DETECTION OF THALASSEMIA USING BLOOD SMEAR IMAGES: A MACHINE LEARNING APPROACH

MUHAMMAD HAMMAD
01-241221-012

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 Hammad Registration No. 01-241221-012

Program of Study: MS (Software Engineering)

Thesis Title: Detection of Thalassemia Using Blood Smear Images: A Machine Learning Approach

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 Hammad hereby state that my MS thesis titled “Detection of Thalassemia Using Blood Smear Images: A Machine Learning Approach” 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 Hammad (01-241221-012)

Date: ___________________________________

PLAGIARISM UNDERTAKING

I, Muhammad Hammad, solemnly declare that research work presented in the thesis titled "Detection of Thalassemia Using Blood Smear Images: A Machine Learning Approach" 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 Hammad (01-241221-012)

DEDICATION

To my beloved mother and father

ACKNOWLEDGEMENT

All praise to **Allah Almighty**, Lord of all creations and His **Prophet Muhammad (S.A.W)** whose blessings enabled us to carry out this project and who bestowed us with the knowledge and courage to complete it successfully.

I wish to express my gratitude to Dr. Joddat Fatima and Dr. Madiha Khalid. Its been an absolute pleasure to work under their supervision. Special thanks to both my supervisors who have given their support and guidance support in completion of my thesis and in my educational career.

I would like to dedicate this thesis to the source of my strength and inspiration throughout my academic journey, my parents Mr. Farman and Shakeela Baigum for their uncountable sacrifices.

Special thanks goes to my family and friends M Rafeh Latif, Saqib Saeed, Farooq Khan and Mehran Ahmed for their support through the years, in times both good and bad to keep me focused on my goals.

Special thanks to Izza Rehman and Hira Rehman for their constant support, encouragement, and understanding.

Finally, I would like to thank Bahria University for enabling me to carry out this thesis and forwarding me to practical life with full strength. To sum it up, it was a total effort between all of the mentioned people that led to my accomplishments.

ABSTRACT

Thalassemia is one of the most common genetic disorders worldwide, particularly prevalent in populations like Asian countries and African descent. In Pakistan thalassemia trait ranges from 5.0% to 7.0%, indicating the presence of more than 10 million carriers in the country. Moreover, the annual incidence of β-thal major (β- TM) in Pakistan is around 5000 children. Early detection of thalassemia carriers is crucial for effective management and prevention of severe forms of the disease. In this study, we propose a machine learning approach for the detection of thalassemia carriers using blood smear images. Our methodology involves preprocessing the images to extract relevant features, including color, texture, and shape. We then employ deep learning models, including Convolutional Neural Networks (CNNs), to classify the images into thalassemia and non-thalassemia categories. The dataset consists of nearly 7108 blood images, including nine cell types associated with thalassemia. The highest accuracy in terms of machine learning models achieve d from Random Forest of 91.1%. While on the other side the highest accuracy of 90% achieved from MobileNetV2 model in applying deep learning algorithms. Our results show promising accuracy rates, with the potential for real-world application in thalassemia screening programs. This research contributes to the advancement of diagnostic methodologies for thalassemia and could lead to improved healthcare outcomes in affected population.

TABLE OF CONTENTS

CHAPTER **TITLE** **PAGE**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Artificial Intelligence (AI), Machine Learning (ML), Image Processing and Medical Sciences have have witnessed remarkable advancements in recent years, revolutionizing the way diseases are diagnosed and treated. AI and ML algorithms have shown exceptional capabilities in analyzing complex medical data, including images, to guide medical professionals in making accurate and timely decisions. Image processing techniques have enabled the extraction of meaningful information from medical images, leading to more precise diagnoses. This thesis focuses on leveraging these cutting-edge technologies to develop a novel approach for the detection of thalassemia using blood images, aiming to enhance the efficiency and accuracy of thalassemia screening processes.

1.1 Thalassemia

Thalassemia is a group of blood disorders caused by abnormal hemoglobin production, resulting in fewer or decreasing red blood cells. It is a common genetic disorders worldwide, particularly prevalent in populations of Mediterranean, Middle Eastern, Southeast Asian, and African descent. It poses a significant healthcare burden worldwide, particularly in countries like Pakistan with a high prevalence of thalassemia carriers.

1.2 Types of Thalassemia

The two main types of thalassemia are as follows:

1. Alpha thalassemia.

2. Beta thalassemia.

Each type has different subtypes or variations.

1.2.1 Alpha thalassemia

When the production of alpha globin chains is disturbed then alpha thalassemia occurs. The four subtypes of alpha thalassemia are:

- Silent Carrier: In this subtype, a person carries one altered alpha globin gene but usually does not experience any symptoms.
- Alpha Thalassemia Trait: Individuals with alpha thalassemia trait carry two altered alpha globin genes and may have mild anemia.
- Hemoglobin H Disease: This subtype is characterized by the deletion of three alpha globin genes, leading to moderate to severe anemia and potential complications.
- Hydrops Fetalis: The most dangerous and crucial form of alpha thalassemia, where all four alpha globin genes are deleted or severely mutated. It usually leads to fetal death or severe symptoms in newborns.

1.2.2 Beta thalassemia

When the production of beta globin chains is disturbed then beta thalassemia occurs. The three subtypes of beta thalassemia are:

- Beta Thalassemia Minor: People with beta thalassemia minor have one altered beta globin gene and usually do not exhibit severe symptoms.
- Beta Thalassemia Intermedia: This subtype is caused by the abnormal production of beta globin chains, leading to moderate to severe anemia that may require occasional blood transfusions.
- Beta Thalassemia Major (Cooley's Anemia): Individuals with beta thalassemia major have two altered beta globin genes, resulting in severe anemia that requires regular blood transfusions and ongoing medical management.

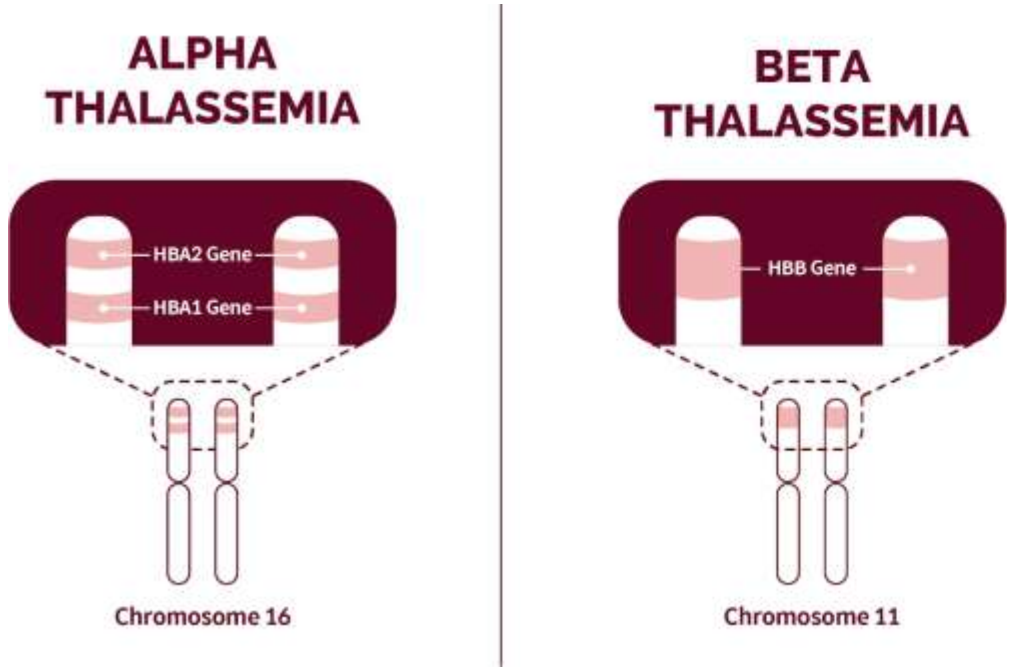


Figure 1.1:Types of Thalassemia, Alpha thalassemia and Beta thalassemia

The severity of thalassemia symptoms can vary widely, ranging from mild and manageable to severe and life-threatening. Treatment options may include regular blood transfusions, iron chelation therapy, and, sometimes bone marrow transplants, which is expensive and available in only three medical facilities in Pakistan. In order to comprehend the inheritance pattern and make informed decisions, genetic counselling is also advised for thalassemia affected individuals and their families.

1.3 Thalassemia in Pakistan

The estimated population of Pakistan is approximately 225,633,392 (225 million), and the frequency of β-thalassemia (β-thal) trait ranges between 5.0% to 7.0%, indicating the presence of more than 10 million carriers in the country. Moreover, the annual incidence of β-thal major (β- TM) in Pakistan is around 5000 children. The management of thalassemia remains challenging due to the lack of standardized protocols and the heavy reliance on blood transfusions [1].

The complicated healthcare delivery system in Pakistan, which is run by both the federal and provincial governments, makes the problem even more difficult to handle. This fragmentation creates challenges in implementing cohesive strategies for early detection, prevention, and treatment of thalassemia.

Additionally, the socioeconomic landscape of Pakistan poses additional hurdles in managing thalassemia. The majority of the population belongs to lower socioeconomic strata, making it financially burdensome for families to afford the long-term treatment and management required for thalassemic children. This situation is exacerbated by the large family units prevalent in the country.

Despite these challenges, there is a pressing need to address thalassemia on a national level and implement effective prevention programs. Currently, there is a lack of comprehensive thalassemia prevention initiatives at the national level in Pakistan. At the provincial level, premarital screening laws have been approved in Sindh, Khyber Pakhtunkhwa (KPK), and Baluchistan, but there is still a long way to go before these measures are successfully implemented.

To mitigate the burden of thalassemia in Pakistan, there is a crucial need to develop reliable and efficient methods for the early detection of thalassemia carriers. In recent years, advancements in medical imaging and analysis techniques have shown promising results in the detection and diagnosis of various diseases. The application of blood imaging analysis for thalassemia detection holds significant potential, as it can provide non-invasive and cost-effective means for identifying individuals carrying thalassemia traits.

The importance of thalassemia in the field of medical diagnostics lies in the early detection and accurate diagnosis of the disorder. Prompt identification of thalassemia allows for appropriate management strategies to be implemented, resulting in improved patient outcomes and quality of life. When it comes to severe forms of thalassemia, early detection is very important since it allows for the timely initiation of treatments including regular transfusions of blood, chelation therapy to eliminate excess iron, and potentially curative bone marrow transplants.

Currently, a combination of clinical assessment, blood testing, and genetic analysis is used to diagnose thalassemia. Hemoglobin levels and red blood cell indices (mean corpuscular volume and mean corpuscular hemoglobin) are measured by blood tests. Hemoglobin electrophoresis is also sometimes used to find aberrant hemoglobin patterns. Genetic analysis is often performed to confirm the specific thalassemia mutation and provide accurate genetic counseling.

1.4 Use of Medical Imaging Techniques

The use of medical imaging techniques, such as analyzing blood cell images, offers a promising avenue for thalassemia detection. By leveraging image analysis and machine learning algorithms, it may be possible to develop a non-invasive, efficient, and cost-effective diagnostic tool. Such a tool could assist in early identification, aid in treatment decisions, and potentially reduce the need for invasive diagnostic procedures.

1.5 Normal Vs Thalassemia Blood Cells

1.5.1 Normal Blood Cells

Normal blood cells exhibit different characteristics. They are typically of regular size and shape, appearing as biconcave discs. This unique structure allows for increased surface area, facilitating efficient exchange of gases, particularly oxygen and carbon dioxide. The red color of normal blood cells is a result of the presence of hemoglobin. Red blood cells contain a complex protein called hemoglobin, which binds to oxygen in the lungs and carries it throughout the body to different tissues and organs. The iron component of hemoglobin binds to oxygen molecules to generate oxhemoglobin, which gives cells their distinctive red color. Unlike thalassemic cells, normal mature red blood cells do not contain nuclei, as they undergo the expulsion of nuclei during the maturation process.

1.6 Thalassemia Blood Cells

In thalassemic cells, the synthesis of hemoglobin is impaired, leading to various abnormalities in the size, shape, and function of the red blood cells. In individuals with thalassemia, certain distinct characteristics can be observed in their blood cells. Thalassemic red blood cells are generally smaller in size, a condition known as microcytosis. Additionally, these cells tend to have reduced hemoglobin content, resulting in a paler appearance, which is referred to as hypochromia. Another notable feature is the presence of anisocytosis, where there is a variation in the size of red blood cells, leading to an uneven distribution. Target cells are frequently observed in thalassemic blood, which have a central area of hemoglobin surrounded by a clear ring and a peripheral dark rim. Insevere cases of thalassemia, immature red blood cells with nuclei, known as nucleated red blood cells, can also be present in the peripheral blood.

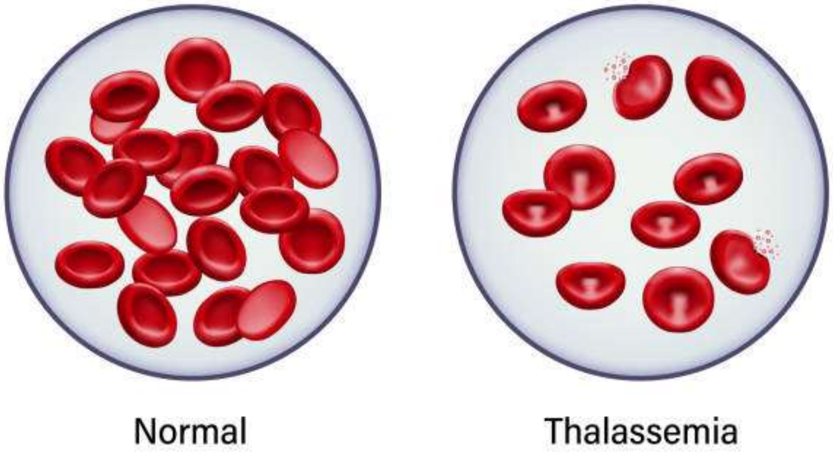


Figure 1.2:Difference between normal blood cells and thalassemic blood cells

1.7 Motivation

A common genetic blood condition affecting a lot of people worldwide is thalassemia. Early and precise detection of thalassemia carriers is essential for executing proper administration techniques, improving patient care, and reducing the burden on healthcare systems. Current diagnostic methods, such as laboratory tests, can be time-consuming, expensive, and require specialized equipment and expertise. By developing an automated system for thalassemia carrier detection using blood images, this research seeks to overcome these limitations and provide a cost-effective and accessible solution that can be easily integrated into existing healthcare practices. Such a system has the potential to revolutionize thalassemia screening, enabling timely interventions, genetic counseling, and personalized treatment plans for individuals at risk. Additionally, the research aims to contribute to the broader field of medical image analysis and machine learning by advancing the state-of-the-art in image-based diagnostics and demonstrating the practical application of these technologies in improving healthcare outcomes.

1.8 Research Problem

The detection of thalassemia using blood smear images poses a significant challenge due to the need for accurate and efficient analysis techniques. Traditional methods of thalassemia detection rely heavily on manual examination by trained experts, which can be time-consuming, subjective, and prone to errors. Automating this process through image analysis algorithms could streamline diagnosis, improve accuracy, and facilitate early intervention, but existing methods may not fully exploit the potential of advanced image processing and machine learning techniques.

1.9 Problem Statement

The study focuses on enhancing patient care and management strategies for thalassemia by developing and optimizing machine learning algorithms for automated detection and classification of the disease from blood smear images. The research aims to advance automated thalassemia detection technology, improving diagnostic efficiency and accuracy for timely interventions.

1.10 Research Questions

RQ 1: How can we effectively preprocess and enhance blood images to ensure the reliability and accuracy of thalassemia carrier detection?

RQ 2: What are the most informative and discriminative features that can be extracted from blood images to distinguish between thalassemia carriers and non-carriers?

RQ 3: How does the performance of the proposed blood image-based thalassemia carrier detection system compare to traditional laboratory tests and other diagnostic methods commonly used for thalassemia screening?

1.11 Aim and Objectives

The aim is to explore the detection of thalassemia carriers using blood images as a diagnostic tool. Some goals and objectives are listed below:

- Early detection of thalassemia carrier using blood smear images for enabling timely interventions to improve patient care.
- The endeavors is to develop an accurate and reliable approach to identify individuals carrying thalassemia traits from blood images.

- The outcomes will contribute to the advancement of diagnostic methodologies, paving the way for early detection, targeted intervention, and improved management of thalassemia in Pakistan.

1.12 Scope and Purpose

The main purpose this thesis is to address the challenges in the early detection of thalassemia carriers, particularly in the context of Pakistan. By developing an accurate and reliable approach for thalassemia detection using blood images, this research aims to improve the efficiency of thalassemia screening programs. Ultimately, the goal is to contribute to the improvement of thalassemia management strategies, leading to better health outcomes for individuals affected by this genetic disorder in Pakistan.

The scope is to explore different machine learning and image processing techniques for the detection of thalassemia using blood smear images. The process includes preprocessing of images, feature extraction, selection of appropriate machine learning models, and evaluation of the proposed approach. The scope also encompasses the development of an algorithm that can efficiently analyze blood images and provide accurately detects the thalassemia traits.

1.13 Challenges

The recent advancements in image processing specially in the field of medical image processing make it challenging for researchers to come up with an effective and efficient methodology. Some of the challenges faced during this research are listed below:

1.13.1 Image Quality

Some factors like lighting conditions, resolution, and focus vary the quality of blood images used for thalassemia detection. Ensuring consistent image quality across the dataset is crucial for the accuracy of the detection model.

1.13.2 Dataset Size and Diversity

A large and diverse dataset of blood images was required for building a robust detection model. However, acquiring and annotating such a dataset, especially with rare conditions like thalassemia, can be challenging.

1.13.3 Feature Extraction

Identifying relevant features from blood cell images that can effectively differentiate between normal and thalassemic cells was another challenge faced in this research. For getting the high accuracy, the model requires most informative features which was a crucial task.

1.13.4 Model Selection and Optimization

Choosing the appropriate machine learning model and optimizing its parameters for thalassemia detection can be challenging. Different models perform differently based on the dataset and feature representation. For detection of thalassemia selecting an appropriate machine learning model and optimizing its parameters was one of the challenges faced during this research.

1.14 Thesis Structure

This thesis consists of 6 chapters:

- In chapter 2, the identical work has been discussed performed in the past 4 years. Along with this, we have discussed the limitations of those works.
- Chapter 3 demonstrate different architectures and datasets in detail which are used in our proposed work. Also, proposed models have been presented.
- Chapter 4 describes the classification techniques.
- In chapter 5, all the results are explained that are obtained from experiments.
- In chapter 6 the conclusion and future work are explained.

Chapter 2

LITERATURE REVIEW

Several key approaches and methodologies were highlighted from the literature on thalassemia diagnosis. Some researchers have focused on combining clinical reports with blood smear images for detection, using both clinical tools and deep learning algorithms feature extraction, and applying machine learning algorithms such as Naive Bayes, random forest, and K-NN for classification. Some studies have also explored image segmentation using deep learning architectures like U-net, transfer learning techniques, and data engineering methods to enhance accuracy. Other methodologies include image preprocessing, erythrocyte segmentation, and feature extraction using various statistical methods and classifiers, achieving significant accuracy in classifying different types of abnormal and normal erythrocytes. Additionally, research has been conducted on automated assessment using hemoglobin electrophoresis images, utilizing deep convolutional neural networks and transfer learning to achieve high accuracy and efficiency in detecting thalassemia. These studies collectively demonstrate the diverse approaches and advancements in thalassemia diagnosis, showcasing the integration of clinical and technological innovations to improve detection and management of this genetic disorder.

2.1 Literature Review

In order to overcome the difficulty in diagnosing thalassemia, a hereditary illness, Shikha Purwar et al.'s research suggests a method that combines clinical reports with blood smear images for identification. Using a blood analyzer, the authors extract clinical features, and using a deep convolutional neural network (CNN), they extract visual features. After fusing these features to produce a useful feature set, principal component analysis (PCA) is used to cut down on computational cost and redundancy. For classification, machine learning methods with high levels of sensitivity, specificity, and accuracy are utilised, such as Naive Bayes, random forest (RF), and K-NN. Using the thalassemia picture characteristics, classification may be done with up to 81.5% accuracy and 0.85 AUC on the ROC curve [2].

Amira J. Zaylaa* et al.'s research described the creation and assessment of a supervised semantic picture segmentation model utilising the U-net architecture. In addition to data engineering techniques like data annotation, augmentation, pre-processing, and preparation, transfer learning techniques were applied. Additionally, Prediction Time Augmentation (PTA) was used to increase prediction accuracy. The study's quantitative findings demonstrated that, for thalassemia prediction, the mean Intersection Over Union (IoU) score was 88% with PTA and 82% without PTA. It was discovered that there was an inverse relationship between the combined loss score metric and the Thalassemia prediction. The qualitative findings demonstrated that the final Thalassemia prediction designated other unknown cells as the background and concentrated on Codocytes, or target cells. Compared to the initial annotated ground truth, the produced images were more streamlined and less dense [3].

The Izyani Ahmad, et al. proposed a methodology that involves image preprocessing techniques, erythrocyte segmentation, and extraction of morphological features. The Fourier descriptor, Hue's moment, Zernike moment, and geometrical features are among the morphological features taken into consideration. Images from blood smears taken from healthy people, IDA patients, and Thalassemia patients were used in the study. The images were captured under a light microscope and digitized for analysis. The erythrocytes were segmented using connected component labeling, and morphological features were extracted from the segmented cells.Two separate experiments were conducted, one with 24 morphological features and another with 31 features. The features were analyzed and compared using statistical methods. Using the features that were retrieved, the effectiveness of several classifiers was assessed, such as logistic regression, radial basis function network, multilayer perceptron, Naïve Bayes classifier, and classification and regression tree. The results showed that the best subsets of features achieved an accuracy of 83.5%, sensitivity of 83.5%, and positive predictive value of 83.3% using logistic regression. The study concluded that the proposed methodology could effectively classify different types of abnormal and normal erythrocytes in IDA and Thalassemia [4].

The paper proposed by Wishwas Sharma, et al. addresses the detection of sickle cell anemia and thalassemia, two prevalent genetic disorders which affect approximately 3.2 million people worldwide. They proposed a method which acquires thin blood smear microscopic images, then preprocess these images using a median filter and segmenting overlapping erythrocytes using marker-controlled watershed segmentation. They used morphological techniques to improve the images, and then they retrieved attributes like aspect ratio, radial signature, metric value, and variance. To identify the three distinct erythrocyte morphologies—elliptocytes, dacrocytes, and sickle cells—that cause sickle cell anaemia and thalassemia, the K-nearest neighbour classifier is trained on 100 images. The algorithm enhances the speed, effectiveness, and efficiency of training and testing, achieving an accuracy of 80.6% and sensitivity of 87.6% [6].

The study proposed by Saima Sadiq, el al. addresses β-Thalassemia carriers detecting challenge. β-Thalassemia a prevalent inherited blood disease with significant implications for offspring if both parents are carriers. Prenatal screening after counseling of couples is crucial for controlling this disease. They suggested utilising the complete blood count test's red blood cell indices to quickly and affordably screen patients. They contrast their suggested model with expensive, time-consuming, and equipment-required high-performance liquid chromatography (HPLC) tests. They present an ensemble model called "SGR-VC" by combining the Random Forest, Gradient Boosting Machine, and Support Vector Machine algorithms. With an accuracy of 93%, comparative study shows that the SGR-VC model is quite successful in β-Thalassemia carrier screening [7].

The paper proposed by Salman Khan et, al. presents a novel method for automated assessment of thalassemia using hemoglobin (Hb) electrophoresis images. The proposed model aiming to assist expert hematologists, particularly in developing countries where their numbers are limited and workload is high. The study includes a large dataset of Hb electrophoresis images from 824 subjects, with a total of 524 images from 103 strips, ensuring clear consensus on the quality of electrophoresis. The two primary components of the suggested methodology are as follows: (1) lane extraction technique-based segmentation of single-patient electrophoresis images, and (2) state-of-the-art deep convolutional neural networks (CNNs) and transfer learning-based binary classification (normal or abnormal) of the images. After a variety of CNN models were assessed, the best models for thalassemia detection were InceptionV3 and MobileNetV2, followed by ResNet18, ResNet50, ResNet101, DenseNet201, SqueezeNet, and MobileNetV2. Accuracy, precision, recall, f1-score, and specificity for InceptionV3 were 95.8%, 95.84%, 95.8%, 95.8%, and 95.8%, respectively. In contrast, MobileNetV2 showed comparable performance with 95.72%, 95.73%, 95.72%, 95.7%, and 95.72%, for accuracy, precision, recall, and specificity. Because MobileNetV2 is a shallow network, it can process single-patient images with low latency, which makes it appropriate for mobile applications. The proposed approach offers high classification accuracy and is expected to facilitate rapid and robust detection of thalassemia using Hb electrophoresis images [8].

To accomplish high-quality segmentation, Nabeel J. et al. suggested combining pretreatment procedures and image processing approaches. Eleven color spaces, six filters, three techniques for enhancing contrast, and the fuzzy c-means and K-means segmentation algorithms are all explored. To assess the research segmentation performance, they employed five assessment metrics derived from ground truth photographs. This work focuses on the automatic red blood cell segmentation from microscopic blood smear images, specifically examining how thalassemia affects red blood cell shape. Photoshop is used to provide ground truth in novel ways for multi-object sensing (RBC cells). To optimize image processing steps, local image datasets from thalassemia patients and samples of normal blood cells were employed. Images were taken in various light conditions and with and without a yellow filter. With medium light intensity and no yellow filter, the study produced the greatest results, with an accuracy of 0.91±0.14 and a performance of 95.34% while capturing images on a microscope slide [9].

The paper proposed by J. Rodellar et, al. presents a methodology for automatic recognition of blood cells, that focuses on malignant lymphoid cells and blast cells. The methodology involves segmentation, extracting quantitative features for each region of interest (nucleus and cytoplasm), and using supervised machine learning for classification. The segmentation process achieves 98.9% efficiency, extracting 150 relevant descriptors from an initial 2464. The SVM classifier used for classification and achieves an overall accuracy of 90.3% for classifying seven groups of abnormal lymphoid cells and normal lymphocytes. This approach integrates segmentation, feature extraction, and classification to accurately recognize blood cell images, particularly targeting specific cell types [10].

Table 2.1:Comparison table of previous papers along with the methodologies applied and the results achieved from the methodologies.

Study	Methodology	Key Findings
Shikha Purwar, et al.	Proposed a model which combines clinical reports with blood smear images, extracted features using a CNN, fused features, and used ML algorithms for classification.	Achieved 81.5% accuracy and 0.85 AUC in ROC curve.
Amira J. Zaylaa* et al.	Developed a supervised semantic image segmentation model using U-net, transfer learning, and data engineering methods.	Mean IoU score for Thalassemia prediction was 88% with PTA and 82% without PTA.
Izyani Ahmad, et al.	Used image preprocessing, erythrocyte segmentation, and extraction of morphological features.	Achieved an accuracy of 83.5% using logistic regression.
Wishwas Sharma, et al.	Preprocessed images, segmented erythrocytes, extracted features, and used K-nearest neighbor classifier for detection.	Achieved an accuracy of 80.6% and sensitivity of 87.6%.
Saima Sadiq, el al.	Proposed a cost-effective screening technique using red blood cell indices and an ensemble model named “SGR-VC”.	Achieved an accuracy of 93% in β-Thalassemia carrier screening.
Salman Khan et, al.	Developed an automated assessment method using Hb electrophoresis images, involving segmentation and binary classification with CNNs.	InceptionV3 achieved 95.8% accuracy, while MobileNetV2 achieved 95.72% accuracy in detecting thalassemia.

Nabeel J. et al.	Explored various image processing techniques for segmentation and used Fuzzy c-means and K-means algorithms.	Achieved an accuracy of 0.91±0.14 and a performance of 95.34% in image slide capture utilizing a microscope.
J. Rodellar et, al.	Developed a methodology for automatic recognition of blood cells, involving segmentation, feature extraction, and SVM classification.	Achieved an overall accuracy of 90.3% for classifying seven groups of abnormal lymphoid cells and normal lymphocytes.

2.2 Discussion

The above table shows implementation of different types of machine learning and deep learning algorithms and the results obtained from these models. Some researchers extract more relevant feature from blood smear images and apply ML algorithms for classification. Image processing techniques were applied for image segmentation and morphological features were extracted. More often it is observed that random forest model works efficiently in the classification.

Chapter 3

RESEARCH METHODOLOGY

In this chapter the dataset is described in detail also the proposed model is explained briefly. In this chapter the proposed system is also explained along with the nine types of blood cell shapes briefly. These nine shapes are used to in the proposed system as a dataset for the detection of thalassemia carrier.

3.1 Available Datasets

There are two datasets available related to blood smear images in thalassemia case and these are as follows:

3.1.1 Erythrocyte (red blood cell) dataset in thalassemia case

The dataset comprised a total of 7108 images of individual red blood cells, representing nine different cell types. Size of the images are 800 x 600 pixels. The data set is publically available.

3.1.2 Blood Cell Images

12,500 augmented images of blood cells (JPEG) with corresponding cell type labels (CSV) are included in this dataset. The dataset is publically available on Kaggle.

I choose Erythrocyte (red blood cell) dataset in thalassemia case dataset because this data is already pre-processed, explained below in the pre-processing section.

3.2 Erythrocyte (red blood cell) dataset in thalassemia case

The data collection included 7108 images of single red blood cells, which corresponded to nine distinct cell types. The images have a resolution of 800 by 600 pixels. The dataset can be accessed by everyone. The Olympus CX21 microscope is used to extract microscopic image data from peripheral blood smears using an Optilab Advance Plus camera. This is the first step in the pre-processing of the data. To lessen the computational strain, images are downsized from 4100 × 3075 pixels (RGB images) to 800 × 600 pixels after they are acquired [5]. For grayscale, the data is formatted in.png. This dataset contains nine different types of cells, listed in order of the number of images:

Table 3.1:Nine types of cells in Thalassemia case

Cell Type	No. of Images	Percentage (%)
Elliptocyte cell (elliptocyte, ovalocyte)	1211	17.04
Pencel Cell	24	0.34
Tear Drop Cell	2076	29.02
Acanthocyte Cell	354	4.98
Stomatocyte Cell	382	5.37
Target Cell	851	11.97
Spherocyte	562	7.91
Hypochromic Cell	222	3.12
Normal Cell	1426	20.06
Total	7108	100

3.3 Different Types of Blood Cells in Thalassemia Case

The following nine different cell types are often seen in thalassemia that can provide valuable diagnostic information when evaluating blood smear samples under a microscope.

3.3.1 Elliptocyte cell (elliptocyte, ovalocyte)

The red blood cells that are elongated and elliptical in shape, rather than the typical round shape are known as Elliptocyte cell (elliptocyte, ovalocyte). These cells can be often seen in various conditions, including thalassemia.

3.3.2 Pencil Cell

The red blood cells are known as Pencil cells that are elongated and thin, resembling a pencil in shape. They can be a characteristic feature of certain types of thalassemia.

3.3.3 Tear Drop Cell

The red blood cells are known as Teardrop cells which shaped like teardrops. Due to the abnormal shape of red blood cell precursors, they are often associated with bone marrow disorders, including thalassemia.

3.3.4 Acanthocyte Cell

These are also known as spur cells. Acanthocytes are the red blood cells that have spiny projections on their surface. Owing to changes in the cell membrane's lipid composition, they are often seen in certain types of anemia, including thalassemia.

3.3.5 Stomatocyte Cell

Stomatocyte cells are red blood cells with a mouth- or slit-like morphology. These cells are found due to the changes in the cell membrane, and these cells can be seen in various conditions, including thalassemia.

3.3.6 Target Cell

These cells also known as codocytes. The red blood cells that have a bull's-eye or target-like appearance when viewed under a microscope. They can be seen in various conditions, including thalassemia. These cells appears with changes in the hemoglobin content of the cell.

3.3.7 Spherocyte

The red blood cells that are smaller and more spherical than normal are known as Spherocyte. These cells are often seen in certain types of anemia, including thalassemia. These cells supposed to be due to changes in the cell membrane that make the cell more fragile.

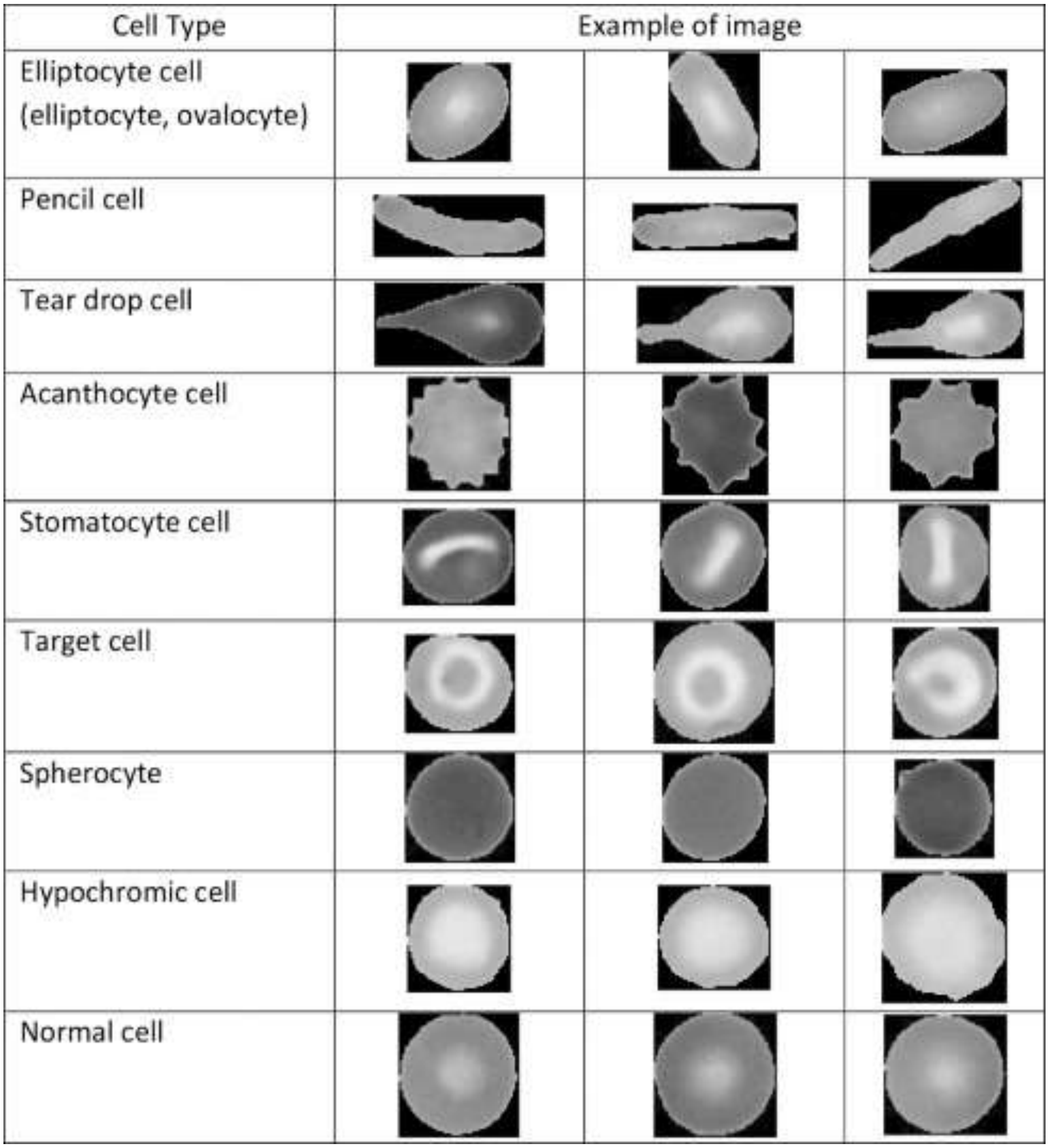
3.3.8 Hypochromic Cell

The red blood cells that have a lower than normal concentration of hemoglobin (the molecule that carries oxygen in the blood) are known as Hypochromic. They are often seen in various types of anemia, including thalassemia, where there is a decreased production of hemoglobin.

3.3.9 Normal Cell

Normal red blood cells have a round, biconcave shape and a uniform appearance. In thalassemia, the number of normal red blood cells may be decreased, leading to anemia and the appearance of various abnormal cell types.

The following figure shows nine different types of cells along with their shapes.



Cell Type	Example of image		
Elliptocyte cell (elliptocyte, ovalocyte)			
Pencil cell			
Tear drop cell			
Acanthocyte cell			
Stomatocyte cell			
Target cell			
Spherocyte			
Hypochromic cell			
Normal cell			

Figure 3.1:Nine types of cells in Thalassemia case with their shapes

3.4 Proposed Solution

The proposed system will automatically and accurately classify whether the patient is normal or affected by thalassemia. The suggested model's architecture is as follows:

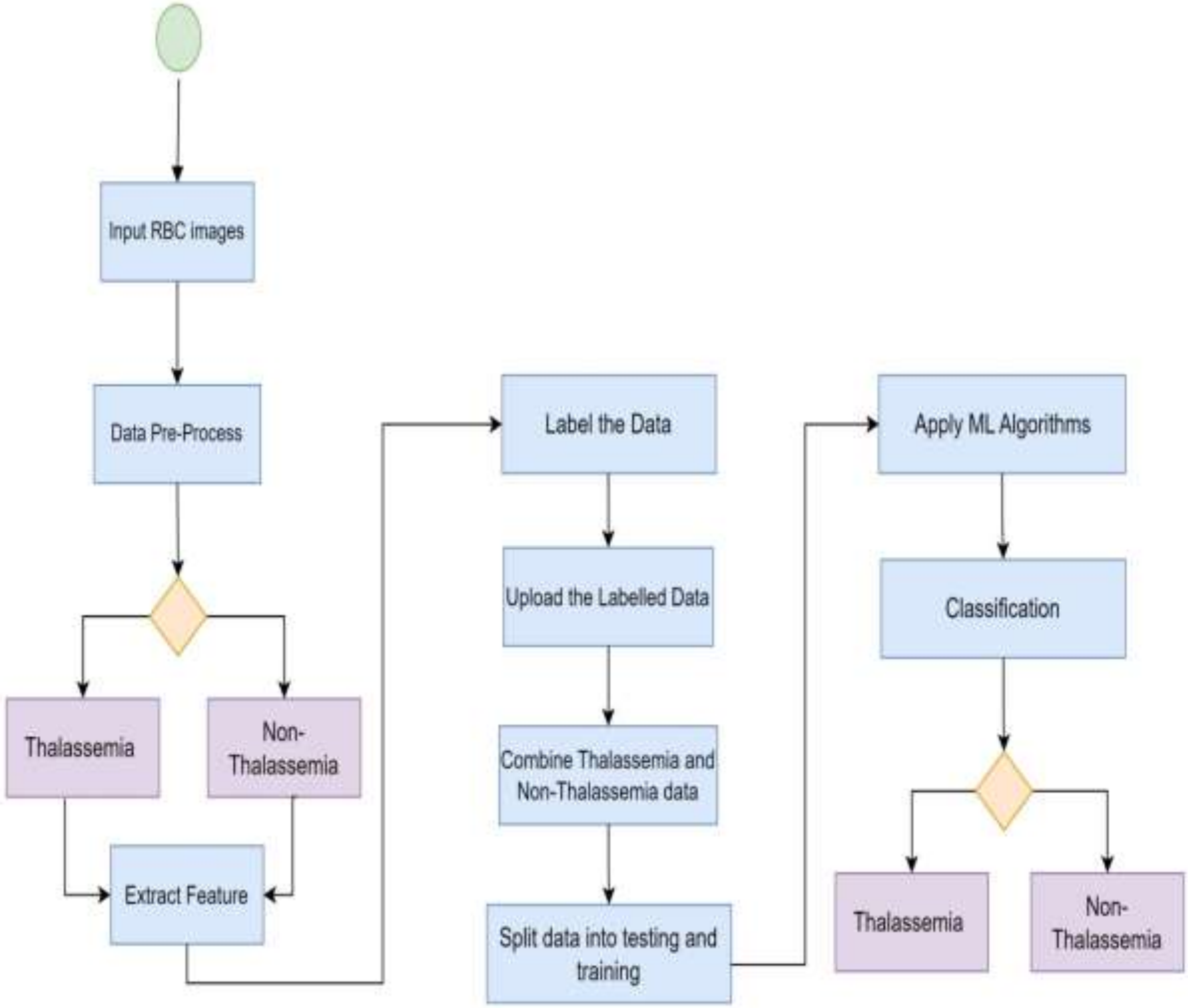


Figure 3.2:Proposed System methodology which takes RBC images as an input, pre-process the images to extract features and apply ML models to classify whether the input is thalassemia carrier or normal

3.4.1 Input

The system takes a red blood cells image as an input. This image serves as the primary data source for analyzing and classifying thalassemia carrier.

3.4.2 Import necessary libraries:

- cv2 (OpenCV): To perform image processing tasks.
- numpy (np): To perform numerical operations.
- warnings: To suppress warnings.
- skimage.feature: For texture feature extraction (GLCM - Grey Level Co-occurrence Matrix).
- pandas (pd): For creating and manipulating data frames.
- google.colab.files: For uploading and downloading files in Google Colab.

3.5 Pre-Processing

The pre-processing step for the red blood cell images in this study involved several key stages to prepare the images for segmentation and analysis. First of all the images were reduced in size from 4100 x 3075 pixels to 800 x 600 pixels. The purpose of this downsizing was to lighten the computational burden in the later phases of the procedure. Subsequently, the RGB images' green color component, or green channel, was isolated and utilized for additional processing. A number of image enhancement methods, including median filtering, canny edge recognition, dilatation, and hole filling, were also incorporated in the pre-processing to raise the image quality and make the red blood cells more visible. In addition, the images were subjected to erosion and the removal of small items whose area was less than 500 pixels, after a WDT (Watershed) procedure to separate overlapping erythrocytes. Finally, cells at the edge of the images, which had incomplete cell shapes, were deleted from the dataset. These pre-processing steps were essential to ensure the accuracy and reliability of the segmentation process, which is crucial for the subsequent analysis and classification of the red blood cells [5]. These pre-processing steps are already annotated in the dataset.

3.6 Feature Extraction

First the thalassemic images were uploaded to extract the features simultaneously followed by non-thalassemic grayscale images. From the input image, the following three categories of features are extracted:

3.6.1 Color Features:

The system determines the grayscale image's mean and standard deviation, which are basic statistical measures of the pixel intensities in the image.

3.6.2 Texture Features:

The Grey Level Co-occurrence Matrix is used to extract texture features (GLCM). The frequency of various pixel intensity combinations in an image is described by GLCM. From the GLCM, the code computes four texture features: contrast, dissimilarity, homogeneity, and energy. These features provide information about the texture patterns in the image.

3.6.3 Shape Features:

Shape features are extracted by calculating the contour area of the objects in the binary image. Contours are the boundaries of objects in an image, and the contour area provides information about the size and shape of these objects.

The feature extraction working is as follows:

Define a function extract_features(image):

- Converts the input image to grayscale.
- Computes color features (mean and standard deviation) of the grayscale image.
- Computes texture features (contrast, dissimilarity, homogeneity, energy) using GLCM.
- Computes shape features (contour area) of the grayscale image.
- Returns a list of extracted features.

Upload images from the local computer to Google Colab:

- Uses files.upload() to upload image files.

Extract features for each uploaded image:

- Iterates over each uploaded image, decodes it, and extracts features using the extract_features function.
- Appends the extracted features to a list (data).

Create a DataFrame to store the features:

- Defines column names for the DataFrame.
- Creates a DataFrame (df) using the extracted features and column names.

Save the DataFrame to a CSV file:

- Saves the DataFrame (df) to a CSV file named features.csv without an index.

Download the CSV file:

- Uses files.download() to download the CSV file (features.csv) to the local computer.

Once the features are extracted, a feature set for thalassemic and non-thalassemic case are derived. Two csv labeled files for thalassemia and normal case are generated respectively. Now different machine learning (ML) models are applied to improve the efficiency and effectiveness of the subsequent classification process.

3.7 Upload the Data

Multiple ML models are applied on the feature set to automate the process of thalassemia classification using erythrocyte (red blood cell) images. Follwing is the code to upload the feature set csv files:

```
from google.colab import files
```

Importing library for file uploading in google colab.

```
import pandas as pd
```

Importing library for working with data in dataframes.

Upload the CSV files

```
uploaded_thalassemia = files.upload()
```

This line used to upload csv feature set file of thalassemia patients.

```
uploaded_non_thalassemia = files.upload()
```

This line used to upload csv file containing non-thalassemia feature set data.

Load the CSV files into pandas DataFrames.

```
thalassemia_df = pd.read_csv(next(iter(uploaded_thalassemia)))
non_thalassemia_df = pd.read_csv(next(iter(uploaded_non_thalassemia)))
```

Loading both the csv files into DataFrames.

Combine the DataFrames.

```
combined_df = pd.concat([thalassemia_df, non_thalassemia_df], ignore_index=True)
```

Combining the two DataFrames into single DataFrame.

Split the data into features (X) and labels (y)

```
X = combined_df.drop('Label', axis=1)
y = combined_df['Label']
```

Splitting the data into features (X) and labels (y).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

The model will be trained using 80% of the data, with the remaining 20% designated for testing. In order to partition the data consistently each time the code runs, the random state is set to 42.

3.8 Classification Techniques

Based on the patterns recognized in the data when an algorithm learns to assign labels or categories to input data this process is known as classification. Classification is a machine learning and deep learning process. Previous researchers used different types of classification models for the detection of thalassemia.

3.9 Apply ML Models

Now once the data is ready for testing and training the four following five ML models were applied. These ML models were selected based on the previous researcher's work, they also used these models for the thalassemia classification purpose.

- Logistic Regression.
- Decision Tree.
- Random Forest.
- Support Vector Machine (SVM).
- Neural Network (CNN).

3.10 Logistic Regression

For binary classification tasks, the most suitable and powerful model is Logistic Regression, where the output variable takes on only two distinct values. It models the probability that a specific input instance falls into a specified category. The odds calculated by the model when an event occur and logistic regression converts these odds to probability. It's a simple yet effective algorithm that is easy to interpret and implement, making it a popular choice for baseline classification tasks.

3.11 Decision Tree

Because decision trees are a non-parametric supervised learning technique, they are employed for both regression and classification applications. Decision Tree potentially improve accuracy when working with complex data. The data is divided into subsets according to the input features, and each node indicates a choice depending on the value of the feature set. The predicted output is represented by the last node. Decision trees are helpful for comprehending the model's decision-making process since they are simple to read and visualize.

3.12 Random Forest

Several decision trees are used in the Random Forest ensemble learning technique to increase the accuracy of classification (or regression). It produces a forest of trees, each of which is trained using a random feature set and a portion of the data. For regression, the final forecast is calculated by averaging the predictions of each individual tree; for classification, the final prediction is determined by a majority vote. Random forests work effectively on a wide range of datasets and are resistant to overfitting.

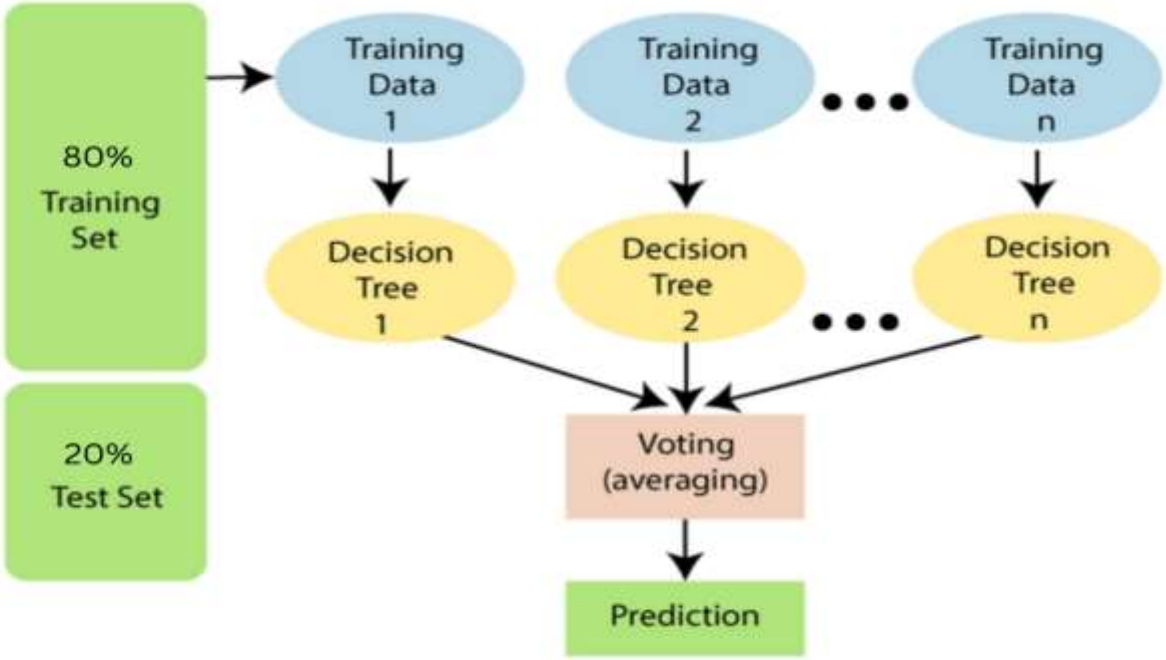


Figure 4. 1:Random Forest model with 80% of training data and 20% of testing data.

3.13 Support Vector Machine (SVM)

A powerful supervised learning approach for regression, outlier identification, and classification is called Support Vector Machine (SVM). In order to maximize the margin between the classes, it locates the hyperplane in the feature space that best divides the various classes. SVM best fits for handling high-dimensional data and find the optimal hyperplane for separating classes.

3.14 Neural Network (CNN)

A deep learning method called a neural network, more especially a convolutional neural network (CNN), was inspired by the way the human brain is organized. The main applications of CNNs are in image processing and recognition. They are made up of convolutional, pooling, and fully connected layers, among other layers of interconnected neurons. CNNs are very good at tasks like object detection, image classification, and facial recognition because they can automatically learn features from the input data and recognize intricate patterns in images.

By applying these diverse models, it is aimed to explore different approaches to achieve the best performance in detecting thalassemia from blood cell images. The most effective model is Random Forest with the highest accuracy achieved, which is explained in the Results chapter.

Now by considering this machine learning problem as an image classification problem, the following image classification algorithm were applied.

- VGG16
- ResNet50
- MobileNetV2
- EfficientNetB4
- DenseNet121

3.15 VGG16

The deep convolutional neural network architecture known as Visual Geometry Group 16 (VGG16) is renowned for its efficiency and simplicity. It has sixteen weight layers, three fully connected layers, max-pooling layers, and convolutional layers in order of precedence. The consistent architecture of VGG16, which uses 2x2 filters with a stride of 2 for the max-pooling layers and tiny 3x3 filters with a stride of 1 for the convolutional layers, is what makes it so popular. Despite its simplicity, VGG16 has shown strong performance on image classification tasks, particularly in the early days of deep learning research.

3.16 ResNet50

The concept of residual learning was first proposed by Residual Network 50, a deep neural network architecture that aids in resolving the issue of disappearing gradients in extremely deep networks. It has fifty layers total, comprising identity blocks, batch normalization layers, and convolutional layers. The key innovation in ResNet50 is the use of skip connections, or shortcuts, that skip one or more layers, allowing the network to learn residual functions. This enables training of much deeper networks, leading to improved performance on image classification tasks.

3.17 MobileNetV2

A lightweight deep neural network architecture called MobileNetV2 was created for embedded and mobile devices with constrained processing power. It is based on depthwise separable convolutions, which reduce the number of parameters and computing cost by separating the spatial and channel-wise convolutions. MobileNetV2 uses inverted residuals and linear bottlenecks to improve performance while maintaining efficiency. On devices with limited resources, it has proven to be successful for tasks including semantic segmentation, object detection, and image classification.

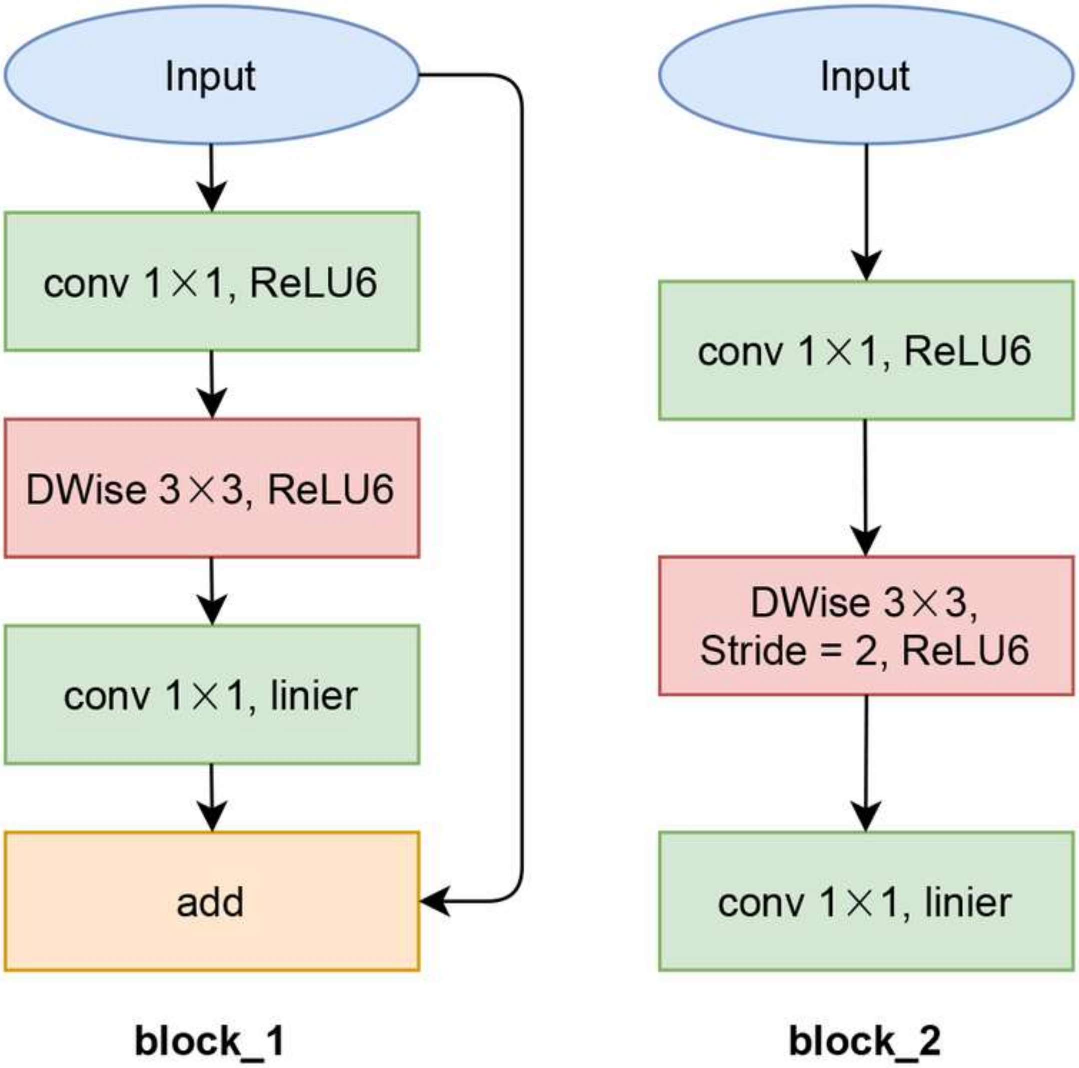


Figure 4. 2:MobileNetV4 deep learning model with activation function relu.

3.18 EfficientNetB4

Designed to deliver state-of-the-art performance with much fewer parameters and FLOPS (floating-point operations per second) compared to other models, EfficientNetB4 is a member of the EfficientNet family of neural network architectures. To maximize speed, EfficientNetB4 employs a compound scaling technique that balances the network's breadth, depth, and resolution. This allows EfficientNetB4 to achieve high accuracy on image classification tasks while being more computationally efficient than larger models.

3.19 DenseNet121

The dense connectivity pattern of DenseNet121, a convolutional neural network design, is well-known. DenseNet creates densely connected blocks by feeding forward connections between each layer and every other layer. Because of its high connectivity, gradient flow is enhanced and feature reuse is made easier, which helps to address the issue of vanishing gradients and permits the training of very deep networks. DenseNet121 has shown strong performance on image classification tasks, often outperforming other architectures with similar numbers of parameters.

By applying these diverse deep learning models, it is aimed to explore different approaches to achieve the best performance in detecting thalassemia from blood cell images. The most effective model is MobileNetV2 with the highest accuracy achieved, which is explained in the Results chapter.

Chapter 4

RESULTS AND DISCUSSION

In this chapter the results obtained from the ML models are discussed in details. Every model is discussed briefly along with the result obtained from the model.

4.1 Confusion Matrix Parameters

Following are the equations of 4 terms that were set as parameters for determining the performance of models.

Precision

When the input is positive and the system also predicts positive this ratio is known as precision.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall

Another name for recall is sensitivity. It demonstrates the model's capacity to locate every positive sample. It displays the proportion of all observations in the real class to the positive projected observations.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1 Score

The harmonic mean of recall and precision is the F1 score. Precision and recall are balanced by the f1 score. The best case for f1 score is when the value of f1 score is 1 and worst when the value is 0.

$$F1\ Score = 2\ X\ \frac{Precision * Recall}{Precision + Recall}$$

Accuracy

The proportion of accurately predicted occurrences to all instances in the dataset is known as accuracy. It provides an overall measure of how often the classifier is correct.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

4.2 Machine Learning Algorithm Results

4.2.1 Logistic Regression model

Logistic Regression is the most suitable and powerful model when working with the binary classification. In the proposed model the Logistic Regression gives the following results:

Table 5.1:Logistic Regression model results with Precision 85.6%, Recall 86.2%, F1 score 85.7% and accuracy 86.2%

Number	Parameters	Score
1	Precision	0.856022
2	Recall	0.862166
3	F1 Score	0.857752
4	Accuracy	0.862166

The table shows that the precision calculated by Logistic Regression model is approximately 0.856 which means that 85.6% the model correctly predicts a positive class of the time on average. The model calculated the sensitivity or recall by 0.862 approximately, indicating that about 86.2% of the actual positive instances identifies by the model correctly. The mean of precision and recall is F1 score and is calculated by the model 0.857. With an accuracy of 0.862, the model can accurately classify around 86.2% of the cases in the test set.

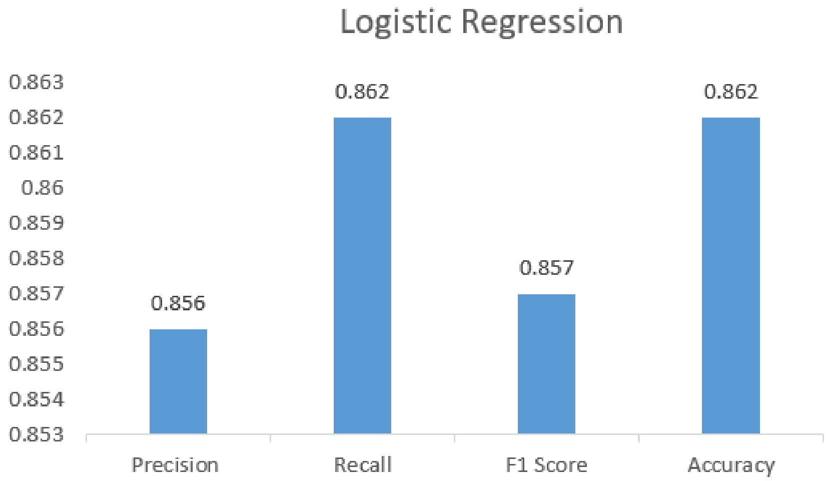


Figure 5.1:Logistic Regression with Precision 85.6%, Recall 86.2%, F1 score 85.7% and accuracy 86.2%

4.2.2 Decision Tree model

Decision tree model is used to improve classification accuracy when the the relationship of the data is complex. Following table shows the derived result from decision tree model:

Table 5.2:Decision Tree model results Precision 86.2%, Recall 85.9%, F1 score 86% and accuracy 85.9%

Number	**Parameters**	**Score**
1	Precision	0.862562
2	Recall	0.859353
3	F1 Score	0.860791
4	Accuracy	0.859353

The decision tree model gives precision of approximately 0.863, indicating 86.3% correctly prediction of positive class. The model achieved recall score of around 0.859, suggest that 85.9% of the actual positive instances is accurately identified. The model yielded a nearly 0.861 F1 score. The model's accuracy of 0.859 indicates that 85.9% of the test set instances are accurately specified by the model.

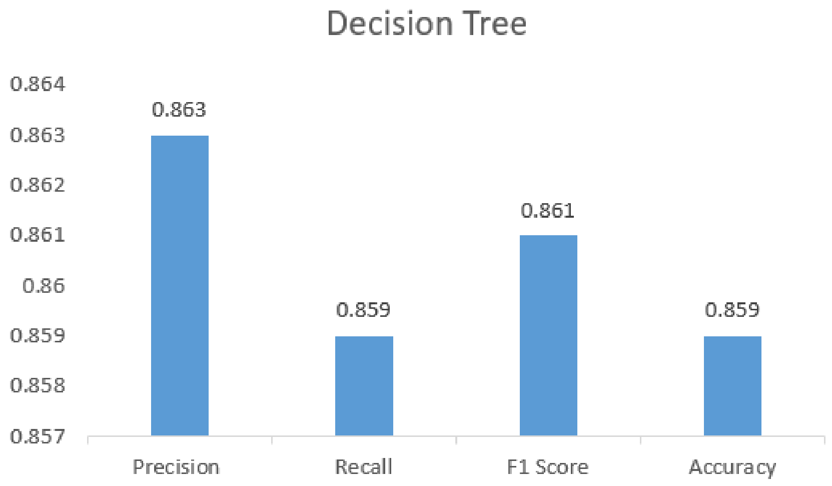


Figure 5.2:Decision Tree with Precision 86.2%, Recall 85.9%, F1 score 86% and accuracy 85.9%

4.2.3 Random Forest model

Random forest works better when the data is unbalance or when the complexity of the data is high. The random forest model gives the following results:

Table 5.3:Random Forest model results Precision 90.9%, Recall 91.0%, F1 score 90.9% and accuracy 91.1%

Number	Parameters	Score
1	Precision	0.909174
2	Recall	0.910689
3	F1 Score	0.909751
4	Accuracy	0.910689

The table shows that the precision calculated by Random-forest model is approximately 0.91 which means that 91% the model correctly predicts a positive class of the time on average. The model calculated the sensitivity or recall by 0.911 approximately, indicating that about 91.1% of the actual positive instances identifies by the model correctly. The mean of precision and recall is F1 score and is calculated by the model approximately 0.91. The model shows accuracy of 0.911, indicating that the model correctly classifies about 91.1% of the instances in the test set.

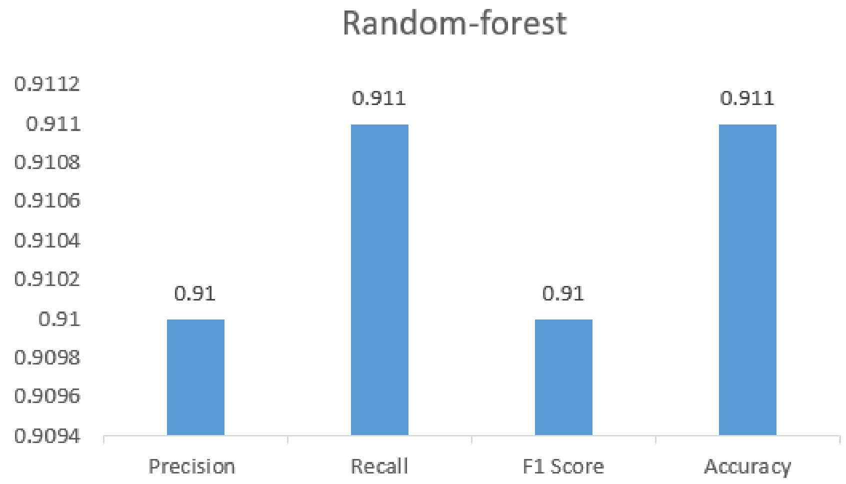


Figure 5.3:Random Forest with Precision 90.9%, Recall 91.0%, F1 score 90.9% and accuracy 91.1%

4.2.4 Support Vector Machine (SVM) model

SVM best fits for handling high-dimensional data and find the optimal hyperplane for separating classes. The following table shows the results obtained from SVM model:

Table 5.4:Support Vector Machine (SVM) model results Precision 83.4%, Recall 84.3%, F1 score 83.6% and accuracy 84.3%

Number	Parameters	Score
1	Precision	0.834452
2	Recall	0.843882
3	F1 Score	0.836688
4	Accuracy	0.843882

In the propose system the SVM gives precision of approximately 0.834, indicating 83.4% correctly prediction of positive class. With a recall score of about 0.843, the model is able to correctly identify 84.3% of the real positive events. The model yielded an approximate F1 score of 0.837. The model accurately specifies 84.3% of the test set instances, according to its accuracy of 0.843.

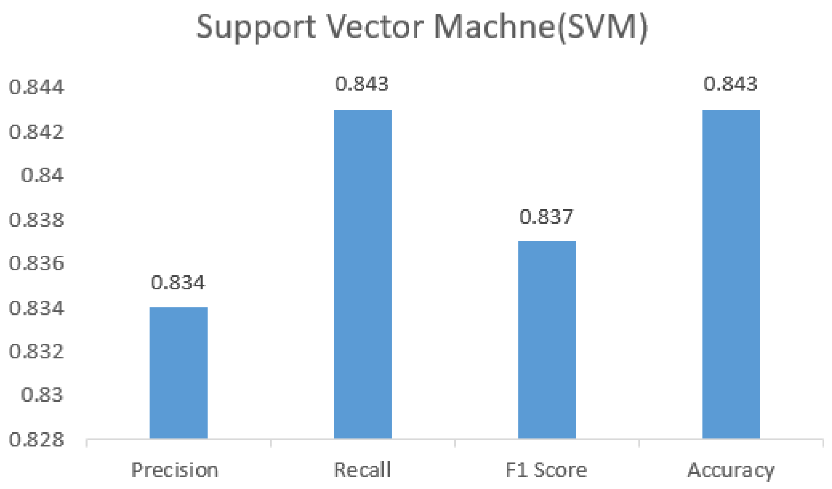


Figure 5.4:Support Vector Machine (SVM) Precision 83.4%, Recall 84.3%, F1 score 83.6% and accuracy 84.3%

4.2.5 Neural Network (CNN) model

After applying the CNN model the following results were obtained.

Table 5.5:Neural Network (CNN) model results Precision 87.5%, Recall 87.8%, F1 score 86.7% and accuracy 87.8%

Number	Parameters	Score
1	Precision	0.875522
2	Recall	0.878340
3	F1 Score	0.867246
4	Accuracy	0.878340

The precision calculated by CNN model is approximately 0.875 which means that 87.5% the model correctly predicts a positive class of the time on average. The model calculated the sensitivity or recall by 0.878 approximately, indicating that about 87.8% of the actual positive instances identifies by the model correctly. The mean of precision and recall is F1 score and is calculated by the model approximately 0.867. The model shows accuracy of 0.878, indicating that the model correctly classifies about 87.8% of the instances in the test set.

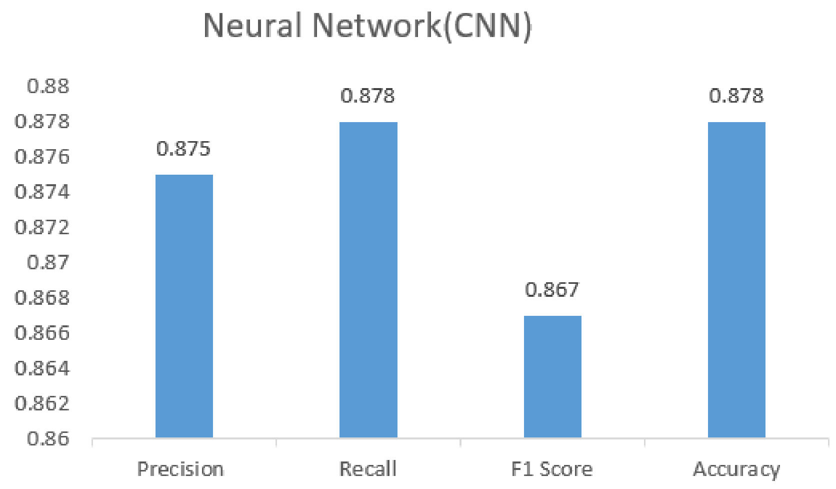


Figure 5.5:Neural Network (CNN) Precision 87.5%, Recall 87.8%, F1 score 86.7% and accuracy 87.8%

Table 5.6:Comparison of different machine learning classifier, the random forest shows the highest accuracy of 91.1%, followed by CNN with 87.8% accuracy and the minimum accuracy of 84.3% gained by SVM

Numbe r	**Parameters**	**Logistic Regression**	**Decision Tree**	**Random-forest**	**SVM**	**CNN**
1	Precision	0.856022	0.862562	0.909174	0.834452	0.875522
2	Recall	0.862166	0.859353	0.910689	0.843882	0.878340
3	F-measure	0.857752	0.860791	0.909751	0.836688	0.867246
4	Accuracy	86.2%	85.9%	**91.1%**	84.3%	87.8%

4.3 Deep Learning Algorithm Results

The parameters are set same for all the models. To train the deep learning models effectively these parameters are very cruicial. Following are the parameters set for all the deep learning models:

- weights='imagenet': The model is initialised using pre-trained ImageNet weights.
- include_top=False: It does not include the topmost completely connected levels of the network.
- input_shape=(224, 224, 3): The input shape for the images, with a height and width

of 224 pixels and 3 channels (RGB).

4.3.1 Activation Function

relu: The Dense layer's 256-unit activation function, or Rectified Linear Unit.
sigmoid: The output layer uses the sigmoid activation function to generate binary classification probabilities.

4.3.2 Optimizer

adam: Adaptive Moment Estimation optimizer, used for updating network weights based on training data to minimize the loss function.

4.3.3 Loss Function

binary_crossentropy: The difference between expected and actual values is measured in binary classification tasks using the binary cross-entropy loss function.

4.3.4 Metrics

accuracy: Metric used to assess how well the model performs during testing and training, measuring the proportion of correct predictions.

4.3.5 Data Augmentation

Generates batches of augmented data to increase the diversity of the training set and reduce overfitting. Parameters include:
Rotation_range=20, the degree range for random rotations, is one of the parameters.
Fraction of total width for horizontal shifts is width_shift_range=0.2.
Fraction of total height for vertical shifts is height_shift_range=0.2.
shear_range = 0.2: Shear intensity, expressed as a radian shear angle.
zoom_range = 0.2 is the random zoom range.
horizontal_flip =True: Flip inputs horizontally at random.
fill_mode='nearest': Strategy for filling in newly created pixels.

4.3.6 Learning Rate Scheduler

ReduceLROnPlateau: When a metric ceases to improve, it lowers the learning rate. Among the parameters are:
monitor='val_loss': The amount (val_loss for validation loss) that has to be tracked.

factor=0.2: The factor that will lower the learning rate.

patience=3: The number of periods in which there is no improvement, following which the learning rate is lowered.

Lower bound on the learning rate is min_lr=1e-6.

4.4 Deep Learning Models Result

The following table shows the result of five deep learning models applied in this research:

Table 5.7:Comparison table of different DL models with the highest accuracy gained by MobileNetV2 of 90%, followed by DenseNet121 with 84% accuracy and the minimum accuracy gained by EfficientNetB4 of 49%

DL Model	**Parameter (Accuracy)**
VGG16	69%
ResNet50	71.6%
MobileNetV2	**90%**
EfficientNetB4	49%
DenseNet121	84%

The maximum accuracy obtained by MobileNetV2 which is 90% followed by DenseNet121 model with the second highest accuracy achieved of 84%. ResNet50 model achieved accuracy of approximately 72%. The minimum accuracy gained by the EfficientNetB4 model which is 49%.

Table 5.8:Comparison table of different ML and DL models applied in this study along with results, it is observed that in ML models random forest has the highest accuracy of 91.1% while in DL models the MobileNetV2 has the highest accuracy of 90%.

Model	**Type**	**Accuracy Score**
Logistic Regression	ML	86.2%
Decision Tree	ML	85.9%
Random Forest	ML	**91.1%**
SVM	ML	84.3%

CNN	ML	87.8%
VGG16	DL	69%
ResNet50	DL	71.6%
MobileNetV2	DL	**90%**
EfficientNetB4	DL	49%
DenseNet121	DL	84%

Comparing the results among ML models and DL models, it is observed in DL models the highest accuracy (90%) achieved by that the MobileNetV2 model. While on the other hand among the traditional machine learning models the Random Forest model achieved the highest accuracy (91.1%). The CNN model also performed well with an accuracy of 87.8%. However, the EfficientNetB4 model had a lower accuracy of 49%, indicating that it performed poorly compared to the other models.

Table 5.9:Results comparison table of proposed model with previous models

Author	**Model/Technique**	**Parameter (Accuracy)**
Shikha Purwar, et al.	Random Forest	Max. 81.5% accuracy, 0.85 AUC in ROC curve
Izyani Ahmad, et al.	Logistic Regression	83.5% accuracy, sensitivity, and positive predictive value
Proposed System	Random Forest, Logistic Regression	91.1% accuracy with RF model and 86.2% accuracy with Logistic Regression.

The above table shows that the proposed system improves the accuracy in both the RF model and Logistic Regression model. The model proposed by Shikha Purwar, et al. using the RF model has achieved maximum accuracy of 85% while the proposed system achieved accuracy of 91.1%. On the other hand applying the Logistic Regression, the model proposed by Izyani Ahmad, et al. has achieved 83.5% accuracy while the proposed model achieved accuracy of 86.2% by applying the Logistic Regression model.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this thesis we have performed classification of thalassemia using blood smear images. First of all the erythrocyte (RBC) image is processed to extract features. Three types of features were extracted; texture features, area and color features for both thalassemic blood cells and non-thalassemic blood cells. After the feautre extraction, five machine-learning models were applied on these feature set including; Logistic Regression, Decision Tree, Random-forest, support vector machine (SVM) and neural networks (CNN). The maximum accuracy of 91.1% was achieved by the Random-forest in the applied ML models. The second highest accuracy was obtained from CNN respectively. The five image classification algorithm were also applied including; VGG16, ResNet50, MobileNetV2, EfficientNetB4 and DenseNet121. The MobileNetV2 achieved the highest accuracy of 90% in the applied DL models.

5.2 Future Work

The proposed system reaches 91.1% accuracy. For future work different types of features should be combined to achieve high accuracy.

- By combining the morphological and texture features can also increase the accuracy of different ML and DL models.
- By applying hybrid model of CBC features and blood smear images features can also increase the accuracy of the models.

REFERENCES

[1] Zaheer, H. A., Waheed, U., Abdella, Y., & Konings, F. (2020). Thalassemia in Pakistan: A forward-looking solution to a serious health issue. Global Journal of Transfusion Medicine, 5(1), 108.

[2] Purwar, S., Tripathi, R. K., Ranjan, R., & Saxena, R. (2021). Classification of Thalassemia Patients Using a Fusion of Deep Image and Clinical Features.

[3] Zaylaa, A. J., Makki, M., & Kassem, R. (2022). Thalassemia Diagnosis Through Medical Imaging: A New Artificial Intelligence-Based Framework.

[4] Ahmad, I., Abdullah, S. R. S., & Sabudin, R. Z. a. R. (2018). Morphological Features Analysis for Erythrocyte Classification in IDA and Thalassemia. International Journal of Advanced Computer Science and Applications.

[5] Tyas, D. L., Hartati, S., Harjoko, A., & Ratnaningsih, T. (2020). Morphological, Texture, and Color Feature Analysis for Erythrocyte Classification in Thalassemia Cases. IEEE Access, 8, 69849–69860.

[6] Sharma, Vishwas, et al. “Detection of sickle cell anaemia and thalassaemia causing abnormalities in thin smear of human blood sample using image processing.” *2016 International Conference on Inventive Computation Technologies (ICICT)*, Aug. 2016,

[7] Sadiq, S., Khalid, M. U., Mui-Zzud-Din, Ullah, S., Aslam, W., Mehmood, A., Choi, G. S., & On, B. W. (2021). Classification of β-Thalassemia Carriers From Red Blood Cell Indices Using Ensemble Classifier. *IEEE Access, 9*, 45528–45538.

[8] Salman Khan, M., Ullah, A., Khan, K. N., Riaz, H., Yousafzai, Y. M., Rahman, T., Chowdhury, M. E. H., & Abul Kashem, S. B. (2022, October 3). Deep Learning Assisted Automated Assessment of Thalassaemia from Haemoglobin Electrophoresis Images. *Diagnostics*, *12*(10), 2405.

[9] Automated Thalassemia cell image segmentation using hybrid Fuzzy C-Means

and K-Means. (2023, August 29). *ZANCO JOURNAL OF PURE AND APPLIED SCIENCES*, *35*(4).

[10] Rodellar, J., Alférez, S., Acevedo, A., Molina, A., & Merino, A. (2018, May). Image processing and machine learning in the morphological analysis of blood cells. International Journal of Laboratory Hematology, 40(S1), 46–53. https://doi.org/10.1111/ijlh.12818

[11] Lee, S. Y., Chen, C. M., Lim, E. Y., Shen, L., Sathe, A., Singh, A., Sauer, J., Taghipour, K., & Yip, C. Y. (2021, January). Image Analysis Using Machine Learning for Automated Detection of Hemoglobin H Inclusions in Blood Smears - A Method for Morphologic Detection of Rare Cells. *Journal of Pathology Informatics*, *12*(1), 18

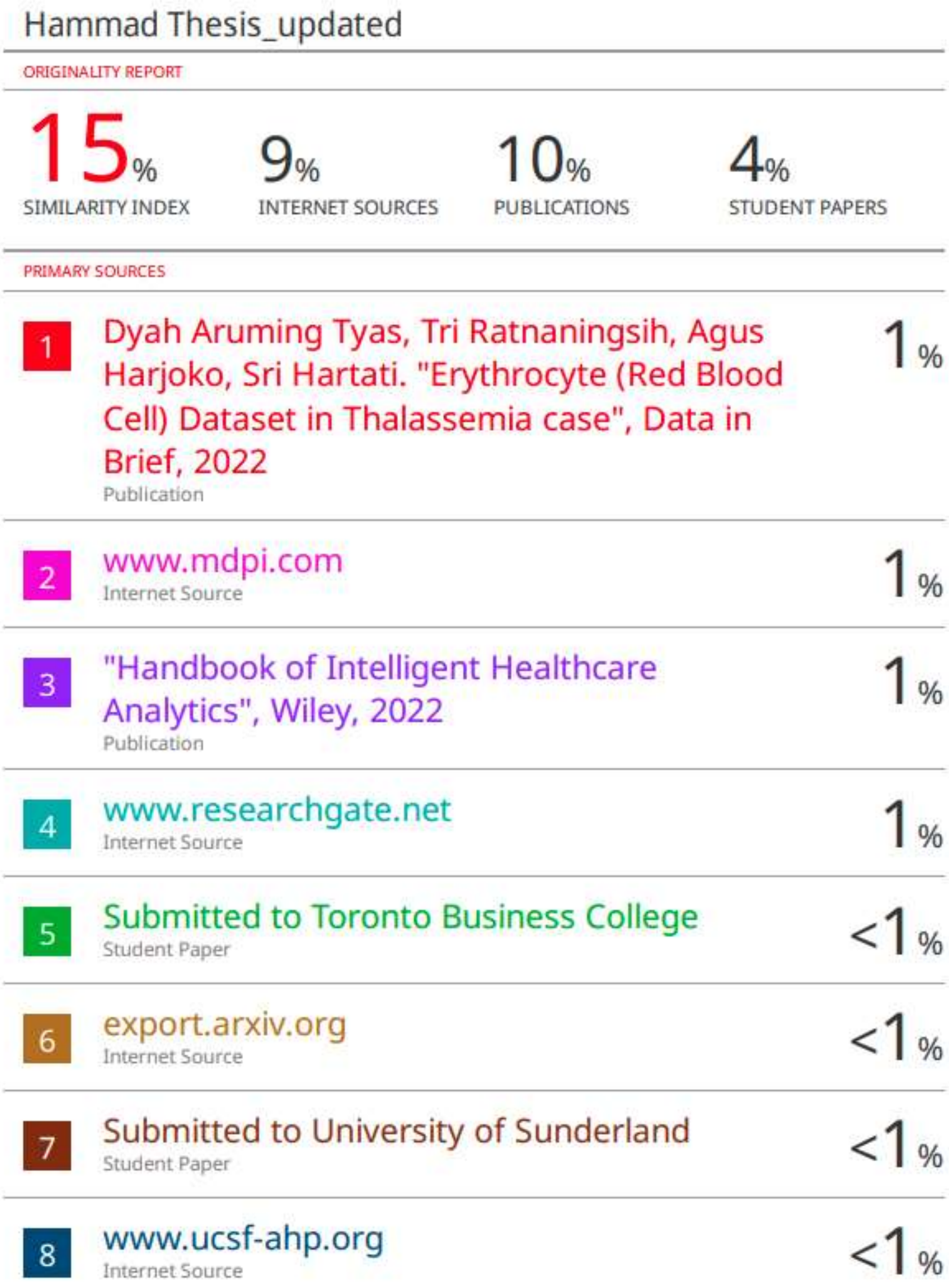
[12] Lin, Y. H., Liao, K. Y. K., & Sung, K. B. (2020, November 13). Automatic detection and characterization of quantitative phase images of thalassemic red blood cells using a mask region-based convolutional neural network. *Journal of Biomedical Optics*, *25*(11).

[13] Ferih, K., Elsayed, B., Elshoeibi, A. M., Elsabagh, A. A., Elhadary, M., Soliman, A., Abdalgayoom, M., & Yassin, M. (2023, April 26). Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review. *Diagnostics*, *13*(9), 1551.

[14] Gonzalez-Hidalgo, M., Guerrero-Pena, F. A., Herold-Garcia, S., Jaume-i-Capo, A., & Marrero-Fernandez, P. D. (2015, July). Red Blood Cell Cluster Separation From Digital Images for Use in Sickle Cell Disease. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1514–1525.

[15] Bhowmick, S., Das, D. K., Maiti, A. K., & Chakraborty, C. (2012, September 1). Computer-Aided Diagnosis of Thalassemia Using Scanning Electron Microscopic Images of Peripheral Blood: A Morphological Approach. *Journal of Medical Imaging and Health Informatics*, *2*(3), 215–221.

[16] Lachover Roth, I., Lachover, B., Koren, G., Levin, C., Zalman, L., & Koren, A. (2018, January 1). DETECTION OF Β THALASSEMIA CARRIERS BY RED CELL PARAMETERS OBTAINED FROM AUTOMATIC COUNTERS USING MATHEMATICAL FORMULAS. *Mediterranean Journal of Hematology and Infectious Diseases*, *10*(1), 2018008.



Hammad Thesis_updated

ORIGINALITY REPORT

15%	9%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dyah Aruming Tyas, Tri Ratnaningsih, Agus Harjoko, Sri Hartati. "Erythrocyte (Red Blood Cell) Dataset in Thalassemia case", Data in Brief, 2022 Publication	1%
2	www.mdpi.com Internet Source	1%
3	"Handbook of Intelligent Healthcare Analytics", Wiley, 2022 Publication	1%
4	www.researchgate.net Internet Source	1%
5	Submitted to Toronto Business College Student Paper	<1%
6	export.arxiv.org Internet Source	<1%
7	Submitted to University of Sunderland Student Paper	<1%
8	www.ucsf-ahp.org Internet Source	<1%