Exploration of Anemia in Hematology Patients: Using Artificial Intelligence Present and Future Perspective

By
Mahamudul Hasan
212-15-4221

FINAL YEAR PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Dr. Arif Mahmud

Associate Professor & Associate Head
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by

Chayti Saha
Lecturer
Department of Computer Science and EngineeringDaffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

May 14, 2025

APPROVAL

This Project titled **"Exploration of Anemia in Hematology Patients: Using Artificial Intelligence Present and Future Perspective"**, submitted by Mahamudul Hasan, ID No: **212-15-4221** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **14 May, 2025**.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) — **Chairman**
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mohammad Jahangir Alam (MJA) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Mohammad Masum Bakaul (MB) — **Internal Examiner**
Sr. Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Arshad Ali (DAA) — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Arif Mahmud, Associate Professor & Associate Head**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Arif Mahmud
Associate Professor & Associate Head
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Chayti Saha
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Mahamudul Hasan

Mahamudul Hasan
Student ID: 212-15-4221
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

The assistance and participation of several people during the previous two semesters were essential to the completion of this task. We really appreciate all of the people who have helped us in any form.

First, we would like to sincerely thank God for His wonderful grace, which enabled us to successfully finish the **Final Year Project (FYDP).**

Dr. Arif Mahmud, Associate Professor and Associate Head of the Department of Computer Science and Engineering at Daffodil International University in Dhaka, Bangladesh, has our sincere gratitude and gratitude. To complete this assignment, our supervisor has extensive understanding of and a strong interest in **artificial intelligence**. The completion of this assignment has been made possible by his unending patience, academic guidance, unceasing encouragement, vigorous and frequent monitoring, constructive criticism, insightful counsel, reviewing several subpar versions, and fixing them at every level.

In addition to other academic members and personnel of Daffodil International University's Department of Computer Science and Engineering, we would like to sincerely thank the Head of the Department for his helpful assistance in completing our research.

We would like to express our gratitude to all of our Daffodil International University classmates who participated in this conversation throughout the course of their studies.

Lastly, we must respectfully thank our parents for their unwavering patience and support.

ABSTRACT

Anemia is a public health disease that reflects the lack of red blood cells or hemoglobin to carry the requisite amount of oxygen to the body, causing weakness, fatigue, and diminished cognition. It is a gigantic public health concern in developing countries like Bangladesh, where the timely diagnosis is still hampered because of less availability of resources and low awareness amongst common people. The study begins to create an intelligent and accurate anemia prediction model using various machine learning models with patient data from a Bangladesh general hospital. The methodology involved a systematic preprocessing of the dataset, including dealing with missing values, normalization, and categorical encoding, and splitting the dataset into training data and test data. Ten historical classifiers were utilized: Random Forest (RF), Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Quadratic Discriminant Analysis (QDA), Ridge Classifier (RC), Passive Aggressive (PA), XGBoost (XGB), and Grid Search CV (GS). Bagging, boosting, and voting were carried out to ensemble and improve classifiers. Accuracy, precision, recall, and F1 score were used as the measurement metrics. According to the findings, the Voting classifier performed better than all others in performance with the highest accuracy of 93.12%, followed by KNN, GS, and RF under bagging and the default setting. Boosting generally delivered a mixed performance with overfitting shortening some of the models. The study concludes.

Keywords- Anemia Prediction, Machine Learning, Multiclass Classification, Ensemble learning, Voting Classifier, Bagging and Boosting, Bangladesh Healthcare, Medical Diagnosis AI.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

The majority of people all over the world are impacted by anemia, which is presented in this chapter. Less effort is needed to diagnose anemia with the help of artificial intelligence.

1.1 Introduction

Anemia is a global health issue that is impacting individuals of all ages but the most vulnerable are children under the age of five. It is a condition where there is a reduction in the amount of hemoglobin, a reduction in red blood cells, or a reduction in hematocrit levels in blood. Hemoglobin, which is a component of red blood cells, is responsible for transporting oxygen from the lung to the body tissues. When hemoglobin falls below normal, the oxygen-carrying capacity of the body is decreased, and manifestations are fatigue, weakness, shortness of breath, and compromised mental and physical development in children.

If a child's hemoglobin level falls below 110 grams per liter (g/L) or 11.0 grams per deciliter (g/dL), the World Health Organization (WHO) diagnoses the child with anemia [2]. There are several other causes of anemia, and they can differ according to the public health infrastructure, geographic region, and socioeconomic status. Vitamin B12, folate, or iron deficiency, severe blood loss from menstruation or trauma, infectious parasitic infestations like hookworm or malaria, chronic diseases, and inherited conditions like sickle cell anemia or thalassemia can all lead to anemia [1].

In Bangladesh, anemia is a priority public health concern, especially among children aged five years and below and reproductive-age women. A large proportion of the population suffers from micronutrient deficiency, and the most common cause of anemia is iron deficiency. According to recent national surveys, approximately 33%–40% of Bangladeshi children aged five years and below are anemic, and the prevalence is more in rural and disadvantaged areas. Risk factors include poverty, ignorance about nutrition, a poor diet, too high an incidence of parasitic disease, and limited access to quality healthcare.

Anemia also has different categories with differing causes and cure. The most prevalent, iron deficiency anemia, occurs when not enough iron is present to create hemoglobin. Vitamin deficiency anemia arises as a result of the insufficiency of crucial nutrients such as folate and vitamin B12 that are necessary for red blood cell formation. Sickle cell anemia and thalassemia are inherited forms that result in the distortion and the capacity of red blood cells. Aplastic anemia is the result of the failure of the bone marrow to form sufficient amounts of blood cells, while hemolytic anemia is the result of the premature breakdown of red blood cells [3].

Management of anemia, particularly in children up to five years, must be timely, i.e., diagnostically timely and intervention-wise timely, and strategic public health intervention. In low-resource zones like most in Bangladesh, it may not, however, always be feasible considering the lack of trained staff, infrastructure, and diagnostic facilities.

In recent decades, the coalescence of artificial intelligence and machine learning technological developments in medicine has shown to have the possibility of

overcoming all such constraints. Machine learning is a field of AI that enables computers to learn from experience and improve efficiency over time without being programmed specifically for this purpose [4]. ML algorithms can identify patterns, diagnose illness, and make predictions with increased accuracy as more data is added. This capability makes ML a critical tool in medical diagnosis and treatment planning.

Machine learning software, if applied to medical information, bring new information and new perspectives through which one can view, and these perspectives and information can significantly enhance the decision-making of doctors. Such software can be taught using previous histories of health, laboratory results, demographics, and environmental factors to detect early warning signs of illnesses like anemia and predict outcomes with very high accuracy [5]. For Bangladesh, with its medical care system usually suffering from shortages of trained medical professionals, coupled with heavy disease burdens, such smart systems can prove to be invaluable in supporting front-line medical care workers and streamlining the diagnostic process.

Data mining, the process of extracting useful information from enormous datasets, has a direct relation with ML. It encompasses two main categories of tasks: descriptive data mining and predictive data mining. Descriptive data mining tries to establish the general characteristics or properties of a dataset, and this can prove useful in identifying common trends or patterns. Predictive data mining, on the other hand, employs past data to predict future or unknown events. This predictive approach is highly useful in the healthcare sector, where early diagnosis and prevention save lives and reduce costs [6, 7].

The application of ML is expanding rapidly across various industries, and the health sector is among the most promising sectors to implement ML. ML techniques are utilized in disease diagnosis, treatment recommendation, patient outcome prediction, drug discovery, and public health surveillance. Disease diagnosis and outcome prediction are two areas where ML has proved particularly promising [8]. ML algorithms can process vast amounts of patient information, identify subtle patterns between variables, and present real-time results that might not be easily discernible to human doctors.

In Bangladesh, the healthcare sector retains massive volumes of patient data in the shape of hospital records, lab reports, demographic information, and even health surveys of the community. However, in the majority of cases, such data remain latent as the industry is devoid of advanced analytical software. ML can fill this void by extracting usable intelligence from existing data sets. Trends in hemoglobin levels in various zones can, for example, be evaluated with ML models to inform policymakers where high-risk areas exist and resource optimization can take place.

Besides, anemia diagnosis has traditionally relied on laboratory tests, which may not always be within reach in rural or underserved regions of Bangladesh because of cost, logistics, and infrastructure constraints. In such situations, ML-based prediction systems offer a viable alternative. By using features such as dietary patterns, socio-economic status, age, gender, infection history, and geographical location, ML models can predict the likelihood of anemia without the need for instant blood tests. These types of tools can be applied to mobile devices or community health worker tablets, making them more accessible to people.

Effective management of childhood anemia not only involves diagnosis but also timely treatment. Machine learning can assist in monitoring the response to treatment, assessing the effectiveness of nutrition programs, and personalizing care plans based on patient profiles. Such functions are particularly critical in a country like Bangladesh, where delivery of healthcare is still uneven across urban and rural settings, and personalized care has yet to get into full swing.

It is also important to mention the significance of ML strength in public health planning and policymaking. Owing to having access to huge datasets such as the Bangladesh Demographic and Health Survey (BDHS), complex interactions between determining variables of anemia can be detected by using ML models. Such variables could include maternal educational status, food security status, hygiene behavior, and household formation. Such results can be applied by policymakers to design evidence-based intervention programs directed at the most critical risk factors, supporting more strategic utilization of scarce resources.

In general, anemia is a significant public health issue in Bangladesh among children under the age of five. Its pathogenesis is multifactorial and requires a multi departmental approach that includes nutrition, health education, infection prevention, and access to healthcare. Fundamental paradigms of diagnosis and treatment, though primitive, would be significantly enhanced by the application of machine learning technology. By applying data-driven analysis, ML has the potential to revolutionize the detection and treatment of anemia and other diseases. By doing so, it has the potential to improve the children's health outputs, halt mortality, and improve the resilience of the healthcare system of Bangladesh.

Lastly, the use of machine learning techniques for anemia prediction and diagnosis can not only enhance clinical outcomes but also further the general goals of the country's health and development strategies. It facilitates a shift from reactive to proactive, predictive, and preventive care, which is in line with the attainment of universal health coverage and child mortality reduction as per the Sustainable Development Goals (SDGs).

1.2 Motivation

The urgency for the research stems from the long-standing and high prevalence of anemia among Bangladeshi groups, a condition that is still a severe public health concern despite ongoing intervention. Classical statistical techniques have been constrained by inability to identify complex patterns and causality due to small sample sizes and location-based data. With the devastating health, cognitive, and development consequences of childhood anemia, there is an imperative need for more scalable, effective, and data-informed solutions. Motivated by the potential of machine learning approaches to precisely predict anemia and infer its most relevant contributing factors from large hospital-derived datasets of the. Capitalizing on these advanced analysis techniques offers a promising potential of improving early detection, optimizing resource allocation, and guiding policy decisions to avert anemia-associated morbidity and mortality in the Bangladeshi populace.

1.3 Objectives

The aim of this research is to overcome the limitations of existing diagnosis protocols through developing a machine learning anemia prediction model that is quick, accurate, and stable. Alternative methods are needed owing to the paucity of anemia and the limitations of traditional blood cell analysis, which is time-consuming and collection method-dependent. This study's objective is to improve anemia prediction accuracy by deploying sophisticated computer models, such as decision trees, neural networks, and Bayesian probabilistic models. The study aims to enhance early diagnosis, enhance screening techniques, and eventually lead to better and timely treatment plans by examining clinical symptoms and cellular patterns.

1.4 Methodology

This study shall utilize a training and assessment-based supervised learning process. The classification model will be developed using the training set, wherein the algorithm identified the patterns and the relationships. On test data, the trained model will subsequently classify new incidents or make predictions. We will develop classifiers based on the RF, LR, GNB, SVM, KNN, QDA, RC, PA, XGB, and GS approaches, among others. As shown by the flowchart, the proposed solution to anemia prediction follows a structured and systematic strategy. Raw data will first be obtained and preprocessed in order to prepare it for analysis. It entails the conversion of categorical data into numerical form, numerical normalization of features, and missing data management. In order to exclude any biased determination of the models, the data will then be separated into the training and test subsets, generally in a 80:20 ratio. RF, LR, GNB, SVM, KNN, QDA, RC, PA, XGB, and GS are a few of the basic machine learning models that are trained on the prepared dataset. The predictions of basic models are combined by ensemble techniques like bagging, boosting, and voting to improve generalization and overall accuracy. The models will be evaluated using F1-score, AUC-ROC, recall, accuracy, and precision metrics. In order to find the most appropriate model or ensemble technique for predicting mental health, a comparative study will be performed at the end, with focus on how well it can produce strong and stable results.

1.5 Project Outcome

The main goal of the research is to use actual data from a general hospital in Bangladesh to build a consistent and effective machine learning model for early anemia detection in children under five. From the hospital-based dataset, the study was able to discern significant risk factors and patterns that are involved in childhood anemia within the Bangladesh context. The best-performing machine learning model had high accuracy and predictive capability, and therefore it is a valuable tool for use in clinical decision-making and public health planning. This discovery is expected to help healthcare providers in early detection and timely intervention, and consequently improved child health outcomes and reduction in the prevalence of anemia among Bangladesh children.

1.6 Organization of the Report

There are six main chapters in this study. The aim, objectives, methodology, and anticipated outcomes of the project are all described in Chapter 1. Background is described in Chapter 2, together with literature review and applications. Gap analysis comes next to identify the necessity for this study. System design, requirement analysis, context and data flow diagrams, user interface design, and job planning are all described in detail under Chapter 3's research methodology. The implementation, environment configuration, testing procedures, performance evaluation, and result dialogue are all discussed in Chapter 4.

The technical standards being followed, impact on society and the environment, ethical considerations, sustainability strategies, and advanced problem-solving methodologies are all discussed in Chapter 5.

A synopsis of the conclusions, project limitations, and recommendations for future studies are provided in Chapter 6, which concludes the study.

Chapter 2
Background

This chapter discusses the literature review of Anemia. The researchers recommend their study and new techniques to identify Anemia more efficiently.

2.1 Introduction

To improve accuracy and effectiveness, the latest developments in disease diagnosis employed a variety of computational approaches. To forecast and track diseases, some studies have invested in machine learning, neural networks, and Internet of Things-based frameworks. This section provides a background for comprehending current trends and evaluating which areas require more improvement by citing the major contributions, techniques, and results of such related research.

2.2 Literature Review

Anemia remains a serious global public health issue, particularly affecting low- and middle-income countries. Several studies have reported that anemia poses significant health risks in Bangladesh [9–15]. The high prevalence of anemia in children under five years of age is largely attributed to their rapid growth and increased physiological demand for iron. In particular, children from socioeconomically disadvantaged households are more susceptible to iron-deficiency anemia due to poor nutrition, limited access to healthcare, and exposure to disease. However, anemia is not exclusive to children—it can affect individuals across all age groups, including adolescents, adults, and the elderly [16].

Globally, anemia affects approximately 1.62 billion people, making it one of the most widespread nutritional disorders [17]. The burden is especially heavy on children, with around 9.6 million suffering from severe anemia worldwide [18]. In 2017 alone, an estimated 293.1 million children under five were anemic, accounting for a staggering 47.4% of the global under-five population. Of these, a significant 67.6% resided in Africa, underlining the continent's disproportionate share of the problem [19].

In Sub-Saharan Africa, the situation is particularly alarming. Countries such as Tanzania, Kenya, and Mali report anemia prevalence rates among children under five of 79.6% [22], 48.9% [21], and 55.8% [20], respectively. More broadly, regional estimates indicate that between 36.4% and 61.9% of children in this age group in Sub-Saharan Africa are anemic [23]. As of 2021, anemia remained a severe concern in the region, affecting over 83.5 million children and maintaining a high prevalence rate of 67% [24]. Importantly, anemia is not confined to under-resourced nations; it is also an urgent issue in wealthy countries [25], where dietary choices, healthcare disparities, and immigration patterns contribute to its persistence.

Despite the universal nature of the problem, there exist significant disparities in the prevalence of anemia, not only between countries but also within different regions of the same country. According to various studies, anemia rates among children under five range from 12% to 59%, demonstrating a wide variation based on environmental, cultural, dietary, and socioeconomic factors.

In Ethiopia, the situation is reflective of the broader regional context. According to the Ethiopian Demographic and Health Survey (EDHS), approximately 57% of Ethiopian children under five suffer from anemia [26]. This figure underscores the

magnitude of the issue and reveals that despite governmental efforts, much work remains to be done. The Bangladesh government, recognizing the severity of the condition, set a target to reduce childhood anemia from 39% in 2016 to 24% by 2020 [27]. Unfortunately, the findings from various studies suggest that this goal remains unmet, and anemia continues to be a formidable challenge.

The consequences of anemia are profound and far-reaching. Globally, the condition adversely impacts economic and social development [28]. For children under five, the implications are even more dire. Anemia in early childhood can lead to weakened immune function, impaired cognitive and motor development, and in extreme cases, heart failure and increased mortality risk [29, 30]. In many developing countries, including Bangladesh, anemia is a significant contributor to child mortality. These facts call for urgent, multi-sectoral interventions supported by scientific research.

The risk factors for anemia are varied and context-specific. They include but are not limited to nutritional deficiencies, parasitic infections, HIV, malaria, and chronic diseases such as sickle cell disease and blood cancers [24, 31]. In Bangladesh and similar low-income contexts, anemia in children is often exacerbated by maternal ignorance of nutritional needs [32], poor dietary diversity [33], unhealthy feeding practices [34], and lifestyle factors such as reduced physical activity [30]. Additionally, parasitic infections like hookworm contribute to intestinal blood loss, resulting in depleted iron reserves and compromised erythropoietin production, further exacerbating anemia [36]. These parasites not only impair iron absorption but also suppress appetite and hinder nutrient utilization, creating a cycle of malnutrition and anemia [29].

Low socioeconomic status, larger family size, and parental illiteracy are also closely correlated with higher anemia prevalence among children. These socio-demographic variables interact with biological and environmental factors, making anemia a complex and multifactorial issue that resists simple solutions.

Traditionally, researchers have employed regression models to identify the factors associated with childhood anemia in Bangladesh. Cross-sectional studies using logistic regression (both bivariate and multivariate) have been widely applied to examine anemia prevalence and its determinants across different regions of Bangladesh. While these approaches provide important insights, they suffer from inherent limitations. Specifically, cross-sectional analyses cannot establish causality and are often constrained by small datasets, typically limited to particular districts or cities. These datasets often include a limited number of variables, restricting the ability to uncover deeper or hidden patterns.

For the analysis of health data and the prediction of illness risks, ML has become a potent substitute in recent years. Numerous studies have demonstrated how ML models outperform conventional statistical techniques in predicting childhood anemia and determining risk variables [37–40]. The advantage of ML lies in its ability to process large-scale, high-dimensional datasets and discover complex, nonlinear relationships between variables. Moreover, ML techniques are particularly well-suited for predictive tasks, offering opportunities for early detection and intervention.

The integration of ML with large datasets, such as those provided by EDHS, offers policymakers and public health officials the opportunity to derive actionable insights [5, 6]. By employing supervised and unsupervised learning methods, researchers can identify the most influential variables contributing to anemia and develop risk prediction tools that help allocate limited resources more efficiently. This is especially important in resource-constrained settings like rural Bangladesh, where clinical infrastructure is sparse and access to laboratory testing is limited [41].

Furthermore, ML-driven predictive models can significantly reduce the burden on healthcare systems by enabling non-invasive screening and personalized treatment recommendations. The implementation of predictive analytics can optimize healthcare workflows and assist clinicians in making timely and accurate diagnoses,

ultimately improving patient outcomes [6, 42].
In contrast to relying on a single algorithm, the study discussed here employed a comprehensive approach by evaluating multiple ML algorithms to identify the best performer for early-stage anemia prediction. Techniques such as Decision Trees, Random Forest, Support Vector Machines, Gradient Boosting, and Neural Networks were benchmarked using performance metrics such as accuracy, precision, recall, and F1 score. This ensemble-like approach ensures that the selected model is robust, generalizable, and capable of handling real-world data variability.
Notably, this study applied its methodology to a dataset from Bangladesh, another low-income country where anemia remains a significant concern. The insights derived from this application are not only relevant for Bangladesh but also transferable to similar contexts, including Bangladesh. The shared socioeconomic and health challenges between the two nations suggest that successful ML models developed in one setting can be adapted and scaled in the other, enhancing their utility and impact.
In conclusion, while anemia remains a deeply entrenched public health issue globally—and especially in Sub-Saharan Africa—advances in machine learning offer new pathways for understanding, predicting, and ultimately mitigating its effects. By moving beyond the constraints of traditional statistical methods and embracing data-driven approaches, researchers and policymakers can gain a more nuanced understanding of anemia's determinants and develop effective interventions. The integration of ML into public health strategies represents a paradigm shift that has the potential to revolutionize child healthcare in Bangladesh and beyond.

2.2.1 Similar Applications

Big data are used to establish key risk factors and prediction and early diagnosis patterns for the majority of diseases like diabetes, blood pressure, and cardiovascular disease. These are only some of the other similar uses of machine learning in medicine. Machine learning models, for instance, have been used effectively in high-risk groups in predicting the development of diabetes based on factors like age, weight, and diet. For the condition of high blood pressure, predictive models from electronic health records have been utilized to predict the at-risk population to receive timely interventions. Machine learning has been employed in forecasting the heart attack, stroke, and even cancer risk, which has benefited the health experts significantly in making suitable decisions regarding intervention. These studies highlight the ability of machine learning to transform healthcare practice through improved prediction, diagnosis, and management of disease.

2.2.2 Related Research

The reviewed literature highlights unprecedented progress in disease diagnosis through a variety of state-of-the-art technologies, such as machine learning, neural networks, and IoT-based systems. A number of predictive models like Naïve Bayes, decision trees, random forests, Convolutional Neural Networks (CNNs), and K-Nearest Neighbors (KNN) have shown a high level of accuracy in disease prediction and diagnosis from symptoms. Among these, IoT-based frameworks have been found to be very effective in monitoring anemia, typically using probabilistic models and artificial neural networks to aid diagnosis accuracy. Also, the efficiency of neural networks in solving complex medical problems is exemplified, with such systems as genetic algorithms and multi-agent systems with heuristic-based models returning solid and correct results. Also, application of screening technologies, such as the application of machine learning algorithms using flow cytometry, has proved to have promising outputs, particularly when applying the random forest model with more than 95% sensitivity levels. This convergence of emerging technologies not only

increases diagnostic accuracy but also achieves significant cost reductions, thereby reinforcing even more the capacity of intelligent systems to enable early diagnosis, detection, and general disease control. Together, these researches demonstrate how advanced machine learning and IoT-based technologies are transforming healthcare by providing more efficient, effective, and scalable means of disease monitoring and control, highlighting their central position in improving the provision of public health benefits.

2.3 Gap Analysis

Drawing on the comparison of specifications of existing systems, gap analysis highlights their weaknesses that are intended to be addressed in this project. Table 2.1 is the comparative table that has been utilized in detailing differences between designed system and systems already implemented.

Table 2.1 Gap Analysis

Features	[19]	[20]	[21]	[22]	[23]	[24]	**Ours**
Data collection template	No	No	No	No	No	No	Yes
Raw data collection	No	No	No	No	No	No	Yes
Data cleaning	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data pre-processing	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Feature selection	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Scaling	Yes	No	No	No	No	Yes	Yes
Fit model	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ensemble Technique	No	No	No	No	No	Yes	Yes
Evaluation Matrix	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hyper Parameter	No	No	Yes	No	Yes	No	Yes
Comparison with existing	Yes	Yes	Yes	Yes	Yes	Yes	Yes

2.4 Summary

Summary of previous disease detection literature using various modes of computation can be read by reading the related works section. It describes the use of deep models like CNN and KNN and more conventional machine learning methods like Naïve Bayes, decision trees, and random forests in disease prediction from symptoms. It discusses sophisticated methods for anemia detection via infection index models, temporal neural networks, and IoT-based platforms. It also emphasizes the application of multiagent systems for behavior analysis and the use of genetic algorithms and artificial neural networks in the optimization of clinical features.
Overall, the section highlights the way that intelligent systems are increasingly being applied to improve health decision-making, cost, and diagnostic effectiveness.

Chapter 3
Research Methodology

Research methodology of anemia is being discussed in this chapter. The research design, proposed methodology, and data preparation techniques are all being described in this chapter.

3.1 Methodology

The research design, proposed methodology, testing procedure, and vice versa are few of the prime elements which make up the methodology section.

3.1.1 Overview

In this paper, a supervised learning method based on training and testing was utilized. The algorithm extracted trends and correlations from the training data used to construct the model for classification. The learned model was then utilized to make predictions or classify new occurrences in the testing data. Additional information about the machine learning and deep learning model utilized in the study was provided in the subsequent sections.

3.1.2 Used Algorithms

Random Forest (RF): Random Forest is a bagging algorithm that creates a lot of decision trees where each tree is fit on a random subset of data points and features. The final decision is made based on voting the output of the forest by majority or mean. It is the algorithm which is renowned for resisting overfitting, handling missing data quite effectively, and doing well on large and high-dimensional datasets.
Logistic Regression (LR): Logistic Regression is a robust and simple binary classification algorithm. It predicts the likelihood of an outcome being in one of two classes based on input variables using a logistic function that squishes outputs between 0 and 1. Although it has linear relationship assumptions, it's still a good, low-computation model for problems with easily visualizable class separations.
Gaussian Naive Bayes (GNB): GNB is a naive Bayes algorithm that employs Bayes' theorem under the belief that features are independent, and each class is normally distributed. This is simpler in favor of speed and efficacy for certain purposes, especially when used for text classification or instances where features are conditionally independent.
Support Vector Machine (SVM): SVM is a supervised algorithm to learn an optimal hyperplane (border) that classifies data points into different classes. Using kernels, it can also handle complex, non-linear data and is thus widely used for most classification tasks. SVM is particularly stable in high-dimensional feature spaces and is known for being able to identify strong solutions with lesser data points.
K-Nearest Neighbors (KNN): KNN is an intuitive, straightforward algorithm used for classification and regression tasks. It assigns a new point to the most common class among its closest neighbors. Simple and intuitive as it is, KNN could be computationally costly, especially when dealing with large datasets since it must calculate distances for every query.
Quadratic Discriminant Analysis (QDA): QDA is a classifier that represents every class by a distinct Gaussian distribution with the covariance matrix defined for each of them.

Due to this, it can represent more complex decision boundaries than linear classifiers. QDA is a preferable option whenever classes vary by spread or variance of features so that it becomes more suitable to certain classification tasks than simpler models.

Ridge Classifier (RC): Ridge Classifier applies L2 regularization to linear classification, which prevents overfitting by adding a penalty for large feature coefficients. The regularization technique improves model performance for very correlated or large feature sets, hence making it a good choice for high-dimensional data.

Passive Aggressive Classifier (PA): The Perceptron is an extremely ancient neural network algorithm that incrementally learns weights from misclassifications within the training set. It continues to learn until it hits a decision boundary that classifies the data points. Although simple, it is not highly effective at being able to separate complex data relationships from more advanced models.

Extreme Gradient Boosting (XGB): XGB is a powerful machine learning algorithm that is gradient boosting-based, in which a sequence of decision trees are built, each attempting to correct errors of the preceding tree. XGB is efficient and accurate and has been widely applied in industry as well as competitive machine learning settings, particularly for complex classification tasks.

Grid Search (GS): Grid Search is a hyperparameter search technique for finding the best model settings using exhaustive search of a set of hyperparameters. Though it is computationally intensive, it is an exhaustive approach towards improving model performance by utilizing the best setting possible for each machine learning algorithm.

Bagging: Bagging is an ensemble learning method designed to reduce the variance and improve the credibility of prediction models.

Bagging involves the training of multiple copies of a single model, each on a different random subset of the data generated using bootstrapping (sampling with replacement). Final prediction is made by taking the average of the predictions of all the individual models either by majority vote in classification problems or by averaging in regression. Bagging is particularly useful for highly variant models, such as decision trees, in order to prevent overfitting. Boosting: Boosting is a different ensemble method, but instead of reducing variance, it tries to enhance the precision of a model by iteratively training new models.

Each successive model tries to correct the mistakes of the past ones, assigning greater weight to misclassified instances. This sequential approach is used to improve the overall performance of the model by reducing both bias and variance. These boosting algorithms like AdaBoost and XGBoost are widely utilized due to their ability to produce highly accurate models, especially when dealing with complex data.

3.1.3 Proposed Methodology

The approach proposed for anemia prediction is based on a step-by-step and well-defined process that integrates varying machine learning approaches to achieve a practical and robust result. Data collection and preprocessing are the initial phases of the process, which play a crucial role in transforming data into appropriate form to process. Raw data are gathered from Bangladesh general hospital with particular emphasis on the suspected patients for anemia. These data will possess numerical as well as categorical variables among them and some having missing values, outliers, or inconsistency. In them, missing values are dealt with either by imputation or deletion based on missing data amount and type. Numerical attributes have normalization processes in which the data is scaled to ensure all attributes have the same contribution to model training and ensure that one attribute does not dominate due to size. Categorical attributes are translated into numerical forms, normally through encoding functions such as one-hot encoding or label encoding, so that machine learning algorithms can process them cost-effectively. After preprocessing, the data is divided into two sets: the training set and the test set, typically in the proportion of 80:20. It is divided in such a manner that the model is trained on one set of data and then tested on an unseen subset so that overfitting is not

performed and an unbiased measure of the performance of the model is achieved.
Following data preprocessing, various traditional machine learning models are utilized for anemia prediction.
Some of the models used in this study are RF, LR, GNB, SVM, KNN, QDA, RC, PA, XGB, and GS.
These models have varying characteristics and applications that have been utilized in anemia prediction. RF, a robust and powerful ensemble model, is utilized because it can handle both classification and regression tasks with ease and provide robust results even with noisy data. LR, a lightweight but powerful model for binary classification problems, is utilized because it is interpretable and easy to use. GNB, as per Bayes' theorem, is utilized since it can do probabilistic classification very well if the features are normally distributed. SVM is utilized since it does well in high-dimensional space and can handle challenging decision boundaries very efficiently. KNN is used because it's simple and effective on non-linear classification tasks, where the instance's class is estimated by the majority of its neighbors' classes. QDA is used because it can work with data whose classes have different covariance structures, a more general solution than LDA. RC is used for regularized linear classification, providing a solution to avoid overfitting when dealing with high-dimensional data. PA models, as online learners, are employed because of their efficiency in handling large sets of data and rapid learning on new data. XGB, as a gradient boosted algorithm, is employed because of its efficiency and scalability with the application of large and complicated sets of data. GS is also utilized for tuning the hyperparameters of the models in an attempt to use the optimal set of parameters in model performance optimization. Once the individual models have been trained, improving their prediction accuracy using ensemble methods is the following step in the methodology.
Ensemble methods are stable methods that learn by aggregating the predictions of multiple models to generate better generalization and reduced overfitting risk.
Bagging (Bootstrap Aggregating), a method that trains multiple models on various random subsets of data and then combines their predictions, is used to fight variance and make the model more stable. Bagging is particularly useful for models like decision trees, where variance is high. Boosting is a technique that trains models sequentially where each new model attempts to learn from errors of earlier models. This is designed to reduce bias and increase overall accuracy of the ensemble. The boosting algorithms like AdaBoost and XGBoost are primarily useful in cases of high-complexity data and possess a strong potential for improving weak model performance. Another group approach used in this methodology is the voting method where multiple base models predict and the ultimate prediction is given in the form of majority vote or mean of their predictions. It merges the strengths of more than one model, so the output is stronger and less prone to errors. These ensemble techniques play a crucial role in improving the performance of anemia prediction models to make them reliable and accurate. In order to evaluate the performance of the individual models and ensemble techniques, various metrics are employed. These are F1-score, a weighted average of precision and recall and of special interest in the case of imbalanced data, as in the case of anemia prediction, where the positive cases (anemic patients) may be less than the negative ones.
The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is another important measure that is used to quantify the discriminative ability of the model between the positive and negative classes, and it provides information about the performance of the model at different classification thresholds.
Note that because measurement of the model's ability to pick up true positive instances (anemic patients), is even more important in the context of medical diagnosis due to the fact that failing to pick up an anemic patient would have serious health consequences. Accuracy, which measures the overall prediction accuracy of the model, is also computed, although it will not always be the best measure to use when dealing with unbalanced datasets. Precision, the proportion of positive model predictions that are true, is also considered, specifically to avoid false positives, which might lead to unnecessary

treatment or intervention. Finally, the entire comparative analysis is conducted to find the best-performing model or ensemble method for anemia prediction. Not only are the performances of all models evaluated on the aforementioned criteria, but also the computational cost and interpretability of the model are considered, which are of high importance in the clinical setting where healthcare professionals should be capable of understanding and relying on the predictions made by the model. Comparative evaluation adds credibility to the selection of the top-performing model or ensemble method, which is then evaluated for implementation in a live clinical setting for anemia prediction and diagnosis.

The ultimate objective of the strategy is to develop a reproducible, efficacious, and precise predictor for anemia prediction to support prompt and evidence-based decision-making on patient care among healthcare practitioners. The research process exhibits the usefulness of ensemble methodology and machine learning methods to re-engineer the field of health diagnosis, particularly in poor-resource settings, due to increased precision and efficacy in disease predictive models. Figure 3.1 is a representation of the flow diagram.

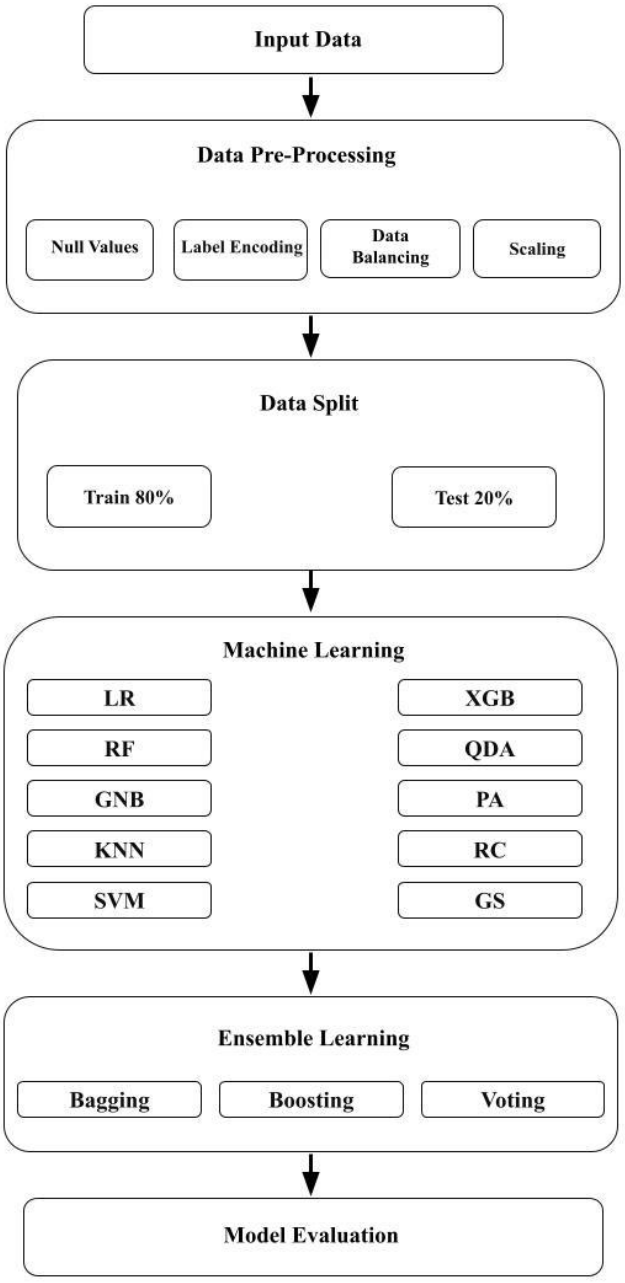

Figure 3.1: Methodology of Anemia Prediction

3.1.4 Data Collection

In order to investigate the many factors impacting anemia, the information utilized in this study was gathered from general hospitals. To make sure it addressed pertinent anemia issues and concerns unique to the participant demographic, the questionnaire was meticulously created in cooperation with a general hospital anemia specialist physician. To make sure that the anemic features of the patients were appropriately recorded and categorized, the data was also clustered under the doctor's supervision. Using the clustering technique, the data was labeled and grouped into meaningful groups. By dividing the patients into discrete groups according to their answers, this unsupervised learning method made it possible to better comprehend the patterns and trends associated with anemia and its effects on health.

3.1.5 Data Cleaning

The general hospital provided the dataset. This means that the dataset must be clean, devoid of anomalous values and situations. We have eliminated any duplicate entries from the dataset after determining whether they exist. The dataset was then examined for negative values, and the category values were converted to numerical representation using label encoding. then balanced the unbalanced data using the ADASYN approach. To make the model's assumptions easier, the values were scaled using the standard scalar approach.

3.1.6 Analysis Techniques

The dataset will be divided into two parts for this study: training and testing. An 80:20 ratio will be used to divide the dataset. Since the shortest classification time is acceptable, both the compilation time and the training accuracy will be measured. The actual values from the testing dataset will be compared to the values predicted by the model. Accuracy, precision, recall, F-1 score, sensitivity, AUC-ROC score, curve, and compilation time will all be used in processing the assessment.

3.2 Detailed Methodology and Design

We looked at a number of different approaches to get around the difficulties we encountered during the study. We first considered utilizing standardized, previously validated questionnaires to generate the questionnaire rather than creating it from scratch, but we decided to modify research papers and hospital test results to make sure the questions were particularly pertinent to our goals. We thought about buying datasets from open-access medical databases to address the challenge of gathering authentic data from hospitals, but we placed more emphasis on establishing daily goals for ongoing actual data collection to guarantee authenticity and relevance to our research context. We thought about employing form-based inputs to automate the laborious process of manually entering data into Excel files, but ultimately opted to use manual entry along with a rigorous data cleaning procedure to ensure correctness and eliminate missing entries. The practicality, affordability, and capacity to maintain the accuracy and dependability of our study data were the main criteria used to choose each option.

3.3 Project Plan

The project plan is shown in Table 3.1 below.

Table 3.1: Project Plan

Tasks	Weeks																	
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Data Collection																		
Pre-process																		
Development																		
Documentation																		

Estimated Work Period	
Actual Work Period	

3.4 Summary

This study used machine learning techniques to forecast anemia disorders in a systematic manner. To improve model performance, a sick and healthy dataset was first gathered and cleaned using data cleaning and missing value treatment. Besides ensemble techniques like bagging and boosting, some machine learning algorithms like RF, LR, GNB, SVM, KNN, QDA, RC, PA, XGB, and GS were utilized. The performance metrics such as precision, accuracy, recall, and F1-score were used to train and test the models. Cross-validation and hyperparameter tuning were used to improve the performance. In trying to choose the best model in high-precision early sickness prediction of anemia disease, the results were finally analyzed.

Chapter 4
Implementation and Results

The results of anemia disease prediction are being discussed in this chapter.
The evaluation methods, algorithm comparison methods, resource usage, and evaluation matrix are all given in this chapter.

4.1 Environment Setup

Several tools and libraries were utilized in this research to make data processing and visualization easier. With the Python interpreter, Conda package management, and Anaconda Navigator tools, Anaconda was the major platform for package and environment management. with its cell and kernel structure, Jupyter Notebook offered an interactive environment for coding and running code, marking up results in markdown, and displaying results. With its mathematical functions and n-dimensional arrays, NumPy was used for fast numerical computation. Pandas supported a variety of data manipulation tasks, including cleaning, merging, and reshaping datasets, and provided structured data management using DataFrames and Series. While Seaborn, which is based on Matplotlib, made it easier to create nice-looking statistical plots with high-level customisation capabilities, Matplotlib provided a range of plotting capabilities for visualisation. Combined, these packages created a powerful data exploration, analysis, and scientific computing pipeline.

4.2 Comparative Analysis

Accuracy: Accuracy is the most widely used metric and the overall correctness of the model. It measures the proportion of correct predictions made to the number of predictions both correct positives and correct negatives. Accuracy is not difficult to understand or intuitive and yet it's a misleading metric in which data is unbalanced. For instance, if 95% of the cases come from one class, a model can be 95% accurate by predicting the majority class on every occasion, even though it will fail to pick up any of the minority class cases. Therefore, while high accuracy is usually to be preferred, it is not invariably the best statistic to be employed, especially in medical or fraud detection contexts where it is important to forecast the minority class.
Precision: Precision is a measure of how many of the examples that the model labeled as positive were actually correct.
That is, it tells us the proportion of true positives out of all positive predictions. Precision is particularly useful in situations where false positives are costly. For instance, in medical diagnosis, diagnosing a healthy patient with anemia might lead to unnecessary stress, further testing, and even inappropriate medication. High precision indicates low false positive rate and implies that the model is conservative in terms of making positive predictions but does so accurately. Recall: Recall, also known as sensitivity or true positive rate, estimates to what extent the model identifies all the cases that are applicable in a dataset.
Recall: It computes the proportion of actual positive instances that were accurately predicted. High recall is critical in situations where failure to detect a positive case could have serious consequences. In healthcare, for instance, when a model fails to detect anemia patients (false negatives), it may delay diagnosis and treatment, hence leading to

worsening health conditions. Therefore, recall is a critical measure when the cost of false negatives is high. F-1 Score: The harmonic mean of recall and accuracy is the F1-score. It balances the false positive and false negative trade-off by combining the two metrics into a single score.

F-1 Score: A good model by all standards is reflected by a high F1-score, which indicates that accuracy and recall are both very good. This is particularly handy in situations where the classes in the dataset are imbalanced and in balancing the cost of eliciting false alarms with the danger of failing to detect positive instances.

In application areas like medicine, where the precise diagnosis and extensive coverage of all positive instances are crucial, the F1-score is widely utilized.

Confusion Matrix: A graphical summary of the outcome of predictions is offered by the confusion matrix. True positives, or positively predicted correctly; false positives, or wrongly predicted positives; true negatives, or negatively predicted correctly; and false negatives, or negative instances failed, are the four categories into which it classifies the performance. It provides a comprehensive picture of the model's problems apart from helping to compute the above measures.

For instance, a model with a large number of false negatives may require optimization for recall, while a model with a large number of false positives may require better precision.

The confusion matrix is typically employed by analysts to manually assess and adjust model behavior to better meet domain-specific needs.

Table 4.1: Traditional Algorithm results

Algorithm	Accuracy	Precision	Recall	F-1 Score
LR	85.87	88.43	85.87	86.02
RF	90.83	91.95	90.83	91.08
SVM	90.07	90.66	90.07	90.19
KNN	**91.98**	**92.09**	**91.98**	**92.02**
GNB	57.25	47.71	57.25	50.09
QDA	89.31	89.87	89.31	88.99
RC	70.99	73.58	70.99	66.52
PA	88.54	89.76	88.54	88.70
GS	91.60	92.35	91.60	91.70
XGB	86.25	87.17	86.25	86.59

Table 4.1 shows the comparison of the performance of some of the traditional machine learning models implemented in anemia disease prediction. The models were tested on the basis of four important parameters: Accuracy, Precision, Recall, and F1 Score. Each measure gives different types of information regarding the performance of the model to recognize the patients correctly, particularly in a sensitive area like health where both the false negatives and false positives are crucial.

Of all the models tried, K-Nearest Neighbors (KNN) performed best in most of the metrics. With an accuracy of 91.98%, precision of 92.09%, recall of 91.98%, and F1 score of 92.02%, KNN has a balanced and stable capacity to identify anemia cases correctly. Its high recall indicates that it captures most of the true anemia cases, and its high precision indicates that when it is predicting anemia, it is most likely to be correct. The almost equal values of all the metrics also indicate the stability and reliability of the model in predicting positive and negative cases without any bias.

Is immediately followed by KNN is Grid Search (GS) optimized model with accuracy 91.60%, precision 92.35%, recall 91.60%, and F1 score 91.70%. Grid Search is not a model but rather a hyperparameter optimization algorithm applied with base learners in order to maximize their performance. The scores were sufficiently high to suggest that hyperparameter tuning contributed significantly towards achieving improved prediction outcomes, therefore, model optimization being equally important as model selection.

Random Forest (RF) and Support Vector Machine (SVM) also performed very well, with RF scoring 90.83% accuracy and F1 of 91.08%, and SVM scoring 90.07% accuracy and F1 of 90.19%. These results continue to support the suitability of ensemble models like Random Forest and margin-based classifiers like SVM for medical diagnosis tasks. Random Forest is particularly lucky to possess an ensemble model, in which multiple decision trees are aggregated to prevent overfitting and variance reduction, while SVM excels in handling high-dimensional space and performs well even for small samples.

The Passive Aggressive (PA) algorithm also did well with accuracy of 88.54%, precision of 89.76%, and F1 of 88.70%. PA is best at large learning problems and works optimally in online learning environments. That it did well here suggests that it is potentially feasible in real-time or streaming data conditions in healthcare settings.

Quadratic Discriminant Analysis (QDA), which is also a probabilistic classifier, also did well with 89.31% accuracy and 88.99% F1 score. QDA assumes that features are Gaussian distributed and have their own covariance matrix for each class, which may be the reason why it is able to learn complex relationships in the data.

On the contrary, Gaussian Naïve Bayes (GNB) was the worst among all models with a significantly lower accuracy of 57.25%, precision of 47.71%, and F1 score of 50.09%. The poor performance can be attributed to the model's feature independence assumption, which is seldom true in real-world health data. The result suggests that despite the fact that Naïve Bayes is computationally cheap, it may not be the ideal model for the prediction of anemia where feature interactions are likely to be important.

Logistic Regression (LR) and Ridge Classifier (RC) performed medium. LR performed well with 85.87% accuracy and 86.02% F1 score, which means it remains a good linear classifier for binary classification tasks. RC, on the other hand, lagged behind with 70.99% accuracy and a significantly low F1 score of 66.52%, which may be due to its high regularization that may have over-limited the model's flexibility to fit the data.

XGBoost (XGB), a top performing boosting algorithm, was slightly higher than LR at 86.25% accuracy and 86.59% F1 score. XGB is famous for its management of missing values, scalability, and speed in relation to structured data and is thus a workhorse in prediction modeling. While not the best here, it is good and may be further improved with hyperparameter tuning or feature engineering.

From the comparison, it can be seen that model selection highly dictates the performance

of anemia prediction systems. KNN, GS, RF, and SVM were the top-performing models with high consistency in all the evaluation metrics. This means they have high potential in the detection of anemia with little error, and thus they are appropriate to be incorporated into medical diagnosis systems. Conversely, GNB and RC models founded on more robust statistical assumptions or over-regularization performed below optimally and perhaps are not ideally appropriate for such application unless additionally tailored. These findings emphasize the importance of comparative model evaluation and the use of more than one figure of merit in addition to accuracy. Precision and recall, in particular, are of utmost importance in clinical environments where the cost of an erroneous prediction is great. In addition, the use of hyperparameter optimization via Grid Search indicates that baseline models can achieve greater performances when properly tuned.

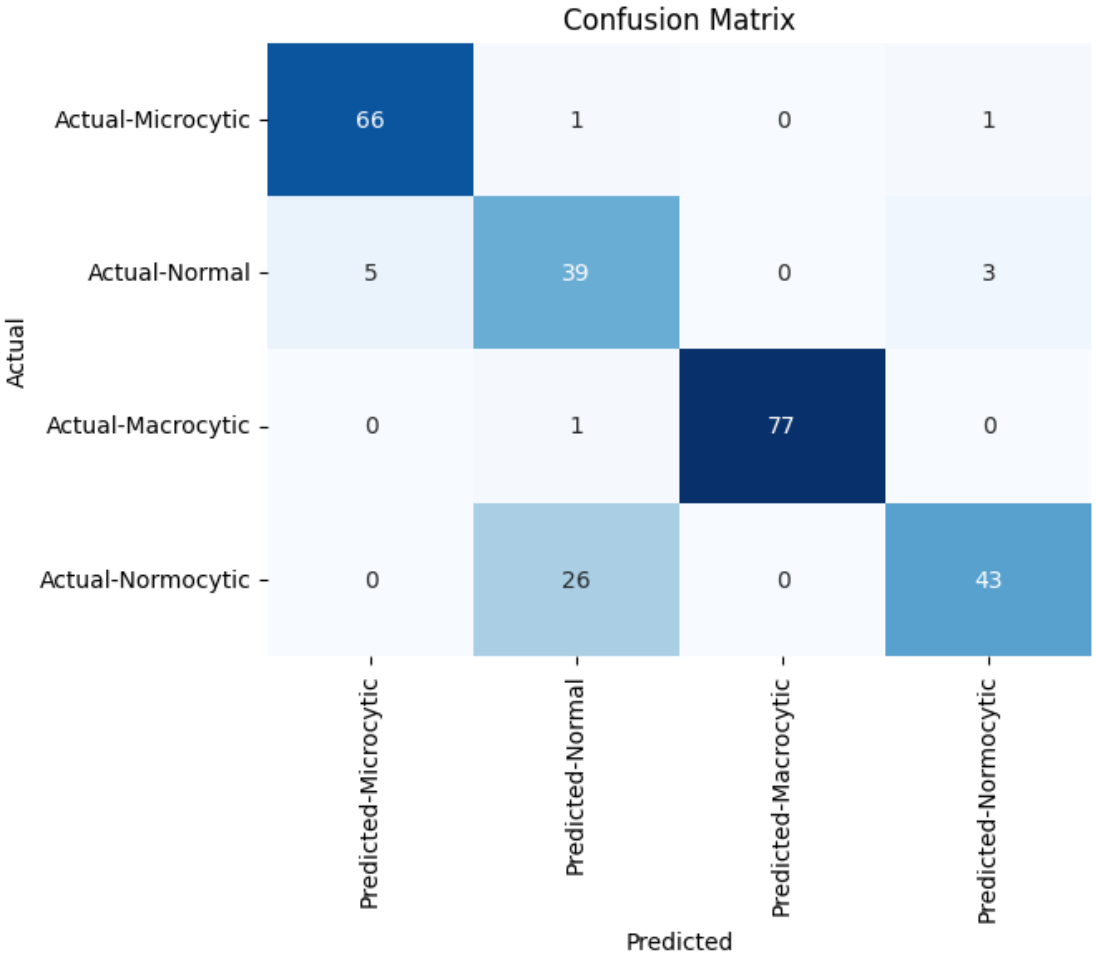

Figure 4.1: Confusion matrix of Logistic Regression

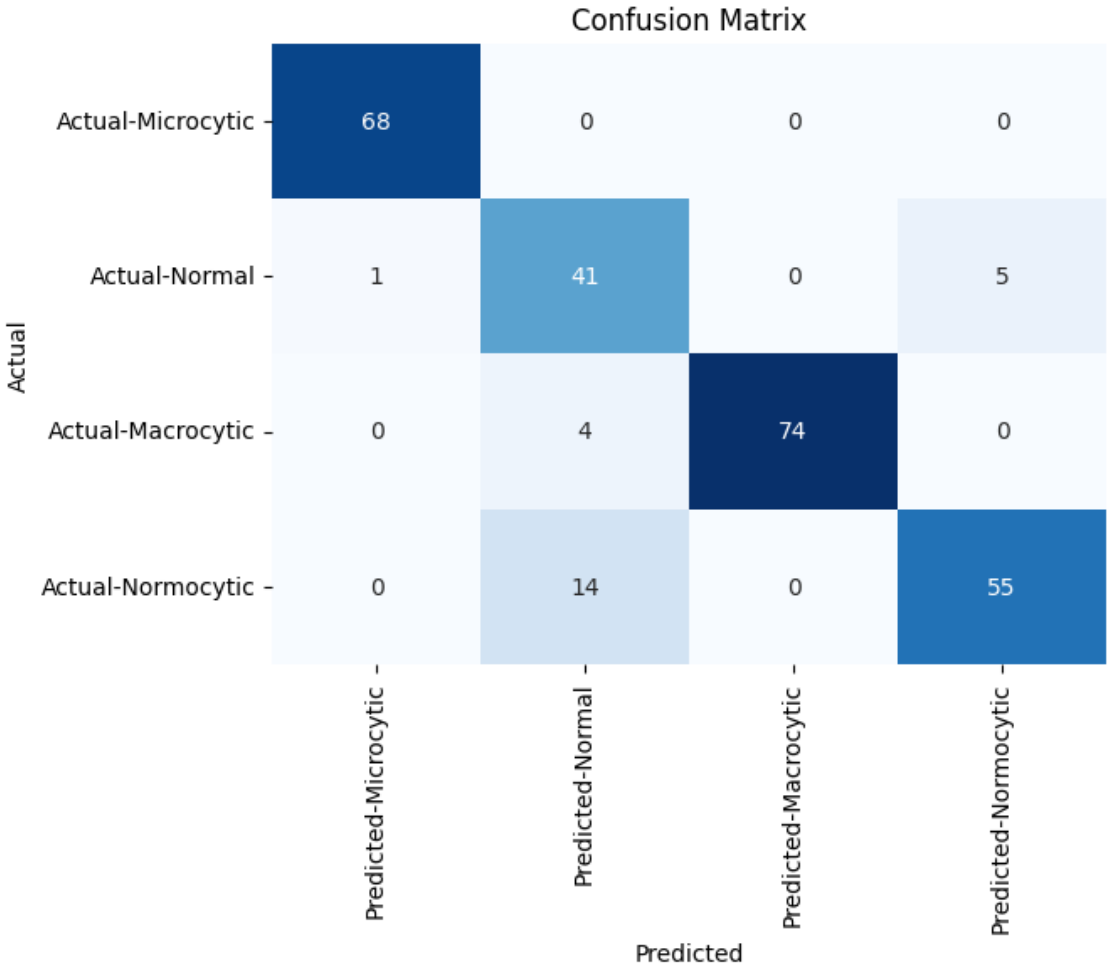

Figure 4.2: Confusion matrix of Random Forest

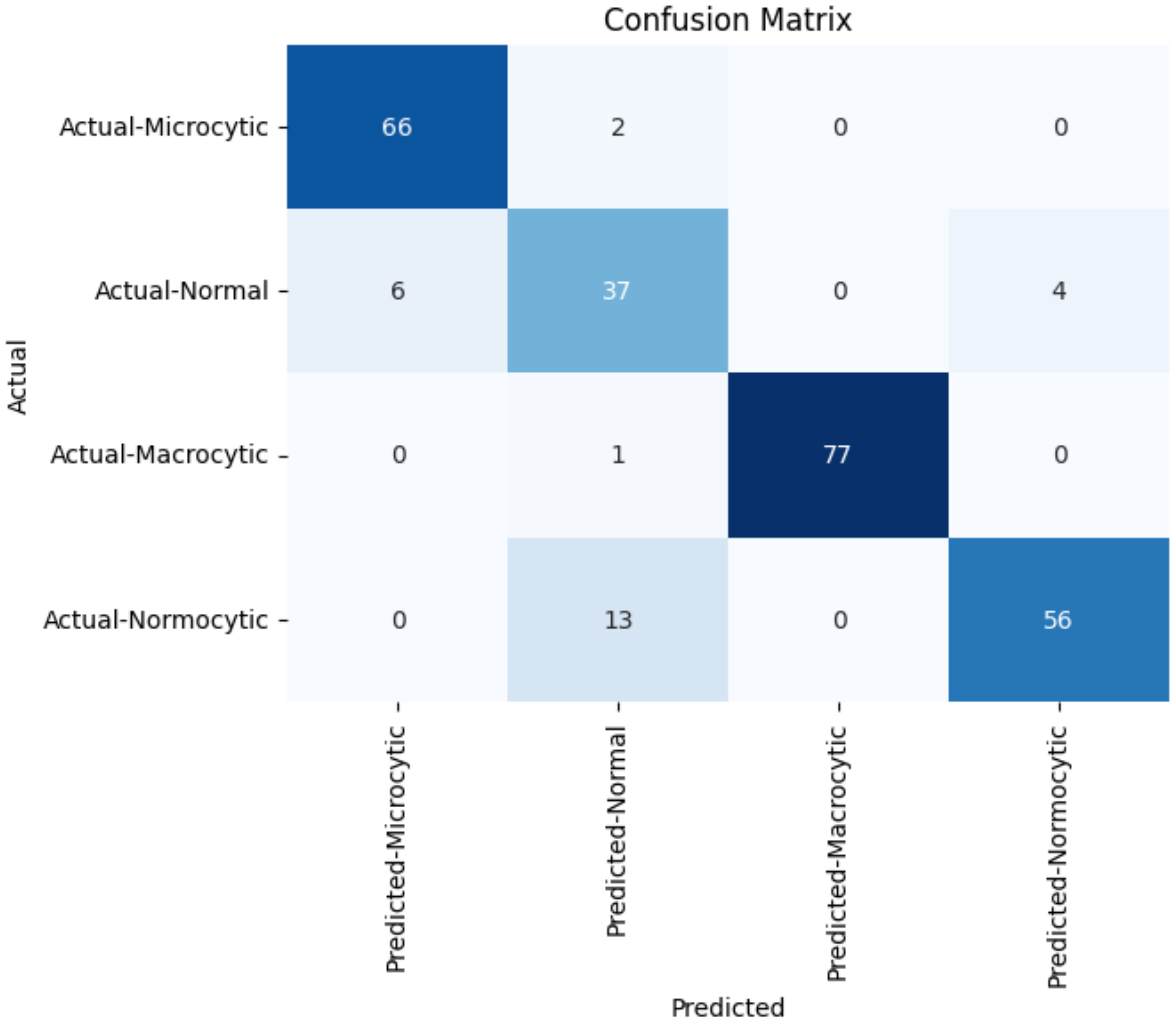

Figure 4.3: Confusion matrix of Support Vector Machine

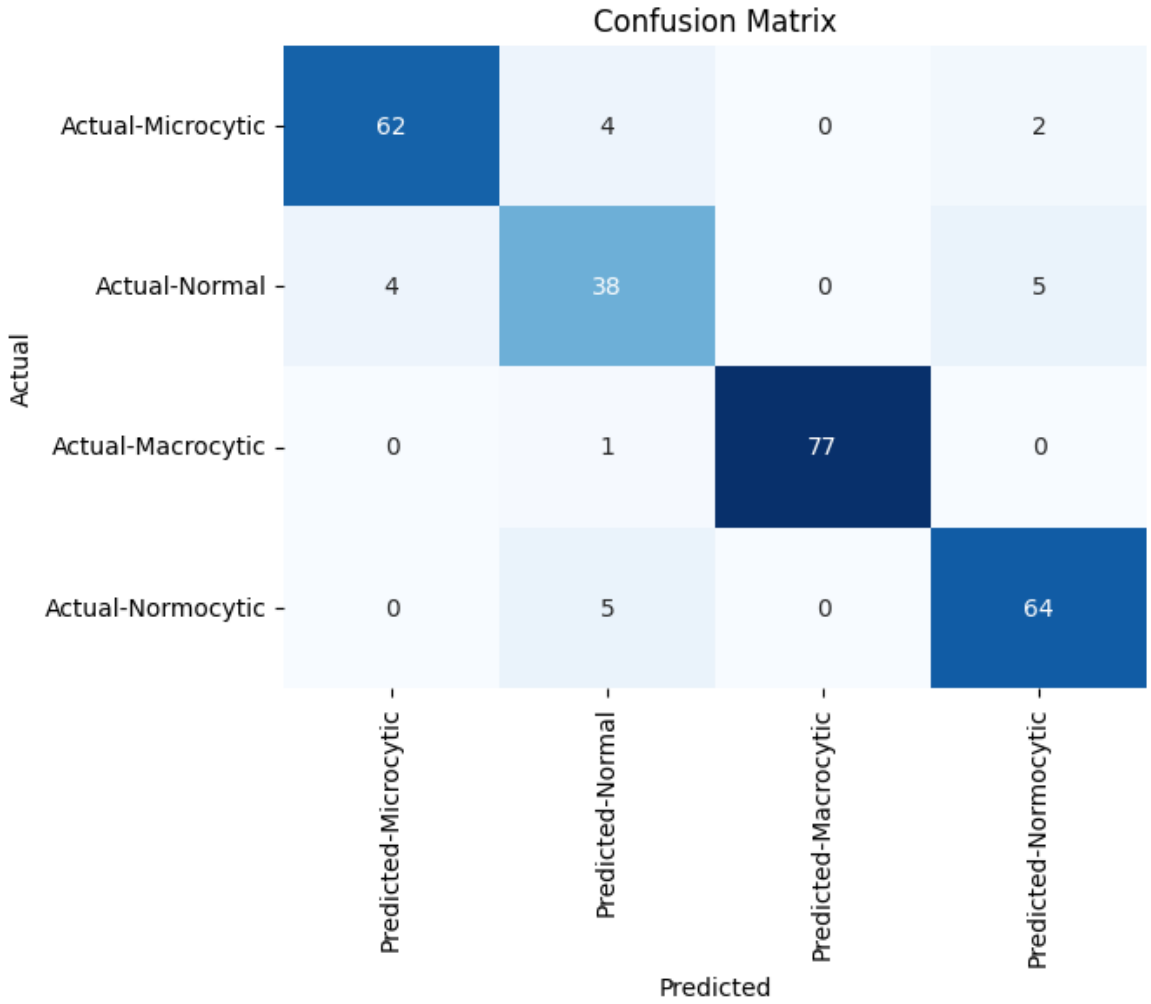

Figure 4.4: Confusion matrix of K-Nearest Neighbors

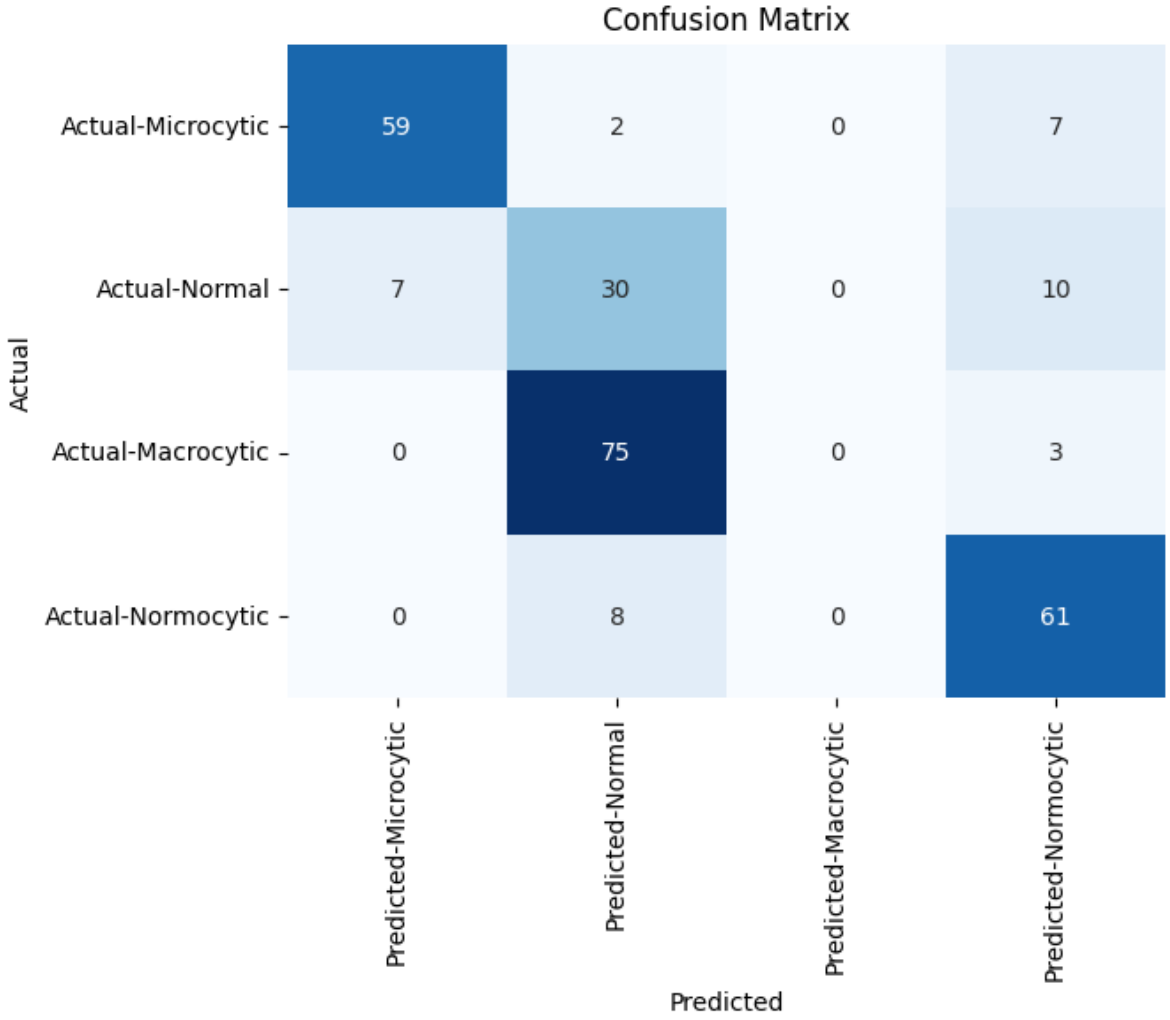

Figure 4.5: Confusion matrix of Gaussian Naïve Bayes

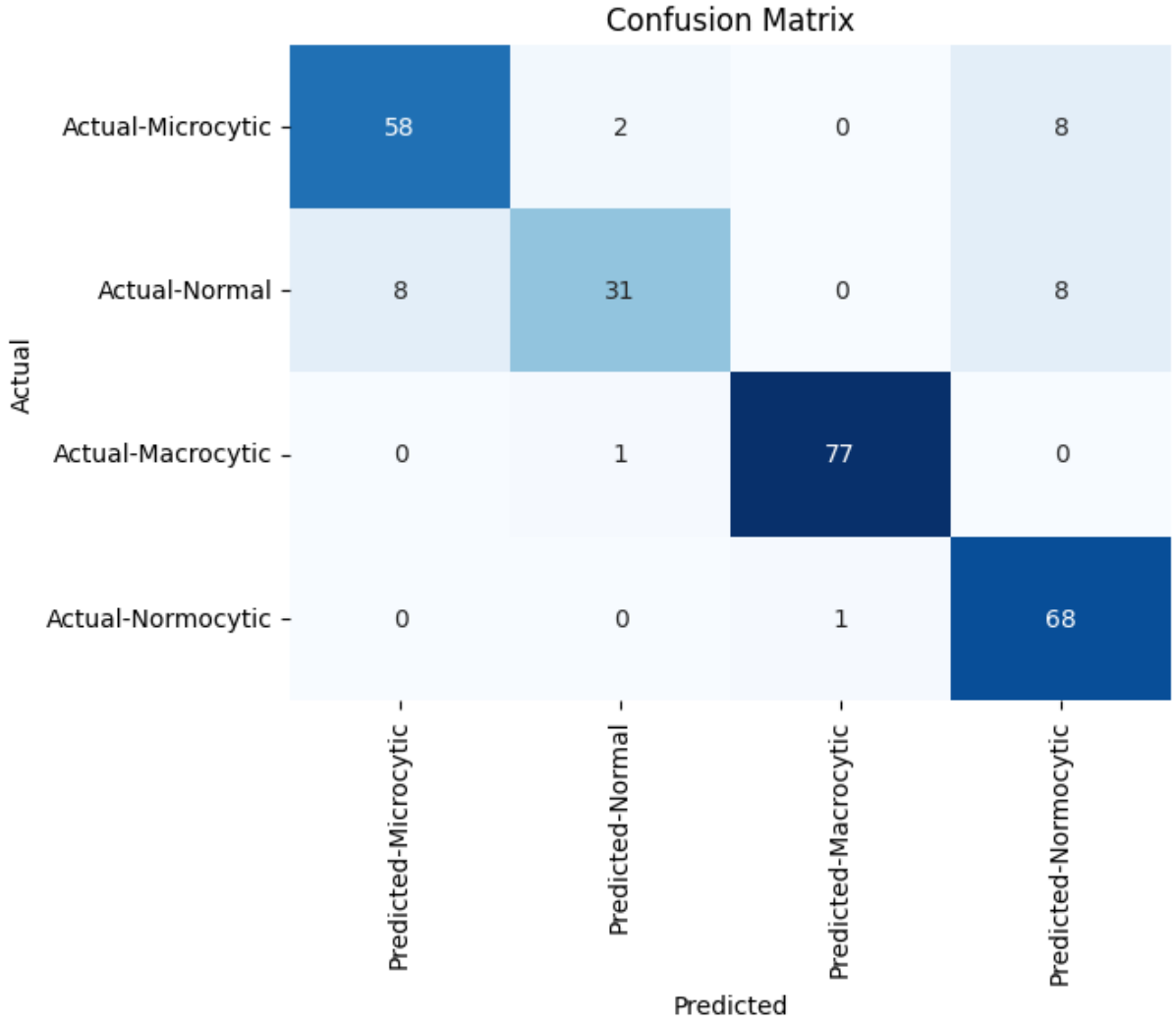

Figure 4.6: Confusion matrix of Quadratic Discriminant Analysis

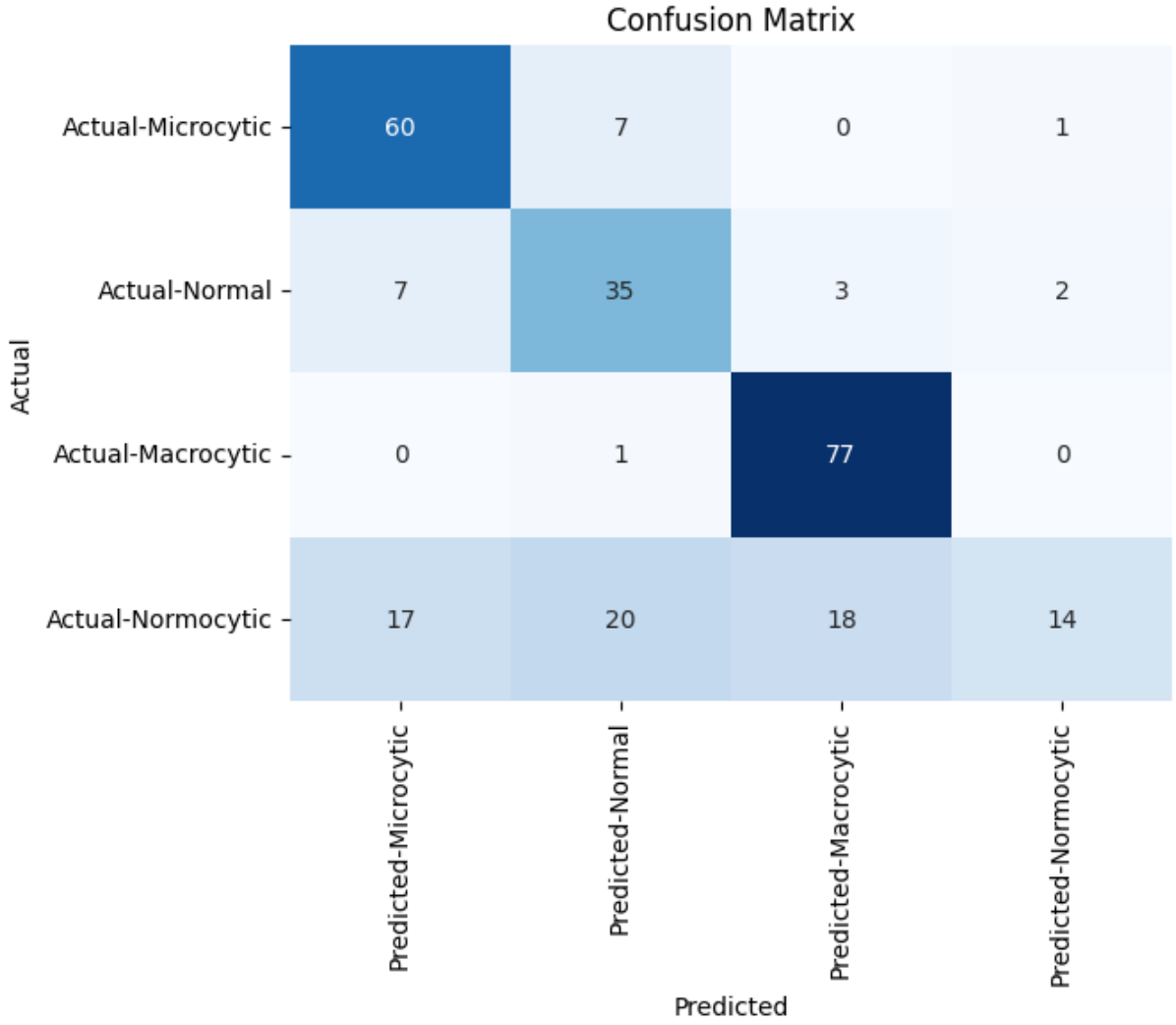

Figure 4.7: Confusion matrix of Ridge Classifier

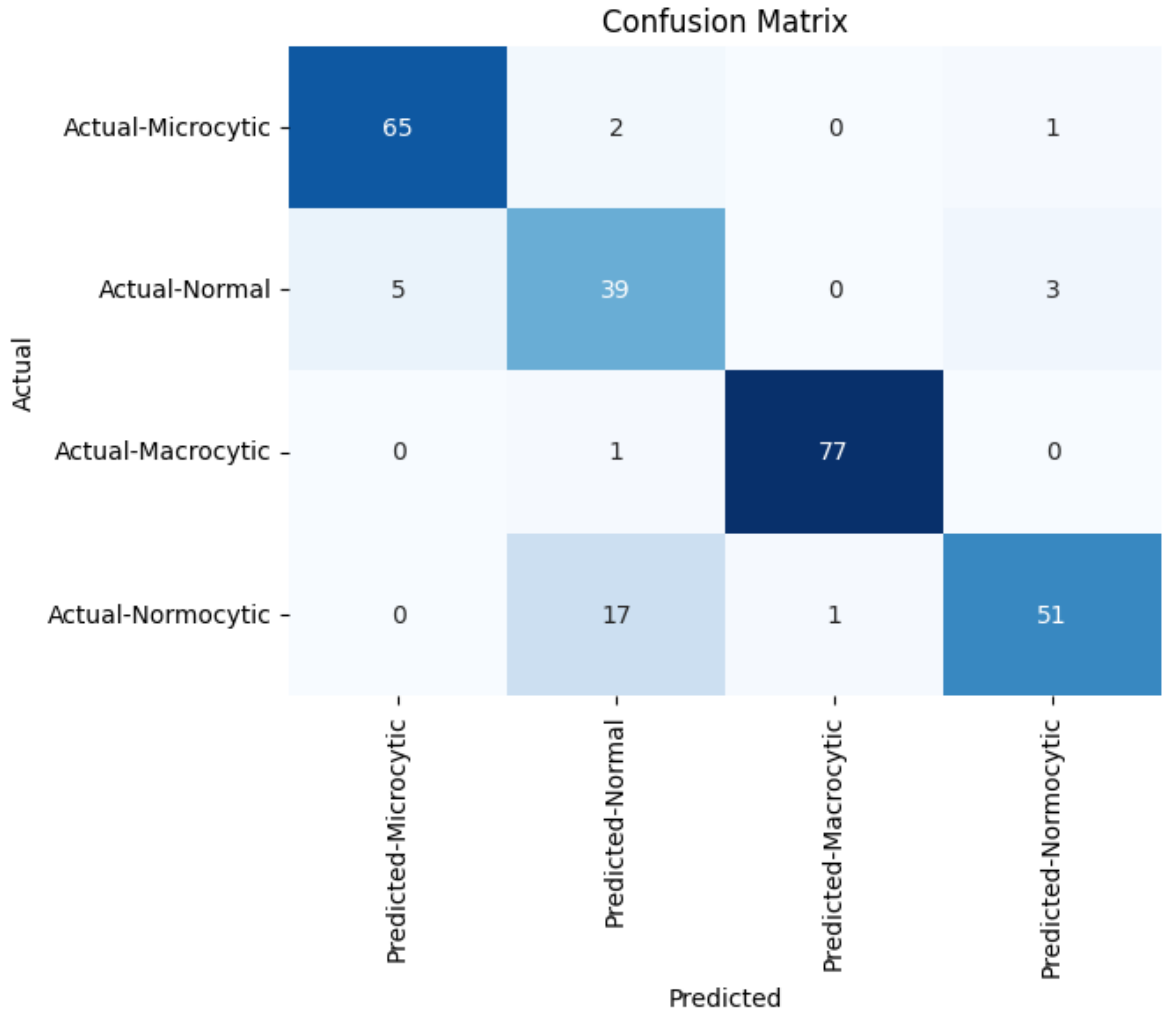

Figure 4.8: Confusion matrix of Passive Aggressive Classifier

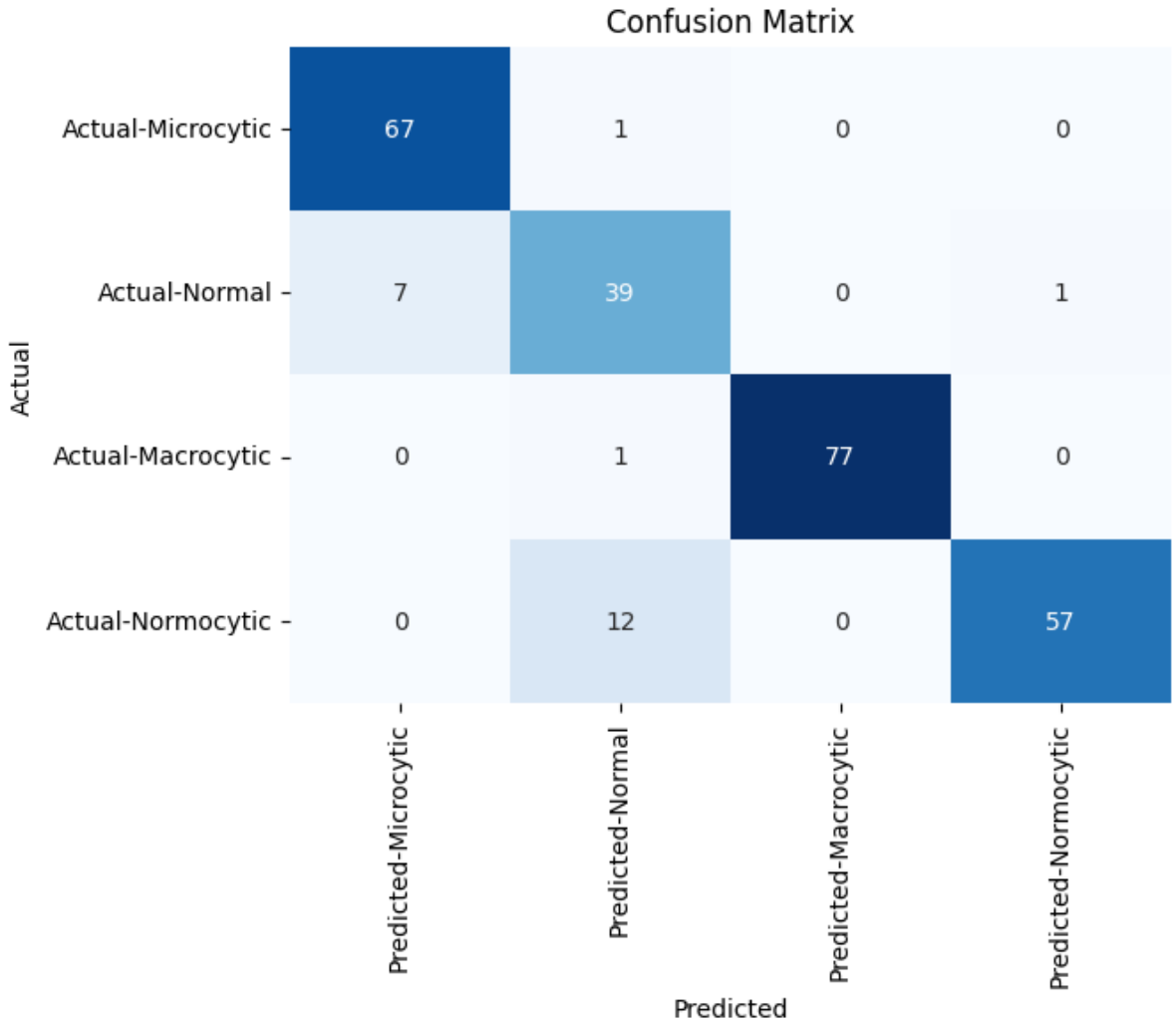

Figure 4.9: Confusion matrix of Passive Aggressive Classifier

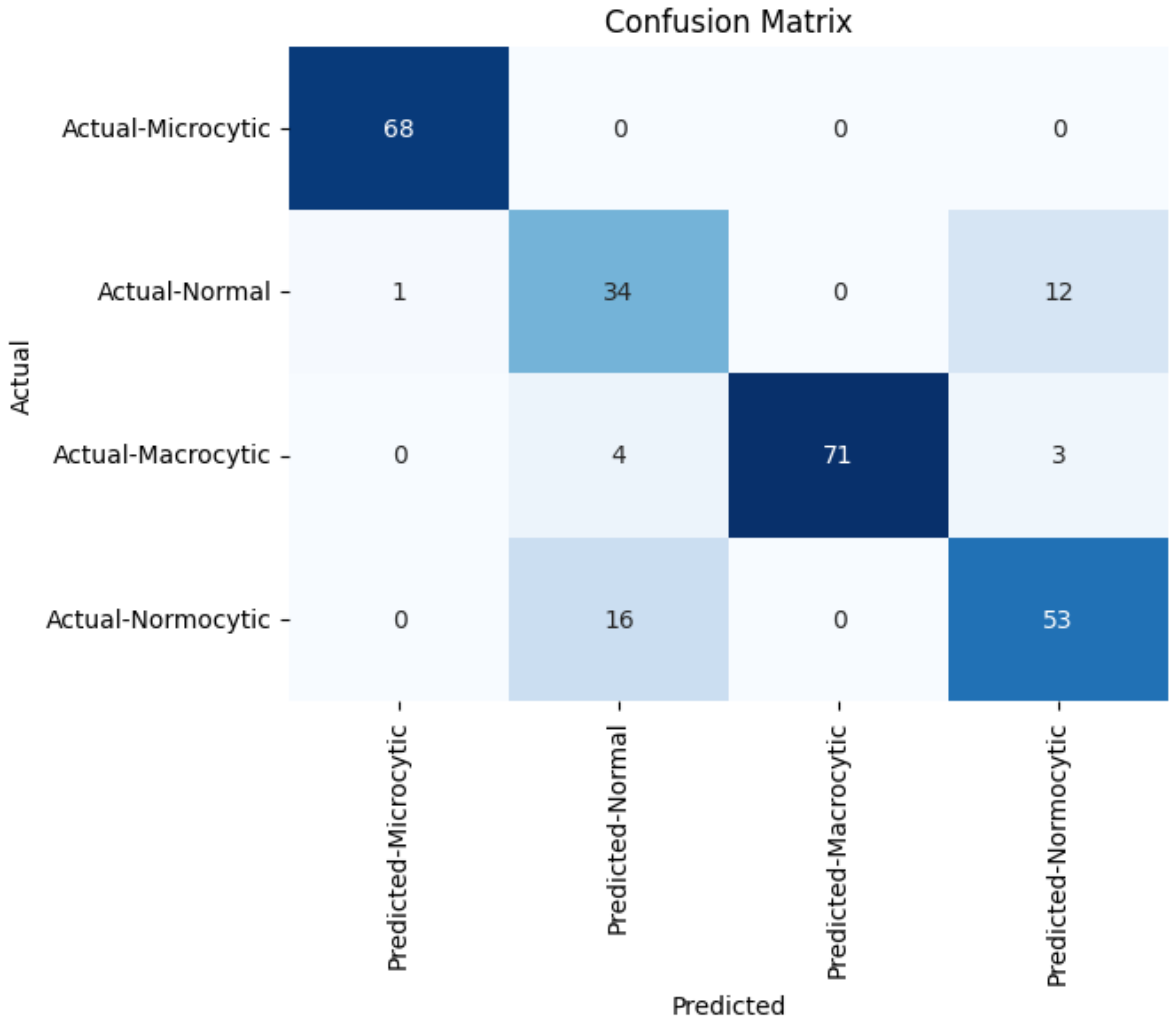

Figure 4.10: Confusion matrix of XGBClassifier

LR, RF, SVM, KNN, GNB, QDA, RC, PA, GS, and XGB confusion matrix are displayed in Figures 4.1, 4.2, 4.3, and 4.6, respectively.

The performance results of different traditional machine learning algorithms implemented in an ensemble bagging framework for anemia disease prediction are displayed in the following table 4.2. The metrics taken into consideration are Accuracy, Precision, Recall, and F1 Score, which all provide a comprehensive idea of the predictive capability of every model. The models were then ensembled with a bagging ensemble method, which minimizes variance and overfitting by combining predictions of multiple base learners trained on varying subsets of the training data.

Table 4.2: Ensemble Bagging Algorithm results

Algorithm	Accuracy	Precision	Recall	F-1 Score
LR	85.87	88.43	85.87	86.02
RF	91.60	92.35	91.60	91.70
SVM	91.22	92.05	91.22	91.38
KNN	90.83	91.11	90.83	90.95
GNB	55.34	44.71	55.34	48.55
QDA	90.07	90.33	90.07	89.96
RC	70.99	72.89	70.99	66.68
PA	87.02	89.17	87.02	87.17
GS	**91.98**	**92.65**	**91.98**	**92.08**
XGB	88.54	89.28	88.54	88.70

Grid Search (GS), which is used here to hyperparameter tune the base learners in the bagging ensemble, was the best performing among all the models. It achieved 91.98% accuracy, 92.65% precision, 91.98% recall, and an F1 score of 92.08%. These results confirm that the base models, when fine-tuned with meticulous hyperparameter tuning and ensembled under bagging, produce extremely stable and precise predictions. That the F1 score is extremely high, a trade-off between precision and recall, indicates that the model is not just correctly predicting anemia cases but also not susceptible to false positives.

Next in sequence, Random Forest (RF), a bagging-based ensemble of decision trees, also fared well with 91.60% accuracy, 92.35% precision, and an F1 score of 91.70%.

RF performance within a bagging context is expected to be good due to its design itself. By averaging predictions of many decision trees trained on bootstrapped samples, it reduces the likelihood of overfitting and offers good generalization—critical strengths when working with medical datasets where data quality and distribution can be poor.

Support Vector Machine (SVM) also did well when used within the bagging ensemble with 91.22% accuracy, 92.05% precision, and 91.38% F1 score. SVM, being a margin-based classifier, is assisted by bagging through the variance reduction, particularly when the model is susceptible to training data fluctuations.

K-Nearest Neighbors (KNN) also performed well at 90.83% accuracy, 91.11% precision, and 90.95% F1 score. KNN is usually troubled by high variance, especially with noisy

data, but bagging appears to have stabilized its performance very well. The proximity-based nature of KNN combined with bagging leads to smoothed-out predictions over varied test cases.

Quadratic Discriminant Analysis (QDA) and XGBoost (XGB) also fared quite well. QDA achieved 90.07% accuracy and F1 score of 89.96%, showing even statistically inclined models can be enhanced by ensemble techniques. XGBoost, although a boosting algorithm by design, did better in a bagging setup as well—achieving 88.54% accuracy and 88.70% F1 score—a indication that maybe a hybrid ensemble was implemented to manage variance and bias effectively.

Logistic Regression (LR) also did quite well with an accuracy of 85.87% and an F1 score of 86.02%, showing its utility in binary classification issues even when utilized in a bagging ensemble. Since it is linear in nature, its power grows when executed on different subsets of data and aggregated.

Passive Aggressive (PA) classifier, which can be scaled up to online learning on large scale, gave 87.02% accuracy and 87.17% F1 score. While updating weights aggressively and operating incrementally, bagging regularizes the typically oscillating decision boundaries of the latter to produce more stabilized predictions.

Ridge Classifier (RC) and Gaussian Naïve Bayes (GNB), however, performed relatively worse than the remainder.

RC performed with 70.99% accuracy and 66.68% F1 score, while GNB was the worst performing of all with 55.34% accuracy, 44.71% precision, and a low F1 score of 48.55%. This is likely because these models make strict assumptions. RC's L2 regularization may penalize complex patterns too heavily, and GNB's feature independence assumption does not typically hold for medical data when variables are intertwined. In spite of bagging, these problems were not alleviated entirely. This comparative analysis reveals the significant benefits of using a bagging ensemble approach to the prediction of anemia. By aggregating multiple base learners trained on subsets of resampled data, bagging enhances model stability, avoids overfitting, and has a tendency to enhance general predictive precision. Random Forest, SVM, and KNN classifiers—especially when optimized through Grid Search—are shown to have excellent performance under this setting.

The ensemble approach appears to be particularly suitable for medical tasks like anemia detection, where the goal is not just to obtain top accuracy but also to have high recall (to identify most of the actual anemia cases) and high precision (to not have false alarms). The fact that the top-performing models have high F1 scores shows that they are capable of trading off these concerns nicely, and thus are viable for incorporation into clinical decision-support systems.

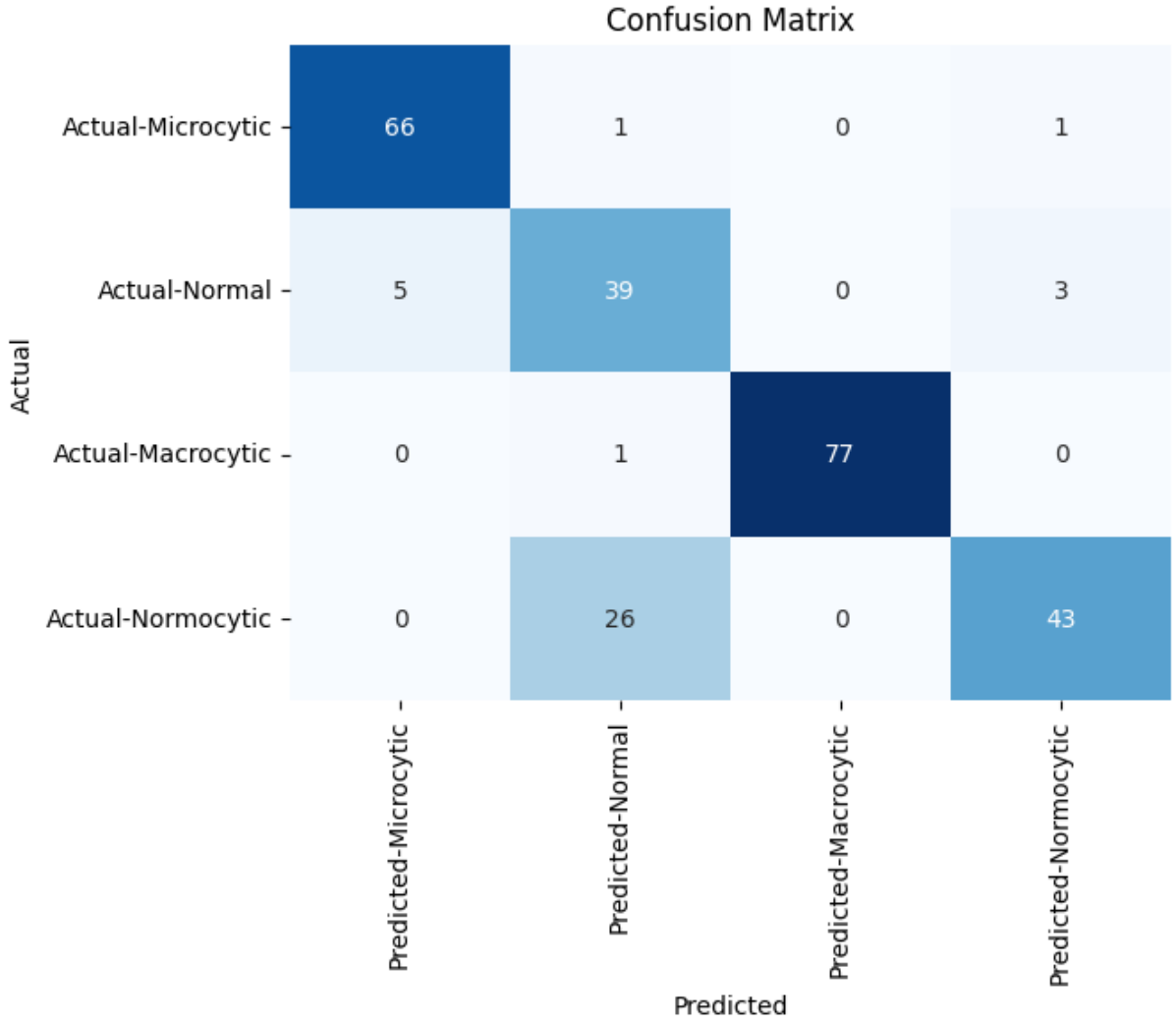

Figure 4.11: Confusion matrix of Bagged Logistic Regression

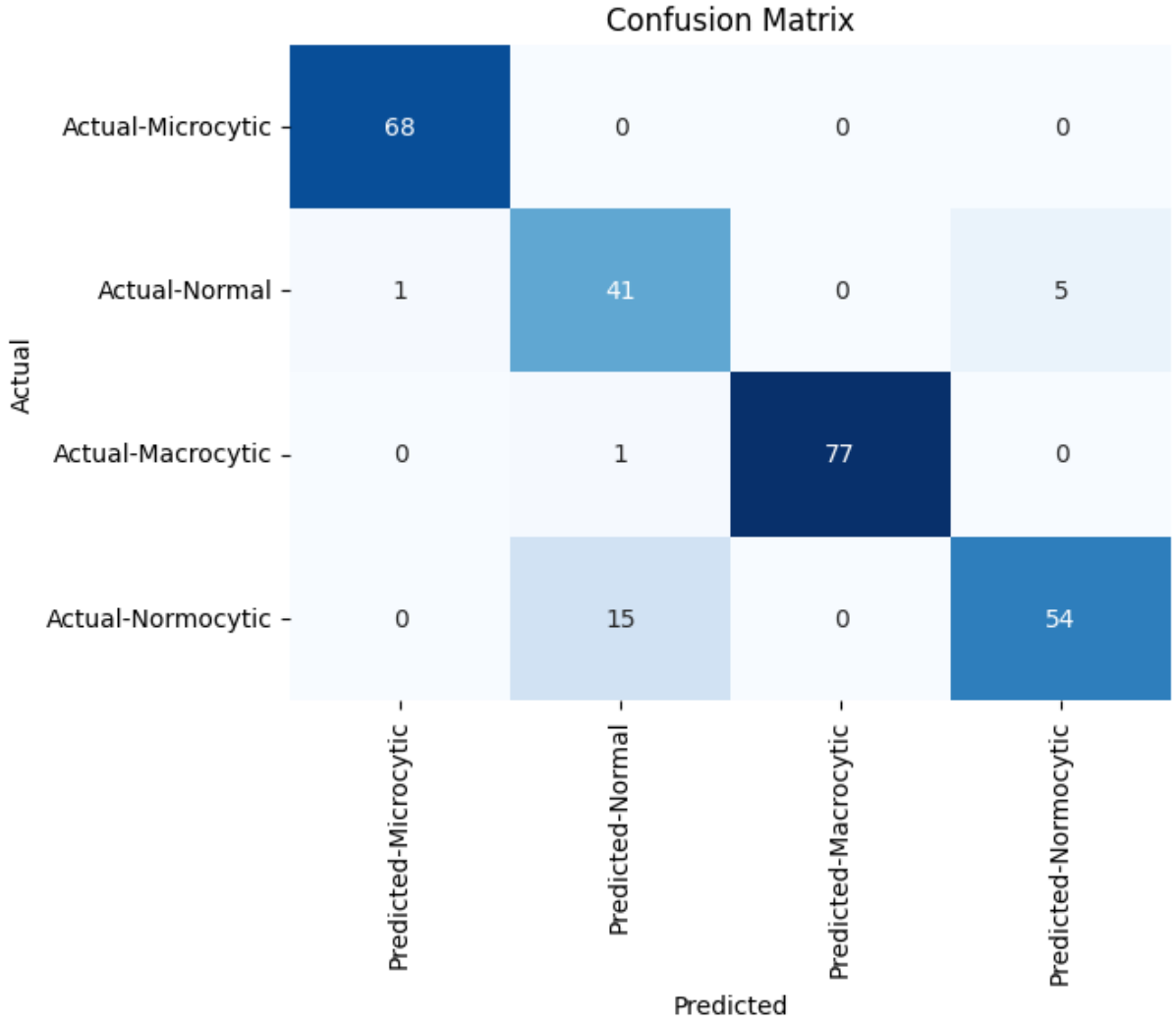

Figure 4.12: Confusion matrix of Bagged Random Forest

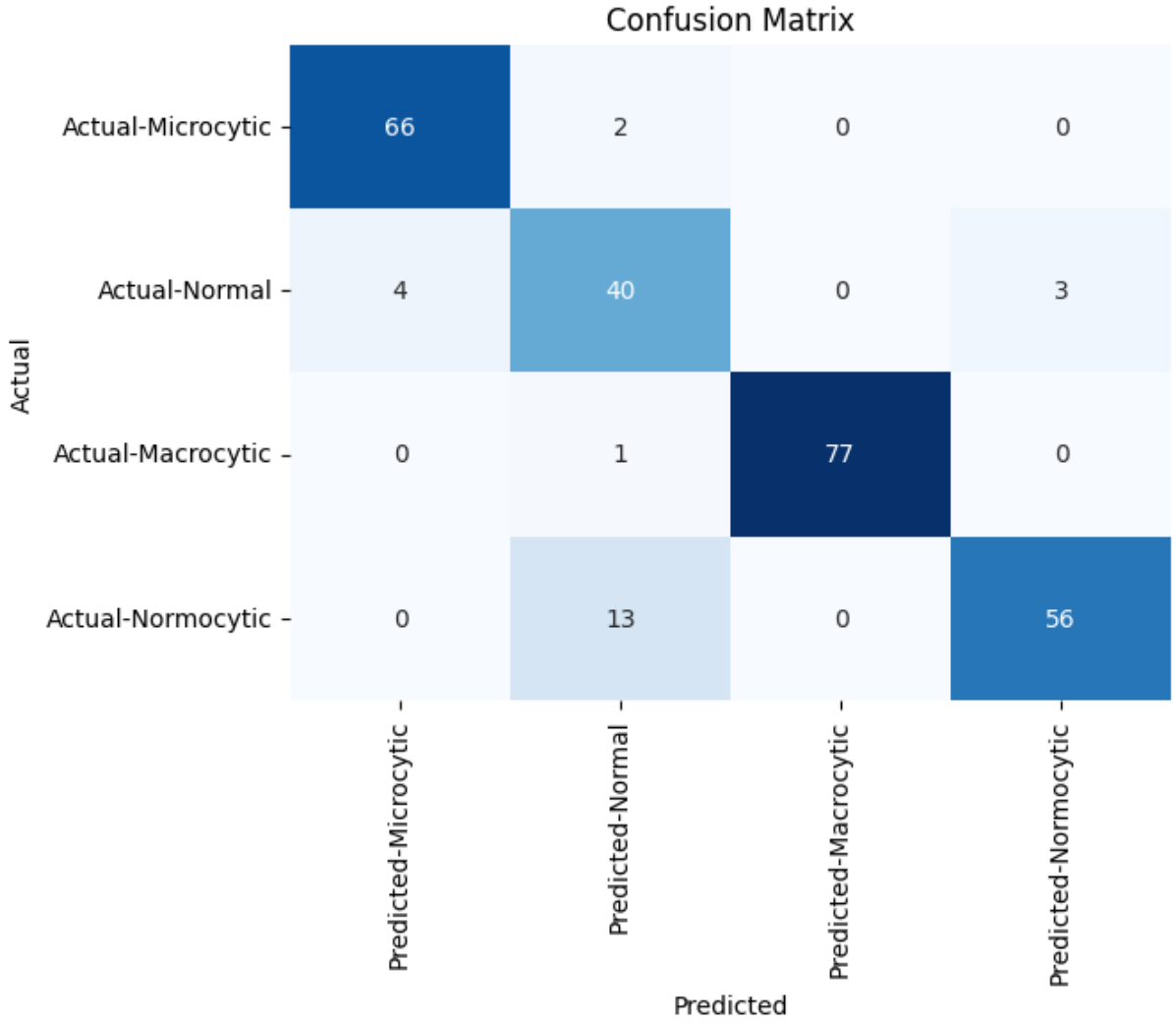

Figure 4.13: Confusion matrix of Bagged SVM

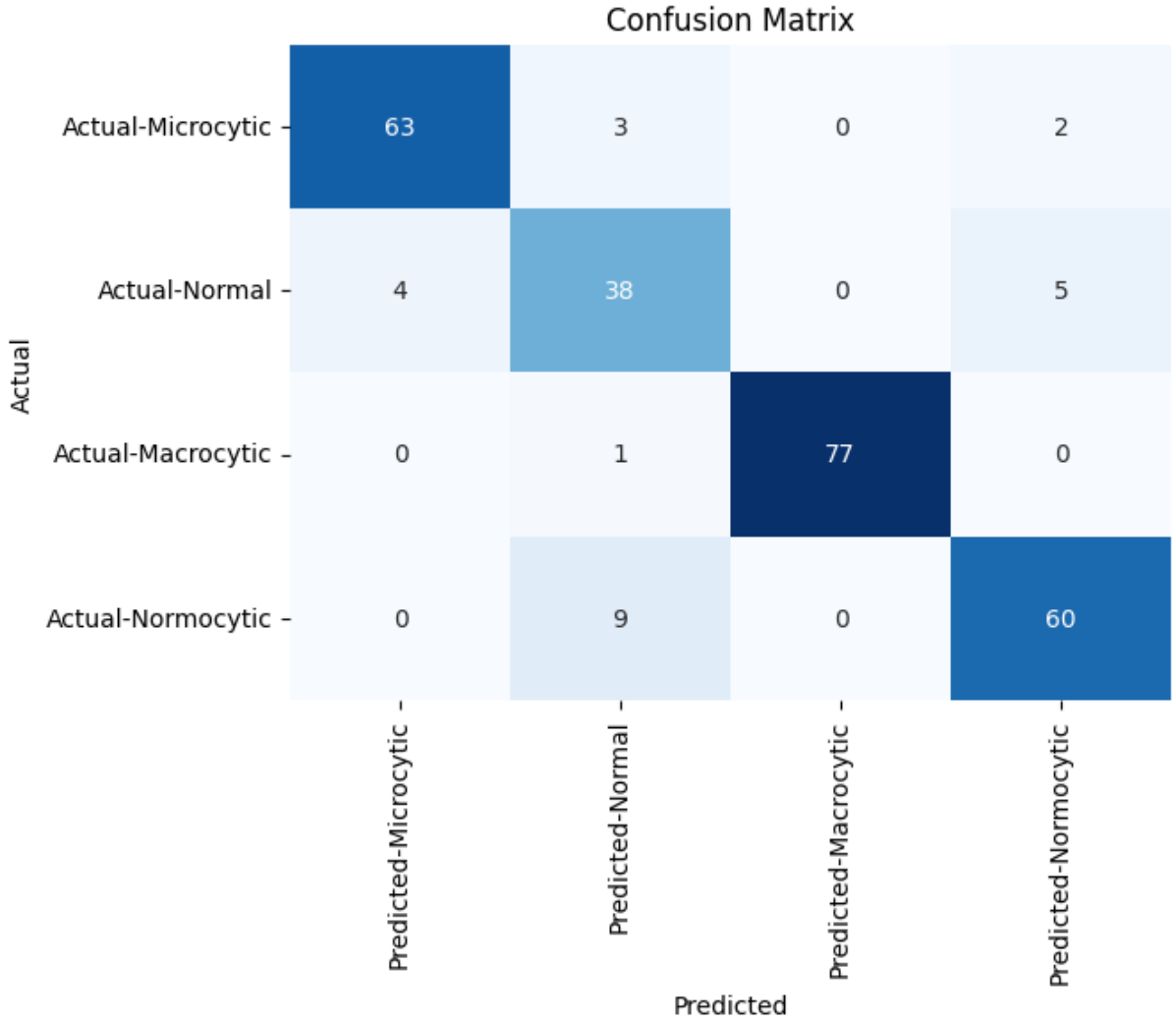

Figure 4.14: Confusion matrix of Bagged K-Nearest Neighbors

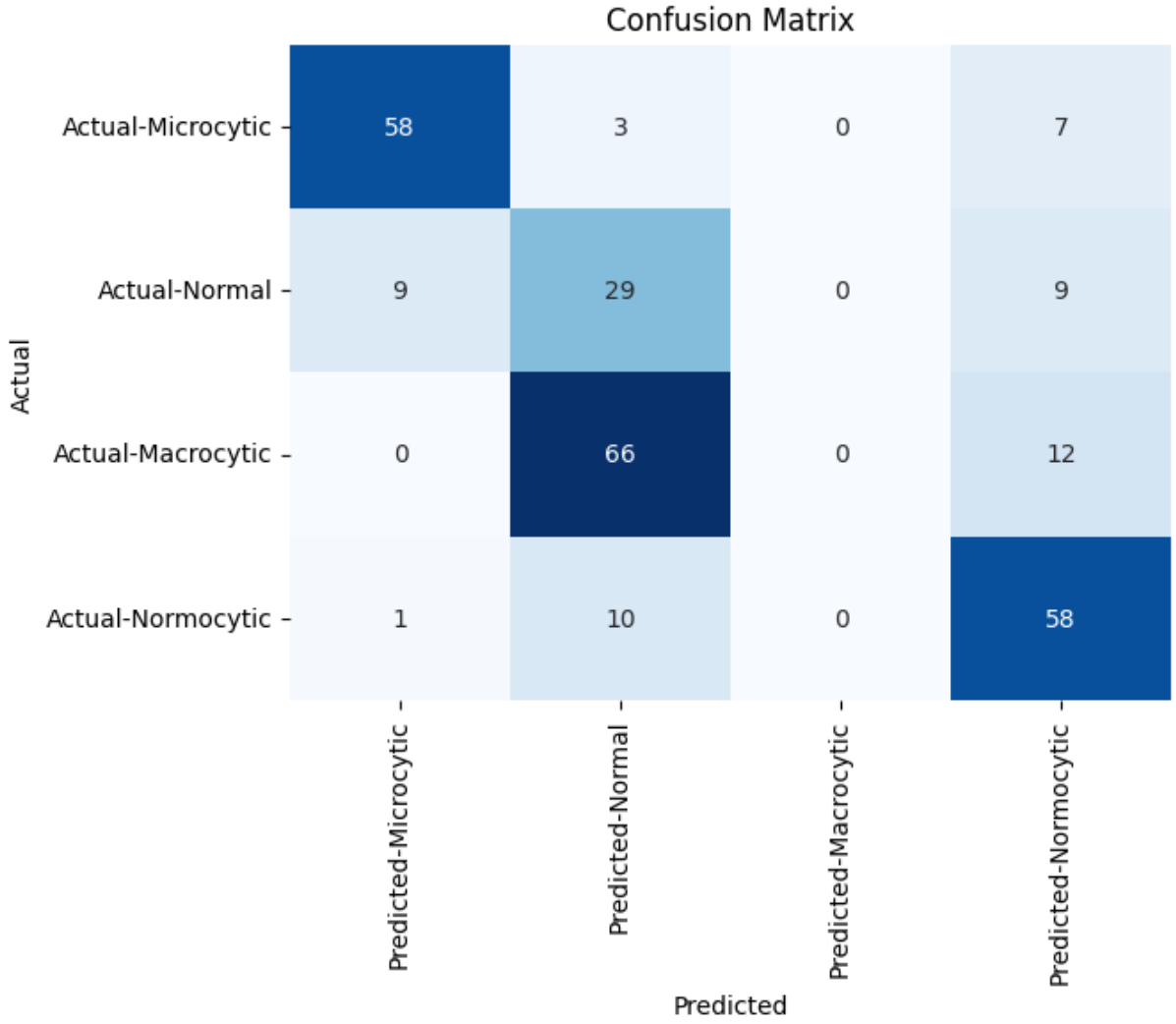

Figure 4.15: Confusion matrix of Bagged Gaussian Naïve Bayes

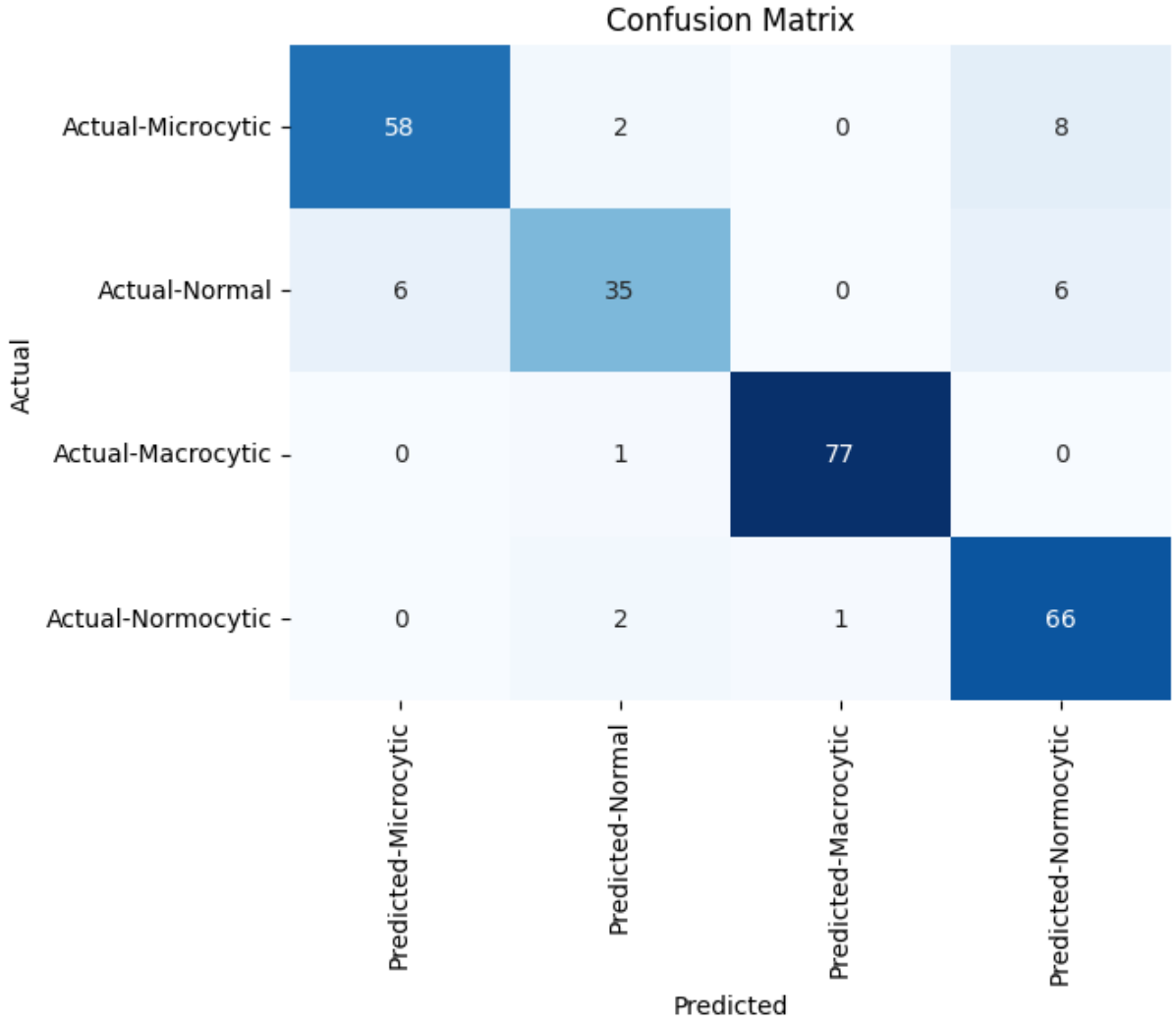

Figure 4.16: Confusion matrix of Bagged Quadratic Discriminant Analysis

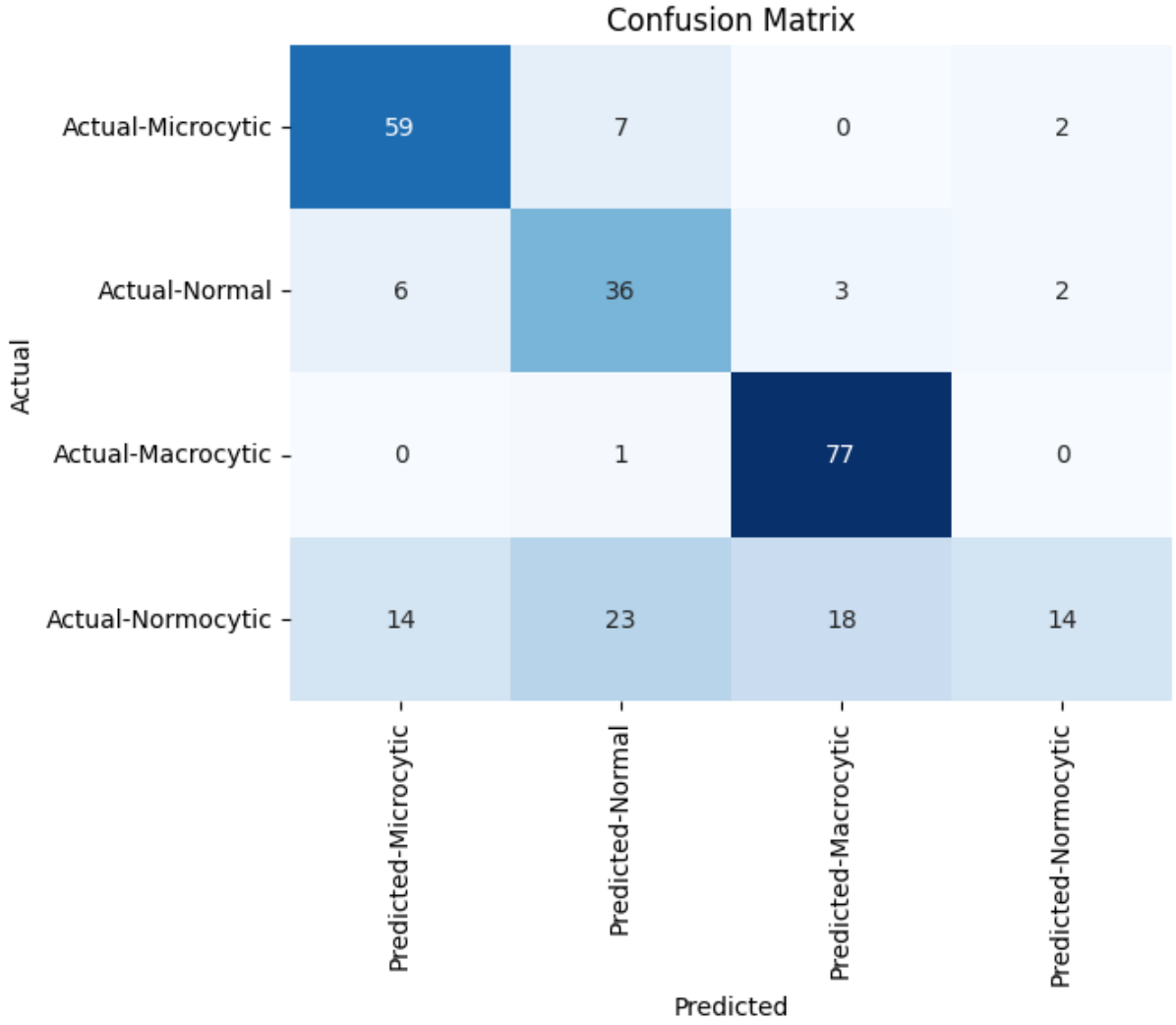

Figure 4.17: Confusion matrix of Bagged Ridge Classifier

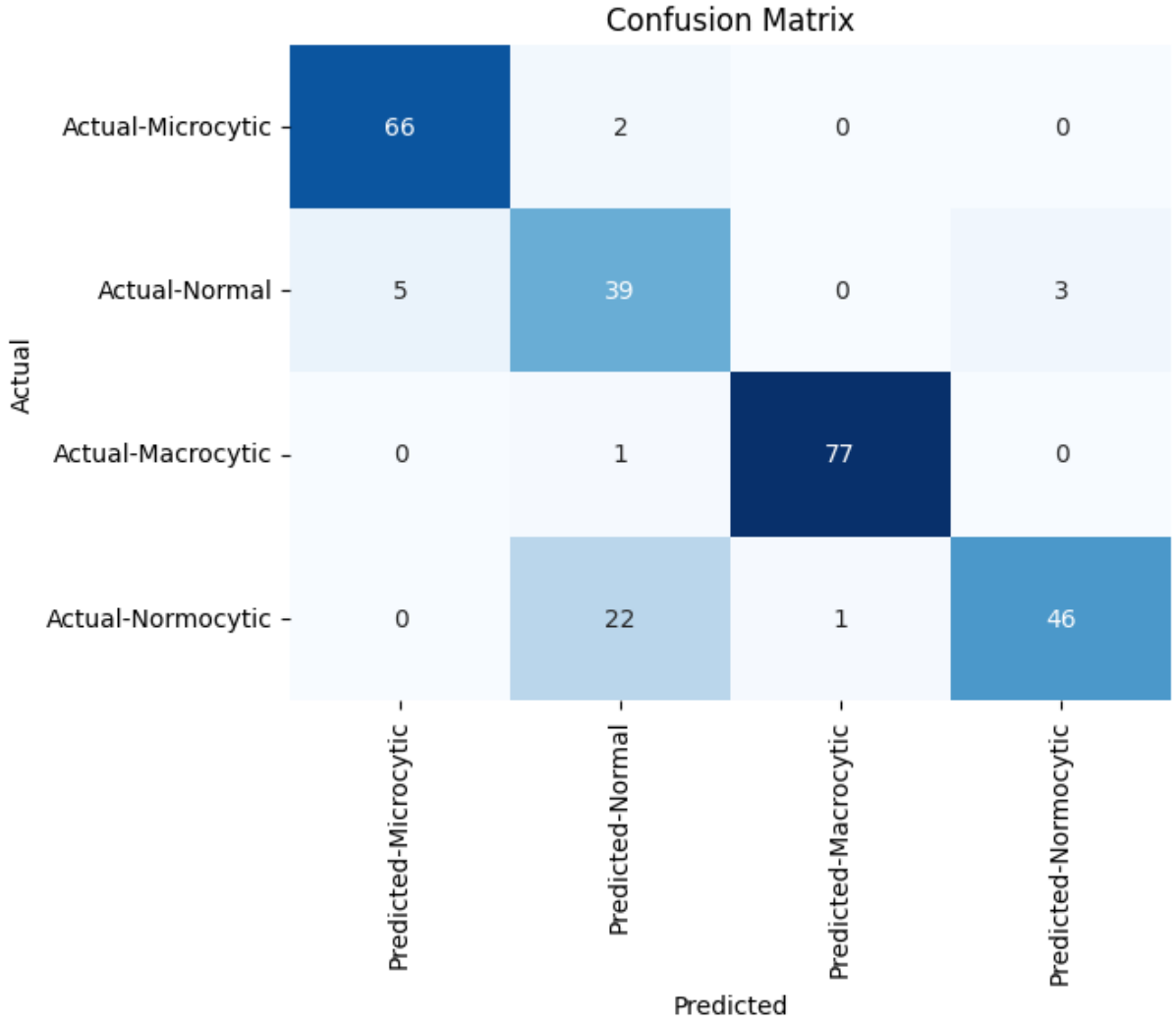

Figure 4.18: Confusion matrix of Bagged Passive Aggressive Classifier

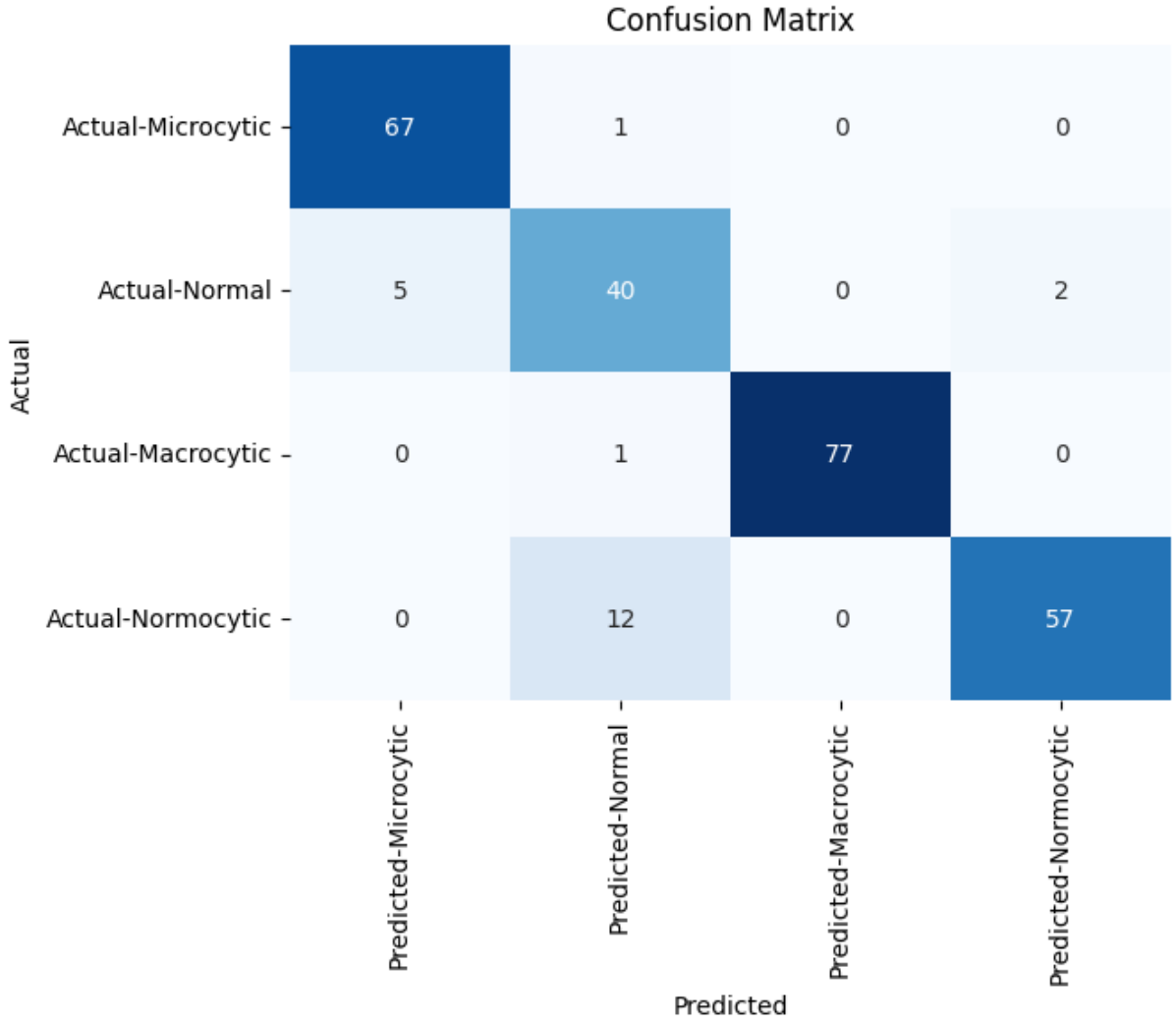

Figure 4.19: Confusion matrix of Bagged Grid Search CV

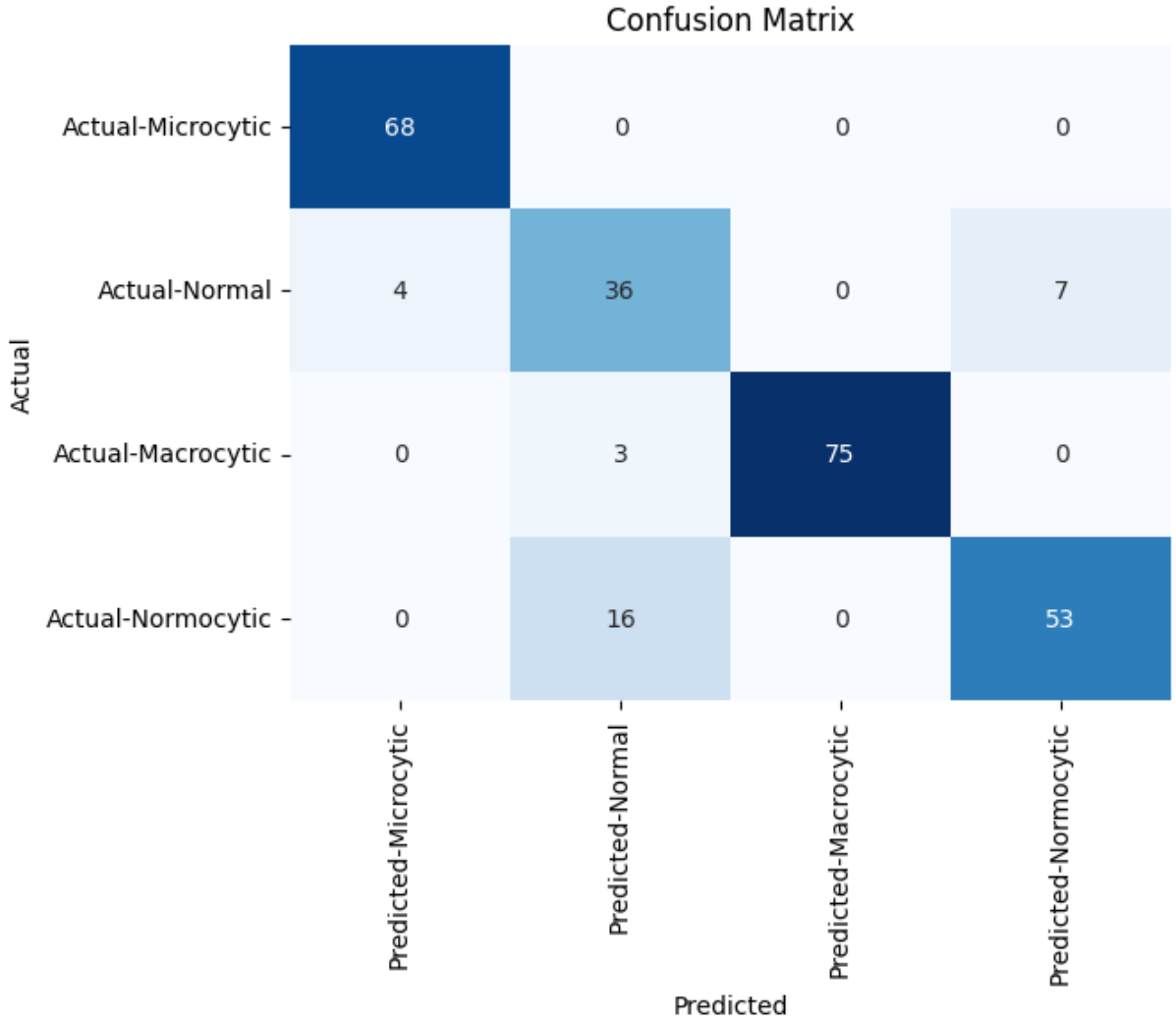

Figure 4.20: Confusion matrix of Bagged Grid Search CV

Bagged LR, Bagged RF, Bagged SVM, Bagged KNN, Bagged QDA, Bagged RC, Bagged PA, Bagged GS, and Bagged XGB confusion matrices are all depicted in Figures 4.11, 4.12, 4.13, and 4.16, respectively.

The above-mentioned table 4.3 illustrates the performance of some of the traditional machine learning algorithms when hybridized with an ensemble boosting paradigm for the anemia disease prediction task. The evaluation metrics employed to contrast each model are Accuracy, Precision, Recall, and F1 Score, which together provide an overall picture of the performance of each classifier. Boosting differs from bagging in that it attempts to train a sequence of weak learners one after another, each of them improving upon the previous one's mistakes, hence enhancing performance incrementally. Boosting is oriented towards hard cases, allowing the final model to be highly focused on high accuracy and low generalization error.

Table 4.3: Ensemble Boosting Algorithm results

Algorithm	Accuracy	Precision	Recall	F-1 Score
LR	75.57	77.69	75.57	74.82
RF	89.69	91.64	89.69	90.10
SVM	17.93	3.21	17.93	5.45
GNB	58.01	49.97	58.01	52.23
RC	83.58	85.00	83.58	83.83

Random Forest (RF), though initially a bagging-based classifier, was transformed to the boosting framework and resulted in the best performance among all classifiers in this study. RF resulted in accuracy of 89.69%, precision of 91.64%, recall of 89.69%, and F1 score of 90.10%. The results indicate that the model not only identifies anemia cases accurately (high recall), but also avoids over-diagnosis (high precision), which is of high significance in clinical decision-making. The better F1 score also shows an ideal balance of precision and recall, and this is the most accurate model with this boosting setting.
Ridge Classifier (RC), another linear classifier employing L2 regularization to prevent overfitting, performed very well at 83.58% accuracy, 85.00% precision, 83.58% recall, and F1 score of 83.83%.
This means that the boosting has significantly improved the performance of this relatively simple model. The error correction and sequential training mechanism inherent in boosting likely allowed RC to learn more subtle patterns in the data that it would otherwise have missed. In medical applications, where underfitting can be dangerous, boosting provides a significant benefit by being capable of making even linear models competitive. Logistic Regression (LR) performed moderately with an accuracy rate of 75.57%, precision rate of 77.69%, recall rate of 75.57%, and an F1 value of 74.82%.
These performances are good for a linear model, particularly in boosting, where its predictive strength is boosted by layered learning methodology.
Boosting facilitates LR to be more accurate with its decisions since it highlights its mistakes through iteration during training. But its relatively lesser output compared to RF and RC suggests that it may still be beset with the non-linearity and feature interaction inherent in anemia-related data. Gaussian Naïve Bayes (GNB), with 58.01% accuracy, 49.97% precision, and 52.23% F1 score, was not very effective even when boosted. Naïve Bayes makes the conditional independence assumption across features, something that rarely holds in real-world medical data where symptoms and laboratory results tend to be correlated. While boosting does enhance model generalization, it cannot entirely correct for defective base assumptions, something that might account for GNB's poor performance here. Surprisingly, Support Vector Machine (SVM), or otherwise viewed as a very strong classifier in binary classification scenarios, has performed extremely poorly within this boosting system, with mere 17.93% accuracy, 3.21% precision, 17.93% recall, and a highly low F1 measure of 5.45%.

This difference may be owing to numerous reasons. Secondly, SVM is a margin classifier and not necessarily easily incorporated into the majority of standard boosting methods like AdaBoost or Gradient Boosting that employ decision tree stumps as the base learners. SVMs also exhibit sensitive parameter tuning and scaling dependence and how their ensemble performance relies fairly heavily on their combination and configuration. The sudden drop in accuracy means that the model generated an awful lot of false positives, and that is just not acceptable within a medical setting. The boosting strategy has the effect of improving weak learners' performance by concentrating on hard-to-classify examples. For anemia prediction, when training data can possess subtle patterns, noisy features, or unbalanced classes, boosting provides a helpful mechanism for emphasizing learning from such hard cases. This is confirmed by the significant improvements in performance for such models as RF and RC. High precision and recall are especially crucial in clinical diagnosis. High precision is to ensure patients with anemia diagnosed actually have the illness, minimizing unjustified treatments or alarm. High recall is to ensure most genuine cases of anemia are correctly identified, which is extremely crucial to avoid missed diagnoses. The F1 Score, the harmonic mean of precision and recall, best serves in describing a model's performance in such well-balanced measures.

The results also indicate that all models are not equally benefited by boosting. Whereas the regularized linear model and tree models like RF and RC work very well here, high assumption or high input transformation sensitivity models like GNB and SVM can get hurt if not properly tuned.

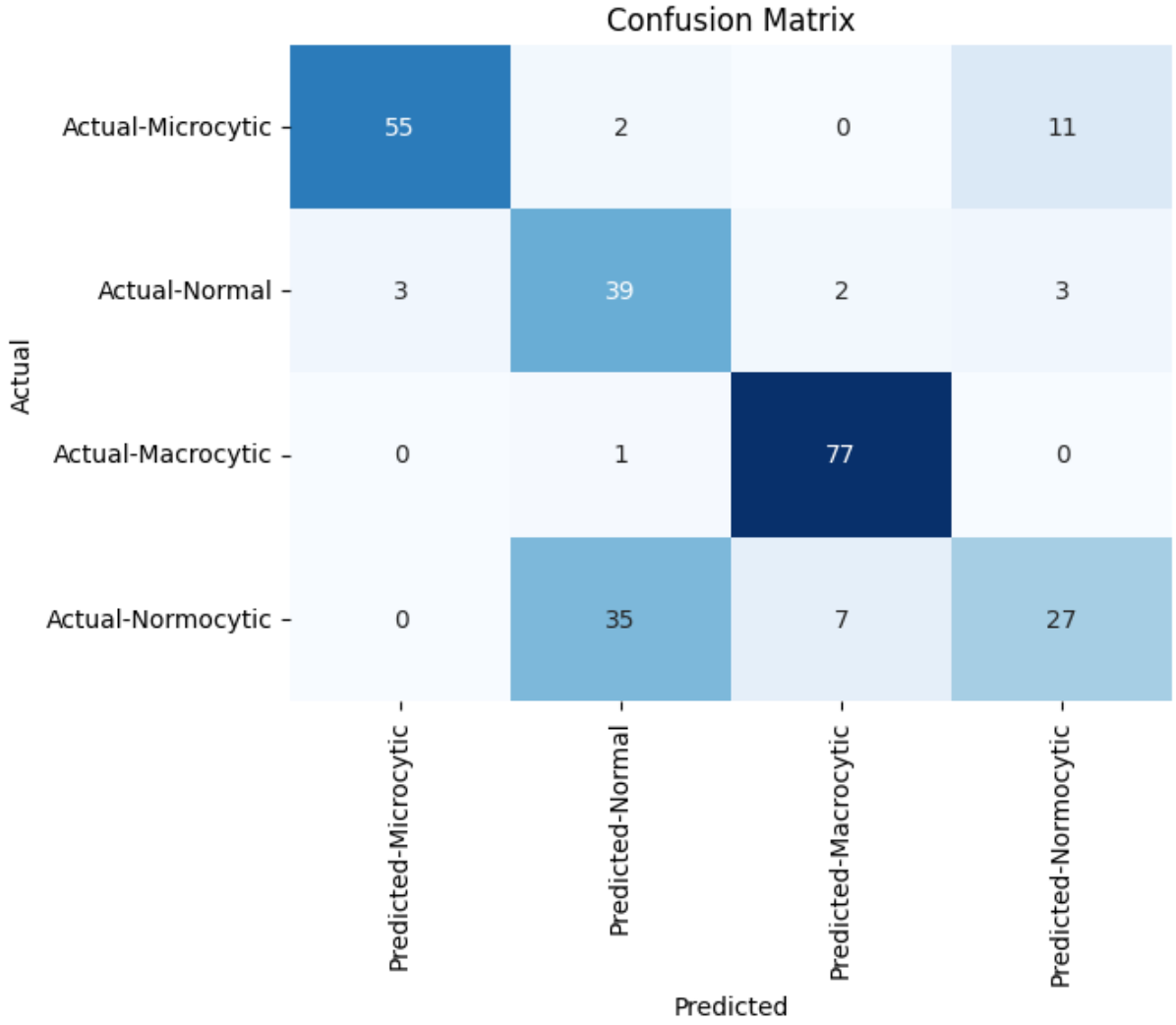

Figure 4.21: Confusion matrix of Boosted Logistic Regression

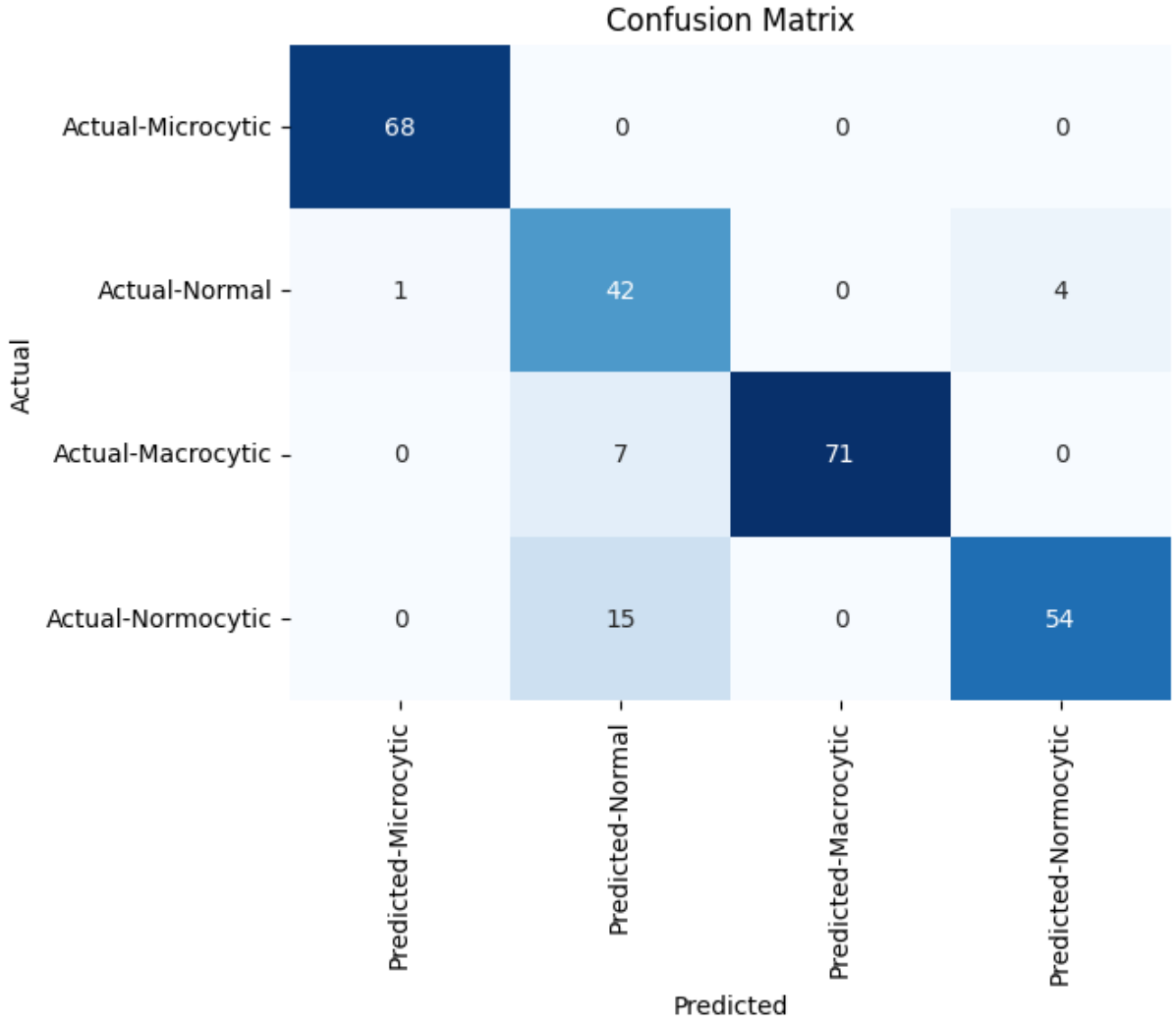

Figure 4.22: Confusion matrix of Boosted Random Forest

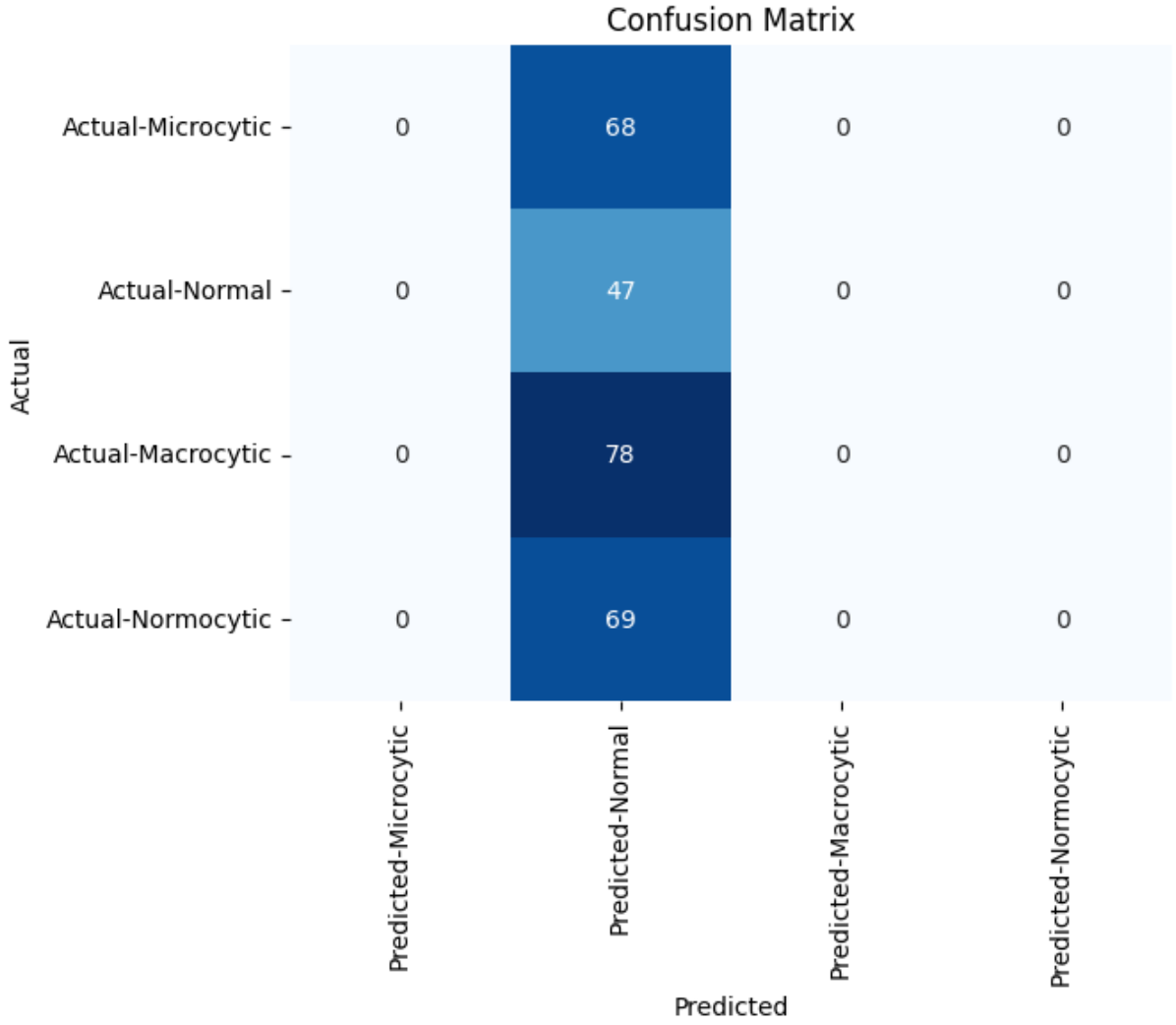

Figure 4.23: Confusion matrix of Boosted SVM

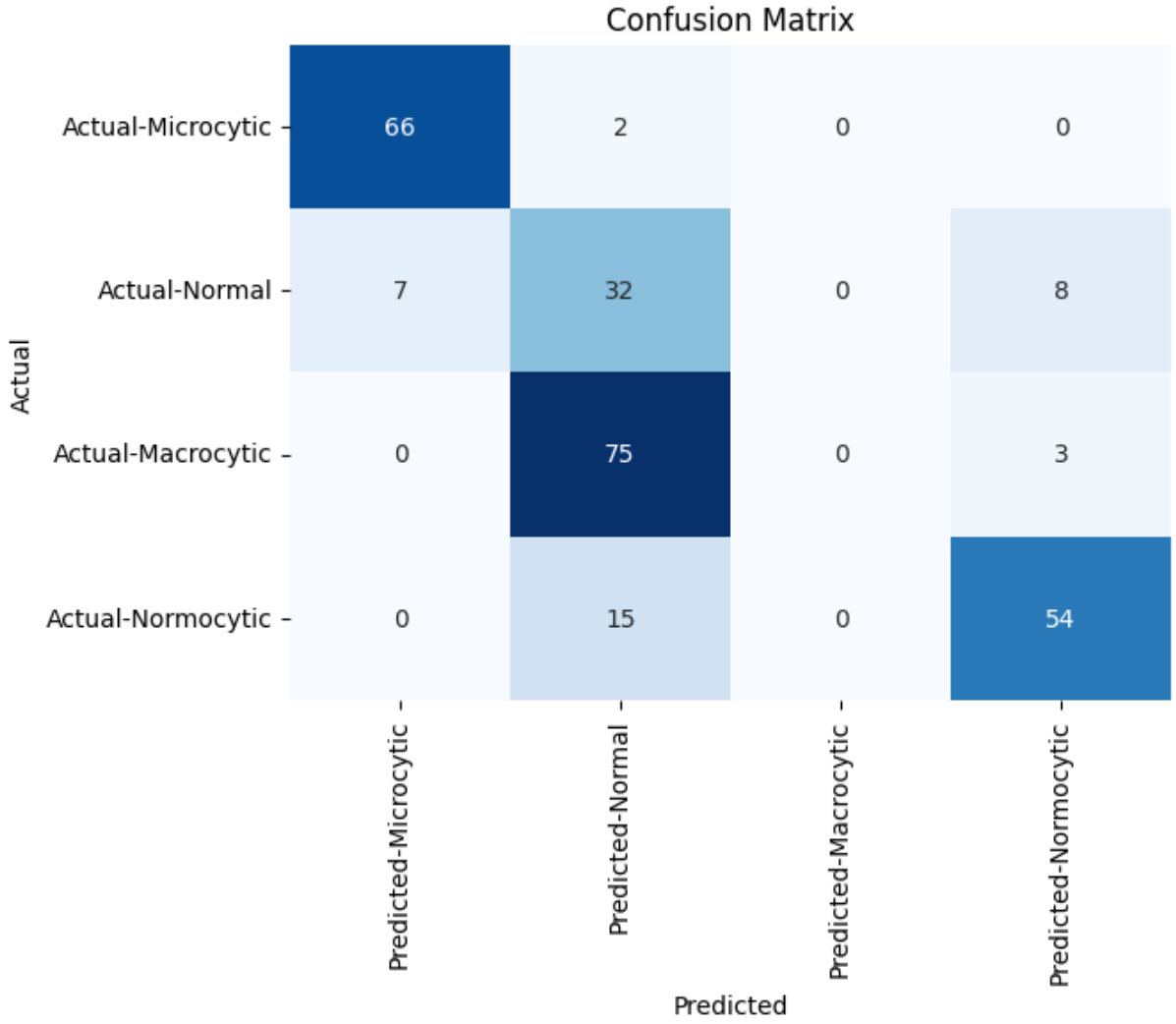

Figure 4.24: Confusion matrix of Boosted Gaussian Naïve Bayes

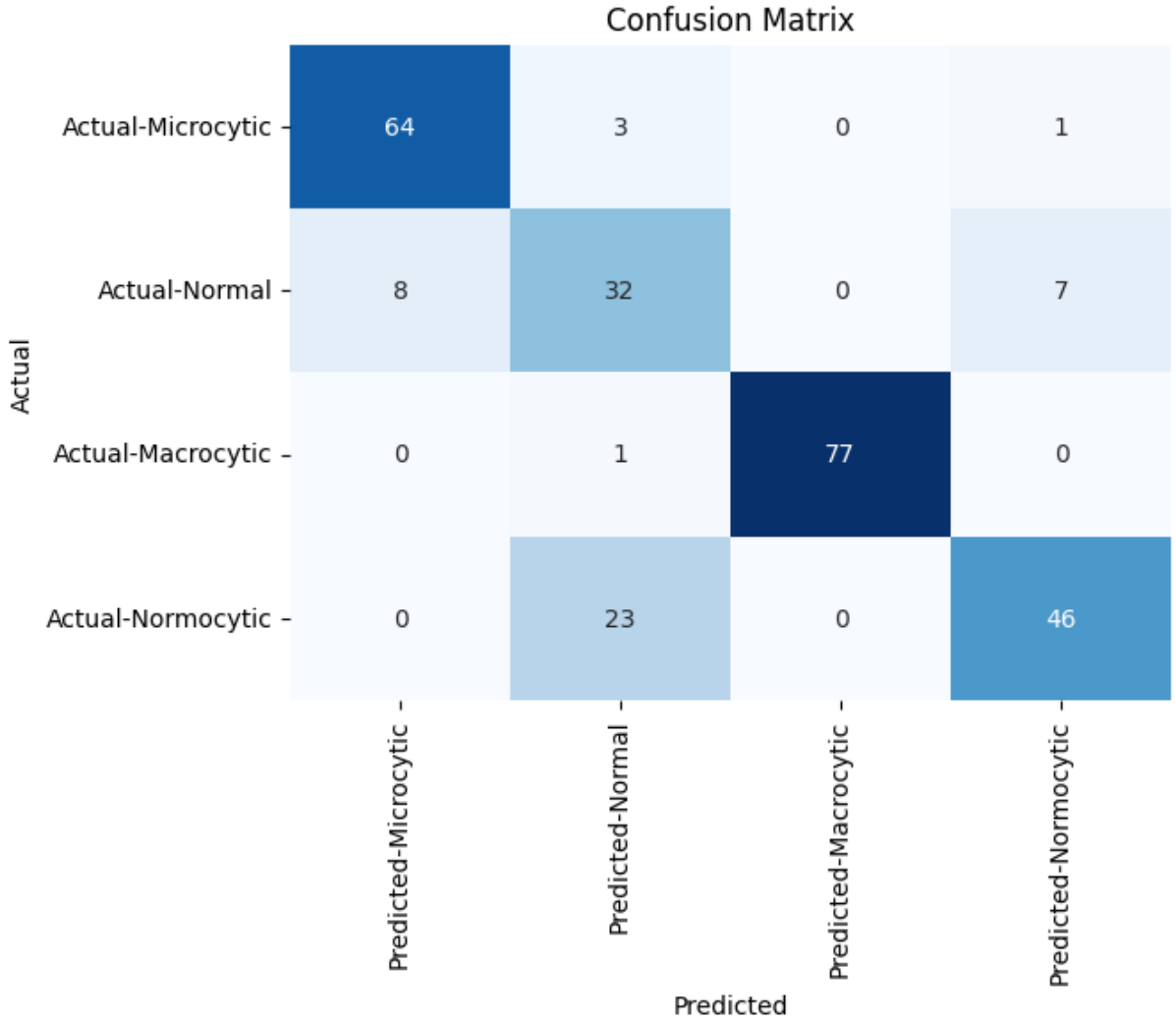

Figure 4.25: Confusion matrix of Boosted Ridge Classifier

Figure 4.21 presents the confusion matrix for Boosted LR, Figure 4.22 for Boosted RF, Figure 4.23 for Boosted SVM, Figure 4.24 for Boosted GNB, and Figure 4.25 for Boosted RC.

Performance measures of the Voting ensemble classifier for anemia disease prediction exhibit highly promising results with accuracy of 93.12%, precision of 93.21%, recall of 93.12%, and F1 value of 93.13%. These values indicate that the voting model, which outputs an aggregate of multiple base classifier predictions (in most instances using majority voting or averaging probabilities), has a very balanced and higher diagnostic accuracy. Its high precision indicating that it has extremely low false positive predictions, and the equally high recall indicating that it correctly identifies nearly all cases of anemia,. The high F1 score, which is balanced between precision and recall, also speaks to the reliability and suitability of the model for clinical decision-making. The ensemble vote leverages the maximum strengths of individual models, compensates for their shortcomings, and provides a more generalized and accurate output, hence being one of the best performing methods in this research.

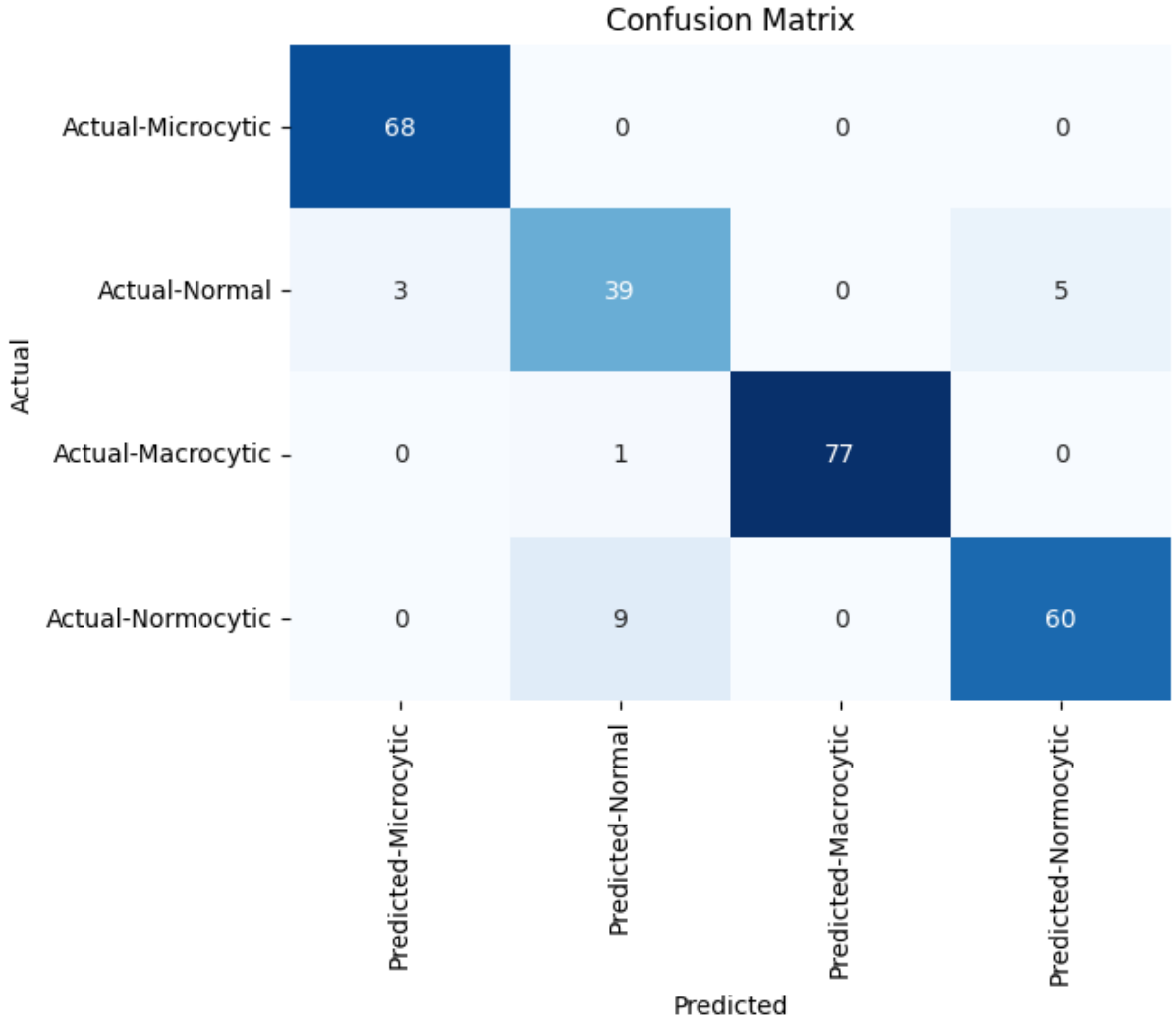

Figure 4.26: Confusion matrix of Voting Classifier

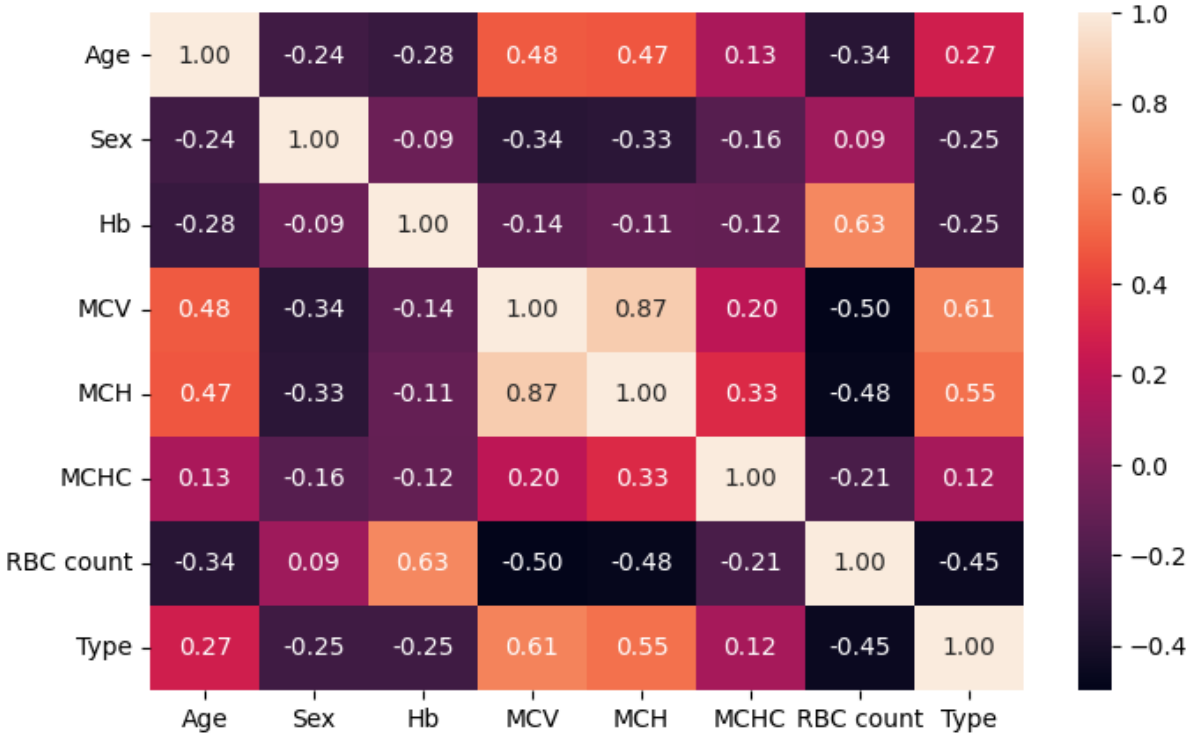

Figure 4.27: Correlation graph of the anemia dataset

Figure 4.27 shows the correlation graph of the anemia disease dataset. It means the internal relationship between data fields.

4.3 Results and Discussion

The comparison and analysis of the traditional, bagging, boosting, and voting classifiers for the prediction of anemia disease offer important insights into the strengths and weaknesses of each classifier. The traditional machine learning classifiers, KNN and GS, were two of the top performers with accuracies of 91.98% and 91.60%, respectively. The models exhibited balanced performances across all measures of evaluation—precision, recall, and F1-score—demonstrating that they are able to make predictions with reliability. Nonetheless, classical models such as RC and GNB performed below par, with GNB demonstrating a remarkably low accuracy rate of 55.34%, indicating its assumptions were not as aligned with the distribution of the dataset.

On the other hand, ensemble bagging techniques, particularly RF, scored consistently better with an accuracy of 91.60% and a good F1-score of 91.70%, which beats most traditional models. Bagging is effective in taming overfitting by taking the average of several weak learners and their predictions to boost overall robustness. Similarly, GS and SVM under bagging also lost minimal predictive capability, cementing the ensemble's ability to generalize well across diverse samples.

On the other hand, boosting classifiers showed contradictory results. While Random Forest (RF) and Logistic Regression (LR) had reasonably good performance, SVM with boosting had a very poor accuracy of only 17.93%, along with very low precision and F1-score, indicating that boosting in this case may have enhanced model weaknesses or overfitted noise in the data. Boosting generally achieves this by paying greater weights to mislabeled samples from previous time, but can introduce even poorer within-class performance under class imbalance and noisy datasets unless properly tuned. Finally, the Voting classifier was the best among all ensemble methods with 93.12% accuracy,

precision, recall, and F1-score. This indicates that combining heterogeneous models—specifically high-performance models like RF, SVM, and GS—using majority or soft voting can lead to improved generalization. The voting ensemble exploits the complementary strengths of individual algorithms, overriding the effect of individual errors and making stable, consistent predictions

The study reaffirms that ensemble techniques, and specifically bagging and voting, far outshine individual traditional models in anemia prediction, with superior accuracy, stability, and generalizability.

4.4 Summary

Boosting, although very strong, must be implemented with caution and sensitive parameter setting, as the performance may significantly depend on model compatibility and dataset type. The findings determine that the Voting ensemble model is the best and most stable at predicting early anemia in the observed population, and it can be considered a potential option for the implementation of decision-support systems in health facilities.

Chapter 5
Engineering Standards and Design Challenges

The engineering standards used in the design of the anemia disease prediction system are explained under this chapter, along with the major design challenges faced. It shows how compliance with the set procedures ensured system reliability as new solutions had to be adapted to counter technology, computation, and data issues at the project level.

5.1 Compliance with the Standards

To keep the system secure, reliable, and ethically process data, different requirements were strictly followed during the development of the anemia disease prediction system. Adherence to these specifications ensured communication processes' reliability and safety, effectiveness in the use of hardware, and software quality.

5.1.1 Software Standards

Functional sufficiency, reliability, usability, and maintainability were taken into account in the project while adhering to ISO/IEC 25010 software quality standards. IEEE 730 (Software Quality Assurance Plans), with explicit procedures for quality assurance, was another candidate in view. But since it addresses a broader set of quality qualities that are essential to healthcare prediction systems, ISO/IEC 25010 was utilized. It offered a structured balance appropriate for the project's research and development stage without enforcing draconian documentation requirements.

5.1.2 Hardware Standards

For hardware compliance, ISO/IEC 60950 (Information Technology Equipment Safety) standards were considered for safe use of computing equipment while rolling out the system. Another alternative was UL 60950, mainly used in North America for safety of information technology equipment. ISO/IEC 60950 was used because it was internationally accepted and usable, such that in the event of future roll-out of the anemia prediction system to other regions, hardware compliance for safety would not be an issue.

5.1.3 Communication Standards

Use of HTTPS (SSL/TLS protocols) to secure communication channels was required since the collection and analysis of data required patient data to be transmitted. Another choice that had been widely used for secure network-level encryption was IPSec. Nevertheless, HTTPS was utilized because it finds widespread usage to secure medical data within healthcare portals and is simple to implement with web applications. With low system overhead, it provided data privacy and integrity that was exchanged.

5.2 Impact on Society, Environment and Sustainability

Enhanced health outcomes and responsible technical progress are anticipated benefits from the anemia disease prediction system for people and society as a whole.

5.2.1 Impact on Life

By facilitating early medical intervention, the method improves early diagnosis of anemia and minimizes the risk of side effects like kidney damage. This enhances the quality of life of patients considerably, particularly for vulnerable groups like the elderly and infants.

5.2.2 Impact on Society & Environment

The process reduces the burden on healthcare systems through promoting early detection, which is less costly and enhances the allocation of medical resources. Environmentally, a computer-based prediction system promotes greener medical practices through minimizing paper-based documentation and employing electronic health records (EHR).

5.2.3 Ethical Aspects

The main focus of the project was to manage patient data ethically. In accordance with privacy legislation and ethical research practice under the Declaration of Helsinki, data were anonymized and no individual participant identifying data were preserved.

5.2.4 Sustainability Plan

Owing to its lightweight and scalable nature, the system can be interfaced with mobile health (mHealth) systems to reach out to underserved or rural areas. In order to address environmental sustainability, future releases will aim at constraining energy consumption for model training and deployment.

5.3 Project Management and Financial Analysis

Project planning, resource planning, and risk management were a few of the formal management methods utilized in the project.
Data gathering, cloud computing services, software licensing, and little marketing were all included in the first budget of $5,000. A second budget of $35,000 that utilized only open-source, free platforms and no marketing was an option, but it was eliminated because the increased budget guaranteed higher quality data, safe hosting, and greater viewership.
A freemium model was suggested for the revenue model, where users can access free basic anemia risk assessments and pay for more expensive services like telemedicine consultations and complete diagnostic reports. This strategy provides public access as well as long-term financial viability for system adjustments and upkeep.

5.4 Complex Engineering Problem

The research undertaken in this thesis, Anemia disease prediction constitutes a Complex Engineering Problem. It delves into the frontiers of machine learning and artificial intelligence, requiring sophisticated analytical skills, management of highly complex and often incomplete datasets, and the development of novel computational methods.

5.4.1 Complex Problem Solving

This table 5.1 and 5.2 is designed to map the EP and CO.

Table 5.1: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent of Stakeholder Involvement	EP7 Inter-dependence
✓	✓	✓	✓		✓	✓

Mapping with Knowledge Profile for EP1.

Table 5.2: Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

5.4.1.1 Justification for EP Attributes Mapping

- **EP1 - Depth of Knowledge Required:**
  Requires advanced knowledge in machine learning (specifically machine learning models), statistics (for model validation and handling uncertainty), graph theory, and confusion matrix (for model evaluation).

- **EP2 - Range of Conflicting Requirements:**
  - **Model Complexity vs. Interpretability:** More complex machine learning models might achieve higher accuracy but become some other than deep learning models can be effective.
  - **Data Heterogeneity vs. Integration:** Integrating diverse data types (graphs, sequences, numerical profiles) effectively into a single model is challenging and may require trade-offs in how much information from each source is retained.

- **Prediction Accuracy vs. Generalizability:** Models highly tuned to existing data might not generalize well to novel anemia diseases.
- **Computational Cost vs. Scalability:** Training machine learning models on large numeric data can be computationally intensive, conflicting with the need for scalable solutions.

- **EP3 - Depth of Analysis:**

  Numeric data is inherently noisy, incomplete, and contains biases. Analyzing experimental results from machine learning requires sophisticated statistical techniques to account for this uncertainty. Understanding failure modes of the models and identifying sources of error involves deep, often iterative, analysis.

- **EP4 - Familiarity of Issues:**

  While machine learning models are an active research area, designing novel architectures specifically tailored for the nuances of anemia repurposing, incorporating interpretability mechanisms, and addressing severe data imbalance in this domain are at the research frontier. Standard machine learning models may not perform optimally without significant novel modifications.

- **EP6 - Extent of Stakeholder Involvement and Conflicting Needs:**
  - **Academic Researchers:** Focus on novelty, methodological rigor, and publication.
  - **Pharmaceutical Companies:** Interest in practical, high-accuracy predictive tools that can reduce R&D costs and time.
  - **Medical Practitioners/Patients:** Ultimate beneficiaries, needing safe and effective treatments.

- **EP7 - Inter-dependence:**

  The machine learning model model itself is a complex system of interacting layers and components. Its predictions can have profound implications for anemia development pipelines, public health (by identifying new treatments faster), and economic impacts on the pharmaceutical industry.

5.4.1.2 Justification for Knowledge Profile Mapping (linked to EP1):

- **K3 - Engineering Fundamentals:**

  Strong grounding in algorithms, data structures, probability, and linear algebra is essential for understanding and implementing machine learning model.

- **K4 - Specialist Knowledge:**

  Deep expertise in machine learning (machine learning model theory, attention mechanisms, optimization), bioinformatics (network architecture, Anemia-target interaction databases), and statistical modeling.

- **K5 - Engineering Design:**

  Designing novel machine learning model model architectures, crafting new loss functions to handle imbalance, developing feature engineering strategies for

heterogeneous numeric data, and designing robust experimental setups for model evaluation.

- **K6 - Engineering Practice:**

  Rigorous experimental design, statistical validation of results, version control for code and experiments, writing research papers, and presenting findings.

- **K8 - Research Literature:**

  This thesis is heavily reliant on, and aims to contribute to, cutting-edge research literature in machine learning model, and confusion matrix discovery. Continuous literature review is paramount.

5.4.2 Engineering Activities

Mapping with Knowledge Profile for EP1

This table 5.3 is designed to map the EP1 to the Knowledge Profile.

Table 5.3: Mapping with complex engineering activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

5.4.2.1 Justification for Engineering Activities Mapping:

- **EA1 - Range of Resources:**

  Utilizes public numeric databases, highperformance computing clusters for model training, specialized software libraries (Python, anaconda), and extensive academic literature.

- **EA2 - Level of Interaction:**

  Collaboration with thesis supervisor, discussions with peers in research labs, interacting with domain experts in diagnosis and imaging.

- **EA3 - Innovation:**

  The core of the thesis is the development and evaluation of novel machine learning model and deep learning model architectures or methodologies for Anemia repurposing, aiming to push beyond existing state-of-the-art.

- **EA4 - Consequences for Society and Environment:**
  - **Societal:** Potential to accelerate the discovery of new treatments for diseases, reduce healthcare costs, and improve public health. Ethical considerations regarding data privacy and equitable access to discovered treatments.

- **Environmental:** Minimal direct environmental impact, though computational resources consume energy.

- **EA5 - Familiarity:**

  While some aspects involve established procedures (e.g., standard data preprocessing steps, common evaluation metrics), the design of novel machine learning model architectures and their application to complex, multi-modal numeric data involves significant departures from well-trodden paths and requires creative problem-solving.

5.5 Summary

Technological standards that were adhered to in designing the anemia disease prediction system have been addressed in this chapter with a focus on software, hardware, and communications compliance to maintain system effectiveness, security, and reliability. The project's ability to enhance access to healthcare, minimize its ecological footprint, and uphold high standards of ethics on patient data was highlighted by exploring its sociological, environmental, and ethical influences. For purposes of long-term sustainability, a sustainability plan was suggested. A review of finances and project management practices were also provided, comparing budgets with each other and creating a freemium financial model to provide for future development as well as user access. The chapter as a whole demonstrates how deliberate standards compliance and planning helped craft a sound, ethical, and lasting healthcare solution.

Chapter 6
Conclusion

This chapter contains the study's findings, limits, and upcoming projects.

6.1 Summary

In order to guarantee quality, security, and ethical integrity, this project concentrated on creating a machine learning-based anemia illness prediction system while following crucial hardware, software, and communication standards. By facilitating early diagnosis, lessening the strain on hospital infrastructure, and encouraging ecologically friendly behaviors, the system was intended to have a favorable effect on healthcare delivery. Financial planning and an organized project management methodology aided in the system's practical development, guaranteeing its scalability, dependability, and long-term viability.

6.2 Limitation

Notwithstanding the encouraging outcomes, the initiative had certain drawbacks. Due to limited access to real-world medical information, people have to rely on publically accessible or artificial datasets, which might not fully represent the range of variances found in real-world clinical settings. Despite its accuracy, the prediction model may still encounter difficulties when used on varied and invisible patient populations because of variations in data quality, healthcare procedures, and demography. Furthermore, the system could not be tested in various deployment contexts, including low-resource and mobile ones, due to resource limitations.

6.3 Future Work

Future research will concentrate on improving the system through partnerships with healthcare organizations to get more extensive, real-world information for validation and training. The model will be further optimized for real-time use on mobile platforms, enabling accessibility in remote locations. In order to make the system's predictions more visible and reliable for patients and healthcare practitioners alike, explainable AI (XAI) approaches will also be given top priority. Important future avenues to increase the model's effect include incorporating it into telemedicine systems and extending its prediction capabilities to other relevant anemia diseases.

References

[1] E. DeMaeyer and M. Adiels-Tegman, "The prevalence of anaemia in the world," World Health Statistics Quarterly. Rapport Trimestriel De Statistiques Sanitaires Mondiales, vol. 38, no. 3, pp. 302–316, 1985, Available: https://pubmed.ncbi.nlm.nih.gov/3878044/.

[2] C. R. K. Zuffo, M. M. Osório, C. A. Taconeli, S. T. Schmidt, B. H. C. da Silva, and C. C. B. Almeida, "Prevalence and risk factors of anemia in children," Jornal de Pediatria, vol. 92, no. 4, pp. 353–360, Jul. 2016, doi: https://doi.org/10.1016/j.jped.2015.09.007.

[3] J. T. Vieth and D. R. Lane, "Anemia," Emergency Medicine Clinics of North America, vol. 32, no. 3, pp. 613–628, Aug. 2014, doi: https://doi.org/10.1016/j.emc.2014.04.007

[4] Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition," O'Reilly | Safari, 2019. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

[5] D. J. Feaster, Y. Pan, M. Nelson, J. Sorensen, and L. R. Metsch, "Predicting sexually transmitted infections in sexually transmitted disease clinics in U.S.: A machine learning approach," Drug and Alcohol Dependence, vol. 156, p. e67, Nov. 2015, doi: https://doi.org/10.1016/j.drugalcdep.2015.07.1100.

[6] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, Dec. 2019, doi: https://doi.org/10.1186/s12911-019-1004-8

[7] "Prediction of Diseases in Smart Health Care System using Machine Learning," International Journal of Recent Technology and Engineering, vol. 8, no. 5, pp. 2534–2537, Jan. 2020, doi: https://doi.org/10.35940/ijrte.e6482.018520

[8] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Medical Research Methodology, vol. 19, no. 1, p. 64, 2019, doi: https://doi.org/10.1186/s12874-019-0681-4

[9] Y. M. Alamneh, T. Y. Akalu, A. A. Shiferaw, and A. Atnaf, "Magnitude of anemia and associated factors among children aged 6–59 months at Debre Markos referral hospital, Northwest Ethiopia: a hospital-based cross-sectional study," Italian Journal of Pediatrics, vol. 47, no. 1, Aug. 2021, doi: https://doi.org/10.1186/s13052-021-01123-3

[10] A. Aliyo and A. Jibril, "Anemia and Associated Factors Among Under Five Year Old Children Who Attended Bule Hora General Hospital in West Guji zone, Southern Ethiopia," Journal of Blood Medicine, vol. Volume 13, pp. 395–406, Jul. 2022, doi: https://doi.org/10.2147/jbm.s363876

[11] D. Kebede, F. Getaneh, K. Endalamaw, T. Belay, and A. Fenta, "Prevalence of anemia and its associated factors among under-five age children in Shanan gibe hospital, Southwest Ethiopia," BMC Pediatrics, vol. 21, no. 1, Dec. 2021, doi: https://doi.org/10.1186/s12887-021-03011-5.

[12] B. G. Malako, M. S. Teshome, and T. Belachew, "Anemia and associated factors among children aged 6–23 months in Damot Sore District, Wolaita Zone, South Ethiopia," BMC Hematology, vol. 18, no. 1, Jul. 2018, doi: https://doi.org/10.1186/s12878-018-0108-1

[13] Andersen, Christopher T et al. "Anemia Etiology in Ethiopia: Assessment of Nutritional, Infectious Disease, and Other Risk Factors in a Population-Based Cross-Sectional Survey of Women, Men, and Children." The Journal of nutrition vol. 152,2 (2022): 501-512. doi:10.1093/jn/nxab366

[14] Bamlaku Enawgaw, Yaregal Workineh, S. Tadesse, E. Mekuria, Ayenew Addisu, and

Meaza Genetu, "Prevalence of Anemia and Associated Factors Among Hospitalized Children Attending the University of Gondar Hospital, Northwest Ethiopia," EJIFCC, vol. 30, no. 1, p. 35, Mar. 2019, Accessed: Apr. 30, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6416809/

[15] Y. Gebereselassie, M. BirhanSelassie, T. Menjetta, J. Alemu, and A. Tsegaye, "Magnitude, Severity, and Associated Factors of Anemia among Under-Five Children Attending Hawassa University Teaching and Referral Hospital, Hawassa, Southern Ethiopia, 2016," Anemia, vol. 2020, pp. 1–6, Aug. 2020, doi: https://doi.org/10.1155/2020/7580104

[16] J. Glory and S. Indradevi, "Determinants of Nutritional Status of the Children," Shanlax International Journal of Economics, vol. 12, no. 1, pp. 78–84, Dec. 2023, doi: https://doi.org/10.34293/economics.v12i1.6770

[17] R. T. Gaston, S. Ramroop, and F. Habyarimana, "Joint modelling of malaria and anaemia in children less than five years of age in Malawi," Heliyon, vol. 7, no. 5, p. e06899, May 2021, doi: https://doi.org/10.1016/j.heliyon.2021.e06899

[18] J. C. Ruel-Bergeron et al., "Global Update and Trends of Hidden Hunger, 1995-2011: The Hidden Hunger Index," PLOS ONE, vol. 10, no. 12, p. e0143497, Dec. 2015, doi: https://doi.org/10.1371/journal.pone.0143497

[19] E. McLean, M. Cogswell, I. Egli, D. Wojdyla, and B. de Benoist, "Worldwide prevalence of anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993–2005," Public Health Nutrition, vol. 12, no. 04, p. 444, May 2008, doi: https://doi.org/10.1017/s1368980008002401.

[20] A. Hall et al., "A randomised trial in Mali of the effectiveness of weekly iron supplements given by teachers on the haemoglobin concentrations of schoolchildren," Public Health Nutrition, vol. 5, no. 3, pp. 413–418, Jun. 2002, doi: https://doi.org/10.1079/phn2001327.

[21] C. G. Neumann et al., "Animal Source Foods Improve Dietary Quality, Micronutrient Status, Growth and Cognitive Function in Kenyan School Children: Background, Study Design and Baseline Findings," The Journal of Nutrition, vol. 133, no. 11, pp. 3941S3949S, Nov. 2003, doi: https://doi.org/10.1093/jn/133.11.3941s.

[22] S. Tatala, C. Kihamia, L. Kyungu, and U. Svanbergrhaaa, "Risk factors for anaemia in schoolchildren in Tanga region, Tanzania," Tanzania Journal of Health Research, vol. 10, no. 4, Aug. 2009, doi: https://doi.org/10.4314/thrb.v10i4.45074.

[23] D. J. Roberts, G. Matthews, R. W. Snow, T. Zewotir, and B. Sartorius, "Investigating the spatial variation and risk factors of childhood anaemia in four sub-Saharan African countries," BMC Public Health, vol. 20, no. 1, Jan. 2020, doi: https://doi.org/10.1186/s12889-020-8189-8.

[24] N. J. Kassebaum et al., "A systematic analysis of global anemia burden from 1990 to 2010," Blood, vol. 123, no. 5, pp. 615–24, 2014, doi: https://doi.org/10.1182/blood-2013-06-508325

[25] A. Mainasara et al., "Prevalence of Anaemia among Children Attending Paediatrics Department of UDUTH, Sokoto, North-Western Nigeria," International Blood Research & Reviews, vol. 7, no. 1, pp. 1–10, Jan. 2017, doi: https://doi.org/10.9734/ibrr/2017/29225

[26] A. S. Mohammed Mujib, A. S. Mohammad Mahmud, M. Halder, and C. M. Monirul Hasan, "Study of Hematological Parameters in Children Suffering from Iron Deficiency Anaemia in Chattagram Maa-o-Shishu General Hospital, Chittagong, Bangladesh," Anemia, vol. 2014, pp. 1–10, 2014, doi: https://doi.org/10.1155/2014/503981

[27] Y. Glazer and N. Bilenko, "Effect of iron deficiency and iron deficiency anemia in the first two years of life on cognitive and mental development during childhood," Harefuah, vol. 149, no. 5, pp. 309–14, 335, May 2010, Available:

https://pubmed.ncbi.nlm.nih.gov/20929071/
[28] W. H. Organization, Worldwide prevalence of anaemia 1993-2005 : WHO global database on anaemia. World Health Organization, 2008. Available: https://iris.who.int/handle/10665/43894
[29] Y. Z. Ginzburg and J. Glassberg, "Inflammation, Hemolysis, and Erythropoiesis Lead to Competitive Regulation of Hepcidin and Possibly Systemic Iron Status in Sickle Cell Disease," EBioMedicine, vol. 34, pp. 8–9, Aug. 2018, doi: https://doi.org/10.1016/j.ebiom.2018.07.023.
[30] D. Djokic et al., "Risk factors associated with anemia among Serbian school-age children 7-14 years old: results of the first national health survey," Hippokratia, vol. 14, no. 4, p. 252, 2025, Accessed: Apr. 30, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3031319/
[31] K. C. Bremner, "Pathogenetic factors in experimental bovine oesophagostomosis," Experimental Parasitology, vol. 24, no. 2, pp. 184–193, Apr. 1969, doi: https://doi.org/10.1016/0014-4894(69)90156-8
[32] H. Alaofè, J. Zee, R. Dossa, and H. T. O'Brien, "Education and Improved Iron Intakes for Treatment of Mild Iron-Deficiency Anemia in Adolescent Girls in Southern Benin," Food and Nutrition Bulletin, vol. 30, no. 1, pp. 24–36, Mar. 2009, doi: https://doi.org/10.1177/156482650903000103
[33] M. Hashizume et al., "Anaemia in relation to low bioavailability of dietary iron among school-aged children in the Aral Sea region, Kazakhstan," International Journal of Food Sciences and Nutrition, vol. 55, no. 1, pp. 37–43, Feb. 2004, doi: https://doi.org/10.1080/09637480310001642466
[34] J. K. Kikafunda, F. B. Lukwago, and F. Turyashemererwa, "Anaemia and associated factors among under-fives and their mothers in Bushenyi district, Western Uganda," Public Health Nutrition, vol. 12, no. 12, pp. 2302–2308, Apr. 2009, doi: https://doi.org/10.1017/s1368980009005333
[35] Y. Lufungulo Bahati, J. Delanghe, G. Bisimwa Balaluka, A. Sadiki Kishabongo, and J. Philippé, "Asymptomatic Submicroscopic Plasmodium Infection Is Highly Prevalent and Is Associated with Anemia in Children Younger than 5 Years in South Kivu/Democratic Republic of Congo," The American Journal of Tropical Medicine and Hygiene, vol. 102, no. 5, pp. 1048–1055, May 2020, doi: https://doi.org/10.4269/ajtmh.19-0878
[36] K. C. Bremner, "Pathogenetic factors in experimental bovine oesophagostomosis," Experimental Parasitology, vol. 25, pp. 382–394, Jan. 1969, doi: https://doi.org/10.1016/0014-4894(69)90085-x
[37] Abdulaziz Kebede Kassaw, A. Yimer, W. Abey, Tibebu Legesse Molla, and Alemu Birara Zemariam, "The application of machine learning approaches to determine the predictors of anemia among under five children in Ethiopia," Scientific reports, vol. 13, no. 1, Dec. 2023, doi: https://doi.org/10.1038/s41598-023-50128-x.
[38] A. B. Zemariam et al., "Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia," Scientific Reports, vol. 14, no. 1, p. 9080, Apr. 2024, doi: https://doi.org/10.1038/s41598-024-60027-4
[39] J. R. Khan, S. Chowdhury, H. Islam, and E. Raheem, "Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh," Journal of Data Science, vol. 17, no. 1, pp. 195–218, Feb. 2021, doi: https://doi.org/10.6339/jds.201901_17(1).0009
[40] A. Jiran Meitei, A. Saini, Bibhuti Bhusan Mohapatra, and Kh Jitenkumar Singh, "Predicting child anaemia in the North-Eastern states of India: a machine learning approach," International Journal of Systems Assurance Engineering and Management, vol. 13, no. 6, pp. 2949–2962, Sep. 2022, doi: https://doi.org/10.1007/s13198-022-01765-4

[41]H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, “Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment,” Medicina, vol. 56, no. 9, p. 455, Sep. 2020, doi: https://doi.org/10.3390/medicina56090455

check.docx

ORIGINALITY REPORT

18%	10%	13%	9%
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	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	1%
4	Submitted to Universiti Teknologi Malaysia Student Paper	1%
5	Ali Yimer, Hassen Ahmed Yesuf, Sada Ahmed, Alemu Birara Zemariam et al. "Optimizing machine learning models for predicting anemia among under-five children in Ethiopia: insights from Ethiopian demographic and health survey data", BMC Pediatrics, 2025 Publication	1%
6	Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advances in Networks, Intelligence and Computing - International Conference on Networks, Intelligence and Computing (ICONIC-2023)", CRC Press, 2024	1%