CHAPTER 1

INTRODUCTION

1.1. OVERVIEW TO THE RESEARCH TOPIC

Early detection and diagnosis is important in healthcare industry, in order to avoid diseases from getting worse and becoming life-threatening. Early detection of diseases increase the chance of successful treatment plan that can lead a person to complete recovery and is especially very important with malignant diseases like cancer. Cancer is considered as one of the main global health problem and has been found to be the second leading cause of death (Siegel *et al.*, 2023). Cancer comes in various forms and examples include breast cancer, cervix cancer, mouth cancer, colon cancer, skin cancer and blood cancer (http://www.cancer.gov/cancertopics/types/common cancers). Among the various types of cancer, the most common and dangerous type is the blood cancer, also known as leukemia.

Leukemia is an abnormal phenomenon that damages red blood cells, bone marrow and the body's defense system (Ghaderzadeh *et al.*, 2021). Leukemia is the result of the rapid overproduction of abnormal white blood cells. Leukemia occurs when abnormal white blood cells in the bone marrow quickly increase and destroy normal blood cells. It is considered to be the 11^{th} top cancer type worldwide (Lin *et al.*, 2021). According to the report published by Uniyal (2022), more than 1.24 million cases related to Leukemia has been reported worldwide annually. This accounts to six percent approximately of all cancer cases.

Anchan (2023), reported that, approximately, one lakh cases have been reported in India, with a 30% survival rate. The report further stated that the mortality rate has gone up by 20%-30% over the past decade, thus making Leukemia as the primary cause of increased death rate. This rate is predicted to grow alarmingly in the forth-coming years by 21% increase in new cancer cases by 2040 globally. The mortality rate related to Leukemia is expected to grow in an alarming fashion in the forth-coming years by 2030 and may increase by 3.5%/1,00,000 (males) and 2.06% (females) 1,00,000 on average (Li *et al.*, 2020).

In order to reduce this predicted figures, early diagnosis and effective treatment plan is needed. The survival rate is high if detected at early stage. Several procedures are being used to detect the disease, which include physical examination, conventional blood tests and bone marrow tests. A frequently used diagnostic method is the examination of blood smear in the form of microscopic blood images. Malformations identified during analysis are a clear indication of the presence of lymphoblasts (Starza *et al.*, 2019).

To identify malformations in white blood cells, generally, a manual inspection is carried out by an expert. Manual inspection and identification has several drawbacks like being time consuming, high cost as experts are expensive and diagnosis accuracy depends on the experience and workload of the expert. Hence, almost all hematologists, at this stage, use a computer aided diagnostic systems that can help them to detect the disease and its stage accurately. These systems tremendously help to overcome the demerits of late or miss-diagnosis and also help to determine its subgroups (stages).

In the current scenario, both patients and hematologists, dream of an automatic Leukemia detection system that can different normal and malignant blood cells with 100 per cent accuracy and in a fast manner. However, inspite of numerous researches that focus on achieving this goal (Shimony *et al.*, 2023; Ranjitha and Duth, 2021), this dream is still not yet a reality. Hence, research in this domain is still very active. To reduce the gap between the reality and dream, this research work proposes algorithms that enhance each step of Leukemia identification, which when applied collectively, will improve the performance of the detection system. For this purpose, this research work uses image processing and machine learning algorithms.

The proposed automatic Leukemia detection system is part of the clinical decision support system and uses image enhancement, segmentation, feature engineering and classification algorithms to improve its detection efficiency. This chapter provides the introductory materials related to the research topic along with the formulated research objectives.

1.2. BLOOD TYPES AND LEUKEMIA TYPES

Leukemia, is a group of heterogeneous cancers related to blood, where each type differs in its pathogenesis, prognosis and response to treatment (Dong *et al.*, 2020). In

order to obtain a clear understanding of leukemia, this section presents a brief discussion on types of blood and types of Leukemia (Figure 1.1).

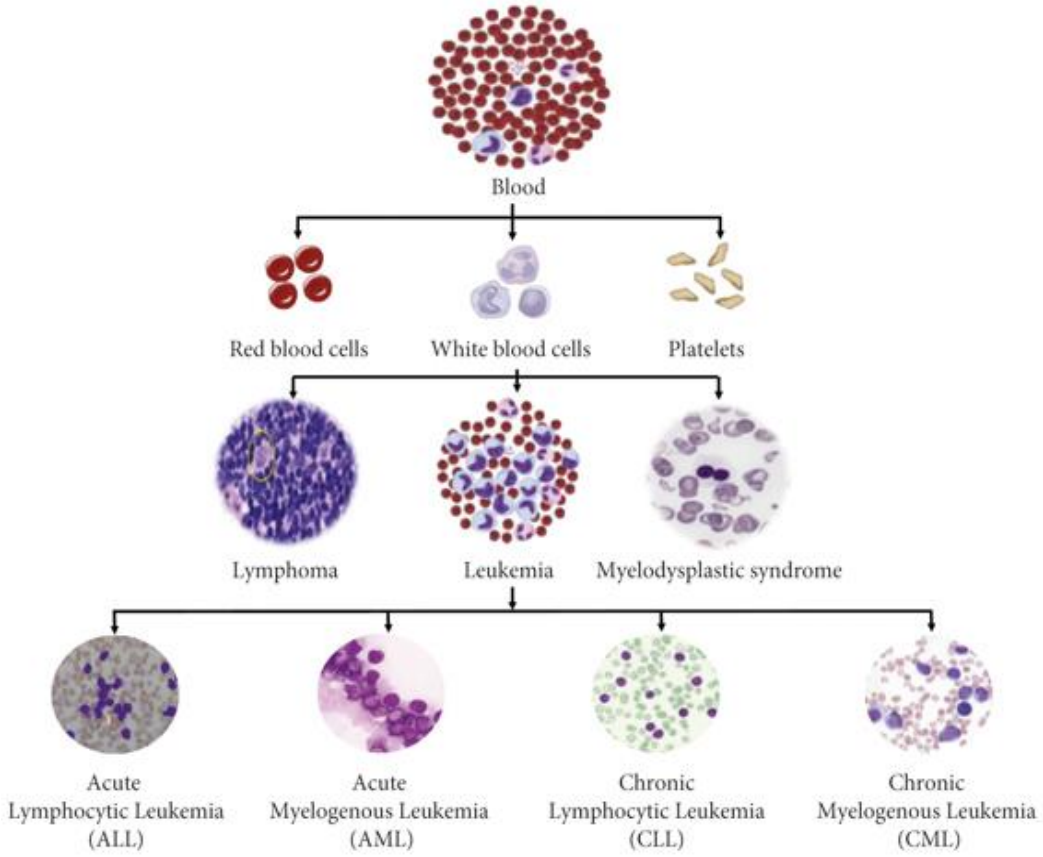


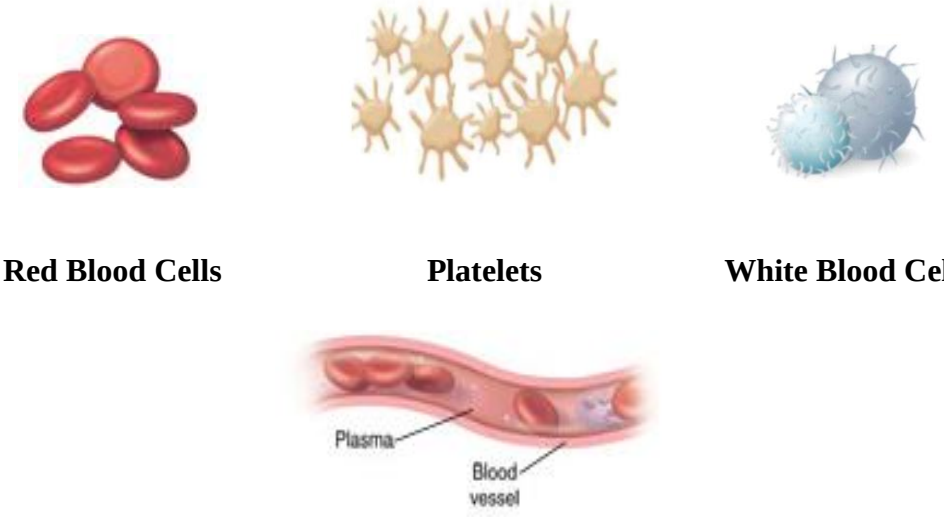
Figure 1.1 : Blood Types and Leukemia Types

1.2.1. Blood Types

Blood, one of the most important part of life, is a specialized body fluid connective tissue. It circulates all over the body and performs various functions that includes delivering oxygen and nutrients to lungs and tissues, forming blood clots that can help to prevent blood loss, carrying cells that have antibodies to fight against infections, taking wastes to kidneys and liver, whose job is to filter and clean blood and regulating body temperature. In an average adult human body 7%-8% of body weight is blood or approximately there is 5-6 liters of blood (https://byjus.com/biology/blood).

The blood has several elements including plasma, blood cells and platelets (Figure 1.2). Plasma, the liquid component of the blood, encompasses ~55% of the blood volume (Bailey, 2022). Plasma consist of water, fat, protein, salt and sugar. The plasma element is

responsible for transmitting blood cells throughout the body combine with nutrients, wastes, antibodies, proteins, hormones and proteins.



Red Blood Cells **Platelets** **White Blood Cells**

Plasma and Blood Vessels

Figure 1.2 : Elements of the Blood

Blood consists of cells, that circulate through the body and has its own functionality and roles. The cells in the blood can be grouped into two main types, namely, Red Blood Cells (RBC) and White Blood Cells (WBC) (Weiqing *et al.*, 2021). RBCs form the major volume of blood cells (~45%) and is named because of its bright red color (Al-Hafiz *et al.*, 2018). The shape of the RBCs is a biconcave disk with a flattened center and has no nucleus. They are also called as erythrocytes. RBCs are produced in bone marrows and transports oxygen to various tissues and organs. The RBC rate in a healthy human body ranges from 40,00,000 to 60,00,000/microliter of blood (Taraconat, *et al.*, 2023).

WBCs, also called Leucocytes, are colourless cells, as it is without haemoglobin. The WBCs are responsible for body immunity and resistance. WBCs in a normal adult human range from 4,500 to 11,000 per microliter of blood (Rahadi *et al.*, 2018). There are two types of WBCs, namely, granulocytes and agranulocytes. Granulocytes are granulated cells and be further classified as eosinophil, basophil and neutrophil. Agranulocytes, on the other hand, do not have granulates in their cytoplasm and can be grouped as monocytes and lymphocytes. Figure 1.3 shows the five types of WBCs present in the blood.

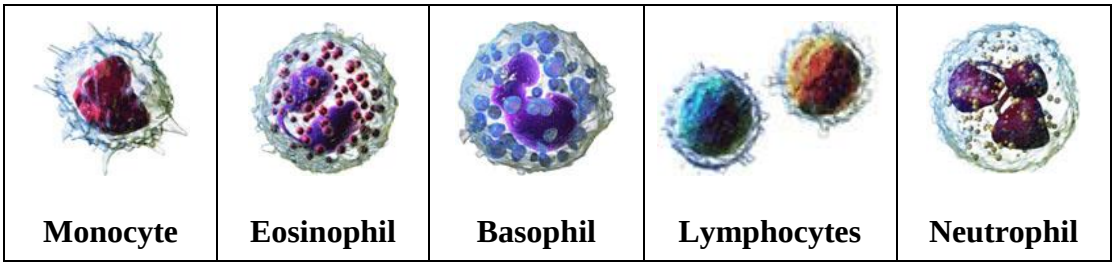


Figure 1.3 : Types of WBCs

The final element of the blood is platelets, also called as thrombocytes. They are specialized blood cells produced from bone marrow and are useful during bleeding or hemorrhage. They are responsible for blood clotting and coagulation during a wound. They range from 1,50,000 to 4,50,000/microliter of blood (Meijden and Heemskerk, 2019).

Increase or decrease in any one or more of these blood elements results in health problems like leukemia, anemia and thalassemia blood disorder (https://www.hematology.org/education/patients/blood-basics). The scope of this research work, as mentioned earlier, is the identification of leukemia. The proposed system uses the microscopic image of the blood smear, where the RBCs in normal blood smear appear as regular, round cells with pale center. Any other variations in size and shape, are identified as blood disorders. In many cases, the blood disorder identified may refer to the detection of Leukemia and hence, requires further careful investigation.

1.2.2. Types of Leukemia

Leukemia is a disease that originates from bone marrow, from blood cells along with platelets can be found. Among the two types of blood cells, Leukemia is detected by examining the WBCs only (Ruberto *et al.*, 2022). It often refers to the condition when a large number of immature WBCs are produced.

Leukemia is generally divided into two types known as acute Leukemia or chronic Leukemia based on the speed with the immature cells or blasts proliferate (Bain, 2010). Acute Leukemia grows rapidly and becomes severe within a short period, while chronic spreads slowly and takes longer to reach the advanced stage. Alternatively, based on the type of the affected WBC, Leukemia can be grouped as myeloid or lymphoid

(Chennamadhavuni *et al.*, 2022). Thus, there exists four types of Leukemia based on WBC and speed of progression (Figure 1.4). The four types are ,

(i) Acute Lymphocytic Leukemia (ALL) (Figure 1.5a),
(ii) Acute Myeloid Leukemia (AML) (Figure 1.5b),
(iii) Chronic Lymphocytic Leukemia (CLL) (Figure 1.5c), and
(iv) Chronic Myeloid Leukemia (CML) (Figure 1.5d).

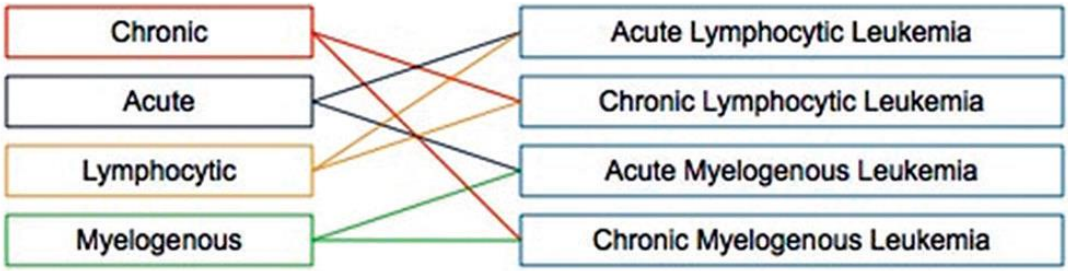
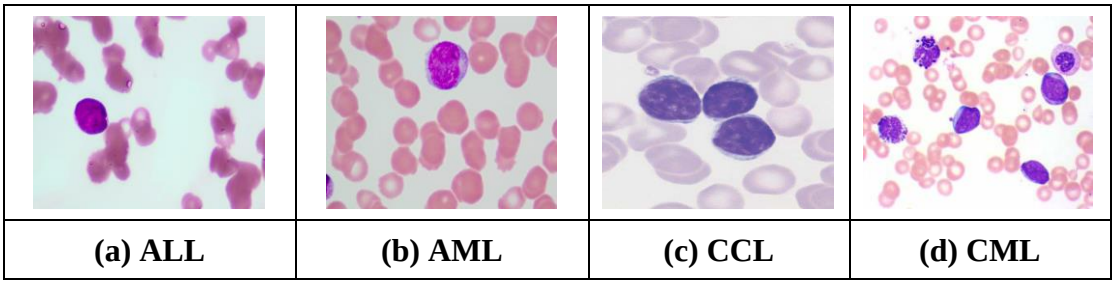


Figure 1.4 : Types of Leukemia



(a) ALL	(b) AML	(c) CCL	(d) CML

Figure 1.5 : Types of Leukemia – Example Microscopic Images

Among the four, ALL is the common type of Leukemia, whose detection and classification are the focal point of this research work.

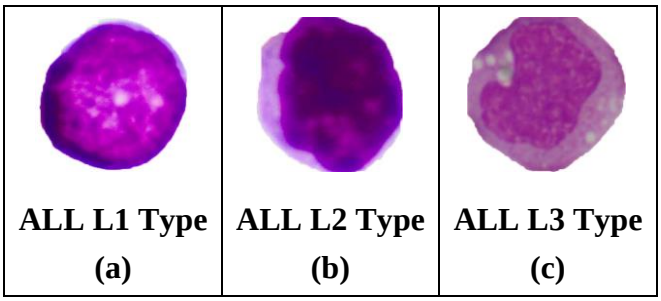
1.3. ACUTE LYMPHOBLASTIC LEUKEMIA

The ALL refers to a type of disease that affects the blood cells. This disease spreads rapidly in the human body and if left undiagnosed and/or untreated, often results in fatal consequences. It is more common with children and adults aged above 50 years (Terwilliger and Abdul-Hay, 2017). It can affect the bone marrow all over the body and can also quickly spread to lymph nodes, liver and spleen. A person with ALL may recover if the disease is identified at an early stage. In the blood smear microscopic image, the ALL appears as a large cell with round to oval nucleus, with coarse, clumped chromatic

and inconspicuous nucleoli and scanty cytoplasm (Figure 1.5a) (http://ilovepathology.com/microscopy-of-leukemia-illustrated)

The ALL has been further classified into three sub-types, namely, L1, L2 and L3, based on the way the appearance of the blood cells in the microscopic image (https://www.cancer.org/cancer/acute-lymphocytic-leukemia/detection-diagnosis-staging/how-classified.html). This classification is based on FAB classification, which was brought forward by a group of Leukemia haematologists in 1970s (Bennett *et al.*, 1976, 1980, 1981, 1985a, b, 1991). FAB classification is widely used as they technically simple, has high diagnostic reliability and are very cost effective.

ALL L1 (Figure 1.6a) are small blasts with little cytoplasm, little cell-to-cell variation. The nucleus is round and homogeneous and chromatin is slightly reticulated with perinucleolar clumping, while cytoplasm is scant blue in color. ALL L2 (Figure 1.6b) can be described as larger cells with greater amount of cytoplasm, greater cell-to-cell variation; irregular nuclei with multiple nucleoli. The nucleus of L2 type is irregular and homogenous with fine chromatin. The cytoplasm of L2 is moderately pale. ALL L3 (Figure 1.6c) are large cells with strong basophilic cytoplasm, often with vacuoles and the nucleoli, which are round to oval shaped and homogeneous, are often in multiples. The chromatin of L3 is coarse with clear parachromatin, while the cytoplasm is moderate blue prominently vacuolated.



ALL L1 Type (a)	**ALL L2 Type (b)**	**ALL L3 Type (c)**

Figure 1.6 : Subtypes of ALL (FAB Classification)

1.3.1. Symptoms

Leukemia can be identified using various common symptoms as presented in Figure 1.7 (Shephard *et al.*, 2016). However, even with these several symptoms, identifying Leukemia is very challenging, as many of them coincide with normal ailments

like viral infection. Hence, careful scrutiny is required during symptom analysis, which can be made easy while using automated systems.

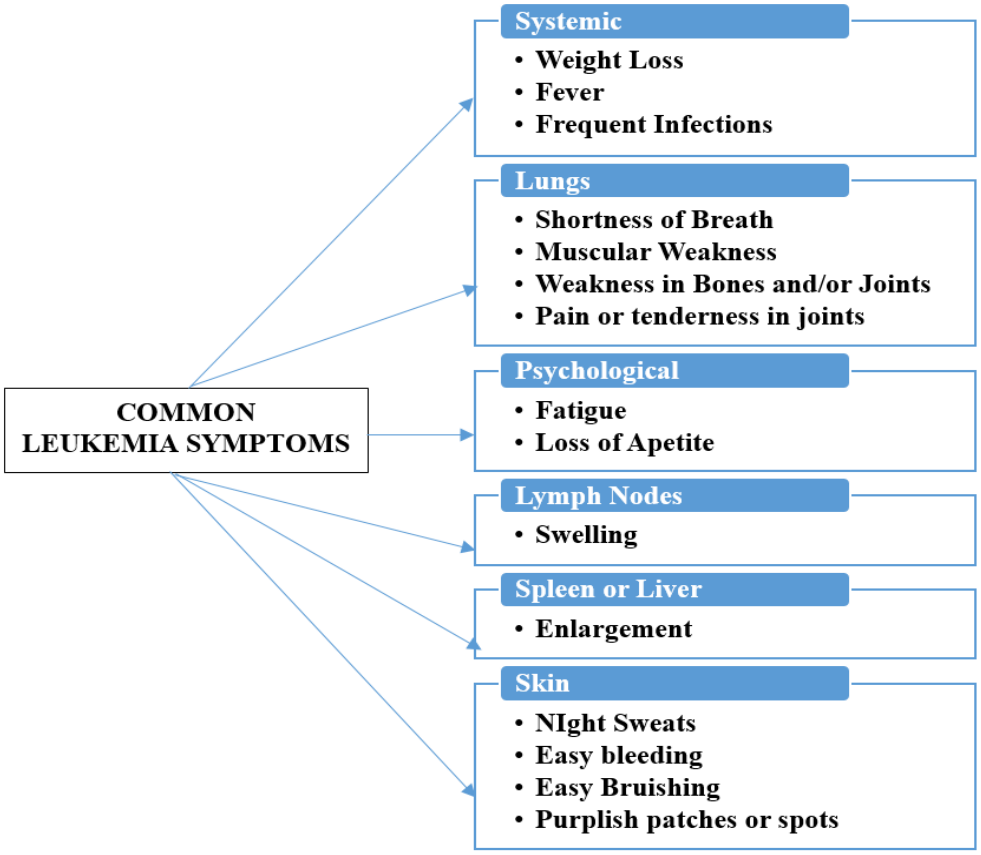


Figure 1.7 : Common Symptoms of Leukemia

1.3.2. Diagnosis Process

ALL diagnosis starts with the physician studying medical history of a patient to understand about the various symptoms and their duration. Physical examination to find abnormalities like bleeding areas, skin rashes and enlarged lymph nodes are also conducted. A complete blood count with differential is needed to find if abnormalities in blood count is present. The count of WBC, RBC and platelets is obtained and a patient is considered healthy if the percentage of WBC satisfies the following (Dean, 2005) :

Neutrophils range 50%-70%,

Eosinophils range 1%-4%,

Basophils range 0%-1%,

Monocytes range 2%-8% and

Lymphocytes range 20%-40%.

If the above conditions are not satisfied by the blood count, then the case is considered suspicious and a microscopic analysis of peripheral blood smear is performed. This analysis finds the presence of blast cells, if any, in the smear. If no evidence of blast cells is found, then the patient is declared cancer free. On the other hand, in case of blast cells are present, the type of Leukemia (myeloid or lymphoid) is determined. With the help of bone marrow test, the percentage of blood blast is determined and if it is greater than 30%, then the patient is free of cancer (DiNardo *et al.*, 2016). Else, Leukemia is confirmed and several other labouratory tests like immunophenotyping, cytogenetic, molecular studies are performed to determine the nature of treatment.

In spite of the above-mentioned advanced diagnosing techniques, the microscopic examination of the blood smear, still remains as the standard technique during Leukemia identification. Frequently, in order to identify malformations in WBCs, a manual inspection of the microscopic image of blood smear is carried out by an expert. Manual inspection and identification has several drawbacks like the following :

- Poorly prepared or stained blood smears,
- Time consuming as the number of available trained personal who can analyze blood smears is less, and thus, may lead to backlog and delay,
- Lack of standardization and inter-observer variability,
- Susceptible to human errors and accuracy depends on the experience and workload of the expert and
- high cost as experts are expensive.

According to Dasariraju *et al.* (2020), the manual investigation has an error rate is in the range of 30%-40% and the accuracy of correct identification is directly related to the experience of the hematologist. Thus, in order to avoid the influence of operator experience and fatigue on Leukemia identification, automated diagnosis systems are preferred. As a consequence, several automated Leukemia detection systems that use blood smear images, to analyze WBCs, have been proposed. These system report the presence of ALL if malformations were detected. Several studies have proved that the usage of automatic systems has increased the accuracy and speed of disease identification (Reem *et al.*, 2022; Sharma *et al.*, 2022). Moreover, automated systems help to avoid/reduce human intervention during diagnosis and are cost-effective.

1.4. AUTOMATIC ALL IDENTIFICATION SYSTEM

ALL-C (Acute Lymphocytic Leukemia Classification) has an important role during treatment plan and treatment. The risk grouping, in case of positive leukemia, requires sub-classification of ALL. The prognosis and treatment outcome is directly related on the correct classification of ALL. An automatic system performs detection in three major steps, namely, preprocessing, identification of WBCs and classification. In general, the ALL-C is designed as a binary classification system, which declares the input microscopic image as either normal or Leukemia present. In the case of malignant blood cells, this step may also classify the malignant blood cell into three of its subtypes, namely, L1, L2 or L3. Thus, an automated ALL-C system works to teach a learning algorithm to recognize malignant regions in a blood smear and identify the type of ALL. Compared with other diagnostic methods, the use of learning classifiers has been the most successful and proved method (Bukhari *et al.*, 2022; Mustafa *et al.*, 2022). The general methodology used by the existing automated ALL-C system, while using Machine Learning Classifier (MLC) and Deep Learning Classifier (DLC), are presented in Figure 1.8a and Figure 1.8b.

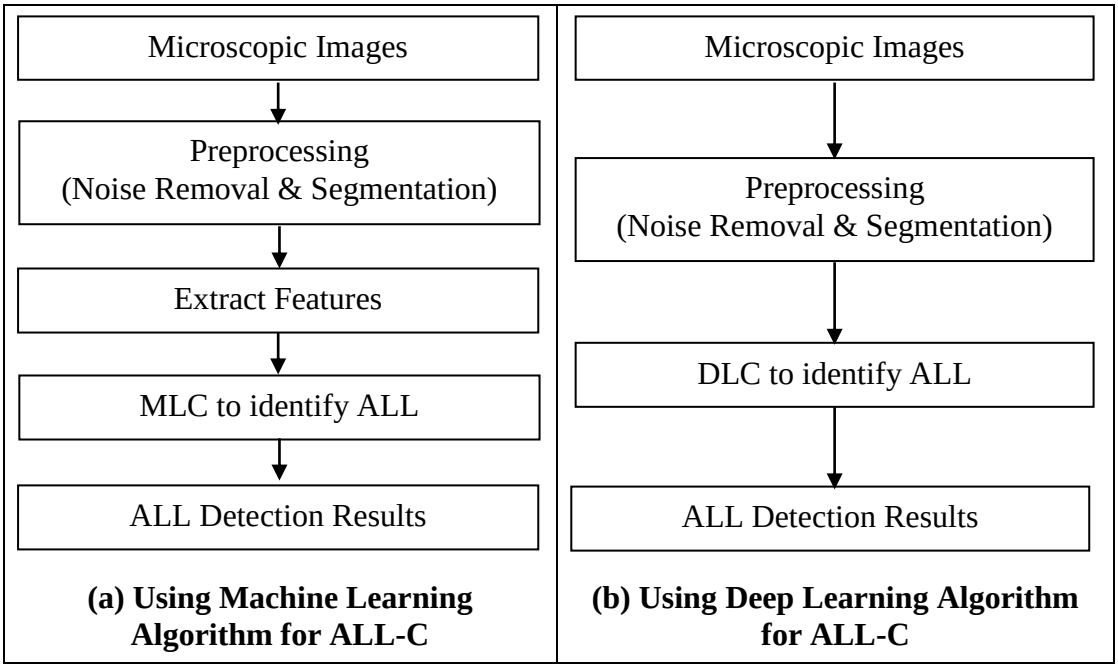


Figure 1.8 : General Steps in ML-Based and DL-Based ALL-C Systems

ALL-C system (both MLC and DLC) starts with enhancement operations that improve the visual quality of the microscopic image. This step improves the quality and

interpretability of the microscopic images and help to improve the performance of the subsequent steps of the automatic ALL-C system, which in turn improve the overall performance of the ALL-C system. Next, the blood cell identification step extracts the region of interest, WBCs, from the preprocessed microscopic image. This step uses segmentation algorithm for this purpose.

With MLC, a third step, performs feature engineering that in turn performs feature extraction to extract significant characteristics of the WBCs identified and then selects optimal features from the extracted set. This step creates a feature vector constructed using the various features extracted. The final step identifies normal and malignant blood cells using a machine learning classifier trained using the constructed feature vector. The success of MLC depends on the optimal feature vector. This dependency is solved with the use of DLC, where the optimal features are identified and used automatically and used to classify the WBCs. The working of MLC and DLC are described in the following subsections along with a brief description on the common step, preprocessing.

1.4.1. Preprocessing

The preprocessing step involves two main tasks, namely, enhancement and identification of WBC. During image acquisition, the microscopic images almost always are degraded by the presence of noise and non-uniform contrast (Laine *et al.*, 2021). Apart from noise, the image edges may also be corrupted. The enhancement step handles all these types of degradations in the microscopic image. The enhancement algorithm include operations that can remove or reduce noise, improve or reduce edge degradation, adjust contrasts. These operations improve the image properties, which help to improve the accuracy of WBC identification.

In the second step, frequently, a segmentation algorithm is used to detect the WBCs (Abhishek *et al.*, 2023). The segmentation algorithm subdivides the microscopic image into several parts that can act as meaningful entities. The segmentation algorithm uses several features like colour intensity to separate the various regions of the microscopic image. With ALL-C, the segmentation algorithms, aims to group the microscopic image into three regions, namely, RBC region, WBC region and background. This section introduces the basic concepts of image enhancement and segmentation algorithms.

(i) Enhancement Operations

The quality of a microscopic image is determined by the imaging method, the characteristics of the equipment and the imaging variables selected by the operator. Image quality is not a single factor but is a composite of three factors, namely, contrast, edge distortions and noise (Hu *et al.*, 2020). Thus, in order to be effective, ALL-C system uses a mandatory step called enhancement. In this research work, three enhancement operations, namely, contrast adjustment, noise removal and edge enhancement are performed in the enhancement step of pre-processing.

(a) Contrast Adjustment

Contrast refers to the difference between the luminance values in the microscopic image (Rahimi-Nasrabadi *et al.*, 2021). It is defined as the ratio of the maximum intensity to the minimum intensity over a microscopic image. The contrast ratio has a strong feature on the resolving power and detectability of the ROI region. When this ratio is large, it is easy to perform segmentation and feature extraction. During acquiring process, sometimes, the image may lack adequate contrast and adjusting contrast is an important part of enhancement step. Improving contrast results with expanded brightness value range consequently expands the density values over a large range. Contrast adjustment algorithms increase the visual contrast between two regions that has dissimilar uniform densities. This can help the classification step to discriminate regions that have very small difference in density.

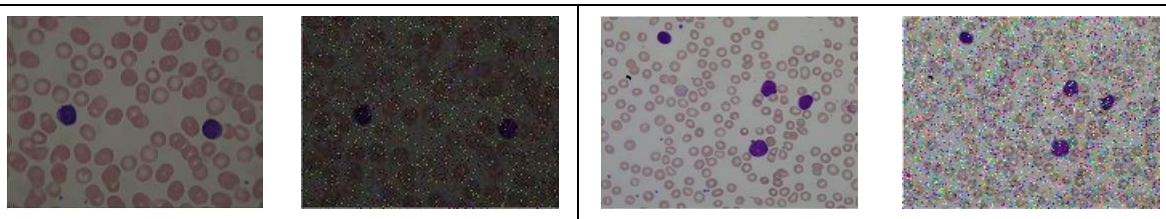
(b) Noise Removal

Noise in microscopic images are defined as the random variations of brightness and/or color characteristics produced by the acquiring device and is considered as an undesirable by-product (https://en.wikipedia.org/wiki/Image_noise). The microscopic image (M), in general, is encoded as a 2-dimensional matrix of gray or color values. The elements of the matrix is composed of a pair of values of the form (i, v(i)), where v(i) is the value of the ith pixel. The pixel i refers to a point in M and v(i) represents its gray level value or triplet values representing the red, green and blue color components.

Shannon and Weaver (1998) found that an image accuracy is limited by the presence of noise perturbation and an observed value in M can be represented as in Equation (1.1).

$$V(i) = v(i) + n(i) \quad (1.1)$$

where, v(i) is the true value of pixel i and n(i) is the noise contaminating M. The amount of n in M is signal dependent and is large when v(i) is large. Figure 1.9 shows some example of microscopic images corrupted by noise.



Original	Corrupted	Original	Corrupted
Microscopic Image with Normal Blood Cells		**Microscopic Image with Cancerous Blood Cells**	

Figure 1.9 : Original and Noise Corrupted Microscopic Images

Presence of noise in a microscopic image can be interpreted as a pattern, which may inversely effect the ALL detection process. Therefore, a noise removal algorithm is an essential task of ALL-C system. The noise removal algorithm require careful handling and should be make sure the algorithm does not remove any significant characteristics or edge details.

(c) Edge Enhancement

Edges in an image represent the objects of an image in a higher level of abstraction and are one of the fundamentally important primitive image characteristics (Rui *et al.*, 2022). It is one of the most vital components of an image and is considered as a colour independent feature that has extreme importance during ALL detection. Edges are the outline of an object in the microscopic image. It is also defined as the boundary between an object and its background. It the defined as the intensity transition of the degraded microscopic image. Edges play an important role during feature extraction and classification. If the edges in an image can be identified accurately, all of the objects can be located and basic features can be measured more accurately.

The usage of edge enhancement algorithms with microscopic images is to highlight the fine details and/or restore blurred edge details. These algorithms sharpen the outline or borders of the various objects in the microscopic image with respect to its background. It is a process that calculates the magnitude at each pixel and enhances the local discontinuities at the boundaries of different objects (edges) in the microscopic image. Several researchers defined edge enhancement as a filter that first identifies sharp edge boundaries, like edge between the object and its background with contrasting color and then increase its contrast around that edge (Zhu *et al.*, 2023). Edge enhancement involves sharpening the boundaries of the various objects in the image with respect to their background. This has the effect of creating subtle bright and dark highlights on either side of any edges in the image, leading the edge to look more defined when viewed from a typical viewing distance.

Edge enhancement algorithms are capable of improving the perceived sharpness or acutance of a microscopic image. Edge enhancement is not completely reversible and therefore significant details lost during edge enhancement is permanently lost. Therefore, an edge enhancement algorithms has to be designed in a careful manner, so that it improves edge quality while not removing important details.

(ii) Segmentation

The second part of preprocessing is segmentation. Segmentation represents the art of partitioning an image into two or more non-overlapping regions of pixels, such that each region is homogeneous and the union of no two adjacent regions is homogeneous, with respect to some characterisitcs of an image (Jyothish *et al.*, 2020). It can also be used to separate objects of interest from the rest of the image. Segmentation can simply be considered as a task that assigns labels to pixels in a 2-dimensional image. Let M be the enhanced microscopic image.

The segmentation is then stated as a problem that determines a set of regions ($S_k \subset M$, k is the number of segments) in a manner that $\cup(S_k)$ produces the whole image M. Ideally, a segmentation method finds those sets that correspond to distinct blood cells or region of interest in the image. The ALL detection is performed by analyzing the WBC in the microscopic image and there segmentation step is considered as an essential process in ALL-C. An example of segmentation used to detect WBC is shown in Figure 1.10.

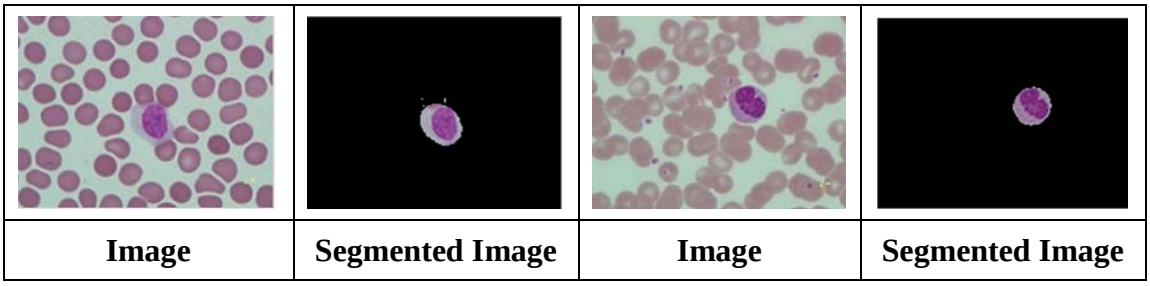


Image	Segmented Image	Image	Segmented Image

Figure 1.10 : Example of Segmentation of Microscopic Image

1.4.2. MLC to Identify ALL

Identification of ALL using MLC consists of two steps, namely, feature extraction and classification.

(i) Feature Extraction

An efficient ALL-C system requires a set of quality features that can best describe the microscopic image. Feature vector extraction is highly responsible for the success of the subsequent step of ALL-C, that is, MLC. The main motivation of using feature extraction is to obtain a microscopic image representation that is compact when compared to the original image.

Kalashami *et al.* (2022) defined feature extraction as an abstraction or compact form of the microscopic image and achieves significant compression on image representation. On the other hand, https://en.wikipedia.org/wiki/Feature_extraction, describes feature extraction as a process that builds an initial set of measured data (called as feature vector) using values derived from the microscopic image and which exhibit the following properties:

(i) The values are both informative (that is, relevant) and non-redundant ,
(ii) The values facilitate subsequent learning and generalization steps,
(iii) The values lead to better interpretations, and
(iv) The values best represent the microscopic image and whose size is much smaller than the original image.

Kumar and Bhatia (2014), viewed feature extraction as a part of dimensionality reduction, since the algorithm reduces the number of pixels to be analyzed and this smaller

set is more convenient and manageable. Alternatively, feature extraction can be considered as a technique that identifies significant characteristics or attributes of an image (Latif *et al.*, 2019). The quality of the feature vector is defined as a set that has maximum discriminating power and has high ability to differentiate between different regions of the microscopic image into different classes. A set of features are said to be good features, if the regions from the same type of cells have similar feature vector, while regions from different type of cells have entirely different set of vector.

(ii) Classification

The classification algorithm is used to determine the presence and absence of ALL. All the preceding steps, namely, preprocessing, segmentation and feature extraction, were fine tuned to improve the accuracy of this step. The classification step is a process that simply transforms the quantitative input feature vector to a qualitative output. The output of the classifier is a discrete selection of one of the pre-defined target classes. The classification, also called as supervised learning algorithm, prediction algorithm, recognition algorithm, is defined as a task that involves construction of a procedure that matches the input feature set into any one of the pre-defined target class (Sarker, 2021). Figure 1.11 presents a basic classification model and consists of three steps :

(i) Defining training and test sets,

(ii) Creation of the classification model, and

(iii) Classification of new/unknown documents.

Along with feature extraction, the next crucial step is the classification algorithm that determines the presence and absence of ALL. All the preceding steps, namely, preprocessing, segmentation and feature extraction, were fined for improving the accuracy of this step. The classification step is a process that simply transforms the quantitative input feature vector to a qualitative output. The output of the classifier is a discrete selection of one of the pre-defined target classes. The classification, also called as supervised learning algorithm, prediction algorithm, recognition algorithm, is defined as a task that involves construction of a procedure that matches the input feature set into any one of the pre-defined target class (Montejo-Raez, 2005).

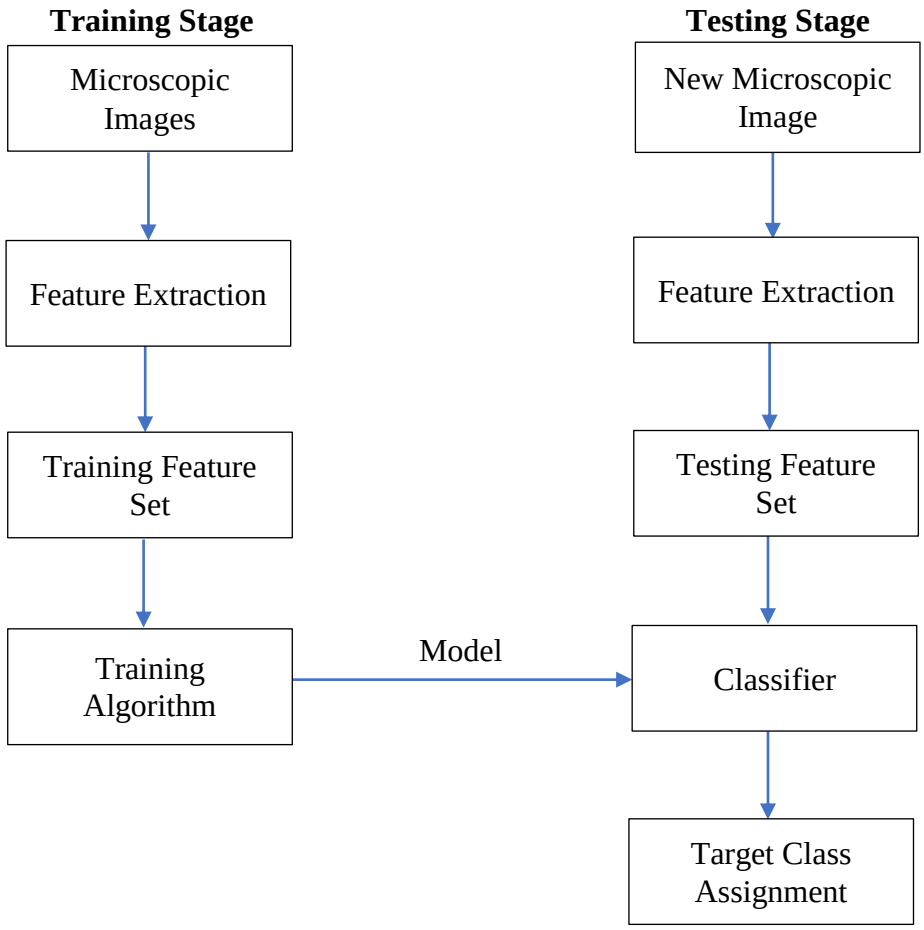


Figure 1.11 : Schematic of ALL Detection Using Machine Learning Classifier

The first step in the design of a classifier starts with partitioning of a feature set, F, into two sets, namely. training set (Tr) and testing set (Te). A constant x is used to indicate the percentage of features that will be used for training and testing tasks. Care should be taken to build the training set size big when compared to the size of the testing set. For example, when x = 80, 80% of feature dataset is used to train the classifier, while the rest 20% is used for testing.

A classifier can be designed in three ways, depending on the type of output required by an application. They are, binary case, multi-class case and multi-label case (Er *et al.*, 2016). The binary case classification classifies samples into exactly two predefined classes. In case of ALL detection, the classifier has to determine to which of the two sets the new microscopic images belongs, namely, normal or cancerous. In mutli-class case, a microscopic image belongs exactly to just one target class of a set of 'm' classes. For

example, the microscopic image may belong to Normal, L1, L2 or L2 class. With the multi-label case, the new image may belong to several classes at the same time, that is, classes may overlap. For example, through very rate, a patient might get two primary cancers at the same time. Several classifiers have been used to classify leukemia, each using different methods to learn and classify input data. The choice of classifier depends on how the learning algorithm best fits the relationship between the feature vector and class label of the input feature set.

In the final step, the classifier designed is used to classify a new microscopic image into one of the pre-defined target label, using the knowledge learned from the training set.

1.4.3. DLC to Identify ALL

Usage of deep learning methods to analyse medical images is gaining wide popularity because of its high performance (Das *et al.* 2022). Several researches that compare the performance of DLC with human experts and have proved that DLCs outperform human expertise have been conducted (https://en.wikipedia.org/wiki/Deep_ learning). This new machine learning paradigm has also been proved to work better than machine learning in any applications including image processing, speech analysis and medicine (Nazari *et al.*, 2020).

DL is a subset of ML that is concerned with algorithm inspired by the structure and function of the brain called Artificial Neural Networks (ANN) (Figure 1.12). Artificial intelligence makes the applications to perform tasks that require human intelligence to complete. On the other hand, ML gives the applications learning ability without explicitly programmed. DL are ML algorithms that have a logical structure similar to human brain and can also learn automatically with explicit programming.

DL, according to Patterson and Gibson (2017), is a NN having a large number of parameters and layers uses a hierarchical non-linear structure in multiple layers to extract features, perform transformation and classification. The main advantage of DL is that it can automatically extract features which transforms low-level features to high-level abstraction (Schmidhuber, 2015) and can integrate small momentary and indirect changes that can improve classification accuracy (Rost and Sander, 1994).

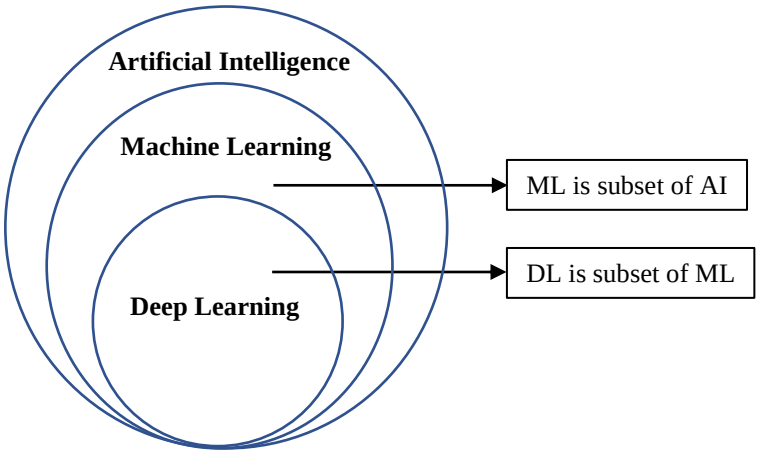


Figure 1.12 : Position of Deep Learning

It has three or more layers, designed to optimize classification accuracy (https://machinelearningmastery.com/what-is-deep-learning). DLC uses mathematical functions to map input feature to any of the pre-defined target label. The mathematical functions have the ability to extract non-redundant features from microscopic images and, during training, enable them to build a relationship between input and output. A deep learning uses multiple layers to progressively extract features from microscopic images. Almost all DL methods use ANN architecture, and hence the name deep NN. Here, deep refers to the number of hidden layers in the NN. In ML scenario, the number of hidden layers is normally 2-3, while with deep NN it might go as high as 150 (https://in.mathworks.com/discovery/deep-learning.html). In DL, the NNs are organized in layers, with each layer having interconnected nodes. An example of such a Deep NN is shown in Figure 1.13.

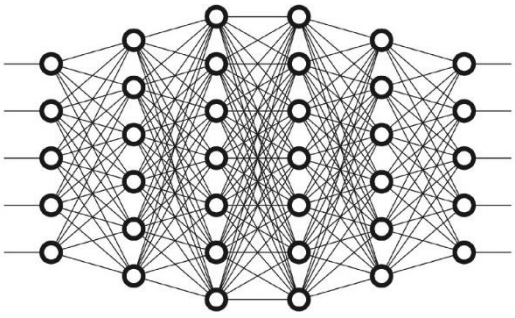


Figure 1.13 : Deep Neural Network

There are several DL architectures, which include deep neural networks, recurrent neural networks, conventional neural networks and deep belief networks (Alzubaidi *et al.*,

2021). All these network use input, multiple hidden and output layers. The importance of an input feature is determined using a weight which are associated with the connection between neurons. The neurons use an activation function on the features to standardize the results from the same neurons. DL iterates through the feature set and compare the results using a cost function, which indicate how much the Artificial Intelligence (AI) deviates from the original results. After each iteration, the weights are adjusted using gradient descent automatically, which also updates and reduces the cost function. Finally, the last layer compiles the weighted input to produce the classification result.

The working of machine learning and deep learning are different in several manners (Almadhor *et al.*, 2022). Figures 1.14 pictorially demonstrate the difference between MLC and DLC and a detailed description of the difference is consolidated in Table 1.1.

1.5. ISSUES IN EXISTING ALL-C SYSTEMS

As seen in previous section, the ALL-C system uses three steps, preprocessing, feature extraction and classification, during classification. In order to have space for betterment of any system, there is always room for improvement. With this in mind, the following points were identified as places which if addressed carefully can be used to further improve ALL-C performance.

- Lack of systems that operate efficiently on the extremes of three requirements simultaneously, namely, high accuracy, scalability and ease of usability,
- Lack of techniques to deliver the required image quality for ALL identification because of image degradation,
- Lack of methods that extract optimal features that have maximum discriminating power to identify the presence and absence of ALL efficiently, and
- Lack of single framework with techniques that focus on improving the performance of various steps of ALL identification system, so as to ensure more accurate identification. Existing solutions use off-the shelf conventional algorithms to perform identification, which may have various issues that are not addressed properly.

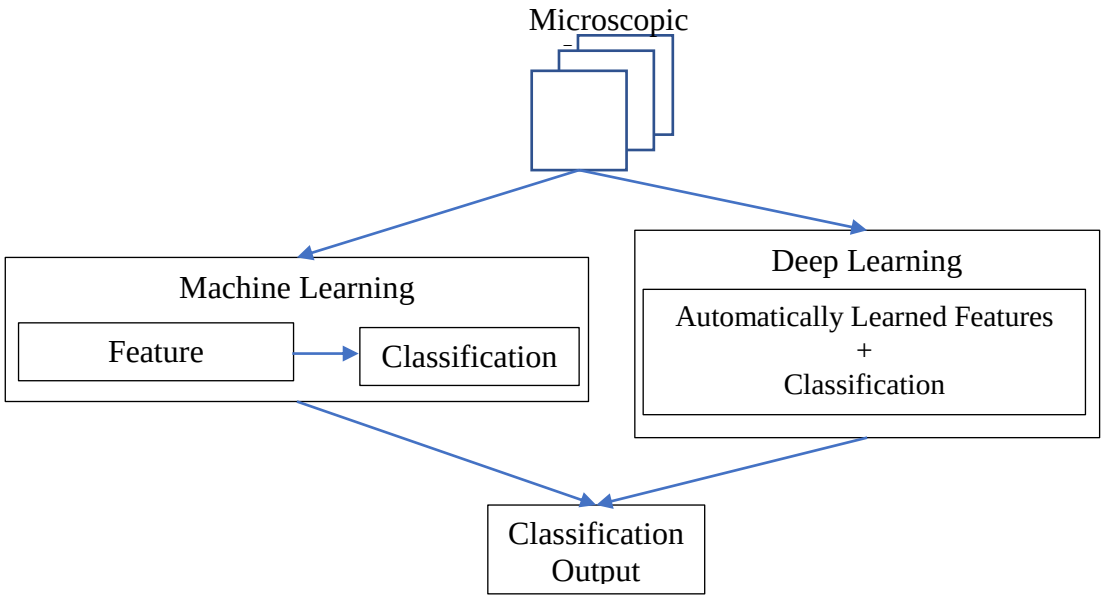


Figure 1.14 : Difference Between MLC and DLC

TABLE 1.1

DIFFERENCE BETWEEN ML AND DL

S.No.	ML	DL
1.	Superset of DL	Subset of ML
2.	Data representation uses structured form	Data representation is different as it uses ANN
3.	Evolution of AI	Evolution of ML
4.	Use various algorithms that automate the various functions of the model	Use NN that iterates data through processing layer in order to identify, interpret data features and its relations
5.	Algorithms are identified by data analysts to determine attributes in datasets	Algorithms are self-depicted on data analysis
6.	Needs human intervention	Require less human intervention
7.	Simple to implement and their effectiveness may be constrained	Complex, but can produce results immediately
8.	Less training time	High training time
9.	Manual feature engineering	Do not require feature engineering, as features are automatically detected by NNs

This research work designs an ALL identification and classification system from microscopic images as an answer to the above issues. The proposed system is designed using enhanced image processing and classifiers.

1.6. MOTIVATION

Increasing ALL incidence makes automated systems that can identify and predict ALL has become essential (Kumar *et al.*, 2022). These systems are much preferable as they are required to offer patients correct care and reduce risks. The treatment is heavily dependent on the timely and accurate detection where the use of classifiers are very powerful. Detection of ALL is a highly sensitive issue as it associated with both health and life of humans. Early detection of ALL can open door to successful future care, treatment and can offer best chance of cure.

Increase in the number of patients with ALL systems increases the size of data that has to be examined and analylzed, and identifying an appropriate and effective system that can accurately analyze this growing data is very cumbersome. Manual examination and analysis of data takes longer duration and poses several limitations in the diagnosis ability, which increases the need for automated systems.

Classification algorithms that predict and classify ALL and its stage are popular scientific area that is based on the concepts of AI (Kumar *et al.*, 2022a, Lalotra *et al.*, 2021). Several researchers have proposed MLC and DLC as solutions to early detection of ALL. Moreover, as described in the previous section, automatic systems consists of various steps and several solutions exist for performing these steps. However, most of the solutions present analyze the algorithms separately, and only very few studies have focused on the problem of finding methods on how to integrate the best algorithms of each step, in order to achieve better solution to ALL classification.

Moreover, the system performance is often degraded by the quality of microscopic images. Similarly, the feature vector used to train the classifier, also have high connection with the classification performance. Identifying correct set of features with maximum discrimination power, low redundancy and maximum relevancy is very difficult. Apart from this, the size of the feature set has a direct impact on the classifier time complexity.

Inspite of several proposals to improve the performance of ALL-C systems, it could be understood that there is still room available for research for the improvement of Leukemia detection accuracy. The ALL-C system should be designed in a manner that it replaces manual operators and analysis completely without any difficulty. The current scenario, also points out that very few works have focused on the classification of ALL subtypes (L1, L2 and L3). It is difficult to detect and classify L1, L2 and L3 because of its high intraclass variability and interclass similarity. It is very important to diagnosis correctly, as ALL subtype plays a significant role during treatment plan.

All the above challenges motivated this research work to focus on providing solutions that can enhance the process of ALL identification and its subtype classification.

1.7. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

Let the template dataset (denoted as D) have M microscopic images collected from N persons. Let m_i refer to the m^{th} microscopic image from the i^{th} person. Let $\oplus$ be the enhancement operation, $\wp$ be the segmentation operation and ς be the optimal feature vector generation operations. The research problem is stated as follows.

"To design an Automatic ALL Classification System of the form

$$\Re' = DLC + MLC$$

where $\Re'$ is the output indicating the classification of ALL using hybrid enhanced DLC and enhanced ensemble MLC. The classification operation, $\Re$, is of the form

$$\Re = \oplus(m_i) \rightarrow \wp(m_i) \rightarrow \varsigma(m_i) \rightarrow \Re'$$

where $\rightarrow$ denotes the sequential application of the three operations. The classification output is any one of the pre-defined target label set {Normal, L1, L2, L3}."

To accomplish this problem statement, the primary research objective was set to strengthen the clinical decision support system by designing an automatic system that enhances the operation of each step involved during ALL-C in order to increase the overall accuracy and speed of Leukemia classification. The specific objectives designed to meet this primary objective are

- To design and develop preprocessing techniques that enhance the visual quality of the microscopic images and identify the white blood cells in them.
- To design and develop an enhanced ensemble machine learning classifier to improve the process of ALL classification, and
- To design and develop a deep learning classifier that is combined with ensemble SVM machine learning classifier.

1.8. LAYOUT OF THE THESIS

The main objective of this research work is to develop automatic ALL detection and classification system and this chapter (**Chapter 1 – Introduction**), presented the underlying basic concepts behind Leukemia with emphasis to ALL-C along with the research objectives. The rest of the thesis is arranged as below.

Literature study is conducted to understand the current status of the research topic. In the domain of cancer research, several studies have addressed the problem ALL detection. A critical look at these available literatures connected to ALL detection and topics related to ALL are given in **Chapter 2, Review of Literature.**

Chapter 3, Methodology, presents the overall description of the research methodology used by the research work and introduces the various techniques proposed at the various stages of ALL-C. As mentioned previously, the ALL-C system, consists of three main steps, namely, preprocessing, WBC identification and cancer classification. The first step, preprocessing, deals with two tasks, namely, enhancement and segmentation (Identification of WBC). Detailed description of the algorithms proposed to perform both these tasks are presented in **Chapter 4, Design of Preprocessing Algorithms**.

The most important step of ALL-C is classification. This research work proposes two classifiers. The first classifier used is the SVM classifier which is enhanced in its working and the second classifier is the CNN deep learning classifier which is also enhanced in its working. **Chapter 5, ALL-C System Using Machine Learning Classifier**, presents details regarding the enhancement algorithms incorporated with SVM to improve ALL detection. **Chapter 6, ALL-C System Using Deep Learning Classifier**, describes the CNN classifier and the various performance improving algorithms proposed to improve the CNN classifier.

The algorithms proposed in each step of the ALL-C system and their effect on ALL detection was evaluated using various performance metrics and a standard acute lymphoblastic Leukemia image database. The performance evaluation results are presented in **Chapter 7, Results and Discussion.** The research work is summarized and concluded with future research directions in **Chapter 8, Summary and Conclusion**. The work of several researchers are quoted and used as evidence to support the concepts explained in this dissertation. All such evidences used are listed in the **Bibliography** section of the dissertation.

1.8. CHAPTER SUMMARY

Owing to the importance of the healthcare domain, several research works can be found in the literature that focus on cancer prediction. As the focus of this research work is ALL detection and classification, it is necessary to have a clear understanding of the existing systems and their working. The following chapter (Chapter 2, **Review of Literature**), presents the result of the literature survey conducted to better understand the research domain area.