CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Acute Lymphoblastic Leukaemia Identification through Blood Smear Images

by

Muhammad Saad Hashmi

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2025

I want to dedicate this achievement to my parents,teachers and friends who always encourage and support me in every crucial time

CERTIFICATE OF APPROVAL

Acute Lymphoblastic Leukaemia Identification through Blood Smear Images

by

Muhammad Saad Hashmi

(MCS213022)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ayyaz Hussain	QAU, Islamabad
(b)	Internal Examiner	Dr. Abdul Basit Siddique	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Masroor Ahmed	CUST, Islamabad

Dr. Muhammad Masroor Ahmed
Thesis Supervisor
February, 2025

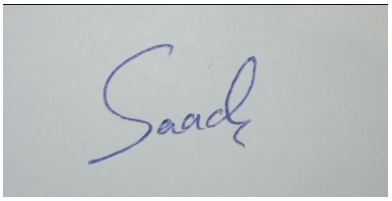
Dr. Abdul Basit Siddiqui
Head
Dept. of Computer Science
February , 2025

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
February, 2025

Author's Declaration

I, **Muhammad Saad Hashmi** hereby state that my MS thesis titled "**Acute Lymphoblastic Leukaemia Identification through Blood Smear Images**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

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



(Muhammad Saad Hashmi)

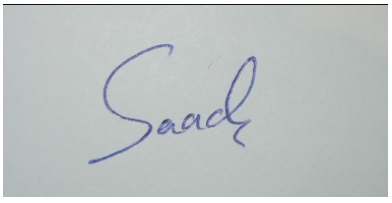
Registration No: MCS213022

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Acute Lymphoblastic Leukaemia Identification through Blood Smear Images** " is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology 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.

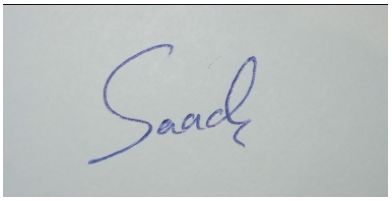
I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of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



(Muhammad Saad Hashmi)
Registration No: MCS213022

Acknowledgement

First and Foremost, I express my gratitude to the most merciful, powerful, and benevolent ALLAH Almighty for bestowing upon me the skills, knowledge, and unwavering determination to reach this point and achieve my research goal. I extend my sincere appreciation to my supervisor, Dr Masroor Ahmed, whose guidance and support were instrumental in completing my research thesis. I was through her inspiration that my interest in research work was ignited.



(Muhammad Saad Hashmi)

Abstract

Leukaemia, a formidable blood cancer affecting white blood cells in the bone marrow, demands prompt identification for effective intervention. Conventional methods for diagnosing acute Leukaemia rely on manual microscopic examinations, specifically assessing white blood cell percentages in peripheral blood. However, these methods encounter challenges associated with inaccuracies, error susceptibility, and time consumption due to human factors such as fatigue, stress, and limited experience.

In medical image analysis, providing a Sample number of labeled images for training deep learning models poses a significant challenge. This arises from restrictions on image availability and the associated costs of expert time required for meticulous ground truth annotations. Notably, Acute Lymphoblastic leukemia (ALL) emerges as the most prevalent childhood cancer, claiming the unfortunate distinction of being the leading cause of death among children.

This study critically examines the landscape of computer-aided systems for acute Leukaemia detection, spanning various stages such as image acquisition, preprocessing, segmentation, feature extraction, and classification. Furthermore, the paper addresses the contemporary hurdles in medical image analysis, emphasizing the challenges in sourcing labeled images for deep learning models. Given the critical significance of early detection, the study underscores the need for innovative approaches and explores the potential of deep learning in revolutionizing the landscape of Leukaemia diagnosis.

Contents

List of Figures

List of Tables

Abbreviations

ALL	Acute Lymphoblastic Leukaemia
CMYK	Cyan, Magenta, Yellow, and Key (Black)
DRLBP	Dynamic Recurrent Local Binary Patterns
GLCM	Gray Level Co-occurrence Matrix
HMRF	Hidden Markov Random Field
PBS	Peripheral blood smears
PPCA	Probabilistic Principal Component Analysis
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
WBC	White Blood Cell

Symbols

σ	Sigma
α	Alpha
δ	Delta
λ	lambda
∇	Nabla
I_{EPSF}	Image Encapsulated PostScript File
$\sum$	Sum
μ	mu

Chapter 1

Introduction

1.1 Background

White blood cell malignancy is known as acute lymphoblastic leukemia (ALL). It mostly affects youngsters, although in rare instances, adults may also be impacted. The incidence of ALL peaks between the ages of 2 and 5 years. The causes of ALL are multifactorial, involving a combination of exogenous (environmental) and endogenous (genetic) factors, as well as random chance. Significant improvements in the survival rates of pediatric ALL have been achieved, with current clinical trials reporting survival rates of approximately 90%. This success is attributed to risk stratification based on the biological characteristics of leukemic cells and patients' responses to treatment, treatment regimens based on pharmacodynamics and pharmacogenomics modifications, and supportive care advancements. [1]

Diagnosing leukemia early is crucial but challenging due to the nonspecific nature of its initial symptoms, such as enlarged lymph nodes, pallor, fever, and weight loss. These symptoms can easily be mistaken for other illnesses, complicating early detection efforts.[2] In many cases, these symptoms are subtle and not easily recognizable in the early stages of leukemia, which further hinders timely diagnosis.Despite advancements in treatment, timely diagnosis remains a challenge, emphasizing the need for improved diagnostic techniques.

1.2 Diagnostic Methods

The most common method for diagnosing leukemia involves the microscopic evaluation of peripheral blood smears (PBS). In this procedure, a small sample of blood is examined under a microscope to detect abnormal cells [3]. Although PBS is widely used, the definitive diagnosis of leukemia often requires the analysis of bone marrow samples. This more invasive method allows for a comprehensive assessment of the bone marrow, where leukemia typically originates [4].

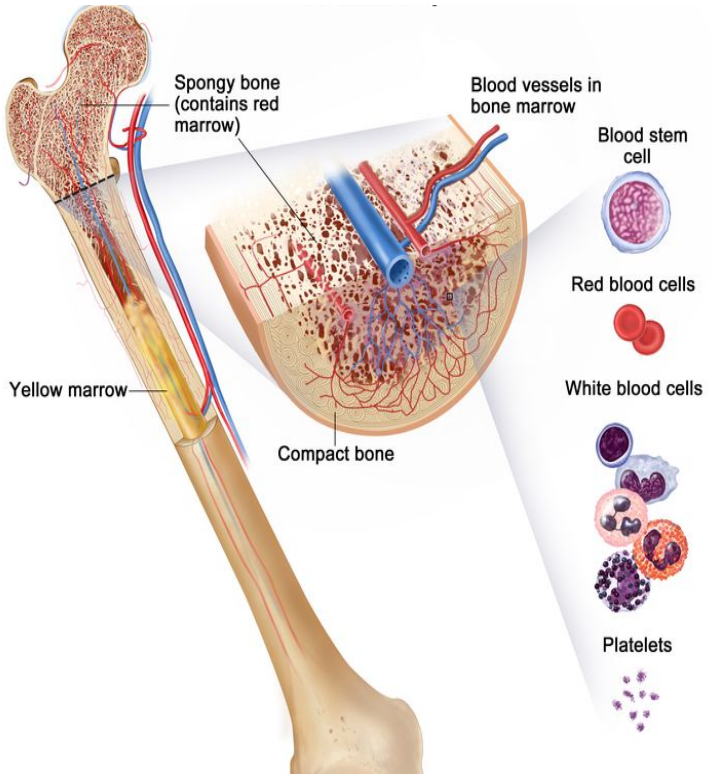


FIGURE 1.1: Structure of Bone Marrow and Blood Cells.

Medical imaging provides valuable insights into the characteristics of both normal and abnormal blood cells. Currently, the microscopic examination of blood smears is the most prevalent method for diagnosing acute leukemia. These images include both normal and abnormal blood cells, and despite extensive research, accurately detecting and classifying acute leukemia using image-processing techniques remains challenging. The primary difficulties stem from the variability in the size, shape, location, and intensity of images representing different types of acute leukemia.The process of detecting leukemia through image processing and

computer vision techniques comprises five stages: image acquisition, image preprocessing, white blood cell (WBC) segmentation, feature extraction, and the classification of leukemia cells. Fig. 1.2 illustrates these stages [5].

FIGURE 1.2: Stages

1.3 Advances in Automated Leukemia Detection

A new automated system for detecting leukemia from blood images employs machine learning and image processing approaches to achieve high precision in identifying cases of leukemia [6]. The effectiveness and dependability of ALL diagnosis might be greatly improved by this technique, thereby improving patient care.

Advances in automated ALL detection and WBC categorization demonstrates the vast potential of deep learning and machine learning in the medical domain. As research continues to address the remaining challenges, these techniques promise to revolutionize the diagnosis and treatment of ALL, ultimately leading to better patient outcomes.

1.4 Key Points

Incidence and Survival Rates: ALL is most prevalent in young children, with incidence peaking between 2 and 5 years of age. Significant improvements in pediatric ALL survival rates, reaching approximately 90%, have been achieved through advancements in risk stratification, treatment modification, and supportive care [1].

Challenges in Early Diagnosis: Early detection of leukemia remains a significant challenge due to the nonspecific nature of initial symptoms, which can easily be

confused with other illnesses. This makes it difficult for researchers, doctors, and hematologists to diagnose leukemia at an early stage [2].

Diagnostic Methods: While the microscopic evaluation of peripheral blood smears is commonly used for initial leukemia diagnosis, the definitive diagnosis often requires bone marrow analysis, which provides a more comprehensive assessment [4].

Automated Detection Systems: Automated techniques for identifying leukemia from blood smear pictures have been made possible by recent developments in machine learning and image processing. By increasing the precision and dependability of ALL diagnosis, these systems should enhance patient care and results [6].

1.5 Problem Statement

Acute Lymphoblastic Leukemia (ALL) diagnosis traditionally relies on manual microscopic examination of blood smears, a process that is time-consuming and prone to inter-observer variability. To address these limitations, automated machine learning-based diagnostic systems have gained significant attention. This study aims to develop a robust leukemia classification framework by integrating precise image segmentation and dimensionality reduction techniques.

The research focuses on achieving accurate segmentation of leukemia cells to closely match expert-labeled ground truth, ensuring reliable feature extraction for classification. Additionally, Principal Component Analysis (PCA) is employed for dimensionality reduction to optimize computational efficiency while maintaining classification accuracy. The impact of PCA on leukemia image classification is assessed in terms of processing time and predictive performance, providing insights into the trade-offs between reduced feature space and model accuracy.

By implementing Fuzzy C-Means (FCM) for segmentation and Support Vector Machines (SVM) for classification, this study explores an effective approach to

improve leukemia diagnosis. The findings aim to contribute to the development of efficient, scalable, and accurate computer-aided diagnostic systems for ALL detection.

1.6 Research Objectives

1. To develop an automated machine learning-based diagnostic system for Acute Lymphoblastic Leukemia (ALL) using advanced image processing techniques, including Fuzzy C-Means (FCM) for segmentation and Support Vector Machines (SVM) for classification.

2. To assess how precise image segmentation enhances the accuracy of leukemic cell detection and differentiation from normal cells.

3. To evaluate the effectiveness of Principal Component Analysis (PCA) for dimensionality reduction, focusing on its impact on computational efficiency and classification accuracy in leukemia image classification.

4. To compare the performance of the proposed machine learning-based diagnostic system with traditional manual methods, assessing its diagnostic accuracy, efficiency, and robustness.

5. To explore the trade-offs between reduced feature space via PCA and improved model performance in the context of leukemia detection, highlighting key insights for optimizing diagnostic systems.

1.7 Research Question

1. How can precise segmentation of leukemia images be achieved to closely match the gold standard?

2. How does dimensionality reduction impact the classification of leukemia images in terms of computational efficiency and accuracy?

1.8 Research Contribution

This research makes several key contributions to the field of automated leukemia diagnosis, particularly for Acute Lymphoblastic Leukemia (ALL) detection from blood smear images:

1. **Development of an Automated Diagnostic System:** The research introduces a machine learning-based system that combines image preprocessing, segmentation (Fuzzy C-Means clustering), and classification (Support Vector Machines) to automate ALL detection from blood smear images, reducing reliance on manual examination.

2. **Enhanced Image Preprocessing and Segmentation:** The study applies Fuzzy C-Means (FCM) segmentation for precise detection of leukemic cells, enhancing their separation from other blood cell types and improving the diagnostic quality of processed images.

3. **Feature Extraction for Improved Classification:** The study extracts key features (color, shape, texture) from blood smear images to improve classification accuracy, enabling the SVM classifier to better distinguish between normal and leukemic cells.

4. **Validation and Performance Evaluation:** The system's performance is evaluated through accuracy, precision, and recall comparisons, demonstrating its effectiveness in matching or surpassing traditional manual diagnostic methods.

5. **Automation and Efficiency in Leukemia Detection:** The research supports a faster and more consistent ALL diagnosis by automating detection, contributing to improved efficiency and reduced diagnostic errors in clinical settings.

Chapter 2

Literature Review

2.1 Introduction

Acute Lymphoblastic Leukemia (ALL) stands as a formidable adversary in the realm of hematologic malignancies, characterized by the uncontrolled proliferation of immature lymphoid cells. The urgency associated with ALL diagnosis and intervention underscores the need for advanced and efficient diagnostic methodologies. Traditional manual examination of blood smears, while a longstanding practice, faces challenges related to subjectivity and time constraints. The field of leukemia diagnosis has seen a radical change mainly due to the convergence of advanced technologies, including data science, supervised and unsupervised learning, machine learning, and artificial intelligence (AI) in recent years. In the field of blood smear image analysis for ALL detection, this development shows potential. Leukemia diagnosis and categorization are being revolutionized by machine learning and deep learning, which provide more accuracy and efficiency than conventional techniques[1].

Among the various ML algorithms, Support Vector Machines (SVMs) have consistently shown superior performance in leukemia classification [7].The standard procedure for using image processing techniques to diagnose Acute Lymphoblastic Leukemia (ALL) is pre-processing blood smear pictures, which are then segmented,

feature extracted, and classified. Commonly employed methods for segmentation include Otsu's threshold approach, Watershed transform, and K-means clustering, all of which have produced positive outcomes. Other techniques like the Distance transform, Zack's algorithm, and fuzzy-C-means clustering have also shown good results.

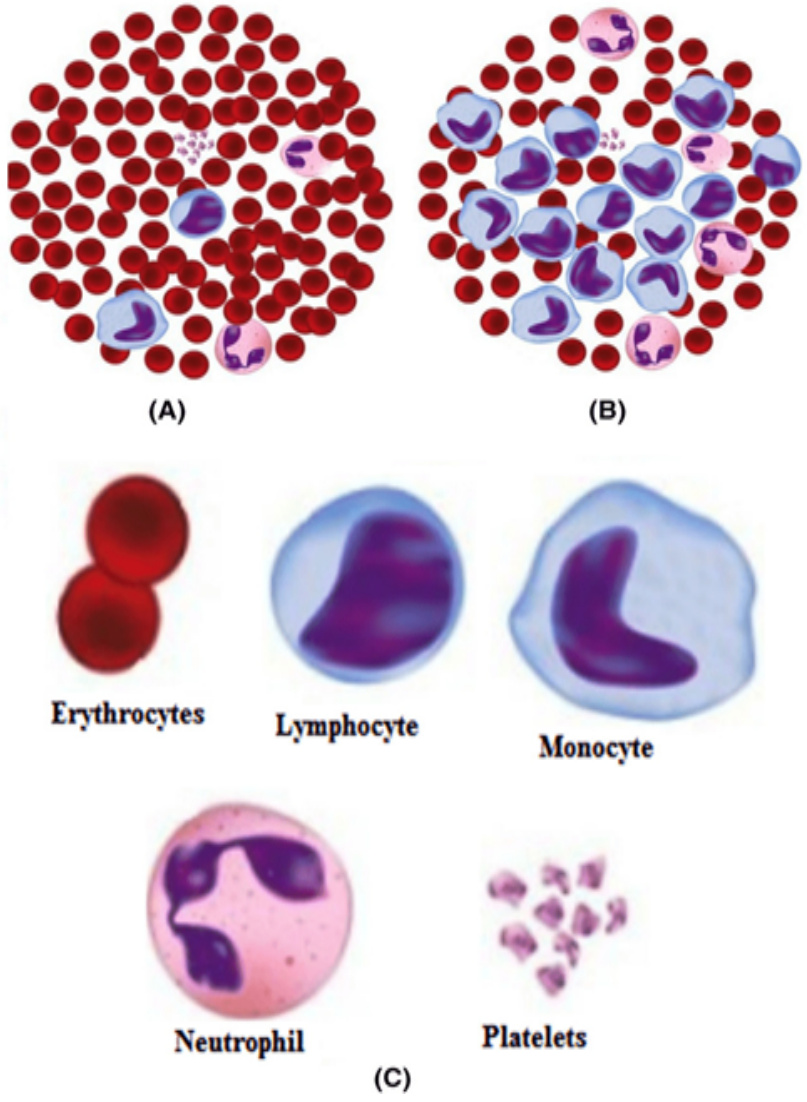


FIGURE 2.1: Structure of Bone Marrow and Blood Cells.

Analyzing geometrical, statistical, color, and textural properties is part of feature extraction.Granulometry and Gray Level Co-occurrence Matrix are commonly employed to capture texture details, while Fisher's Discrimination Ratio aids in feature selection.When it comes to classification, the Support Vector Machine (SVM)

classifier has demonstrated superior accuracy compared to other approaches [8].

Beyond supervised learning, the integration of unsupervised learning techniques becomes imperative for uncovering hidden patterns within unlabelled datasets. Clustering algorithms, like hierarchical clustering or K-means, contribute to a more comprehensive understanding of the heterogeneity inherent in ALL [8].

Once cells are segmented, the classification process comes into play, distinguishing normal from leukemic cells. Supervised learning techniques, incorporating Support Vector Machines (SVMs) and neural networks,leverage labeled datasets to recognize distinctive features associated with ALL [9].Cell size, shape irregularities, and nuclear morphology are among the critical parameters facilitating accurate classification.

2.2 Literature Review

The Machine learning (ML) application in the diagnosis of leukemia has gained significant traction in recent years. ML algorithms, particularly Support Vector Machines (SVMs), have shown promising results in the classification of leukemic cells. SVMs are powerful classifiers that can handle nonlinear classification problems by creating hyperplanes that separate different classes with maximum margin and minimal error.

Various studies have demonstrated the efficacy of SVMs in leukemia classification. For instance, a study by Kurniawan et al. [10] reported an accuracy of 95% using an SVM classifier for ALL detection. Combining color and texture data taken from blood smear pictures, the study demonstrated how crucial feature selection is to attaining high classification accuracy.

A deep learning approach were used in another study by Akram et al.[11] to categorize leukemic cells. With a 97% accuracy rate, the CNN model showed that deep learning could be used to automate the diagnosing process. The study

emphasized the advantage of deep learning in learning complex patterns from large datasets, which is critical for accurate leukemia detection.

2.2.1 Image Processing Techniques for Leukemia Detection

Automated leukemia detection relies heavily on image processing techniques. To improve the quality of blood smear photographs, pre-processing techniques including noise reduction, contrast enhancement, and normalization are crucial. Leukemic cells are frequently separated from the background using segmentation techniques as Otsu's thresholding, Watershed transform, and K-means clustering. Each segmentation method has its strengths and weaknesses. For instance, K-means clustering is efficient and simple to implement, but it may struggle with overlapping cells. Watershed transform provides good separation of touching cells but can be sensitive to noise. Otsu's thresholding is effective for bimodal distributions but may not perform well with images having complex backgrounds [12].

Feature extraction is another critical step in leukemia detection. Techniques such as Granulometry and Gray Level Co-occurrence Matrix (GLCM) are used to capture texture details, while geometric features like cell size and shape are also important. Fisher's Discrimination Ratio is often employed to select the most relevant features for classification. These features are then fed into classifiers such as SVMs or neural networks to distinguish between normal and leukemic cells. A study by Rehman et al. [13] demonstrated the effectiveness of combining texture and geometric features, achieving an accuracy of 96% with an SVM classifier.

2.2.2 Supervised vs. Unsupervised Learning

Supervised learning techniques, including SVMs and neural networks, rely on labeled datasets to train models. These techniques have shown high accuracy in

leukemia classification due to their ability to learn distinctive features from labeled data. However, obtaining labeled datasets can be challenging and time-consuming. Unsupervised learning techniques, such as clustering algorithms, offer an alternative by uncovering hidden patterns within unlabeled data. K-means and hierarchical clustering are commonly used unsupervised techniques that help in understanding the heterogeneity of leukemic cells. A study by Albarqouni et al. [14] applied hierarchical clustering to categorize leukemic cells based on their morphological characteristics, providing insights into the diversity of cell types in ALL.

2.2.3 Advances in Automated Detection Systems

Automated leukemia detection methods have been developed as a result of recent developments in image processing and machine learning. These systems combine a number of methods to attain great efficiency and precision. A work by Abbas et al.[15], for instance, suggested an automated system that used deep learning to integrate pre-processing, segmentation, feature extraction, and classification phases. The system demonstrated the promise of automated ways to improve leukemia diagnosis with a 98% accuracy rate[16, 17].

Some of the Image preprocessing technique used by researchers are:

2.3 Image Preprocessing Technique

2.3.1 Image Normalization

Normalization refers to the process of scaling the pixel values of an image to a specified range, typically $[0, 1]$ or $[-1, 1]$. Let $I(x, y)$ represent the pixel intensity in the original image, and let $I_{\text{norm}}(x, y)$ denote the intensity of the pixel after normalization[18].

For min-max normalization to the range $[0, 1]$:

$$I_{\text{normalized}}(x, y) = \frac{I(x, y) - \min(I)}{\max(I) - \min(I)} \tag{2.1}$$

For z-score normalization:

$$I_{\text{normalized}}(x, y) = \frac{I(x, y) - \mu}{\sigma} \tag{2.2}$$

where μ denotes the average pixel intensity in the image, and σ represents the standard deviation of the pixel intensities within the image [19].

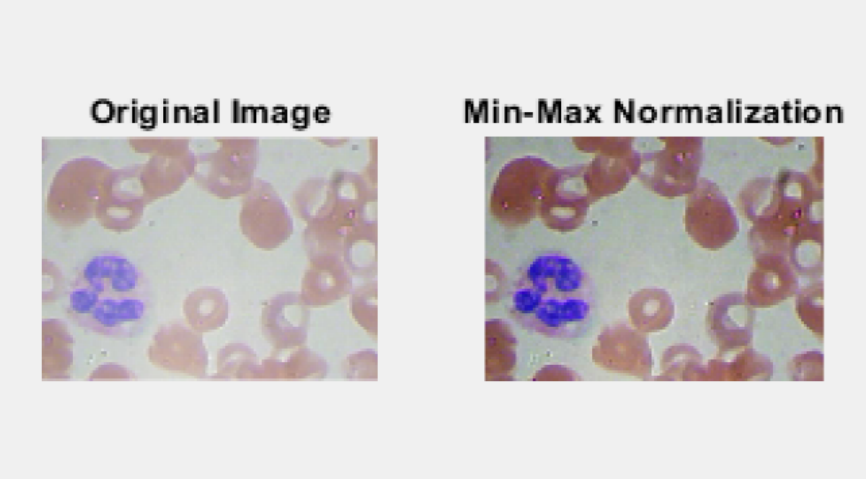


FIGURE 2.2: PreProcessing Of Blood Seamer Images.

2.3.2 Threshold-based Segmentation

Threshold-based segmentation is a technique for separating areas of interest from the background in image processing by comparing pixel intensities to a threshold value [20]. The process converts a grayscale image into a binary one, with pixels either part of the foreground or background, depending on their intensity compared to the threshold.

$$I_{bw}(x,y) = \begin{cases} 1 & \text{if } I_{gray}(x,y) \geq T \\ 0 & \text{if } I_{gray}(x,y) < T \end{cases} \tag{2.3}$$

Here, the grayscale image $I_{gray}(x,y)$ is represented, the resulting binary image is given by $I_{bw}(x,y)$, the pixel coordinates are denoted by (x,y), and the threshold value is represented by T. [21, 22].

Automated methods for selecting the threshold value include:

i. **Histogram shape-based methods**: Analyzing the shape of the histogram to identify peaks and valleys. Methods like the double-peak threshold and Otsu's method maximize variance between objects and background.

ii. **Clustering method**: Treating the image as clusters and iteratively reducing them to determine background and foreground.

iii. **Maximum-entropy-based method**: Maximizing the histogram's foreground and background areas' entropy.

iv. **Local thresholding method**: Choosing thresholds based on local image characteristics.

v. **Watershed method**: Using gradient magnitude to simulate flooding and segmenting objects.

These methods can be extended to color images by determining thresholds for each color channel and combining them. Segmentation using color information may offer improved accuracy over grayscale images[23, 24].

2.3.3 Edge-based Segmentation

Edges in images mark the transition between different areas, indicating changes in properties like intensity or color. Detecting edges helps define the boundaries

of objects within images, as these transitions often occur where one object or surface meets another. This process is crucial for understanding image features and segmenting images into discrete regions that represent various objects[25, 26].

The majority of edge-detection techniques depend on computing the image's derivatives, especially the first- and second-order derivatives. The gradient, which gauges how rapidly pixel values vary throughout the image, is the main first-order derivative employed in image processing. By detecting these changes, gradients can identify the presence of edges [27]. Second-order derivatives, such as the Laplacian operator, delve deeper into the rate of change of gradients. They help determine the exact location of edges and whether a pixel belongs to the darker or lighter side of an edge boundary [28, 29].

The vector is expressed as the gradient of an image $f(x, y)$ at a point (x, y):

$$\nabla f(x, y) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{2.4}$$

where the gradients in the x and y directions are represented by g_x and g_y, respectively.

The gradient vector's magnitude is established by:

$$|\nabla f(x, y)| = \text{mag}(\nabla f(x, y)) = \left(g_y^2 + g_x^2\right)^{1/2} = \left(\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right)^{1/2} \tag{2.5}$$

In order to make the calculation easier, this size is frequently estimated as:

$$\nabla f \approx |g_x| + |g_y| \tag{2.6}$$

At the point (x, y), the direction of the gradient vector, $\theta(x, y)$, indicates the direction in which the image $f(x, y)$ experiences the highest rate of change.

$$\theta(x, y) = \arctan\left(\frac{g_y}{g_x}\right) \tag{2.7}$$

The following equation can be used to estimate the gradient of a discrete image:

$$|\nabla f(x, y)| = \left[(f(x, y) - f(x + 1, y))^2 + (f(x, y) - f(x, y + 1))^2\right]^{1/2} \tag{2.8}$$

This can be further simplified as:

$$\nabla f \approx |f(x, y) - f(x + 1, y)| + |f(x, y) - f(x, y + 1)|$$

Edges in images can be identified using various first-order derivative operators such as:

- **Sobel Edge Detector**: Involves applying convolutional masks to compute gradients along the horizontal and vertical axes, making it effective for edge detection in these directions[30].
- **Roberts Cross Operator**: Determines edges by calculating the magnitude of the gradient using convolution, with a focus on diagonal edge detection at 45° and 135° angles[31].
- **Prewitt Operator**: Employs gradient-based filters in multiple directions, identifying edges as regions with the highest gradient response[32].
- **Kirsch Edge Detector**: Utilizes eight distinct filters to evaluate gradients in various compass directions, detecting edges with significant intensity changes[33].
- **Frei-Chen Operator**: Focuses on edge detection using simplified row and column gradient filters for computational efficiency[34].

These methods are designed to locate edges by identifying extremes in the first derivative of an image's intensity. However, their sensitivity to noise can sometimes result in inaccurate edge detection [35, 36].

The Laplacian of an image $f(x, y)$ is defined as:

$$\nabla^2 f(x, y) = \frac{\partial^2 f(x, y)}{\partial x^2} + \frac{\partial^2 f(x, y)}{\partial y^2} \tag{2.9}$$

where $\frac{\partial^2 f(x,y)}{\partial x^2}$ is the second derivative of the image with respect to x, and $\frac{\partial^2 f(x,y)}{\partial y^2}$ is the second derivative with respect to y.

The second derivative with respect to x can be approximated as:

$$\frac{\partial^2 f(x, y)}{\partial x^2} = f(x + 1, y) + f(x - 1, y) - 2f(x, y) \tag{2.10}$$

Similarly, the second derivative with respect to y is:

$$\frac{\partial^2 f(x, y)}{\partial y^2} = f(x, y + 1) + f(x, y - 1) - 2f(x, y) \tag{2.11}$$

Thus, the Laplacian operator is:

$$\nabla^2 f(x, y) = f(x + 1, y) + f(x - 1, y) + f(x, y + 1) + f(x, y - 1) + 4f(x, y) \tag{2.12}$$

The Laplacian operator, being a second-order derivative, is often avoided for direct edge detection due to its high sensitivity to noise, tendency to produce double edges, and inability to determine edge direction [37, 38]. To address these limitations and improve edge detection accuracy, two main algorithms are commonly used:

i. **Laplacian of Gaussian (LoG)**: This technique uses the Laplacian operator in conjunction with Gaussian filtering to efficiently detect edges by smoothing the image and then identifying regions where intensity changes occur sharply. However, it doesn't provide information about edge orientation [29, 39].

ii. **Canny edge detector**: This algorithm employs a multistage approach to detect edges in images. Initially applying Gaussian convolution for smoothing the image and minimize noise[40]. Then, the gradient magnitude is then calculated using a first-derivative operator (such as the Robert or Sobel operators). Non-maximum suppression is then applied to eliminate false replies and narrow down the edges. After identifying possible edges using a double thresholding approach,through the suppression of unwanted edges, edge tracking via hysteresis aids in improving detection accuracy. Although the performance of the Canny edge detector is good, it is computationally intensive and more complex compared to other methods [41, 42].

2.3.4 Image Segmentation using Morphology Operations

Detection of abnormalities in blood smear images, particularly for Leukaemia diagnosis, is crucial in medical analysis. Just as signs and notices provide important information in everyday scenarios, text detection in natural scene images is analogous to identifying relevant features in medical images. Algorithms developed for this purpose vital role is played in various real-world applications, including assisting medical professionals in navigating through images, aiding visually impaired individuals, and enhancing diagnostic accuracy [43, 44]. Although several techniques have been put forth to detect text in scene images, many rely on analyzing the entire image. However, for Leukaemia detection in blood smear images, where abnormalities are often localized,segmenting the picture into homogenous sections first is a more efficient method[45]. By focusing on these regions, which may correspond to abnormal cell clusters, the detection process becomes more efficient. Thus, an algorithm is provided by this study as its purpose for dividing

blood smear images into uniform areas, followed by the identification of abnormal cell clusters, akin to detecting text on signboards in scene images. The objective of this technique is to increase the precision and effectiveness of leukemia identification in blood smear images by targeting specific regions of interest rather than analyzing the entire image indiscriminately. Further details of the proposed method and its implications for Leukaemia diagnosis will be discussed in subsequent sections[46, 47].

2.3.5 Extracting the Features of Images

Extracting relevant features from pre-processed images involves calculating specific metrics that capture various aspects such as color distribution and morphology.

The distribution of pixel intensities across several color channels is shown by a color histogram (e.g., Red, Green, and Blue)[48, 49]. For a given color channel c, the histogram H_c is computed as: The color histogram for a given color channel c is computed as:

$$H_c(k) = \sum_{i=1}^{N} \sum_{j=1}^{M} \delta(I(i,j,c) = k) \tag{2.13}$$

Where: the pixel intensity at position (i, j) in the color channel c is denoted by $I(i, j, c)$, and k ranges from 0 to 255 (for an 8-bit image). $\delta(\cdot)$ is the Kronecker delta function[19, 50]

The spatial relationships between pixel intensities are captured using the Gray-Level Co-occurrence Matrix (GLCM). Given a grayscale image, the GLCM P is defined as:

$$P(d,\theta)(i,j) = \frac{\text{Number of occurrences of pixel pairs at distance } d \text{ and angle } \theta}{\text{Total pixel pairs at } d \text{ and angle } \theta} \tag{2.14}$$

Where: d refers to the distance separating the pixels, and θ represents the angle formed between the pairs of pixels. Common texture features derived from GLCM include contrast, energy, and homogeneity [50, 51]

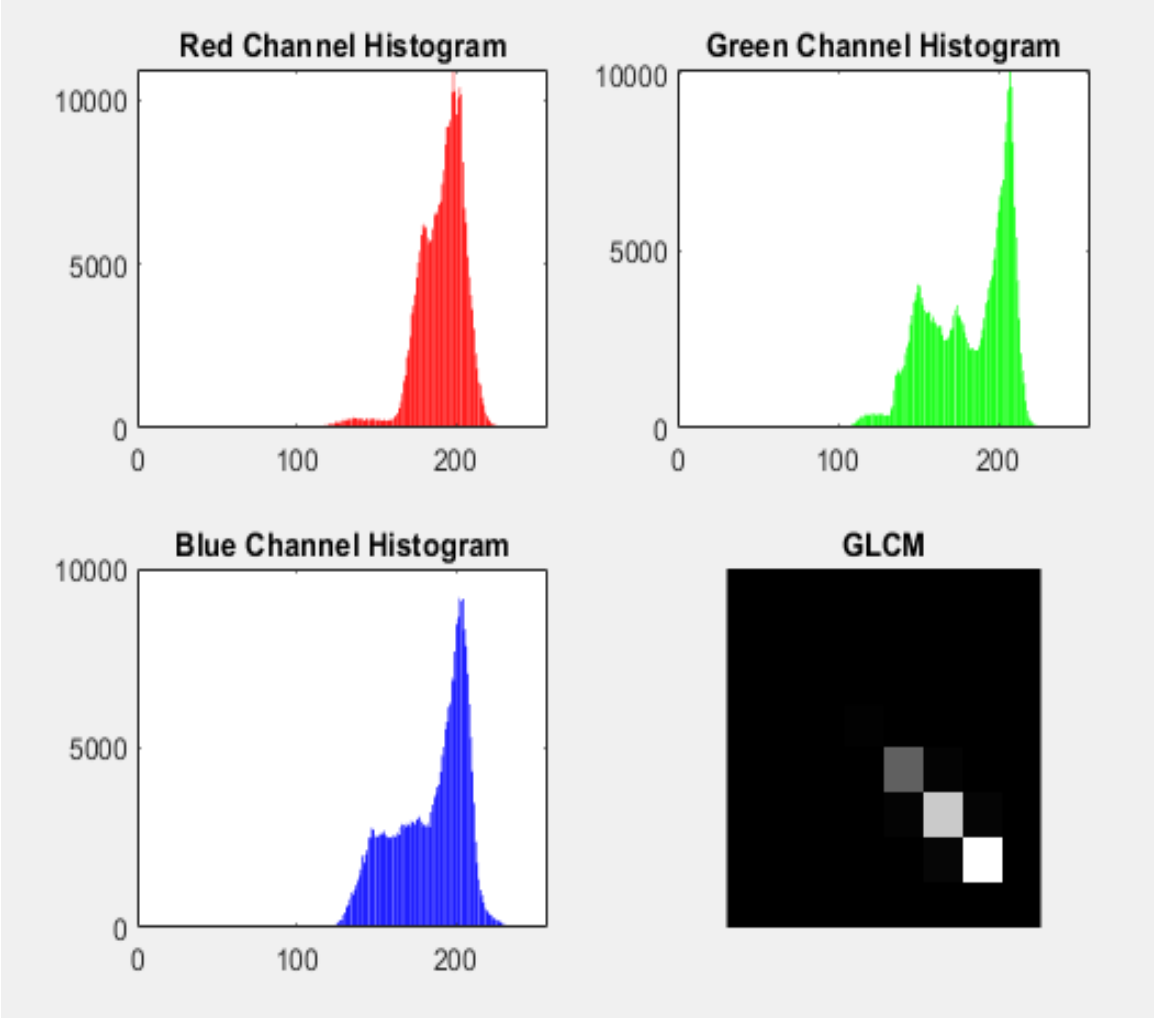


FIGURE 2.3: Image PreProcessing

Morphological features involve analysing the shape and structure of objects in the image. Common morphological features include area, perimeter, compactness, etc. [50] the area of an object in a binary image is given by:

$$\text{Area} = \sum_{i=1}^{N} \sum_{j=1}^{M} \delta(I(i,j) = \text{Object}) \tag{2.15}$$

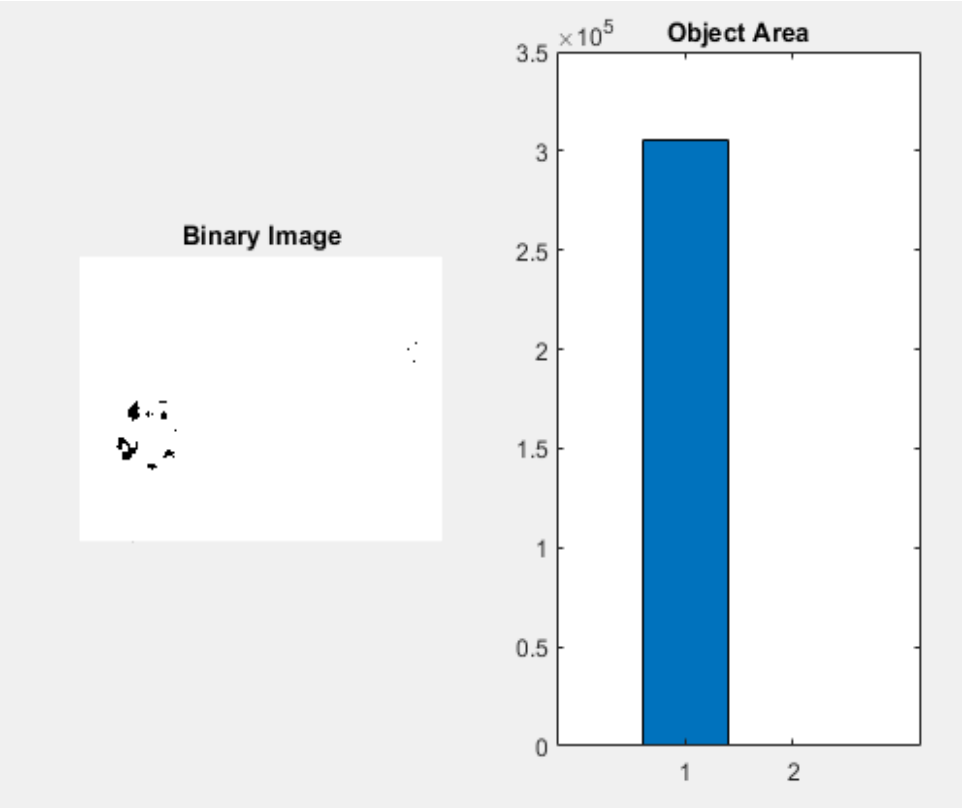


FIGURE 2.4: Image PreProcessing

A binary image is defined as one in which each pixel is assigned a value of either 0 (background) or 1 (object). In the context of this equation, the object is represented by the value 1. The dimensions of the image are defined as $N \times M$, where N represents the number of rows (height) and M represents the number of columns (width).

The pixel intensity at position (i, j) within the image is given by $I(i, j)$.

The Delta Function $\delta(I(i, j) = \text{Object})$ checks whether the pixel at position (i, j) is part of the object (i.e., whether $I(i, j) = 1$). If the pixel is part of the object $I(i, j) = 1$, the delta function returns 1.

If the pixel is not part of the object $I(i, j) = 0$, the delta function returns 0.

The double summation $\sum_{i=1}^{N} \sum_{j=1}^{M}$ indicates that we sum the values of the delta function over all pixels in the image [52, 53]

2.3.6 Leukocyte Segmentation

The segmentation phase was carried out following the application of the pre-processing step. The most important and difficult process is thought to be the segmentation step, which deals with separating discrete object components that contain crucial information. The goal of this work is to decrease the size of the high-resolution photographs, which significantly strain the hospital server's storage capacity, as well as the computing complexity of the following processes (Chen et al., 2020).

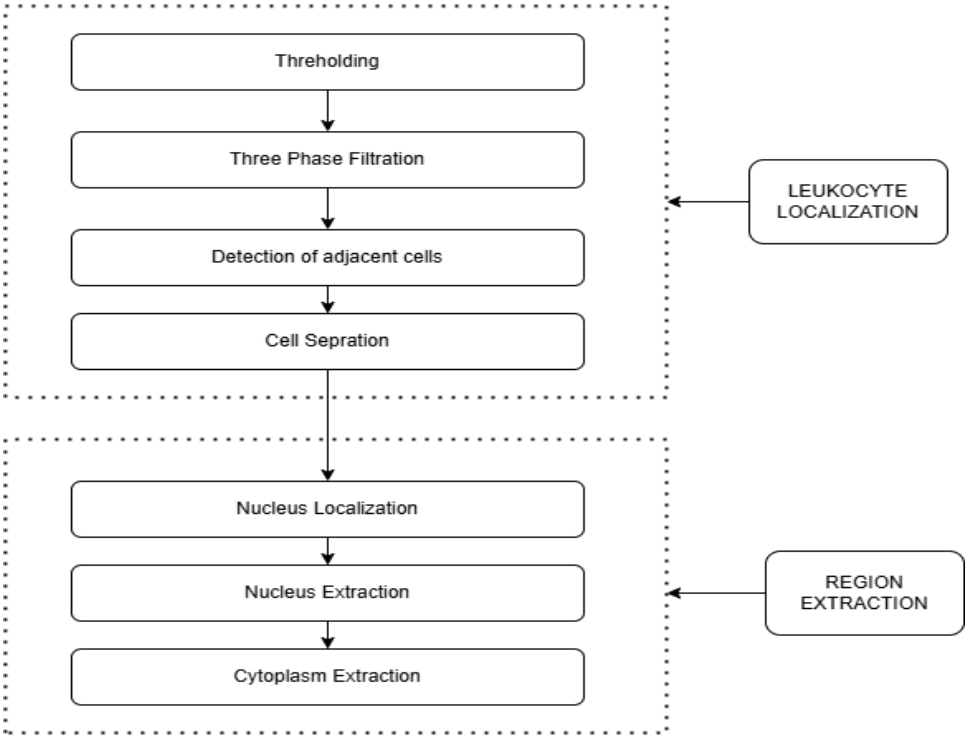


FIGURE 2.5: Image Segmentation

Leukocyte localization and region extraction, which concentrate on separating cell components like the nucleus and cytoplasm, are crucial steps in the process of identifying leukocytes in blood samples. From a morphological perspective, leukemic cells can be differentiated from mature leukocytes by having a large nucleus with fine lines. This process typically involves converting the image to a binary format, a crucial step to isolate the white blood cells (WBCs) from the

background and other blood components, including erythrocytes and platelets [54, 55]. As shown in Fig. 2.5.An effective algorithm used for leukocyte localization and region extraction includes four stages. Initially, a histogram-based thresholding method is applied to transform the image into a binary format, which helps to reduce the background and retain the full size of the WBCs. However, this method can introduce noise due to the overlapping nature of erythrocytes and the varying color spectrum of platelets [56]. To address these issues, additional noise-reduction techniques are required. A three-phase filtration process has been suggested to clean up the image by removing unwanted residual parts of cell components and any blood film defects (Bodzas, 2019)., [51, 57]).

Some of the model that are used by researchers are:

2.4 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) for detecting ALL from blood smear images, we need to define the key components, and the operations performed in the network [58, 59]. Here is a detailed breakdown of the mathematical formulation. The input layer represents the blood smear images as a multidimensional array. For an image of size 128128 with 3 color channels (RGB), the input X [60, 61]. An activation function σ (usually $ReLU$) is applied following the convolution operation:

$$A_{i,j,k} = \sigma\left(Z_{i,j,k}\right) = \max\left(0, Z_{i,j,k}\right) \tag{2.16}$$

The pooling layer reduces the spatial dimensions. For max pooling with pool size:

$$P_{i,j,k} = \max_{m=1}^{p} \max_{n=1}^{p} A_{i\cdot p+m-1, j\cdot p+n-1, k} \tag{2.17}$$

The flattening layer converts the 3D matrix into a 1D vector:

$$F \in \mathbb{R}^N \tag{2.18}$$

where N is the result of multiplying the pooled feature maps' dimensions.[62, 63] An activation function σ is computed by each neuron in a fully connected layer after a weighted sum of inputs and a bias is added:

$$O_i = \sigma\left(\sum_{j=1}^{N} W_{ij} \cdot F_j + b_i\right) \tag{2.19}$$

Where O_i is the output of the i-th neuron, W_{ij} are the weights, and b_i is the bias term.

For binary classification (detecting ALL or not), the output layer has a single neuron that generates a probability score using a sigmoid activation function p:

$$P = \sigma\left(\sum_{i=1}^{N} W_i \cdot O_{i+b}\right) \tag{2.20}$$

$$\sigma\left(x\right) = \frac{1}{1+e^x} \tag{2.21}$$

2.4.1 Cost Function

In machine learning, a mathematical function called a cost function measures the error or disparity that exists between expected values (y_{pred}) and actual values (y_{true}). It helps the model adjust its parameters to minimize this error.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \tag{2.22}$$

- $J(\theta)$: Represents the cost function, which needs to be minimized.
- m: Denotes the total number of training samples.
- $x^{(i)}$: The input features associated with the i-th training example.
- $y^{(i)}$: Represents the true output associated with the i-th training sample.

- $h_\theta(x^{(i)})$: The hypothesis function, representing the predicted value for the i-th example.
- θ: Model parameters (weights and bias).

2.4.2 Optimization of Cost Function

The process of determining the values of θ that minimize the cost function $J(\theta)$ is referred to as **optimization**. A widely used technique for this purpose is the *Gradient Descent* algorithm.

2.4.3 Gradient Descent

Gradient Descent is a technique used for optimization, where the parameters θ are progressively adjusted across the gradient's opposite direction to lower the value of the cost function $J(\theta)$[64, 65]. The gradient shows which way the greatest growth in $J(\theta)$, so moving in the reverse direction reduces $J(\theta)$[33].

The partial derivative of the cost function $J(\theta)$ with regard to each parameter θ_j tells us how much $J(\theta)$ will change if we slightly adjust θ_j.

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} \tag{2.23}$$

- $J(\theta)$: The cost function to minimize.
- θ_j: The parameter (weight or bias) being updated.
- $\frac{\partial J(\theta)}{\partial \theta_j}$:The gradient of $J(\theta)$ with respect to θ_j represents the rate at which $J(\theta)$ changes as θ_j varies.
- m: Refers to the number of training samples in the dataset.
- $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$: The hypothesis function, which gives the predicted output for the i-th sample.

- $y^{(i)}$: The true output value associated with the i-th training sample.
- $x_j^{(i)}$: Represents the j-th feature of the i-th training sample.

The gradient is averaged over all m training examples to ensure that the updates are robust and not influenced by a single training example [66, 67].

Once the gradient is calculated, the parameters θ_j are updated using the following rule:

$$\theta_j := \theta_j - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta_j} \tag{2.24}$$

- α: The learning rate, which regulates the update's step size. The procedure may overshoot the minimum if α is very big. Convergence will be sluggish if α is too tiny.
- $\frac{\partial J(\theta)}{\partial \theta_j}$: The gradient calculated in step 1.

This update ensures that θ_j moves in the direction that reduces the cost $J(\theta)$[68, 69].

For linear regression, substituting the gradient:

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} \tag{2.25}$$

The process is repeated for all parameters θ_j until the cost function $J(\theta)$ no longer decreases significantly. This indicates that the algorithm has likely found the minimum[70, 71].

The difference between the actual label y and the projected probability p is measured by the binary cross-entropy loss L.:

$$L = -\left(y \cdot \log(p) + (1 - y) \cdot \log(1 - p) \right) \tag{2.26}$$

This model maps the input image to a probability score indicating the presence of ALL [72, 73].

2.5 K-Nearest Neighbors (KNN)

A popular instance-based, non-parametric learning technique for classification and regression problems is K-Nearest Neighbors (KNN). In essence, KNN uses the majority class of a data point's 'k' nearest neighbors in a feature space to classify it. This approach is simple and assumes that comparable instances are located close to one another [74, 75].

To categorize a point of data using kNN, the following steps are typically taken:

Select 'k': Determine the number of closest neighbours to consider for classification. The optimal value of 'k' is usually determined through experimentation or cross-validation[76].

Calculate Distances: Determine how far the target data point is from every other point in the dataset. One popular metric is the Euclidean distance, which has the following mathematical definition:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.27}$$

where n is the number of characteristics and x and y are the feature vectors of two points.Determine Nearest Neighbors: Using the calculated distances, determine the 'k' data points that are nearest to the target location.Voting/Prediction: For classification, determine the class by a majority vote among the 'k' nearest neighbours. The target data point is allocated to the class with the greatest count [77, 78]. KNN has several advantages, including simplicity, flexibility with different distance metrics, and no need for explicit model training. However, it has limitations, such as sensitivity to noise, computational cost when the dataset is large, and the need for careful selection of 'K' and distance metric. Its performance can

also be affected by imbalanced data or high-dimensional spaces due to the "curse of dimensionality"[79, 80].

2.6 Linear Regression

The connection between a dependent variable (y) and one or more independent variables (x) can be represented statistically using linear regression. It presumes a linear association, indicating that variations in the independent variables correspond to proportional changes in the dependent variable, forming a straight-line relationship is formed [81, 82]

$$y = a + bx \tag{2.28}$$

- y is the dependent variable,
- a is the y-intercept of the regression line,
- b is the slope of the regression line,
- x is the independent variable.

The slope of the regression line, denoted by b, is calculated as the ratio of the covariance between x and y to the variance of x [83]. The formula is:

$$b = \frac{\gamma_{xy}}{s_x} \tag{2.29}$$

Where:

- γ_{xy} is the covariance between x and y,
- s_x is the standard deviation of x.

The y-intercept a of the regression line can be calculated using the formula:

$$a = \bar{y} - b\bar{x} \tag{2.30}$$

- The average value of the dependent variable y is represented by $\bar{y}$,
- The average value of the independent variable x is represented by $\bar{x}$,
- The slope of the regression line is denoted by b.

The covariance γ_{xy} between x and y is computed using the formula:

$$\gamma = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{2.31}$$

Where:

- γ represents the covariance,
- $\bar{x}$ and $\bar{y}$ denote the means of x and y, respectively,
- We take the sums over all data points.

Standard Deviation of y

We refer to the standard deviation of the dependent variable y as S_y, and it is computed using the following formula[84]:

$$S_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}} \tag{2.32}$$

Where:

- S_y is the standard deviation of y,
- $\bar{y}$ is the mean of y,
- n is the number of data points.

Standard Deviation of x

The standard deviation of the independent variable x, denoted as S_x, is calculated using the following formula[85]:

$$S_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \tag{2.33}$$

Where:

- S_x is the standard deviation of x,
- $\bar{x}$ is the mean of x,
- n is the number of data points.

2.7 Logistic Regression

Logistic regression extends linear regression to classification tasks, particularly binary classification, where the output is a probability that a given input belongs to a particular class (e.g., leukemic or healthy)[86, 87].The logistic regression model for binary classification is given by:

$$P(Y' = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{2.34}$$

Where:

- $P(Y' = 1 \mid X)$ denotes the likelihood that the output $Y' = 1$, given the input X,
- β_0 is the intercept of the model (the bias term),
- β_1 represents the coefficient associated with the independent variable X,
- X denotes the independent variable (input feature),
- e is the base of the natural logarithm (Euler's number).

TABLE 2.1: Summary of Current Methods in Leukemia Detection

Author, Year	**Methods**	**Performance**	**Strengths and Weaknesses**
Sajida & Alourani, 2024	Preprocessing: K-means clustering with HMRF	Segmentation accuracy: 96%	Strengths: Improved segmentation in whole and cropped images.
	Feature extraction: Texture, color features, and SVM		Weaknesses: K-values difficult to estimate; not suitable for non-convex clusters.
Akram et al., 2022	Preprocessing: Histogram green component from RGB	Accuracy: 99.517%	Strengths: High precision in identifying ALL cells and distinguishing cells that overlap.
	Segmentation: K-means clustering	Sensitivity: 99.348%	Weaknesses: K-values are not appropriate for non-convex clusters and are challenging to determine.
	Feature extraction: Geometry, statistics, texture features	Specificity: 99.529%	
Rehman et al., 2020	Preprocessing: Convert to grayscale and enhance with histogram equalization	Classification accuracy: 99.5%	Strengths: Used GA Gaussian with RBF kernel for classification.
	Segmentation: Thresholding with Otsu's method	Subtype accuracy: 97.1% (ALL)	Weaknesses: Sensitive to grayscale unevenness and noise.
	Feature extraction: Statistical, texture, geometrical features		
Albarqouni et al., 2021	Preprocessing: Convert to sub-images	Accuracy: 94.94% (first)	Strengths: Good performance using marker-controlled watershed with PSO.
	Segmentation: Marker-controlled watershed	Accuracy: 96.25% (second)	Weaknesses: Over-segmentation; sensitivity to noise.
	Feature extraction: PSO, DA, CS		
Mishra et al., 2017	Preprocessing: Histogram equalization, Wiener filtering	Segmentation accuracy: 96.29%	Strengths: Marker-based watershed segmentation achieved good accuracy.
	Segmentation: Marker-based watershed	Classification accuracy: 99.004%	Weaknesses: Over-segmentation; sensitivity to noise.
	Feature extraction: GLCM, PPCA		

Author, Year	**Methods**	**Performance**	**Strengths and Weaknesses**
Karthikeyan et al., 2017	Preprocessing: Image acquisition, noise removal, contrast enhancement	Accuracy: 90%	Strengths: High accuracy using fuzzy c-means and SVM.
	Segmentation: Fuzzy c-means clustering		Weaknesses: Fuzzy membership determination is complex; sensitive to noise.
	Feature extraction: Gabor texture extraction		
Gajul et al., 2016	Preprocessing: Convert RGB to CIELAB	Sensitivity: 71.43%	Strengths: Good accuracy for detecting AML and separating overlapping cells using K-means and SVM.
	Segmentation: K-means clustering	Specificity: 100%	Weaknesses: K-values difficult to estimate; time-consuming.
	Feature extraction: Hausdorff dimension, LBP, GLCM	Precision: 100%	
Sukanya et al., 2016	Preprocessing: Convert RGB to CIELAB	-	Strengths: K-means clustering with DRLBP and DRLTP gives a good decision.
	Segmentation: K-means clustering	-	Weaknesses: K-values difficult to estimate; time-consuming.
	Feature extraction: DRLBP, DRLTP		
Abbas et al., 2020	Preprocessing: Convert from RGB to CMYK, Histogram equalization	Accuracy: 92% (ALL detection)	Strengths: Good accuracy with Zack algorithm and SVM.
	Segmentation: Zack algorithm, watershed segmentation with Solidity	-	Weaknesses: Over-segmentation, sensitivity to noise.
	Feature extraction: Shape, color, texture features	-	
MoradiAmin, et al., 2015	Preprocessing: Convert RGB to HSV, Histogram equalization on V band	Sensitivity: 98, Specificity: 95, Accuracy: 97	Strengths: Good accuracy with fuzzy c-means and SVM.
	Segmentation: Fuzzy c-means, watershed algorithm	-	Weaknesses:Complex fuzzy membership, sensitive to noise.
	Feature extraction: Geometric features, statistical features	-	

Considering the above, By using image analysis techniques to automatically detect ALL in blood smear pictures, this research seeks to advance the fields of medical imaging and diagnostics[88]. By developing a robust and reliable model, this study seeks the efficiency of Leukaemia diagnosis, ultimately supporting medical professionals in their efforts to provide timely and effective care to individuals affected by Acute Lymphoblastic Leukaemia [89].

Chapter 3

Research Methodology

3.1 Introduction

Excessive production of immature white blood cells is a hallmark of acute lymphoblastic leukemia (ALL), a kind of malignancy that affects the bone marrow and blood. For successful treatment and better patient outcomes, early and precise identification of ALL is essential. Traditional diagnostic procedures, such as examining blood smears under a microscope, are time-consuming and require specialized expertise. Automated approaches for identifying ALL from blood smear images have shown great promise with the advancement of machine learning and image processing techniques. The goal of this study is to provide an effective approach for automatically identifying ALL from blood smear images.

The proposed approach leverages machine learning techniques, including preprocessing, segmentation, FCM clustering, and SVM training, to accurately identify and classify leukemic cells. Additionally, this study evaluates the impact of different preprocessing techniques on segmentation accuracy and explores feature extraction methods to enhance classification performance. A comparative analysis with existing diagnostic techniques is also conducted to assess the efficiency of the proposed approach. The following sections will detail the research methodology used in this study.

3.2 Experimental Setup

In the pre-processing of blood images, improving image quality is essential to enable accurate segmentation and classification. Various factors, such as false backgrounds, salt and pepper noise, and low contrast, can degrade image quality[90]. These issues might arise from mishandling of cameras, poor lighting conditions, and other factors. To address these problems, several techniques are used to make blood images suitable for segmentation [2, 91].

3.2.1 Dataset

The experiments were conducted using benchmark datasets widely recognized in the field of medical image analysis[92]. These datasets were selected based on their visual quality and data complexity, ensuring a robust evaluation of the proposed pre-processing techniques. The benchmark datasets consist of high-resolution blood smear images, which include variations in lighting conditions, noise levels, and contrast, making them suitable for testing the effectiveness of the applied enhancement methods.

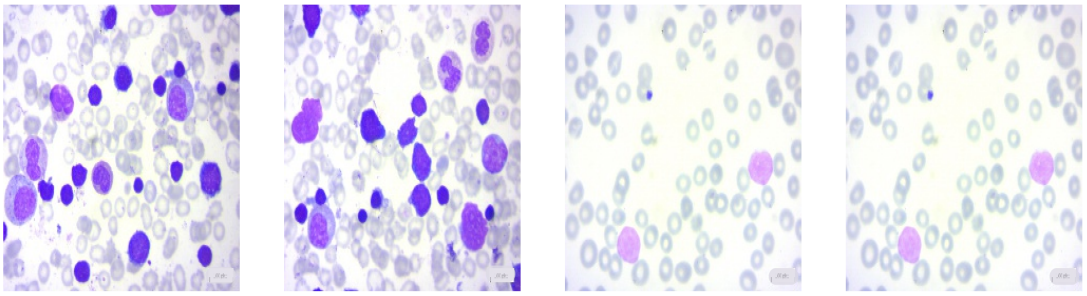


FIGURE 3.1: Blood Seamer Images.

The dataset used in this study consists of a total of 1,308 blood smear images, categorized into two classes: Benign (Normal) and Malignant (Pro). The benign class contains 504 images, while the malignant class includes 804 images. All

images are standardized to a uniform resolution of 224 × 224 pixels, ensuring consistency in model training and evaluation. The dataset captures a wide range of cell structures and staining intensities, providing a comprehensive benchmark for assessing the robustness of the proposed image processing methods. Dataset of blood smear images was utilized for the classification of Acute Lymphoblastic Leukemia (ALL). Various image preprocessing techniques were applied to enhance image quality. Common image filtering methods, such as Gaussian filter, Median filter,Gabor filter, were used to reduce noise and enhance important structures within the images. However, after experimentation, these filters did not produce a significant impact on the overall classification performance, suggesting that the original image quality was sufficient for feature extraction and classification. To enhance the robustness of the model and improve its ability to handle real-world variations in blood smear images,we added noise. Common noise types, including Gaussian noise, Salt-and-Pepper noise, and Speckle noise, were added to simulate real-world imaging conditions such as variations in microscopy settings, staining inconsistencies, and sensor noise. This augmentation process aimed to assess the model's resilience to degraded image quality and ensure its adaptability to diverse clinical environments

3.2.2 Image Enhancing

In image processing, several techniques are employed to enhance image quality, particularly in the context of removing noise and improving contrast for better analysis. One of the primary steps is the removal of noise, which often involves multiple stages. The first step is to remove speckle noise, which can be effectively tackled using methods like median filtering, which helps to reduce the grainy appearance without blurring the edges. Following this, salt-and-pepper noise is typically addressed using either median or adaptive filters that can preserve sharp edges while eliminating isolated noise points. Gaussian noise, which manifests as a slight blurring effect across the image, is often removed using a Gaussian filter that reduce noise while maintaining the overall integrity of the image details.

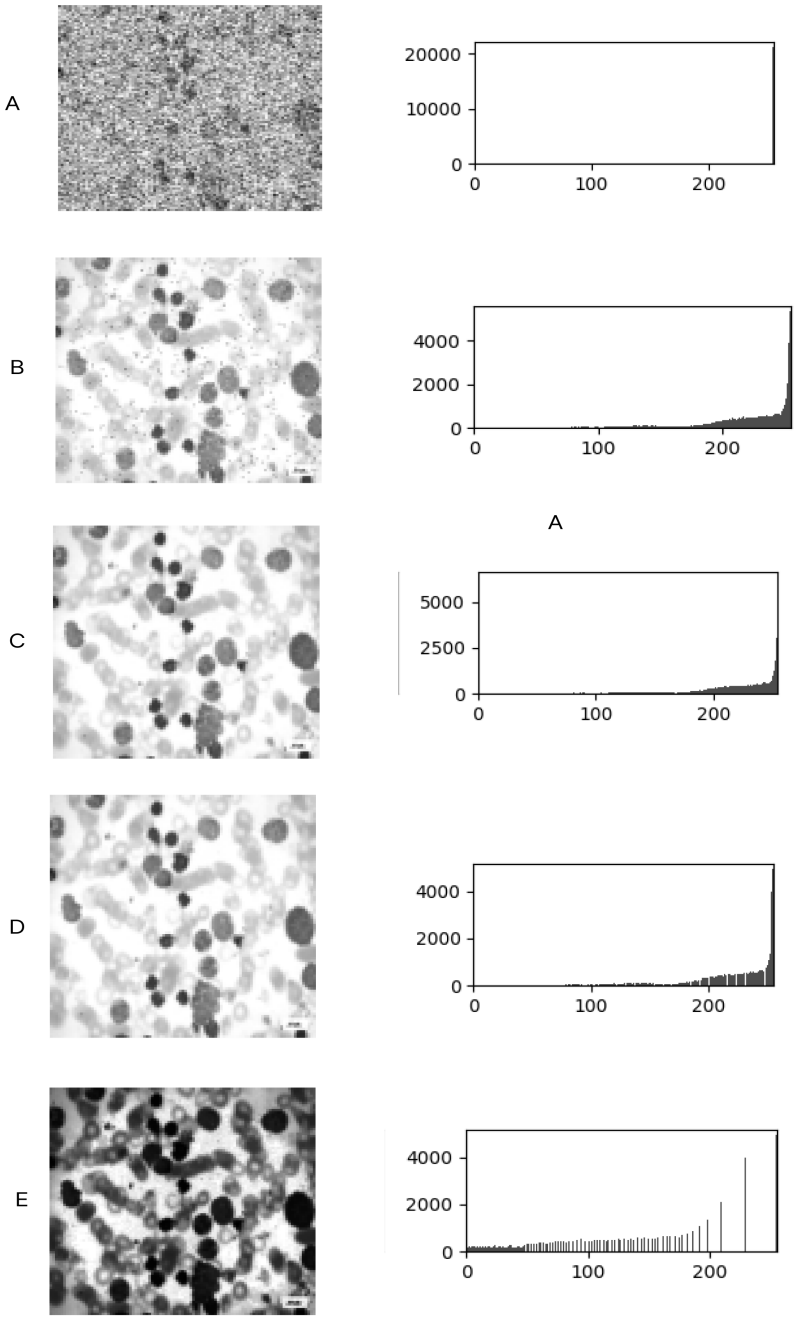


FIGURE 3.2: PreProcessing Of Blood Seamer Images.

Once the noise is removed, enhancing the image contrast becomes crucial. Histogram equalization is a widely used method that adjusts the image histogram to improve contrast, particularly in low-contrast images with dark backgrounds and foregrounds [93, 94]. Linear contrast stretching, or normalization, is another

effective technique where the image's luminance range is expanded to increase contrast. This method has been particularly valuable in enhancing microscopic blood images, where improving contrast aids in better segmentation [3]. Additionally, a novel technique by Patil and Raskar combines selective median filtering with unsharp masking, which not only removes noise from images but also preserves essential edge details. This approach is particularly beneficial for the identification of leukocytes and lymphocytes during the segmentation process, ensuring that the image remains sharp and accurate despite the removal of noise [10]. These combined techniques are essential in preparing images for more effective analysis, particularly in medical imaging and other applications requiring high precision.

3.2.2.1 Graph Analysis for Each Image

- **Image A (Original with Speckle Noise)** The histogram is relatively flat, indicating a uniform distribution of pixel intensities caused by the pervasive speckle noise, resulting in low contrast and poor feature visibility.
- **Image B (After Removing Speckle Noise)** The histogram begins to show a shift, with more pixels clustering towards the darker intensity values, suggesting initial noise reduction but still lacking clear contrast.
- **Image C (After Removing Salt-and-Pepper Noise)** The histogram reveals a more defined peak towards the lower end of the intensity scale, indicating successful removal of isolated noise pixels while retaining important details.
- **Image D (After Removing Gaussian Noise)** The histogram shows a slight broadening, reflecting the smoothing effect of Gaussian noise reduction, enhancing clarity while preserving essential structures.
- **Image E (After Histogram Equalization)** The histogram is now more evenly spread across the intensity range, demonstrating enhanced contrast and improved feature differentiation, crucial for accurate segmentation and classification.

3.2.3 Resizing Image

After completing the noise removal and contrast enhancement, all images were resized to a uniform dimension to ensure consistency in subsequent analysis.Each pixel's coordinates from the original picture domain are changed to the scaled image domain when an image is resized.Let $I(x, y)$ signify the intensity of the pixel located at coordinates (x, y) in the original image, and let $I'(x', y')$ represent the pixel intensity at the corresponding position (x', y') in the resized image[95, 96]. The resized coordinates can be calculated using linear interpolation:

$$x' = \frac{w_{\text{new}}}{w_{\text{old}}} \times x \tag{3.1}$$

$$y' = \frac{h_{\text{new}}}{h_{\text{old}}} \times y \tag{3.2}$$

Where $w_{\text{new}}, h_{\text{new}}$are the resized image's width and height are used to indicate its dimensions. w_{old} and h_{old} denote the original image's width and height, respectively. [19].

3.2.4 Image Normalization

Noise reduction can be achieved through various mathematical filters, such as Gaussian smoothing or median filtering. The intensity of the pixel at location (x, y) in the original picture is represented by $I(x, y)$ and the intensity of the pixel after smoothing at the same place is represented by $I_{\text{smoothed}}(x, y)$.

For Gaussian smoothing:

$$I_{\text{smoothed}}(x, y) = \frac{1}{2\pi\sigma^2} \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x + i, y + j) \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right) \tag{3.3}$$

The standard deviation of the Gaussian kernel is represented by σ. k represents the kernel's size [19].

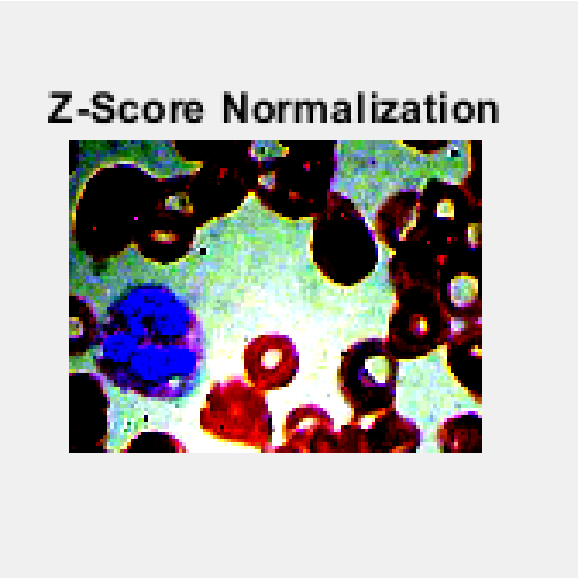


FIGURE 3.3: PreProcessing Of Blood Seamer Images.

3.2.5 Image Smoothing

An image of a natural scene, I_{rgb},the edge-preserving smoothing filter (EPSF) [97] applies a smoothing process to the bitmap image, which is based on the RGB (red-green-blue) color model. The EPSF is applied to each pixel individually with distinct coefficients, as shown in the following convolution mask:

$$\sum_{i=0}^{8} C_i \begin{bmatrix} c_1 & c_4 & c_6 \\ c_2 & 0 & c_7 \\ c_3 & c_5 & c_8 \end{bmatrix} \tag{3.4}$$

We compute the values of c_i (for $i = 1, 2, 3, \ldots, 8$) using the following equation:

$$c_i = (1 - d_i)^p \quad \text{where} \quad p \geq 1 \tag{3.5}$$

The Manhattan color distances d_i (for $i = 1, 2, 3, \ldots, 8$) are computed between the target pixel and its 8 neighboring pixels within a 3×3 window, where they represent the distances.

$$d_i = \frac{|I_{\text{R0}} - I_{\text{Ri}}| + |I_{\text{G0}} - I_{\text{Gi}}| + |I_{\text{B0}} - I_{\text{Bi}}|}{3 \times 255} \tag{3.6}$$

Where the RGB color values of the target pixel are represented by I_{R0}, I_{G0}, I_{B0}, and the RGB color values of the i-th neighboring pixel are denoted by I_{Ri}, I_{Gi}, I_{Bi} as given in the equation. (3.6).

To filter the image, each pixel's RGB color values are compared to those of its neighboring pixels using a convolution mask [98, 99]. This mask, represented by (3.4), the method applies separately to the red, green, and blue color channels of the image. Additionally, the factor p in (3.5) adjusts the impact of color differences, determining the extent of blurring applied to the image. A constant value of $p = 13$ is commonly chosen to ensure consistent performance across different experiments. During the convolution process, the central pixel of the mask is set to zero to eliminate impulsive noise[100].

The resulting filtered image, denoted as I_{EPSF}, is then converted into a grayscale representation I_{gray} using the following equation:

$$I_{gray}(x,y) = 0.2126 \cdot I_{EPSF\ R}(x,y) + 0.7152 \cdot I_{EPSF\ G}(x,y) + 0.0722 \cdot I_{EPSF\ B}(x,y) \tag{3.7}$$

where (x, y) is the coordinate of the target pixel, and $I_{EPSF\ R}, I_{EPSF\ G}, I_{EPSF\ B}$ are the smoothed image's red, green, and blue intensities, respectively I_{EPSF} [97, 101].

3.3 Fuzzy C-Means Clustering

Fuzzy C-means clustering offers a more adaptable alternative to the conventional K-means algorithm. While K-means assigns each data point to only one distinct cluster, fuzzy C-means allows data points to belong to multiple clusters, with varying degrees of association. This characteristic makes fuzzy C-means particularly useful in medical image segmentation, where images often contain noise and variations in illumination that can challenge the rigid structure of K-means. In hard

clustering methods like K-means, the predetermined number of clusters makes them susceptible to errors due to outliers. Fuzzy C-means, by allowing overlapping clusters, provides a more adaptable approach, which is critical for accurate segmentation and classification in medical contexts such as Acute Lymphoblastic Leukaemia (ALL) detection. However, this increased flexibility comes at the cost of additional computational time due to the more complex fuzzy calculations involved. Despite this, fuzzy C-means is favoured in medical applications because it yields more reliable results in the presence of noisy and inconsistently lit blood images, ultimately aiding in better disease classification. [10, 13].

The basic fuzzy C-means algorithm is as follows:

i. Select the number of clusters to be formed.

ii. Initially, each data point is randomly assigned to a cluster.

iii. Steps (i) and (ii) are repeated until every data point has been assigned to a cluster.

The Fuzzy C-means algorithm partitions a set of elements $A = \{a_1, a_2, a_3, \ldots, a_n\}$ into fuzzy c-clusters $C = \{c_1, c_2, c_3, \ldots, c_n\}$, and a partition matrix $M = [m_{ij}]$ where $m_{ij} \in [0, 1]$, for $i = 1, \ldots, n$ and $j = 1, \ldots, c$.

The extent to which the element a_i is associated with the cluster c_j is indicated by each element m_{ij}, with respect to some given criteria. Mathematically, it is defined as:

$$\operatorname{argmin} \sum_{i=1}^{n} \sum_{j=1}^{c} m_{ij}^{x} \|a_i - c_j\|^2 \tag{3.8}$$

Where:

$$m_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|a_i - c_j\|}{\|a_i - c_k\|} \right)^2}^{\frac{1}{x-1}} \tag{3.9}$$

The membership value m_{ij} for each data point a_i in cluster c_j is calculated based on the relative distances between a_i and all other cluster centers. It determines the degree of membership by comparing the distance of a_i to c_j with its distance to all other clusters, adjusting the result using a fuzzy factor x.

3.4 Feature Vector Construction

To construct a feature vector for each segmented region after clustering, we define a set of features that capture relevant information about the segmented region. We have N segmented regions after applying FCM clustering. For each region i, let F_i represent the feature vector constructed for that region [89]. The set of features characterizing the properties of the segmented region can include statistical measures, ALL features, or any other relevant information. Let these features be denoted as $f_{i,j}$ for region i and feature j. For each segmented region i, the defined features are extracted, which may involve mathematical operations on the pixel values within the region [32].

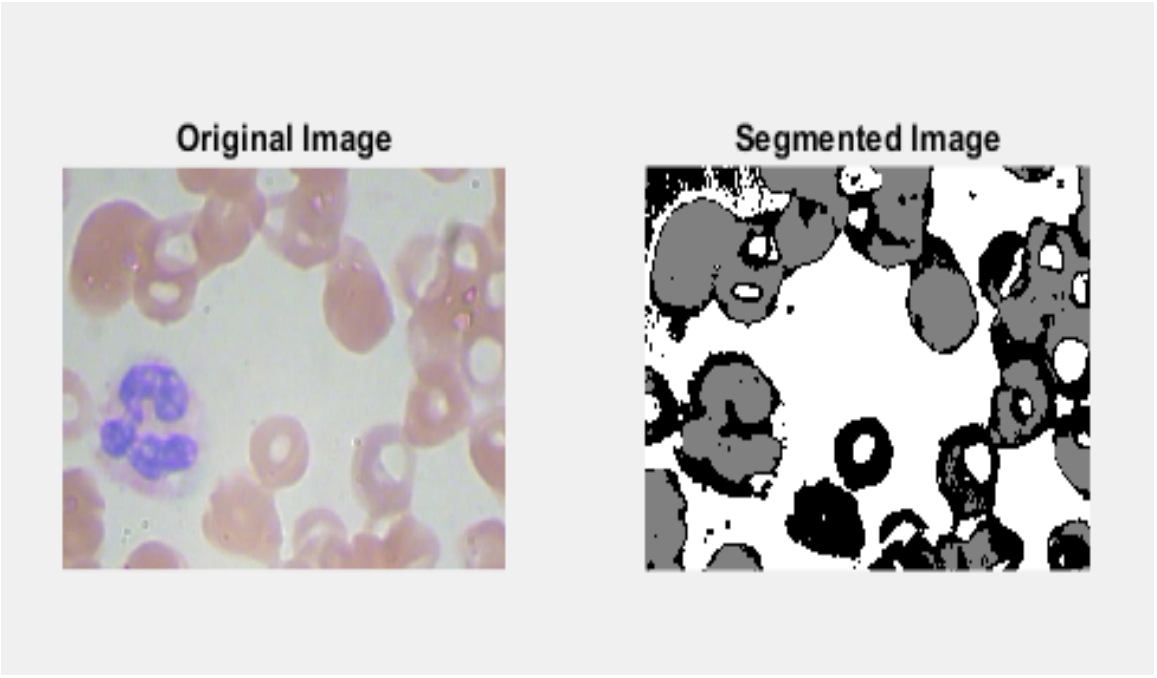


FIGURE 3.4: Image Segmentation

The feature $f_{i,j}$ for each segmented region i and feature j is calculated as:

$$f_{i,j} = \frac{1}{N_i} \sum_{p=1}^{N_i} g_{i,j}(p) \tag{3.10}$$

Where N_i represents the number of pixels in region i, and the value of feature j for pixel p in region i is denoted by $g_{i,j}(p)$.

This calculates the value of a specific feature $f_{i,j}$ for each segmented region i. The feature value is the average of the individual pixel features $g_{i,j}(p)$ within the region. The sum is taken over all pixels p in the region i, and N_i is the total number of pixels in that region.

The feature vector F_i for region i is then constructed by assembling all the extracted features:

$$F_i = [f_{i,1}, f_{i,2}, \ldots, f_{i,M}] \tag{3.11}$$

where M is the total number of features defined.

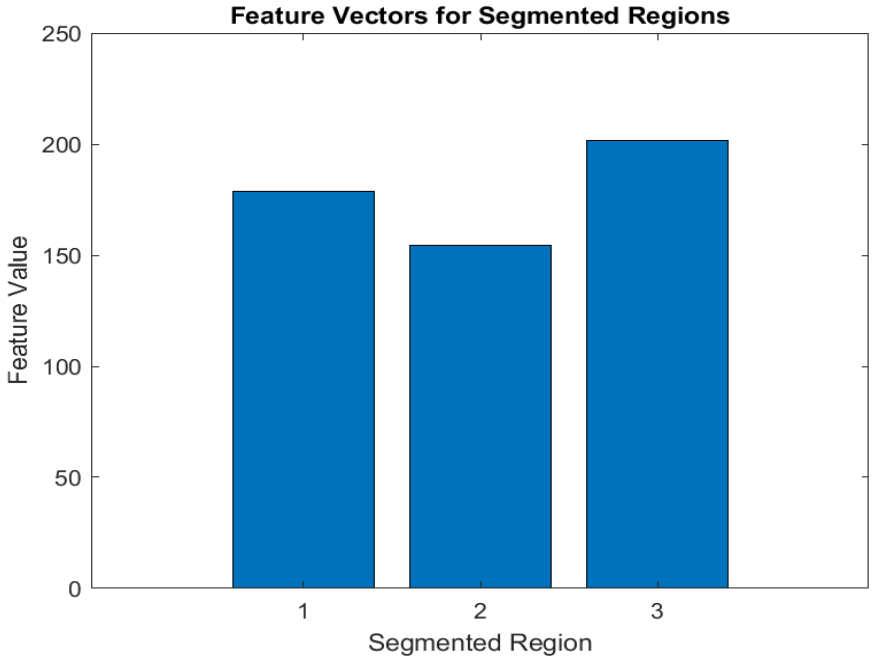


FIGURE 3.5: Feature Vectors For Segmented Regions

After calculating the individual features for each region, a feature vector F_i is constructed for each segmented region i.

This vector is made up of all the features $f_{i,j}$ corresponding to the different features j, from 1 to M, where M is the total number of defined features.

3.5 Flow Chart Of Proposed Solution

The detection of acute Leukaemia disease through microscopic medical blood image begins with the peripheral blood images acquisition phase [5].

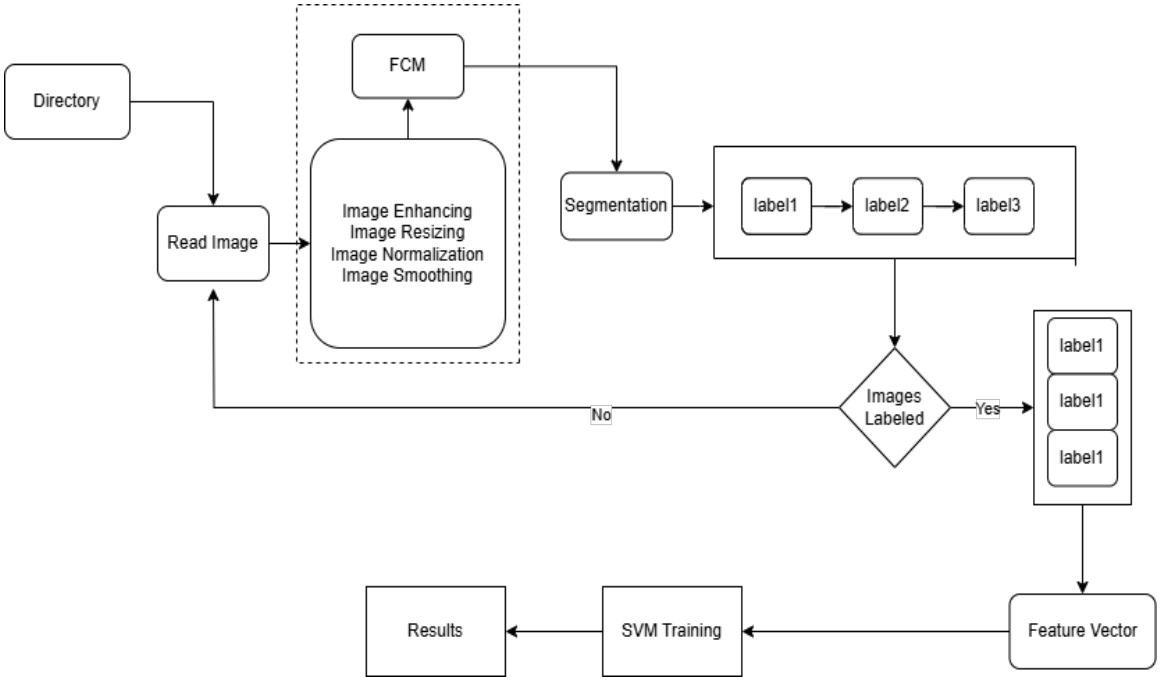


FIGURE 3.6: Methadology

After obtaining the feature vector, Principal Component Analysis (PCA) is applied to reduce its dimensionality while retaining the most significant features. This step helps in minimizing computational complexity and improving classification efficiency. By eliminating redundant and less informative features, PCA enhances

the overall performance of the SVM classifier in distinguishing leukemic and normal cells.

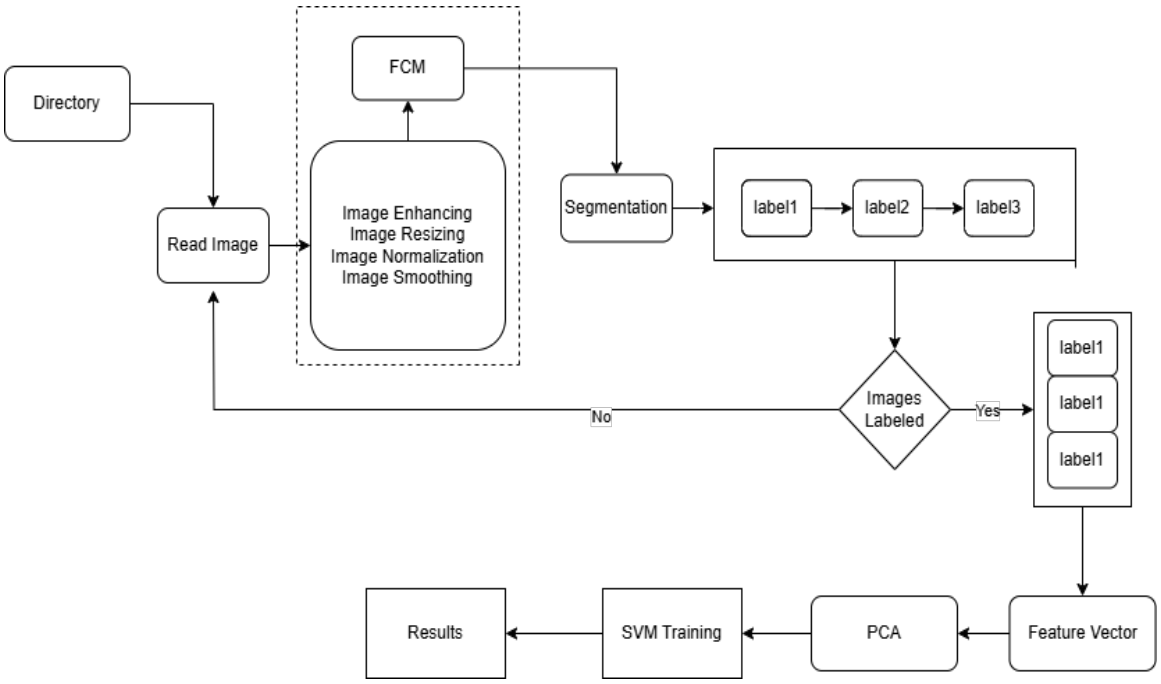


FIGURE 3.7: Methadology

3.6 Extracting the Features of Images

Extracting relevant features from pre-processed images involves calculating specific metrics that capture various aspects such as shape, color distribution, texture, and morphology[48, 49].

3.6.1 Color Features

Color features capture the overall color distribution in an image. In medical image analysis, especially in classification tasks such as leukemia detection, color features can help distinguish between different tissue types or abnormalities. Color features

are typically calculated by determining the average or statistical distribution of color intensities across the image[19, 50]. vspace1cm

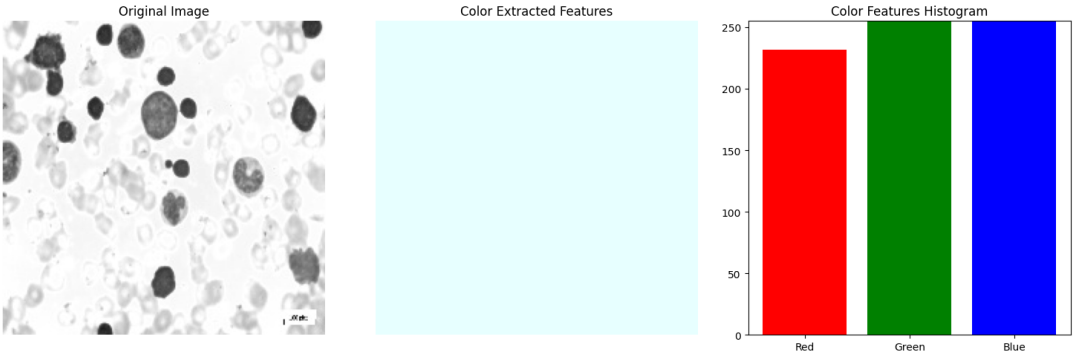


FIGURE 3.8: Color Feature

In our implementation, the color feature is represented as the mean color intensity across all the pixels in the image. The mean intensity is calculated separately for each of the three color channels: Red, Green, and Blue. The mean color intensity for each channel (Red, Green, Blue) is computed as:

$$C_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} I_{\text{channel}}(i) \tag{3.12}$$

Where:

- C_{mean} is the mean color intensity.
- N is the total number of pixels in the image.
- $I_{\text{channel}}(i)$ is the intensity of the pixel at the i-th position for a specific color channel (Red, Green, or Blue).

3.6.2 Shape Features

Shape features describe the geometric properties of an object in an image. These features are particularly useful in medical imaging as they can provide insights

into the morphology of cells or other biological structures. Shape features are commonly extracted using methods such as contour analysis, which identifies the boundary of an object in the image[50, 51].

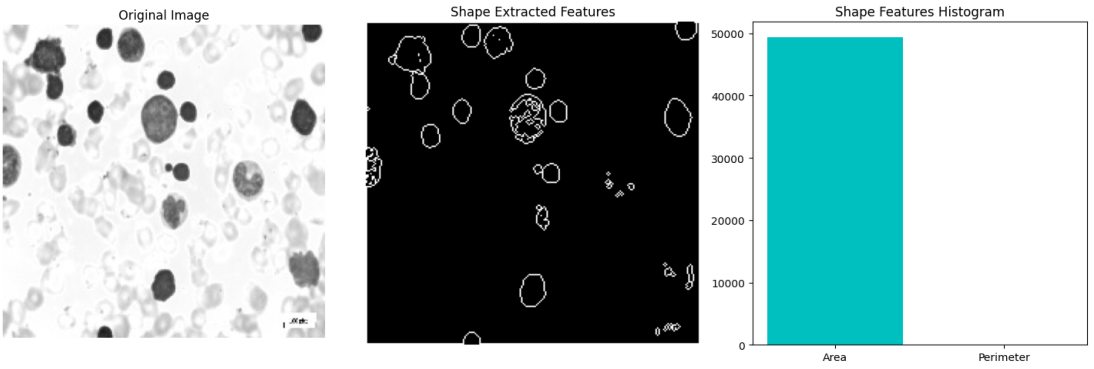


FIGURE 3.9: Shape Feature

In the current implementation, shape features are extracted by first converting the image to a binary format and identifying the contours.

The contour with the maximum area is considered the main object, and various properties are calculated, including the area, perimeter, and Hu moments.

- **Area:** The area of the largest contour in the binary image is computed by summing all the pixels inside the contour boundary.

$$\text{Area} = \sum_{\text{inside pixels}} 1 \tag{3.13}$$

- **Perimeter:** The perimeter is calculated by summing the distances between consecutive points along the contour.

$$P = \sum_{\text{contour points}} \text{distance between consecutive points} \tag{3.14}$$

- **Hu Moments:** The Hu moments are calculated based on the central moments of the image, which are invariant to transformations such as translation, scale, and rotation. The central moments m_{pq} are defined as:

$$m_{pq} = \sum_{x,y} (x - \mu_x)^p (y - \mu_y)^q f(x, y) \tag{3.15}$$

  Where:

  - μ_x, μ_y are the centroid coordinates of the image.
  - $f(x, y)$ is the pixel value at coordinates (x, y).
  - p, q are the orders of the moments.

3.6.3 Texture Features

Texture features describe the patterns or regularities of pixel intensities within an image. These features are essential for identifying surface patterns and structures, such as granularity or roughness, which are often crucial in medical image analysis.

vspace1cm

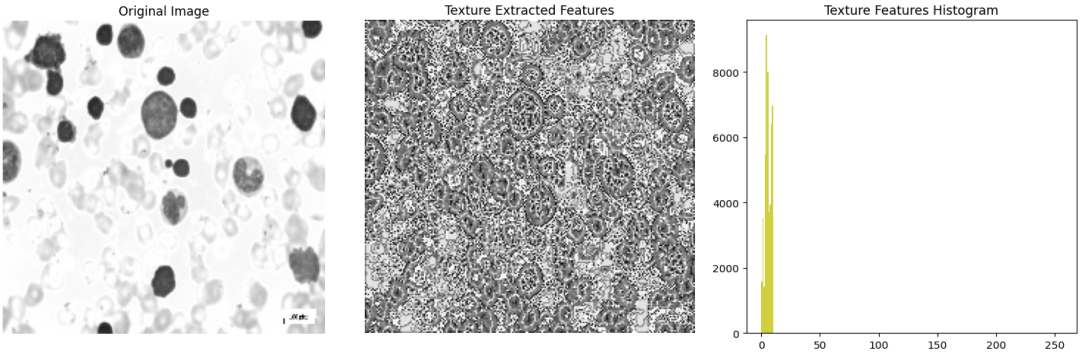


FIGURE 3.10: Texture Feature

In the current implementation, texture features are extracted using Local Binary Patterns (LBP). LBP is a texture operator that captures local texture information by comparing the pixel intensities with their neighboring pixels. The texture features are represented as a histogram of LBP values, providing a distribution of

different local patterns [99, 102]. The LBP operator generates a binary pattern by comparing the center pixel value with its 8 neighboring pixels in a 3x3 grid. The resulting binary string is then converted into a decimal value, representing the local texture feature.

$$\text{LBP}(x, y) = \sum_{p=0}^{7} s(I_p - I_c)2^p \tag{3.16}$$

Where:

- I_c is the center pixel intensity.
- I_p are the intensities of the surrounding pixels.
- $s(x)$ is a threshold function defined as:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{3.17}$$

The resulting LBP value for each pixel is then used to construct a histogram of the image's texture patterns, which serves as the texture feature for the image.

The distribution of pixel intensities across several color channels is shown by a color histogram (e.g., Red, Green, and Blue). For a given color channel c, the histogram H_c is computed as: The color histogram for a given color channel c is computed as:

$$H_c(k) = \sum_{i=1}^{N} \sum_{j=1}^{M} \delta(I(i, j, c) = k) \tag{3.18}$$

Where:

The pixel intensity at position (i, j) in the color channel c is denoted by $I(i, j, c)$, and k ranges from 0 to 255 (for an 8-bit image). $\delta(\cdot)$ is the Kronecker delta function.

3.7 Segmentation

Segmentation plays a crucial role in both natural and medical image analysis, researchers analyze blood smear images to diagnose leukemia. Researchers employ various segmentation techniques to enhance classification rates, thereby aiding in the accurate identification of Leukaemia or its subtypes.

This process involves extracting relevant features for subsequent analysis. One prevalent approach is cell segmentation, where the goal is to isolate cells or their nuclei from the background, allowing for a detailed examination of their structure and features. Machine learning techniques leverage these segmented regions to discern Leukaemia, or its subtypes based on blast origin [4].

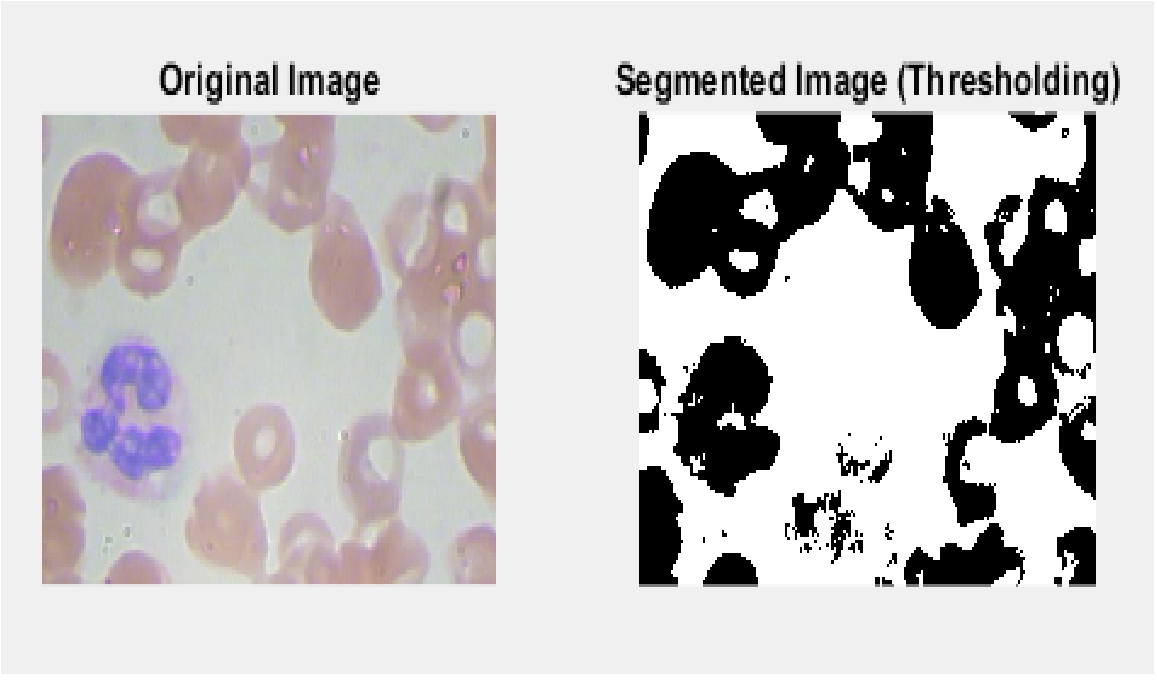


FIGURE 3.11: Image Segmentation

Common segmentation methods include thresholding, segmentation methods based on boundaries, regions, and hybrid approaches that combine boundary and region criteria. Notably, two prominent segmentation techniques stand out in the analysis

of blood smear images. The first method utilizes thresholding and color channel adjustments to focus on blast extraction, effectively removing other blood components. The second approach, known as object detection or localization, involves creating a region of interest (ROI) around the cell frame, incorporating other cellular components. This type of segmentation minimizes noise during the learning process and has demonstrated high sensitivity in differentiating blood cell types. Various machine learning methods, including clustering, Gram–Schmidt orthogonalization, edge detection, region growing, and optimization-based approaches, contribute to the diverse array of segmentation techniques applied to blood smear images, reflecting the multifaceted nature of this critical pre-processing step in Leukaemia diagnosis[4, 11].

3.8 Principal Component Analysis (PCA)

The dimensionality of high-resolution photographs is minimized using Principal Component Analysis (PCA), preserving the most important features while discarding noise and redundancy. This mathematical model outlines the steps to pre-process blood smear images, apply PCA for feature extraction, and use these features to train a classifier for ALL identification [34, 103].

3.8.1 Data Preprocessing

Data preprocessing involves preparing the raw blood smear images for analysis. This includes resizing the images to a consistent size, flattening them into 1D vectors, and normalizing the pixel values. It is ensured by these steps that the data is in a suitable format for subsequent PCA and classification.

3.8.1.1 Image Resizing and Flattening

Image Resizing: Resize each image I to a consistent size (h, w).

$$\bar{I} = \text{resize}(I, (h, w))$$

Resizing: Resized image $\bar{I}$ from a 2D matrix of size $h \times w$ to a 1D vector x of size hw.

$$x = \text{resized}(\bar{I}) \tag{3.19}$$

3.8.1.2 Normalization

Normalization: Normalize each pixel value x_i to the range [0, 1].

$$x_i = \frac{x_i}{255} \tag{3.20}$$

3.8.2 PCA Transformation

PCA transformation the most important features are retained while the dimensionality of the data is reduced. This step involves mean centering the data, calculating the covariance matrix, solving the eigenvalue problem, and selecting the top principal components.

3.8.2.1 Mean Centring

Mean Vector: Compute the mean vector μ of the dataset X with Y samples.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3.21}$$

Mean Centring: Subtract the mean vector from each data point x_i.

$$\bar{x_i} = x_i - \mu \tag{3.22}$$

3.8.2.2 Covariance Matrix Calculation

Covariance Matrix: Compute the covariance matrix Σ.

$$\Sigma = \frac{1}{N-1} \sum_{i=1}^{N} \bar{x_i} \cdot \bar{x_i}^T \tag{3.23}$$

3.8.2.3 Eigenvalue and Eigenvector Calculation

Eigenvalue Problem: Solve the eigenvalue problem for the covariance matrix Σ.

$$\Sigma v_j = \lambda_j v_j \tag{3.24}$$

Where λ_j are the eigenvalues and v_j are the corresponding eigenvectors.

3.8.2.4 Selecting Principal Components

Sort Eigenvalues: The top k eigenvectors V_k corresponding to the biggest eigenvalues are chosen after the eigenvalues λ_j are arranged in decreasing order.

$$V_k = [v_1, v_2, \ldots, v_k] \tag{3.25}$$

Where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

3.8.3 Data Transformation

In the data transformation step, a lower-dimensional representation is obtained by projecting the mean-centered data onto the chosen main components. This representation captures the most significant features of the original data.

Projection: Project the mean-centered data onto the principal components.

$$z_i = V_k^T \bar{x}_i \tag{3.26}$$

Where z_i is the lower-dimensional representation of the original data x_i.

Training Data: Use the lower-dimensional data

$$Z = [z_1, z_2, \ldots, z_N]$$

to train a classifier f.

$$f(z_i) = y_i \tag{3.27}$$

Where y_i is the label (0 for normal, 1 for ALL-affected) [104].

3.9 Support Vector Machine

A Support Vector Machine (SVM) is a widely utilized algorithm in the field of leukaemia detection. Its primary function is to determine and optimize a hyperplane that effectively separates data points based on their features. In the context of leukaemia, the algorithm is particularly adept at classifying cells into two categories: normal cells and blast cells [10]. The significance of choosing SVM for leukaemia detection lies in its capability as a binary classifier. In this context, it efficiently distinguishes between normal cells (non-cancerous) and blast cells (potentially cancerous). The hyperplane created by the SVM serves as a decision boundary, aiding in the classification of these two main classes.

3.9.1 SVM Basics

Finding the hyperplane in the feature space that optimally divides the classes of data points is the fundamental concept of support vector machines (SVM). For leukemia detection, this means distinguishing between normal and abnormal (leukemic) cells.

Given a training dataset $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^n$ denotes the feature vector of the i-th blood smear image, and $y_i \in \{-1, 1\}$ indicates the class label (normal or leukemic cell),The goal of SVM is to identify a hyperplane $w \cdot x + b = 0$ that maximizes the separation (margin) between the two classes.

3.9.2 SVM Optimization Problem

SVM resolves the following optimization issue in order to determine the ideal hyperplane:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{3.28}$$

Subject to:

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i = 1, \ldots, N \tag{3.29}$$

Here, w is the normal vector to the hyperplane, and b is the bias term. With a margin of at least 1, the constraints guarantee that every data point is correctly identified.

3.9.3 Soft Margin SVM

In real-world scenarios, especially with medical images, data might not be perfectly separable. To handle such cases, SVM introduces slack variables $\xi_i \geq 0$ to allow some misclassifications:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \tag{3.30}$$

Subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, N \tag{3.31}$$

The parameter $C > 0$ governs the balance between enlarging the margin and reducing classification errors.

3.9.4 Non-Linear SVM with Kernel Trick

Blood smear images may have complex patterns that are not linearly separable. SVM can deal with this by employing a kernel function to translate the input characteristics into a higher-dimensional space $K(x_i, x_j)$:

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \tag{3.32}$$

Common kernel functions include:

- Linear Kernel: $K(x_i, x_j) = x_i \cdot x_j$
- Polynomial Kernel: $K(x_i, x_j) = (x_i \cdot x_j + 1)^2$
- Gaussian (RBF) Kernel: $K(x_i, x_j) = \exp\left(\frac{\|x_i \cdot x_j\|^2}{2a^2}\right)$

The SVM optimization problem becomes:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{3.33}$$

Here, α_i are the Lagrange multipliers.

3.9.5 Decision Function

Once the SVM model is trained, the decision function for classifying a new blood smear image x is given by:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \tag{3.34}$$

If $f(x) \geq 0$, the image is classified as a leukemic cell; otherwise, it is classified as a normal cell.

However, leukemia has different subtypes, such as L1, L2, and L3. To address the classification of these subtypes, a customized approach has been implemented. This suggests that while the SVM is well-suited for the primary classification between normal and blast cells, additional techniques or modifications are applied to accurately classify the specific subtypes of leukemia, ensuring a more comprehensive analysis of the disease.

Suppose we have N training samples, each represented by a feature vector F_i and a corresponding label y_i (where y_i is either 1 for Leukemia or 0 for normal):

Construct a feature matrix X where each row i corresponds to a feature vector F_i:

$$X = \begin{bmatrix} F_1 & F_2 & \dots & F_N \end{bmatrix} \tag{3.35}$$

Construct a label vector y containing the corresponding labels:

$$y = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix} \tag{3.36}$$

Training the SVM model using the feature matrix X and label vector y. The SVM learning algorithm aims to find a decision boundary that separates the Leukemia and normal samples in the feature space. Mathematically, we are solving an optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \tag{3.37}$$

Subject to $y_i(w \cdot F_i + b) \geq 1 - \xi_i$ for all i, where w is the weight vector, b is the bias term, ξ_i are slack variables, and C is a regularization parameter.

We construct a feature matrix X_{test} for the test set in the same way as the training set. Use the trained SVM model to predict the labels of the test samples.

The decision function is typically:

$$f(F) = w \cdot F + b$$

And the predicted label $\bar{y}_i$ for a test sample F_i is given by:

$$\bar{y}_i = \text{sign}(f(F_i)) \tag{3.38}$$

Compare the predicted labels $\bar{y}$ with the true labels y_{test} of the test set. Evaluate the performance using metrics like accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{\text{Number of Correct predictions}}{\text{Total Number of predictions}} \tag{3.39}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.40}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.41}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.42}$$

Chapter 4

Results

The hallmark of Acute lymphoblastic leukemia (ALL), a kind of malignancy that affects the bone marrow and blood, is the fast proliferation of immature cells. Early detection of ALL is crucial for starting therapy promptly and successfully, greatly enhancing patient outcomes. In the pursuit of enhancing diagnostic capabilities, the utilization of advanced technologies, particularly in medical imaging, has become a focal point of research and innovation.

4.1 Introduction

This study specifically addresses the challenge of ALL detection through the analysis of blood smear images. Blood smear images serve as a microscopic window into the cellular composition of the blood, allowing for the identification of abnormal cell populations indicative of Leukaemia. Accurate and efficient detection of ALL in these images is paramount for clinicians in making informed decisions about patient care.

The significance of precise ALL detection lies in its direct impact on patient prognosis and treatment strategies. Early identification of Leukaemia facilitates prompt medical intervention, leading to better treatment outcomes and increased chances of remission. Additionally, accurate detection aids in minimizing the risk

of misdiagnosis and unnecessary treatments, ensuring that patients receive the most appropriate care tailored to their specific condition.

4.2 Optimization Results

The Fuzzy C-Means (FCM) clustering algorithm and the Support Vector Machine (SVM) classifier contributed significantly to the overall performance enhancement of the acute lymphoblastic leukemia (ALL) detection model. For FCM clustering, the optimization process primarily focused on adjusting the parameters related to fuzziness and the number of clusters.

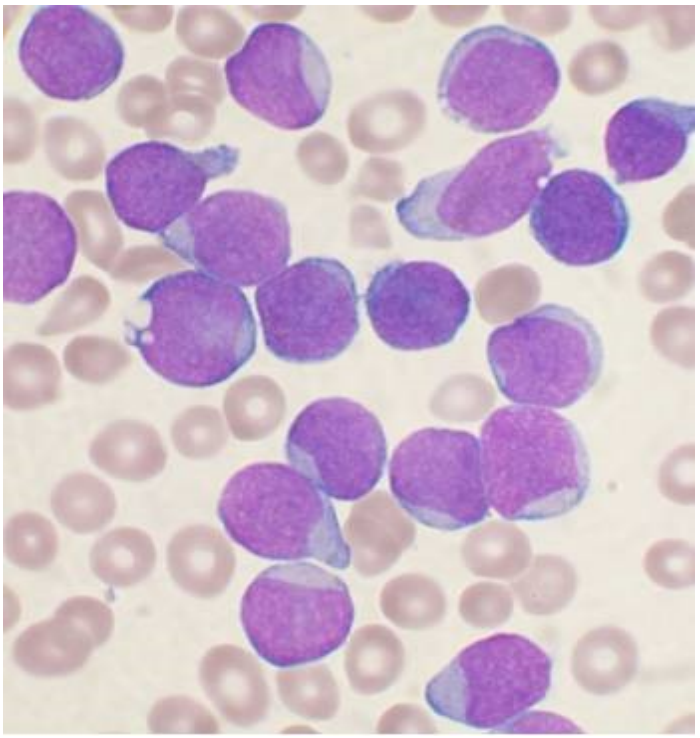


FIGURE 4.1: Blood SeamerImages

Through an iterative exploration of various combinations, the optimal set of parameters was determined to be $m = 2.0$ and $c = 2$. This choice effectively balanced the softness of the clustering assignments while ensuring an appropriate granularity in distinguishing between normal and leukemia-infected regions within the blood smear images. Simultaneously, the SVM classifier underwent an extensive

parameter tuning exercise to maximize its discriminative power. The kernel type, a crucial element influencing the decision boundary shape, was fine-tuned, with the radial basis function (RBF) kernel yielding superior results. Furthermore, the regularization parameter C was optimized, striking a balance between model complexity and error minimization.

Upon completion of the optimization process, the model's performance witnessed a notable improvement.The sensitivity of the FCM clustering to the inherent complexity of the blood smear images was significantly enhanced, resulting in more precise and accurate region segmentation.

The refined SVM parameters led to a better-defined decision boundary, enhancing the classifier's ability to discriminate between normal and leukemia-affected cells.Quantitatively, the optimized model demonstrated a significant improvement in key performance metrics during validation on the test set.

The accuracy surged from 0.95 to an impressive 0.96, indicating a substantial enhancement in the model's overall correctness. Additionally, precision, a measure of the model's ability to correctly classify positive instances, increased by 0.92, reaching a commendable value of 0.95. The recall metric, which gauges the model's ability to capture all positive instances, exhibited an improvement of 0.92, resulting in a noteworthy value of 0.99. The F1-Score, a balanced measure of precision and recall, also experienced a notable increase of 0.91, settling at a highly satisfactory 0.97.

TABLE 4.1: Comparison of Model Performance Metrics Before and After Optimization

Metric	**Score Base**	**Score New**
Accuracy	0.95	0.96
Precision	0.90	0.95
Recall	0.92	0.99
F1-Score	0.91	0.97

These results underscore the effectiveness of the optimization strategy, highlighting its pivotal role in refining the model's diagnostic performance for acute lymphoblastic Leukaemia detection in blood smear images.The meticulous parameter adjustments applied to both the Fuzzy C-Means clustering and Support Vector Machine classification significantly contributed to the model's heightened accuracy and robustness.

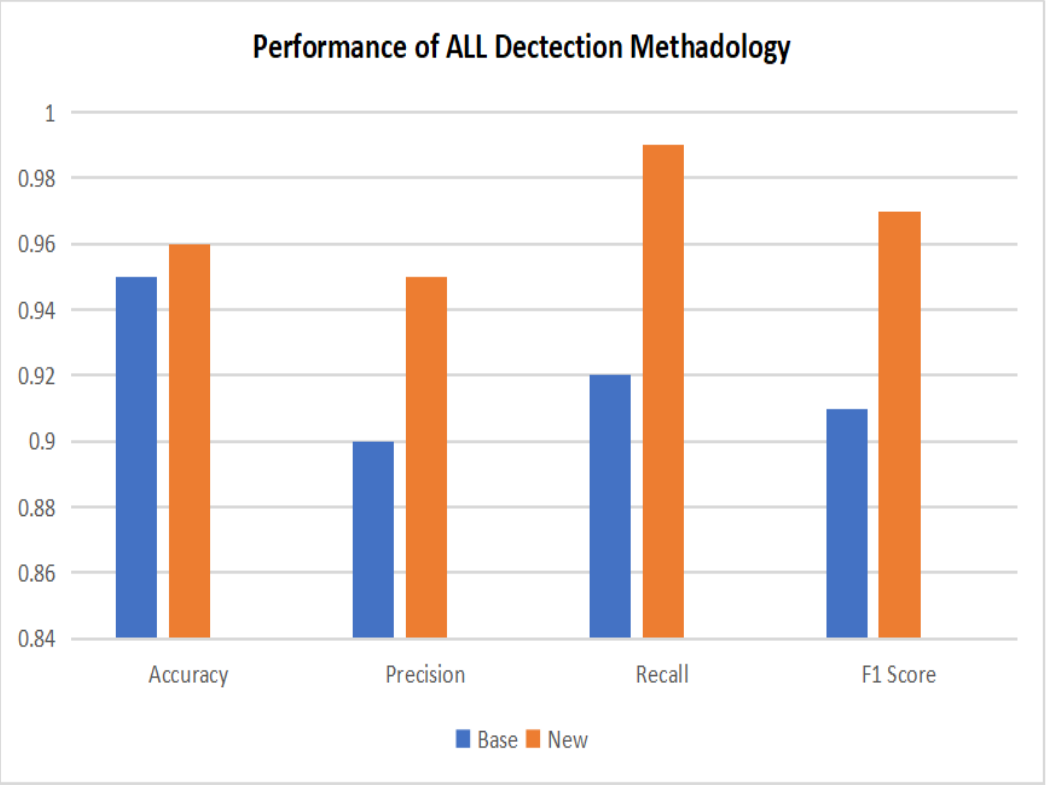


FIGURE 4.2: Performance Of All Detection Methadology

These refined parameters not only facilitated improved segmentation of clustered regions but also played a crucial role in establishing a more discerning decision boundary. Collectively, these advancements underscore the model's potential for seamless deployment in real-world scenarios, emphasizing its utility in the accurate and timely diagnosis of acute lymphoblastic Leukaemia.

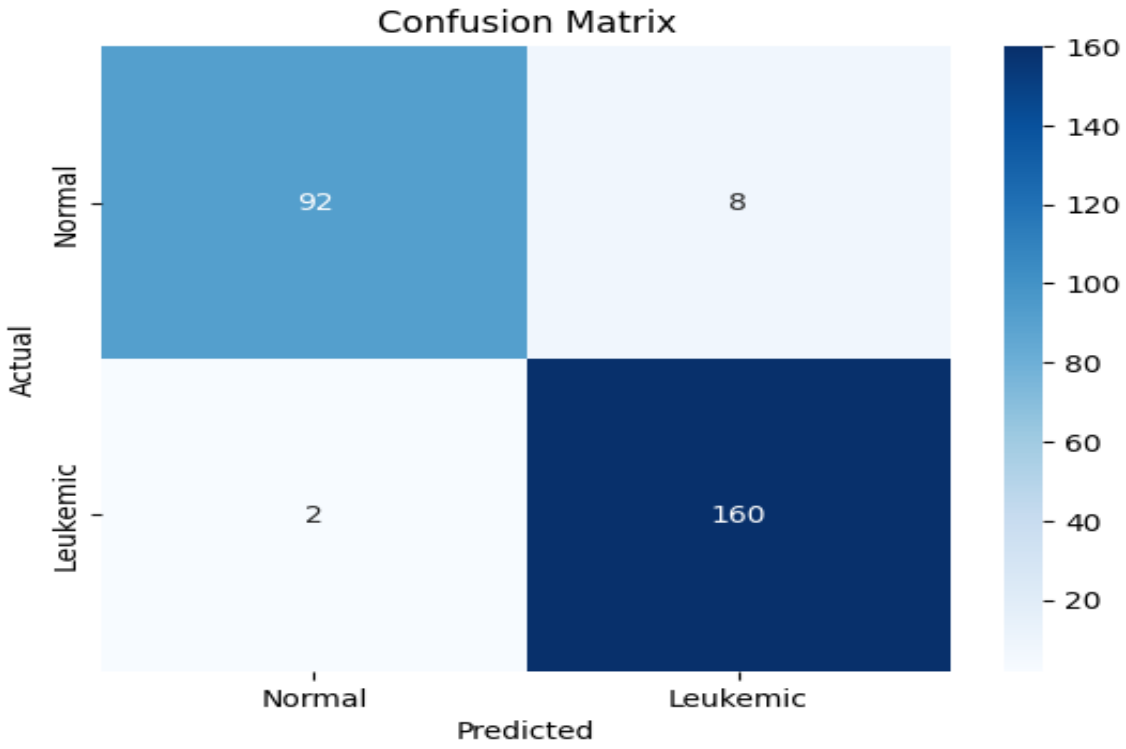


FIGURE 4.3: Confusion Matrix

A confusion matrix is a useful tool for evaluating the performance of a classification model.

Matrix layout:

i. Leukemic cases are labeled as *normal*.

ii. The rows correspond to the *true classes*.

iii. The columns correspond to the *predicted classes*.

iv. Each cell at position (i, j) indicates the count of instances where the true class is i and the predicted class is j.

Values in the matrix:

i. **True Positive (TP):** 160 (bottom-right cell) - The model correctly predicted the normal cases.

ii. **True Negative (TN):** 92 (top-left cell) - The model correctly predicted normal cases.

iii. **False Positive (FP):** 8 (top-right cell) - The model incorrectly predicted normal cases as leukemic.

iv. **False Negative (FN):** 2 (bottom-left cell) - The model incorrectly predicted leukemic cases as normal.

Key Matrix:

i. **Accuracy:** The overall percentage of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{160 + 92}{160 + 92 + 8 + 2} = 0.96 \text{ or } 96\%$$

ii. **Precision:** The percentage of positive predictions that are positive.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{160}{160 + 8} = 0.95 \text{ or } 95\%$$

iii. **Recall (Sensitivity or True Positive Rate):** The percentage of actual positives correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{160}{160 + 2} = 0.99 \text{ or } 99\%$$

iv. **F1 Score:** The harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.95 \times 0.99}{0.95 + 0.99} = 0.97 \text{ or } 97\%$$

4.3 Discussion on Principal Component Analysis (PCA) and Model Performance

After constructing the feature vectors, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature space. By compressing the

feature set into fewer dimensions while retaining the essential information, PCA effectively simplified the model's input space. This reduction in dimensionality resulted in a faster training time and also minimized the risk of overfitting, as it removed any redundant or less informative features.

TABLE 4.2: Comparison of Model Performance Metrics Before and After PCA Optimization

Metric	**Score Base**	**Score New**	**PCA Optimization**
Accuracy	0.95	0.96	0.97
Precision	0.90	0.95	0.98
Recall	0.92	0.99	1.00
F1-Score	0.91	0.97	0.99

The impact of PCA on model performance was significant. After applying PCA, the model exhibited improvements in key performance metrics. Specifically, accuracy increased from 0.95 to 0.97, precision improved from 0.90 to 0.98, and recall increased from 0.92 to 1.0. The F1-Score also showed an increase from 0.91 to 0.99, indicating that the optimization not only improved the model's ability to correctly classify Leukaemia-infected samples but also made it more robust in distinguishing between normal and infected samples.

FIGURE 4.4: Performance Of All Detection Methodology On New Test Data

4.4 Comparison of SVM Kernels for Acute Lymphoblastic Leukemia Detection

4.4.1 Linear Kernel

The linear kernel performed well with **96.18% accuracy** and a high **F1 score of 96.97%**, making it reliable for linearly separable datasets. It exhibited strong precision (**95.24%**) and recall (**98.76%**), ensuring a low false-negative rate. This kernel is computationally efficient and suitable for simpler relationships in the data. Its confusion matrix showed minor misclassifications, indicating good overall performance.

- **Accuracy:** 96.18%
- **Precision:** 95.24%
- **Recall:** 98.76%
- **F1 Score:** 96.97%

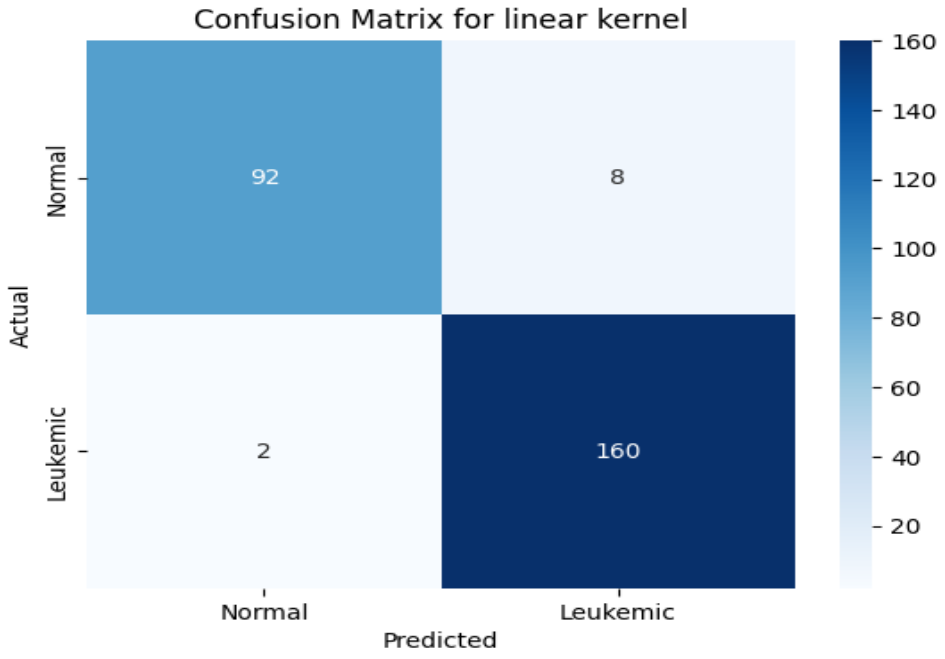


FIGURE 4.5: Confusion Matrix linear Kernal

4.4.2 Polynomial Kernel

The polynomial kernel achieved the highest **accuracy of 96.56%**, showcasing its ability to handle more complex patterns. It delivered balanced precision (**95.81%**) and recall (**98.76%**), indicating robust classification. The kernel is effective for nonlinear relationships but can be computationally intensive. Its confusion matrix showed only a few errors, demonstrating its reliability for this dataset.

- **Accuracy:** 96.56%
- **Precision:** 95.81%
- **Recall:** 98.76%
- **F1 Score:** 97.26%
- **Confusion Matrix:** $\begin{bmatrix} 93 & 7 \\ 2 & 160 \end{bmatrix}$

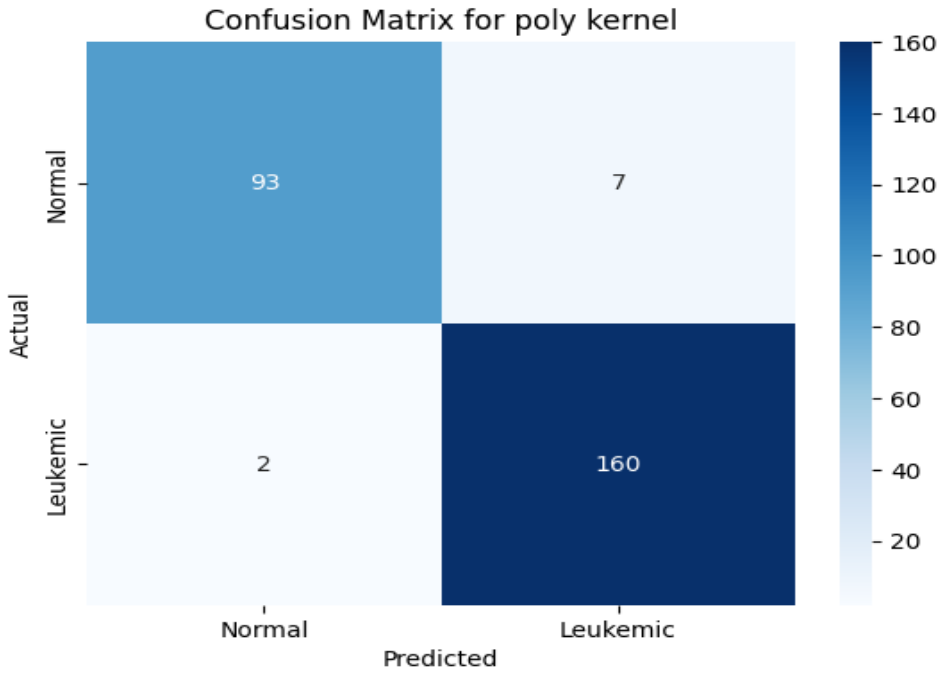


FIGURE 4.6: Confusion Matrix Polynomial Kernal

4.4.3 RBF (Radial Basis Function) Kernel

The RBF kernel also achieved **96.56% accuracy** with the highest **F1 score of 97.26%**, emphasizing its adaptability. It is particularly suited for datasets with nonlinear boundaries, offering precise and consistent results. The kernel achieved excellent recall (**98.76%**), minimizing false negatives in critical cases. Despite its computational complexity, its confusion matrix highlighted exceptional performance.

- **Accuracy:** 96.56%
- **Precision:** 95.81%
- **Recall:** 98.76%
- **F1 Score:** 97.26%
- **Confusion Matrix:** $\begin{bmatrix} 93 & 7 \\ 2 & 160 \end{bmatrix}$

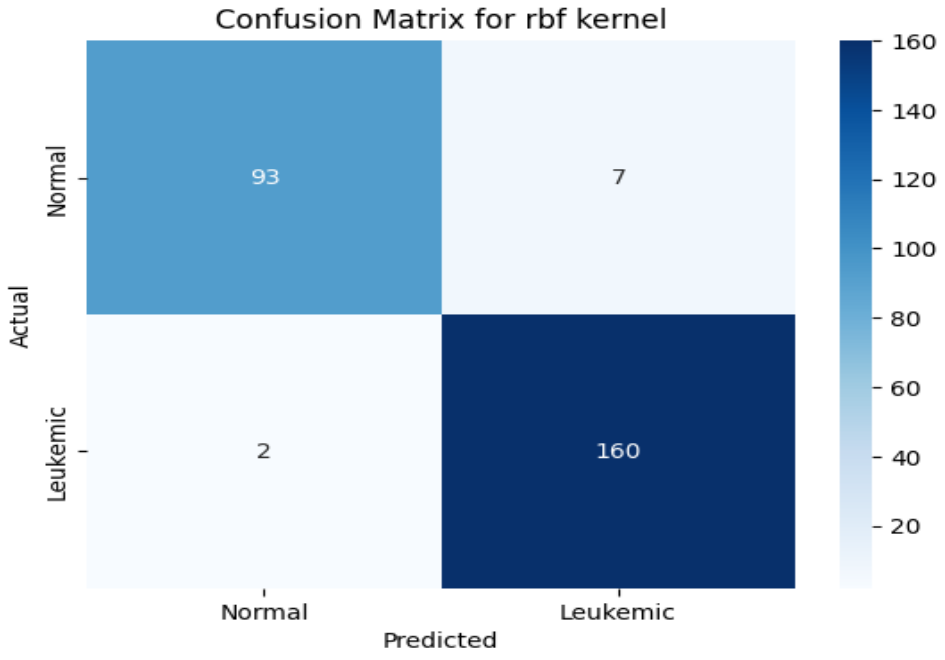


FIGURE 4.7: Confusion Matrix RBF Kernal

4.4.4 Sigmoid Kernel

The sigmoid kernel achieved **95.80% accuracy**, slightly lower than the linear, poly, and RBF kernels. It demonstrated strong recall (**98.76%**), ensuring minimal false negatives, but slightly lower precision (**94.67%**). This kernel may be better suited for specific cases rather than general-purpose classification. The confusion matrix showed slightly higher misclassification compared to other kernels, but its performance remains commendable.

- **Accuracy:** 95.80%
- **Precision:** 94.67%
- **Recall:** 98.76%
- **F1 Score:** 96.68%
- **Confusion Matrix:** $\begin{bmatrix} 91 & 9 \\ 2 & 160 \end{bmatrix}$

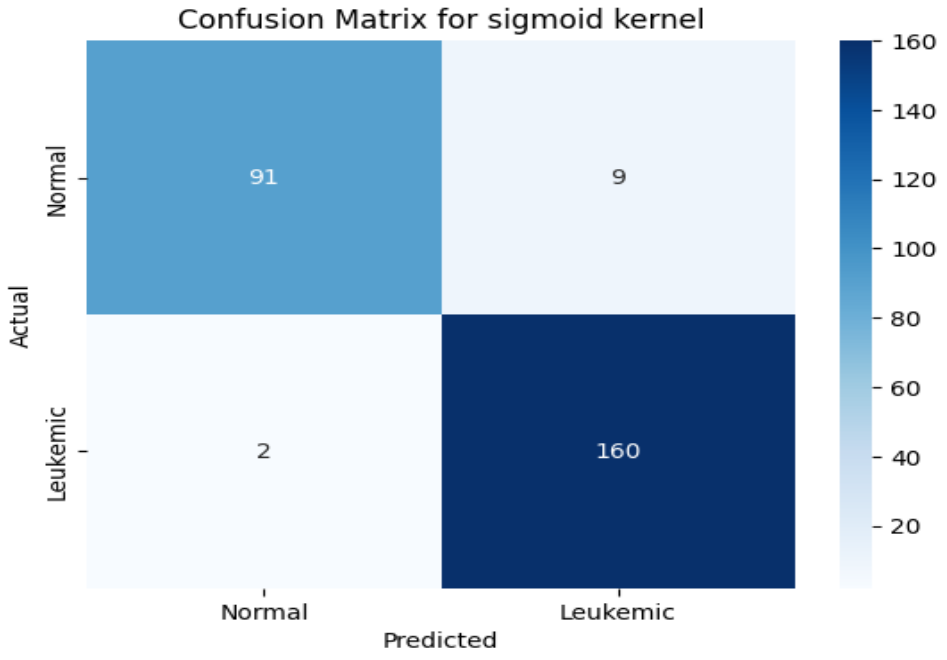


FIGURE 4.8: Confusion Matrix Sigmoid Kernal

4.5 Evaluation and Performance with Noise

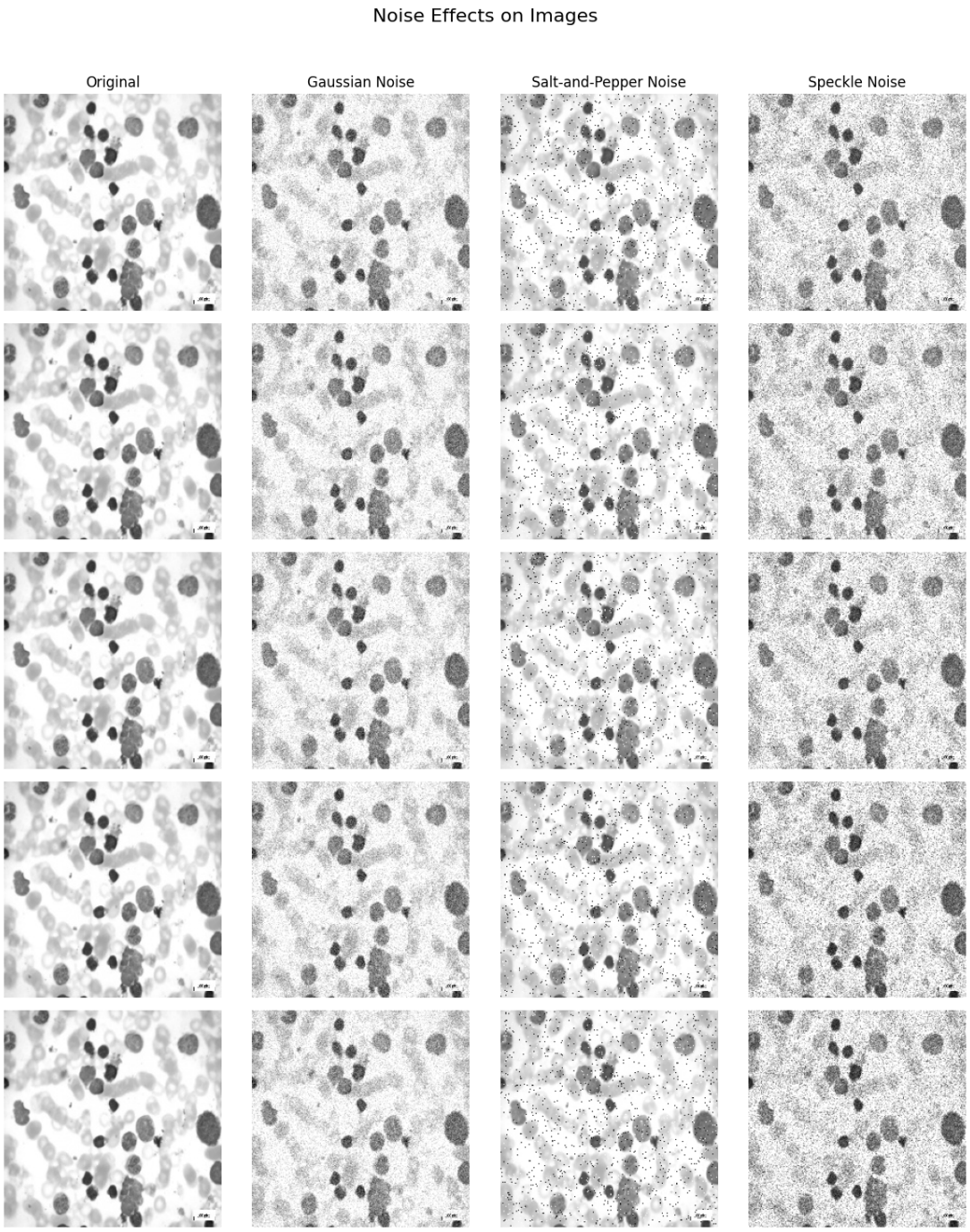


FIGURE 4.9: Noise effect On Blood Smear Images

Figure 4.9 illustrates the impact of various types of noise on blood smear images. The first column represents the original images, while the second, third, and fourth columns show images affected by Gaussian noise, salt-and-pepper noise,

and speckle noise, respectively. It can be observed that salt-and-pepper noise introduces sharp black-and-white pixel distortions, whereas Gaussian and speckle noise result in more uniform grainy patterns

4.5.1 Initial Evaluation (No Noise)

The initial evaluation of the SVM model was conducted on the original dataset without noise interference. The model achieved high accuracy and performed well across precision, recall, and F1 score metrics. The confusion matrix also demonstrated that the classifier made only a small number of misclassification [105].

- **Accuracy:** 96.18%
- **Precision:** 95.24%
- **Recall:** 98.77%
- **F1 Score:** 96.97%

The confusion matrix for the no-noise scenario is shown below:

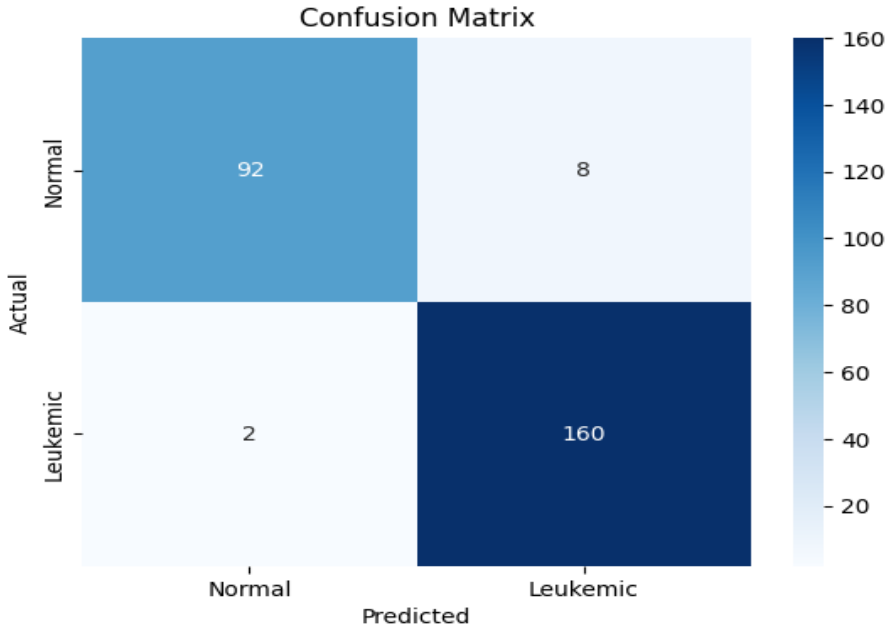


FIGURE 4.10: Confusion Matrix (No Noise)

4.5.2 Adding Gaussian Noise

Gaussian noise was added to simulate variability and noise in the image dataset[106]. This kind of noise is commonly found in medical images due to imaging equipment limitations or environmental factors. We experimented with different noise levels ranging from 0.01 to 0.3.

Interestingly, adding Gaussian noise did not significantly affect the SVM model's performance across all levels of noise. This robustness demonstrates the classifier's ability to handle smooth, continuous perturbations in image features.

- **Accuracy:** Remained at 96.18%
- **Precision:** Remained at 95.24%
- **Recall:** Remained at 98.77%
- **F1 Score:** Remained at 96.97%

The confusion matrix for Gaussian noise (at any noise level) was identical to the no-noise condition.

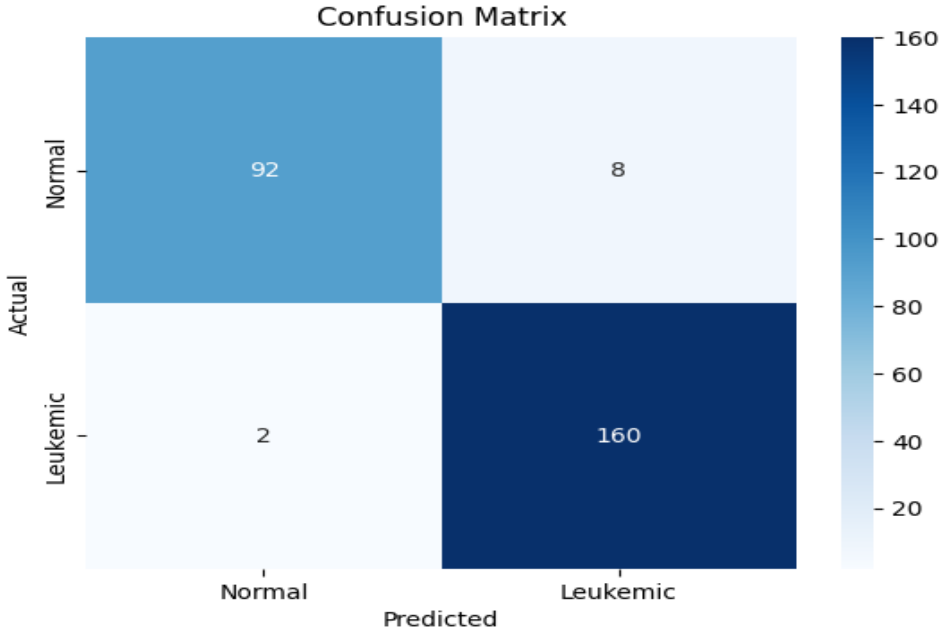


FIGURE 4.11: Confusion Matrix Gaussian Noise)

4.5.3 Adding Salt-and-Pepper Noise

Salt-and-pepper noise was added to mimic sharp, sparse noise, which can be caused by transmission errors or sensor malfunctions [107]. Unlike Gaussian noise, this form of noise had a more pronounced effect on the model's accuracy and other metrics as the noise level increased.

At a noise level of 0.01, the model's performance slightly declined:

- **Accuracy:** 94.27%
- **Precision:** 93.56%
- **Recall:** 98.76%
- **F1 Score:** 96.09%

Noise level: 0.01

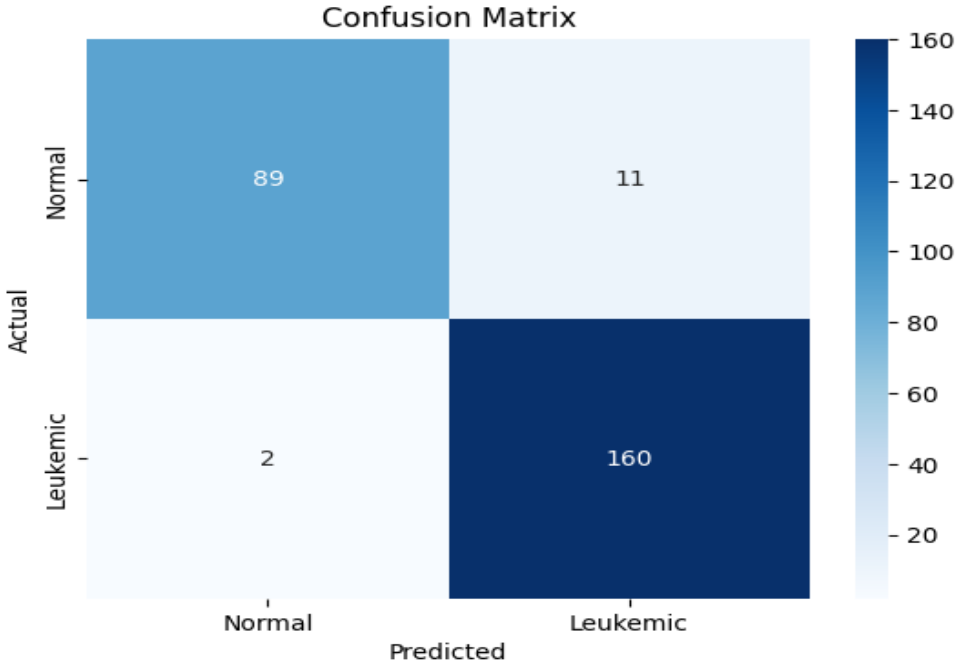


FIGURE 4.12: Confusion Matrix Salt-and-Pepper Noise)

With increasing noise levels, performance dropped further. At a noise level of 0.3, the model's accuracy declined significantly:

- **Accuracy:** 81.68%
- **Precision:** 77.51%
- **Recall:** 100%
- **F1 Score:** 87.33%

The confusion matrices illustrate the degradation in performance, with more mis classifications occurring at higher noise levels.

Noise level: 0.3

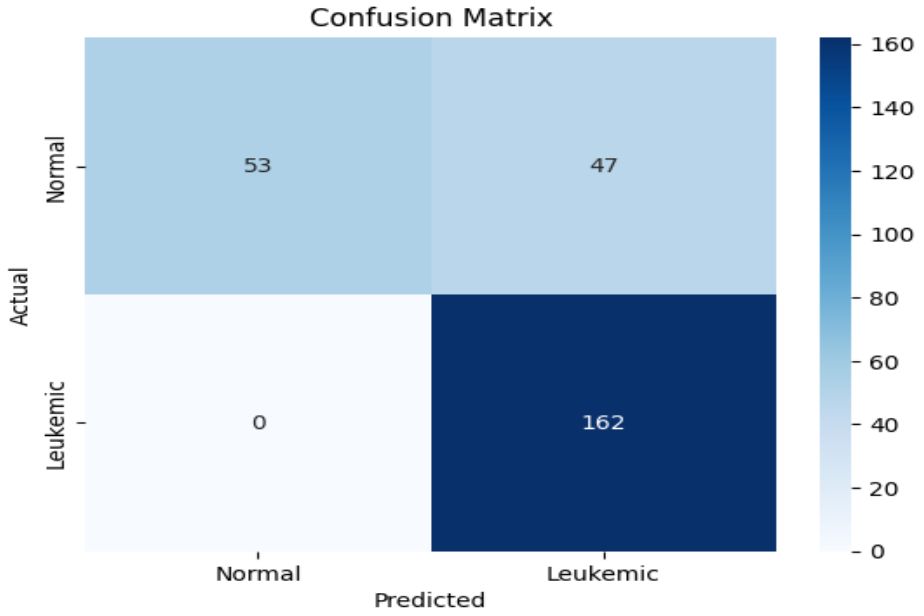


FIGURE 4.13: Confusion Matrix Salt-and-Pepper Noise)

4.5.4 Adding Speckle Noise

Speckle noise, typically found in medical ultrasound images, was also tested to evaluate the classifier's robustness to this type of noise [108]. This noise caused a significant drop in the model's performance, even at low noise levels.

At a noise level of 0.01, the performance metrics were as follows:

- **Accuracy:** 68.70%
- **Precision:** 66.39%
- **Recall:** 100%
- **F1 Score:** 79.80%

Noise level: 0.01

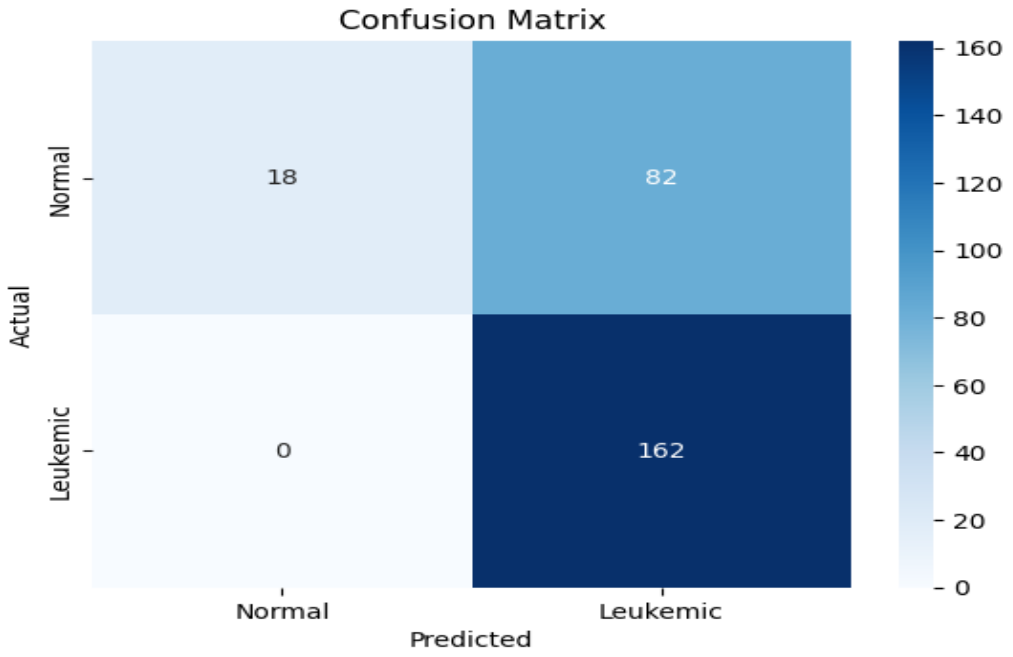


FIGURE 4.14: Confusion Matrix Speckle Noise)

As the noise level increased to 0.3, the model's performance further declined:

- **Accuracy:** 69.64%
- **Precision:** 66.94%
- **Recall:** 100%
- **F1 Score:** 80.19%

The confusion matrices for speckle noise reveal that the classifier's precision dropped, with an increase in false positives, leading to a reduction in overall performance metrics.

Noise level: 0.3

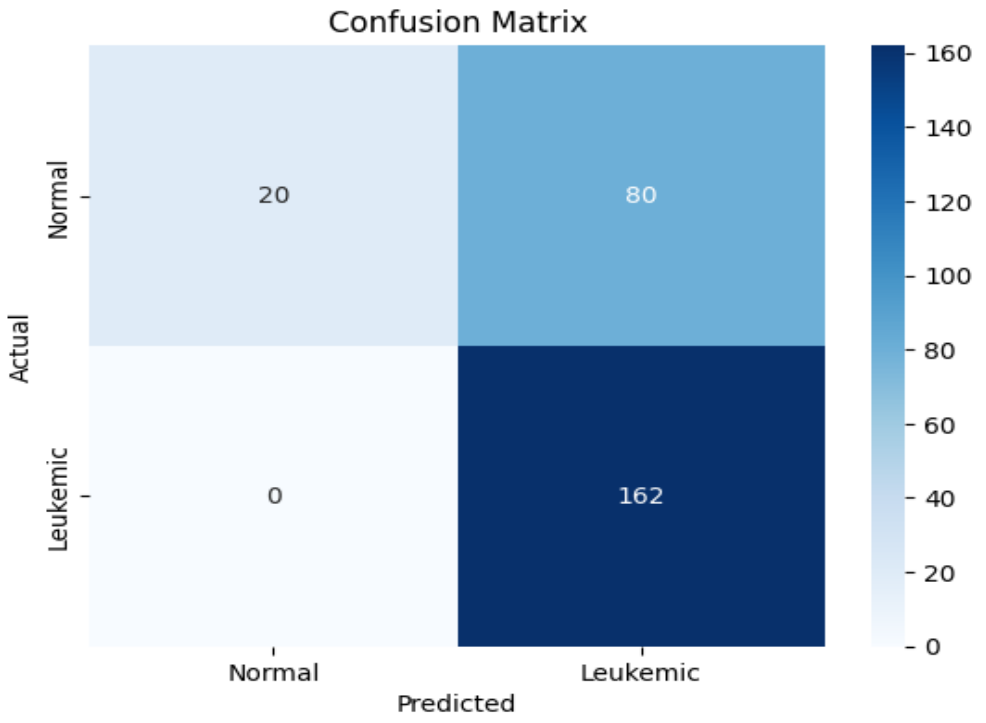


FIGURE 4.15: Confusion Matrix Speckle Noise)

The results indicate that the SVM classifier is highly effective for detecting ALL in blood smear images under normal conditions. The classifier maintains high performance under Gaussian noise, which suggests robustness to smooth, continuous noise present in imaging systems. However, salt-and-pepper and speckle noise, which introduce sharp perturbations, degraded the classifier's performance significantly as noise levels increased.

This behaviour can be explained by the nature of these noise types. Salt-and-pepper and speckle noise introduce extreme pixel values that obscure the subtle

patterns necessary for leukemia detection, which makes it challenging for the SVM model to differentiate between classes.

4.6 Analysis and Discussion

The presented overview explores image analysis and machine learning techniques applied to acute lymphoblastic leukaemia (ALL) detection, emphasizing their advantages over manual diagnostic methods. Various studies have employed distinct pre-processing, segmentation, feature extraction, and classification techniques. For instance, Karthikeya and Purnima utilized histogram equalization, median filtering, and Gabor texture extraction with an SVM classifier, achieving an accuracy of 90% [15]. Putzu and Ruberto improved accuracy to 92% by incorporating triangle thresholding and GLCM-based features with SVM [109]. Mohapatra et al. achieved 94.73% accuracy using contrast enhancement, selective median filtering, shadowed k-means clustering, and an ensemble classifier [110]. Mishra et al. attained 96.29% accuracy with histogram equalization, Wiener filtering, improved watershed segmentation, and a random forest classifier [111]. This analysis highlights the need for continued research to enhance leukaemia detection accuracy, particularly in classifying ALL subtypes, which has been somewhat overlooked in the existing literature despite its crucial role in precise medical treatment.

Chapter 5

Conclusion and Future Work

5.1 Introduction

In summary, the goal of this work was to improve the diagnosis of Acute Lymphoblastic Leukemia (ALL) by analyzing blood smear pictures utilizing a thorough approach that included segmentation, classification, and pre-processing techniques. Fuzzy C-Means (FCM) clustering and Support Vector Machine (SVM) classification showed encouraging outcomes in terms of increasing ALL detection efficiency and accuracy.

5.2 Preprocessing

The pre-processing stage played a crucial role in enhancing the quality of blood smear images by addressing issues such as noise, artifacts, and variations in illumination. Employing appropriate pre-processing techniques ensured that the subsequent stages of the methodology were conducted on clean and standardized images, contributing to the overall robustness of the system.

In addition, image enhancement techniques such as contrast adjustment and histogram equalization were applied to improve cell visibility. These steps helped in

preserving critical morphological features, ensuring more accurate segmentation and feature extraction.

5.3 Methodology

egmentation, an integral step in isolating relevant regions of interest within the blood smear images, facilitated the extraction of key features for subsequent analysis. The segmentation process aimed to accurately delineate normal and abnormal cell populations, providing a foundation for the subsequent classification steps.

The application of FCM clustering enabled the extraction of subtle patterns and relationships within the segmented regions, contributing to the characterization of different cell populations. FCM demonstrated its ability to handle uncertainties and partial membership, making it a valuable tool for analyzing complex and heterogeneous medical images like those associated with ALL.

The SVM classifier, known for its efficacy in handling high-dimensional data and achieving optimal separation between classes, was employed for the final classification of cells into normal and leukemic categories. The combination of FCM and SVM leveraged the strengths of both methods, enhancing the overall accuracy and reliability of the ALL-detection system.

5.4 Results

The experimental results obtained in this study demonstrated the effectiveness of the proposed methodology in accurately identifying and classifying Leukaemia cells in blood smear images. The integrated approach showcased improved performance compared to traditional methods, highlighting the potential for its application in clinical settings.

As we move forward, further refining the methodology and exploring additional features could enhance the system's robustness and sensitivity. This study's findings

contribute to the growing body of research aimed at leveraging advanced image processing and machine learning techniques for the early and accurate detection of Acute Lymphoblastic Leukaemia, ultimately improving patient outcomes through timely intervention.

Bibliography

[1] E. Özbay, F. A. Özbay, and F. S. Gharehchopogh, "Retracted article: Peripheral blood smear images classification for acute lymphoblastic leukemia diagnosis with an improved convolutional neural network," *Journal of Bionic Engineering*, vol. 10, no. 4, pp. 1–5, 2023, retracted. [Online]. Available: https://doi.org/10.1007/s42235-023-00441-y

[2] M. Bukhari, S. Yasmin, S. Sammad, and A. A. A. El-Latif, "A deep learning framework for leukaemia cancer detection in microscopic blood samples using squeeze and excitation learning," *Mathematical Problems in Engineering*, vol. 2022, p. 18, 2022. [Online]. Available: https://doi.org/10.1155/2022/2801227

[3] K. M. Hosny, T. Magdy, N. A. Lashin, K. Apostolidis, and G. A. Papakostas, "Refined color texture classification using cnn and local binary pattern," *Mathematical Problems in Engineering*, vol. 2021, p. 5567489, 2021. [Online]. Available: https://doi.org/10.1155/2021/5567489

[4] S. Perveen, A. Alourani, M. Shahbaz, M. U. Ashraf, and I. Hamid, "A framework for early detection of acute lymphoblastic leukemia and its subtypes from peripheral blood smear images using deep ensemble learning technique," *IEEE Access*, vol. 12, pp. 29 252–29 268, 2024.

[5] D. T. Johnson, J. Zhou, A. V. Kroll, R. H. Fang, M. Yan, C. Xiao, X. Chen, L. Zhang, and D. E. Zhang, "Acute myeloid leukemia cell membrane-coated nanoparticles for cancer vaccination immunotherapy," *Leukemia*, vol. 36, no. 4, pp. 994–1005, 2022.

[6] H. Inaba, M. Greaves, and C. G. Mullighan, "Acute lymphoblastic leukaemia," *Lancet*, vol. 381, no. 9881, pp. 1943–1955, 2013, epub 2013 Mar 22.

[7] P. Wang, M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A. Roshanpour, "Machine learning in detection and classification of leukaemia using smear blood images: A systematic review," *Scientific Programming*, vol. 2021, p. 9933481, 2021.

[8] S. Aljaboriy, N. Sjarif, and S. Chuprat, "Segmentation and detection of acute leukaemia using image processing and machine learning techniques: A review," *AUS*, vol. 26, no. 2, p. 511, 2019, september 15, 2019.

[9] F. E. Al-Tahhan, M. E. Fares, A. A. Sakr, and D. A. Aladle, "Accurate automatic detection of acute lymphatic leukaemia using a refined simple classification," *Microsc Res Tech*, vol. 83, no. 10, pp. 1178–1189, Oct 2020, epub 2020 Jun 4, PMID: 32497337.

[10] G. Atteia, R. Alnashwan, and M. Hassan, "Hybrid feature-learning-based pso-pca feature engineering approach for blood cancer classification," *Diagnostics*, vol. 13, no. 2672, 2023.

[11] A. T. Philip, S. Shifaana, S. Sunny, and P. Manimegalai, "Detection of acute lymphoblastic leukaemia in microscopic images using image processing techniques," *Journal of Physics: Conference Series*, vol. 1937, no. 1, p. 012022, 2021.

[12] C. Mondal, M. K. Hasan, M. Ahmad, M. A. Awal, M. T. Jawad, A. Dutta, M. R. Islam, and M. A. Moni, "Ensemble of convolutional neural networks to diagnose acute lymphoblastic leukaemia from microscopic images," *Informatics in Medicine Unlocked*, vol. 27, p. 100794, 2021.

[13] R. Wang, S. Shafique, and S. Tehsin, "Computer-aided diagnosis of acute lymphoblastic leukaemia," *Computational and Mathematical Methods in Medicine*, vol. 2018, p. 6125289, 2018.

[14] J. Zhao, M. Zhang, Z. Zhou, J. Chu, and F. Cao, "Automatic detection and classification of leukocytes using convolutional neural networks," *Medical & Biological Engineering & Computing*, vol. 55, no. 8, pp. 1287–1301, 2017.

[15] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, pp. 35–39, 2013.

[16] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, A. M. Umar, O. U. Linus, H. Arshad, A. A. Kazaure, U. Gana, and M. U. Kiru, "Comprehensive review of artificial neural network applications to pattern recognition," *IEEE Access*, vol. 7, pp. 158 820–158 846, 2019.

[17] J. Kufel, K. Bargieł-Łaczek, S. Kocot, M. Koźlik, W. Bartnikowska, M. Janik, Czogalik, P. Dudek, M. Magiera, A. Lis, I. Paszkiewicz, Z. Nawrat, M. Cebula, and K. Gruszczyńska, "What is machine learning, artificial neural networks and deep learning?—examples of practical applications in medicine," *Diagnostics*, vol. 13, no. 15, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/15/2582

[18] Y. Matsuda, S. Omachi, and H. Aso, "String detection from scene images by binarization and edge detection," *Institute of Electronics, Information and Communication Engineers, D*, vol. J93-D, no. 3, pp. 336–344, 2010.

[19] S. Mohapatra, D. Patra, and S. Satpath, "An ensemble classifier system for early diagnosis of acute lymphoblastic leukaemia in blood microscopic images," *Neural Computing and Applications*, pp. 1887–1904, 2014.

[20] L. H. Vogado, R. M. Veras, F. H. Araujo, R. R. Silva, and K. R. Aires, "Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 415–422, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197618301039

[21] N. Sampathila, K. Chadaga, N. Goswami, R. Chadaga, M. Pandya, S. Prabhu, M. Bairy, S. Katta, D. Bhat, and S. Upadya, "Customized deep

learning classifier for detection of acute lymphoblastic leukaemia using blood smear images," *Healthcare*, 2022.

[22] A. Saeed, S. Shoukat, K. Shehzad, I. Ahmad, A. Eshmawi, A. Amin, and E. Tag-Eldin, "A deep learning-based approach for the diagnosis of acute lymphoblastic leukaemia," *Electronics*, 2022.

[23] C. Mondal, M. Hasan, M. Ahmad, M. Awal, M. Jawad, A. Dutta, M. Islam, and M. Moni, "Ensemble of convolutional neural networks to diagnose acute lymphoblastic leukaemia from microscopic images," *Inform. Med. Unlocked*, 2021.

[24] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 1, 2004.

[25] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technology in cancer research & treatment*, vol. 17, p. 1533033818802789, 2018.

[26] T. Thanh, C. Vununu, S. Atoev, S.-H. Lee, and K.-R. Kwon, "Leukemia blood cell image classification using convolutional neural network," *International journal of computer theory and engineering*, vol. 10, no. 2, pp. 54–58, 2018.

[27] D. Kumar, N. Jain, A. Khurana, S. Mittal, S. C. Satapathy, R. Senkerik, and J. D. Hemanth, "Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks," *IEEE Access*, vol. 8, pp. 142 521–142 531, 2020.

[28] R. Saini, M. Dutta, and R. Kumar, "A comparative study of several image segmentation techniques," *Journal of Information and Operations Management*, vol. 3, no. 1, pp. 21–22, 2012.

[29] K. N. Plataniotis and A. N. Venetsanopoulos, *Color Image Processing and Applications*. Berlin, Heidelberg: Springer, 2000.

[30] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proceedings of the Asian Conference on Computer Vision*, 2010, pp. 770–783.

[31] J. Chen, H. Shao, and C. Hu, "Image segmentation based on mathematical morphological operator," in *Colorimetry and Image Processing*. Intech, 2018, [Internet]. [Online]. Available: http://dx.doi.org/10.5772/intechopen.72603

[32] U. Arab, P. Zhoena, B. Julien, H. A. Ndreu, and E. V. F. Skuka, "Effect of preprocessing on performance of neural networks for microscopy image classification," in *2020 International Conference on Computing, Electronics Communications Engineering (ICCECE)*, October 2020, pp. 162–165.

[33] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, pp. 1–20, 2019. [Online]. Available: https://doi.org/10.1109/TGRS.2019.2907932

[34] S. Marukatat, "Tutorial on pca and approximate pca and approximate kernel pca," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5445–5477, 2023.

[35] A. Sulaiman, S. Kaur, S. Gupta, H. Alshahrani, M. S. A. Reshan, S. Alyami, and A. Shaikh, "Resrandsvm: Hybrid approach for acute lymphocytic leukemia classification in blood smear images," *Diagnostics*, vol. 13, no. 12, p. 2121, 2023.

[36] G. M. Bai and P. Venkadesh, "Taylor–monarch butterfly optimization-based support vector machine for acute lymphoblastic leukemia classification with blood smear microscopic images," *Journal of Mechanics in Medicine and Biology*, vol. 21, no. 06, p. 2150041, 2021.

[37] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis, and F. Scotti, "Acute lymphoblastic leukemia detection based on adaptive unsharpening and deep learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1205–1209.

[38] J. Rawat, A. Singh, H. S. Bhadauria, J. Virmani, and J. S. Devgun, "Classification of acute lymphoblastic leukaemia using hybrid hierarchical classifiers," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 19 057–19 085, 2017.

[39] N. Pham, A. Morrison, J. Schwock *et al.*, "Quantitative image analysis of immunohistochemical stains using a cmyk color model," *Diagnostic Pathology*, pp. 2–8, 2007.

[40] P. K. Das, D. V. A, S. Meher, R. Panda, and A. Abraham, "A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia," *IEEE Access*, vol. 10, pp. 81 741–81 763, 2022.

[41] N. Senthilkumaran and R. Rajesh, "Edge detection techniques for image segmentation: A survey of soft computing approaches," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 250–254, 2009.

[42] P. P. Acharjya, R. Das, and D. Ghoshal, "A study on image edge detection using the gradients," *International Journal of Scientific and Research Publications*, vol. 2, no. 12, pp. 1–5, 2012.

[43] R. Yogamangalam and B. Karthikeyan, "Segmentation techniques comparison in image processing," *International Journal of Engineering and Technology*, vol. 5, no. 1, pp. 307–313, 2013.

[44] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[45] A. Muntasa and M. Yusuf, "Modeling of the acute lymphoblastic leukemia detection based on the principal object characteristics of the color image," *Procedia Computer Science*, vol. 157, pp. 87–98, 2019.

[46] S. Lakshmi and V. Sankaranarayanan, "A study of edge detection techniques for image segmentation computing approaches," *Special Issue on Computer Aided Soft Computing Techniques for Image and Biomedical Applications*, pp. 35–41, 2010.

[47] S. Wang, C. Fu, and Q. Li, "Text detection in natural scene image: A survey," in *Proceedings of the International Conference on Machine Learning and Intelligent Communications*, 2017, pp. 257–264.

[48] S. Ansari, A. H. Navin, A. B. Sangar, J. V. Gharamaleki, and S. Danishvar, "A customized efficient deep learning model for the diagnosis of acute leukemia cells based on lymphocyte and monocyte images," *Electronics*, vol. 12, no. 2, p. 322, 2023.

[49] S. Mishra, B. Majhi, and P. K. Sa, "Glrlm-based feature extraction for acute lymphoblastic leukemia (all) detection," in *Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Volume 2*. Springer, 2018, pp. 399–407.

[50] S. Mishra, B. Majhi, P. K. Sa, and L. Sharma, "Gray level co-occurrence matrix and random forest based acute lymphoblastic leukaemia detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 272–280, 2017.

[51] L. Putzu and C. D. Ruberto, "White blood cells identification and counting from microscopic blood images," in *Proceedings of the IWBBIO*, 2013, p. 363.

[52] Z. Moshavash, H. Danyali, and M. S. Helfroush, "An automatic and robust decision support system for accurate acute leukemia diagnosis from blood microscopic images," *Journal of digital imaging*, vol. 31, no. 5, pp. 702–717, 2018.

[53] N. Ahmed, A. Yigit, Z. Isik, and A. Alpkocak, "Identification of leukemia subtypes from microscopic images using convolutional neural network," *Diagnostics*, vol. 9, no. 3, p. 104, 2019.

[54] W. M. A. Dayata, S. Y. C. Yap, and C. D. Bandalan, "Automated blast cell detection for acute lymphoblastic leukemia using a stacking ensemble of convolutional neural networks," in *2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*. IEEE, 2023, pp. 95–102.

[55] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for diagnosis of acute lymphoblastic leukemia based on vit-cnn ensemble model," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 7529893, 2021.

[56] R. Baig, A. Rehman, A. Almuhaimeed, A. Alzahrani, and H. T. Rauf, "Detecting malignant leukemia cells using microscopic blood smear images: a deep learning approach," *Applied Sciences*, vol. 12, no. 13, p. 6317, 2022.

[57] P. K. Das, A. Pradhan, and S. Meher, "Detection of acute lymphoblastic leukemia using machine learning techniques," in *Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication: Proceedings of MDCWC 2020*. Springer, 2021, pp. 425–437.

[58] A. Yadav *et al.*, "An intelligent model for the detection of white blood cells using artificial intelligence," *Computer methods and programs in biomedicine*, vol. 199, p. 105893, 2021.

[59] D. Umamaheswari and S. Geetha, "Segmentation and classification of acute lymphoblastic leukemia cells tooled with digital image processing and ml techniques," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2018, pp. 1336–1341.

[60] H. Eivazi, M. Tahani, P. Schlatter, and R. Vinuesa, "Physics-informed neural networks for solving reynolds-averaged navier–stokes equations," *Physics of Fluids*, vol. 34, no. 7, 2022.

[61] S. Kumar, S. Mishra, P. Asthana, and Pragya, "Automated detection of acute leukemia using k-mean clustering algorithm," in *Advances in Computer and Computational Sciences: Proceedings of ICCCCS 2016, Volume 2*. Springer, 2018, pp. 655–670.

[62] K. A. Tareke and A. G. Awoke, "Hydrological drought forecasting and monitoring system development using artificial neural network (ann) in ethiopia," *Heliyon*, vol. 9, no. 2, 2023.

[63] T. Guillod, P. Papamanolis, and J. W. Kolar, "Artificial neural network (ann) based fast and accurate inductor modeling and design," *IEEE Open Journal of Power Electronics*, vol. 1, pp. 284–299, 2020.

[64] P. K. Das, P. Jadoun, and S. Meher, "Detection and classification of acute lymphocytic leukemia," in *2020 IEEE-HYDCON*. IEEE, 2020, pp. 1–5.

[65] M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A. Roshanpour, "Machine learning in detection and classification of leukemia using smear blood images: a systematic review," *Scientific Programming*, vol. 2021, no. 1, p. 9933481, 2021.

[66] A. S. Negm, O. A. Hassan, and A. H. Kandil, "A decision support system for acute leukaemia classification based on digital microscopic images," *Alexandria engineering journal*, vol. 57, no. 4, pp. 2319–2332, 2018.

[67] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, p. 114161, 2021.

[68] A. Bodzas, "Diagnosis of malignant haematopoietic diseases based on the automation of blood microscopic image analysis," Master's thesis, Technical University of Ostrava, Ostrava, CZ, 2019, master's thesis. [Online]. Available: https://scholar.google.com

[69] J. Zhao, M. Zhang, Z. Zhou, J. Chu, and F. Cao, "Automatic detection and classification of leukocytes using convolutional neural networks," *Medical & Biological Engineering & Computing*, vol. 55, no. 8, pp. 1287–1301, 2017.

[70] S. Mishra, B. Majhi, and P. K. Sa, "Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection," *Biomedical Signal Processing and Control*, vol. 47, pp. 303–311, 2019.

[71] S. Afshar, F. Abdolrahmani, T. F. VAKILI, S. M. ZOHDI, and K. Taheri, "Recognition and prediction of leukemia with artificial neural network (ann)," 2011.

[72] A. Shah, M. Shah, A. Pandya, R. Sushra, R. Sushra, M. Mehta, K. Patel, and K. Patel, "A comprehensive study on skin cancer detection using artificial neural network (ann) and convolutional neural network (cnn)," *Clinical eHealth*, 2023.

[73] V. Acharya and P. Kumar, "Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms," *Medical & biological engineering & computing*, vol. 57, pp. 1783–1811, 2019.

[74] Y. Gao, S. Ma, J. Liu, Y. Liu, and X. Zhang, "Fusion of medical images based on salient features extraction by pso optimized fuzzy logic in nsst domain," *Biomedical Signal Processing and Control*, vol. 69, p. 102852, 2021.

[75] M. G. Lanjewar, J. S. Parab, and A. Y. Shaikh, "Development of framework by combining cnn with knn to detect alzheimer's disease using mri images," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 12 699–12 717, 2023.

[76] L. Boldú, A. Merino, S. Alférez, A. Molina, A. Acevedo, and J. Rodellar, "Automatic recognition of different types of acute leukaemia in peripheral blood by image analysis," *Vol. 72, No. 11*, pp. 755–761, 2019.

[77] S. Zhang and J. Li, "Knn classification with one-step computation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2711–2723, 2021.

[78] A. S. Negm, O. A. Hassan, and A. H. Kandil, "A decision support system for acute leukaemia classification based on digital microscopic images," *Alexandria engineering journal*, vol. 57, no. 4, pp. 2319–2332, 2018.

[79] L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with k-nearest-neighbors (knn) algorithm," *Building and Environment*, vol. 202, p. 108026, 2021.

[80] P. K. Das, A. Pradhan, and S. Meher, "Detection of acute lymphoblastic leukemia using machine learning techniques," in *Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication: Proceedings of MDCWC 2020*. Springer, 2021, pp. 425–437.

[81] B. Mahaboob, J. P. Praveen, B. A. Rao, Y. Haranadh, C. Narayana, and G. B. Prakash, "A study on multiple linear regression using matrix calculus," *Advancecs in Mathematics Scientifc journal*, vol. 9, no. 7, pp. 1–10, 2020.

[82] R. Efendi, N. M. Nawi, M. M. Deris, S. A. Burney *et al.*, "Cleansing of inconsistent sample in linear regression model based on rough sets theory," *Systems and Soft Computing*, vol. 5, p. 200046, 2023.

[83] M. Aboal-Somoza and R. M. Crujeiras, "Misuse of linear regression technique in analytical chemistry?" *Journal of Chemical Education*, vol. 101, no. 3, pp. 1062–1070, 2024.

[84] P. Viswanathan, "Fuzzy c means detection of leukaemia based on morphological contour segmentation," *Procedia Computer Science*, vol. 58, pp. 84–90, 2015.

[85] A. Bodzas, P. Kodytek, and J. Zidek, "Automated detection of acute lymphoblastic leukaemia from microscopic images based on human visual perception," *Frontiers in Bioengineering and Biotechnology*, vol. 8, pp. 1005–1013, 2020.

[86] L. Möckl, A. Roy, and W. Moerner, "Deep learning in single-molecule microscopy: Fundamentals, caveats, and recent developments [invited]," *Biomed. Opt. Express*, 2020.

[87] R. A. Hazarika, A. K. Maji, S. N. Sur, B. S. Paul, and D. Kandar, "A survey on classification algorithms of brain images in alzheimer's disease based on feature extraction techniques," *IEEE Access*, vol. 9, pp. 58 503–58 536, 2021.

[88] C. Huerga, A. Morcillo, L. Alejo, A. Marín, A. Obesso, D. Travaglio, J. Bayón, D. Rodriguez, and M. Coronado, "Role of correlated noise in textural features extraction," *Physica Medica*, vol. 91, pp. 87–98, 2021.

[89] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical Image Analysis*, vol. 43, pp. 98–111, 2018.

[90] A. Al-Ghraibah, M. Altayeb, and F. A. Alnaimat, "An automated system to distinguish between corona and viral pneumonia chest diseases based on image processing techniques," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 7, p. 2261575, 2024.

[91] A. Al-Ghraibah, M. Al-Ayyad, and H. Elkhalil, "Tibia fracture detection using a modified edge detection method based on bone length," *Journal of Engineering Science and Technology*, vol. 15, no. 1, pp. 249–260, 2020.

[92] A. Bodzas, P. Kodytek, and J. Zidek, "Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 1005, 2020.

[93] P. G. F. Desai and G. Shet, "Detection of leukemia using image processing," *International Journal of Advance Research in Science and Engineering (IJARSE)*, vol. 7, no. 3, pp. 149–156, 2018.

[94] M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A. Roshanpour, "Machine learning in detection and classification of leukemia

using smear blood images: a systematic review," *Scientific Programming*, vol. 2021, no. 1, p. 9933481, 2021.

[95] P. Jagadev and H. Virani, "Detection of leukemia and its types using image processing and machine learning," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. IEEE, 2017, pp. 522–526.

[96] N. Z. Supardi, M. Y. Mashor, N. H. Harun, F. A. Bakri, and R. Hassan, "Classification of blasts in acute leukemia blood samples using k-nearest neighbour," in *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, 2012, pp. 461–465.

[97] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: A survey," *International Journal of Document Analysis and Recognition*, vol. 7, no. 2-3, pp. 84–104, 2005.

[98] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis, and F. Scotti, "Histopathological transfer learning for acute lymphoblastic leukemia detection," in *2021 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. IEEE, 2021, pp. 1–6.

[99] S. N. M. Safuan, M. R. M. Tomari, W. N. W. Zakaria, M. N. H. Mohd, and N. S. Suriani, "Investigation of white blood cell biomaker model for acute lymphoblastic leukemia detection based on convolutional neural network," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 2, pp. 611–618, 2020.

[100] P. K. Das, S. Meher, R. Panda, and A. Abraham, "An efficient blood-cell segmentation for the detection of hematological disorders," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 615–10 626, 2021.

[101] G. Vieira and M. E. Valle, "Acute lymphoblastic leukemia detection using hypercomplex-valued convolutional neural networks," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[102] R. Khandekar, P. Shastry, S. Jaishankar, O. Faust, and N. Sampathila, "Automated blast cell detection for acute lymphoblastic leukemia diagnosis," *Biomedical Signal Processing and Control*, vol. 68, p. 102690, 2021.

[103] J. R. Beattie and F. W. Esmonde-White, "Exploration of principal component analysis: deriving principal component analysis visually using spectra," *Applied Spectroscopy*, vol. 75, no. 4, pp. 361–375, 2021.

[104] C. Marzban, U. Yurtsever, and M. Richman, "Principal component analysis for equation discovery," *arXiv preprint arXiv:2401.04797*, 2024.

[105] B. Goyal, S. Agrawal, and B. Sohi, "Noise issues prevailing in various types of medical images," *Biomedical & Pharmacology Journal*, vol. 11, no. 3, p. 1227, 2018.

[106] K. Chithra and T. Santhanam, "Hybrid denoising technique for suppressing gaussian noise in medical images," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, pp. 1460–1463.

[107] J. Azzeh, B. Zahran, and Z. Alqadi, "Salt and pepper noise: Effects and removal," *JOIV: International Journal on Informatics Visualization*, vol. 2, no. 4, pp. 252–256, 2018.

[108] A. Maity, A. Pattanaik, S. Sagnika, and S. Pani, "A comparative study on approaches to speckle noise reduction in images," in *2015 International Conference on Computational Intelligence and Networks*, 2015, pp. 148–155.

[109] S. Mohapatra, S. S. Samanta, D. Patra, and S. Satpathi, "Fuzzy based blood image segmentation for automated leukaemia detection," in *Proceedings of the 2011 International Conference on Devices and Communications, ICDeCom 2011*, India, February 2011.

[110] J. Laosai and K. Chamnongthai, "Acute leukaemia classification by using svm and k-means clustering," in *Proceedings of the 2014 International Electrical Engineering Congress, iEECON 2014*, Thailand, March 2014.

[111] D. R. Karthikeyan and N. Poornima, "Microscopic image segmentation using fuzzy c means for leukaemia diagnosis," *International Journal of Advanced Research in Science, Engineering and Technology*, vol. 4, no. 1, 2017.