THALASSEMIA DETECTION USING CBC REPORT THROUGH MACHINE LEARNING

Muhammad Rafeh Latif
01-241221-005

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Software Engineering)

Department of Software Engineering

BAHRIA UNIVERSITY ISLAMABAD

April 2024

APPROVAL FOR EXAMINATION

Scholar's Name: Muhammad Rafeh Latif Registration No. 01-241221-005

Program of Study: MS (Software Engineering)

Thesis Title: Thalassemia Detection Using CBC Report through Machine Learning

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor's

Signature:____

Date:____

Name:____

AUTHOR'S DECLARATION

I, Muhammad Rafeh Latif hereby state that my MS thesis titled "Thalassemia Detection Using CBC Report through Machine Learning" is my own workand has not been submitted previously by me for taking any degree from this university Bahria University Islamabad or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of scholar: Muhammad Rafeh Latif (01-241221-005)

Date: ____________________________

PLAGIARISM UNDERTAKING

I, Muhammad Rafeh Latif, solemnly declare that research work presented in the thesis titled "Thalassemia Detection Using CBC Report through Machine Learning" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and thatcomplete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the University has the right to publish my name on the HEC/University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar / Author's Sign: ________________________

Name of the Scholar: Muhammad Rafeh Latif (01-241221-005)

DEDICATION

To my beloved mother and father

ACKNOWLEDGEMENT

I would start by thanking ALLAH Almighty, with gratitude for giving me strengthin every aspect of life and helping me in this thesis as well.

I wish to express my sincere appreciation to my thesis supervisor, Dr. Joddat Fatima and Co-Supervisor Dr. Madiha Khalid, for her support, guidance, and valuable feedback throughout this thesis. Her expertise and encouragement have been instrumental in shaping my research and helping me to overcome the challenges that I faced. I would like to express my sincere appreciation to Dr. Amina Risalat and Dr. Hafsa Hashmi for her invaluable contribution to annotating and labeling the dataset.

I want to acknowledge my parents, who supported me and. This thesis work is dedicated to my parents, who have been a constant source of support during the challenges of life and prayed for my success. I am truly thankful for the support in every aspect whether that is financial, emotional, or mental. I would like to thank my brothers Maj. Muhammad Bilal latif and Capt. Muhammad Mueez Latif for their constant support that kept me motivated.

Lastly, I would like to acknowledge the contributions of all the participants who took part in my study, without whom this research would not have been possible.

ABSTRACT

Thalassemia, which is a hereditary blood disorder that impacts millions all over the world, require early identifying of carriers for reduces its prevalence and the associated complications. Carriers, even though they often don't show symptoms themselves, may pass the genetic mutation to their children, potentially causes thalassemia in future generations. Detecting carriers at an early stage allowing for important interventions like genetic counseling, family planning, and education on thalassemia risk.

In recent time, machine learning algorithms have become valuable tools in healthcare, capable to analyzing large datasets for predictive insights. This thesis aims for exploring the use of machine learning to identifying thalassemia carriers based on Complete Blood Count (CBC) results. The project involves data collections, preprocessing, feature selections, and model training. Specifically, we prioritize features which are most relevant to thalassemia detection, including utilizing the Mentzer index. Our approach uses the Random Forest Model for the detecting of thalassemia carriers.

Model performance will be evaluated rigorously using appropriate metrics for reliability and accuracy. The outcomes of this studies hold the potentials to significantly contribute to the field of thalassemia diagnosis's and managements. Through developing an accurate and efficient machine learning model based on CBC results, clinicians and researchers will gain valuable insights that could improves patient outcomes and inform future researches and treatment strategies. That research could ultimately lead to better-targeted interventions and personalized cares for individuals affected by thalassemia.

Keywords: Genetic counseling, Complete Blood Count (CBC), feature selection, Mentzer index, Random Forest Model,patient outcomes, personalized care.

TABLE OF CONTENTS

CHAPTER **TITLE** **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WBC - White Blood Cell count, reflecting the immune function.

LYMp - The proportion of Lymphocytes, which are a type of white blood cell that are involved in the immune system.

MIDp - The percentage of mid-sized cells in the blood.

NEUTp - The percentage of Neutrophils, which is a type of white blood cell essential for fighting infections.

LYMn - Absolute count of Lymphocytes.

MIDn - Absolute number of Mid-sized cells.

RBC - Red Blood Cell count, which transports Oxygen to tissues.

HGB - The concentration of hemoglobin that carries oxygen in the blood.

HCT - Hematocrit, the volume fraction of red blood cells in the blood.

MCV - The Mean Corpuscular Volume, which is the average volume of red blood cells.

MCHC - MCV is the concentration of hemoglobin in a given volume of red blood cells.

RDWSD - Red Cell Distribution Width - Standard Deviation, which is an index of variability of red blood cell size.

RDWCV - Red Cell Distribution Width - Coefficient of Variation, another coefficient of variation indicating variability in red blood cell size.

PLT - Platelet count is necessary for blood clotting.

MPV - Mean Platelet Volume, the average size of platelets.

PDW - Platelet Distribution Width (PDW), an indicator of platelet size variation.

PCT - Plateletcrit or the volume percentage of platelets in the blood.

PLCR - Platelet-Lymphocyte Ratio, a calculated ratio that serves as an inflammatory marker.

CHAPTER 1

INTRODUCTION

1.1. Background

Thalassemia is a hereditary blood disorder characterized by abnormal hemoglobin production, which can result in insufficient red blood cell formation and subsequent anemia. It poses a very significant global health concern, affecting millions worldwide. In many different countries, including Pakistan, there are reports showing that blood disorders like thalassemia are on the rise.

Table 1-1 Comprehensive Overview of Thalassemia Statistics in Pakistan

Parameter	**Value**
Estimated Population of Pakistan	225,633,392 (225 million)
Frequency of β-thalassemia trait	5.0-7.0%
Estimated number of carriers	>10 million
Number of children diagnosed with β-thal major/year	~5,000

The above table indicates that Pakistan, which is estimated to have 225,633,392 (225 million people) population, is facing a huge disease burden of β-thalassemia. The frequency of β-thalassemia traits is between 5.0 - 7.0%; 10 million or more carriers have been estimated across the nation. Annually, 5,000 kids are officially diagnosed with major β-thal more emphatically illustrating the size of the problem. This poses a great challenge to health care system that mostly depends on blood transfusions to control the disease.

The high rate of thalassemia is often a result of inadequate genetic counseling and prenatal screening [1]. Limited access to these services may aggravate the cycling of thalassemia from one generation to the next [2], emphasizing the importance of general population genetic education and screening programs. The lack of proper intervention may lead to increased workload for affected individuals and families [3], as they attempt to manage the disease as well as its effect on quality-of-life.

Thalassemia is usually only identified after birth, with noticeable symptoms appearing. However, delaying diagnosis until a child is about two years old can cause serious risks as urgent treatment is vital [4]. With regards to its clinical course, thalassemia encompasses the forms of thalassemia major, thalassemia minor or trait, and thalassemia intermedia [5]. The way the condition is classified represents the severity and clinical presentations, with thalassemia major serving as the most severe form that requires regular blood transfusions and complex medical treatment. This form of the disease, known as thalassemia minor or trait, is usually characterized by milder symptoms [6] and can be found unintentionally during routine blood tests. Thalassemia intermedia is an intermediate type [7], high in symptoms but low in manifestations and may necessitate regular transfusions or other supportive measures. Thalassemia major calls for lifelong blood transfusions, while thalassemia intermedia may necessitate less frequent transfusions. Conversely, thalassemia minor patients do not need transfusions but carry the thalassemia trait [8].

Early discovery and precise diagnosis of thalassemia are pivotal for prompt action and effective management. It's crucial to emphasize the considerable risk posed to offspring born to thalassemia carriers. When both parents are carriers, there is a 25% chance that their child will inherit thalassemia major, requiring lifelong management, frequent blood transfusions, exhaustive medical care, and significant emotional and financial strain on the family.

In countries like Pakistan, where the average number of children per woman is four, the impact of thalassemia is notably pronounced. Carrier couples confront the challenging reality that, on average, one of their offspring will be born with thalassemia. Therefore, there is an urgent need for increased awareness about thalassemia prevention, genetic counseling, and prenatal testing. Providing accurate information and promoting family planning options can empower individuals and couples to make informed decisions about family size, thus decreasing the incidence of thalassemia in future generations.

In recent years, machine learning algorithms have emerged as potent tools in healthcare for automating and enhancing disease detection and diagnosis. These algorithms can analyze large datasets, identify patterns, and make predictions based on input data, such as Complete Blood Count (CBC) results. Leveraging machine learning techniques can boost thalassemia detection, enabling early identification and providing valuable insights for clinicians and researchers.

Research hints that most thalassemia patients have a shorter lifespan, approximately around 25 to 30 years. However, with appropriate medical care and support, their lifespan can be extended to approximately around 60 years. Proper medication plays a crucial role in increasing the life expectancy of thalassemia patients. Thalassemia can also cause abnormalities in bone structure, further impacting patient health. Heart-related issues, especially in beta thalassemia major cases, can be life-threatening and may arise before reaching the age of 30. Thus, early prediction and treatment are key to ensuring timely intervention and the necessary care to enhance patient quality of life [9].

Aside from medical challenges, thalassemia patients in Pakistan also face social stigma and psychological burdens due to limited awareness and misconceptions about the condition. This can lead to discrimination and social exclusion, significantly affecting overall well-being and quality of life. Therefore, efforts should be made to educate the public, reduce stigma, and promote inclusivity for thalassemia patients.

Furthermore, addressing the economic burden faced by thalassemia patients and their families is crucial. The costs associated with regular blood transfusions, medications, and supportive care can be substantial, straining household finances and impeding access to necessary treatments. Government initiatives offering financial assistance, health insurance coverage, and support for thalassemia treatment are essential to tackle these economic challenges.

Premarital screening and genetic counseling are vital components of thalassemia prevention in Pakistan. Carrier couples have a higher likelihood of having children with thalassemia major, underscoring the significance of comprehensive premarital screening programs and genetic counseling services. These services aid individuals in making informed decisions about their marriage partners, lessening the risk of transmitting thalassemia to future generations. Raising awareness about genetic testing and encouraging screening before marriage are fundamental steps in preventing thalassemia.

1.2. Impact of thalassemia:

Thalassemia, it's a blood disorder. It matters a lot to doctors and health workers. Knowing how big thalassemia is for us can help a lot. We can catch it early, tell people about it, and take good care of those who have it. The sickness touches more than the ones who have it. It changes the lives of families, friends, and everyone else. By grasping its effects, we see it's more than a health issue. It's also about our lives, money, and minds.

Thalassemia, a blood illness that's in our genes, matters big time in health and care. Seeing its big role is a must. This way, we can spread the news, catch it sooner, and care better for those with it. Thalassemia isn't just a health issue, it has a big financial impact too. The expenses it brings. for example, doctor appointments, tests, and lifelong treatments, can really hurt a family's finances. When you add in indirect costs, like being less productive and job and schooling issues, the burden grows. There's also a larger social issue. Many people don't understand thalassemia. They may have wrong ideas and even discriminate against people with the disease and their loved ones. It's important to work on changing these views and getting better acceptance. This way, we can create a more supportive environment for those affected by thalassemia.

Thalassemia's existence proves the need for genetics-based guidance and plans for families. By teaching people about thalassemia's inherited traits and advising them on family growth, we equip them to decide wisely about their reproductive health. With pre-pregnancy checks and counseling, couples can figure out if they're likely to pass the disease to their babies. This aids in future prevention and handling of thalassemia.

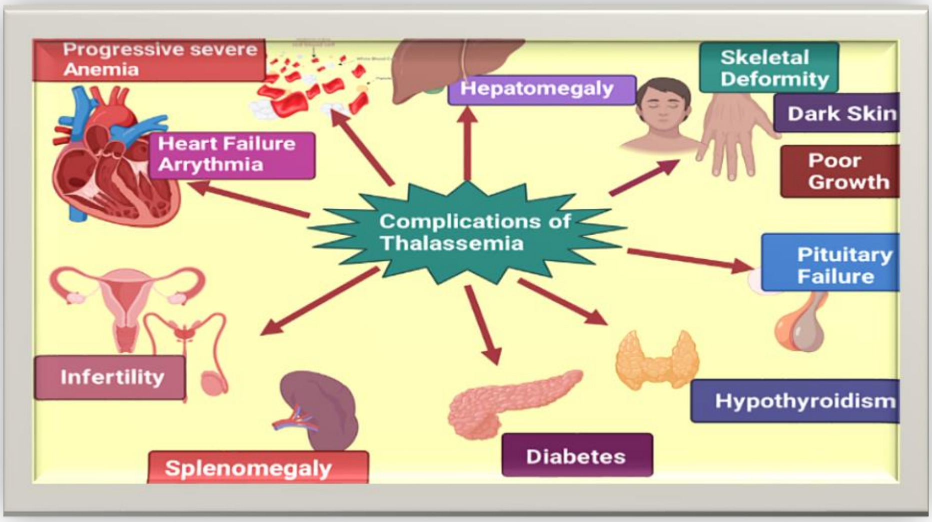


Figure 1-1 (Haque, 2022) The Impact of Thalassemia: Understanding its Broad and Significant Effects

To sum up, thalassemia is more than a diagnosis or treatment. It touches individual health, family peace, society views and even public health plans. Knowing and tackling the many issues that thalassemia brings is key to improving affectee's life quality and reducing its far-reaching impact on society.

1.3. Identifying the Problem Area

This section tackles the challenge of using machine learning to spot thalassemia. Three main issues block the path to accurate detection. The first is the limited use of machine learning in using Complete Blood Count (CBC) results. Next, we don't know which CBC parameters hint at thalassemia. Finally, few have studied how to pick the right features for thalassemia detection with machine learning.

1.4. Research Motivation

The motivation behind this research is that we need to fill an existing gap between thalassemia diagnosis and detection, particularly using the highly accurate machine learning techniques that rely on the essential results of Complete Blood Count (CBC). Even though CBC is a standard procedure in clinical practice, we haven't fully explored the opportunities it brings to improve the understanding of thalassemia. We can substantially increase the sensitivity and specificity of the diagnosis by determining certain CBC parameters or their joints that reliably indicate the presence of thalassemia.

Moreover, research on characteristic choosing ways for CBC reports in thalassemia diagnosis enhanced by machine learning algorithms is not yet available. CBC data can be enhanced by applying feature extraction and selection methods, which will, in turn, boost the precision of diagnostic models. As a result, our study is intended to fill these gaps and add to the development of more accurate and specific diagnostic techniques for thalassemia.

1.5. Problem Statement

Develop a predictive model to identify individuals who may be carriers of thalassemia based on their Complete Blood Count (CBC) reports. It is a genetic blood disorder characterized by abnormal hemoglobin production, leading to anemia and other health complications.

1.6. Research Questions

RQ 1: How to do data annotation and labeling?

RQ 2: How does ML approaches and features in the thalassemia detection model impact diagnostic accuracy

RQ 3: If the parameters like Mentzer Index, using MCV and RDW parameters, contribute to the thalassemia diagnosis?

1.7. Thesis Objective

The aim of this research is to transform thalassemia diagnosis and treatment techniques, and handle the existing problems within the health services. With healthcare emphasizing precision and reliability in diagnosis, the existing shortage of specialized medical professionals and the high costs associated with testing amplify the necessity for accurate detection methodologies. Moreover, there is an increase in cases of thalassemia as a result of consanguineous marriages which makes it necessary to have efficient detection strategies.

This study attempts to blaze new trails such as automated detection models and affordable diagnostic solutions to mitigate these challenges. The broader objective is therefore on raising patient care standards, reducing healthcare disparities and easing burdens on individuals as well as health systems managing patients suffering from Thalassemia. In doing so, we hope to redefine paradigms of Thalassemia diagnosis and treatment that shall lead ultimately to better patient quality of life together with healthcare outcome.

1.8. Contribution to the field and future directions

In the course of the study the Mentzer Index and the Random Forest models are displayed for improving thalassemia diagnosis. These studies have remarkably resulted in obtaining new diagnosticians to thalassemia using the approaches. The application of Mentzer Index, together with the help of machine learning methodology, brings a new level of accuracy and productivity for diagnostics in this field. Although the approach is limited by the specific datasets used and the need to generalize it for more applications, this method still holds significant potential in driving future data analysis. In the future research, collections of bigger as well as more diverse datasets should be taken into account. More attention should be also given to the machines that integrate machine learning with legacy methods. To achieve this, we will target these obstacles and, in the process, assist in the diagnosis of thalassemia that will in turn enhance the patient management operations.

1.9. Outline of this thesis

The organization of this paper is as follows: **Chapter 1** “Introduction” section includes the introduction of the study. **Chapter 2** “Literature Review” section summarizes the related works on embedding techniques. **Chapter 3** “Research Methodology” section explains the whole methodology of this study. **Chapter 4** “Classification Techniques” section shows different techniques used. And **Chapter 5** “Results and Evaluation” section shows the results obtained. Finally, **Chapter6** “Conclusion” section highlights the key findings and conclusion of this study.

CHAPTER 2

LITERATURE REVIEW

This part of the research paper is the introduction and presents about the earlier research that has been done in this field. In a literature review, conducted for this research, we have established a number of studies which have applied machine learning and artificial intelligence (AI) methods for the purposes of diagnostics and detection of various types of blood-related issues.

2.1. Application of Machine Learning Models on Complete Blood Count (CBC) Reports

Fu, Y.-K. et al. In 2021, [11], in-depth research was done on the informational capability of AI models mainly considering SVM to monitor essential blood parameters including Hb, RBC, MCV, MCH, MCHC, and RDW to distinguish between positive and negative patients. The work has revealed the high accuracy result of 95% on the testing dataset by using ROC curve analysis in us to measure model performance.

Besides that, a survey in 2007[12] was done. Yousefian, Fatemeh et al testing the ability of detecting blood figures like RBC, HGB, MCV, and HTC using SVM and KNN, the artificial intelligence models. In this study, the perception achieved was 88.89% for SVM and 85.19% for KNN, which have proved that AI is able to identify some condition by the features of blood.

Another study in medical science 2020 [13] E. R. Susanto et al addressed the implementation of the AI models such as Fuzzy C Mean on the characteristics of blood components like HGB, MCV, and MCH. The index metrics of precision, recall, and accuracy scored very high, with precision at 99%, recall at 96.6%, while accuracy was at 96.5%.

Additionally, yet another conference dated 2022 [14] A. Devanathet al portrayed the recognition of diverse blood elements through the use of AI models among them LR, KNN, SVM, DT, and NB. These models achieved an impressive accuracy of 97% and also did very well in the measures of precision, recall, and F1 score, which demonstrated their robust performance in the classification task of predicting or not a certain condition.

A recent study published in 2023 [15] Saleem M et al investigated feature selection techniques for thalassemia detection, analyzing blood parameters such as MCV, PCV, MCH, MCHC, RDW, PLT, TLC, and HGB. Machine learning algorithms including KNN, DT, GBC, LR, AdaBoost, XGB, RF, LGBM, and SVM were evaluated on a dataset comprising records from 6000 patients. Among these algorithms, SVM emerged as the top performer with an accuracy, recall, and F1-score all above 90%, emphasizing the efficacy of machine learning approaches in improving thalassemia detection.

Collectively, these studies underscore the growing role of AI and machine learning in analyzing blood parameters for accurate detection and diagnosis of various medical conditions, highlighting the significance of feature selection and algorithmic performance evaluation in achieving reliable results.

Table 2-1 Reviewed Research Work on Machine Learning Techniques Applied to Complete Blood Count (CBC) Reports

Ref.	Year	Techniques / Method / Model	Results
[11]	2021	SVM (Support Vector Machine)	SVM performed best with (ROC=95%)
[12]	2017	SVM (Support Vector Machine), KNN (K-Nearest Neighbour)	SVM performed best with (Accuracy=88.8%) KNN performed best with (Accuracy=85.1%)

[13]	2020	Not a AI Model something similar to Fuzzy C Mean	Manual testing(Using medical specialist) performed best with (Precision=99%
			Recall=96%
			Accuracy=96%)

[14]	2022	LR
		KNN
		SVM
		DT(Decision Tree)
		NB (Naive Baye's)
			Algorithm performed best with (Accuracy=97%
			Precision=87%
			Recall=97%
			F1-score=93%)

[15]	2023	K-Nearest Neighbors (KNN)
		Decision Trees (DT)
		Gradient Boosting Classifier (GBC)
		Linear Regression (LR)
		AdaBoost, Extreme Gradient Boosting (XGB)
		Random Forest (RF)
		Light Gradient Boosting Machine (LGBM)
		Support Vector Machine (SVM)
			Algorithm performed best with (Accuracy=93%
			Recall=93%
			F1-score=92%)

2.2. Utility of the Mentzer Index in Thalassemia Screening and Diagnosis

Muhammad Idrees et al did a study in 2023 [16] on the Mentzer Index that was conducted at the Hematology unit of Hayatabad Medical Complex was designed to determine the sensitivity and specificity of the Mentzer Index when it comes to differentiating between beta thalassemia minor and iron deficiency anemia. This cross-sectional study of 860 cases with hemoglobin concentration below 11 g/dL, demonstrated that the index, calculated as the middle corpuscular volume (MCV) divided by the count of the red blood cells (RBC), was a potent diagnostic tool. The examine discovered that the Mentzer Index displayed high sensitivity and specificity, with 56.86% of patients diagnosed with iron deficiency anemia and 43.14% with a higher suspicion of beta thalassemia based on the index parameters.

These results demonstrate the effectiveness of the Mentzer Index as a simple and accurate differential diagnosis between thalassemia minor and iron deficiency anemia and, therefore, aiding in the early diagnosis and confirmation of suspected cases by Hb Electrophoresis.

W. Siswandari et al. in 2019 [17] investigated the diagnostic capacity of the Mentzer Index in the prediction of beta-thalassemia carriers compared to Hb electrophoresis. Thalassemia diagnosis is usually accomplished by PCR or Hb electrophoresis, which are specialized technologies that may not be readily available in every hospital. The study obtained Mentzer Index odds ratio (OR) of 2.4 (0.5 - 11.5, CI95%) among 37 anemia patients at Prof. Dr Margono Soekarjo Regional Public Hospital. The sensitivity (Sn) is 0.36 while the specificity (Sp) is 0.81. Also, the positive predictive value (PPV) was 0.44, and the negative predictive value (NPV) was 0.75. It implies that the Mentzer Index could provide a more straightforward and readily available alternative to the costly and specialized methods. Thus the index can serve as a useful predictor of beta-thalassemia carrier diagnosis.

The research led by Dr. Shubhi Saxena et al [18] in 2020 was a retrospective observational study with 1236 patients to assess the diagnostic accuracy of the Mentzer index to distinguish between iron deficiency anemia and β thalassemia trait compared to HPLC. The results showed that out of 741 patients, 59.9% had iron deficiency anemia and 40.1% of the patients had β thalassemia trait. The Mentzer index was shown to be 89.0% sensitive and 87.9% specific for the detection of β thalassemia trait with the positive predictive value of 83.2% and the negative predictive value of 92.3%. The Youden's Index was 76.9%. This discovery emphasizes the usefulness of Mentzer index as screening tool particularly in poor countries like India where HPLC equipments may not be readily available. A HPLC validation is advisory in cases of doubt.

Table 2-2 Exploring the Utility and Efficacy of the Mentzer Index as a Diagnostic Method:A Comprehensive Review of Research Findings

Ref.	**Year**	**Techniques / Method**	**Results**
[16]	2023	Mentzer Index	Sensitivity:high Specificity:high Iron Deficiency Anemia 56.8% B-thalasemia 43.14%
[17]	2019	Mentzer Index	Sensitivity:36% Specificity:81%
[18]	2020	Mentzer Index	Sensitivity:89% Specificity:87.9% Iron Deficiency Anemia 59.9% B-thalasemia 40.1%

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Proposed Methodology

The study employs a systematic and robust research design to explore the research questions and accomplish the objectives. It is done according to the following:

Data Collection: In this research, the dataset comprises CBC reports obtained from Al Zahra Hospital in Iraq, consisting of records from 500 patients. It's important to note that the dataset**.** Initially lacked annotations for thalassemia classification. To address this, we sought Assistance from medical specialists for labeling purposes. Following expert annotation, the dataset was categorized into 350 records indicating the absence of thalassemia and 150 records indicating the presence of thalassemia. This meticulous labeling process ensured the integrity and accuracy of the dataset, enabling us to effectively train and evaluate our classification model.

Feature Selection: Analysis of CBC reports aim at identifying significant features related to thalassemia like MCV, RDW, and newly derived Mentzer index. Statistical methods including correlation analysis are used in selecting informative features that give an all-round evaluation of possible indicators for detection of thalassemia.

Model Development: Machine learning algorithms like decision trees, random forests or support vector machines are employed in coming up with automated thalassemia detection model. Input variables for this stage are those selected from the previous one.

Model Training and Evaluation: A part of the datasets is used in training the designed model while its performance is assessed using appropriate measures such as accuracy, sensitivity, specificity, area under receiver operating characteristic (ROC) curve.

Comparison Results with Existing Features: The performance of the developed model with new feature is compared with existing features of CBC report, including HB, MCV etc. to assess its superiority in terms of accuracy and reliability.

Validation and Generalizability: The developed model is validated using an independent dataset to assess its generalizability and robustness.

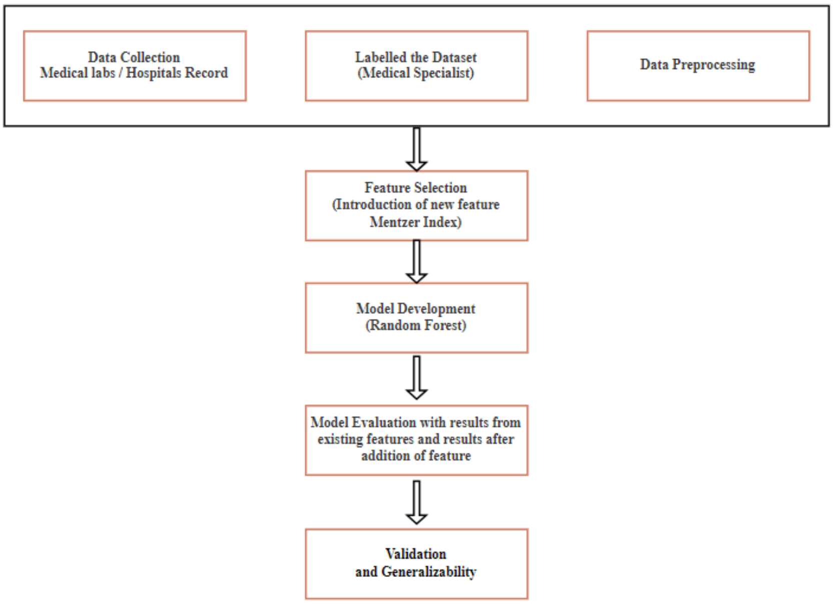


Figure 3-1 Navigating Methodological Approaches: An In-depth Exploration of Methodology Steps

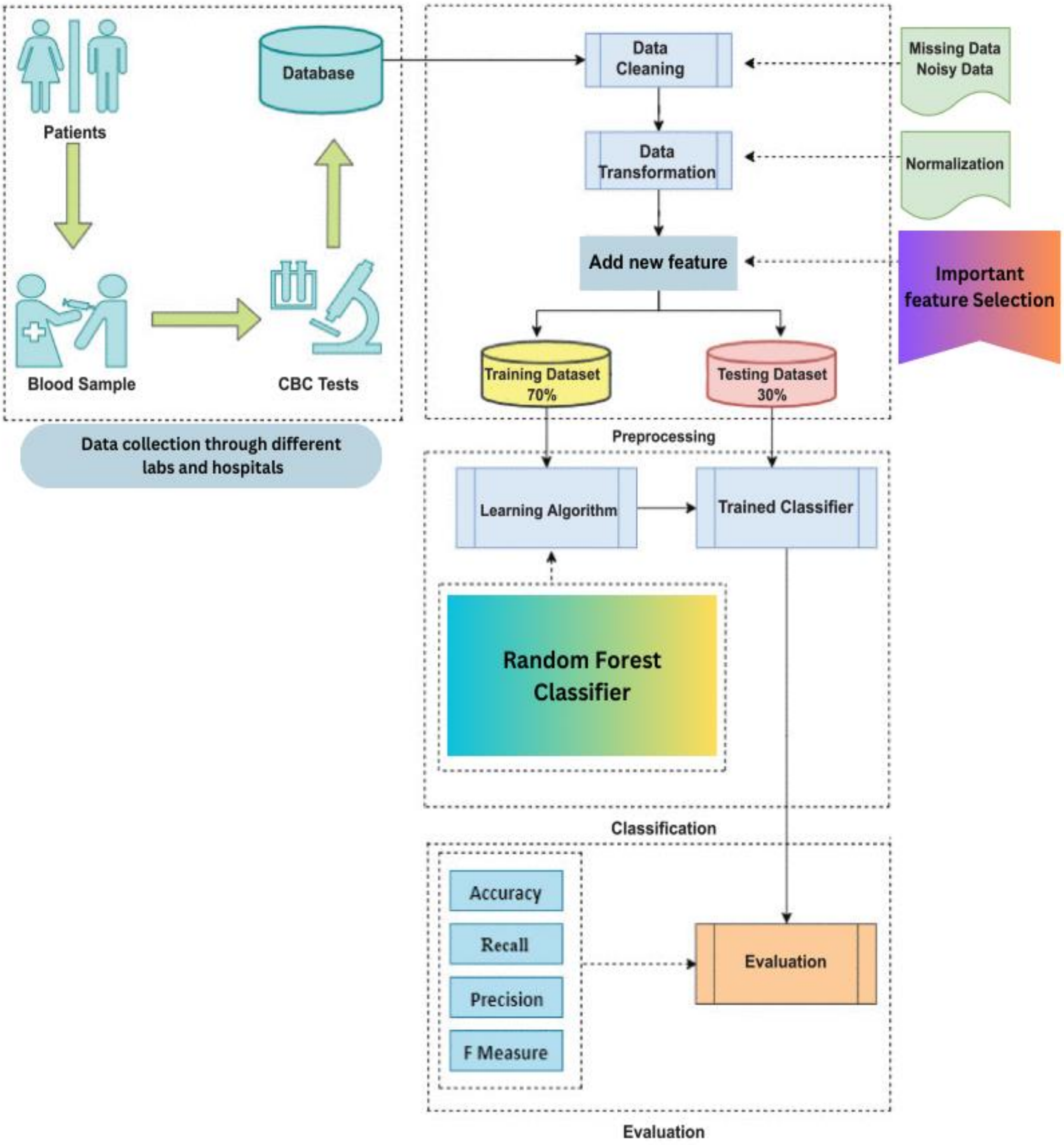


Figure 3-2 Unveiling the Layers: A Comprehensive and In-depth Exploration of Research Methodology

The image above highlights the fact that data extraction starts with the procurement of appropriate data from trustworthy sources the dataset utilized in this research comprises records from Alzahra Hospital Iraq, consisting of data from 500 patient records. The annotation and labeling process has been conducted by medical specialist Dr. Amina Risalat, with verification overseen by Dr. Hafsa Hashmi, both experts from the thalassemia center. Secondly, feature selection is performed by investigating CBC reports to select significant features related to thalassemia, like MCV, RDW, and Mentzer index. The statistical techniques such as correlation analysis assist in identifying the informative features that are holistic and incorporate multiple thalassemia detection marks. Then, machine learning algorithms, including decision trees, random forests, or support vector machines, are implemented in order to create automated thalassemia detection model with input variables that are identified in the previous stage. Model training is done on part of the dataset and the evaluation is performed using measures like accuracy, sensitivity, specificity and the area under the ROC curve. The performance of the model with the new feature is then checked by comparing it with other features of CBC reports like hemoglobin (HB) and MCV, to identify the best one in terms of accuracy and reliability. Finally, the constructed model goes through a validation process with the help of an independent dataset to measure its applicability and strength.

3.2. Data Gathering

In this research, we have assembled a dataset from Alzahra Hospital Iraq consisting of records from 500 patients, the dataset is publicly available at this link “https://data.mendeley.com/datasets/28s2bhdjfd/1”. The annotation and labeling of this dataset were meticulously carried out by medical specialist Dr. Amina Risalat, with verification overseen by Dr. Hafsa Hashmi. This dataset includes approximately 500 Complete Blood Count (CBC) reports from both thalassemia and non-thalassemia patients, reflecting a concerted effort to gather comprehensive data for our research analysis.

We followed the strict guidelines for keeping the data clean of any misinformation and patient’s privacy, all the data was anonymized. We did this for a diverse range of patients in all ages, from both genders. Patients with different clinical background were kept in mind to keep the diversity of our dataset so our findings can be representative.

There are a total of 350 thalassemia patients whereas 150 are non-thalassemia patients. We made sure to keep patients from urban as well as rural areas in the count.

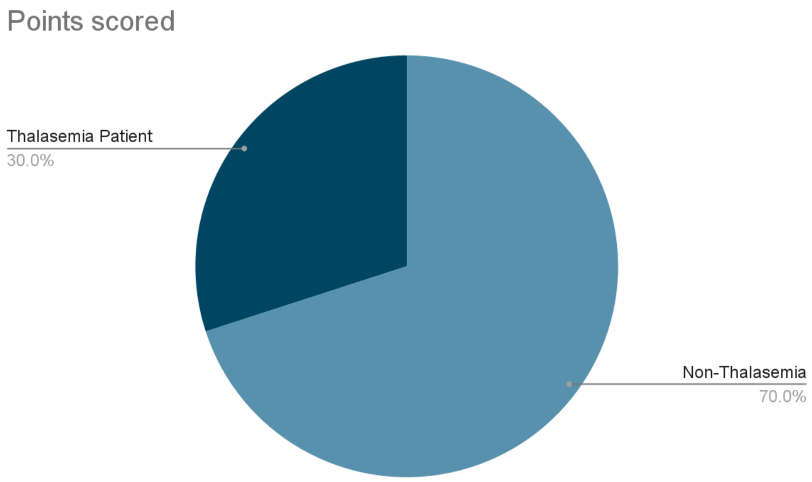


Figure 3-3 Mapping the Patient Landscape: A Detailed Exploration of Patient Distribution

3.2.1 Dataset Variables:

The data set consists of different variables, collected from Complete Blood Count (CBC) reports, except the Mentzer index, a new feature introduced specifically for identifying the thalassemia carriers. All of these variables are essential for the diagnosis of beta-thalassemia and other hematologic diseases.

WBC: White Blood Cell count, reflecting the immune function.

LYMp: The proportion of Lymphocytes, which are a type of white blood cell that are involved in the immune system.

MIDp: The percentage of mid-sized cells in the blood.

NEUTp: The percentage of Neutrophils, which is a type of white blood cell essential for fighting infections.

LYMn: Absolute count of Lymphocytes.

MIDn: Absolute number of Mid-sized cells.

NEUTn: Absolute count of Neutrophils.

RBC: Red Blood Cell count, which transports Oxygen to tissues.

HGB: The concentration of hemoglobin that carries oxygen in the blood.

HCT: Hematocrit, the volume fraction of red blood cells in the blood.

MCV: The Mean Corpuscular Volume, which is the average volume of red blood cells.

MCH: MCH, or the average amount of hemoglobin in a red blood cell.

MCHC: MCV is the concentration of hemoglobin in a given volume of red blood cells.

RDWSD: Red Cell Distribution Width - Standard Deviation, which is an index of variability of red blood cell size.

RDWCV: Red Cell Distribution Width - Coefficient of Variation, another coefficient of variation indicating variability in red blood cell size.

PLT: Platelet count is necessary for blood clotting.

MPV: Mean Platelet Volume, the average size of platelets.

PDW: Platelet Distribution Width (PDW), an indicator of platelet size variation.

PCT: Plateletcrit or the volume percentage of platelets in the blood.

PLCR: Platelet-Lymphocyte Ratio, a calculated ratio that serves as an inflammatory marker.

MI: Mentzer Index, calculated by division of MCV/RBC

These variables collectively provide valuable insights into the hematological profile of individuals, aiding clinicians in diagnosing and monitoring various blood disorders, including thalassemia. The addition of the Mentzer index further enhances the diagnostic capabilities, enabling more accurate identification of individuals with thalassemia carrier status, thereby facilitating timely intervention and management.

3.3. Code Configuration

First the Python and an Integrated development environment (IDE) should be downloaded and installed. Python can be found on the official website and you should select the most recent stable version that is compatible with your operating system of choice. After Python is installed, the next step is to choose your IDE (Integrated Development Environment) that will suit your needs and preferences. Top picks here are PyCharm, Jupyter Notebook, Visual Studio Code, and Google Colab. By having these platforms, programmers can use a user-friendly UI for writing, running, and debugging the code, which increases the development process efficiency and simplifies it. Then the development environment could be set and required libraries and modules imported, as expressed below in the provided code snippet, to allow for data analysis and machine learning activities.

We used a simple yet systematic method for the initial handling of the data as this was acquired. With the help of 'files.upload()' function of Google Colab, we uploaded the

Excel file containing our data to our drive. Then we applied pandas to read the uploaded Excel file into a DataFrame. To be sure that the file is correctly handled, we pulled the file name using the keys in the uploaded dictionary. A preliminary view of dataset's first rows was presented using the 'head() method', which helped to understand the structure and content immediately.

3.4. Data Preprocessing

We developed crucial techniques to arrange the dataset for modelling and analysis. First of all, we did a careful examination for missing values within the dataset with 'isnull().sum()' method proving dataset completeness and reliability. Then a mapping technique was used, coding 'Yes' as 1 and 'No' as 0 to the categorical values of the 'Carrier' column. Such a transformation aids in later analysis and modeling, helping to incorporate the categorical data into machine learning. It was then ensured that the unique values in the 'Carrier' column were printed out in order to check the final success of the mapping, to ensure both consistency and accuracy in data cleaning. Such preparation stage is a prerequisite for later modelling of results ensuring high quality of research.

3.5. Feature Importance Analysis:

The table below indicates the contribution of each feature to the accuracy of the model as a result of the feature importance analysis. Facets like Red Blood Cell Count (RBC), Mentzer Index (MI), and Red Cell Distribution Width - SD (RDWSD) are the most important variables in the discrimination of thalassemia patients from non-thalassemia individuals. Among others, these features offer useful information about the hematological changes linked to thalassemia, which reinforces their significance in the diagnostic algorithms.

Table 3-1 Unveiling Significance: A Comprehensive Examination of Feature Importance Analysis

Number	Feature	Importance
1	RBC	0.082231
2	MI	0.080460
3	RDWSD	0.079362
4	MPV	0.073993
5	PLT	0.062940
6	RDWCV	0.059381
7	PLCR	0.049159
8	MCHC	0.048593
9	MCV	0.043527
10	HCT	0.042994
11	PCT	0.042193
12	MCH	0.039369
13	PDW	0.038219
14	HGB	0.035785
15	LYMn	0.035477
16	MIDn	0.034773
17	MIDp	0.033174
18	NEUTp	0.032531
19	LYMp	0.030442
20	NEUTn	0.028590
21	WBC	0.026807

RBC (Red Blood Cell Count)

In the model, red blood cell count has a major impact on predictions, as the importance score is 0.082231. It becomes clear that fluctuations in RBC count are key for identifying thalassemia patients from those who do not have this disease.

MI (Mentzer Index)

The Mentzer Index (MI) also makes a huge contribution to the model and it has a coefficient of 0.080460. This indicates that MI, which is a calculated parameter from RBC and MCV values, does a lot to help detecting thalassemia.

RDWSD (Red Cell Distribution - SD Width)

RDWSD (Standard Deviation of Red Cell Distribution Width) has a notable part, getting an importance value of 0.079362. Varying values of the RDWSD parameter can give an indication about the abnormalities in red blood cell size distribution, which help in diagnosing thalassemia.

MPV (Mean Platelet Volume)

MPV, which is the mean platelet volume, also is a crucial part of the model with a score of 0.073993. MPV represents a platelet size’s average in the blood and it is linked with thalassemia and other hematologic disorders.

PLT (Platelet Count)

Platelet (PLT) is one of the most important features of the model with an importance weighted score of 0.062940. A decrease in platelet count could indicate the presence of hematological disorders and hence the platelet count is a critical parameter in thalassemia screening.

RDW-CV (Red Cell Distribution Width - CV)

Red Cell Distribution Width - Coefficient of Variation (RDWCV) features in the model as well, ranked the third with the score of 0.059381. Differences in RDW, a metric of red blood cell size variability, arise from heterogeneity in red blood cells and could be indicative of thalassemia.

PLCR (Platelet-Lymphocyte Count Ratio)

Platelet-Lymphocyte Count Ratio (PLCR) contributes to the model significantly, and its weight is 0.049159. Changes in PLCR values may be an indication of systemic inflammatory disorders or blood related disorders such as thalassemia.

MCHC (mean corpuscular hemoglobin concentration)

One of the important parameters (MCHC) has a score of 0.048593. MCHC values are an average of the concentration of hemoglobin in red blood cells, thus allowing for the detection of thalassemia-specific alterations.

MCV (Mean Corpuscular Volume)

Mean Corpuscular Volume (MCV) is an indicator in the model, and its value is 0.043527. MCV determines the average cell volume of red blood cells and is increased in thalassemia patients.

HCT (Hematocrit)

Hematocrit (HCT) is a cardinal feature in the model with a rating of 0.042994. HCT gives the proportion of the blood volume occupied by red blood cells and is a marker of blood abnormalities which are the hallmark of thalassemia.

PCT (Plateletcrit)

Plateletcrit (PCT) plays a vital role in model's prediction with an importance score of 0.042193. PCT serves as an indicator of the level of platelets and may be useful in the diagnosis of thalassemia.

MCH (Mean Corpuscular Hemoglobin)

Mean haemoglobin concentration (MCH) has a role to play with a score of 0.039369 in the model. MCH acts as an indicator of the mean hemoglobin concentration in red blood cells, thus allowing for the detection of thalassemia.

PDW (Platelet Distribution Width)

Platelet Distribution Width (PDW) also takes an attribute in the model with a score of 0.038219. Differences in PDW may signify platelet abnormalities linked with thalassemia and hematological disorders.

HGB (Hemoglobin)

Hemoglobin (HGB) is one of the main factors that affect the model prediction with a score of 0.035785. Normal HGB levels play a key role in oxygen transportation and may signify the complications of thalassemia.

LYMn

Lymphocyte Count (LYMn) score equals 0.035477. Lymphocyte (LYMn) changes suggest immune system problems that can be associated with thalassemia.

MIDn (Mid-Range Neutrophil Count - Absolute)

The model contains MIDn, which is an important Mid-Range Neutrophil Count - Absolute (MIDn) with the score 0.034773. MIDns are markers of changes in the neutrophil count, which may be linked to the presence of systemic inflammation or infections in thalassemia patients.

MIDp (Mid-Range Neutrophil Count - Percentage)

Mid-level Neutrophil Count - Percent (MIDp) is one of the factors considered, having a score of 0.033174. MIDp values give information about neutrophils' distribution and function that may be affected in patients with thalassemia.

NEUTp (Neutrophil Count - Percentage)

Neutrophil Count - Percentage (NEUTp) is an essential feature of the model with a score equal to 0.032531. NEUTp stands for the amount of neutrophils in blood, which can be considered as a sign of inflammation or infection in thalassemia patients.

LYM% (Lymphocyte Count - Percentage)

Lymphocyte Count - Percentage (LYMp) is as significant as 0.030442 in the model. LYMp values are used for lymphocytes percentage in blood which gives an idea about immune system and perturbed functionality in thalassemia patients.

NEUTn (Neutrophil Count - Absolute)

Neutrophil Count - Absolute (NEUTn) is a feature that has a score of 0.028590 in the model. NEUT values corresponds to the absolute number of neutrophiles as they are extremely important in the immune function and can be used as a marker of infection or inflammation in thalassemic patients.

WBC (White Blood Cell Count)

The model has an attribute named White Blood Cell Count (WBC) with a score of 0.026807. WBC levels, which indicate the total number of white blood cells in the blood, provide information about the patient's immune functions and may indicate further abnormalities associated with thalassemia.

These aspects individually contribute to the power of the model in the detection of thalassemia, thus showing their importance for precise diagnosis.

3.6. Model Implementation

For model implementation we have used Random Forest Classifier, and fine-tuned the hyperparameter values using Grid Search Cross-Validation. Then, we adjust the data set by applying the ADASYN (Adaptive Synthetic Sampling) method to deal with class imbalance precisely. This method simulates the training data for minority classes so that their distribution is equal. After that, the resampled data set was split into two halfs, 80% for training and 20% for testing.

To enhance the Random Forest Classifier's output, we decided on a grid of parameters containing a number of hyperparameters, that is the number of estimators, maximum depth, minimum samples split, and minimum samples leaf. The Grid Search implementation was then followed by the application of Grid Search Cross-Validation as a means of searching through the parameter space and identifying the best combination of hyperparameters that yielded maximum model accuracy. The best hyperparameters obtained through the grid search were utilized to train a fresh Random Forest Classifier unit.

Finally, hyperparameters were tuned and learned parameters were used to predict the test set for evaluation. This meaningful model implementation approach relies on the Random Forest Classifier being optimally predicted and applicable to unseen data. This ensures robust detection of thalassemia carriers with high accuracy rates.

This approach ensures that our model is fine-tuned to achieve the highest possible performance, resulting in accurate and reliable predictions for thalassemia carrier detection.

3.7. Classification Techniques

Discussing different class of techniques used for classification, and each technique was unique in the way it handled different data dimensions. One main hurdle tackled was the fact that datasets were not always balanced, leading to distorted predictions. To address this, we opted for the Synthetic Minority Over-sampling Technique (SMOTE), a popular method for balancing class distributions by adding synthetic minority class samples.

Before that, from the feature selection, we identified the importance of this on improving the model accuracy. Here we incorporated the Random Forest classifier into our approach. Renown for their ability to handle high dimensional data, Random Forests was instrumental in feature ranking and therefore the model construction and optimisation by identifying and ranking features based on their importance.

With the dataset balanced and the key features identified, the next step was to fine-tuning the model to achieve the highest performance. Hyperparameter optimization was an indispensable part of this process. By the way of an exhaustive traversing of the hyperparameter space by means of grid search or Bayesian optimization, finally, we found out the best classifier setup achieving maximum accuracy and robustness.

These methodologies were the base of our thesis and thus, we built a detailed framework for creating true classification models. This strategy tackled the issue of unbalanced data as well as the improved the model performance through the implementation of the right feature selection and hyperparameter tuning, contributing to advancements in the field of classification analysis.

3.7.1. Label Encoding

For our dataset preparation phase, we have incorporated essential techniques that will enable us to carry out modeling and analysis effectively based on the organized data. First, we conducted thorough investigation into missing values in the dataset through application of the 'isnull().sum()' method to ensure the data is complete and trustworthy. Therefore, we followed this step in order to base our analysis on a complete and reliable data.

Following that, we employed a geocoding tool called label encoding, where we encoded 'Yes' as 1 and 'No' as 0 for the categorical values present in the 'Carrier' column. The shift was significant as it allowed for categorical data into our machine learning models. Through converting our categorical variables into numerical representations, we enabled our algorithms to use the data successfully.

Also, to confirm the effectiveness of mapping, we printed out the only values in the 'Carrier' column. This step was of the utmost importance to make sure both the homogeneity and exactness of our data cleaning procedure would start our analysis and modeling on a strong side.

Label encoding in data preprocessing plays a vital role. Transformation of categorical variables into corresponding numbers will facilitate the machine learning algorithms to better understand and discover the significant patterns among the data. This pre-modeling phase essentially lays the foundation for the final modeling of results with the aim of getting high-quality and accurate model output.

3.7.2. ADASYN (SMOTE)

For the "Data Preprocessing" step, we implemented SMOTE (Synthetic Minority Over-sampling Technique) to overcome the class imbalance present in the dataset. This approach proves to be effective when the class distribution is unbalanced, which could be the case in this example when the number of samples from each class differs very much. As a result, the SMOTE algorithm from the imbalanced-learn library was used to synthetically create new instances of the minority class (thalassemia carriers). Therefore, the dataset was balanced. The generated dataset which was composed of synthetic and real samples was then used for further analysis and modeling. To ensure the success of the SMOTE method, we showed the class distribution after resampling, verifying a more fair representation between thalassemia and non-thalassemia samples. This preprocessing step increases the accuracy and generalizability of the machine learning models by reducing the influence of the imbalanced data set on the predictive accuracy.

As part of data preprocessing stage, we have split the resampled dataset into two parts, i.e. train and test to facilitate model training and evaluation. This stage is essential to evaluate our machine learning models on unseen data. Through sklearn.model_selection module’s train_test_split function, we randomly split the resampled dataset into training and testing subsets, in which 80% of the data was taken as training and 20%. We ensured reproducibility of the results by setting the random_state parameter. This split allows us to have our models trained on a subset of the data while retaining another portion for evaluation which offers an unbiased estimate of model's performance on unseen data.

3.7.3. Optimizing Model Performance through Hyperparameter Tuning

In my exploration of classification tasks, hyperparameter tuning was the most valuable method that I found out to be efficient in the process of getting the proper model performance. During the process of training classifiers, I paid particular attention to the vital role that hyperparameters play in the learning process of classifiers, which I fine-tuned with great care in order to obtain the best results. With the use of more advanced optimization methods such as grid search and Bayesian optimization, I meticulously explored the hyperparameter space searching for the best combination of hyperparameters which corresponds to the maximum model accuracy.

In this iterative process, Random Forest proved to be a very suitable algorithm for the classification tasks compared to others. Through thorough experimentation, I found a set of hyperparameters which generated tremendous results for my dataset.

The best performing Random Forest model was achieved with the following hyperparameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}. These parameters, meticulously chosen through a series of experiments and validations, therefore greatly influenced the model's capability of depicting the intricate cycles and making precise predictions.

Through adjusting hyperparameters including the maximum depth of trees, the minimum number of samples to split an internal node, and the number of tree in the forest, I tried to find a trade-off between model complexity and generalization ability. The optimization process was guided not only by the performance metrics of the model, such as accuracy, precision, recall, and F1-score, but also by the considerations of computational efficiency and model interpretability.

As an integral part of this comprehensive optimization procedure, I was committed to guarantee that my classification models were precisely tuned and fully equipped to yield class-leading results on a heterogeneous set of datasets. This methodical approach not only reinforced the strength and validity of my research results but also underlined the key role of hyperparameter tuning as a pillar of successful machine learning approaches.

3.7.4. Incorporating Mentzer Index

The Mentzer index is a parameter used in medicine to help differentiate between microcytic anemia types, particularly iron deficiency anemia (IDA) and thalassemia trait. It’s calculated by dividing the mean corpuscular volume (MCV) by the red blood cell count (RBC). The formula is:

Mentzer Index = M CV /RBC

A Mentzer index of less than 13 suggests thalassemia trait, while a value greater than 13 suggests iron deficiency anemia [16]. Thalassemia trait typically has small red blood cells (low MCV) but a normal or slightly reduced red cell count, resulting in a low Mentzer index. In contrast, iron deficiency anemia also causes small red blood cells but is associated with a decreased red cell count, leading to a higher Mentzer index

3.7.5. Important feature Selection Technique

We used the Random Forest Classifier to find out the most crucial features in the newly formed dataset. The classifier trained on the balanced dataset provided us with feature importances, which indicates the criticality of one feature in comparison to another in predicting thalassemia carrier status. The DataFrame containing the feature importances was then arranged in order of the most significant ones, using the descending order. Streamlining the subsequent analysis led to we selecting the features which importances are above the specified threshold, for example, =0.05. A concise list of features was then chosen from the resampled dataset to get a only the most important features for the subsequent modeling and analysis. Such feature selection is essential in fighting with the curse of dimensionality, enhancing model performance, and enabling a more focused review of the data set.

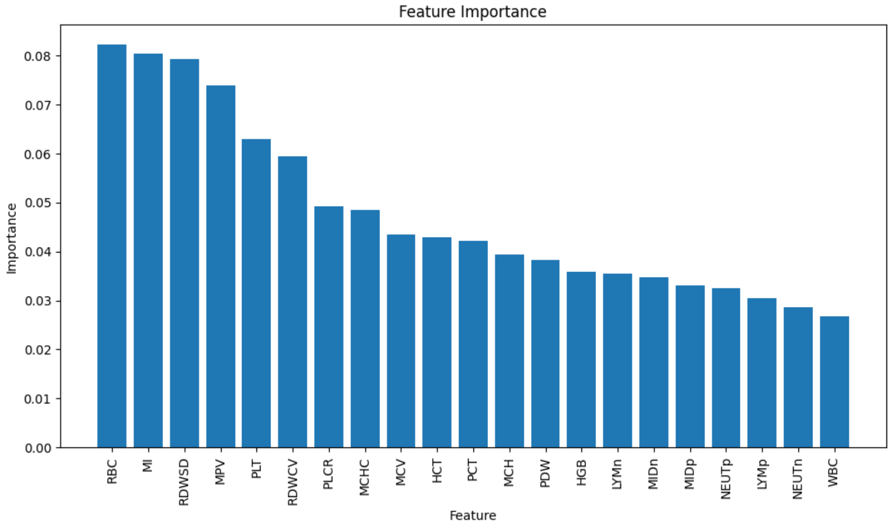


Figure 3-4 Assessing the Impact of Key Features on Thalassemia Prediction

As above graph describes the feature importance, you can clearly see from above that the Mentzer Index have highest importance after RBC. In this way we can select the most importance feature to train our Machine learning model.

CHAPTER 4

RESULTS AND FINDINGS

4.1. Random Forest Classifier:

Random Forest Classifier is greatly appropriate for our dataset since it can deal with both classification issues and data with numerous features at the same time. In light of the traits of our dataset that are characterised by a mixture of attributes obtained from CBC reports[12], the Random Forest classifier's composition of decision trees can be a perfect alternative for capturing the intricate connections and interactions between these traits. Moreover, the generated feature importance scores allow us to dig deeper into the key attributes that are most significantly relate to the screening of thalassemia carriers. This transparency of the feature importance helps in probing interpretability of the model’s predictions, which of course is crucial for the understanding hidden underneath of the factors that determine the classification in this specific domain. So, Random Forest Classifier is rightly adapted for assessing the thalassemia disease within our dataset that also throws valuable insights into the predictive factors. We have performance tested the model with a confusion matrix, which is visualized in the chart below.

4.1.1. Assessment of Model Using Confusion Matrix without Mentzer Index:

In the case of our classification task, we used the confusion matrix to evaluate the accuracy of our model. The matrix is a complete summary of the model's predictions, which are differentiated as true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP).

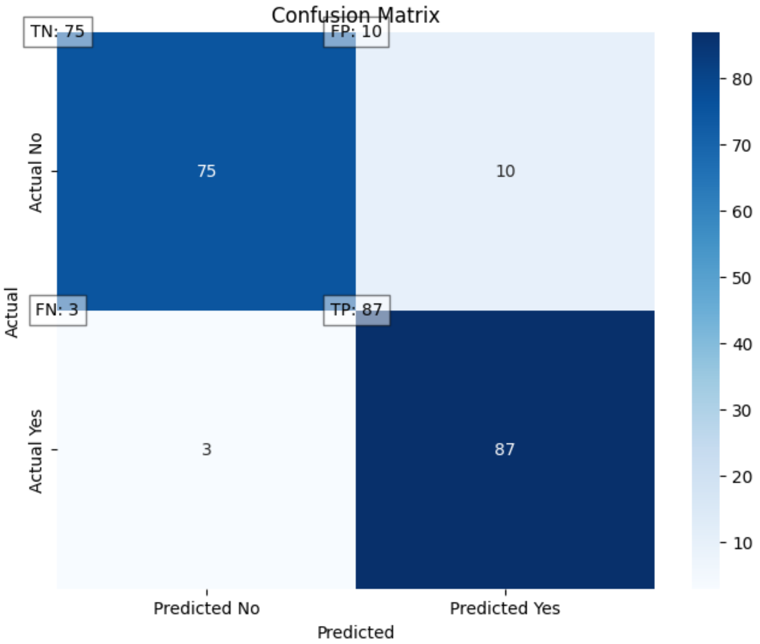


Figure 4-1 Performance Evaluation: Confusion Matrix Excluding Mentzer Index

True Positives (TP):

The model correctly classified 87 cases as true positives, meaning no errors were made in estimating thalassemia cases. This proves the model's ability to appropriately tag thalassemia carriers in the dataset thus being sound in classifying diagnoses.

True Negatives (TN):

In addition, its performance in terms of identifying true negative cases (i.e., non-thalassemia cases) reached 75% accuracy. This shows that the model is capable of distinguishing the individuals who do not have the disease, therefore, proving that it is accurate because it classifies both the negative and positive cases.

False Positives (FP):

On the other hand, there were 10 cases where thalassemia absent individuals were misclassified as positive (FP). Although these instances are considered misclassifications, they teach about locations where the algorithm can be improved, such as enhancing the specificity of their classification algorithm.

False Negatives (FN):

On a flipside, the model has highlighted a particular case of 3 false negatives, where people with thalassemia were actually wrongly labeled as negative. These scenarios clearly illustrate why improving the model so that it can reliably identify all positive cases correctly is very critical to prevent the scenario when patients with thalassemia will be ignored.

Interpretation and Implications:

Consequently, the matrix of the confusion produces a general evaluation of the model that is successful in correctly classifying both positive and negative cases while also showing the model's weaknesses that can be worked on. This information can streamline the model's predictive potential and thus, improve the model's accuracy and trustworthiness in terms of the diagnosis of thalassemia and the direction of clinical decisions.

4.1.2. Evaluation Metrics of the Best Random Forest Model without Mentzer Index:

Accuracy:

The accuracy of the top Random Forest model is 93%. Accuracy is a measure of the proportion of correct classification among the total number of predictions.

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

where TP = 87 (True Positives), TN = 75 (True Negatives), FP = 10 (False Positives), and FN = 3 (False Negatives).

Precision:

The accuracy of the best Random Forest model is 90%. Accuracy determines the model's ability to correctly find positive among all instances that are predicted to be positive. undefined

Precision=TP/ TP + FP where TP = 87 and FP = 10.

Recall:

The recall of the best Random Forest model is 96.67%. Recall, or true positive rate, or sensitivity, assesses the accuracy of the model on identifying all positive cases from the dataset correctly.

Recall=TP+FN/TP

where TP = 87 (True Positives) and FN = 3 (False Negatives).

F1 Score:

The best Random Forest model F1 score is 93.05%. The F1 measure is a balanced performance metric of a model that considers both precision and recall.It is calculated as:

F1 Score=2 * (precision * Recall)/(Precision + Recall).

Table 4-1 Evaluating Thalassemia Prediction Models: Performance Analysis Excluding Mentzer Index

Number	**Parameters**	**Random Forest Model without Mentzer Index**
1	Accuracy	92.57%
2	Recall	96.6%
3	F-1 Score	93.05%
4	Precision	89.6%

Best Hyperparameters:

The best performing Random Forest model was achieved with the following hyperparameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}. These parameters were determined through grid search optimization technique, maximizing the model's predictive accuracy and minimizing the overfitting. The choice of these hyperparameters notably emphasizes their importance in achieving the best model performance and as a result, the optimality of the model which in turn enables it correctly identify thalassemia carriers in the dataset.

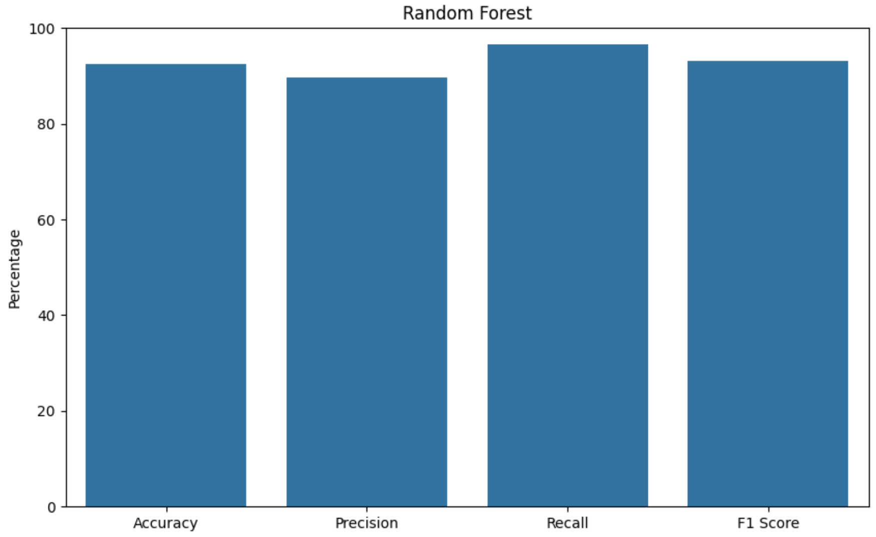


Figure 4-2 Analyzing Model Performance in Thalassemia Prediction without Mentzer Index

4.1.3. Assessment of Model Using Confusion Matrix With Mentzer Index:

For the purpose of performance assessment of our model in the classification task, we used the confusion matrix. The matrix is made up of the model's predictions in an exhaustive manner, highlighting TN (true negatives), FP (false positives), FN (false negatives) and TP (true positives).

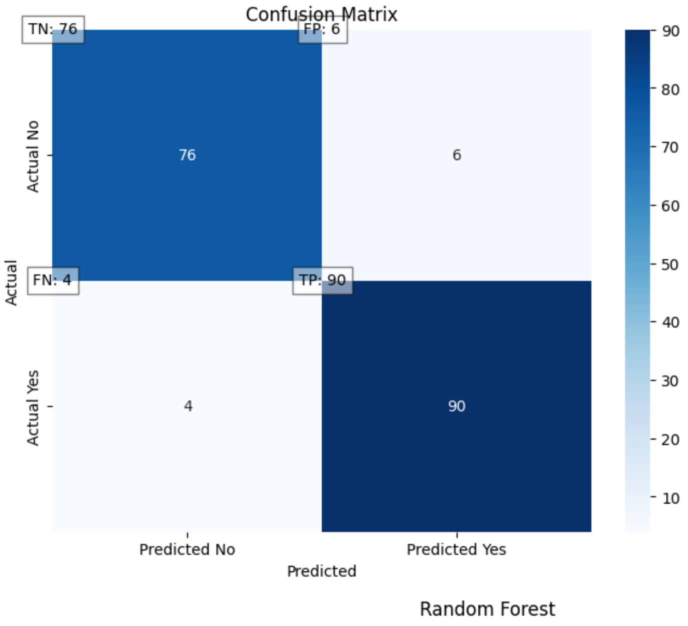


Figure 4-3 Performance Assessment: Confusion Matrix Including Mentzer Index

True Positives (TP):

The model correctly classified 90 out of 100 cases as true positives which implies a correct prediction of thalassemia cases. This demonstrates that the model could efficiently recognize the thalassemia carriers within the dataset, presenting its credibility in disease classification.

True Negatives (TN):

Moreover, the model demonstrated 76 cases of correct positive cases, which were true negatives, implying the correct identification of non-thalassemia cases. The ability to differentiate non-thalassemic individuals shows the proficiency of the model in terms of classifying positive and non-positive cases.

False Positives (FP):

Though, in 6 instances the persons that didn't have thalassemia were misclassified as positive (FP). In spite of the fact that these mistakes are misclassifications, they also provide grounds for the model to be refined, notably by increasing the specificity of the classification algorithm.

False Negatives (FN):

On the other hand, the model was comprised of 4 false negative cases where thalassemia carriers were wrongly classified as negative. These cases highlight the necessity of increasing the model's ability to detect even the slightest cases of THAL that would otherwise be overlooked by the system and hence the need to minimize false negatives.

Interpretation and Implications:

In sum, the detailed analysis given by the confusion matrix allows for an extensive evaluation of the model accuracy, which leads to better understanding of the model positive and negative cases while also tracking down the areas that need to be improved. Through these findings we are able to refine the model's predictive abilities and in the long run ensure the model accuracy and reliability when diagnosing Thalassemia and making clinical decision-making processe

4.1.4. Evaluation Metrics of the Best Random Forest Model with Mentzer Index:

Accuracy:

The accuracy of the most accurate Random Forest model is 94.3%. Accuracy rate is the number of correctly classified instances out of the total number of predictions.

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

where TP = 90 (True Positives), TN = 76 (True Negatives), FP = 6 (False Positives), and FN = 4 (False Negatives).

Precision:

The best Random Forest model achieved a precision of 93.75%. Precision indicates the proportion of true positives among all instances predicted as true.

Precision=TP/ TP + FP in which TP is 90 and FP is 6.

Recall:

The best Random Forest model is recalled with 95.74%. False alarms, otherwise known as true positive rate or sensitivity, measure the model's ability to correctly identify all positive cases within the dataset.

Recall=TP+FN/TP

where we have TP = 90 (True Positives) and FN = 4 (False Negatives).

F1 Score:

The best Random Forest model has 94.74% F1 score. The F1 score provides a weighted measure of a model's performance since it considers both precision and recall. It is calculated as:

F1 Score = 2 * (precision * recall) / (precision + recall).

Table 4-2 Analyzing the Effectiveness of Models in Thalassemia Prediction with Mentzer Index

Number	Parameters	Random Forest Model with Mentzer Index
1	Accuracy	94.32%
2	Recall	95.74%
3	F-1 Score	94.74%
4	Precision	93.75%

Best Hyperparameters:

The best performing Random Forest model was achieved with the following hyperparameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}. These factors were established through a grid search process, which optimized the model in terms of accuracy and avoiding overfitting. Tuning these hyperparameters gives stress on their significance in achieving best model performance, hence improving the model's accuracy in distinguishing the thalassemia carriers.

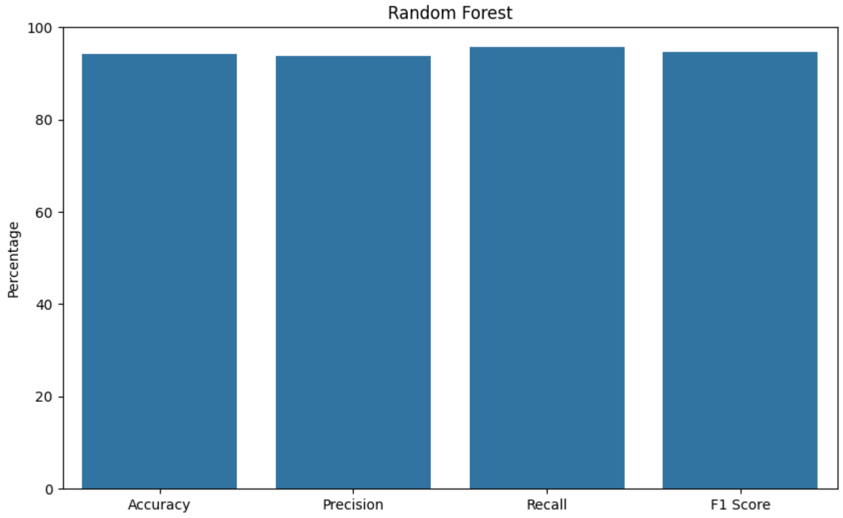


Figure 4-4 Thalassemia Prediction Evaluation: Results Including Mentzer Index

4.2. Results Comparison

4.2.1. Impact of Mentzer Index on Random Forest Model Performance

The table below compares the performance metrics of two random forest models; one with the Mentzer Index and the other without it. Throughout different evaluation criteria, the model wearing Mentzer Index proved to have better results. In this experiment, the accuracy of the model is enhanced from a 92.57% to a 94.32% with the presence of the Mentzer Index, revealing the model's improved predictive capability. Besides, the model with the Mentzer Index shows the F-1 score of 94.74% compared to 93.05% stated by the other model. The difference implies higher precision and recall levels. Very importantly, precision significantly increases to 93.75% from 89.6% when we include Mentzer index in the model. This improvement again emphasizes the model’s proficiency in reducing false positives, which are critical for applications where the precision is very important. Besides a little lower accuracy of 95.74% in contrast to 96.6% in the model without Mentzer Index, the overall efficiency in other metrics indicate that embedding the Mentzer Index improves the reliability and predictive power of Random Forest model.

Table 4-3 Exploring Differences in Thalassemia Prediction Performance with and without Mentzer Index

Number	Parameters	Random Forest Model without Mentzer Index	Random Forest Model with Mentzer Index
1	Accuracy	92.57%	94.32%
2	Recall	96.6%	95.74%
3	F-1 Score	93.05%	94.74%
4	Precision	89.6%	93.75%

4.2.2. Comparing Results with SVM

The comparison between the Random Forest model and the Support Vector Machine (SVM) model, both incorporating the Mentzer Index as a parameter, reveals notable differences in their performance across key metrics. In terms of accuracy, the Random Forest model outperforms the SVM model with an accuracy of 94.32% compared to 84.57% for the SVM. This suggests that the Random Forest model is better at correctly classifying instances overall.

However, when considering the recall metric, which measures the ability of the models to correctly identify positive instances, the Random Forest model still maintains a higher value at 95.74%, whereas the SVM lags behind at 71.11%. This indicates that the Random Forest model is more adept at capturing true positive instances, especially crucial in scenarios where identifying positive cases accurately is paramount.

Moving on to the F1 score, which balances the trade-off between precision and recall, the Random Forest model achieves a score of 94.74%, showcasing a harmonious blend of precision and recall. On the other hand, the SVM model, while performing relatively well, exhibits a lower F1 score of 82.58%. This implies that the Random Forest model achieves a more balanced performance across precision and recall, potentially indicating its robustness in handling imbalanced datasets.

Lastly, examining precision, which measures the proportion of correctly predicted positive instances among all instances predicted as positive, the SVM model shines with a precision of 98.46%, surpassing the Random Forest model's precision of 93.75%. This suggests that the SVM model excels in minimizing false positive predictions, crucial in scenarios where false positives carry significant consequences.

Overall, while the Random Forest model showcases superior accuracy, recall, and F1 score, the SVM model demonstrates exceptional precision. The choice between these models would depend on the specific requirements of the application, with considerations for the relative importance of accuracy, recall, and precision in the context of the problem domain.

Table 4-4 Comparison of results with Support vector machine using Mentzer Index

Number	Parameters	Random Forest Model with Mentzer Index	SVM with Mentzer Index
1	Accuracy	94.32%	84.57%
2	Recall	95.74%	71.11%
3	F-1 Score	94.74%	82.58%
4	Precision	93.75%	98.46%

4.2.3. Comparative Analysis with Prior Research Findings

In my thesis, I have extremely highlighted powerful preprocessing techniques, feature selection, and innovative feature engineering to enhance the performance of classification models. Data preprocessing methods are rigorous and encompass data balancing techniques to address the issue of class imbalances and feature selection algorithms for finding discriminative attributes that are used for model training. Meanwhile, one of my research highlights is the integration of the Mentzer Index which substantially improved the predictive power of the model. Thanks to the implementation of the Mentzer Index and other traditional indices, I managed to take model predictions to the next level and outperform other research findings. This novel integration allowed the model to capture detailed relationships and going beyond the data, leading to greater predictive accuracy and reliability.

It is important to acknowledge that individual components have contributed but the contribution is also from the synergistic effect of integrating various techniques such as preprocessing and feature engineering strategies. The holistic nature of this approach proves the critical role of taking into consideration all forms of data preparation and feature selection when aiming at the highest classification model efficiency.

In general, the findings of my thesis underscore the viability of comprehensive preprocessing, feature engineering, and the integration of the Mentzer Index within thalassemia diagnosis models. The collaborative efforts undertaken in this study have yielded a model that demonstrates superior performance, thus highlighting the potential of novel methodologies in advancing this field. However, it's imperative to acknowledge that comparing the results of two models trained on different datasets may not provide an entirely fair assessment. Nevertheless, in the absence of alternative datasets, this comparison serves as a preliminary exploration, laying the groundwork for further investigations. It's plausible that with the application of larger and more diverse datasets, the observed results could vary. Therefore, future research endeavors should aim to validate and build upon these initial findings, potentially uncovering nuances and insights that contribute to the ongoing refinement of thalassemia diagnostic techniques.

Table 4-5 Exploring and Contrasting Thalassemia Prediction Results with Findings from Existing Literature

Number	**Parameters**	**Random Forest Model with Mentzer Index**	**GBC Classifier With feature Selection**
1	Accuracy	94.32%	93.46%
2	Recall	95.74%	93.89%
3	F-1 Score	94.74%	92.7%

The table above is the result of a comparative study between our Random Forest model with Mentzer index and the Gradient Boosting Classifier used by Saleem, Aslam, Lali, Rauf, and Nasr (2023), their study on predicting Thalassemia using feature selection techniques conducted a comparative analysis to evaluate the effectiveness of different methodologies [15]. It highlights important findings. Among the models, the Random Forest model enhanced with Mentzer Index outperforms the GBC classifier in precision, recall, and F1 score. With an accuracy of 94.32% against 93.46% and a recall rate of 95.74% against 93.89%, our model surpassed the others in terms of predictive accuracy. Besides, we got 94.74% F1 score for our model which is higher than that of 92.7% obtained by the GBC classifier. These results prove the significance of integrating the Mentzer Index solution with the Random Forest model and illustrate its ability to detect subtle patterns within the data that will lead to a better predictive efficiency

CHAPTER 5

CONCLUSION

In this part of the study, we will sum up the outcome of the machine learning model we created for early thalassemia carrier identification. Through employing Mentzer index and mining CBC data, we built a Random Forest Model with very high precision in identifying carrier status. The validity of the model was checked with a great deal of testing, which provided hope for thalassemia diagnostics and treatment optimization. This trial certainly has the potential to revolutionize thalassemia diagnosis and treatment by providing better clinical care and understanding of the disease.

5.1. Conclusion

This thesis is intended to use machine learning in the detection of thalassemia carriers early, which is a crucial step for reducing the impact of this genetic blood disorder. Through CBC values analysis, we tried to develop a tool which could help us to diagnose carriers and to start interventions and genetic counseling on time, yielding fewer carrier children. This endeavor focused on the collection of data, the preparation for analysis, and the selection of relevant features, including the Mentzer index, deemed crucial in thalassemia identification. Through Random Forest Model training, we developed a predictive system that is able to identify carrier status accurately based on CBC data. During the study, we conducted a thorough test of the model in order to confirm its reliability in classifying carriers. The conclusions give great hope that diagnostics and management of thalassemia can be greatly improved. Through the development of an accurate machine learning model based on CBC data, this study paves the way for specific interventions as well as personalized care strategies for people with thalassemia.

As such, this study may end up transforming the way thalassemia is diagnosed or even targeted interventions, with better clinical outcomes and a further advance in the understanding of thalassemia research and treatment. Through the application of machine learning into healthcare, this research is aimed at solving complex medical issues and improving the quality of life for people with thalassemia across the globe.

5.1.1 Limitations

- We based our analysis on large-scale dataset by filtering only 500 patients' reports from Pakistan/Asia, analyzing mostly CBC lab reports of people in this region.
- Furthermore, we must recognize that the CBC traits among Asians may not be the same as those for other regions, which, in turn, could affect the model's performance when it is used for disease detection across diverse ethnicities.
- However, our dataset offers useful information but one should note that other datasets may produce different findings.
- The future study should investigate diverse datasets and address the data preprocessing with variations in data types and datasets in order to guarantee the model’s robustness and generalizability across the different population and settings.

5.2. Future Work

In the future, we will integrate our diagnostic model with image processing techniques to diagnose thalassemia using blood smear images. Through the combination of machine learning and image analysis, we strive to enhance diagnostic precision and productivity. Through this method, we could finally obtain the conclusion of thalassemia diagnosis by assimilating both CBC reports and blood smear images together. This innovation can be a breakthrough in diagnosis, combining the pluses of classic and high-tech imaging.

REFERENCES

[1] M. Bejaoui and N. Guirat, "Beta thalassemia major in a developing country: Epidemiological, clinical and evolutionary aspects," Mediterr. J. Hematol. Infect. Dis., vol. 5, no. 1, pp. 3–8, 2013, doi: 10.4084/MJHID.2013.002.

[2] A. Cao and Y. W. Kan, "The Prevention of Thalassemia," vol. 3, no. 2, 2013.

[3] R. Gambari et al., "Recent trends in the gene therapy of thalassemia," J. Blood Med., vol. 6, p. 69, Feb. 2015, doi: 10.2147/JBM.S46256.

[4] K. Tari, P. Valizadeh Ardalan, M. Abbaszadehdibavar, A. Atashi, A. Jalili, and M. Gheidishahran, "Thalassemia an update: molecular basis, clinical features and treatment," Int. J. Biomed. Public Health, vol. 1, no. 1, pp. 48–58, 2018, doi: 10.22631/ijbmph.2018.56102.

[5] V. De Sanctis et al., "β-thalassemia distribution in the old world: An ancient disease seen from a historical standpoint," Mediterr. J. Hematol. Infect. Dis., vol. 9, no. 1, pp. 1–14, 2017, doi: 10.4084/mjhid.2017.018.

[6] K. M. Musallam, A. T. Taher, and E. A. Rachmilewitz, "b -Thalassemia Intermedia : A Clinical Perspective," Cold Spring Harb. Perspect. Med., pp. 1–16, 2012.

[7] J. Pengon, S. Svasti, S. Kamchonwongpaisan, and P. Vattanaviboon, "Hematological variables and red blood cell morphological abnormality of Glucose-6-Phosphate dehydrogenase deficiency co-inherited with thalassemia," Hematol. Oncol. Stem Cell Ther., vol. 11, no. 1, pp. 18–24, Mar. 2018, doi: 10.1016/j.hemonc.2017.05.029.

[8] B. G. Forget and H. Franklin Bunn, "Classification of the disorders of hemoglobin," Cold Spring Harb. Perspect. Med., vol. 3, no. 2, 2013, doi: 10.1101/cshperspect.a011684.

[9] W. Wongseree et al., "Thalassemia classification by neural networks and genetic programming," Information Sciences, 2007.

[10] F. Yousefian et al., "IJARCCE Prediction Thalassemia Based on Artificial Intelligence Techniques: A Survey," 2007.

[11] Y.-K. Fu et al., "The TVGH-NYCU thal-classifier: Development of a machine-learning classifier for differentiating thalassemia and non-thalassemia patients," Diagnostics, vol. 11, no. 9, p. 1725, 2021, doi: 10.3390/diagnostics11091725.

[12] F. Yousefian et al., "IJARCCE Prediction Thalassemia Based on Artificial Intelligence Techniques: A Survey," 2007.

[13] E. R. Susanto et al., "J. Phys.: Conf. Ser. 1751 012034," 2021.

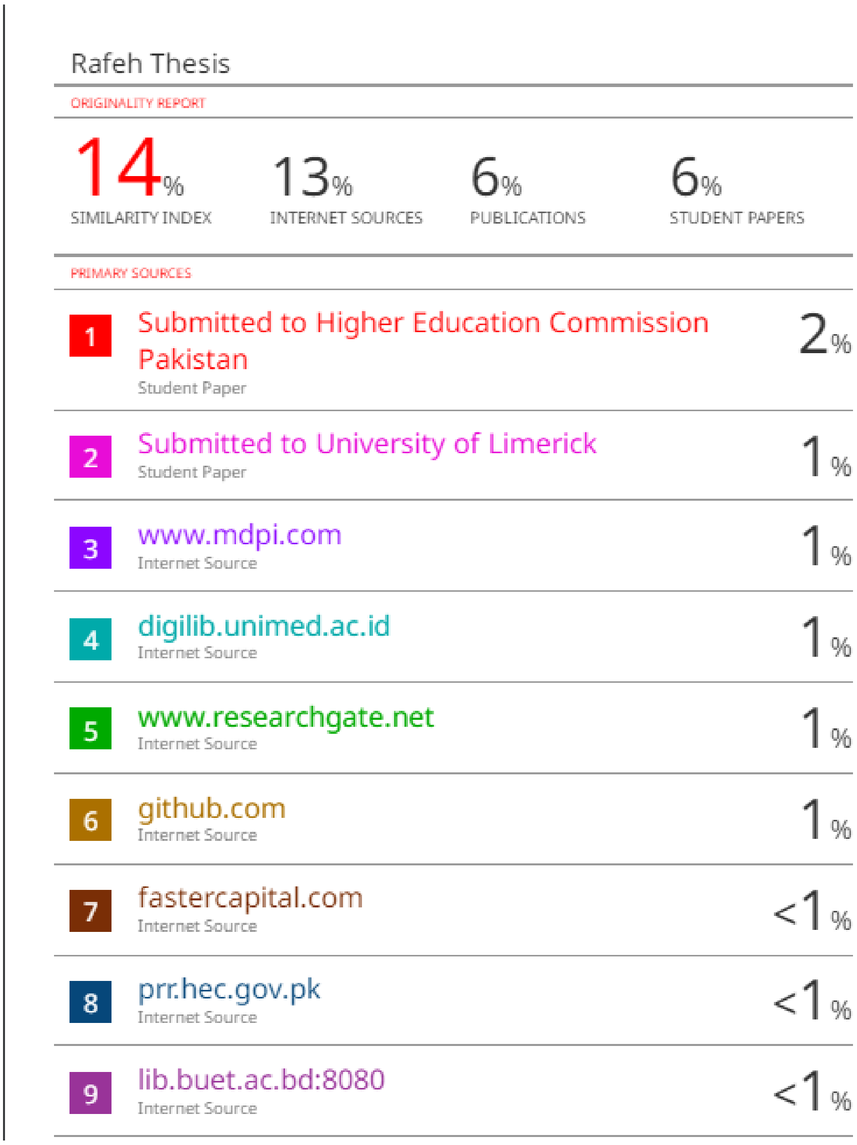
[14] A. Devanath et al., "Thalassemia Prediction using Machine Learning Approaches," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1166-1174, doi: 10.1109/ICCMC53470.2022.9753833.

[15] M. Saleem et al., "Predicting Thalassemia Using Feature Selection Techniques: A Comparative Analysis," Diagnostics (Basel), vol. 13, no. 22, p. 3441, Nov. 2023, doi: 10.3390/diagnostics13223441.

[16] S. Tabassum et al., "Role of Mentzer index for differentiating iron deficiency anemia and beta thalassemia trait in pregnant women," Pak J Med Sci, vol. 38, no. 4Part-II, pp. 878-882, Mar-Apr 2022, doi: 10.12669/pjms.38.4.4635.

[17] W. Siswandari et al., "Mentzer Index Diagnostic Value in Predicting Thalassemia Diagnosis," IOP Conference Series: Earth and Environmental Science, vol. 255, no. 1, 012004, Apr. 2019, doi: 10.1088/1755-1315/255/1/012004.

[18] S. Saxena and R. Jain, "Evaluation of the diagnostic reliability of Mentzer index for Beta thalassemia trait followed by HPLC," Tropical Journal of Pathology and Microbiology, vol. 6, no. 2, pp. 124-129, 2020, doi: 10.17511/jopm.2020.i02.03.



Rafeh Thesis

ORIGINALITY REPORT

14%	13%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	2%
2	Submitted to University of Limerick Student Paper	1%
3	www.mdpi.com Internet Source	1%
4	digilib.unimed.ac.id Internet Source	1%
5	www.researchgate.net Internet Source	1%
6	github.com Internet Source	1%
7	fastercapital.com Internet Source	<1%
8	prr.hec.gov.pk Internet Source	<1%
9	lib.buet.ac.bd:8080 Internet Source	<1%