UNIVERSITÀ POLITECNICA
DELLE MARCHE
FACULTY OF ENGINEERING

MASTER OF SCIENCE IN BIOMEDICAL ENGINEERING

Automated Detection of Acute Lymphoblastic Leukemia from Microscopic Images

Advisor:
Prof. Laura Burattini

Co-advisors:
Dr. Agnese Sbrollini
Dr. Selene Tomassini

Candidate:
Ruba Sharaan

Y. 2021/2022

Thanks and gratitude…

All thanks and appreciation to ***Professor Laura Burattini, Dr. Agnese Sbrollini, Dr. Selene Tomassin*** *and all professors in Università Politecnica Delle Marche for their great support and precious effort.*

- *To the man who gave me life and happiness, who supported me in all stages of my life and who encouraged me to love and seek knowledge, you are forever in my heart...*

  My Dad

- *To the gentle, tender and loving lady, to you who give me strength through your blessed prayers....*

  My mom

- *To my hero, my backbone, my best partner in life and the father of our children in the future...*

  My love Ahmed

- *To my family in Germany who support me like no one else has, there are no words to explain how grateful I am.*

  Gazzeh's family

- *To my second family, you have all my love.*

  Abu Laila's family

- *To the two beautiful partners, the spoiled Superman and the little princess....*

  Mohamad, Hiba, Jafar and Hadeel

- *To my sweet, my soul mate, my heart engineer...*

  My sister Haya

- *To my super hero, the kindest person on earth...*

  My brother Jihad

- *To the smart engineer and my kind sister...*

  Zahra

- *To my friends in Italy, partners in projects and success...*

  Ayham, Sameh

- *To my cute roommates, the days with you were better...* ***Chiara A, P, T.***

CONTENTS

List of Figures

List of Tables

Summary

In acute lymphoblastic leukemia (ALL), the malignant clone arises from hematopoietic progenitors in the bone marrow or lymphatic system resulting in an increase of immature non-functioning leukemic cells and could be detected through screening of blood and bone marrow smears by pathologists. Microscopic image analysis plays a significant role in initial leukemia screening and its efficient diagnostics. Since the present conventional methodologies partly rely on manual examination, which is time consuming and depends greatly on the experience of domain experts, automated leukemia detection opens up new possibilities to minimize human intervention and provide more accurate clinical information. This thesis presented a method for distinguishing between normal and ALL cells based on an artificial neural network. The focus of the research was to process blood images with a simple strategy and take advantage of them by identify and show the differences in the properties of them and obtaining logical numerical results after applying 12 morphological and statistical features and then finding the best six features, the most important of which was the solidity and form factor which can strongly participate in the classification and recognition of the two cases. The processing and classification algorithm was applied to a set of 215 blood cells which confirmed the promising results of the proposed method. The overall accuracy of the network was 81.4%. This thesis describes the algorithm used in the segmentation and identification of leukocytes using simple image processing techniques, and then the classification of ALL using ANN based on morphological and statistical features extracted from the data.

Chapter One

Blood anatomy and functioning

1.1 Introduction

Blood is classified as a connective tissue because it has both fluid and solid (cellular) components. The fluid is plasma, in which plasma proteins and cells (red blood cells, white blood cells, and platelets) are suspended in the watery base [1].

The functions of blood can be divided into three general categories: transportation, regulation, and protection.

1. Transportation includes the movement of gases (oxygen and carbon dioxide), nutrients, and metabolic wastes.
2. Regulation includes stabilizing the body's temperature, blood pH, and fluid volume and pressure of the blood.
3. Protection includes fighting infections and protecting against blood loss.

1.2 Blood components

It has four main components: plasma, red blood cells, white blood cells, and platelets (**Figure 1.1**) [2].

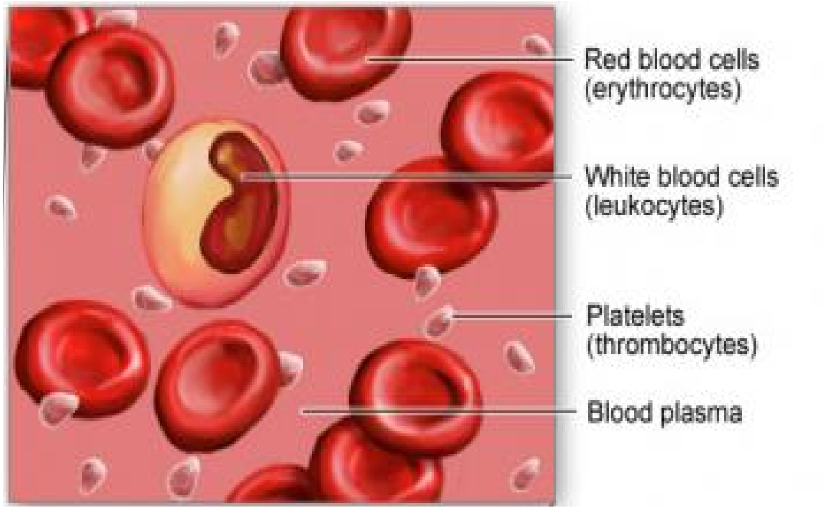


Figure 1.1. Blood components.

Plasma is the pale-yellow liquid part of blood. It accounts for 46 to 63 % of total blood volume, with an average of 55 %. It is mostly water (95%) with a number of dissolved substances that add to its viscosity. The majority (92 %) of the dissolved solutes are plasma proteins. Nonprotein components include metabolic waste products, nutrients, ions, and dissolved gases.

Red blood cells (Erythrocytes or RBCs) are the most specialized cells in the human body. They are a biconcave (donut) shape with a thin central disc. This shape is important because the disc increases the surface-area-to-volume ratio for faster exchange of gases and it allows red blood cells to stack, one on another, as they flow through very narrow vessels. Also, since some capillaries are as narrow, red blood cells can literally squeeze through narrow vessels by changing shape, Erythropoiesis occurs in the red bone marrow located in the vertebrae, sternum, ribs, skull, scapulae, pelvis, and proximal limb bones. Red blood cells begin as large, immature cells (proerythroblasts), and over a seven-day period they change into a much smaller, mature, red blood cell that then enters the blood stream. The red blood cells survive on average only 120 days.

Leukocytes (also called white blood cells or WBCs) are a cellular component of the blood that lacks hemoglobin has a nucleus and is capable of motility. Their main function is immunological defense. They defend the body against infection and disease by: ingesting foreign materials and cellular debris, by destroying infectious agents and cancer cells, or by producing antibodies. WBCs are produced by bone marrow and their levels of production are regulated by organs such as the spleen, liver, and kidneys. The total number of white blood cells is 4,500 to 10,000 per cubic millimeter of blood.

Unlike red and white blood cells, platelets (Thrombocytes) are not actually cells but rather small fragments of cells. They are circular-shaped cytoplasmic substances or bodies that do not have a nucleus that are formed in the bones, found in the blood. Platelets help the blood clotting process (or coagulation) by gathering at the site of an injury, sticking to the lining of the injured blood vessel, and forming a platform on which blood coagulation can occur. This results in the formation of a fibrin clot, which covers the wound and prevents blood from leaking out. Fibrin also forms the initial scaffolding upon which new tissue forms, thus promoting healing. A higher-than-normal number of platelets can cause unnecessary clotting, which can lead to strokes and heart attacks.

1.3 Physical properties of blood

Blood is a type of connective tissue that can be described based on its physical characteristics including color, volume, viscosity, plasma concentration, temperature, and pH (**Table 1.1**) [3-4].

- **Color:** The color of blood depends upon whether it is oxygen-rich or oxygen-poor. Oxygen-rich blood is bright red or almost scarlet. Contrary to popular belief, oxygen-poor blood is not blue; rather, oxygen-poor blood is dark red. appears blue in superficial veins due to the way blue light reflected back to the eye.
- **Volume:** The blood volume is about 5 liters in an adult and males with slightly more blood than females. Normal blood volume essential for maintaining blood pressure.
- **Viscosity:** The viscosity depends on the number of dissolved substances in the blood. Blood is 4 to 5 times more viscous than water (thicker). For example, it increases if red blood cells are increased and also increases if amount of fluid decreased.
- **Plasma Concentration:** The relative concentration of solutes in plasma determines whether fluids move into or out of plasma by osmosis. For example, during dehydration plasma hypertonic, the fluid moves into plasma from surrounding tissues.
- **Temperature:** In general, normal blood temperature is about the same as normal body temperature, but 1 degree (or 1 degree Celsius) above the measured body temperature heats the area through which blood travels. 38°C (100.4°F).
- **Blood PH:** Plasma slightly alkaline (or slightly basic) pH between 7.35 and 7.45 and the normal PH is 7.

Table 1.1. Physical Characteristics of Blood [4].

Characteristics	**Normal Values**
Color	Scarlet (oxygen-rich) to dark red (oxygen-poor)
Volume	4-5 L (females) 5-6 L (males)
Viscosity (relative to water)	4.5-5.5 × (whole blood)
Temperature	38°C (100.4°F)
PH	7.35-7.45

1.4 Leukocyte

A type of blood cell that is made in the bone marrow and found in the blood and lymph tissue. Leukocytes are part of the body's immune system. They help the body fight infection and other diseases. Types of leukocytes are granulocytes (neutrophils, eosinophils, and basophils), monocytes, and lymphocytes (T cells and B cells). Checking the number of leukocytes in the blood is usually part of a complete blood cell (CBC) test. It may be used to look for conditions such as infection, inflammation, allergies, and leukemia. Also called WBC and white blood cell [1]. The (**Figure 1.2**) shows the development of blood cells. A blood stem cell goes through several steps to become a red blood cell, platelet, or white blood cell [5].

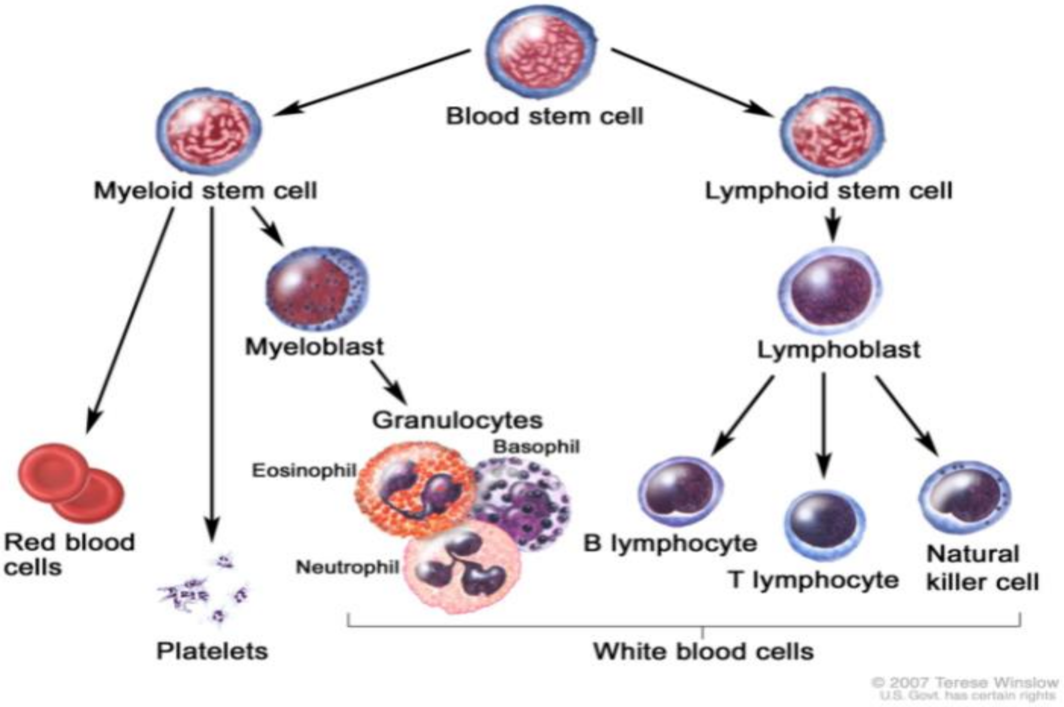


Figure 1.2. Blood cell development. A blood stem cell goes through several steps to become a red blood cell, platelet, or white blood cell [5].

1.4.1 Types of white blood cells

There are several different types of white blood cells, each with varying responsibilities [5]:

- **Lymphocytes:** These are vital for producing antibodies that help the body to defend itself against bacteria, viruses, and other threats.
- **Neutrophils:** These are powerful white blood cells that destroy bacteria and fungi. These cells have a single nucleus with multiple lobes. Neutrophils are the most abundant white blood cell in circulation. They are chemically drawn to bacteria (by cytokines) and migrate through tissue toward infection sites. Neutrophils are phagocytic (i.e., they engulf and

destroy target cells). When released, their granules act as lysosomes to digest cellular macromolecules, destroying the neutrophil in the process.

- **Basophils:** These alert the body to infections by secreting chemicals into the bloodstream, mostly to combat allergies. They are the least numerous type of white blood cells. They have a multi-lobed nucleus, and their granules contain immune-boosting compounds such as histamine and heparin. Heparin thins the blood and inhibits blood clot formation while histamine dilates blood vessels to increase blood flow and the permeability of capillaries so that leukocytes may be transported to infected areas.
- **Eosinophils:** These are responsible for destroying parasites and cancer cells, and they are part of an allergic response. The nucleus of these cells is double-lobed and appears U-shaped in blood smears. Eosinophils are usually found in connective tissues of the stomach and intestines. These are also phagocytic and primarily target antigen-antibody complexes formed when antibodies bind to antigens to signal that they should be destroyed. Eosinophils are most active during parasitic infections and allergic reactions.
- **Monocytes:** These cells are the greatest in size of the white blood cells. They have a large, single nucleus that comes in a variety of shapes but is most often kidney-shaped. Monocytes migrate from blood to tissue and develop into either macrophages or dendritic cells.

1.4.2 Problems affecting white blood cells

A number of diseases and conditions may affect white blood cell levels [6]:

1. **Weak immune system:** This is often caused by illnesses such as HIV/AIDS or by cancer treatment. Cancer treatments such as chemotherapy or radiation therapy can destroy white blood cells and leave you at risk for infection.
2. **Infection:** A higher-than-normal white blood cell count usually means you have some type of infection. White blood cells are multiplying to destroy the bacteria or virus.
3. **Myelodysplastic syndrome:** This condition causes abnormal production of blood cells. This includes white blood cells in the bone marrow.
4. **Cancer of the blood:** Cancers including leukemia and lymphoma can cause uncontrolled growth of an abnormal type of blood cell in the bone marrow. This results in a greatly increased risk for infection or serious bleeding.
5. **Myeloproliferative disorder:** This disorder refers to various conditions that trigger the excessive production of immature blood cells. This can result in an unhealthy balance of

all types of blood cells in the bone marrow and too many or too few white blood cells in the blood.

6. **Medicines:** Some medicines can raise or lower the body's white blood cell count.

Chapter Two

Leukemia

2.1 Introduction

The term leukemia comes from the Greek leukos, which means “white,” and haima, which means “blood,” and it refers to a variety of cancers that affect blood cells. Leukemia is a cancer of the blood that is characterized by an overproduction of abnormal white blood cells (or leukocytes), which flood the bloodstream, crowd out healthy cells, and prevent normal cell death from happening.

In general, leukemia is thought to occur when some blood cells acquire changes (mutations) in their genetic material or DNA. A cell's DNA contains the instructions that tell a cell what to do. Normally, the DNA tells the cell to grow at a set rate and to die at a set time. In leukemia, the mutations tell the blood cells to continue growing and dividing.

When this happens, blood cell production becomes out of control. Over time, these abnormal cells can crowd out healthy blood cells in the bone marrow, leading to fewer healthy white blood cells, red blood cells and platelets (**Figure 2.1**), causing the signs and symptoms of leukemia [7].

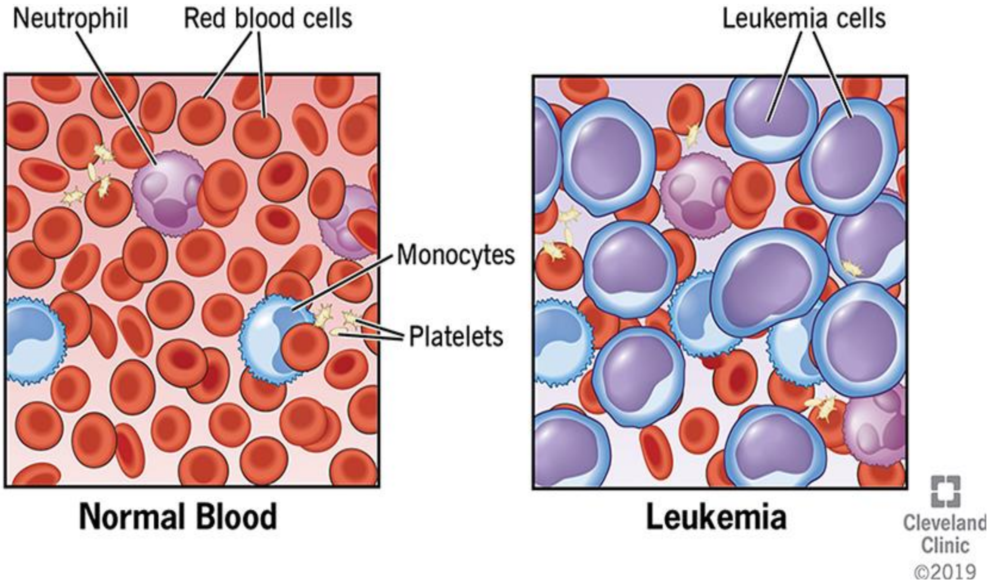


Figure 2.1. Normal blood contains red blood cells, white blood cells and platelets. Leukemia cells outnumber normal cells in leukemia [8].

2.2 Leukemia classification

Classification of leukemia based on its speed of progression and the type of cells involved [9].

The first type of classification is by how fast the leukemia progresses:

Acute leukemia: In acute leukemia, the abnormal blood cells are immature blood cells (blasts). They can't carry out their normal functions, and they multiply rapidly, so the disease worsens quickly. Acute leukemia requires aggressive, timely treatment.

Chronic leukemia: There are many types of chronic leukemias. Some produce too many cells and some cause too few cells to be produced. Chronic leukemia involves more-mature blood cells. These blood cells replicate or accumulate more slowly and can function normally for a period of time. Some forms of chronic leukemia initially produce no early symptoms and can go unnoticed or undiagnosed for years.

The second type of classification is by type of white blood cell affected:

Lymphocytic leukemia: This type of leukemia affects the lymphoid cells (lymphocytes), which form lymphoid or lymphatic tissue. Lymphatic tissue makes up the immune system.

Myelogenous leukemia: This type of leukemia affects the myeloid cells. Myeloid cells give rise to red blood cells, white blood cells and platelet-producing cells.

2.3 Leukemia types

The major types of leukemia are: (**Figure 2.2**) [10].

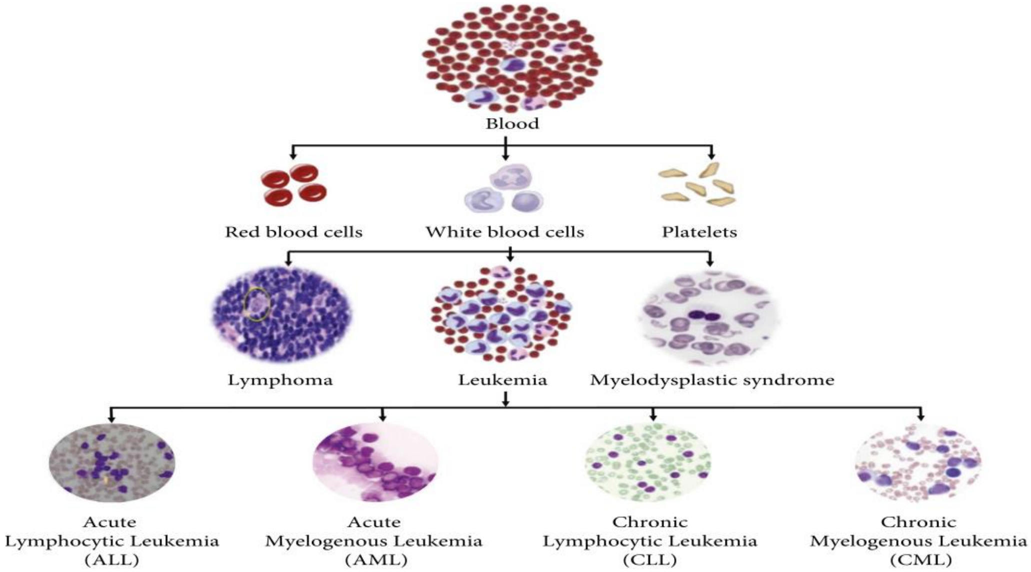

Blood
Red blood cells
White blood cells
Platelets
Lymphoma
Leukemia
Myelodysplastic syndrome
Acute Lymphocytic Leukemia (ALL)
Acute Myelogenous Leukemia (AML)
Chronic Lymphocytic Leukemia (CLL)
Chronic Myelogenous Leukemia (CML)

Figure 2.2. Blood and leukemia types [11].

2.3.1 Acute lymphocytic leukemia

In acute lymphoblastic leukemia (ALL), the malignant clone arises from hematopoietic progenitors in the bone marrow or lymphatic system resulting in an increase of immature non-functioning leukemic cells. Infiltration of bone marrow leads to anemia, granulocytopenia, and thrombocytopenia with the clinical manifestations of fatigue, weakness, infection, and hemorrhages. These symptoms are more often the reason a patient first seeks medical advice rather than consequences of tumor bulk, such as lymph node enlargement, hepatosplenomegaly caused by leukemic infiltration, or symptoms of the central nervous system (meningeosis leukemica).

ALL is the most frequent neoplastic disease in children with an early peak at the age of 3–4 years. The incidence in adult's ranges from 0.7 to 1.8/100,000 per year, being somewhat higher in younger adults (1–1.5 for the age group 15–24 years) and decreasing thereafter, only to increase again in elderly people to 2.3 for age >65 years. The frequency of immunological, cytogenetic, and genetic subtypes changes with age.

2.3.2 Acute myelogenous leukemia

Acute myeloid leukemia (AML) is a neoplasm characterized by infiltration of the blood, bone marrow, and other tissues by proliferative, clonal, poorly differentiated cells of the hematopoietic system. These leukemias comprise a spectrum of malignancies that, untreated, are uniformly fatal. In 2016, the estimated number of new AML cases in the United States was 19,950, comprising ~1.2% of all cancer cases. AML is the most common acute leukemia in older patients, with a median age at diagnosis of 67 years. Long-term survival is infrequent; U.S. registry data report that only 27% of patients survive 5 years.

Most cases of AML are idiopathic. Genetic predisposition, radiation, chemical/other occupational exposures, and drugs have been implicated in the development of AML, but AML cases with established etiology are relatively rare. No direct evidence suggests a viral etiology. Genome sequencing studies suggest that most cases of AML arise from a limited number of mutations that accumulate with advancing age. Indeed, genome sequencing is providing paradigm-shifting advances in our understanding of leukemogenesis. The Cancer Genome Atlas (TCGA) and other databases demonstrate that blood cells from up to 5–6% of normal individuals aged >70 years contain potentially "premalignant" mutations that are associated with clonal expansion. The additional insults that subsequently direct "premalignant" blood cells to leukemia are quite heterogeneous and still poorly understood.

2.3.3 Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) is a monoclonal proliferation of mature B lymphocytes defined by an absolute number of malignant cells in the blood (5×10^9/mL). The presence of malignant B cells under this count in the blood without nodal, spleen, or liver involvement and absent cytopenias is a precursor of this disease called monoclonal B-cell lymphocytosis (MBL) with ~1–2% chance per year of progressing to overt CLL. CLL is a heterogeneous disease in terms of natural history, with some patients presenting asymptomatically and never requiring therapy, whereas others present with symptomatic disease, require multiple lines of therapy, and eventually die of their disease.

CLL is primarily a disease of older adults, with a median age at diagnosis of 71 and an age-adjusted incidence of 4.5/100,000 people in the United States. The prevalence of CLL has increased over the past decades due to improvements in therapy for this disease and also survival of older patients from other medical ailments. In 1980, the 5-year overall survival of patients was 69%, and this increased to 87.9% in 2007 and is likely even higher today. The male: female ratio is 2:1; however, as patients age, the ratio becomes more even, and over the age of 80, the incidence is equal between men and women.

2.3.4 Chronic myelogenous leukemia

Chronic myeloid leukemia (CML) is a clonal hematopoietic stem cell disorder. The disease is driven by the *BCR-ABL1* chimeric gene product, that codes for a constitutively active tyrosine kinase, resulting from a reciprocal balanced translocation between the long arms of chromosomes 9 and 22, t (9;22) (q34.1; q11.2). Untreated, the course of CML is typically biphasic or triphasic, with an early indolent or chronic phase, followed often by an accelerated phase and a terminal blastic phase.

CML accounts for ∼15% of all cases of leukemia. There is a slight male preponderance (male: female ratio 1.6:1). The median age at diagnosis is 55–65 years. It is uncommon in children; only 3% of patients with CML are younger than 20 years although in recent years a higher proportion of young patients seem to be diagnosed. CML incidence increases slowly with age, with a steeper increase after the age of 40–50 years. By extrapolation, the worldwide annual incidence of CML is about 100,000–120,000 cases.

Chapter Three

Technical background

3.1 Computer-aided diagnosis

Computer-aided diagnosis (CAD) is a common tool for the detection of diseases, particularly different types of cancers, based on medical images. CAD systems based on image processing become an interesting topic in medical image processing research area and is a computer-based system that helps medical professionals in diagnosing of diseases from medical images such as X-ray, magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and microscopic images. Digital image processing thus plays a significant role in the processing and analysis of medical images for diseases identification and detection purposes [12].

3.1.1 Rationale for computer-aided diagnosis

In the clinical interpretation of medical images, limitations are posed by the nature of the human eye/brain visual system, reader fatigue, distraction, the presence of overlapping structures in images, and the vast number of normal cases in screening programs. These limitations provide motivation for the use of CAD with the potential to improve detection, diagnostic performance, and ultimately patient care [13].

3.1.2 Development of computer-aided diagnostic methods

development of a computer algorithm appropriate to the medical interpretation task, validation of the algorithm alone using appropriate cases for performance evaluation and robustness assessment, evaluation of radiologists in the relevant diagnostic task with and without the use of the computer aid, and then ultimate performance evaluation with a clinical trial. Two general types of systems for CcAD are being developed by multiple researchers: CADe for computer aided detection and CADx for computer-aided diagnosis [14].

1. Computer-aided detection CADe: involves the use of computer analyses to indicate locations of suspect regions in a medical image. The characterization, diagnosis, and patient management are left to the radiologist.
2. Aided diagnosis CADx: involves the use of computer analyses to characterize a region or lesion, leaving the final diagnosis and patient management to the radiologist.
   Both CADe and CADx are schematically shown in (**Figure 3.1**).

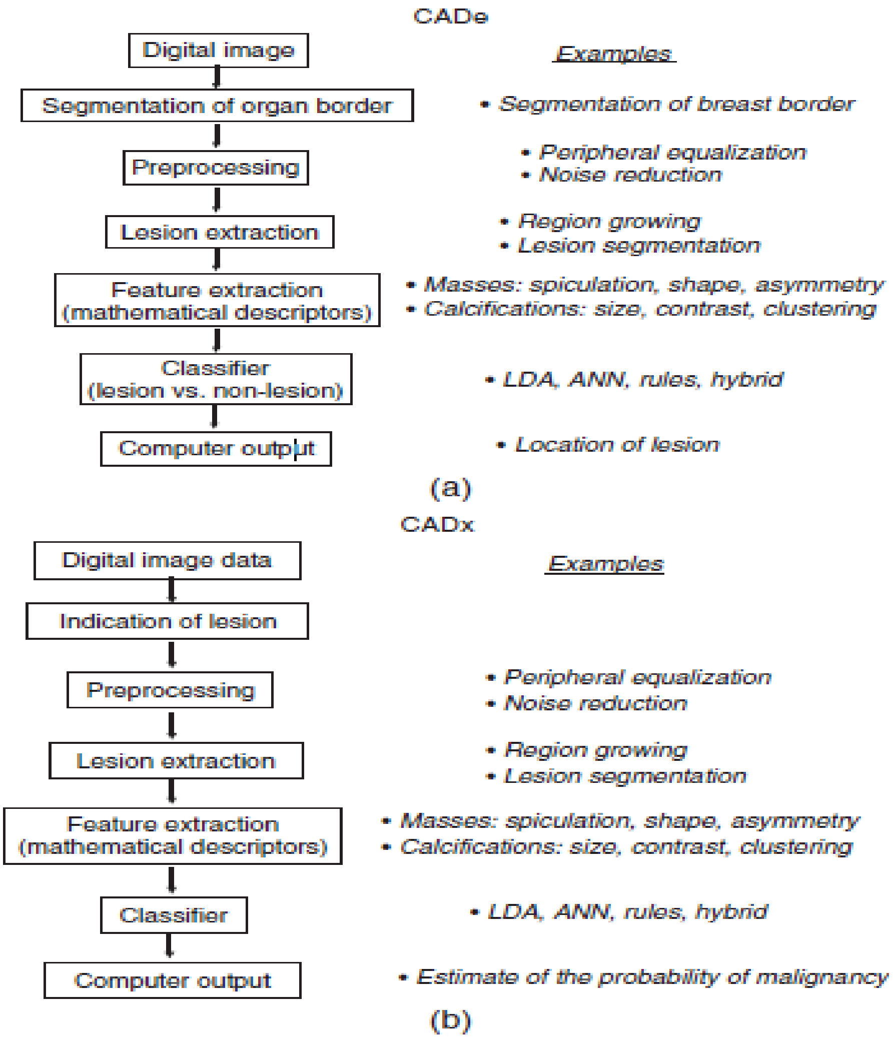

CADe
Digital image
Segmentation of organ border
Preprocessing
Lesion extraction
Feature extraction (mathematical descriptors)
Classifier (lesion vs. non-lesion)
Computer output
Examples
• Segmentation of breast border
• Peripheral equalization
• Noise reduction
• Region growing
• Lesion segmentation
• Masses: spiculation, shape, asymmetry
• Calcifications: size, contrast, clustering
• LDA, ANN, rules, hybrid
• Location of lesion
(a)
CADx
Digital image data
Indication of lesion
Preprocessing
Lesion extraction
Feature extraction (mathematical descriptors)
Classifier
Computer output
Examples
• Peripheral equalization
• Noise reduction
• Region growing
• Lesion segmentation
• Masses: spiculation, shape, asymmetry
• Calcifications: size, contrast, clustering
• LDA, ANN, rules, hybrid
• Estimate of the probability of malignancy
(b)

Figure 3.1. Schematic diagram illustrating the incorporation of CAD into mammographic interpretation [14].

3.2 Principles and basics of image processing

3.2.1 Digital image

An image may be defined as a two-dimensional function, f (x, y), where x and y are spatial (plane) coordinates, and the amplitude off at any pair of coordinates (x, y) is called the intensity or gray level of the image at that point. A digital image in which x, y, and the intensity values of f are all finite, discrete quantities. The field of digital image processing refers to processing digital images by means of a digital computer as a digital image is composed of a finite number of elements, each of which has a particular location and value. These elements are called picture elements, image elements, pixels, and pixels. Pixel is the term used most widely to denote the elements of a digital image [15].

3.2.2 Application of digital image processing technology

Digital image processing technology has made great progress in all walks of life. The application fields of digital image processing are shown in (**Table 3.1**). Digital image processing technology first came from the medical field. Therefore, in the field of biomedical engineering, digital image processing technology has also played a huge role. In addition to the X-ray, MRI and (CT), there are still some microscopic image processing technologies, mainly to identify Red blood cells, white blood cells, and chromosome analysis have played an important role in medical diagnosis and treatment of X-ray image enhancement, electrocardiogram analysis, and ultrasound image processing techniques [16].

Table 3.1. Application of Digital Image Processing Technology [16].

Field	**Application**
Physics and Chemistry	Spectrum Analysis
Biology and Medicine	Cell analysis; CT; X-ray analysis
Environment Protection	Research of atmosphere
Agriculture	Estimation of plants
Irrigation Works	Lake, river and dam
Weather	Cloud and weather report
Communication	Fax; TV; phone
Traffic	Robot; products
Military	Missile guidance; training
Economics	IC-card

3.2.3 Components of image processing system

Image Processing System is the combination of the different elements involved in the digital image processing. Digital image processing uses different computer algorithms to perform image processing on the digital images. It consists of following components **(Figure 3.2) [17]**:

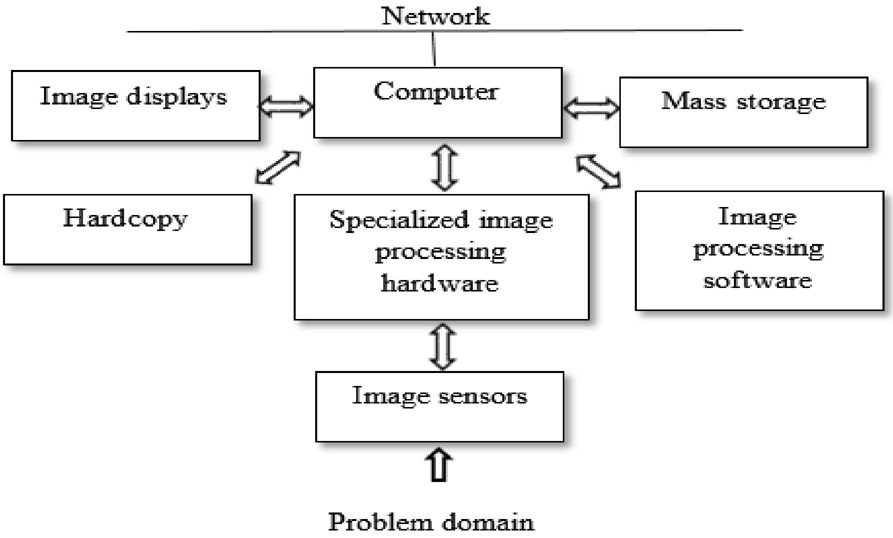


Figure 3.2. Components of a general purpose image processing system [17].

- Image Sensors: senses the intensity, amplitude, co-ordinates and other features of the images and passes the result to the image processing hardware. It includes the problem domain.
- Image Processing Hardware: is the dedicated hardware that is used to process the instructions obtained from the image sensors. It passes the result to general purpose computer.
- Computer: used in the image processing system is the general purpose computer that is used by us in our daily life.
- Image Processing Software: is the software that includes all the mechanisms and algorithms that are used in image processing system.
- Mass Storage: stores the pixels of the images during the processing.
- Hard Copy Device: once the image is processed then it is stored in the hard copy device. It can be a pen drive or any external ROM device.
- Image Display: includes the monitor or display screen that displays the processed images.
- Network: is the connection of all the above elements of the image processing system.

3.2.4 Fundamental steps in digital image processing

There are two categories of the steps involved in the image processing: methods whose input and output are images, and methods whose inputs may be images, but whose outputs are attributes extracted from those images (**Figure 3.3)** [15]:

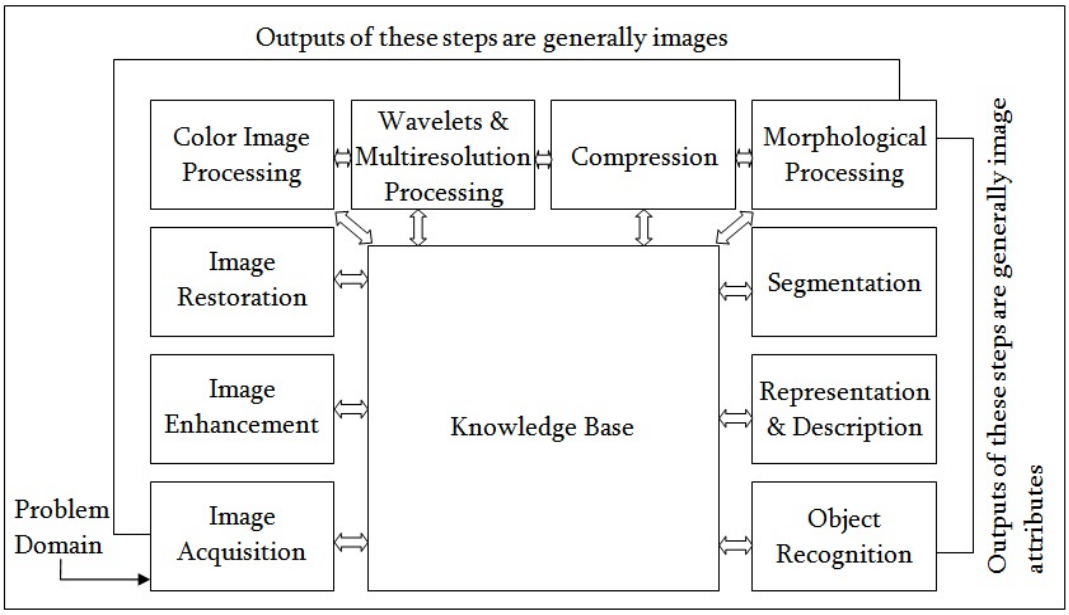


Figure 3.3. Fundamental steps in digital image processing [15].

- Image acquisition: It could be as simple as being given an image that is already in digital form.
- Image Enhancement: The idea behind this is to bring out details that are obscured or simply to highlight certain features of interest in image.
- Image Restoration: It deals with improving the appearance of an image. It is an objective approach, in the sense that restoration techniques tend to be based on mathematical or probabilistic models of image processing. Enhancement, on the other hand is based on human subjective preferences regarding what constitutes a “good” enhancement result.
- Color image processing: It deals with basically color models and their implementation in image processing applications.
- Wavelets and Multi resolution Processing: These are the foundation for representing image in various degrees of resolution.
- Compression: It deals with techniques reducing the storage required to save an image, or the bandwidth required to transmit it over the network.

- Morphological processing: It deals with tools for extracting image components that are useful in the representation and description of shape and boundary of objects. It is majorly used in automated inspection applications.
- Representation and Description: It always follows the output of segmentation step that is, raw pixel data, constituting either the boundary of an image or points in the region itself.
- Recognition: It is the process that assigns label to an object based on its descriptors. It is the last step of image processing which use artificial intelligence of software.
- Knowledge base: Knowledge about a problem domain is coded into an image processing system in the form of a knowledge base. This knowledge may be as simple as detailing regions of an image where the information of the interest in known to be located. Thus limiting search that has to be conducted is in seeking the information. The knowledge base also can be quite complex such interrelated list of all major possible defects in a materials inspection problems or an image database containing high resolution satellite images of a region in connection with change detection application.

3.3 Machine learning algorithms

Machine learning (ML) is a sub-field of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
There are many types of Machine Learning Techniques that are shown in **(Figure 3.4)** [18]. Supervised, Unsupervised, Semi Supervised, Reinforcement Learning [19].

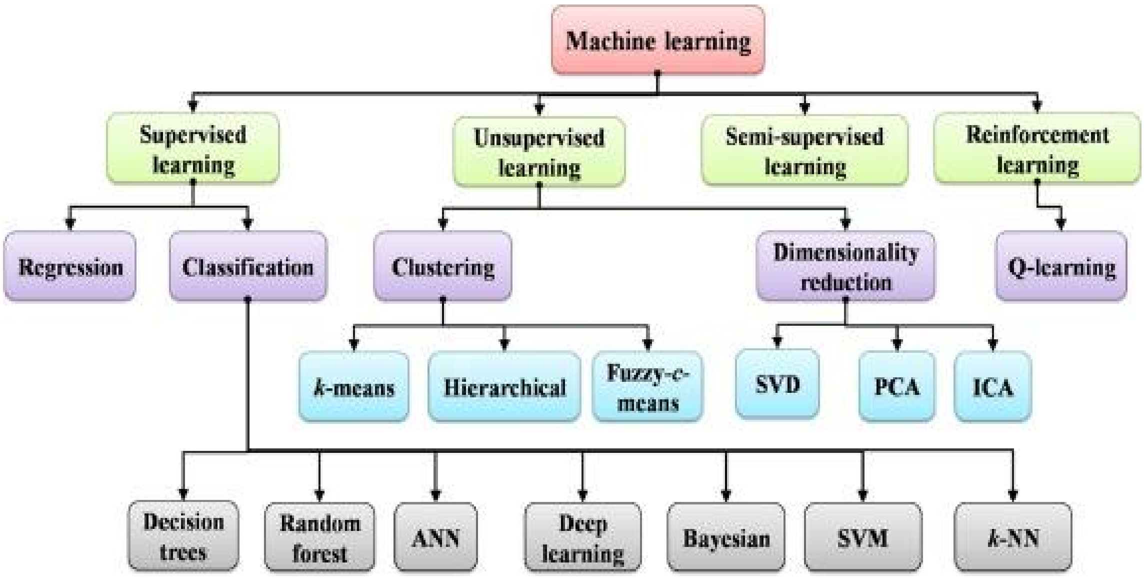


Figure 3.4. Taxonomy of ML techniques [18].

- Supervised Learning: In this type of machine-learning system, the data that you feed into the algorithm, with the desired solution, are referred to as “labels”. The most important supervised algorithms are K-nears neighbors, Linear regression, Neural networks, Support vector machines, Logistic regression, Decision trees and random forests.
- Unsupervised Learning: In this type of machine-learning system, you can guess that the data is unlabeled. The most important unsupervised algorithms are: Clustering: k-means, hierarchical cluster analysis. Association rule learning: Eclat, apriori. Visualization and dimensionality reduction: kernel principal component analysis, t-distributed.
- Reinforcement Learning: This is another type of machine-learning system. An agent “AI system” will observe the environment, perform given actions, and then receive t rewards in return. With this type, the agent must learn by itself. Ties called a policy.

Chapter Four

Literature review

4.1 Introduction

In terms of medical engineering, many research papers have been published, specifically in the field of Bioimage processing, which contributed to the study of microscopic blood images, including normal or pathological images, in order to segment, identify or classify leukemia based on different and advanced techniques. Numerous algorithms have been developed to identify different diseases, e.g., leukemia. A quick, safe, and accurate early-stage diagnosis of leukemia plays a key role in curing and saving patients' lives. However, in terms of effectiveness, practicality and performance, these methods and studies require continuous improvement [11].

4.2 Related studies

- **Leucocyte classification for leukemia detection using image processing techniques**

This paper presents a complete and fully automated method for WBC identification and classification using microscopic images [20].

Methology: The proposed approach isolates the whole leucocyte and then separates the nucleus and cytoplasm. This approach is necessary to analyze each cell component in detail. From each cell component, different features, such as shape, color and texture, are extracted using a new approach for background pixel removal. This feature set was used to train different classification models in order to determine which one is most suitable for the detection of leukemia.

Results: Using this method, 245 of 267 total leucocytes were properly identified (92% accuracy) from 33 images taken with the same camera and under the same lighting conditions (**Figure 4.1**).

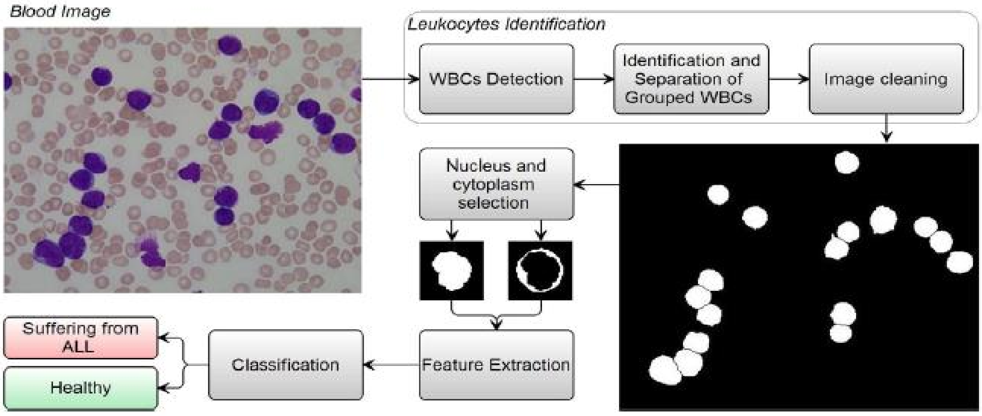

Blood Image
Leukocytes Identification
WBCs Detection
Identification and Separation of Grouped WBCs
Image cleaning
Nucleus and cytoplasm selection
Feature Extraction
Classification
Suffering from ALL
Healthy

Figure 4.1. Diagram of the proposed method; from blood image to ALL classification via the identification of WBCs [20].

Performing this evaluation using different classification models allowed to establish that the support vector machine (SVM) with a Gaussian radial basis kernel is the most suitable model for the identification of ALL, with an accuracy of 93% and a sensitivity of 98%. Furthermore, evaluated the goodness of this new feature set, which displayed better performance with each evaluated classification model. The proposed method permits the analysis of blood cells automatically via image processing technique and could also be used for counting.

- **Robust technique for the detection of acute lymphoblastic leukemia**

This paper demonstrates a method using morphological operations in MATLAB to segment images and compare between different classifiers to perfectly diagnose the presence of ALL in blood smear [21].

Methodology: An automated blood slide image analysis system can be a useful tool for the detection of ALL. The basic method for diagnosis can be divided into following steps (**Figure 4.2**).

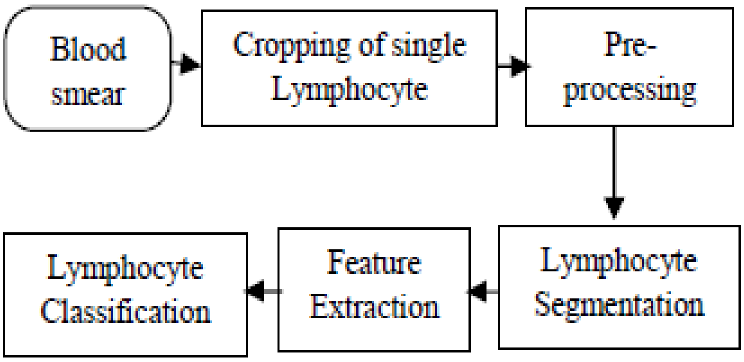


Figure 4.2. Basic block diagram of the system [21].

Results: The segmentation technique used relied on the morphological shape features (Area, Perimeter, Circularity, Axis Length, Form Factor) and the segmentation accuracy was 96.67%. The classification was based on the use of an artificial neural network, and its accuracy was 95.23%.

- **Automatic detection of Acute Lymphoblastic Leukemia using Image Processing**

This study provides high speed, accuracy and scope for early detection of the disease. The algorithm is implemented in MATLAB and an average accuracy greater than 90% is achieved [22].

Methodology: The implementation occurs in five stages; The first stage identifies the lymphoblasts based on its physical characteristics and separates it from the rest of the blood sample, the second stage is the separation of grouped and individual lymphoblasts. The next stage involves the separation of clustered lymphoblasts by application of distance transform of watershed segmentation. The fourth phase involves removing of abnormal and unwanted cell components by shape control. The final stage deals with the counting of detected lymphoblasts and calculating the accuracy of the method (**Figure 4.3**).

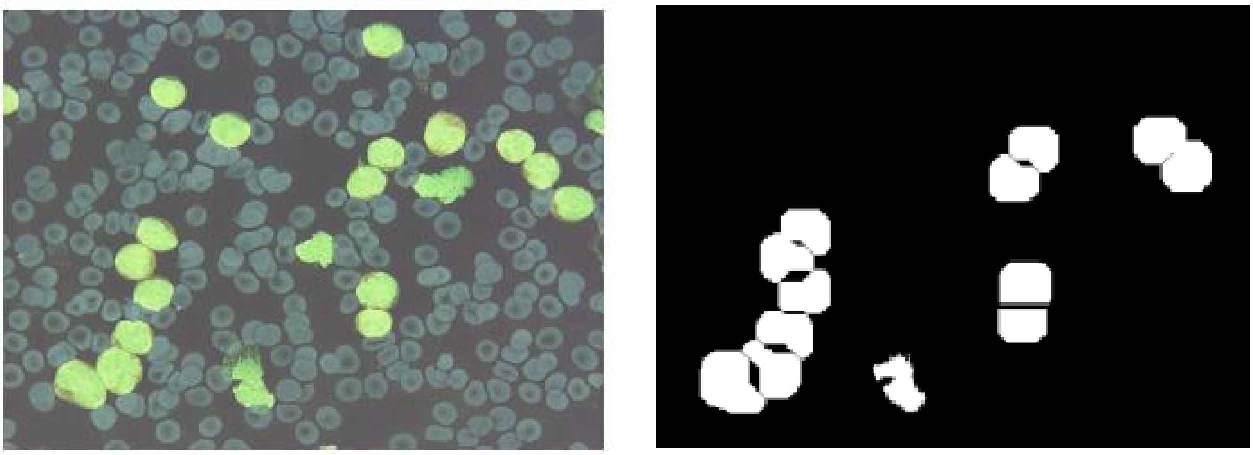


Figure 4.3. Grouped lymphoblasts before and after their separation [22].

Results: The proposed algorithm is validated on the ALL-IDB1 database. The test was carried on 33 blood samples from the database. The algorithm successfully identified 245 lymphoblasts out of the 255 that were given as the input. The algorithm failed to produce accurate results in scenarios where the lymphoblasts were irregular in shape and grouped in large clusters. It also failed to produce the exact count when the shapes or appearance of the other blood cell components was similar to lymphoblasts.

- **Leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks**

In this work, scientists developed deep convolutional neural network for automated detection of acute lymphoblastic leukemia (ALL) and classification of its subtypes into 4 classes, that is, L1, L2, L3, and Normal which were mostly neglected in previous literature **(Figure 4.4).**

Methodology: In contrary to the training from scratch, they deployed pretrained AlexNet (**Figure 4.5**) which was fine-tuned on data set (ALL-IDB1 andALL-IDB2). Last layers of the pretrained network were replaced with new layers which can classify the input images into 4 classes. To reduce overtraining, data augmentation technique was used [23].

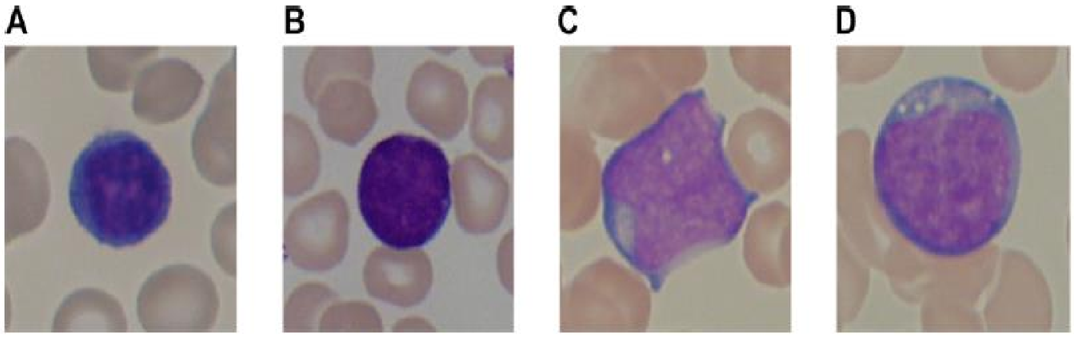


Figure 4.4. Subtypes of ALL according to FAB. (A) A noncancerous cell, (B) L1 type ALL, (C) L2 Type ALL, (D) L3 Type ALL. ALL indicates acute lymphoblastic leukemia; FAB, French American British [23].

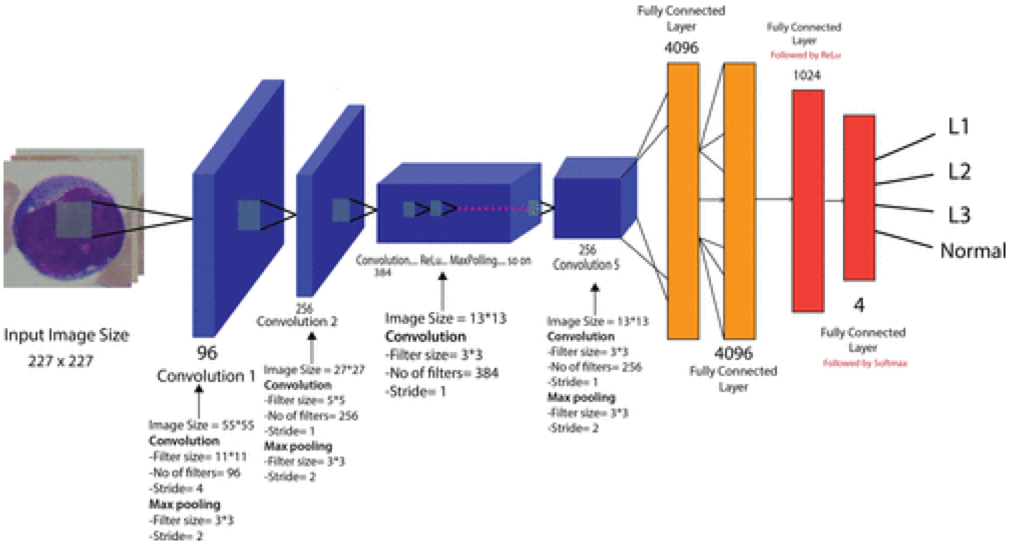


Figure 4.5. AlexNet architecture for acute lymphoblastic leukemia subtype classification. Last 2 layers are newly added [23].

Results: For ALL detection, they achieved a sensitivity of 100%, specificity of 98.11%, and accuracy of 99.50% and for ALL subtype classification the sensitivity was 96.74%, specificity was 99.03%, and accuracy was 96.06%. Unlike the standard methods, this proposed method was able to achieve high accuracy without any need of microscopic image segmentation.

4.3 Comparison of studies

The following table (**Table 4.1)** shows the most prominent differences between previous studies that focused on acute lymphoblastic leukemia (ALL).

Previous studies were selected to discuss their results and benefit from them in our research. Despite the different treatment conditions and algorithms used, the researchers achieved good results in terms of segmentation, identification, and classification of ALL.

Observing the (**Table 4.1**), we find that employing convolutional neural networks (CNN) gave a high accuracy for the detection and classification of ALL. Also, using artificial neural network (ANN) and support vector machine (SVM) technology gave satisfactory results. In addition to using the Hough transform which is a good method but the algorithm needs to be developed. Although the ANN study relied on morphological features only related to the shape of leukocytes, while the SVM study relied on extracting many features (shape, color, texture features), which opens the way for us to integrate and search for many features and explain their difference between normal and leukemia images to obtain get better classification and therefore high accuracy results using deep learning technology.

Table 4.1. Characteristics of the studies that used machine learning algorithms to detect and classify ALL.

First author, year of publication, and country	**Aim of the study**	**Data**	**ML method**	**Validation results**
Putzu et al., 2014, [20]	Leucocyte classification for leukemia detection	ALL-IDB1	SVM Shape, color, and texture features	Accuracy= 93%
Bhattacharjee and Saini., 2015, [21]	ALL detection	ALL-IDB1, and ALL-IDB2	ANN Morphological operation	Accuracy =96.67%

Chaitra and Aditi., 2016,[22]	Automatic detection of ALL	ALL-IDB1	Hough transform. Features not mentioned	Accuracy = 90%
Shafique and Tehsin, 2018, [23]	Detection and classification of ALL	ALL-IDB1 and ALL-IDB2	CNN (AlexNet)	ALL detection accuracy = 99.50%, ALL subtype classification = 96.06%,

Chapter Five
Materials and methodology

5.1 Dataset

Images used in this study were obtained from ALL-Image Data Base (IDB) dataset which is a public dataset available online. The images of the dataset have been captured with an optical laboratory microscope coupled with a Canon PowerShot G5 camera. All images are in JPG format with 24-bit color depth, resolution 2592 x 1944. This dataset was divided into 2 versions [24].

- **Acute lymphoblastic leukemia-IDB 1 (ALL_IDB1):** The ALL_IDB1 version 1.0 collected during September, 2005 (**Figure 5.1**) can be used both for testing segmentation capability of algorithms, as well as the classification systems and image pre-processing methods. This dataset is composed of 108 images where 59 images were from healthy patients and 49 images were from patients affected with leukemia. It contains about 39000 blood elements, where the lymphocytes have been labelled by expert oncologists. The images are taken with different magnifications of the microscope ranging from 300 to 500.

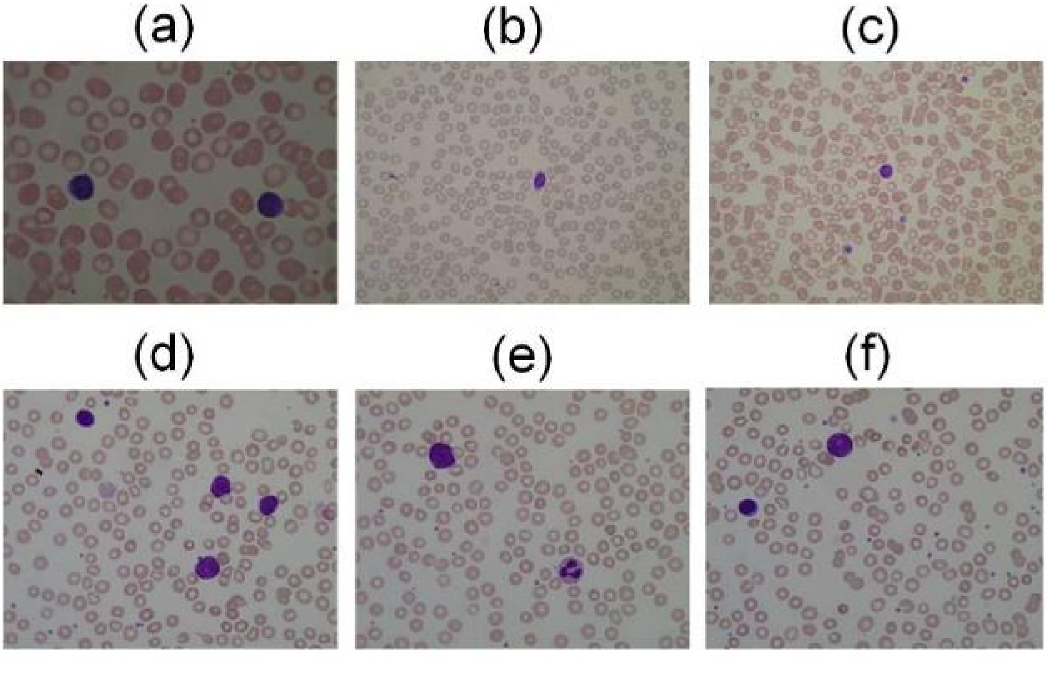


Figure 5.1. Examples of the images contained in ALL-IDB1: healthy cells from non-ALL patients (a-c), probable lymphoblasts from ALL patients (d-f) [24].

- **Acute lymphoblastic leukemia-IDB 2 (ALL_IDB2):** The ALL-IDB2 version 1.0 is a collection of cropped area of interest of normal and blast cells that belongs to the ALL-IDB1 dataset (**Figure 5.2**). This dataset consisted of 260 images having single cell where 130 images were from patients affected by leukemia and 130 were normal images. These images had resolution of 257×257 with 24 bit color depth. ALL-IDB2 images have similar grey level properties to the images of the ALL-IDB1, except the image dimensions.

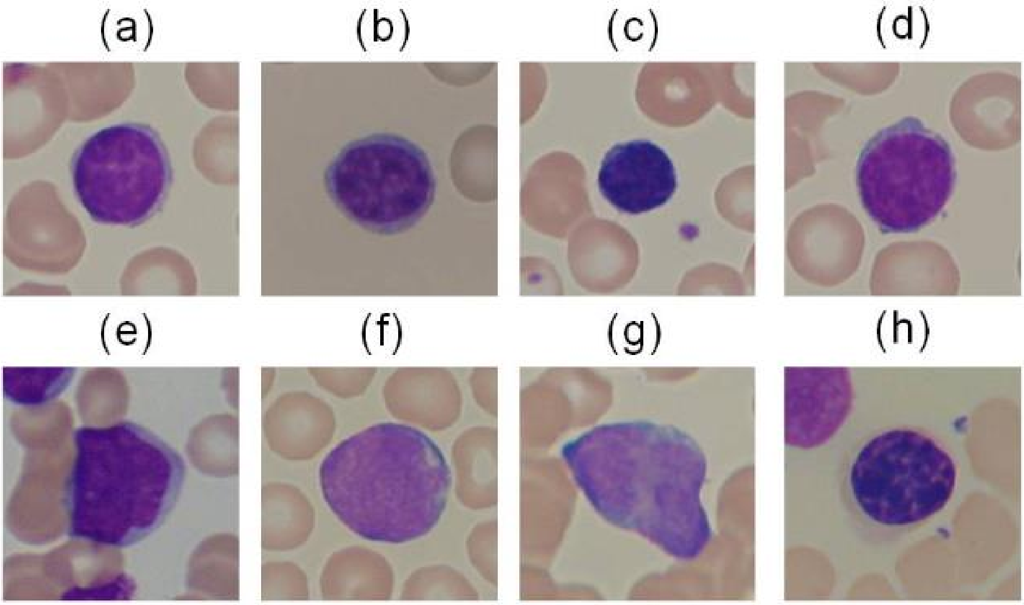


Figure 5.2. Examples of the images contained in ALL-IDB2: healthy cells from non-ALL patients (a-d), probable lymphoblasts from ALL patients (e-h) [24].

5.2 Proposed methodology

This chapter of this study presents a systematic survey of the computational steps in ALL detection based on histopathology. These steps are: (1) Image acquisition to provide an appropriate data set, (2) Image pre-processing to enhance the quality of images, (3) WBCs detection to extract the leukocyte from the image and deal with it only in the later stages (4) Feature extraction to quantify the properties of these leukocytes, and (5) Classifying these leukocytes as cancerous and noncancerous. The proposed method for diagnosis of ALL cells presented in (**Figure 5.3**). These steps are described in details in the following sections.

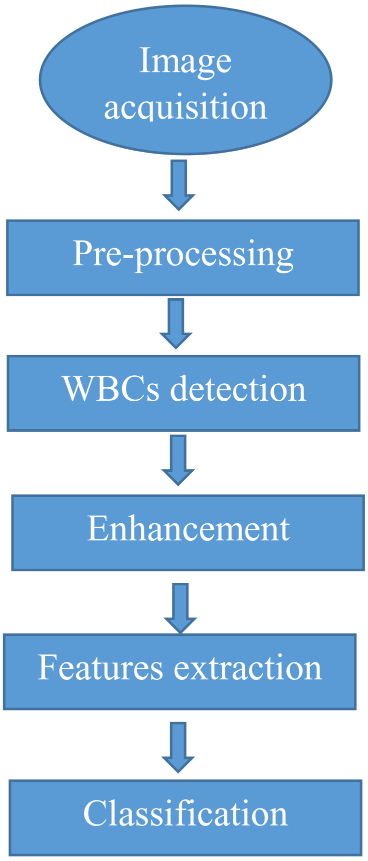


Figure 5.3. Block diagram of the proposed method.

1. Image acquisition

The ALL_IDB2 data was relied on because the aim of this study is to focus on the region of interest, which is WBCs, and to understand their features in the case of ALL, as well as the impact of these features on the classification process later, so that we can directly and accurately process the leukocyte alone.

2. Pre-processing

The initial processing procedures were started after initializing the image path and entering it into Matlab version 2020, which included converting the image to grayscale and enhancing its contrast.

To improve visibility and reduce noise, the complement of the image was applied.

The results of applying this image pre-processing on a sample image are shown in (**Figure 5.4**).

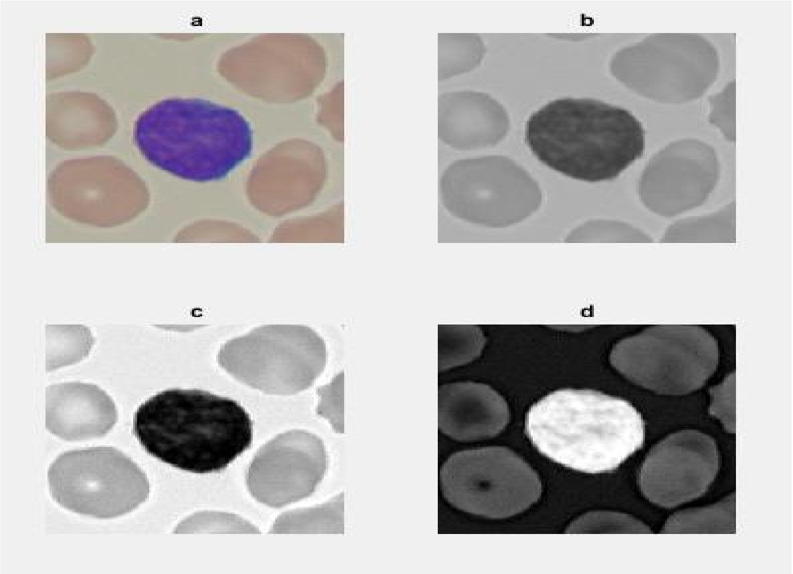


Figure 5.4. Result of pre-processing algorithm. (a) Original image, (b) image in grayscale, (c) adjust image and (d) image complement.

3. WBCs detection

This section discusses the segmentation technique used to extract the leukocyte from the blood smear images. After applying the pre-processing step, the segmentation phase was performed. Segmentation plays a key role since it will directly affect subsequent processing that is feature extraction and classification.

To identify the leukocyte in the image, the automated threshold was applied to the majority of the images in order to determine a single threshold value that is appropriate for all images after that the image was sampled at an acceptable threshold (the manual threshold selected was 0.5) and thus converted to a binary image (**Figure 5.5**).

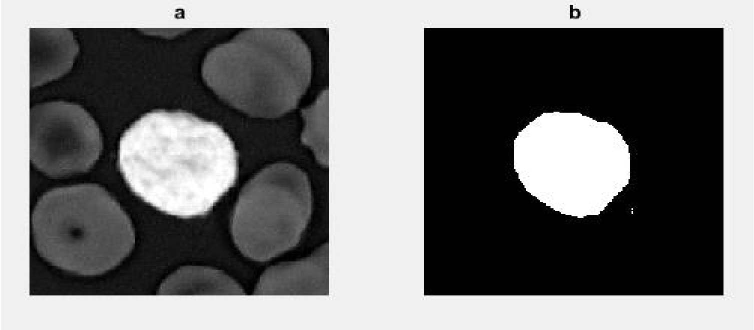


Figure 5.5. Result of WBC detection. (a) image complement, (b) binary image.

4. Enhancement

Because the attention at this phase is on the leukocyte, which is now the region of interest, it was important to enhance the image, especially this region. The optimization procedure in this case included a few simple process. Since several images had more than one leukocyte, it was necessary to eliminate the secondary cells and concentrate on only one cell for each image, therefore all regions with an area less than 1000 were deleted for all images. The undesirable elements around the cell were then removed, and the image's borders were cleaned up (**Figure 5.6**) which might affect the quality of feature extraction later on.

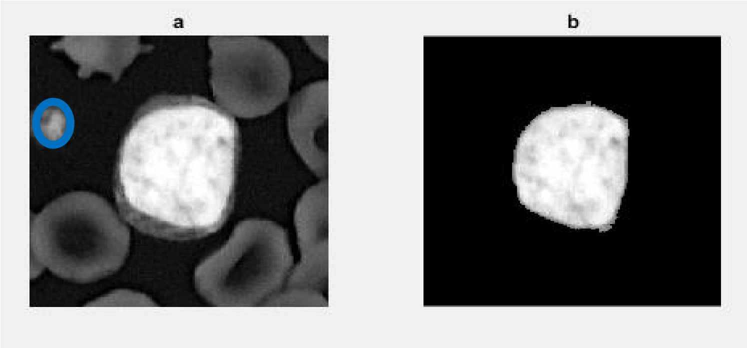


Figure 5.6. Result of WBC enhancement. (a) complement cleaned image, (b) enhanced image.

5. Features extraction

For the feature extraction stage, segmented regions were employed. First, the collected features give important information for classifying cells as cancerous or noncancerous. The success of this section directly affects the classifier's overall performance so the aim of this phase is to extract the descriptive information from an image. The proper selection of the features is considered the second most challenging step in the field of automated identification of leukemic cells.

This section includes the following two steps: The first phase is "Features generation" which involves creating a set of features from segmented leukocytes. "Features selection" is the second step, which involves determining the optimal set of features that will lead in the highest recognition efficiency. These two steps will be thoroughly described in the following sections.

- **Features generation:**

The feature generation step's purpose is to create a set of quantitative features from the image's WBCs. To identify distinct cells, we require specific characteristics from the cells. For this study, two main kinds of features were considered for cell classification: morphological and statistical features. These two main features will provide useful information for classification.

we implemented twelve widely used features, of which eight had morphological characteristics and four had statistical characteristics.

Morphological features: According to hematologists, the geometric of the cells is one of the essential features that is used for characterization (**Table 5.1**). In order to reflect this information in feature vectors, several geometrical features are considered including: Minor axis length, major axis length, perimeter, area, solidity, eccentricity, form factor and compactness [25].

Table 5.1. The morphological features that were applied in this study.

Morphological features	
Minor axis length	Length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region, returned as a scalar
Major axis length	Length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region, returned as a scalar.
Perimeter	The perimeter is computed by calculating the distance between each adjoining pair of pixels around the border of the region.
Area	Actual number of pixels in the region, returned as a scalar. (This value might differ slightly from the value returned by bwarea, which weights different patterns of pixels differently).
Solidity	Proportion of the pixels in the convex hull that are also in the region, returned as a scalar. Computed as Area/Convex Area.
Eccentricity	The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. (0 and 1 are degenerate cases. An ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.)
Form factor	A measure of shape irregularities independent on the object's size. In general, a circular nucleus has the greatest area to perimeter ratio, and this measure is equal to 1 for a perfect circle. Consequently, for the nuclei of leukemic cells, this ratio converges to a value of 1. The form factor is defined as $Form\ factor = \frac{4*\pi*Area}{Perimeter^2}$ (1)

Compactness	The degree to which the form is compressed. The shape of the nucleus changes widely according on the maturity and kind of WBC. Leukemic cell nuclei are typically ovoid or spherical in form, with more overall compactness than mature cell nuclei. The following formula yields the compactness measure: $\text{Compactness} = \frac{Perimeter^2}{Area}$ (2)

Statistical features: Statistical features in image processing give us information about the spatial arrangement of intensities in an image. These features, which include measurements such contrast, correlation, energy and homogeneity are produced from the enhanced images of the cells (**Table 5.2**). These features, which are defined as standard methods and are available in the Matlab image processing toolbox, are simple to measure.

Table 5.2. The statistical features that were applied in this study.

Statistical features	
Contrast	Returns a measure of the intensity contrast between a pixel and its neighbor over the whole image. $Contrast = \sum_{i,j} \lvert i-j \rvert^2 \; p(i,j)$ (3) Range = [0 (size(GLCM,1)-1) ^2] Contrast is 0 for a constant image. The property Contrast is also known as variance and inertia.
Correlation	Returns a measure of how correlated a pixel is to its neighbor over the whole image. $Corr = \sum_{i,j} \frac{(i-\mu j)(i-\mu j)p(i,j)}{\sigma_i \sigma_j}$ (4) Range = [-1 1] Correlation is 1 or -1 for a perfectly positively or negatively correlated image. Correlation is NaN for a constant image.
Energy	Returns the sum of squared elements in the GLCM. $E = \sum_{i,j} P(i,j)^2$ (5) Range = [0 1] Energy is 1 for a constant image. The property Energy is also known as uniformity, uniformity of energy, and angular second moment.
Homogeneity	Returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. $H = \sum_{i,j} \frac{p(i,j)}{1+\lvert i-j \rvert}$ (6) Range = [0 1] Homogeneity is 1 for a diagonal GLCM

- **Features selection:**

The described methods for feature generation produce a rich set of 2580 leukocyte fragmented features that were detected from 215 images (104 leukemia and 111 normal images) distributed over the 12 previously mentioned statistical and morphological features. Although the number of features is not much but it was essential to investigate and analyze these features and their relationships, first from a statistical standpoint and then from a pathological one.

From a statistical point of view, in order to select the features, the (best ANOVA F values) have been applied to the 12 morphological and statistical features used in the study in Python language.

ANOVA is a statistical hypothesis test that allows determining whether there is a statistical difference between the means of three or more groups. While other hypothesis tests for differences in means directly use the means of the data (T-test), ANOVA is unique in that it uses a similar sum of squares approach to variance to determine whether the means are statistically different [26].

ANOVA looks at three sources of variability (**Figure 5.7**):

1) Total: Total variability among all observations.

2) Between: Variation between subgroup means (signal).

3) Within: Random (chance) variation within each subgroup (noise).

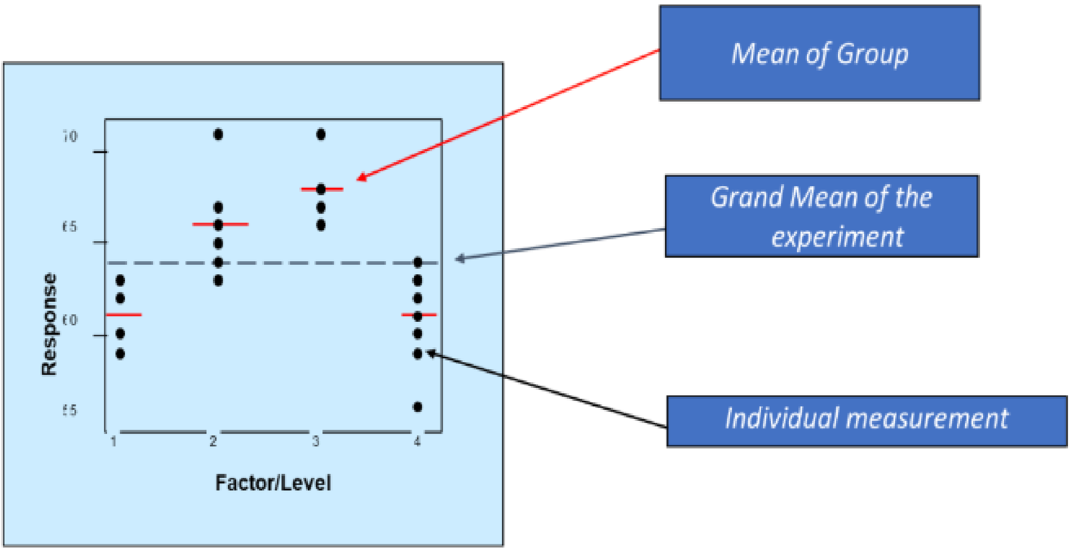


Figure 5.7. Analysis of Variance (ANOVA).

The F-value is the ratio (F-ratio) of the Between and Within variation. The Between variation is calculated by summing the squared differences between the group means and the grand mean. The Within variation is calculated by summing the squares of the difference between the individual values and the group mean.

$$\sum_{j=1}^{g} n_j \, (\bar{y}_j - \bar{\bar{y}})^2 \qquad \text{SS(Between)} \quad (7)$$

$$\sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \qquad \text{SS(Within)} \quad (8)$$

The Between (Factor) and Within (Error) variation are now put in the form of a ratio. This ratio is called the F-ratio, and the resulting value of that ratio is called the F-value.

$$\text{F value (F ratio)} = \frac{Between\ group\ variation}{Within\ group\ variation} \quad (9)$$

6. Classification

After determining an appropriate set of features from WBCs as mentioned above, the next step is to distinguish these WBCs using these features as the inputs of ANN.

ANN is a classification technique, that uses several computing units to imitate neurons in the human brain. All units are connected with each other via a weighted link, which determines the prominence of the respective input to the output. Each neuron in a structure performs a weighted sum of all inputs and finds the output using an activation function. This activation function decides whether the information is relevant or should not pass to the subsequent unit. The whole process of learning is based on altering the values of weights and biases depending on the calculated loss function between the actual and desired output [27].

Due to the fact that there are no specific guidelines on how to determine the optimal neural network architecture parameters, in particular the number of hidden layers and neurons, we decided to select these parameters through a trial and- error process. During this process, several architectures with different numbers of neurons and hidden layers were tried experimentally. The number of neural units in the first and last layers depends on the number of given inputs and desired outputs. In this study, we consider 12 input neurons, where each neuron represents one of the extracted features, and two output neurons, for the leukemic and normal classes. In this phase, we additionally split the dataset into a training, validation and test set.

Chapter Six

Results

Experimental verification and results:

The results of applying the proposed algorithm (**Figure 6.1**) showed pathological images of white blood cells and we were later able to use them to obtain and approve the extracted feature values for classification.

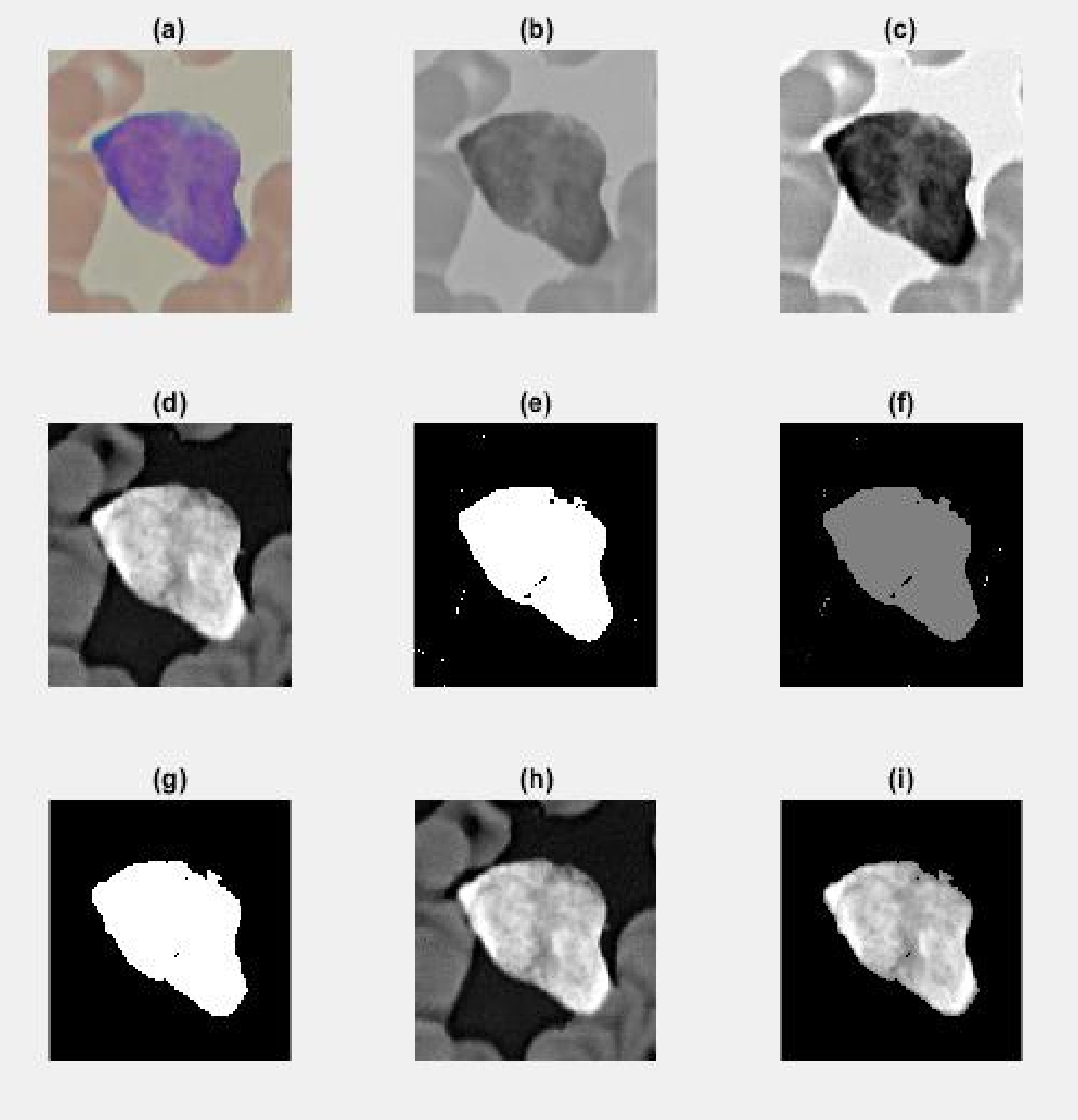


Figure 6.1. Results of proposed algorithm. (a) Original image, (b) image in grayscale, (c) adjust image, (d) image complement., (e) binary image, (f) image Labels, (g) cleaned image, (h) complement cleaned image, (i) enhanced image.

In the final analysis, 215 extracted sub-images of 111 normal WBCs and 104 leukemic cells were used to evaluate the proposed system. 2580 morphological and statistical features were extracted from these images (**Table 6.1)**.

Table 6.1. The mean and standard deviation of each of the morphological and statistical feature values extracted from processed medical images for cancerous and noncancerous states in addition F and P ANOVA values.

	Morphological								Statistical			
	l	L	P	A	S	EC	FF	COM	CON	COR	EN	H
Cancerous	113 ± 16	134 ±24	435 ±97	11865 ±3498	0.95 ±0.03	0.48 ±0.15	0.80 ±0.15	16.31 ±4.30	0.301 ±0.03	0.981 ±0.002	0.25 ±0.05	0.919 ±0.009
Noncancerous	100 ±25	125 ±33	433 ±161	9480 ±4937	0.90 ±0.09	0.54 ±0.17	0.69 ±0.23	21.98 ±12.98	0.335 ±0.07	0.980 ±0.004	0.25 ±0.056	0.920 ±0.014
F-value	18.22	4.74	0.013	16.50	25.16	5.68	17.06	17.93	19.51	7.46	0.001	0.34
P-value	2.952	0.030	0.906	6.808	1.109	0.017	5.181	3.394	1.591	0.006	0.968	0.55

1: l; minor axis length, 2: L; major axis length, 3: P; perimeter, 4: A; area, 5: S; solidity, 6: EC; eccentricity, 7: FF; form factor, 8: COM; compactness, 9: CON; contrast, 10: COR; correlation, 11: EN; energy, 12: H; homogeneity.

After applying (best ANOVA F values) to the 12 morphological and statistical features used in the study, the number of features was reduced and the best 6 features were obtained, which are minor axis length, area, solidity, form factor, compactness and contrast (**Table 6.2)**.

Table 6.2. The mean and standard deviation of best F and P ANOVA features values extracted from processed medical images for cancerous and noncancerous states.

	l	A	S	FF	COM	CON
Cancerous	113 ±16	11865 ±3498	0.95 ±0.03	0.80 ±0.15	16.31 ±4.30	0.301 ±0.03
Noncancerous	100 ±25	9480 ±4937	0.90 ±0.09	0.65 ±0.23	21.98 ±12.98	0.335 ±0.07
F-value	18.22	16.50	25.16	17.06	17.93	19.51
P-value	2.952	6.808	1.109	5.181	3.394	1.591

1: l; minor axis length, 2: A; area, 3: S; solidity, 4: FF; form factor, 5: COM; compactness,6: CON; contrast.

Classification experiments are performed to distinguish between cancerous and noncancerous states for 215 images. For the classification step with an (ANN), we assigned 70% of the data set to a training subset, which was used to build the prediction network, and 30% of the data to validate and test the suggested network.

The classification is based on Levenberg-Marquardt Artificial Neural Networks (LMANNs). LMANNs was designed with hidden layers [15 7 3] and 12 inputs according to the number of features extracted and 1 outputs representing the previous states. The performance was evaluated based on the classification results. The cancerous and non-cancerous cells against the ANN result shows in (**Table 6.3).**

Table 6.3. cancerous and noncancerous cells versus result of ANN.

Output of ANN	Detected Cancerous	Detected noncancerous
Cancerous	81	23
Noncancerous	17	94

The performance of the ANN is evaluated by these parameters: Sensitivity, specificity, and accuracy.

In our study, prementioned parameters in the definition of evaluation terms are as below: True positive (cancerous cell correctly identified), false positive (noncancerous cells identified as cancerous), true negatives (noncancerous correctly identified), false negatives (Cancerous cells identified as noncancerous).

- Sensitivity is the probability of a positive diagnosis test among persons that have the disease and it is defined as:

$$\text{Sensitivity} = \frac{T_p}{T_p+F_N} \qquad (10)$$

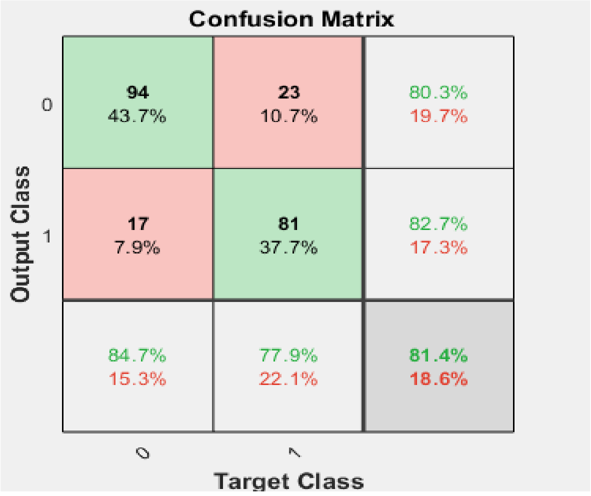
- Specificity is the probability of a negative diagnosis test among persons that do not have the disease and it is defined as:

$$\text{Specificity} = \frac{T_N}{T_N+F_P} \qquad (11)$$

- Accuracy is a criterion that shows the closeness of the output of the classifier and real value and it is defined as:

$$\text{Accuracy} = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \qquad (12)$$

The results of the proposed algorithm for ANN show 77.8%, 84.6%, and 81.4%, sensitivity, specificity, and accuracy, respectively.



Confusion Matrix

Output Class \ Target Class	0	1	
0	**94** 43.7%	**23** 10.7%	80.3% 19.7%
1	**17** 7.9%	**81** 37.7%	82.7% 17.3%
	84.7% 15.3%	77.9% 22.1%	**81.4%** **18.6%**

Figure 6.2. Classification Confusion Matrix.

Chapter Seven

Discussion and conclusion

7.1 Discussion

An ANN was suggested in this study to classify cancerous and non-cancerous cells using just features extracted from the leukocytes.

Regarding the processing results, it should be mentioned that after 12 features were extracted from each of the 215 processed images, and these features were reduced based on the previously explained (best ANOVA F values), the best 6 features were relied on. Not to enhance classification performance, but to investigate the physiological effect of leukemia on leukocytes.

Referring to (**Table 6.1.**), we note that the feature related to the solidity of both cells (cancerous and non-cancerous) has the highest ANOVA F value of 25.16, indicating that there is a significant difference between the two cases. The solidity of most of the cancer cells was greater than that of this nature, as well as the contrast feature and the form factor, Although the difference between the results in the two cases is slight, but it can be noted that the mean values of form factor for leukemia cells is 0.8, while the other 0.6.

On the contrary, the compactness feature was shown that cancer cells have lower values with mean of 16.13, compared to non-cancerous cells, which have a value of 21.98. In view of the minor axis length and area, we find that the numerical results are slightly higher for cancer cells.

On the other hand, and returning to all the features used in this study, it appears clearly that each of the features of perimeter, homogeneity and energy have the lowest F ANOVA values. The values were very close between the two classification conditions, as they did not add any benefits to using them in understanding the medical images better, as well as improving the performance of the network later.

In conclusion, we note that studying all these features together gives a good evaluation of the images.

Regarding the classification results, the results of the proposed algorithm for ANN show 77.8%, 84.6%, and 81.4%, sensitivity, specificity, and accuracy, respectively.

It is clear that although our proposed method is relatively simple, this algorithm has acceptable performance for diagnosing between ALL as cancerous and non-cancerous cells. Thus, the proposed algorithm can be used as an auxiliary diagnostic tool for pathologists. Moreover, the clinical impact of this research is that it will provide the ability for pathologists to examine blood smears to find cancer cells.

Because there are many studies that used the same data by extracting a certain number of features each time, as well as employed multiple techniques in the classification operations. We hope in the near future to make available new data that includes a large number of medical images for various blood diseases through which we can apply the newly used algorithms and compare the results of current studies on them.

The most important advantages of this study are that it has a simple and uncomplicated algorithm and it has included morphological features that depend on the shape and surface of leukocytes, as well as statistical features. It also discussed the differences between these features in both cases of cancerous and non-cancerous cells, unlike other studies that were satisfied each time by mentioning The features used without discussing them from a medical point of view, in addition to the fact that the number of cells included in this study is not small compared to other studies, as for the accuracy of the classification, it is acceptable but not completely satisfactory.

On the other hand, this study had some negative points, the first of which was that it was better to separate the nucleus of the white blood cell from the cytoplasm and study the features of each of them separately, and the second is that this research relied only on ALL_IDB2 data and not on ALL_IDB1 and ALL_IDB2 data together, which may give better results and reliability. It is also suggested to count the red and white blood cells and take the ratio between them, given that leukemia significantly increases the number of white blood cells, but the counting algorithm that was followed was not accurate, so its results were not written.

The following table (**Table 7.1.)** shows the most important differences of this study from the previous studies that we have previously explained in the literature review chapter. It appears clearly that the features used are more, they depend on the morphology and statistical values resulting from the white blood cells, as well as the most important features that that contribute to understanding and classifying the images, depending on the numerical values generated after the application of ANOVA.

Table 7.1. The most important differences between the characteristics of studies that used ML algorithms to detect and classify ALL and the algorithm used in this study.

First author, year of publication, and country	**Type of feature extracted**	**Data**	**ML method**	**Validation results**	**More information**
Putzu et al., 2014, [20]	Shape, color, and texture features	ALL-IDB1	SVM	Accuracy= 93%	SVM with a Gaussian radial basis kernel is the most suitable model for the identification of ALL.
Bhattacharjee and Saini., 2015, [21]	Morphological operation	ALL-IDB1, and ALL-IDB2	KNN SVM ANN K-means	Specificity KNN=95.23% Specificity SVM=90.47% Specificity ANN=95.23% Specificity K-m=85.71%	Not available for practical application because of the small number of sample.
Chaitra and Aditi., 2016,[22]	Features not mentioned	ALL-IDB1	Hough transform.	Accuracy = 90%	This algorithm failed to produce accurate results and exact count.
Shafique and Tehsin, 2018, [23]	/	ALL-IDB1 and ALL-IDB2	CNN (AlexNet)	ALL detection accuracy = 99.50%, ALL subtype classification = 96.06%,	This proposed method was able to achieve high accuracy without any need of microscopic image segmentation.
This study	Morphological (shape and surface) Statistical	ALL-IDB2	ANN (LMANNs)	Accuracy=81.4% Sensitivity=77.8% Specificity= 84.6%	It relied on leukocyte segmentation that was processed and all the features used before the classification were discussed based on ANOVA.

For future research, we believe that in addition of segmentation of leukocytes and extraction features from it, can improved performance of this system, and the proposed method can be applied on more amount of data for improving validation. Another approach that may improve the procedure is the use of dependable methods for feature reduction, such as PCA.

7.2 conclusion

The main aim of this study was to investigate the features that distinguish cancerous and non-cancerous cells to further understand the biological effect of leukemia on leukocytes. The idea of counting both white and red blood cells and calculating the ratio between them was good because it helps in the automatic identification of ALL and influences the classification process. In this study, 215 medical images of 111 normal and 104 leukemia cells were used to apply the proposed algorithm, which initially included the stage of pre-processing the images, then segmentation, identification and enhancement of leukocytes, removing all unwanted cells and extracting 12 morphological and statistical features, but the counting algorithm will not be addressed, because it did not achieve satisfactory results and therefore will affect the accuracy of the classification later. The best features were then selected based on ANOVA. It turns out that the feature of solidity and form factor as well as contrast are important features that gave analyzable numbers that will be really important for the medical team and on the contrary, the perimeter, energy and homogeneity these features that do not help to understand the differences between the two cell states and which can be ignored in the future studies. In the end, an ANN was relied upon, which was recommended by some previous studies for classification, and the network achieved an acceptable classification accuracy of 81.4%. In the future, this study suggests searching for more features and using, for example, the PCA method to choose the best features, as well as deleting some other features to achieve better results. Certainly, the availability of data with more images and the use of more than one method of classification will give more importance to research and its application in practical life.

References

1. P. B. Svarney, Th. E. Svarney, “The handy anatomy answer book included physiology,” *Library of Congress*, china. 2016.
2. N. Menche, “Biologie Anatomie Physiologie,” Urban *and FischerElsevier* , 8th edition, Munich, German 2012.
3. R. L. Fournier, “Basic Transport Phenomena in Biomedical Engineering,” *Taylor & Francis Group, LLC*, 4th edition. 2018.
4. C. Migul., “Shigley’s Mechanical Engineering Design,” *McGraw-Hill Inc.*, 8th edition, Budynas−Nisbett, 2006.
5. V. W. Rodwell, D. A. Bender, K. M. Botham, P. J. Kennelly, P. A. Weil, “Harper’s Illustrated Biochemistry, White Blood Cells.,” *McGraw Hill*, 31 editions. New York, 2019.
6. L. R. Waston, R. Turley, T. Gersten, “UMMC-Health Encyclopedia- What Are White Blood Cells,” *University of Rochester Medical Center Rochester*, New York, 2022.
7. P. J. Parks, “Leukemia Diseases and Disorders,” *ReferencePointPress,Inc*, San Diego, CA 92198, pp. 11–13, 2010.
8. National Cancer Institute. Surveillance, Epidemiology and End Results Program. Cancer Stat Facts:Leukemia, 2019.
9. S. C. Litin, S. Nanda, “ Mayo Clinic Family Health Book,” *The Ultimate Home Medical Reference,* 5th edition, 2018.
10. J. L. Jameson, A. S. Fauci, “Harrison's Principles of Internal Medicine,” Chapter 102: Acute Lymphoid Leukemia, *McGraw-Hill Education,* 20th edition, *2018*.
11. N. Bibi, M. Sikandar, I. U. den, A. ALmogren and S. Ali, “IoMT-Based Automated Detection and Classification of Leukemia Using Deep Learning,” *Research Article, Journal of Healthcare Engineering*, vol. 2020, Article ID 6648574, 12 pages, 2020.
12. H. Arimura, T. Magome, T., Y. Yamashita, “Computer-aided diagnosis systems for brain diseases in magnetic resonance images,” Algorithms, *Machine Learning for Medical Imaging,* pp. 925–952, Japan, 2009.
13. G. Kumar, “Analysis of medical image processing and its applications in healthcare industry,” *Int. J. Comput. Technol. Appl.*, 2014, **5**, pp. 851–860.
14. D. D. Feng, “Biomedical information Technology,” chapter 16, *Academic Press is an imprint of Elsevier*, United States of America, 2018.
15. R.C. Gonzales, R. E. Woods, “Digital Image Processing,” 4th edition, New York, NY, 2018.
16. C. Luo, Y. Hao, Z. Tong, “Research on Digital Image Processing Technology and Its Application,” 8th International Conference on Management, Education and Information, 2018.
17. N. Subash, M. Sucharitha, “Digital Image Processing,” *Approved by AICTE-Accredited,* Telangana State, India, 2021.
18. D. P. Kumar, T. Amgoth, Ch. S. R. Annavarapu, “Machine learning algorithms for wireless sensor networks: A survey,” *Information Fusion 49 (2019) 1–25,* Dhanbad, India, 2019.
19. R. Russell, “Machine Learning Step-by-Step Guide to Implement Machine Learning Algorithms with Python,” 2018.
20. L. Putzu, G. Caocci, and C. Di Ruberto, “Leucocyte classification for leukaemia detection using image processing techniques,” *Artificial Intelligence in Medicine*, vol. 62, no. 3, pp. 179–191, 2014.
21. R. Bhattacharjee and L. M. Saini, “Robust technique for the detection of acute lymphoblastic leukemia,” in Proceedings of the 2015 IEEE Power, Communication and Information Technology Conference (PCITC), Bhubaneswar, India, 2015.
22. N. Chaitra, S Aditi, “Automatic detection of Acute Lymphoblastic Leukemia using Image Processing,” *(2016) IEEE International Conference on Advances in Computer Applications (ICACA),* India, 2016.
23. 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, 2018.

24. Acute Lymphoblastic Leukemia Image Database for Image Processing Department of Computer Science - Università degli Studi di Milano: https://homes.di.unimi.it/scotti/all/#:~:text=Dataset%20ALL

25. S. A. Ahmad, A. S. Morsy, E. A. Mohy, "Microscopic digital image segmentation and feature extraction of acute Leukemia," Int. J. Sci. Eng.Appl. 5, 228–233. doi: 10.7753/IJSEA0505.1001, 2016.

26. Online Learn Six Sigma (LSS) Certificate Program: F-VALUE (ANOVA) https://www.isixsigma.com/dictionary/f-value-anova/?fbclid=IwAR1sP-

27. Z. Amer and B. Nizar, "Neural Network Principles and Applications," in Digital Systems, Ed. R. J. Tocci (London: Pearson), doi: 10.5772/intechopen.80416, 2018.