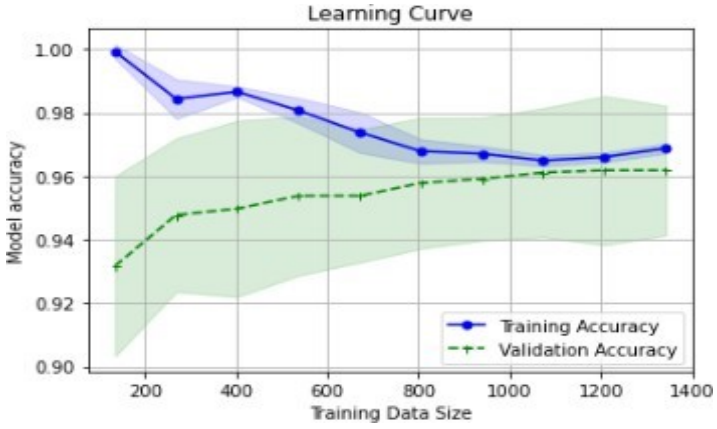

Figure 6:11 Learning curve representing training and validation scores

The average score results of other evaluation performance metrics such as precision, recall, and f1- score of the SVM, RF, and XGB models with RFECV are also illustrated in the table below.

Table 6-5 Performance Metrics for RF and XGB with RFECV models

Models	**Stratified 10-fold CV score of performance Evaluation metrics**		
	Precision (%)	**Recall (%)**	**F1-score (%)**
SVM	95.43	95.63	95.11
RF	96.66	96.98	96.67
XGB	97.02	97.16	97.11
Stacking	**97.21**	**97.38**	**97.30**

Based on tables 6.2, 6.3, 6.4, and 6.5, above, the Stacking ensemble model with Recursive feature elimination with cross-validation outperforms over Stacking ensemble model with original features in all used performance metrics. And the performance of the Stacking ensemble model compared with and without features selection methods.

Table 6.8 shown below points out a summary of the selected model with and without feature selection methods using stratified 10-fold cross-validation with performance evaluation metrics.

Table 6-6 Summary of Model Evaluation with and without FS techniques

Models/FS	**Selected features**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
XGB	19	95.91	96.01	96.13	95.79
XGB with RFECV	12	97.65	97.02	97.16	97.11
RF	19	95.91	95.80	95.70	95.70
RF with RFECV	12	96.43	96.66	96.98	96.67
SVM	19	94.86	95.14	95.68	95.11
SVM with RFECV	12	95.30	95.43	95.63	95.11
Stacking	19	96.69	96.70	97.70	96.90
Stacking with RFECV	**12**	**97.04**	**97.21**	**97.38**	**97.30**

The above table shows the performance of models before and after features selection is applied and the newly developed Stacking model outperforms RFECV with 97.04, 97.21, 97.38, and 97.30 in accuracy, precision, recall, and f-score respectively.

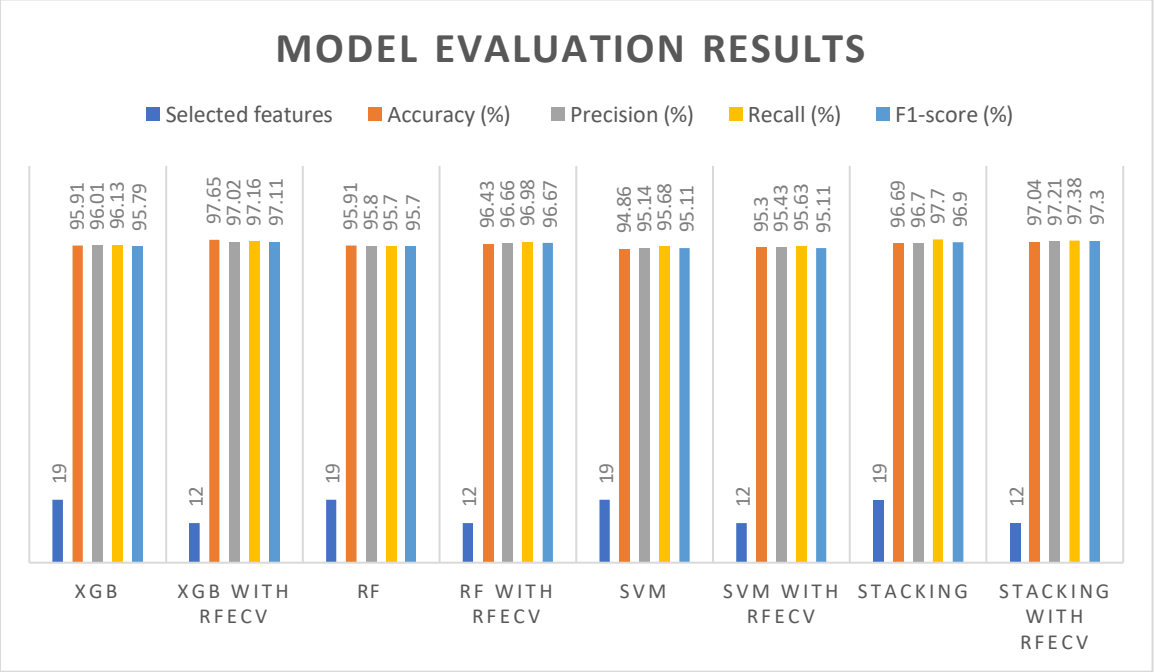

Figure 6:12 Model Evaluation Results

6.5. Hyperparameter Tuning Result

Although the use of automatic hyperparameter tuning is computationally expensive and requires huge memory resources, this study uses random search hyperparameter tuning with cross-validation, as explained in chapter four in section 4.3.1, instead of manual tuning of hyperparameters, which is probably an inefficient optimization strategy. and results in poor and very tedious performance. Both grid search and cv random search were experimented with in this study, and the result revealed that cv random search performs better than grid search, and the suggested parameters are shown in the table below.

Table 6-7 Parameters used in RF as returned by randomized search CV

Parameters	n_estimators	max_depth	criterion	bootstrap	max_features	min_samples_split
Values	1200	15	entropy	True	auto	2

Table 6-8 Parameters used in XGB as suggested by randomized search CV

Parameters	**n_estimators**	**learning_rate**	**Max_depth**	**gamma**	**n_jobs**
Values	100	0.05	6	0.3	1

Table 6.8 above also indicates the list of parameters used in the XGB algorithm, which are returned by randomized search CV.

Result of SVM Model-based Feature Scaling

To train and decide the decision boundary between classes in SVM classifier, it is good to maximize the distance from the nearest point on each data class. For this reason, feature scaling is crucial in support vector machines. Table 6.10 shows how feature scaling affects the classification of neonatal disease using SVM.

Table 6-9 Performance metric result for SVM model with and without Normalization

Models	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SVM without normalization	92.38	91.62	91.93	91.67
SVM with Normalized	95.30	95.14	95.68	95.11

6.6. Discussion

This study increased the performance of the machine learning model by combining the base model as a stacking ensemble. At the beginning of this study, three research questions were encompassed. Hence, this section tries to answer those questions.

- What are the determinant attributes for the prediction of neonatal diseases?

To answer this question, recursive feature elimination with cross-validation(RFECV) is used to select the determinant attributes. As the finding indicated that the following attributes are the major to the prediction of neonatal diseases. APGAR score, CRP (C-reactive protein), Resuscitate, LLVW (Low Lung volume and whiteout), ICSCR (Intercostal Subcostal retractions), Blood Cultures, SpO2 (Oxygen Saturation), GA (Gestational Age), WBC (white blood cells), Seizures, RR (Respiratory rate), weight and Grunting are the determinant attributes.

- Which classification algorithm is the best to construct the classifier model from the selected algorithm?

For this question, the selected machine learning algorithm are combined and form a new model which is a stacking ensemble for neonatal disease classification. In addition, to answer this question the researcher selected three appropriate classification algorithms and features selection methods and compare the performance of the models with and without features selection under stratified 10-fold cross-validation.

Three selected models such as SVM, RF, and XGB compared with each other, and then the new ensemble using logistic regression was developed. According to the experiment the new proposed model outperforms with 97.04% accuracy as the best classifier model of than other three models.

- In what criteria does ensemble machine learning outperform state-of-the-art than single models?

This question was answered by combining three different base machine learning models with optimal parameters and forming an ensemble stacking model. In the sacking ensemble, the machine learning model individual model is trained as level-0 and three individual there is SVM, RF, XGB prediction result are then feed to logistic regression finally the prediction of logistic regression is 97.04 accuracy.

To answer those questions the following activities were done first data were collected and data undergoes preprocessing steps as discussed in chapter three under section 3.4.3. After the data is preprocessed, the dataset is fed to machine-learning models. Then the performance of the models is evaluated using stratified 10-fold cross-validation with other model evaluation matrices such as confusion matrix, accuracy, precision, f-score, and recall with and without feature selection methods. In this study, the proposed model which is a stacking ensemble performs than random forest, support vector machine, and XGB with both feature selection and without feature selections. Models with RFECV perform better than models with original features. The study recommended a Stacking ensemble for the neonatal disease prediction to end users due to its high accuracy, precision, recall, and f1-score of 97.04%,97.21%,97.38%, and 97.30 respectively under stratified 10-fold cross-validation. After the best model with the best feature selection techniques is identified, a porotype is deployed on the webserver to help medical experts as illustrated in the figure below which is the graphical user interface(GUI), that helps models to deploy and implement to end user to interact with interface simply.

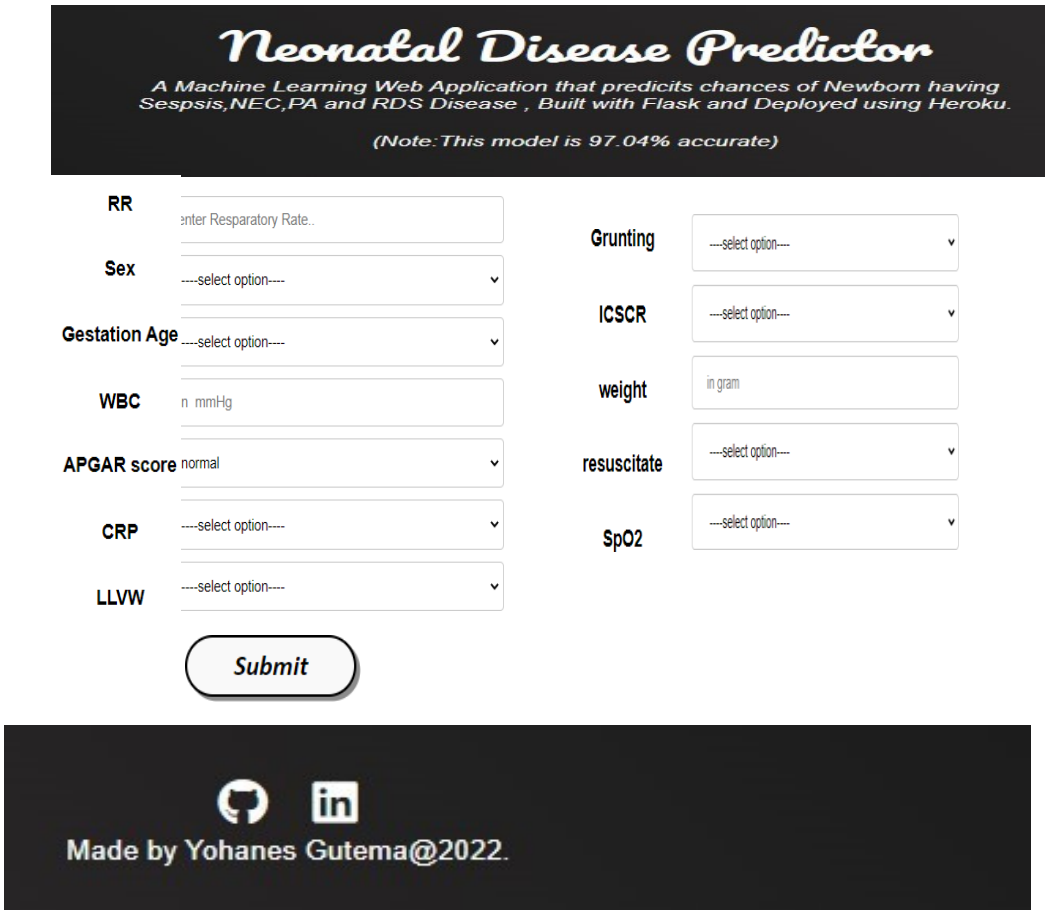

Figure 6:13 GUI for Neonatal Diseases Diagnosis prototype

CONCLUSION, AND FUTURE WORK

Conclusion

In Ethiopia, neonatal mortality is a big problem, which accounts for the lion's share of under-five mortality. As the trends in other countries, the highest contributing age group for under-five mortality is the first 28 days from birth. In Ethiopia, nearly half (43.3%) of under-five deaths are due to deaths in the neonatal period. Hence, neonatal health must be prioritized to sustain rapid progress in the reduction of overall kid mortality. World leaders have launched a renewed effort to reduce child mortality under the Sustainable Development Goals (SDGs). One major target of the SDGs is to "end preventable deaths among newborns and children under the age of five, with all nations aiming for a neonatal mortality rate of at least 12 per 1,000 live births by 2030. To achieve the healthcare service goal machine learning techniques in the diagnosis of neonatal disease play an appropriate contribution to improving the healthcare service outcomes, ameliorating the efficiencies of a healthcare institution, to reduces the economic burden involved in healthcare management, and reducing newborn mortality. To achieve this study neonatal disease data were collected from the Asella Compressive Hospital, Oromia, Ethiopia. To label the dataset into neonatal disease classes namely NEC, PA, RDS, and sepsis, direct communication that was held with the domain expert and neonatal survey report of the study area were used. In addition, the labeled dataset was again reviewed by a medical expert specifically a pediatrician, and finally, it contains 2298 rows with 20 columns including one target class. The labeled dataset contains a slight class imbalance, which was alleviated by setting the class_weight parameter to balance in a random forest algorithm, choosing an ensemble gradient boosting algorithm, which has a built-in approach to combat class imbalance. As well models are evaluated under stratified 10-fold cross-validation, which ensures that there is the same proportion of instances in each fold with a given label. So that a successful result is obtained. To make the dataset ready for the machine-learning algorithm, it goes through several preprocessing steps such as handling missing value, which is resolved by the mean imputation method, and categorical features, which were handled by a label encoder. After preprocessing, steps are completed recursive feature elimination with cross-validation is used to select a relevant set of features from the original features. The performance of three models with original features, and with feature selection methods are compared in terms of confusion matrix, accuracy, precision, recall, and f1-score under stratified 10-fold cross-validation with a newly developed model stack ensemble.

SVM model since it is a distance-based learning algorithm. However, it is the least performing model in all used performance metrics. In addition, the performance of RF with RFECV performs better than RF with original features. Stacking ensemble RF, XGB, and SVM models perform the best with RFECV in all used performance metrics. Stacking ensemble with RFECV feature selection is a final model used for neonatal disease prediction, it is recommended to the end-users, and the prototype is implemented to help health experts, health sectors, and patients due to its high performance in all used performance metrics

Recommendation and Future Works

This study has shown the application of Machine learning for the prediction of neonatal diseases to achieve the main objectives. Based on the results of the research the following recommendations are suggested as future works for healthcare centers and researchers. The recommendation can initiate interested researchers to investigate the further implementation of a prototype machine learning in the other domain.

- The developed machine-learning model provides suggestions for the diagnosis of neonatal diseases. Hence, the researcher recommends that health care institutions use this

  model to improve the experience of health workers and achieve sustainable development goals in Ethiopia.
- This study was focused on identifying the determinant factors that are used for the diagnosis of newborn diseases and building machine learning for the diagnosis of four common neonatal diseases namely NEC, sepsis, RDS, and perinatal asphyxia. Therefore, in the future additional investigation is needed to be done for the severity of the disease and other neonatal diseases such as congenital malformations, jaundice, meningitis, etc.

****

Special Acknowledgment

This research was funded by Adama science and Technology University with grant Number:

ASTU/SM-R/486/22

REFERENCES

(IHME), institute for H. M. and E. (2019). *Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (XGBD 2019)*. http://ghdx.healthdata.org/XGBd-results-tool

Abdo, R. A., Halil, H. M., Kebede, B. A., Anshebo, A. A., & Gejo, N. G. (2019). Prevalence and contributing factors of birth asphyxia among the neonates delivered at Nigist Eleni Mohammed memorial teaching hospital, Southern Ethiopia: A cross-sectional study. *BMC Pregnancy and Childbirth*, *19*(1), 1–7. https://doi.org/10.1186/s12884-019-2696-6

Alsaied, A., Islam, N., & Thalib, L. (2020). The global incidence of Necrotizing Enterocolitis: A systematic revieww and Meta-analysis. *BMC Pediatrics*, *20*(1), 1–15. https://doi.org/10.1186/s12887-020-02231-5

Arcangela Lattari Balest. (2021). *Respiratory Distress Syndrome in Newborns*. https://www.msdmanuals.com/home/children-s-health-issues/lung-and-breathing-problems-in-newborns/respiratory-distress-syndrome-in-newborns

Atun, R. (2015). Transitioning health systems for multimorbidity. *The Lancet*, *386*. https://doi.org/10.1016/S0140-6736(14)62254-6

Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., & Napolitano, A. (2012). A review of the stability of feature selection techniques for bioinformatics data. *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, 356–363. https://doi.org/10.1109/IRI.2012.6303031

Badawi, O., Brennan, T., Celi, L., Feng, M., Ghassemi, M., Ippolito, A., Johnson, A., Mark, R., Mayaud, L., Moody, G., Moses, C., Naumann, T., Pimentel, M., Pollard, T., Santos, M., Stone, D., & Zimolzak, A. (2014). Metadata Correction: Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference. *JMIR Medical Informatics*, *2*, e22. https://doi.org/10.2196/medinform.3447

Bitew, F. H., Nyarko, S. H., Potter, L., & Sparks, C. S. (2020). Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey. *Genus*, *76*(1). https://doi.org/10.1186/s41118-020-00106-2

Cabrera-Quiros, L., Kommers, D., Wolvers, M. K., Oosterwijk, L., Arents, N., van der Sluijs-Bens, J., Cottaar, E. J. E., Andriessen, P., & van Pul, C. (2021). Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning. *Critical Care Explorations*, *3*(1), e0302. https://doi.org/10.1097/cce.0000000000000302

Chen, W., Wang, Y., Cao, G., Chen, G., & Gu, Q. (2014). A random forest model based classification scheme for neonatal amplitude-integrated EEG. *BioMedical Engineering Online*, *13*(Suppl 2), 1–13. https://doi.org/10.1186/1475-925X-13-S2-S4

Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 1211–1215. https://doi.org/10.1109/ICCMC.2019.8819782

Dantas, J. (2020). *The importance of k-fold cross-validation for model prediction in machine learning*. https://towardsdatascience.com/the-importance-of-k-fold-cross-validation-for-model-prediction-in-machine-learning-4709d3fed2ef

Daunhawer, I., Kasser, S., Koch, G., Sieber, L., Cakal, H., Tütsch, J., Pfister, M., Wellmann, S., & Vogt, J. E. (2019). Enhanced early prediction of clinically relevant neonatal hyperbilirubinemia with machine learning. *Pediatric Research*, *86*(1), 122–127. https://doi.org/10.1038/s41390-019-0384-x

Delele, T. G., Biks, G. A., Abebe, S. M., & Kebede, Z. T. (2021). Prevalence of common symptoms of neonatal illness in Northwest Ethiopia: A repeated measure cross-sectional study. *PLoS ONE*, *16*(3 March). https://doi.org/10.1371/journal.pone.0248678

Demisse, A. G., Alemu, F., Gizaw, M. A., & Tigabu, Z. (2017). Patterns of admission and factors associated with neonatal mortality among neonates admitted to the neonatal intensive care unit of University of Gondar Hospital, Northwest Ethiopia. *Pediatric Health, Medicine and Therapeutics*, *Volume 8*, 57–64. https://doi.org/10.2147/phmt.s130309

Desalew, A., Sintayehu, Y., Teferi, N., Amare, F., Geda, B., Worku, T., Abera, K., & Asefaw, A. (2020). Cause and predictors of neonatal mortality among neonates admitted to neonatal intensive care units of public hospitals in eastern Ethiopia: A facility-based prospective follow-up study. *BMC Pediatrics*, *20*(1), 1–11. https://doi.org/10.1186/s12887-020-02051-7

Dhal, P., & Azad, C. (2021). A comprehensive survey on feature selection in the various fields of machine learning. In *Applied Intelligence*. Applied Intelligence. https://doi.org/10.1007/s10489-021-02550-9

Eaton, E. (2008). Introduction to Machine Learning What is Machine Learning. *October*, 1–17.

Eichenwald, E. c, Hansen, R. A., Martin, C. R., & Stark, A. R. (2021a). Cloherty and Stark's Manual of neonatal care. In *Wolters Kluwer: Vol. 8 Ed* (Issue 3).

https://shop.lww.com/Cloherty-and-Stark-s-Manual-of-Neonatal-Care/p/9781496343611

Eichenwald, E. c, Hansen, R. A., Martin, C. R., & Stark, A. R. (2021b). Cloherty and Stark's Manual of neonatal care. In *Wolters Kluwer: Vol. 8 Ed* (Issue 3).

Ershad, M., Mostafa, A., Dela Cruz, M., & Vearrier, D. (2019). Neonatal Sepsis. *Current Emergency and Hospital Medicine Reports*, *7*(3), 83–90. https://doi.org/10.1007/s40138-019-00188-z

Ethiopian Public Health Institute Addis Ababa. (2019). Ethiopia Mini Demographic and Health Survey. In *FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA Ethiopia* (Issue July).

Fattah, M., Alaskar, S. A., Alkhamis, R. I., Othman, F., & Karar, T. A. H. (2019). The Association between Neonatal Sepsis and C-reactive Protein: A Cross-Sectional Study at Tertiary Care Hospital. *International Journal of Medical Research and Health Sciences*, *8*, 54–60.

Forum, A. O. (2021). *Maternal and newborn health in low- and middle-income countries : April*, 20–21.

Getabelew, A., Aman, M., Fantaye, E., & Yeheyis, T. (2018). Prevalence of Neonatal Sepsis and Associated Factors among Neonates in Neonatal Intensive Care Unit at Selected Governmental Hospitals in Shashemene Town, Oromia Regional State, Ethiopia, 2017. *International Journal of Pediatrics*, *2018*, 7801272. https://doi.org/10.1155/2018/7801272

Gizaw, M., Molla, M., & Mekonnen, W. (2014). Trends and risk factors for neonatal mortality in Butajira District, South Central Ethiopia, (1987-2008): A prospective cohort study. *BMC Pregnancy and Childbirth*, *14*(1). https://doi.org/10.1186/1471-2393-14-64

Gupta, H. D., & Choudhury, R. G. (2001). Neonatal disorders and obstetricians. *Journal of the Indian Medical Association*, *99*(5), 262—4, 266. http://europepmc.org/abstract/MED/11676112

Holm, G. (2021). *Necrotizing Enterocolitis*. https://www.healthline.com/health/necrotizing-enterocolitis

Ibrahim, N., Hamid, H. A., Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, *26*, 329–340.

Jayachandiran, K. (2018). *A Machine Learning Approach for Cryptanalysis*. *9028*, 825–830.

Kamiri, J., & Mariga, G. W. (2021). *Research Methods in Machine Learning: A Content Analysis*.

Kaushik, S. (2016). *Introduction to Feature Selection methods with an example (or how to select the right variables?)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Kohli, P. S., & Arora, S. (2018). Application of Machine Learning in Disease Prediction. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1–4. https://doi.org/10.1109/CCAA.2018.8777449

Lashkari, C. (2021). *Diagnosis and Treatment of Sepsis in Newborns*. https://www.news-medical.net/health/Diagnosis-and-Treatment-of-Sepsis-in-Newborns.aspx

Liu, J., & Sorantin, E. (2019). Neonatal respiratory distress syndrome. *Neonatal Lung Ultrasonography*, *September*, 17–39. https://doi.org/10.1007/978-94-024-1549-0_3

McDougall, P., & Hunt, R. (2009). Neonatal Conditions. *Paediatric Handbook: Eighth Edition*, *23*, 431–451. https://doi.org/10.1002/9781444308051.ch32

Mekonnen, S. M., Bekele, D. M., Fenta, F. A., & Wake, A. D. (2021). The Prevalence of Necrotizing Enterocolitis and Associated Factors Among Enteral Fed Preterm and Low Birth Weight Neonates Admitted in Selected Public Hospitals in Addis Ababa, Ethiopia: A Cross-sectional Study. *Global Pediatric Health*, *8*, 2333794X211019695-2333794X211019695. https://doi.org/10.1177/2333794X211019695

Mikhno, A., & Ennett, C. M. (2016). *Prediction of Extubation Failure for Neonates with Respiratory Distress Syndrome Using the MIMIC - II Clinical Database*. 5094–5097.

Mitiku, H. D. (2021). Neonatal mortality and associated factors in Ethiopia: a cross-sectional population-based study. *BMC Women's Health*, *21*(1), 1–9. https://doi.org/10.1186/s12905-021-01308-2

Najafian, B., & Khosravi, M. H. (2020). Neonatal Respiratory Distress Syndrome: Things to Consider and Ways to Manage. In R. M. Barría (Ed.), *Update on Critical Issues on Infant and Neonatal Care*. IntechOpen. https://doi.org/10.5772/intechopen.90885

Neu, J. (2020). Necrotizing Enterocolitis: The Future. *Neonatology*, *117*(2), 240–244. https://doi.org/10.1159/000506866

Oleribe, O. O., Momoh, J., Uzochukwu, B. S. C., Mbofana, F., Adebiyi, A., Barbera, T., Williams, R., & Taylor-Robinson, S. D. (2019). Identifying key challenges facing healthcare systems in Africa and potential solutions. *International Journal of General Medicine*, *12*, 395–403. https://doi.org/10.2147/IJGM.S223882

Phyu, T. Z., & Oo, N. N. (2016). Performance Comparison of Feature Selection Methods. *MATEC Web of Conferences*, *42*. https://doi.org/10.1051/matecconf/20164206002

Rabbi, F. (2019). A review of the recent trends in the use of machine learning in business. *International Journal of Artificial Intelligence and Machine Learning*, *1*(1), 1–6.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*, 103375. https://doi.org/https://doi.org/10.1016/j.compbiomed.2019.103375

Roy Chowdhury, D., Chatterjee, M., & Samanta, R. (n.d.). An Artificial Neural Network Model for Neonatal Disease Diagnosis. In *Samanta International Journal of Artificial Intelligence and Expert Systems (IJAE)* (Issue 2).

Safdari, R., Kadivar, M., Langarizadeh, M., Nejad, A. F., & Kermani, F. (2016a). Developing a fuzzy expert system to predict the risk of neonatal death. *Acta Informatica Medica*, *24*(1), 34–37. https://doi.org/10.5455/aim.2016.24.34-37

Safdari, R., Kadivar, M., Langarizadeh, M., Nejad, A. F., & Kermani, F. (2016b). Developing a fuzzy expert system to predict the risk of neonatal death. *Acta Informatica Medica*, *24*(1), 34–37. https://doi.org/10.5455/aim.2016.24.34-37

Sharew, N. T., Bizuneh, H. T., Assefa, H. K., & Habtewold, T. D. (2018). Investigating admitted patients' satisfaction with nursing care at Debre Berhan Referral Hospital in Ethiopia: A cross-sectional study. *BMJ Open*, *8*(5), 1–8. https://doi.org/10.1136/bmjopen-2017-021107

Sheikhtaheri, A., Zarkesh, M. R., Moradi, R., & Kermani, F. (2021). Prediction of neonatal deaths in NICUs: development and validation of machine learning models. *BMC Medical Informatics and Decision Making*, *21*(1), 1–14. https://doi.org/10.1186/s12911-021-01497-8

Shirwaikar, R. D., Acharya, U. D., Makkithaya, K., Surulivelrajan, M., & Lewis, L. E. S. (2016). Machine learning techniques for neonatal apnea prediction. *Journal of Artificial Intelligence*, *9*(1–3), 33–38. https://doi.org/10.3923/jai.2016.33.38

Shirwaikar, R. D., Mago, N., Dinesh Acharya, U., Makkithaya, K., & Govardhan Hegde, K. (2017). Supervised Learning techniques for analysis of neonatal data. *Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, ICATccT 2016*, 25–31. https://doi.org/10.1109/ICATCCT.2016.7911960

Song, W., Jung, S. Y., Baek, H., Choi, C. W., Jung, Y. H., & Yoo, S. (2020). A predictive model based on machine learning for the early detection of late-onset neonatal sepsis: Development and observational study. *JMIR Medical Informatics*, *8*(7). https://doi.org/10.2196/15965

Sun, P., Wang, D., Mok, V. C., & Shi, L. (2019). Comparison of Feature Selection Methods and Machine Learning Classifiers for Radiomics Analysis in Glioma Grading. *IEEE Access*, *7*, 102010–102020. https://doi.org/10.1109/ACCESS.2019.2928975

UNICEF's Data & Analytics. (2020). *Neonatal mortality*. https://data.unicef.org/topic/child-survival/neonatal-mortality/

Valizadeh, S., Hasankhani, H., & Shojaeimotlagh, V. (2016). Nurses' Immigration: Causes and Problems. *International Journal of Medical Research & Health Sciences*, *5*(9), 486–491. www.ijmrhs.com

Victorio, M. C. (2021). *Neonatal Seizure Disorders*. MSD Manuals Professional Version. https://www.msdmanuals.com/professional/pediatrics/neurologic-disorders-in-children/neonatal-seizure-disorders

WHO. (2020). *Neonatal mortality*. https://data.unicef.org/topic/child-survival/neonatal-mortality/

WHO. (2021). *Newborn health*. https://www.who.int/health-topics/newborn-health#tab=tab_2

Yuan, J., Shi, L., Li, H., Zhou, J., Zeng, L., Cheng, Y., & Han, B. (2022). The Burden of Neonatal Diseases Attributable to Ambient PM 2.5 in China From 1990 to 2019. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.828408

APPENDIX

Appendix A: Dataset Description

Temperature(^{0}C):- degree of hotness or coldness and measured in Celsius(^{0}C)

Respiratory Rate(bpm): is the number of breaths a baby takes per minute

Reflexes: are the inborn behavioral patterns that develop during uterine life

Resuscitate: Emergency procedures performed by a doctor to support newborns at risk

Seizures: caused by sudden, abnormal, and excessive neuronal activity in the brain

Heart Rate(bpm): is the number of times each minute that newborn heartbeats

Seizures: are caused by sudden, abnormal, and excessive neuronal activity in the brain

SpO2: It measures the amount of oxygen-carrying hemoglobin in the blood

C-Reactive Protein(mg/L): level of c-reactive protein in the blood

White Blood Cells(/µL): number of white blood cells in the body

Low Lung Volumes and Whiteout: complication occurs in the lung which is black lung replaced by white lung

Weight: range of health size of a newborn

Intercostal Sub Costal Retractions: lung compliance or airway resistance

Grunting: sound caused by the sudden closure of the glottis

Vomiting: forceful throwing up of stomach contents through the mouth

Blood Cultures: standard to detect the presence of bacteria or fungi in the blood

Antenatal Follow-up: Supervision of humans during pregnancy to monitor the progress of fetal growth

Crying: dropping of tears in response to an emotional state, or pain

Appendix A1: Samples dataset

AF	Temp	Grunti	seizu	WBC	CRP	RR	BC	reflexe	Vomitir	weight	GA	SpO2	HR	ICSCR	resuscit	crying	APGARS	LLVW	Class
no	35.8	yes	yes	4520	abnorma	21	positive	yes	yes	2	Pre-term	abnormal	176	no	no	yes	abnormal	no	NEC
no	34.9	no	yes	14500	normal	21	negative	no	yes	2	Pre-term	normal	102	no	yes	no	abnormal	no	Perinatal asphyxia
yes	34.6	yes	no	4520	abnorma	22	positive	no	yes	2	term	normal	152	no	yes	no	normal	no	Sepsis
no	36.2	yes	no	11000	normal	22	negative	no	yes	2	Pre-term	abnormal	125	yes	no	yes	normal	no	RDS
yes	35	yes	no	9000	abnorma	23	negative	no	yes	2	term	normal	165	no	no	yes	normal	no	Sepsis
no	34.5	yes	no	12000	abnorma	23	positive	no	yes	2	Pre-term	abnormal		yes	no	yes	normal	yes	RDS
yes	38.1	no	no	7850	abnorma	24	negative	no	yes	2	Pre-term	abnormal	115	no	no	yes	abnormal	yes	RDS
yes	36	yes	yes	9500	abnorma	24	positive	yes	yes	2	post-term	normal		yes	yes	no	abnormal	no	Perinatal asphyxia
no	35	no	yes	4520	abnorma	25	positive	yes	yes	2	Pre-term	abnormal	184	no	no	yes	abnormal	yes	NEC
yes	36	yes	yes	12500	abnorma	25	positive	no	yes	2	Pre-term	normal		no	no	yes	abnormal	no	Sepsis
yes	36	yes	no	14600	normal	26	negative	yes	yes	2	Pre-term	normal	114	yes	no	yes	normal	yes	RDS
yes	34	no	yes	17900	abnorma	26	negative	no	yes	2	term	normal	142	no	no	no	normal	no	Sepsis
yes	35	yes	yes	4520	abnorma	27	positive	no	yes	2	post-term	normal		no	no	yes	normal	no	Sepsis
yes	38	no	yes	15600	normal	27	negative	no	yes	2	term	normal	180	no	no	yes	normal	no	Sepsis
yes	36.2	yes	yes	12000	abnorma	28	negative	no	yes	2	Pre-term	abnormal	115	yes	no	no	normal	yes	RDS
no	38.5	yes	yes	17000	abnorma	28	positive	no	yes	2	Pre-term	abnormal	127	no	yes	no	abnormal	no	Perinatal asphyxia
no	34	yes	no	9500	abnorma	29	positive	yes	yes	2	term	normal	170	no	no	yes	normal	yes	RDS
yes	35.3	no	yes	10100	normal	29	negative	yes	no	2	term	normal	161	no	no	no	normal	no	Sepsis
no	38	yes	no	15000	normal	29	positive	no	yes	2	Pre-term	abnormal	150	yes	no	no	normal	yes	RDS
no	37.4	yes	yes	24000	abnorma	29	negative	no	yes	2	Pre-term	abnormal		yes	no	yes	normal	no	RDS
yes	38	no	yes	16000	normal	30	negative	no	yes	2	term	normal	160	no	yes	yes	abnormal	no	Perinatal asphyxia
yes	34	yes	yes	4520	normal	32	negative	no	yes	2	term	normal		no	no	yes	abnormal	no	Sepsis
yes	35	yes	yes	4520	abnorma	34	positive	yes	yes	2	Pre-term	abnormal	84	no	no	yes	abnormal	no	NEC
yes	36	yes	no	4520	abnorma	36	positive	no	yes	2	Pre-term	abnormal	122	no	yes	yes	abnormal	no	Perinatal asphyxia
yes	37.9	no	yes	10400	normal	37	positive	no	no	2	term	normal		yes	no	yes	normal	no	Sepsis
yes	36.2	yes	no	4520	abnorma	38	positive	no	yes	2	term	normal		no	no	no	abnormal	no	Sepsis
yes	35.8	no	yes	14500	normal	42	negative	no	yes	2	Pre-term	abnormal	195	yes	no	no	normal	yes	RDS
no	37.6	yes	yes	15000	normal	43	negative	no	yes	2	term	normal	168	no	no	yes	normal	no	Sepsis

Appendix B: sample code

B1: Stacking ensemble sample code

```
# stacking ensemble code
#importing library
from sklearn.svm import LinearSVC
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import classification_report
from sklearn.model_selection import StratifiedKFold
# startified 10-folds
skfold=StratifiedKFold(n_splits=10,random_state=42,shuffle=True)
#Base machine learning algorthms
SVM = LinearSVC(C=1, class_weight=None,multi_class='ovr', random_state=0)
XGB=XGBClassifier(n_estimators=100,learning_rate=0.1)
rfc = RandomForestClassifier(criterion = 'gini', n_estimators=1000, random_state=0,verbose=1, n_jobs = -1,
                             class_weight = 'balanced', max_features = 'auto')
# model fit
SVM.fit(X_train,y_train)
rfc.fit(X_train, y_train)
XGB.fit(X_train, y_train)

model1 = XGB.predict(X_test)
model2 = rfc.predict(X_test)
model3 = SVM.predict(X_test)

LR = LogisticRegression()
f = [model1, model2,model3]
f = np.transpose(f)
LR.fit(f, y_test)
Stack = LR.predict(f)
# to print in folds
acc=cross_val_score(Stack,X,y,scoring='accuracy',cv=skfold,n_jobs=-1)
print("Accuracy Scores of each Fold per iteration")
print("======================================================")
fold=1
for i in acc:
    print("Accuracy of Fold",fold,"is:",round(i*100,2))
    print("------------------------------------------------------")
    fold=fold+1

print("Mean Accuracy of all Folds",round(acc.mean()*100,2))
print('\n Classification Report:\n', classification_report(y_test,pred))
```

B2: Sample Code for Integrating ML Model with Flask Server

```
# Importing essential libraries
from flask import Flask, escape,request, render_template
import pickle
import numpy as np
# Load the Random Forest CLassifier model
filename = Neonatal-disease-prediction.pkl'
model = pickle.load(open(filename, 'rb'))

app = Flask(__name__)

@app.route('/')
def home():
    return render_template('main.html')

@app.route('/predict', methods=['GET','POST'])
def predict():
    if request.method == 'POST':

        RR = int(request.form['RR'])
        BC  =  request.form.get('BC')
        GA = request.form.get('GA')
        WBC = int(request.form['WBC'])
        APGARS = int(request.form['APGARS'])
        CRP = request.form.get('CRP')
        LLVW = int(request.form['LLVW'])
        ICSCR = int(request.form['ICSCR'])
        resuscitate = request.form.get('resuscitate')
        SpO2 = float(request.form['SpO2'])
        grunting = request.form.get('grunting')
        weight = int(request.form['weight'])
        #thal = request.form.get('thal')

        data =
np.array([[RR,BC,GA,WBC,APGARS,CRP,LLVW,ICSCR,resuscitate,SpO2,grunting,weight
]])
        my_prediction = model.predict(data)

        return render_template('result.html', prediction=my_prediction)

if __name__== '__main__':
    app.run(debug=True)
```

Appendix C Classification Reports

Stacking ensemble

```
Accuracy Scores of each Fold per iteration
==========================================================
Accuracy of Fold 1 is: 97.39
----------------------------------------------------
Accuracy of Fold 2 is: 97.83
----------------------------------------------------
Accuracy of Fold 3 is: 94.78
----------------------------------------------------
Accuracy of Fold 4 is: 97.83
----------------------------------------------------
Accuracy of Fold 5 is: 93.91
----------------------------------------------------
Accuracy of Fold 6 is: 96.96
----------------------------------------------------
Accuracy of Fold 7 is: 97.39
----------------------------------------------------
Accuracy of Fold 8 is: 95.22
----------------------------------------------------
Accuracy of Fold 9 is: 98.25
----------------------------------------------------
Accuracy of Fold 10 is: 96.94
----------------------------------------------------
Mean Accuracy of all Folds   97.39
```

Classification Report:

	precision	recall	f1-score	support
NEC	0.96	1.00	0.98	105
Perinatal asphyxia	0.92	0.95	0.94	127
RDS	0.98	0.96	0.97	163
Sepsis	0.96	0.93	0.94	180
accuracy			0.96	575
macro avg	0.96	0.96	0.96	575
weighted avg	0.96	0.96	0.96	575