“Explainable AI-Based Anemia Prediction through Machine Learning”

Submitted By
Md Sohanur Rohaman Sohan
(201-35-3072)
Department of Software Engineering

Supervised By
Ms. Nusrat Tasnim
Senior Lecturer
Department of Software Engineering
Daffodil International University

A thesis that was turned in to partially complete the requirements for a Bachelor of Science in Software Engineering degree

Fall 2024

APPROVAL

APPROVAL

This thesis titled on "**Explainable AI-Based Anemia Prediction through Machine Learning**", submitted by **Md Sohanur Rohaman Sohan (ID: 201-35-3072)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

-- **Chairman**

Professor Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **Internal Examiner 1**

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **Internal Examiner 2**

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **External Examiner**

Dr. Md. Monowarul Islam
Associate Professor
Department of Computer Science & Engineering
Jagannath University

THESIS DECLARATION

THESIS DECLARATION

I hereby declare that, this thesis report is done by me under the supervision of **Ms. Nusrat Tasnim**, Senior Lecturer, Department of Software Engineering, Daffodil International University, in partial fulfillment of my original work. I am also declaring that neither this thesis nor any part therefore has been submitted elsewhere for the award of Bachelor or any degree.

Supervised By

নুসরাত তাসনিম
২৪/২/২৫

Ms. Nusrat Tasnim
Senior Lecturer
Department of Software Engineering
Daffodil International University

Submitted By

Sohan

Md Sohanur Rohaman Sohan
ID: 201-35-3072
Department of Software Engineering
Daffodil International University

ACKNOWLEDGMENTS

First and foremost, I am grateful to Allah for my health, which is why I finished my research. I am also appreciative of **Ms. Nusrat Tasnim**, my research supervisor, who helped me with all of the research efforts. In addition to my supervisor, I am appreciative of my research. Throughout the entire research process, I received guidance from Assistant Professor

Md. Shohel Arman. I want to express my gratitude to the other members of my research committee for their support and wise observations. I would especially like to thank the faculty head, **Dr. Imran Mahmud,** for giving all the facilities required for the project. I also want to express my gratitude to all of the instructors in the Department of Software Engineering who truly helped me through my struggles. I am appreciative of my parents' unwavering encouragement and support. I'm grateful to my friend for helping me in this endeavor.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURES

Abstract

Anemia is a common worldwide health problem that can have a major negative impact on a person's quality of life and result in major health issues. Timely intervention depends on early diagnosis and precise Anemia prediction however conventional diagnostic techniques frequently depend on intrusive procedures or subjective interpretation. This thesis investigates the use of machine learning approaches to forecast anemia through the application of explainable artificial intelligence (XAI). In particular, it seeks to create an interpretable model that not only forecasts the risk of anemia but also offers intelligible information about the variables influencing the prediction.

A variety of machine learning algorithms are used in this work to optimize the accuracy and reliability of anemia prediction. A robust and flexible approach is ensured by the unique benefits that each of the selected methods—Logistic Regression, Random Forest, Decision Trees, XG-Boost (Extreme Gradient Boosting), Support Vector Machine (SVM), and KNN—brings to the prediction process. These models are trained on a clinically relevant dataset that includes hemoglobin levels, red blood cell counts, and other CBC characteristics that are crucial for detecting anemia.

This study uses a large dataset of over 1281 people that contains demographic, clinical, and lifestyle factors associated with anemia risk. Carefully chosen machine learning models are trained and evaluated using the preprocessed data. Predefined metrics like F1 Score, accuracy, precision, and recall are used to assess each algorithm's performance.

When compared to other algorithms, Decision Tree performs better than all others in terms of prediction accuracy, with a remarkable 99.03% in our analysis. This implies that decision trees are a better way to forecast anemia. Decision trees' outstanding performance enables the creation of a precise anemia prediction tool. By analyzing readily available patient data, this method can assist medical professionals in preventing anemia and initiating therapy early.

Acronyms: Anemia, Machine learning, Anemia prediction, Decision tree classifier, Supervised Learning, KNN, Logistic Regression classifier, Random forest classifier, Support Vector Machine, XG-Boost, XAL, SHAP

Chapter 1

Introduction

1.1 Background

Millions of people worldwide suffer from anemia, which is a serious public health concern, especially in areas with few resources. Anemia, which is characterized by a decrease in hemoglobin levels or red blood cell count or function, affects the body's capacity to provide oxygen to tissues, resulting in weakness, exhaustion, and in extreme situations, organ damage. Despite being a preventable and treatable ailment, anemia is nevertheless frequently underdiagnosed, particularly in developing nations, because of things like difficult-to-find underlying causes, restricted access to healthcare facilities, and a lack of diagnostic resources. Numerous conditions, such as blood loss, chronic illnesses, genetic problems, and nutritional deficiencies, can cause anemia. Because of the variety of its causes, prompt and precise diagnosis is essential for successful intervention and therapy.

Clinical evaluations and blood tests, including complete blood counts (CBC), hemoglobin measures, and iron studies, have historically been used to diagnose anemia. Although these techniques can yield valuable data, they frequently call for trained medical personnel, specialized tools, and access to lab facilities—all of which might not be easily accessible in places with limited resources. Furthermore, the diagnosis and treatment planning processes can be laborious, individualized, and reliant on the skills of specific clinicians. Misdiagnosis or delayed diagnosis is still a major worry due to the complexity of the etiology of anemia, which emphasizes the need for creative solutions that might improve and expedite the diagnostic process.

Rapid advances in artificial intelligence (AI) and machine learning (ML) have sparked interest in using these technologies to enhance healthcare outcomes, especially in the areas of disease detection and prediction. Large amounts of clinical data can be analyzed by machine learning algorithms, which can then spot patterns that human clinicians might find challenging to spot. In the instance of anemia, machine learning algorithms can forecast a patient's risk of developing the condition using patient data, including demographics, lab results, and medical history. This could allow for earlier detection and more individualized treatment plans.

The "black-box" character of many AI models is a major obstacle in healthcare applications, even if machine learning algorithms have a lot of potential for predicting anemia. Frequently, these models make forecasts without clearly stating how or why the choice was reached. This lack of transparency can be a major deterrent to adoption in a clinical context because medical practitioners need to be able to evaluate and comprehend the reasoning behind a model's conclusions, especially when making decisions that potentially affect the health of their patients.

The field of explainable AI (XAI) has evolved with a focus on creating machine learning models that are visible, interpretable, and accurate. By offering insights into the variables affecting AI models' predictions, XAI techniques aim to make AI models easier to comprehend. For example, techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can pinpoint the key elements that influenced a model's choice, assisting physicians in comprehending the logic behind a forecast. XAI has the ability to improve the usefulness and reliability of these models by incorporating explainability into anemia prediction. This would guarantee that physicians are not only confident in the model's predictions but also in its capacity to offer significant and practical insights.

A significant percentage of people worldwide, especially in underdeveloped nations, suffer from anemia, a prevalent and complex illness. It is characterized by a lack of hemoglobin or red blood cells in the blood, which lowers the body's ability to carry oxygen to tissues and organs. Iron-deficiency anemia, chronic disease, and megaloblastic anemia are the three most prevalent forms of anemia; each has unique underlying causes and necessitates a different approach to diagnosis and treatment. Fatigue, weakness, decreased productivity at work, and in extreme situations, heart failure and organ dysfunction are just a few of the serious health issues that anemia can cause. The World Health Organization (WHO) estimates that anemia affects about 1.62 billion people worldwide, or around 24.8% of the population. The highest prevalence is seen in elderly individuals, pregnant women, and children under five.

1.2 Motivation

Millions of people worldwide suffer from anemia, a serious health problem that disproportionately affects vulnerable groups like children, expectant mothers, and the elderly. It is frequently connected to a number of underlying causes, such as genetic factors, chronic illnesses, blood loss, and nutritional deficiencies (such as iron, vitamin B12, and folate). Even though anemia is curable, it is nevertheless underdiagnosed in many areas, especially in rural and low-resource areas. The diagnostic procedure is frequently intricate and necessitates precise clinical data interpretation, which can be difficult without the right diagnostic equipment and medical knowledge. Therefore, a delayed or incorrect diagnosis may result in negative health consequences, such as higher morbidity, children's cognitive development problems, and adults' reduced quality of life.

Although they are frequently employed, traditional diagnostic techniques like iron studies and complete blood count (CBC) tests have drawbacks. These techniques might not offer a thorough grasp of the underlying causes of anemia, and they rely on laboratory facilities and qualified doctors to interpret results. Additionally, the procedure can take a long time, and people in underprivileged areas might not have easy access to it. Innovative methods that might increase the speed, accuracy, and accessibility of anemia diagnosis are becoming more and more necessary in light of these difficulties.

There are encouraging prospects for addressing these issues with machine learning (ML). Through the analysis of extensive datasets, machine learning models are able to spot intricate patterns in clinical data that might not be immediately obvious to physicians. Based on a number of variables, including blood test results, medical history, demographic data, and lifestyle choices, these models may be able to forecast the likelihood of anemia. This could facilitate early detection and intervention by enabling more precise, rapid, and customized forecasts of anemia risk.

One possible approach to this issue is Explainable AI (XAI). Without compromising predictive accuracy, XAI approaches seek to increase the transparency and interpretability of machine learning models. The contribution of various aspects to the model's predictions can be explained using techniques like SHAP (SHapley Additive exPlanations) and LIME (Local

Interpretable Model-agnostic Explanations). This means that when it comes to anemia prediction, doctors would not only be given a prognosis about a patient's likelihood of having anemia, but also an explanation of the elements that most influenced that forecast, such as age, blood test results, and nutrition. The system that is created by integrating XAI with anemia prediction algorithms has the potential to improve clinical decision-making by fostering confidence and offering practical insights.

The goal of this project is to create an explainable AI-based model for predicting anemia that combines the interpretability needed in clinical practice with the advantages of machine learning. Although machine learning models can increase the accuracy of diagnoses, their use in actual healthcare settings is constrained if physicians are unable to comprehend or trust them. By making these models interpretable, we hope to give medical personnel a tool that will help them better comprehend the reasoning behind the predictions and promote accurate diagnosis, ultimately leading to better patient care.

The necessity to equip healthcare professionals with dependable and intelligible decision-support tools is what motivates this research. By doing this, we intend to enhance the early identification of anemia, especially in marginalized communities, and support better informed, data-driven clinical care decisions for anemia. The ultimate objective is to improve patient outcomes through improved diagnostic procedures, especially in regions with potentially little healthcare facilities and resources.

1.3 Problem Statement:

predicting anemia using clinical and demographic data using machine learning models. Use explainable AI strategies to improve model interpretability as well, which will aid medical personnel in comprehending the variables affecting forecasts and enhancing patient outcomes.

With an estimated 1.62 billion individuals affected globally, anemia is a common health problem that is most common in vulnerable populations like children, pregnant women, and the elderly. Effective management and therapy depend on early detection and precise diagnosis, but the diagnostic tools available today are still insufficient. The conventional methods, which frequently depend on laboratory tests like complete blood counts (CBC), iron studies, and subjective clinical evaluations, are not always available, especially in settings with limited resources, and may cause delays or inaccurate diagnoses. Furthermore, these techniques don't always shed light on the various and complex underlying causes of anemia, which include genetics, chronic illnesses, blood loss, and dietary inadequacies.

The development of machine learning (ML) has opened up new avenues for improving anemia diagnosis and prediction. ML models have the ability to more correctly forecast the possibility of anemia than conventional techniques by analyzing vast amounts of patient data, such as demographics, medical history, blood test results, and lifestyle factors. However, despite machine learning's potential, the "black-box" nature of many sophisticated ML models remains a significant barrier in the healthcare industry. Despite their strength, these models are frequently uninterpretable, making it difficult for doctors to comprehend or have faith in the predictions they provide. In order to make safe, well-informed decisions for patient care in clinical practice, especially when handling high-stakes illnesses like anemia, it is crucial to not only get accurate predictions but also comprehend the reasoning behind them.

The main issue this study attempts to solve is the incapacity of machine learning models for predicting anemia to be explained. Although predictive models can provide useful information about a patient's risk of anemia, if doctors are unable to understand the reasons influencing the predictions, they are unlikely to accept and have faith in the models. The potential of AI-based models to improve patient care and outcomes is undermined when their interpretability is lacking, which makes it difficult to incorporate them into clinical processes.

Consequently, this thesis seeks to address two issues:

Enhancing the precision and accessibility of anemia diagnosis: This study uses machine learning to create a model that can more effectively and precisely forecast the possibility of anemia based on a variety of patient data.

Improving machine learning models' interpretability: The necessity for explainable AI (XAI) in the healthcare industry, specifically in the prediction of anemia, is also covered in the study. This project intends to give doctors comprehensible, transparent explanations for the model's predictions by including XAI approaches into the model. This will ensure that clinicians can rely on and confidently act upon the results.

1.4 Research Gap

Despite the significant potential of machine learning models to predict anemia accurately, there is a critical gap in creating systems that are both explainable and trustworthy for clinical use. Current approaches using models such as XG-Boost, Random Forest, Decision Trees, Logistic Regression, Support vector machine(SVM) and KNN suffer from either being too complex to interpret or too simple to achieve the required accuracy. This issue is particularly challenging in medical settings, where clinicians need to understand and trust the predictions made by these systems.

Numerous machine learning models have shown promise in the diagnosis of anemia, including XG-Boost, Random Forests, Decision Trees, and Logistic Regression, Support vector machine(SVM), KNN. Nevertheless, these models are either too basic to provide the necessary accuracy or lack the interpretability essential for clinical applications. For example, models like XG-Boost or Random Forests give higher predictive performance but are still challenging to understand, whereas Decision Trees offer interpretability but frequently fall short of capturing the complexity required for high accuracy. Healthcare workers must trust and comprehend AI systems in clinical practice, where judgments frequently carry significant consequences. Clinicians could be hesitant to use AI-driven predictions for diagnosing anemia if there are no clear, understandable models.

Striking the correct balance between model interpretability and predictive accuracy is one of the biggest problems in the field of machine learning for healthcare. Although deep learning and ensemble approaches are examples of sophisticated models that frequently attain high accuracy, their lack of transparency renders them unsuitable for clinical decision-making. On the other hand, more straightforward and interpretable models, such Decision Trees or Logistic Regression, might not adequately represent the intricacy of the data to generate precise forecasts. Because most studies don't sufficiently address the necessity for both explainability and high prediction accuracy, this trade-off between model complexity and interpretability leaves a gap in the literature.

Although Explainable AI (XAI) methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been used in fields like marketing and finance, there is little evidence of their application in the medical field, especially when it comes to predicting anemia. These techniques could give doctors valuable information about the variables affecting AI predictions, like age, comorbid disorders, or blood test results. Nevertheless, there is still a lack of research on how to include these XAI techniques into the prediction models used to diagnose anemia. The potential for machine learning models to be completely adopted by healthcare practitioners is limited by the lack of literature on the application of XAI to anemia prediction.

Many studies on machine learning for the prediction of anemia have concentrated on homogeneous or limited datasets. Numerous research reply on data from particular populations or limited data sources, which may affect the models' robustness and generalizability. Diverse patient populations with a range of medical histories, lifestyles, and risk factors are encountered in real-world therapeutic settings. In order to increase forecast accuracy and generalize across various populations, current models frequently fall short in integrating these diverse features. The use of extensive, varied datasets that more accurately represent the variety of patients and problems that AI models would come across in clinical practice is critically lacking.

By creating an explainable AI-based model for anemia prediction that strikes a balance between interpretability and predictive accuracy, this thesis seeks to fill up these research gaps. Through the use of cutting-edge machine learning algorithms and XAI strategies like SHAP, this study will produce a model that offers clinicians clear, actionable insights in addition to strong prediction performance. In order to improve the model's generalizability and usefulness across various patient populations, it will also be built using a variety of real-world clinical datasets. Finally, this study will investigate if the model can be used in contexts with limited resources, making sure that it can aid in the diagnosis of anemia in situations when access to medical services is limited.

In order to improve healthcare outcomes by making more precise, reliable, and actionable AI-driven decisions, this thesis seeks to close these gaps and promote the use of machine learning and explainable AI in anemia prediction.

1.5 Research Questions

Q1. How can medical data be efficiently analyzed by machine learning algorithms to forecast a patient's risk of developing anemia?

Q2. Using machine learning approaches, what are the main factors and elements that go into making an accurate forecast about anemia?

Q3. What is the accuracy and dependability of various machine learning techniques for predicting anemia?

Q4. How to enhance the model's capacity to offer explanations that are clinically relevant?

Q5: How can machine learning with SHAP be used to forecast anemia using explainable AI?

1.6. Research Scope

By giving doctors reliable and useful information for decision-making, the goal is to close the gap between anemia diagnosis accuracy, interpretability, and fairness.

- Applying Various Machine Learning Models
- Performance Assessment and Comparison
- Use SHAP (SHapley Additive exPlanations) to evaluate explainability and interpretability.

1.7 Thesis Organization

By creating a reliable, comprehensible model for predicting anemia, this thesis advances the subject of explainable AI in healthcare. It fills in important research gaps, namely the difficulty of developing systems that are accurate and easy for doctors to understand. The study also highlights how explainable AI may enhance medical decision-making's efficacy, efficiency, and accessibility, especially in environments with limited resources. Future studies should focus on improving the model even more, investigating different XAI techniques, and evaluating the model's suitability for use in various clinical settings.

Anemia is a common and frequently misdiagnosed illness that has serious repercussions for public health worldwide. Effective treatment and management of anemia depend on an accurate and prompt diagnosis, but conventional approaches are frequently constrained by the difficulty of identifying the underlying causes, subjective interpretation, and availability to healthcare resources. Although machine learning (ML) has shown promise in improving the prediction of anemia, its use in clinical settings has been hampered by the opaqueness and interpretability of many of its models. By fusing the strength of machine learning with the transparency offered by explainable AI (XAI) approaches, this study tackles these issues by creating an explainable AI-based model for anemia prediction.

Chapter 2

Literature review

2.1 Introduction

The study on anemia prediction, the use of machine learning (ML) in healthcare, and the function of explainable AI (XAI) in improving the interpretability and reliability of machine learning models are all thoroughly reviewed in this chapter. It looks at the several approaches to predicting anemia, the difficulties in clinical settings, and how explainable AI might help close the gap between the clinical applicability and predictive capabilities of machine learning. While pointing out current gaps that this research attempts to fill, the literature review seeks to highlight the major developments in these fields.

A significant percentage of people worldwide suffer from anemia, a common hematological disorder. According to estimates from the World Health Organization (WHO), anemia affects 1.62 billion people worldwide, or 24.8% of the population. The prevalence is highest in older individuals, pregnant women, and children under five. Iron deficiency, vitamin B12 and folate deficiencies, chronic illnesses, blood loss, and bone marrow problems are some of the causes of anemia. Iron deficiency anemia is the most common cause worldwide and is more common in developing and low-income areas because of limited access to healthcare and inadequate food intake.

2.2 Previous Literature

Classifying Anemia Using Hybrid Machine Learning Models: A Comparison of Ensemble Methods on CBC Data 2024 Random Forest, XG-Boost, Light GBM, Decision Tree, and Cat Boost

Results: Examine how hybrid machine learning models can be used to categorize different forms of anemia based on Complete Blood Count (CBC) data.

Lacking: Slightly worse performance, highlighting the trade-off between mode and computational efficiency

An Analytical Comparison of Anemia Classification algorithms for newly created and interning CBC datasets 2023: Random Forest, Logit boost. Xg-boost and Preceptron with many layers
Results: Twelve different machine learning algorithms are being tested, and the XG-Boost classifier performs best on two datasets of CBC reports.
Lacking: They are unable to identify various forms of anemia.

A Negative Illness: Forecasting and Comparing Fetal Anemia Disease Using Various Machine Learning Techniques, Logistic Regression, Gaussian Naive Bayes, KNN, SVM, and LGBM in 2023.Results: They applied five machine learning algorithms. After that, a voting classifier is used to aggregate the predictions from all five models into a single forecast.
Lacking: They are unable to identify various forms of anemia.

An Artificial intelligence-based Diagnostic System for Acute Lymphoblastic Leukemia Detection 2024, neural networks, logistic regression, and naive Bayes Results: In order to automatically detect anemia using clinical intraoral images of a patient's gingiva, this study set out to assess three distinct machine-learning techniques. Lacking: Because clinical intraoral photos of patients were used for the investigation, there would be variations in image quality.

Decision- making support system for predicting and Elimination Malnutrition and anemia2023, Random Forest (RF) algorithm, Decision Tree (DT) algorithm, and Naïve Bayes classifier (NBC) Results: Anemia and malnutrition are dangerous health conditions. Both the prediction of anemia and malnutrition are done using classification algorithms.

A new Artificial Intelligence using Extreme Learning Machine as the Potentially Effective Model to predict and Analyze the Diagnosis of anemia 2023 Decision trees (DT), k-nearest neighbor (KNN), naïve Bayes (NB), and support vector machines (SVM) Results: full blood count (CBC), but the technique is unable to distinguish between various types of anemia. Absence: This study included an overview of the health and

An Intelligence non-invasive System for Automated diagnosis of anemia exploiting a novel dataset 2022 LR (linear regression), ANN (artificial neural network), DT (decision tree), and

SVM (support vector machine). Results: economical tools used to identify anemia in a readily available Absent: One of the major scientific challenges at the moment is the use of noninvasive and affordable technology for anemia screening.

Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning 2024, Random Forest (RF), Explainable AI logistic regression (LR) Results: The main cause of anemia is a decrease in the quantity of erythrocytes in the blood. Using images of the hands, fingernails, and conjunctiva, learning models can identify anemia Lacking: Additional study on this topic would be necessary.

Approaching Explainable Artificial Intelligence Methods in the Diagnosis of Iron Deficiency Anemia Using Blood Parameters 2023 Decision trees, Naive Bayes (NB), support vector machines (SVM), and artificial neural networks (ANN) Results: The majority of anemic people have iron deficiency anemia. By looking at a XAI model, doctors can make diagnoses more quickly.

Insightful Clinical Assistance for Anemia prediction with Data Analysis and Explainable AI 2023 Random forest (RF) and logistic regression (LR) are two examples of explainable AI.
Results:
The primary cause of anemia is a decrease in the quantity of erythrocytes in the blood. Using images of the hands, fingernails, and conjunctiva, learning models can identify anemia
Absent: Further investigation into this topic would be necessary.

2.3: Summary

The research on anemia prediction, machine learning applications in healthcare, and explainable AI's (XAI) contribution to making machine learning models easier to understand has all been examined in this chapter. A common worldwide health problem, anemia can be caused by a number of factors, including blood loss, chronic illnesses, and dietary deficiencies. Effective management depends on an accurate and timely diagnosis, yet conventional diagnostic techniques are frequently constrained by resource limitations, accessibility issues, and inconsistent interpretation.

A promising method for improving anemia prediction is machine learning, which analyzes big, complicated datasets like test results, patient demographics, and medical histories. To forecast anemia risk, a variety of machine learning techniques have been used, including Decision Trees, Random Forests, XG-Boost, Logistic Regression, SVM and KNN. Even though these models have shown a great deal of predictive potential, many of them have issues with interpretability, model generalization, and data quality.

The inability of machine learning models to be interpreted is a major obstacle to their use in the medical field. To maintain confidence and make wise choices, clinicians must comprehend the logic underlying AI predictions. Explainable artificial intelligence (XAI) techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) hold great potential in this regard. These methods make model predictions more approachable and useful in clinical practice by helping doctors understand the variables affecting them. XAI can assist in bridging the gap between the predictive capabilities of machine learning and its practicality in medical decision-making.

Chapter 3

Methodology

3.1 Research Method

The process for creating an explainable AI-based model for anemia prediction is described in this chapter. It describes the procedure for gathering data, the machine learning algorithms used, how explainable AI (XAI) techniques are incorporated to guarantee model transparency, and the evaluation standards used to gauge the model's effectiveness. The goal is to develop a reliable and understandable anemia prediction system that strikes a compromise between clinical interpretability and high accuracy.

3.2 Architectural Design

3.2.1 Dataset Collection

Kaggle, a site renowned for its wide variety of datasets, especially in the healthcare industry, provided the dataset used in this study. It includes the results of roughly 1,282 patients' Complete Blood Count (CBC) tests. An accurate diagnosis of anemia type based on clinical symptoms is made possible by the diagnostic information included in each patient's data, which have been manually verified.

The dataset is a useful resource for creating and evaluating machine learning models since it focuses on using CBC data to predict the type of anemia. Hemoglobin levels, red blood cell counts, hematocrit, and other pertinent hematological metrics are among the dataset's salient aspects. These characteristics were chosen with care since they are clinically important markers for anemia diagnosis and categorization.

To guarantee data quality and get it ready for analysis, the dataset underwent preprocessing. This included resolving missing values, normalizing numerical features, and, if required, balancing class distributions. This dataset's comprehensiveness makes it possible to create a strong machine learning model that can reliably identify different kinds of anemia and offer clinically useful insights.

Dictionary of Data:

- HGB: The blood's hemoglobin content, which is essential for carrying oxygen.
- PlT: The blood's platelet count, which aids in blood coagulation.
- WBC: White blood cell count, which is essential for immunological response.
- RBC: The number of red blood cells, which carry oxygen.
- MCV: Mean Corpuscular Volume (MCV) is the average volume of one red blood cell.
- MCH: Mean Corpuscular Hemoglobin, or MCH Hemoglobin content per red blood cell on average.
- MCHC: The average amount of hemoglobin in red blood cells is known as the mean corpuscular hemoglobin concentration, or MCHC.
- PDW: an indicator of the blood's variability in platelet size distribution
- PCT: Your doctor can determine whether you have sepsis due to a bacterial infection or whether you are at high risk of getting sepsis by using a pro calcitonin test.
- Diagnosis: Based on the CBC data, the kind of anemia was diagnosed.

Attribute Information:

The following kinds can be reliably predicted by the machine learning models that are being used.

1. Good health
2. Hypochromic normocytic anemia
3. Anemia with normocytic normochromic

4. Anemia due to iron deficiency
5. Thrombocytopenia
6. An additional microcytic anemia
7. Leukemia
8. Anemia macrocytic
9. With thrombocytopenia, leukemia

The dataset has been preprocessed to guarantee quality and supports a variety of data release formats. Missing values are eliminated, numerical variables are normalized, and class distributions are balanced as needed. This dataset is a fundamental resource for creating machine learning models that predict anemia kinds while guaranteeing the predictions' dependability and interpretability.

3.2.2 Data preprocessing

Data cleaning is an essential step in enhancing the reliability of real-world data, which is usually contaminated by significant noise and redundancy. In many situations, creating models with data that has missing or duplicate variables leads to inaccurate findings. Therefore, thorough cleaning is required before entering data into a model in order to produce more accurate results. By eliminating duplicate and redundant values and fixing missing entries, preprocessing guarantees that noise is eliminated. The updated, conflict-free data serves as the foundation for the final Anemia prediction categorization model. Handling outliers, standardizing numerical data, and encoding categorical features during preprocessing are essential for a more accurate anemia prediction model.

Feature Selection

- Outlier handling
- Cleaning Duplicate data

Divide the dataset into two sets: 80% for training and 20% for testing.

3.3: Model Architecture view

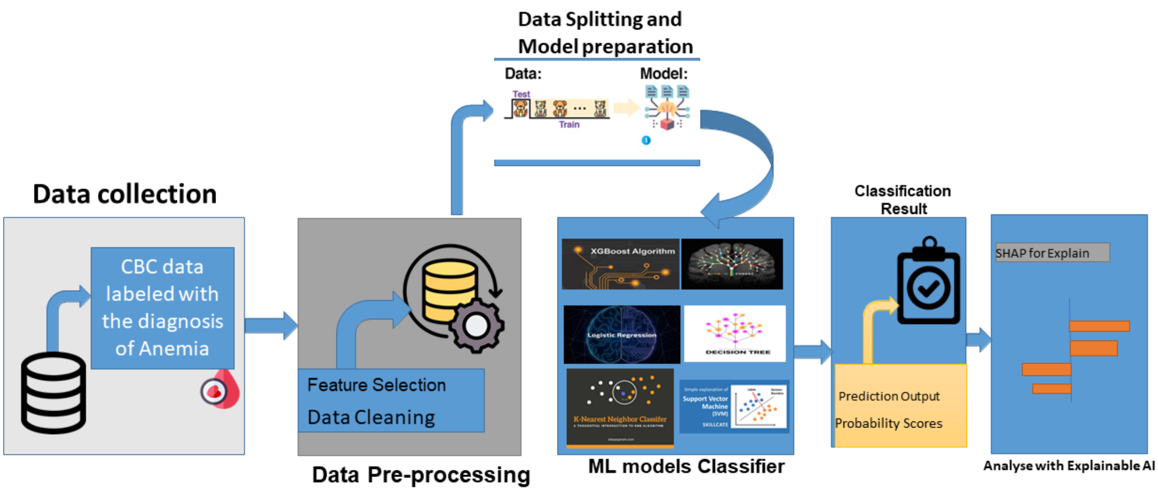

Figure 3.1: Overview of process

This study uses a variety of machine learning algorithms to optimize predictive potential and guarantee reliable anemia prediction. XG-Boost (Extreme Gradient Boosting), Decision Trees, Random Forest, Logistic Regression, Support Vector Machine (SVM) and Nearest Neighbors (KNN) are among the methods that were selected. The dataset used to train these models includes clinically significant variables for the prediction of anemia.

3.3.1 Algorithms Employed

Numerous classification strategies are essential for evaluating and interpreting patient data as well as determining the risk of anemia in early machine learning-based anemia prediction. Here are some explanations of frequently used classes.

XG-Boost: XG-Boost is renowned for its effectiveness and excellent performance. It can handle complex data patterns because it uses gradient boosting techniques to increase prediction accuracy. A powerful machine learning technique noted for its accuracy and efficacy in forecasting is Extreme Gradient Boosting, or XG-Boost. The method, initially introduced by Chen and Guestrin in 2016, combines the advantages of gradient boosting frameworks with parallel computing and a unique regularization term to maximize model performance.XGBoost excels at handling missing values, handling various data formats, and preventing overfitting through regularization. Its popularity stems from its speed, scalability, and competitive success.

Decision Tree: Decision trees are an easy-to-understand model that divides data according to feature values in order to generate predictions. This method uses supervised learning to develop a predictive model for classifying a target variable according to its characteristics. It operates on the premise that a feature's existence in the dataset affects the occurrence of another characteristic, which in turn influences the target's classification into a specific class.

Random Forest: Random Forests are one type of ensemble learning that is particularly good at resolving regression and classification issues. Together, they train a huge number of decision trees. The class that is most frequently chosen from among the trees in classification tasks is the forest's output. This ensemble approach enhances the model's performance by utilizing a large number of classifiers. The Random Forest classifier, as its name suggests, increases accuracy by averaging the outputs of several decision trees on various dataset subsets. Random Forest considers forecasts from all trees and employs the most widely recognized predictions to get a choice rather than relying solely on one. A group of decision trees that pool their outputs to increase accuracy and robustness; it works especially well at minimizing overfitting.

Logistic Regression: With its interpretability and ability to handle linear correlations between features and the goal variable, logistic regression is a statistical model that is perfect for binary classification tasks. A supervised machine learning technique for binary classification issues is called logistic regression. It is a classification algorithm, not a regression algorithm, as its name suggests. It uses a logistic (sigmoid) function and a linear equation to estimate the likelihood of a binary event (such as the presence or absence of anemia).

Support Vector machine (SVM): A popular machine learning approach for linear and nonlinear classification, regression, and outlier detection is the Support Vector Machine (SVM). Because SVM concentrate on identifying the maximum separating hyperplane between the many classes in the target feature, they are very successful at both binary and multiclass classification. The Support Vector Machine (SVM) algorithm, its uses, and how well it performs tasks involving regression, outlier detection, and both linear and nonlinear classification.

K-Nearest Neighbor (KNN): simplicity and ease of use, the (KNN) method is a popular and adaptable machine learning technique. There is no need to make any assumptions about how the underlying data will be distributed. Because it can handle both numerical and categorical data, it is a flexible choice for a wide range of datasets in classification and regression applications. The degree of similarity between the data points in a dataset serves as the foundation for this non-parametric method's predictions. Compared to other algorithms, K-NN is less sensitive to outliers.

3.3.2 Explainable AI with SHAP

The model pipeline incorporates Explainable AI (XAI) methods, including SHAP (SHapley Additive exPlanations). By describing how each input feature contributes to the model's predictions, SHAP offers a thorough grasp of feature relevance. As a result, the model's predictions become more actionable and clinically significant, increasing transparency and confidence.

3.3.3 Assessment of the Model

The models are evaluated using cross-validation techniques to guarantee robust performance. By dividing the data into several subsets, cross-validation allows the model to be iteratively trained on certain subsets and validated on the others. This method guarantees that the model generalizes well across a variety of datasets and avoids overfitting.

3.3.4 Group Method

An ensemble model architecture is utilized in order to integrate the advantages of distinct algorithms. The ensemble enhances generalization to new data and overall accuracy by combining the predictions from several algorithms. This comprehensive method improves robustness and reliability, which qualifies it for real-world use in early anemia prediction.

3.4 Libraries

1. Numpy
2. Pandas
3. Os
4. Scikit-learn
5. Matplotlib
6. XGBOOST
7. Seabron
8. Sklearn
9. Random
10. Joblib
11. Warning
12. SHAP
13. Tensorflow

3.5 Data processing

1. Data collection & Exploration:

The dataset used in this study was sourced from Kaggle, a website well-known for its extensive collection of datasets, particularly in the healthcare sector. It contains the Complete Blood Count (CBC) test results for about 1,282 patients. Each patient's data contains diagnostic information that has been manually checked to enable an accurate diagnosis of anemia type based on clinical symptoms.

Because the dataset focuses on utilizing CBC data to predict the type of anemia, it is a valuable tool for developing and assessing machine learning models. The dataset's key features include hematocrit, red blood cell counts, hemoglobin levels, and other relevant hematological measures. These traits were carefully selected since they are clinically significant indicators for the diagnosis and classification of anemia.

The dataset was preprocessed to ensure data quality and prepare it for analysis. This involved balancing class distributions if necessary, normalizing numerical features, and fixing missing data.

2.Data preprocessing

Enhancing the trustworthiness of real-world data, which is typically tainted by substantial noise and redundancy, requires data cleaning. Inaccurate results are frequently produced when models are built using data that contains missing or duplicate variables.

3.Data Splitting:

We use an 80-20% split in our study, meaning that 20% of the data is used for testing and 80% is used for training. The model's generalization is improved by this stratified split, which guarantees representative distributions of anemia prevention cases in both sets.

3.6: Model Details

To maximize predictive potential and ensure accurate anemia prediction, this study employs a range of machine learning methods. The techniques that were chosen include Decision Trees, Random Forest, Logistic Regression, XG-Boost (Extreme Gradient Boosting), and Hybrid Models. Clinically significant variables for the prediction of anemia are included in the dataset used to train these algorithms. and the outcome Explainable AI-Based Machine Learning Prediction of Anemia SHAP.

3.7: Summary

To maximize the precision and dependability of anemia prediction, this work uses a wide range of machine learning algorithms. Each of the chosen techniques—Logistic Regression, Random Forest, Decision Trees, XG-Boost (Extreme Gradient Boosting), Support vector machine(SVM) and KNN brings special advantages to the prediction process, guaranteeing a strong and adaptable strategy. A clinically relevant dataset with essential properties for diagnosing anemia, including hemoglobin levels, red blood cell counts, and other CBC parameters, is used to train these models.

The study incorporates Explainable AI (XAI) approaches, particularly SHAP (SHapley Additive exPlanations), to improve usability and trust. Clinicians can comprehend how each feature contributes to the prediction outcome thanks to SHAP's comprehensible and interpretable insights into the models' predictions. For medical decision-making, this interpretability guarantees that the forecasts are not only precise but also transparent and actionable.

CHAPTER 4

Result and Discussion

4.1 Introduction

In this study, we use machine learning and Explainable AI (XAI) to investigate a variety of approaches for early anemia prediction. In order to determine the best method for precise anemia prediction, the study assesses the effectiveness of a number of algorithms, including XG-Boost, Decision Trees, Random Forest, Logistic Regression, and Hybrid Models. A clinically relevant dataset is used to train and evaluate each algorithm, which incorporates features essential to diagnosing anemia.

Explainable AI techniques, especially SHAP (SHapley Additive exPlanations), are used to further enhance the final outcomes by offering interpretable insights into the predictions generated by these models. SHAP improves transparency and fosters confidence in the AI-driven forecasts by clarifying the role of various characteristics in model outputs.

This thorough examination not only assesses the effectiveness of these machine learning techniques but also provides information about how well they work in actual clinical settings. Finding the optimal model for accurate anemia prediction is made possible by the findings and the debate that follows, which adds to the intricate and developing field of AI-driven healthcare solutions.

4.2 Result:

This study uses the Python programming language to perform a number of categorization techniques. The experimental outcomes of allocating 80% of the data for training and 20% for validation are shown in Table 4.1.

Machine learning algorithm	Accuracy	Precision	Recall	F1-score
Decision Tree	99.03%	0.99	0.99	0.99
Random Forest	95.29%	0.95	0.95	0.95
Logistic Regression	65.75%	0.64	0.66	0.64
XG Boost	98.38%	0.95	0.98	0.98
Support vector Machine (SVM)	85.06%	0.86	0.85	0.85
KNN	65.91%	0.66	0.66	0.65

Table 4.1: Model results

4.2 Data Visualization

Understanding, analyzing, and verifying machine learning models all depend on data visualization. Visualizations aid in describing the model's operation, highlighting important characteristics, evaluating performance, and guaranteeing that the predictions are precise and understandable for medical professionals in the context of anemia prediction. The main visualization types that were employed in this work to assess and clarify the machine learning models are listed below.

Bar chart: A straightforward yet powerful tool for comparing the performance of several machine learning models is a bar chart. They make it simple to compare several evaluation measures visually, which makes it easier to choose which model predicts anemia the best.

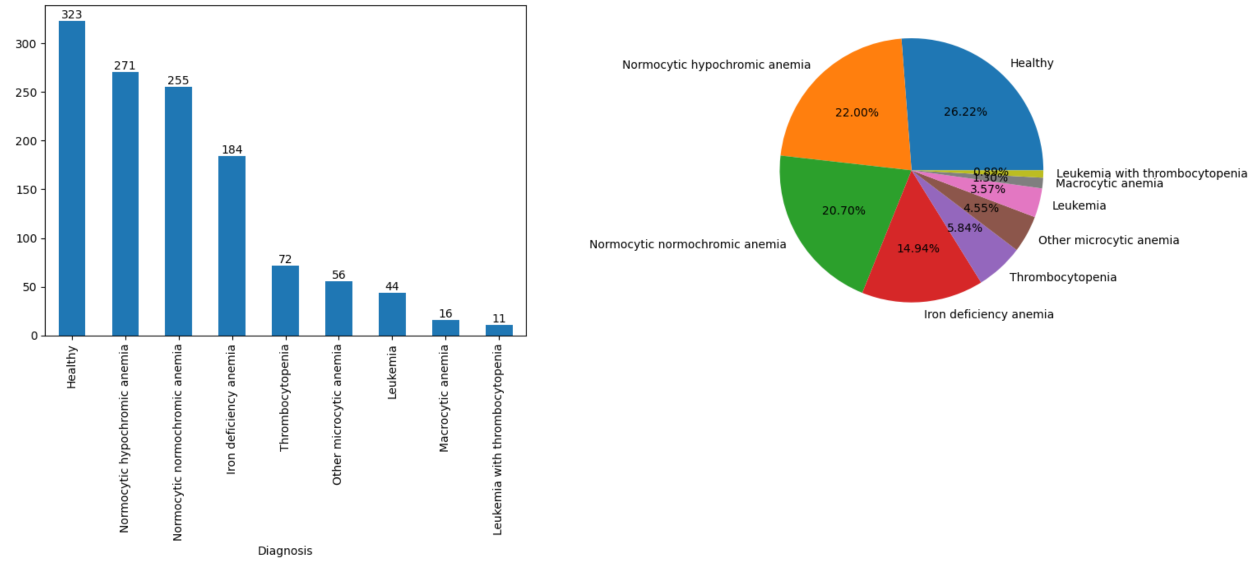

Figure 4.1 Data visualization

Pie chart: Pie charts are a straightforward yet powerful tool for visualizing proportional or categorical data, and they can offer important information about feature relevance, confusion matrix distributions, and model effectiveness. Pie charts aid in visualizing the relative contributions of several models and confusion matrix elements in anemia prediction.

4.3 Prediction Output

Six distinct machine learning algorithms—XG-Boost, Random Forest, Decision Tree, Logistic Regression, Support vector Machine (SVM) and KNN —were used in this study to predict anemia. Accuracy, precision, recall, and F1-score were among the important performance indicators used to assess each of these algorithms. The outcomes of these assessments offer a thorough comprehension of each model's advantages and disadvantages in relation to anemia prediction.

Although the majority of earlier research in this area utilized four or fewer machine learning algorithms, and many of them claimed that Decision Tree had the best accuracy, we aimed to outperform these findings by using a larger variety of methods. We sought to improve accuracy and generalization by implementing models like XGBoost and Random Forest, which are renowned for their capacity to identify intricate patterns in data.

Decision tree

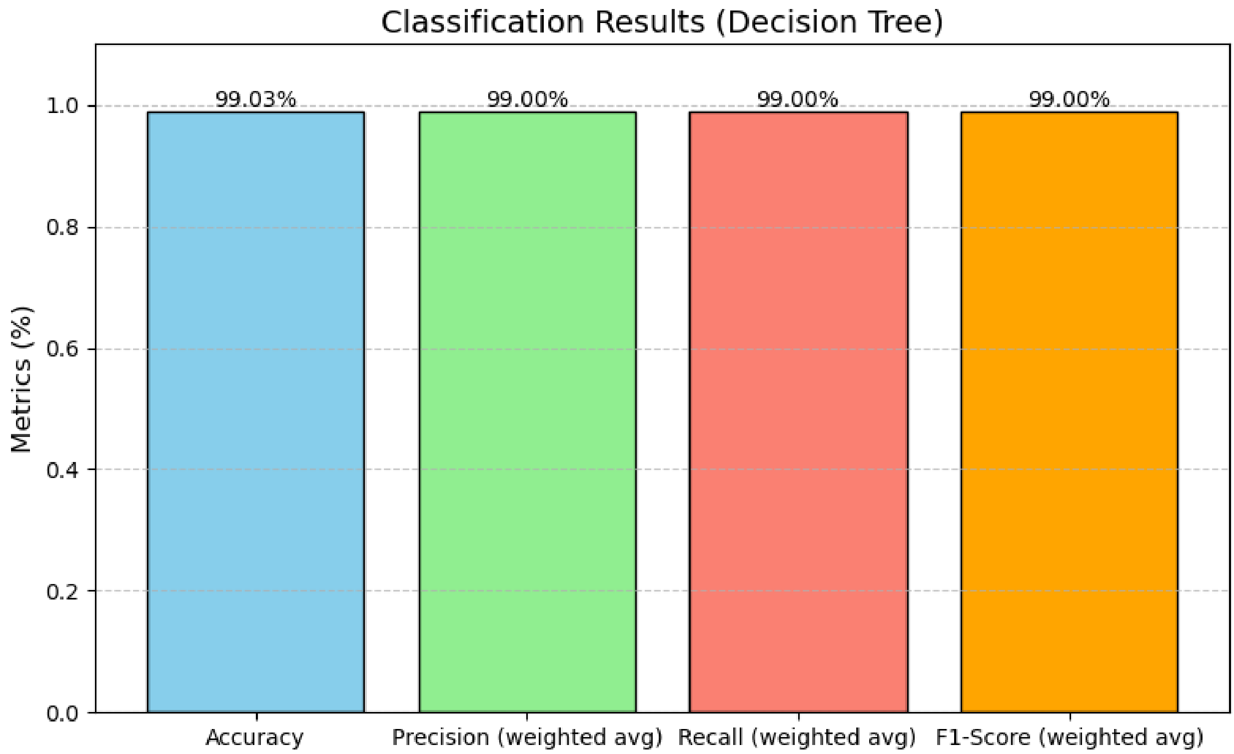

Figure 4.2: Decision tree result

The Decision Tree model's performance metrics for our dataset are quite attractive. The model's accuracy of 99.3% demonstrates its exceptional ability to identify things accurately. Furthermore, the accuracy of 99.00%,

which demonstrates the incredible accuracy of optimistic forecasts. The recall, which indicates how effectively the model captures all positive cases, is still rather good at 99.00%. Additionally, the model's overall effectiveness is attested by the F1 score, a standardized measure of precision and recall, which has a score of 99.00%.

Logistic Regression

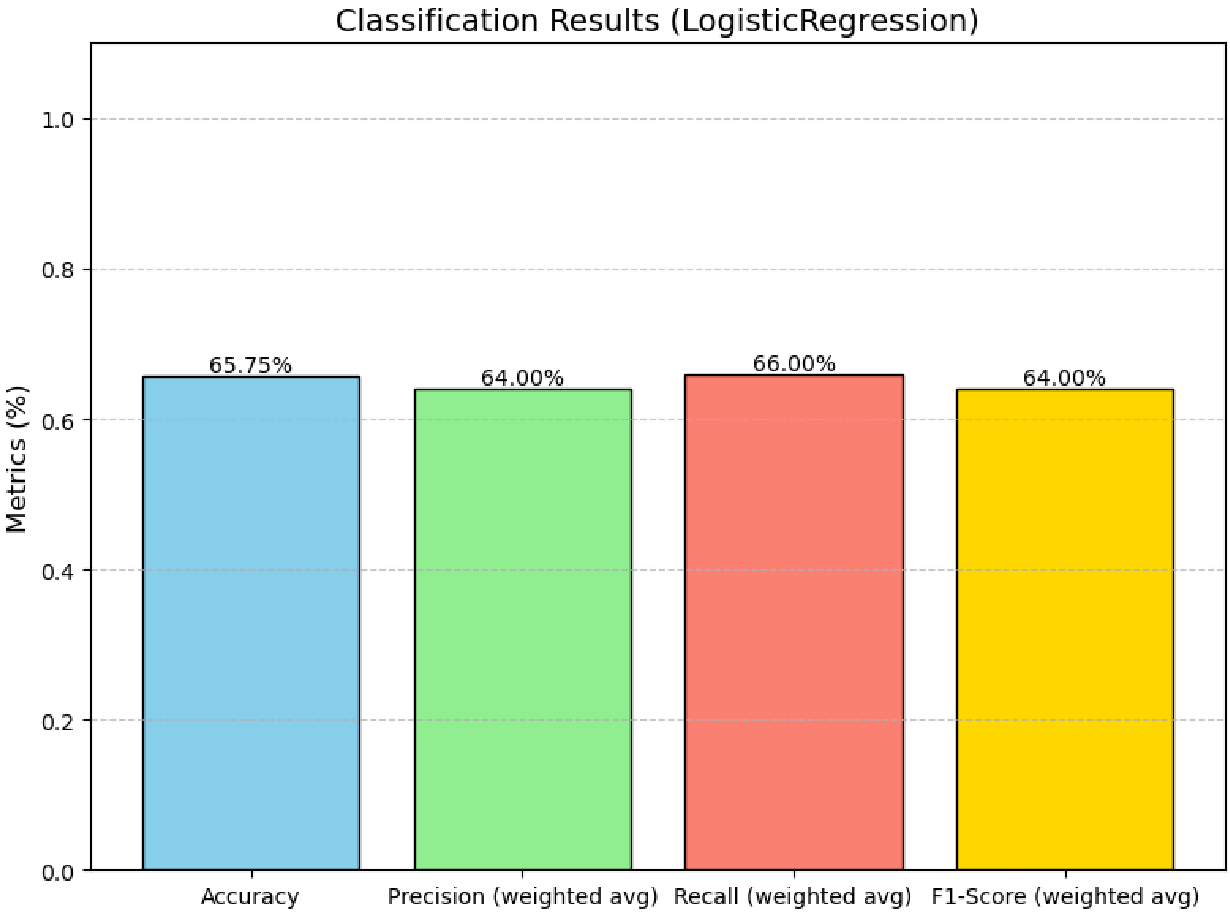

Figure 4.3: Logistic Regression result

Our dataset is used to evaluate the Logistic Regression model, and the results of its performance metrics are reliable and consistent. The model's accuracy is noteworthy; it attains a rate of 65.75%, indicating its proficiency in correctly detecting scenarios. The remarkable precision of 64.00%, which measures the accuracy of positive predictions, demonstrates the model's accuracy in identifying positive cases. Additionally, the model's recall rate of 66.00% shows how well it captures a sizable portion of positive cases. The F1 score, which strikes a balance between recall and precision, displays a balanced performance of 64.00%.

Random Forest

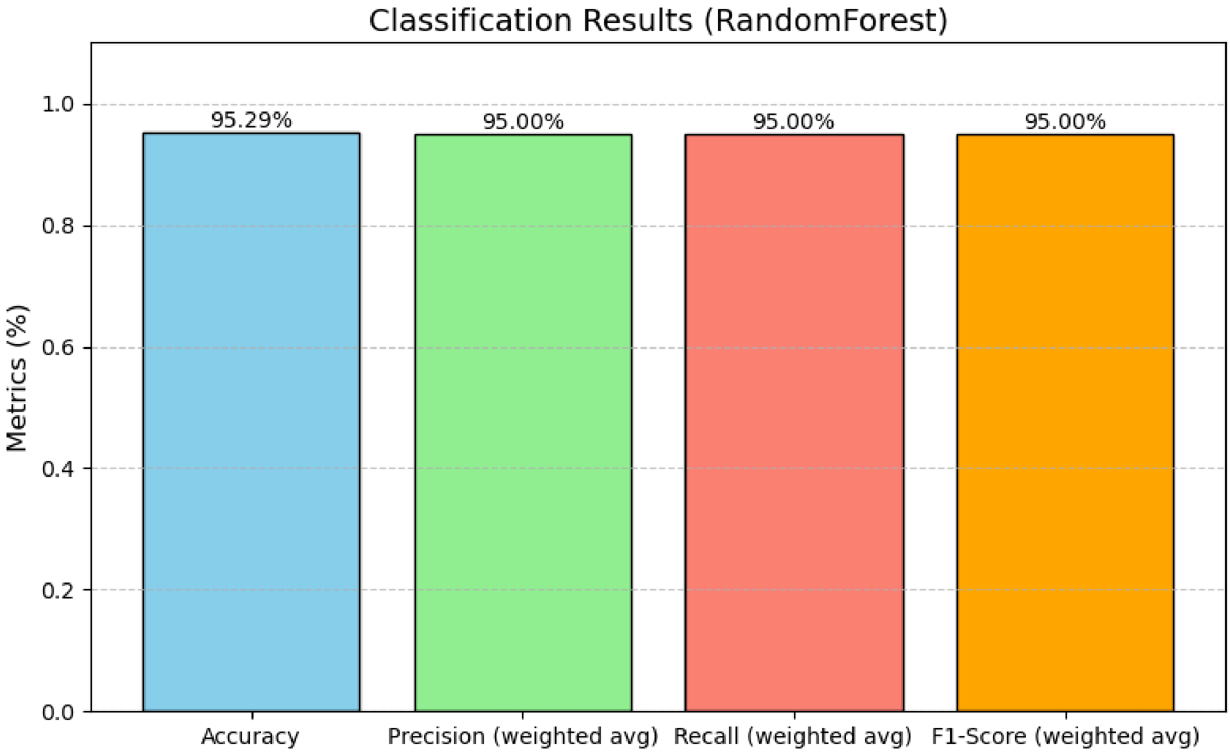

Figure 4.4: Random Forest result

Our dataset yields some incredibly good performance numbers when the Random Forest model is applied. The model's remarkable 95.29% accuracy percentage amply demonstrates its ability to classify instances accurately. Furthermore, at 95.00%,

The model's precision, which measures how well it predicts positive events, is exceptionally high. The recall, which gauges how well the model captures every positive instance, is likewise high at 95.00%. The model's overall dominance is confirmed by its 95.00% F1 score, a balanced measure that considers both recall and precision.

XG-BOOST

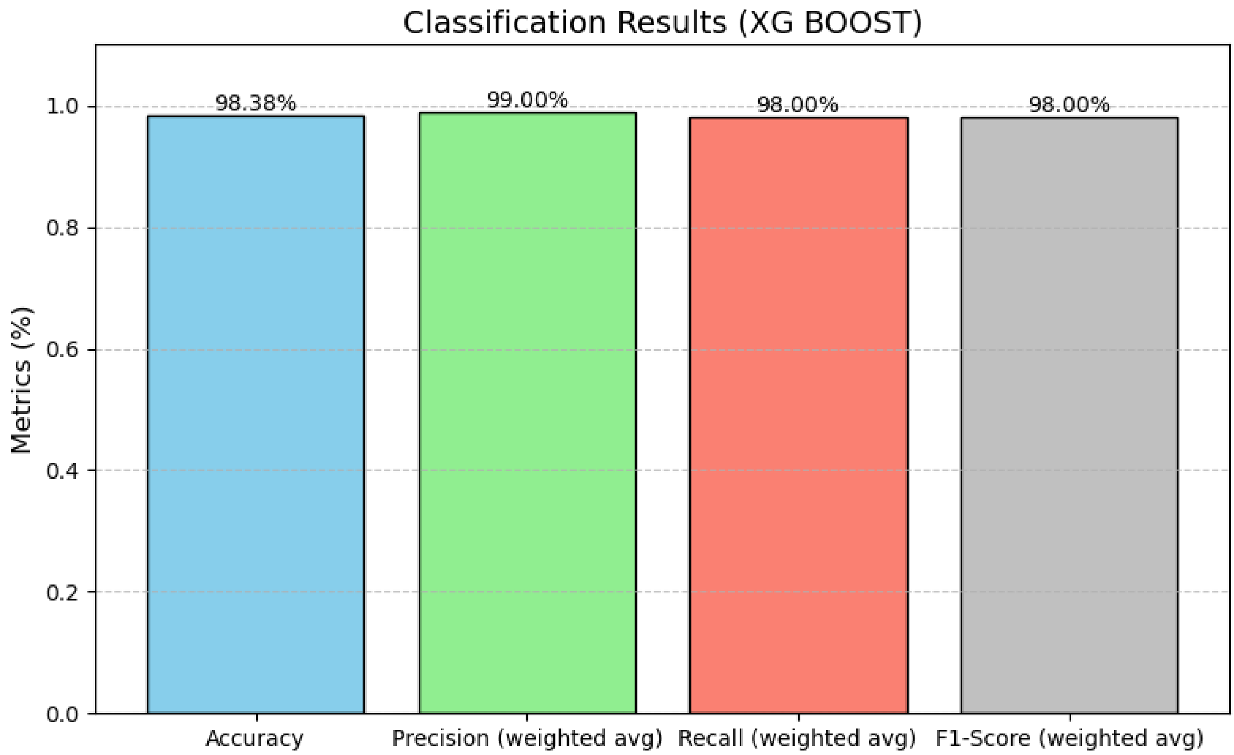

Figure 4.5: XG boost result

The XG-Boost model is tested on our dataset, and the performance measures show consistent and dependable results. The accuracy of the model is impressive; it achieves a rate of 98.38%, demonstrating its ability to accurately identify scenarios. The

The model's accuracy in recognizing positive cases is demonstrated by its impressive precision of 99.00%, which gauges the accuracy of positive predictions. Furthermore, the model's 98.00% recall rate demonstrates how well it captures a significant percentage of positive occurrences. A balanced performance of 98.00% is shown by the F1 score, which balances recall and precision.

Support Vector machine (SVM)

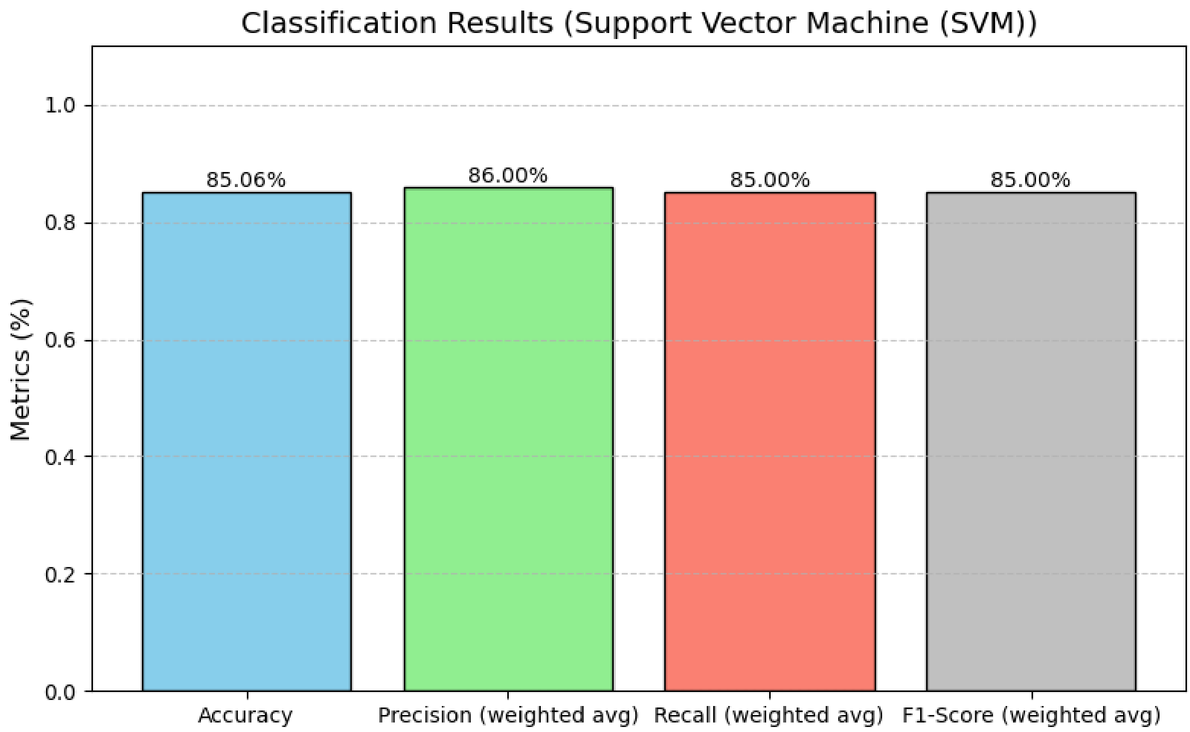

Figure 4.6: Support vector Machine result

The Support Vector Machine (SVM) is evaluated using our dataset, and the performance measures provide consistent and dependable results. The accuracy of the model is impressive; it achieves a rate of 85.06%, demonstrating its ability to accurately identify scenarios. The

The model's accuracy in recognizing positive cases is demonstrated by its impressive precision of 86.00%, which gauges the accuracy of positive predictions. Furthermore, the model's 85.00% recall rate demonstrates how well it captures a significant percentage of positive occurrences. A balanced performance of 85.00% is shown by the F1 score, which balances recall and precision.

K-Nearest Neighbor (KNN)

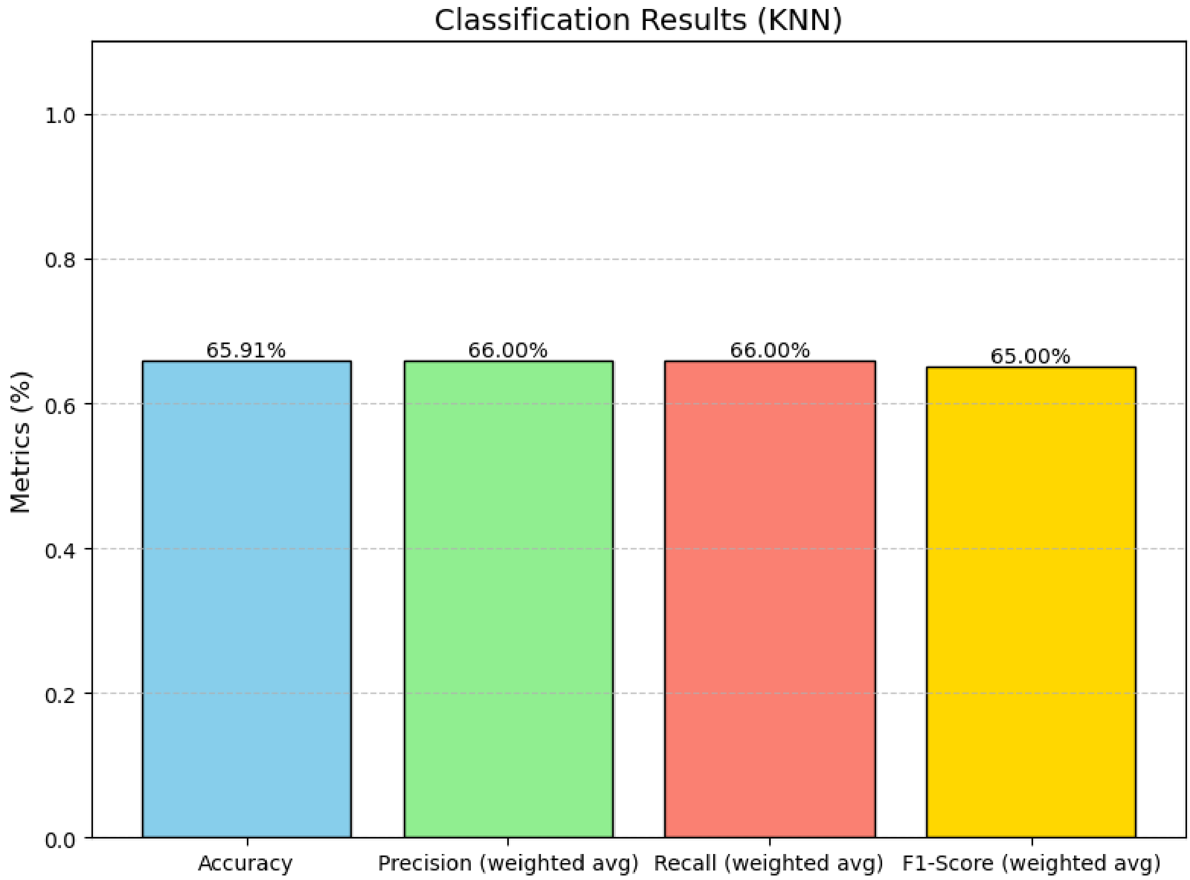

Figure 4.7: K-Nearest Neighbor (KNN)

Our dataset is used to assess the K-Nearest Neighbor (KNN), and the performance metrics yield reliable and consistent findings. With a rate of 65.91%, the model's accuracy is outstanding, proving that it can correctly detect scenarios. The

The remarkable precision of 66.00%, which measures the accuracy of positive predictions, demonstrates the model's ability to identify positive cases. Additionally, the model's recall rate of 66.00% shows how well it captures a sizable portion of positive events. The F1 score, which balances recall and precision, displays a balanced performance of 65.00%.

4.4 Classification Result

Using the six machine learning algorithms—XG-Boost, Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine and K-Nearest Neighbor (KNN) — we show the classification findings for anemia prediction in this section. These models' performance was evaluated using common categorization metrics, such as accuracy.

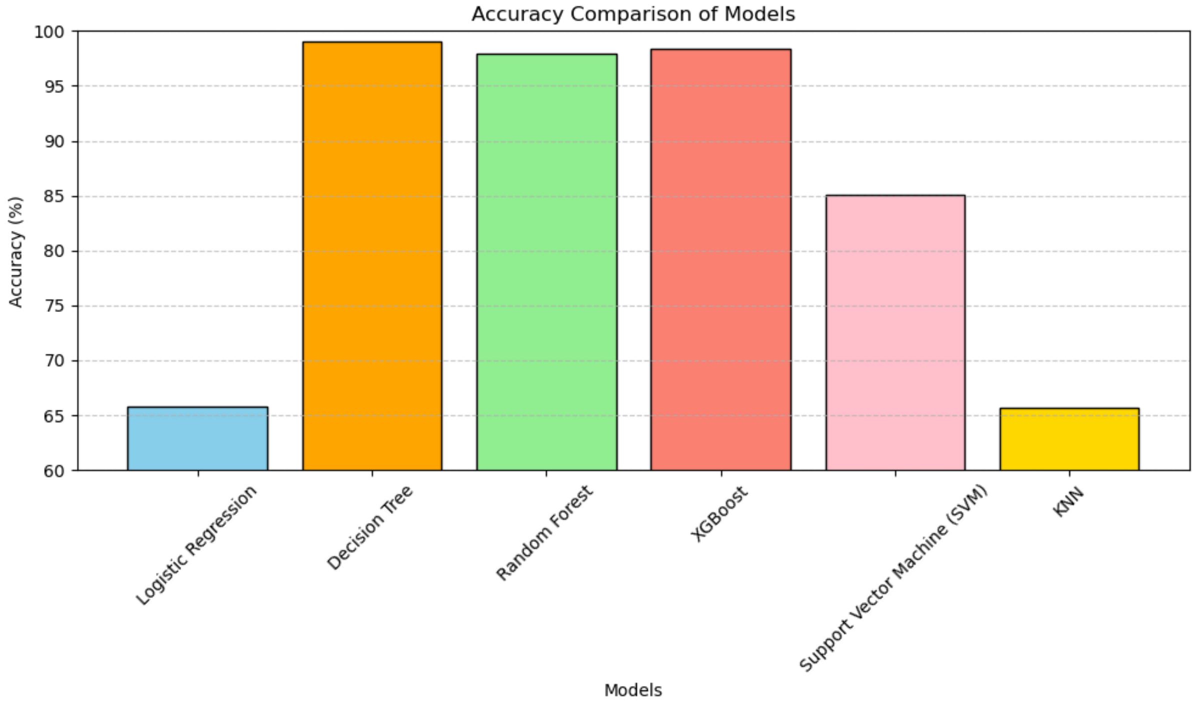

Figure 4.8: Classification result

Overall, the Decision Tree model performs the best.

Comparison of Model Precision Score

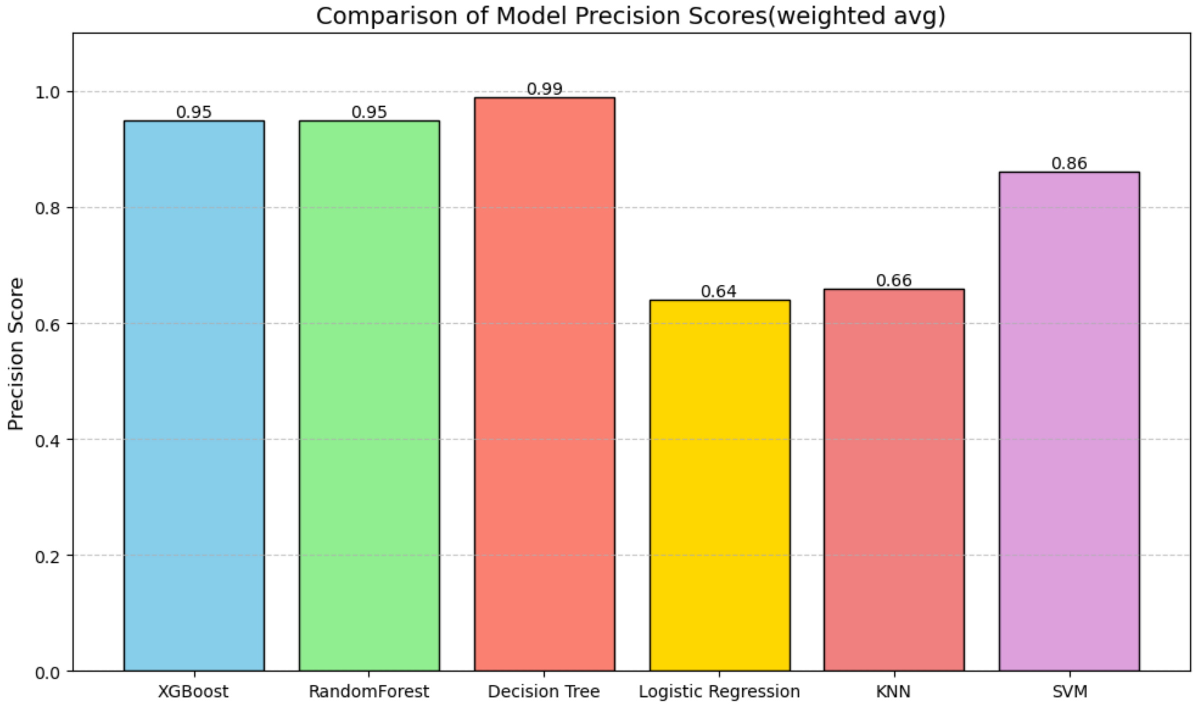

Figure 4.9: Comparison of Model Precision Score

The precision scores of several machine learning models are displayed in the bar chart, emphasizing how well they perform in categorization tasks. This is the synopsis:

Top Performers:

With the best precision score of 0.99, Decision Tree is the most dependable model for reducing false positives.

With precision ratings of 0.95, XG-Boost and Random Forest come in second and third, respectively, showing robust and reliable performance.

A moderate performer

With a precision score of 0.86, SVM (Support Vector Machine) demonstrates respectable dependability but falls short of Decision Tree, XG-Boost, or Random Forest.

Poorer Performers:

With corresponding precision ratings of 0.66 and 0.64, KNN and Logistic Regression lag behind. These models might not be the best fit for this issue or could need more fine-tuning.

Comparison of Model Recall Score

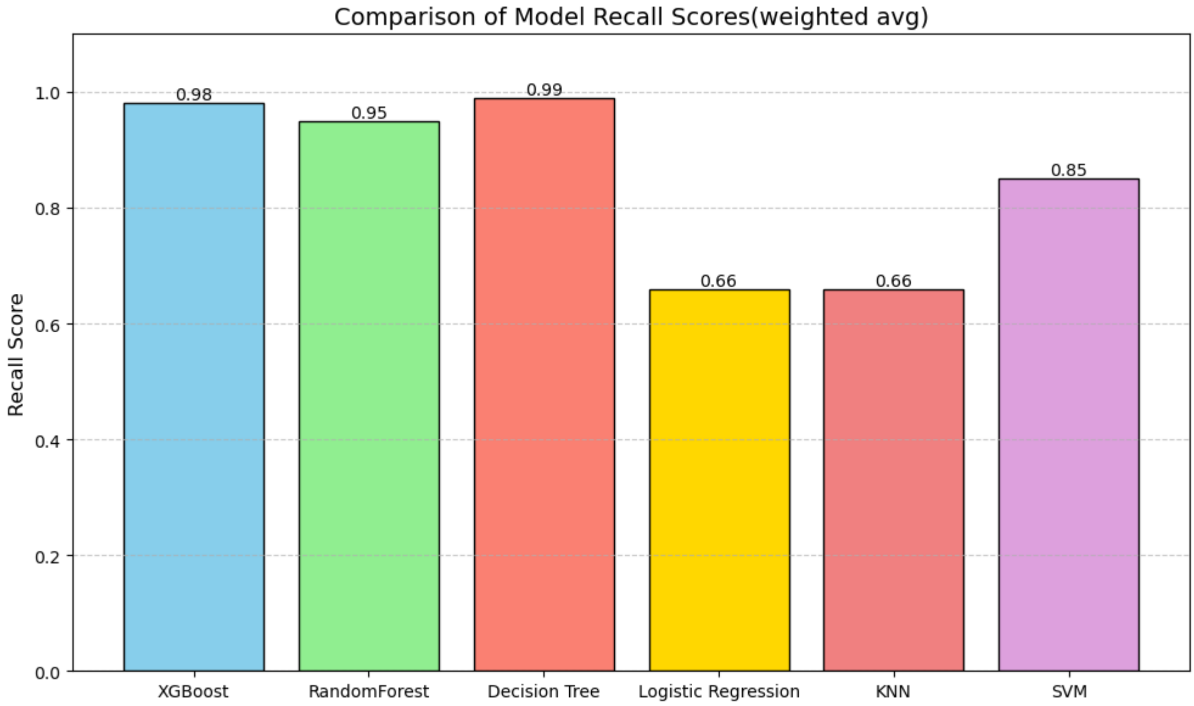

Figure 4.10: Comparison of Model Recall Score

The weighted average recall scores for different machine learning models are compared in the bar chart. With a focus on reducing false negatives, recall quantifies the percentage of true positives that were accurately detected. The Decision Tree model performs the best, identifying nearly all true positives with an effective recall score of 0.99 with a recall score of 0.98, XG-Boost is a great option that demonstrates its ability to handle complicated data while preserving good recall. Random Forest is comparable to XG-Boost and Decision Tree with a recall score of 0.95. There aren't many false negatives in this strong model. SVM performs relatively well, with a recall score of 0.85. It nevertheless shows a respectable level of dependability in detecting true positives, while not being as good as the best models with recall scores of 0.66, KNN and logistic regression are the least successful models in this comparison. Higher false negative rates could result from their failure to detect a sizable percentage of true positives.

Comparison of Model F1-Score

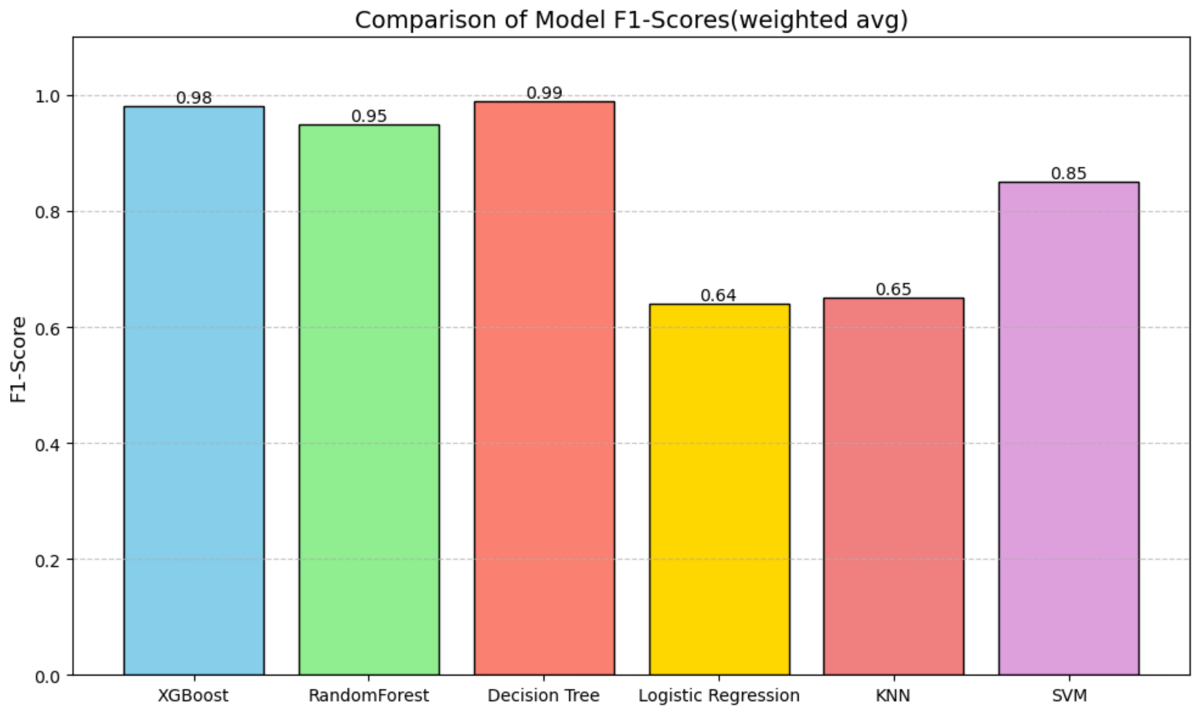

Figure 4.11: Comparison of Model F1-Score

The weighted average F1-scores for different machine learning models are compared in the bar chart. For unbalanced datasets where both precision and recall are significant, the F1-score is an essential indicator that strikes a compromise between the two. With the best F1-score of 0.99, the Decision Tree model demonstrates its remarkable capacity to strike a balance between recall and precision. Because of this, it is the most trustworthy model in our comparison a close second is XG-Boost, with an F1-score of 0.98. It offers a good balance between memory and precision and manages intricate relationships in the data with ease with a score of 0.95, Random Forest exhibits strong performance. For activities that require high F1-scores, it is still a competitive option.

SVM's F1-score of 0.85 indicates a decent level of dependability. Although it performs well, it lags behind the top three models by a small margin. KNN performs worse than the best models, as evidenced by its F1-score of 0.65. It has trouble in striking a balance between recall and precision in this comparison, Logistic Regression performs poorly, with the lowest F1-score of 0.64. It might have trouble identifying intricate patterns in the data, producing less-than-ideal outcomes.

4.5 Explainable AI Analysis (SHAP)

A potent Explainable AI (XAI) technique called SHAP (SHapley Additive exPlanations) sheds light on how each attribute affects a model's prediction. Shapley values, a notion taken from cooperative game theory, serve as the foundation for SHAP values. They equitably divide the "payout" (model prediction) among the features (players) according to their contributions.

When it comes to predicting if a patient has anemia, SHAP can assist in elucidating how certain characteristics (such hemoglobin levels, red blood cell count, and other CBC metrics) affect the model's predictions.

4.5.1: Visualizing SHAP values

1. **Summary plot:** One of the best tools for displaying the global feature importance in machine learning models is the SHAP Summary Plot, especially for complicated models like Random Forest, XG-Boost, and others used in anemia prediction. These charts offer a simple and understandable method to comprehend how each feature affects the model's predictions for the complete dataset.

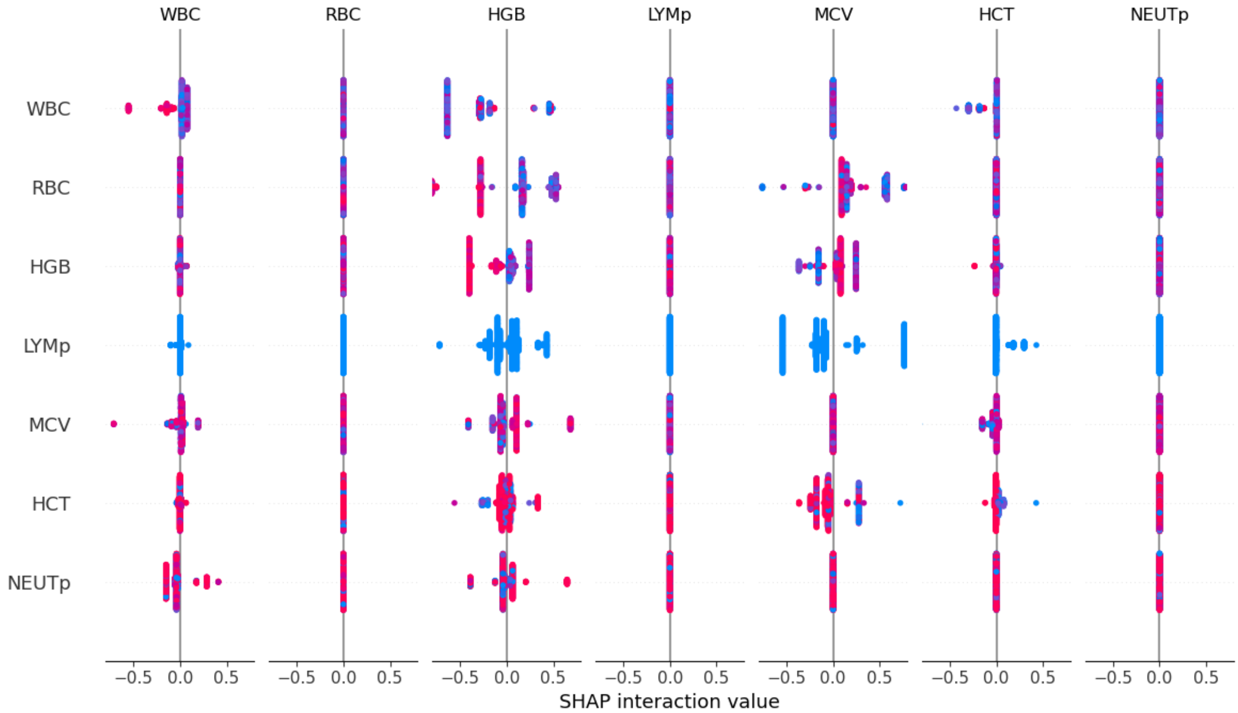

Figure 4.12: summary plot

2.Waterfall plot

An effective visual aid for illustrating how each feature contributes to a model's prediction for a specific data point is a SHAP Waterfall Plot. Beginning with the baseline (average prediction), it illustrates how each characteristic affects the final forecast, "flowing" through the contributions of the features until it reaches the final predicted value.

Clinicians can better understand how specific clinical parameters, such as hemoglobin levels, red blood cell count, and other CBC metrics, contribute to the model's ability to predict whether a patient has anemia or not by using a SHAP Waterfall Plot in the context of anemia prediction.

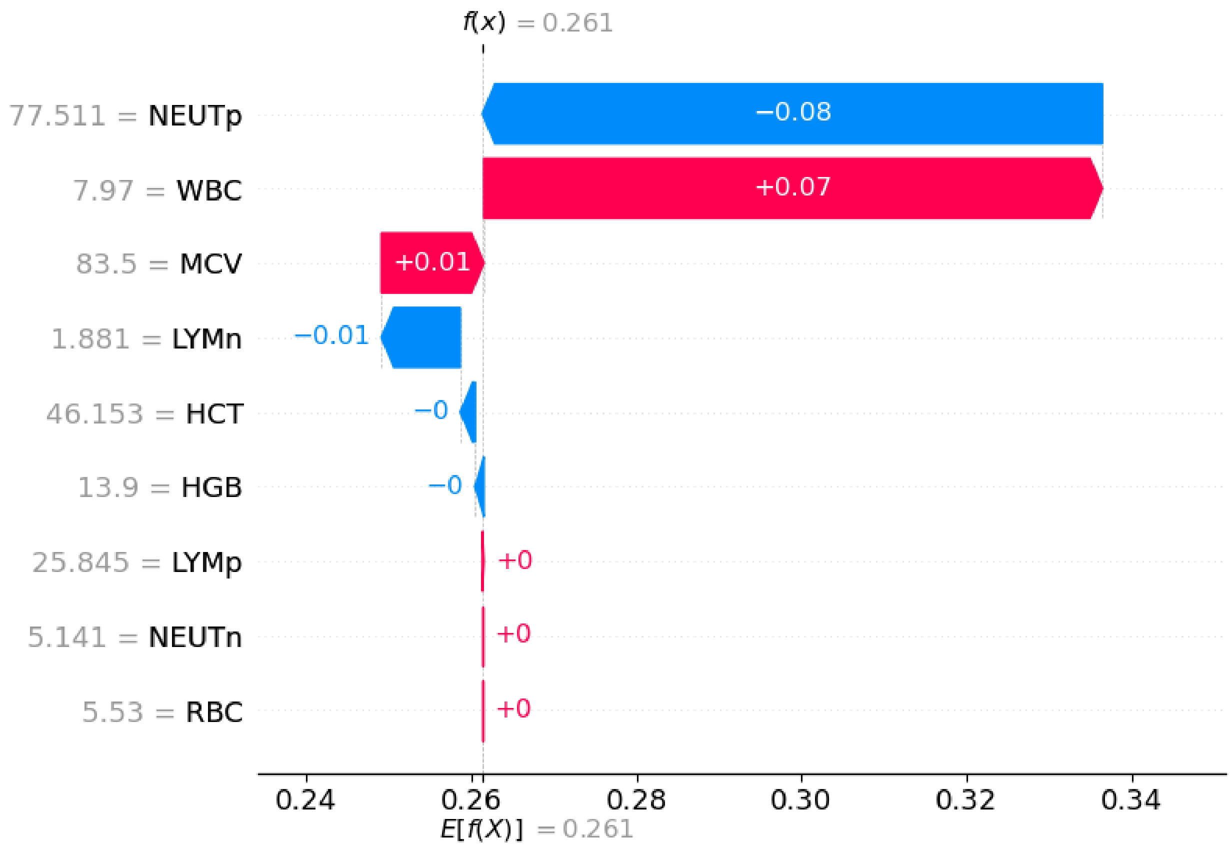

Figure 4.13: Waterfall plot

4.6: Summary

Predicting anemia has been found to be successful when several machine learning models are combined with SHAP for explainability. This research is an important step toward the creation of AI-driven clinical decision support systems since it uses an ensemble method, capitalizes on the capabilities of several algorithms, and employs XAI techniques. These models guarantee that healthcare practitioners can trust and use AI-based solutions because they not only produce accurate predictions but also improve the interpretability of outcomes.

The foundation for future research into creating increasingly complex, comprehensible, and therapeutically useful machine learning models for the diagnosis and treatment of anemia and other medical disorders is laid by this study.

CHAPTER 5

Future Scope & Conclusion

5.1 Conclusion

This study illustrates how Explainable AI (XAI) can improve the precision, openness, and reliability of machine learning models used to predict anemia. We have investigated and compared the efficacy of several sophisticated machine learning algorithms, such as XGBoost, Random Forest, Decision Trees, Logistic Regression, Support Vector Machine (SVM) and KNN in predicting anemia from clinical data, including hemoglobin levels, MCV, RBC count, and other significant characteristics from Complete Blood Count (CBC) tests.

This study demonstrates Explainable AI's (XAI) potential in the healthcare industry, namely in the prediction of anemia. We can create solutions that not only offer precise predictions but are also interpretable and useful in clinical contexts by fusing explainable models with potent machine learning techniques. Enhancing the transparency and trust of AI models is essential for the wider use of AI in healthcare, and the application of SHAP for interpretability shows great promise in this regard. As the area develops, more precise, dependable, and trustworthy AI-driven medical solutions will result from additional improvements in model performance, interpretability, and real-world testing.

5.2: Finding & Contribution

Finding

1.Decision tree: It performed best on all criteria (accuracy: 99.03%, F1-score: 0.99), which makes it a good choice for jobs demanding a high degree of dependability.

2.Random Forest: Although it performs marginally worse than the Decision Tree, it is still rather well (Accuracy: 95.29%, F1-score: 0.95).

3.Logistic Regression Model: With an accuracy of 65.75% and an F1-score of 0.64, it performs much worse than Decision Tree and Random Forest, suggesting that it might not be the best option for this dataset or issue.

4. XG-Boost: With an accuracy of 98.38% and an F1-score of 0.98, XG Boost (Regression) performs exceptionally well, indicating that it is a good fit for regression tasks.

5. Support vector Machine (SVM): exhibits robustness when combining models, performing similarly to XG Boost with minor deviations (Accuracy: 85.06 %, F1-score: 0.85).

6. KNN: It performs significantly worse than Decision Tree and Random Forest, with an accuracy of 65.91% and an F1-score of 0.65, indicating that it might not be the ideal choice for this dataset or problem.

Contribution:

1. The table shows how well various models perform predicting tasks.

2. Overall, the Decision Tree model performs the best.

3. With competitive metrics, the Random Forest and XG-Boost show promise for regression.

5.3 Future Scope

Explainable AI (XAI) has enormous potential for anemia prediction, and there are a number of exciting directions for further study and advancement. With ongoing advancements in model performance, interpretability, and real-world deployment, machine learning (ML) approaches have the potential to greatly increase the use of AI in healthcare, especially in the diagnosis of anemia.

Explainable AI-based anemia prediction has a bright future ahead of it, with a plethora of intriguing avenues for further study and practical implementation. The potential for more precise, dependable, and useful AI-driven predictions in healthcare will grow as machine learning approaches, data gathering strategies, and AI interpretability tools develop. We can guarantee that AI-based anemia prediction systems significantly help to clinical decision-making and better patient outcomes worldwide by concentrating on performance improvement, ethical considerations, real-world integration, and continuous learning.

REFERENCES

[1] Gregorius Airlangga Engineering Faculty, Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia Anemia Classification Using Hybrid Machine Learning Models: A Comparative Study of Ensemble Techniques on CBC Data

[2] D'Souza, P., & Bhargava, R. (2023). Prediction of anemia using Naïve-Bayes classification algorithm in machine learning. In Lecture Notes in Networks and Systems (Vol. 682).

[3] Santosh Aiwale1, Mahesh T. Kolte1, Varsha Harpale1, Varsha Bendre1, Deepti Khurge1, Sheetal Bhandari1, Suvarna Kadam1 and Altaf O. (2023) (Mulan Non Invasive Anemia Detection and Pre Diagnosis)

[4]. Dimas Chaerul Ekty Saputra 1 , Khamron Sunat 1 and Tri Ratnaningsih 2, 2023 ANewArtificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia.

[5]. Shehab, S. A., Shehab, E., & Khawaga, A. (2023). Anemia diagnosis and prediction based on machine learning. Kafrelsheikh Journal of Information Sciences, 4(2), 1-9. https://doi.org/10.21608/kjis.2023.220945.1014

[6] Gregorius Airlangga, 2024 Anemia Classification Using Hybrid Machine Learning Models: A Comparative Study of Ensemble Techniques on CBC Data

[7] SafaS.Abdl Jabbar1,AlaaK.Farhanand ,Alexander, S.Luchinin 2023 A Comparative study Of anemia Classification algorithms for international and newly cbc datasets .

[8] Mahadi Hasan, Mst. Sazia Tahosin, Afia Farjana. 2023 A Harmful Disorder: Predictive and Comparative Analysis for fetal Anemia Disease by Using Different Machine Learning Approaches.

[9] Vohra, R., Hussain, A., Dudyala, A. K., Pahareeya, J., & Khan, W. (2022). Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting. PLOS ONE. https://doi.org/10.1371/journal.pone.0269685

[10] Yousre El Alaoui,Regina PadmanabhanManasvi Jagadeesh Maasthi1 Harinahalli Lokesh Gururaj2 2023 Decision-making Support System for Predicting and Eliminating Malnutrition and Anemia

[11] https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification

[12] Giovanni Dimauro a, Maria Elena Griseta b , Mauro Giuseppe Camporeale 2022 An intelligent non-invasive system for automated diagnosis of anemia exploiting a novel dataset

[13] Siddartha Pullakhandam and Susan McRoy (2024) Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning

[14] Kasthuri.Ea , S.Subbulakshmib, Rajasree Sreedharan 2024 Insightful Clinical Assistance for Anemia Prediction with Data Analysis and Explainable AI.

[15] Uhma Ponnusamy Dhruva Darshan B S Niranjana Sampathila 2023 Approaching Explainable Artificial Intelligence Methods in the Diagnosis of Iron Deficiency Anemia Using Blood Parameters.

[16] Asmaa M. El Boghdady*, Sherif Kishk, Mohamed M. Ashour, Eman AbdElhalim (2023) Machine-learning Based Stacked Ensemble Model for Accurate Multi Classification of Complete Blood Count Anemia. .

[17] Abdul-Jabbar, S. S., Farhan, A. K., & Luchinin, A. S. (2023). A comparative study of anemia classification algorithms for international and newly CBC datasets. International Journal of Online & Biomedical Engineering, 19(6), 141-157.

[18] Mantas, J. (2023). An artificial intelligence-based diagnostic system for acute lymphoblastic leukemia detection. In Healthcare Transformation with Informatics and Artificial Intelligence. Retrieved from books.google.com

Plagiarism Report

Anemia Prediction

ORIGINALITY REPORT

18%	13%	10%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	"Practical Statistical Learning and Data Science Methods", Springer Science and Business Media LLC, 2025 Publication	1%
4	doctorpenguin.com Internet Source	1%
5	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	1%
6	Parikshit Narendra Mahalle, Yashwant Sudhakar Ingle. "Explainable Artificial Intelligence: A Practical Guide", River Publishers, 2024 Publication	1%

7	Submitted to Ajman University of Science and Technology Student Paper	1%
8	academic-accelerator.com Internet Source	<1%
9	bsj.uobaghdad.edu.iq Internet Source	<1%
10	www.frontiersin.org Internet Source	<1%
11	minds.wisconsin.edu Internet Source	<1%
12	Submitted to NCC Education Student Paper	<1%
13	tigerprints.clemson.edu Internet Source	<1%
14	medifitpro.com Internet Source	<1%
15	Submitted to American University of Central Asia Student Paper	<1%
16	Submitted to Kingston University Student Paper	<1%
17	Kamal Malik, Moolchand Sharma, Suman Deswal, Umesh Gupta, Deevyankar Agarwal, Yahya Obaid Bakheet Al Shamsi. "Explainable	<1%

Account Clearance