Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 10
October 2023

Leukaemia/ Blood Cancer Classification, Detection, And Evaluation

Amrut Jadhav
Assistant Professor, Department of Computer Application, Institute of Management and Entrepreneurship Development, Pune-411038, Bharati Vidyapeeth (Deemed to be University), Pune, Mahrashtra, India,
amrutjadhav.bvdu.imed@gmail.com, ORCID-https://orcid.org/0009-0002-4603-396X

Abstract—
Leukaemia is a form of blood cancer that begins in the bone marrow and causes the body to make abnormal blood cells. Acute leukaemia is a kind of blood cancer that is linked to bone marrow malfunction. Young people and children have a higher risk of getting it. Due to the fast cell growth, this kind of leukaemia produces, prompt treatment is essential. The four most frequent types of leukaemia are acute lymphoblastic leukaemia, acute myeloid leukaemia, chronic lymphocytic leukaemia, and chronic myeloid leukaemia (CML). Acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) are the two most common kinds of leukaemia, and they are also the leading causes of mortality from the disease (AML).The computerized approach may help doctors decide on the best course of action by classifying these two forms of leukaemia on pictures of blood slides. This research aims to demonstrate the design of a convolutional neural network (CNN) that can differentiate between blood slides such as those showing ALL, AML, and healthy blood (HBS). A total of 2,415 images from 16 datasets were used in the studies, which achieved an accuracy and precision of 97.18 and 97.23 percent, respectively. The suggested model's efficacy was measured against that of state-of-the-art methods like those based on CNNs. The project's other objective is the development of a system for identifying and categorizing leukaemia. Because of the vast structural differences between leukemic and normal cells, several features are retrieved from the segmented lymphocyte pictures for identification purposes.

Keywords—leukaemia diagnosis, convolutional neural network, computer-aided diagnosis. Comparison of Machine Learning Algorithms, Leukemia Diagnosis, Leukemia Classification, Machine Learning.

I. INTRODUCTION

The uncontrolled growth of aberrant cells that can metastasize to other organs is what we mean when we talk about cancer. It's presently a major killer across the globe. Ten million people will lose their lives to cancer in 2020, while another 19.3 million will be diagnosed, according to research. Different types of cancer have different mortality rates. In 2020, the mortality rates for lung cancer (18.0%) and colorectal cancer (9.4%) will be much higher than those for breast cancer (8.3%), liver cancer (7.5%), and stomach cancer (6.9%). Blood malignancies account for over 10% of all newly diagnosed cancer cases. It has been widely assumed that by early identification and prediction, cancer mortality rates may be lowered globally. Aspects of blood cancer prognosis are the focus of this investigation. According to data collected and analyzed by the Leukemia & Lymphoma Society, about 1,290,773 persons in the United States are now afflicted with a blood malignancy. Myeloma, leukaemia, lymphoma, and myelodysplastic syndromes are all forms of blood cancer. Blood cancers, to be specific, target not only the bone marrow, lymph nodes, and blood cells, but the whole lymphatic system as a whole. Recently, medication has been developed to assist the immune system in recovering from illness and fighting cancer cells. Previous studies have employed a variety of models and algorithms for predicting blood cancer, with mixed results. The accuracy, specificity, and sensitivity of support vector machines (SVM) utilized by Goutam et al., for example, were 85.74, 80%, and 100%, respectively. Using H20 deep learning, the study achieves a 79.45 percent accuracy rate. In addition, Vijayarani and Sudha obtained accuracies of 78%, 75%, and 86%, respectively, while employing K Means, Fuzzy Means, and Weighted K Means. Using k-nearest neighbours (KNN), support vector machines (SVM), decision trees (DT), random forests (RF), and gradient boosting decision trees, Xiao et al. achieved an accuracy of 99.20 percent, 98.78 percent, and 98.51 percent, respectively. But Subhan et al.8 used KNN and the Hough transform to get 93% accuracy. A total of 84%, 74%, and 81% accuracy were attained by Gal et al. when they used KNN, SVM, and RF classifiers, respectively. The efficacy of machine learning and deep learning classifiers in predicting blood cancer has not been as good as was hoped.

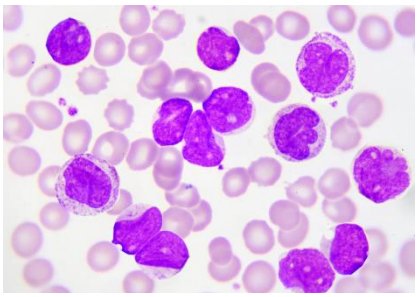


Fig. 1. Microscopic view of Leukaemia

Early detection and identification of disease significantly improve prognosis and treatment outcomes. Because photos are cheaper, don't require expensive testing and lab equipment, and deliver findings more rapidly than manual testing, image processing is a better and easier way for identifying leukaemia. Cancer of the white blood cells is called leukaemia, however it can occur in any blood cell.

Leukemia, or cancer of the blood cells, can be either chronic or sudden in onset (chronic).There are various subtypes of leukaemia, each with its outlook and treatment options. Following are the four most common types of leukaemia. Oncologists at Cancer Treatment Centers of America (CTCA) have years of expertise with cancer staging, diagnosis, and creating individualized treatment programmes for each subtype of leukaemia.

The rates of advancement and the locations at which the disease develops are the fundamental variations between the four main kinds of leukaemia. Because they never fully develop into adult lymphocytes, "chronic" leukaemia cells are less effective than healthy lymphocytes at fighting off infections. Cell division in "acute" leukaemia begins before the immune system has had time to mature.

These items are:

1. Acute myeloid leukemia (AML)
2. Chronic myeloid leukemia (CML)
3. Acute lymphocytic leukemia (ALL)
4. Chronic lymphocytic leukemia (CLL)

Acute myeloid leukemia (AML): Acute myeloid leukaemia (AML) is by far the most prevalent kind of acute leukaemia, yet it is also the most treatable. Blasts, which are immature cells, are produced in the bone marrow. These blasts mature into white blood cells under typical settings. Due to the absence of normal development of these cells, AML renders the body vulnerable to infection. In AML, the bone marrow may be the source of the aberrant platelets and red blood cells. Leukemia cells outweigh healthy white blood cells, red blood cells, and platelets due to their fast multiplication. One of AML's defining features is that it may be further subdivided into eight distinct subtypes based on the cell type from which the leukaemia originated, making it distinct from the other major forms of leukaemia.

Chronic myeloid leukemia (CML): Cancer of the blood and marrow, known as chronic myeloid leukaemia (CML), only manifests in these two organs. This process originates in the bone marrow, where new blood cells are created, then travels throughout the body. The illness spreads from one organ to another over time. Because of its classification as chronic, this kind of leukaemia tends to metastasize and progress very slowly. However, chronic myeloid leukaemia (CML) can progress into an acute type of leukaemia with rapid development and impact nearly every organ in the body. CML stands out from the other three main types of leukaemia due to its unique set of characteristics. The Philadelphia chromosome is a kind of chromosomal aberration that has been associated with chronic myeloid leukaemia.

Acute lymphocytic leukemia (ALL): Acute lymphocytic leukaemia (ALL) is a kind of leukaemia that develops when white blood cells in the bone marrow multiply uncontrollably (leukaemia cells). Rapid replacement of healthy cells responsible for producing functional lymphocytes by immature leukaemia cells is a hallmark of the rapid course of ALL. Leukaemia cells can travel via the bloodstream and proliferate in distant organs. The brain, liver, lymph nodes, and testes are all included in this category. Various symptoms might develop as a result of these leukaemia cells growing, dividing, and spreading. Having more B lymphocytes than T lymphocytes is a common risk factor for Acute Lymphoblastic Leukemia (ALL). Both B and T cells work hard to keep the body healthy and free of sickness and infection by identifying and destroying harmful microorganisms and defective cells. When it comes to protecting against pathogen invasion, B cells are particularly helpful, as T cells are responsible for eradicating the infected cells.

Blood and bone marrow are two of the places where acute lymphocytic leukaemia, a kind of malignancy, can manifest (ALL). Blood cells are produced in the marrow, which is the spongy tissue in the heart of the bones. For children, the most frequent kind of malignancy is acute lymphocytic leukaemia, also known as acute lymphoblastic leukaemia. Adults are also vulnerable to acute lymphocytic leukaemia, but their prognosis is significantly worse. The rapid progression of the disease and the formation of immature blood cells rather than mature ones are what give acute lymphocytic leukaemia its name.

Chronic lymphocytic leukemia (CLL):

Lymphocytes in the bone marrow are the origin of chronic lymphocytic leukaemia (CLL), a malignancy that can spread to the blood. Most of the time, its rate of growth is as slow as molasses. The liver and spleen, among others, might be impacted along with the lymph nodes. When the number of aberrant lymphocytes in the body reaches to harmful levels, the body's ability to fight off infection is compromised. Over time, a chronic condition gets worse. The proliferation and development of the aberrant cells are slowed down. This means that the severe manifestation of an illness like CLL may not occur for some time. In contrast, acute lymphocytic leukaemia (ALL) tends to advance rapidly. Rapid progress toward a cure for ALL is possible. Common symptoms in children include fever, pale complexion, enlarged organs, and easy bruising. After a bone marrow injury, immature white blood cells replace the normal bone marrow cells, resulting in a shortage of blood platelets, which are necessary for the clotting process. It has been suggested that leukaemia patients are more likely to sustain injuries and bleed than healthy people (petechiae).Disease-fighting capabilities of the body might be compromised by suppressed or faulty white blood cells. Overuse of the immune system can lead to its exhaustion, at which point it no longer effectively defends the body against even the most minor of infections, or it might begin to target healthy cells in

the body. Because leukaemia hinders the immune system's normal functioning, certain individuals with the disease are subject to recurring infections, such as infected sores in the mouth or diarrhea, as well as potentially lethal pneumonia or opportunistic infections. Lastly, a shortage of red blood cells causes anemia, which can cause fatigue and shortness of breath.

II. LITERATURE REVIEW

Mughal et al. (2018) Learn how to use a discrete differentiation operator to modify mammography in a new way and get rid of the pectoral muscle. The recommended method is evaluated using the standard MIAS dataset, which consists of 322 mammograms and 20 contrast-enhanced digital mammographic photos, in order to achieve high accuracy in variable pectoral muscle size.

Bibin, Nair, and Punitha (2017) Create a system that uses a Deep Beliefs Network to detect malaria parasites. The primary objective is to single out parasites as an independent species. The authors conclude having developed a model for the HSV colour space conversion, segmented the cells using region-based contours, and classified DBN based on colour and texture attributes.

Jayoti et al. (2017) The nucleus of the blast cell may be removed from the preprocessed picture by employing morphological opening, histogram equalization, and a global threshold. Recovered geometric, colour, statistical, and other data are classified using PCA-kNN, PCA-PNN, PCA-SVM, and PCA-SSVM classifiers. PCA-ANFIS may achieve a precision of up to 97% when used in a tree-like structure. When using too many classifiers in tandem, however, the classification process becomes unacceptably sluggish.

Vidhya. K, et.al. (2015) After K-means clustering had broken the problem down into more manageable pieces, data was categorized using SVM and features retrieved from the Local Directional Pattern (LDP).

Amin. K, et.al. (2015) cluster lymphocytes using k-means, classify acute lymphoblastic leukaemia (ALL) and its subtypes using extracted geometric and statistical characteristics, then apply a multiclass support vector machine to label the collected data (SVM).The author claims 97% accuracy for the categorization, although this is predicated only on nucleus features.

Goutam and Sailaja (2015) Acute myeloid leukaemia cells may be isolated from grayscale pictures by utilizing k-means clustering to isolate and isolate nuclei from normal cells. When combined with SVM, LDP's textual feature extraction yields a 98.1% classification accuracy.

Bhattacharjee and Saini (2015) suggest employing watershed transforms for segmentation, followed by morphological operations, as a means of spotting cases of acute lymphoblastic leukaemia. The suggested technique uses a hidden Markov model (GMM) and binary search trees to extract morphological attributes for classification, and it has a 95.56 percent success rate.

Mahopatara et al. (2014) Blood smear photographs may be analyzed to pinpoint the area of interest with the help of k-means clustering. After extracting RGB color features from the whole image, the technique segments L*a*b* (CIELAB) using the leukocytes shadowed C means methodology. Next, we apply SCM clustering in the color space to deconstruct the sub-image (nucleus, cytoplasm, and backdrop).

Mughal. Muhammad, et.al. (2017). We use morphology, texture, and colour to extract, normalise, and select characteristics. Some examples of classifiers are the Naive Bayesian, K-nearest neighbour (KNN), Multilayer Perceptron, Radial Basis Function Neural Networks, and the Support Vector Machines (SVM). Finally, it has been found that the classification of mature cells and lymphoblasts has an accuracy rate of 94.73 percent.

Kulkarni-Joshi and Bhosale (2014) Detected ALL explosions and differentiating nuclear properties using a thresholding-based technique is suggested. Otsu thresholding is used to reclaim shape attributes for blast identification after the background has been removed.

Abbas and Mohamad (2014) Give a technique for separating the nuclei of lymphocytes that can be utilized to identify cancer in the blood. After convolving the image with a 226 mask to decrease the image's high RGB values, nuclei are first extracted using the Otsu method. It is possible to diagnose leukaemia in 96.5% of cases by segmenting the nuclei, which involves eliminating tiny sections and then expanding the nuclei.

Again. M (2014) Converting the picture to CIELAB and analyzing the L and components are advised for detecting myelogenous leukaemia in the blood. Next, we utilize k-means clustering to create subregions inside the study area. Extracting color, shape, and texture features with qualities associated with cell energy and Hausdorff dimensions, SVM is used to classify dangerous bursts. Approximately 98 percent accuracy was found.

Jagadeesh, et.al. (2013) offer a means of recognizing malignant cells in blood samples. At first, the picture is transformed into grayscale and binary. For further refinement and correction of any distortions, morphological erosion, morphological closure, and a distance map between black and white pixels are subsequently used. The watershed transform is used to create the segments. SVM uses geometric, statistical, and textured features to categorize data. Joshi, Karode, and Suralkar (2013) use the Otsu threshold technique for segmentation, and KNN for classification, based on features related to texture and shape.

Hayan et al. (2012) Binary representations of RGB images are created for use in the H and S bands. Fifteen disk-shaped structuring components are used to expand the H band while dampening the S-band. Last but not least, a morphological operator is

used to reconstruct the images to classify the blast cells according to their centroid and axis length. This method easily segments and localizes the lymphoblasts with a hundred percent success rate.

Nee et al. (2012) Identification and segmentation of edges are essential steps in the watershed transformation, which may be accomplished with the help of the S component of the HSV colour model, as well as erosion, magnification, and magnification gradient. Acute myeloid leukaemia and its subtypes are the only ones for which this method, which has an accuracy of 94.5, is acceptable.

Pan, et.al. (2012) Leukocytes may be separated using the ELM with a sampling threshold set at the pixel with the highest gradient. Then, the greatest entropy is used as a criterion for how to separate the colorful object. Before applying the Otsu method to the cytoplasm, we first witness the edges classifying on ELM, then convert the image to HIS for leukocytes, and lastly notice the edges classifying on ELM.

Abd Halim, et.al. (2011) Finding a new approach to the segment nucleus problem is necessary for distinguishing between acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). RGB images can have their global contrast increased to better highlight the classification area. The photos are then altered to use the HSI color standard. Accurate segmentation of the nucleus was achieved by applying S component processing to the color space and utilizing a threshold. The detection of nuclei and blasts employs region-growing methods.

Rezatofighi, et.al. (2011) Make a three-step process that can automatically identify five distinct kinds of blood cells. The nucleus is segmented using the Gram-Schmidt technique, and then basophils are identified using a co-occurrence matrix of characteristics and LBP. Once the images have been converted to grayscale using the S component of the HIS color model, the remaining four categories of WBC may be recognized using morphological criteria with an accuracy of 93.09%.

Sadeghian, et.al. (2009) Grayscale images can be converted into sub-images of WBCs by isolating leucocytes from the rest of the blood's components. The positions of individual atoms are calculated using a gradient vector flow model. After zack thresholding is used to distinguish the nucleus from the cytoplasm in a grayscale picture, hole filling is utilised to produce the nucleus. The proposed method achieves 92% accuracy when isolating the nucleus and 78% when isolating the cytoplasm.

Adollah et al., (2008) Examine the division of blood cells for clues to the causes of disease and potential treatments. The Otsu technique is another strategy for white blood cell segmentation; it is based on a circular histogram. One indicator of threshold over a two-dimensional histogram on the RGB and HIS color models is a higher degree of entropy. Grey level threshold, morphological processes, various filtering approaches, colour match, and colour threshold are only few of the segmentation methods analysed and compared in this study. An alternate strategy involves first generating maximum intensity, and then using binarization to extract shape information. Last but not least, we use GVF to identify cells and seeded watershed to partition them.

Theera-Umpon, et.al. (2007) Using the pattern spectrum of each nucleus, explain how white blood cell (WBC) subtypes may be determined. At first, two granulometric metrics—area and high spectral position—are selected. Classifiers such as Bayesian and neural networks are implemented.

Scotti (2005) Gives a way to tell the difference between blast cells and typical lymphocytes. Segmentation of cell nuclei using Otsu's thresholding allows for further classification of blasts as L1, L2, or L3. The cells are then sorted using a KNN classifier trained on geometric characteristics. But the method can only distinguish between normal and blast cells.

III. PROPOSED METHODOLOGY

Step1: *Input Images*

In the proposed model, both pictures of cancerous and healthy cells are used as input.

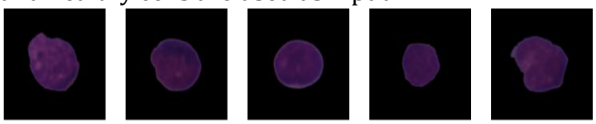


Fig. 2. Input Images fed to the model

Step2*: Import Libraries/ Necessary Packages*

Import all the libraries and necessary packages and all the requirements that are used for the execution of the proposed model.

```
import numpy as np
import imutils
import cv2
import os

from sklearn.neighbors import KNeighborsClassifier as KNN
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB as gnb
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

Step3: *Import Datasets*

Although there are some staining noise and illumination difficulties with these cells, which were considerably fixed throughout the collection process, the overall quality of the photographs is high, making them an excellent depiction of real-world images.

An expert oncologist annotated the ground truth labels to help distinguish between immature leukemic blasts and normal cells, which can seem quite similar under a microscope.

There are 15,135 photographs representing 118 patients.

- Normal cell;
- Leukemia blast.

```
[4]: Datapath = 'Data/ALL_IDB2/img/'
imagePaths = os.listdir(Datapath)

# loop over the input images
for file in imagePaths:
    # load the image and extract the class label (assuming that our
    # path as the format: /path/to/dataset/{class}.{image_num}.jpg
    image = cv2.imread(Datapath + file)
    label = int(file[6])
    # extract raw pixel intensity "features", followed by a color
    # histogram to characterize the color distribution of the pixels
    # in the image
    pixels = image_to_feature_vector(image)
    hist = extract_color_histogram(image)
    # update the raw images, features, and labels matricies,
    # respectively
    rawImages.append(pixels)
    features.append(hist)
    labels.append(label)
```

Step4: *Train Test and Split*

The Training Set

For the model to train and discover any latent features or patterns, it must first learn from the data set. Each iteration of the neural network model is given the same training data, so it may learn from the consistency of the input across time. The model needs to be trained under a wide variety of conditions, with inputs that can account for any future data sample that hasn't been seen yet.

The Test Set

When the training phase is complete, the model is put to the test with a whole new set of data.

The Validation Set

While our model is being trained, its efficacy is checked against a separate set of data known as the validation set. Using the results of this validation, we may potentially modify the model's hyperparameters and other variables. It acts like a critic would, letting us know if our training is headed in the right direction. The model is continuously tested on the validation set while it is being trained on the training set. The primary reason for creating a separate validation set is to prevent overfitting, which occurs when a model performs exceptionally well at identifying samples in the training set but struggles to generalise to new data and make appropriate classifications.

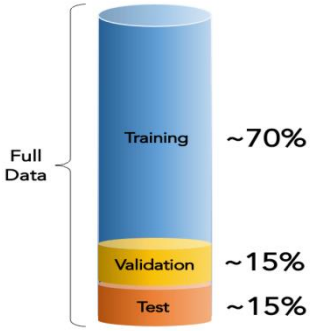


Fig. 3. Train, test, and validate splitting

```
In [7]: # partition the data into training and testing splits, using 75%
        # of the data for training and the remaining 25% for testing
        (trainImage, testImage, trainImageLabel, testImageLabel) = train_test_split(rawImages,
                                                                    labels,
                                                                    test_size=0.25,
                                                                    random_state = 123)
        (trainFeat, testFeat, trainFeatLabel, testFeatLabel) = train_test_split(features,
                                                                    labels,
                                                                    test_size=0.25,
                                                                    random_state = 123)
```

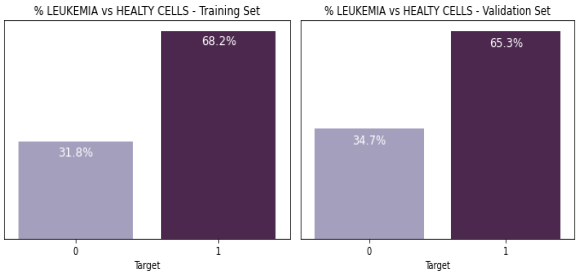


Fig. 4. Dataset training and validation set

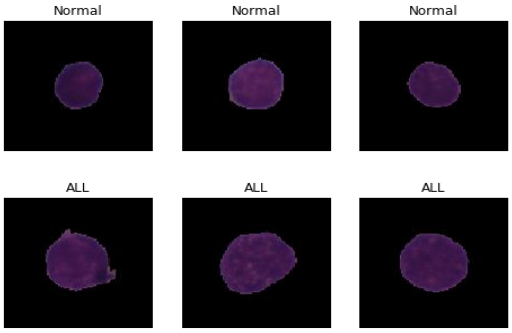


Fig. 5. Data Augmentation method

Step5: *Data Augmentation*

When more information is needed than is currently available, data augmentation can be used to generate new data points from existing data. To increase the size of the dataset, it may be useful to either slightly alter the data or use machine learning models to produce new data points inside the latent space of the original data.

Step6*: Image Preprocessing*

The phrase "pre-processing" is used to describe elementary operations on images, with intensity images serving as both input and output. These well-known images have the same format as the raw data the sensor captured Intensity images, for instance, are often represented as a matrix of picture function

values (brightness). Geometric image transformations (including rotation, scaling, and translation) are grouped together in this article as pre-processing techniques due to their commonalities. However, pre-real processing's goal is to enhance the picture data by reducing artefacts like accidental distortions or boosting the quality of characteristics that will be used in later processing.

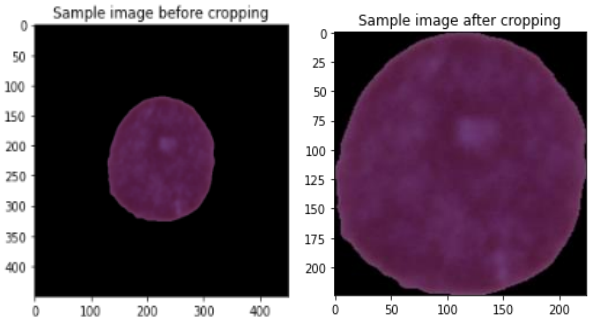


Fig. 6. Image Preprocessing applied for more accuracy

Step7: *Image Segmentation*
Segmentation is a common technique in the domains of digital image processing and analysis, and it is used to break down images into their individual parts. In many cases, the properties of individual pixels are what ultimately decide this segmentation.

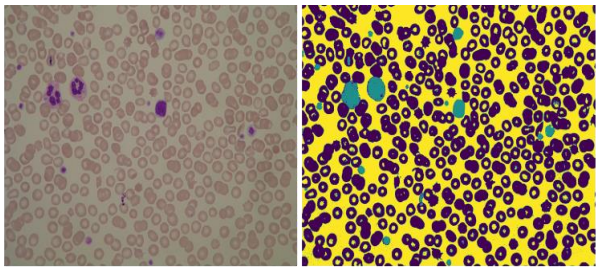


Fig. 7. Image Segmentation process

Picture segmentation, in the disciplines of digital image processing and computer vision, is the process of dividing a big digital image into smaller, more manageable bits termed "image segments," "image regions," or "image objects." Segmentation's goal is to simplify or otherwise alter an image's representation so that it's easier to grasp. Objects and boundaries in images may be easily identified via image segmentation (such as lines, curves, etc.). Accurate picture segmentation involves labeling each image pixel so that pixels with the same label have common characteristics. Segmentation is a method for extracting shapes from images; the result might be a collection of connected segments or a set of contours. When it comes to a certain attribute or calculated characteristic, like colour, brightness, or texture, all pixels inside an area are the same.

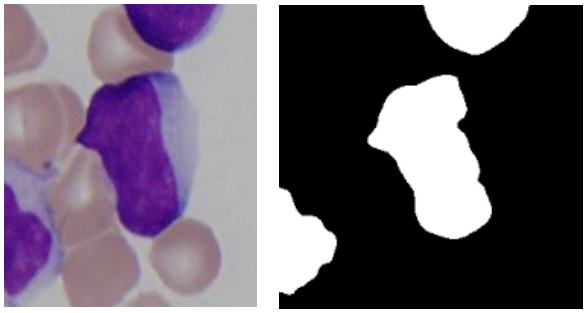


Fig. 8. Image Before Segmentation(Left) Image after Segmentation(Right)

Step8: *CLAHE Method*
CLAHE, a kind of AHE designed to counteract excessive contrast, has been developed. CLAHE processes the image in small pieces called tiles rather than the entire image. The next step is to use bilinear interpolation to merge close tiles and remove the phoney borders. You may use this method to make your photographs stand out more. CLAHE is not just useful for black and white images, but also for colour ones. which is commonly applied to the luminance channel and yields far better results for an HSV image after altering only the luminance channel compared to a BGR image after adjusting all channels.

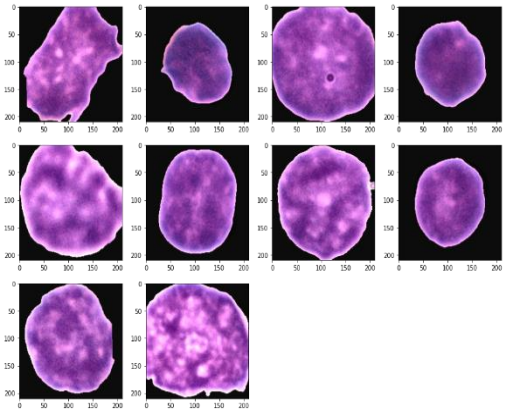


Fig. 9. CLAHE Method for improving the contrast

Step9*: Enhancement*
Using a computer programme, you may improve a digital photo so it looks better on a screen or can be used in further image processing. Noise reduction, sharpening, and increasing contrast are just a few methods for improving the visibility of images' finer features.

Step10: *Normalization*
By eliminating unnecessary information and redundancy, normalization boosts the table's data integrity. Normalization also helps with database organization. There is a multi-step process involved in setting the data into tabular form and eliminating duplicates from relational tables.

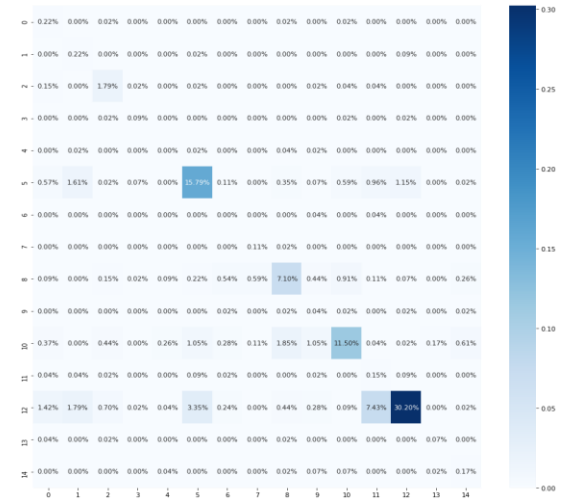


Fig. 10. Normalization of the data

Step11: *Coorelation*

The phrase "correlation matrix" is used to describe a table that displays the relationship between two or more variables through their respective correlation coefficients. Each cell in the table displays the relationship between a given pair of variables. Between -1 and 1, the value is calculable. Correlation matrices are used to do data summarization, in-depth diagnostic analysis, and high-level analytical input.

The correlation's two primary determinants are:
Magnitude: higher the magnitude, the stronger the association.

Sign: If it is positive, a predictable association exists. An inverse correlation exists if the value is negative.

The following two libraries have been used to construct a correlation matrix: Python Library Pandas Libraries.

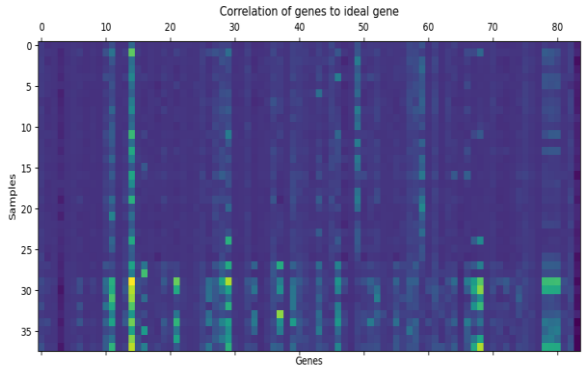


Fig. 11. Correlation of Values

Step12: *Feature Extraction*

By "feature extraction," we mean the procedure of converting unstructured data into structured numerical features in a way that preserves the original dataset's meaning and use. Machine learning on this refined data set outperforms the original data set.

Fig. 12. Feature Extraction using Heatmap

Feature extraction is a method for reducing the dimensionality of a huge dataset, making it more manageable. These massive datasets have a characteristic of having numerous variables, which necessitates substantial computer resources to analyse. Feature extraction refers to methods for selecting and/or combining variables into features, which greatly decreases the amount of data that must be processed while still accurately and completely characterising the original data set. Feature extraction is useful when less computational resources are available but no relevant information is to be lost. A further way that feature extraction aids analysis is by helping cut down on redundant information. The machine's attempts to produce variable combinations (features) and the reduction of data both help to speed up the learning and generalisation phases of machine learning.

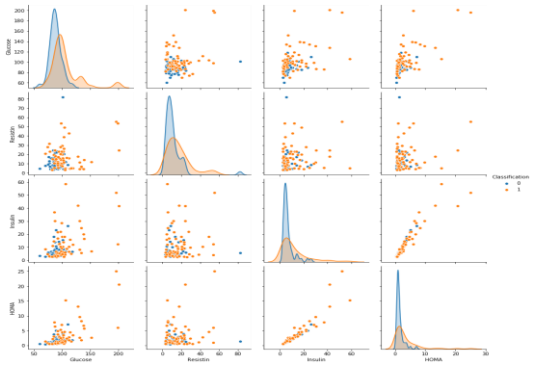


Fig. 13. Features Extraction from the matrix

Step13: *Feature Selection*

Features are selected to reduce the number of input variables while building a predictive model. In some cases, reducing the number of input variables might improve model performance and save on computational resources. Statistical feature selection methods examine the correlations between each input variable and the target variable, then prioritise those that have the strongest ties. On the other hand, depending on the input and output data types, these procedures can be fast and accurate.

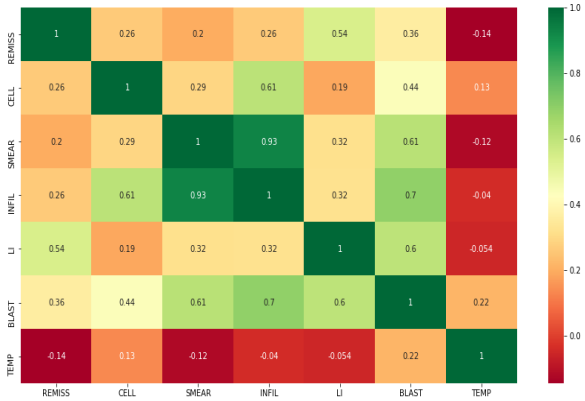
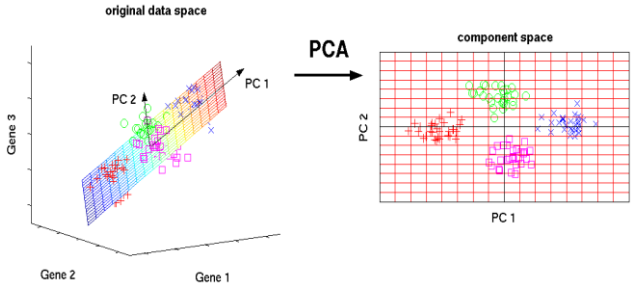


Fig. 14. Feature Selection

Step14: *PCA (Principal Component Analysis)*

Principal component analysis (PCA) is a well-liked method for evaluating large datasets with numerous dimensions or features per observation because it simplifies the interpretation of multidimensional data while retaining as much information as possible. While there are a number of excellent resources available that provide an explanation of principal component analysis (PCA), the vast majority of these sources are rather technical for the average reader. An effective PCA analysis may be broken down into five basic steps. I'll take you by the hand and show you how to do a principal component analysis (PCA) without getting bogged down in the specifics of how to compute the components.



PCA steps:

1. Obtain authentic data
2. Make the data uniform by averaging each characteristic. then we have a new coordinate origin.
3. Calculate the data's covariance matrix.
4. Linear transformation should be used.
5. Eigen Vectors are calculated using Eigen Values.
6. Identify the N biggest Eigen Values.
7. original data onto Eigen Vectors in a projection.

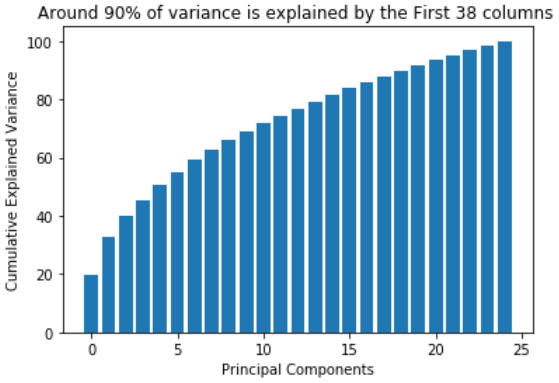


Fig. 15. Principal Component Analysis

Step15*: CNN (Convolution Neural Network)*

A Convolutional Neural Network (ConvNet/CNN) is a form of Deep Learning system that can read in an image, assign varying levels of importance (learnable weights and biases), and identify distinct areas and objects within the picture. A ConvNet needs far less preparation than other classification strategies. ConvNets have the capacity to learn these filters and attributes, while in traditional methods filters are constructed by hand. A ConvNet's architecture mimics that of the human brain's neural connection network and takes design cues from the anatomy of the visual cortex.The Receptive Field is the region of the visual field where individual neurons respond to stimuli. A series of overlapping visual fields span the whole visual field.

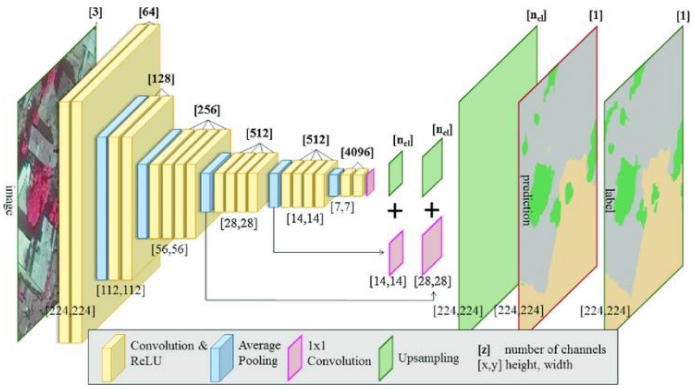


Fig. 16. Convolutional Neural Network Flow Chart

If the right filters are applied to a ConvNet, it may be able to efficiently capture the spatial and temporal relationships present in a picture. The design provides a better fit to the image collection since there are fewer factors to analyze and the weights may be reused. Thus, it is possible to train the network to more accurately assess the level of complexity present in a picture.

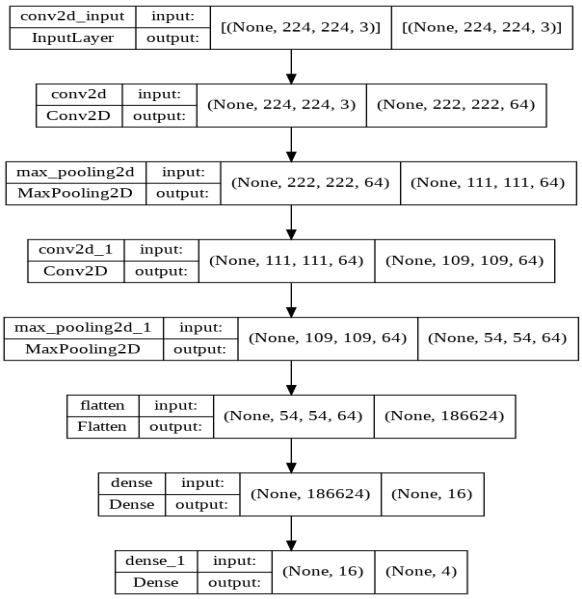


Fig. 17. Convolution Neural Network Derived by our Proposed Methodology

Step16: *Data Exploration*

Data exploration is the first step in data analysis, and it consists of reviewing and visualizing the data to either get quick insights or highlight areas or trends that require further investigation. Users may gain a deeper understanding of the larger picture with the help of interactive dashboards and point-and-click data exploration. Being visual learners, humans can process visual information substantially more quickly than numerical information. As a result, it

could be challenging for data scientists to examine hundreds of rows of data points and form

conclusions on their own. It is possible to discover connections or anomalies through the use of data visualization tools and components like colors, shapes, lines, graphs, and angles.

Step17: *Data Analysis*

The goal of data analysis is to acquire understanding, validate hypotheses, and improve decision-making through the process of collecting, sorting, cleaning, manipulating, and modeling data. [1] There are a wide variety of applications for data analysis in fields including business, research, and the social sciences. Multiple treatments fall under its umbrella, and it is referred to by a wide variety of labels. [2] In today's competitive business world, data analysis is the key to making educated decisions and running an efficient firm. [3]

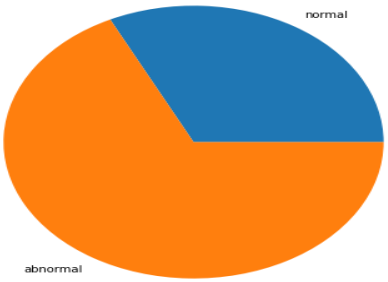


Fig. 18. Classification of Cancerous and Normal Cells

Step18: *Error Analysis*

Through observation, testing, and measurement, we have gained knowledge about the physical world. It's vital to know how to present such data, assess it, and draw effective conclusions from it. Keep in mind that there is always some degree of error associated with every measurement of a physical quantity. Anything cannot ever be measured precisely. Even though it's best to cut down on mistakes as much as possible, they're inevitable. Additionally, the error has to be acknowledged and effectively handled to draw valid conclusions.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity}=\frac{TN}{TN+FP} \quad (4)$$

$$\text{F1-Score}=\frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (5)$$

Classification reporting

	precision	recall	f1-score	support
0.0	0.97	0.92	0.94	30
1.0	0.81	0.87	0.84	15
accuracy			0.89	45
macro avg	0.88	0.88	0.89	45
weighted avg	0.89	0.89	0.89	45

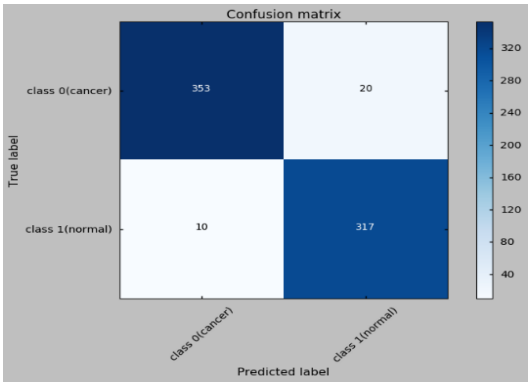


Fig. 19. Confusion Matrix for Cancerous and Normal Cells

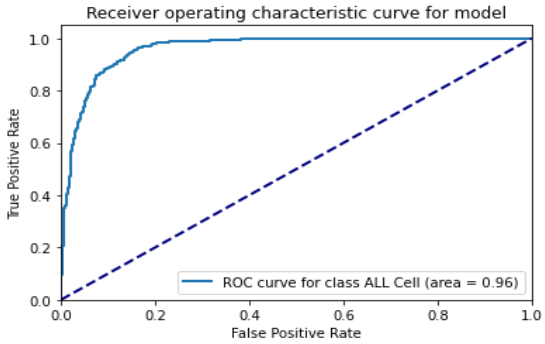


Fig. 20. ROC Curve for Evaluation

Step19: *Classification*

The term "data categorization" refers to the act of organizing information into meaningful groups for retrieval, management, and long-term storage. Data retrieval is facilitated by an organized database, which can be quickly accessed upon need.

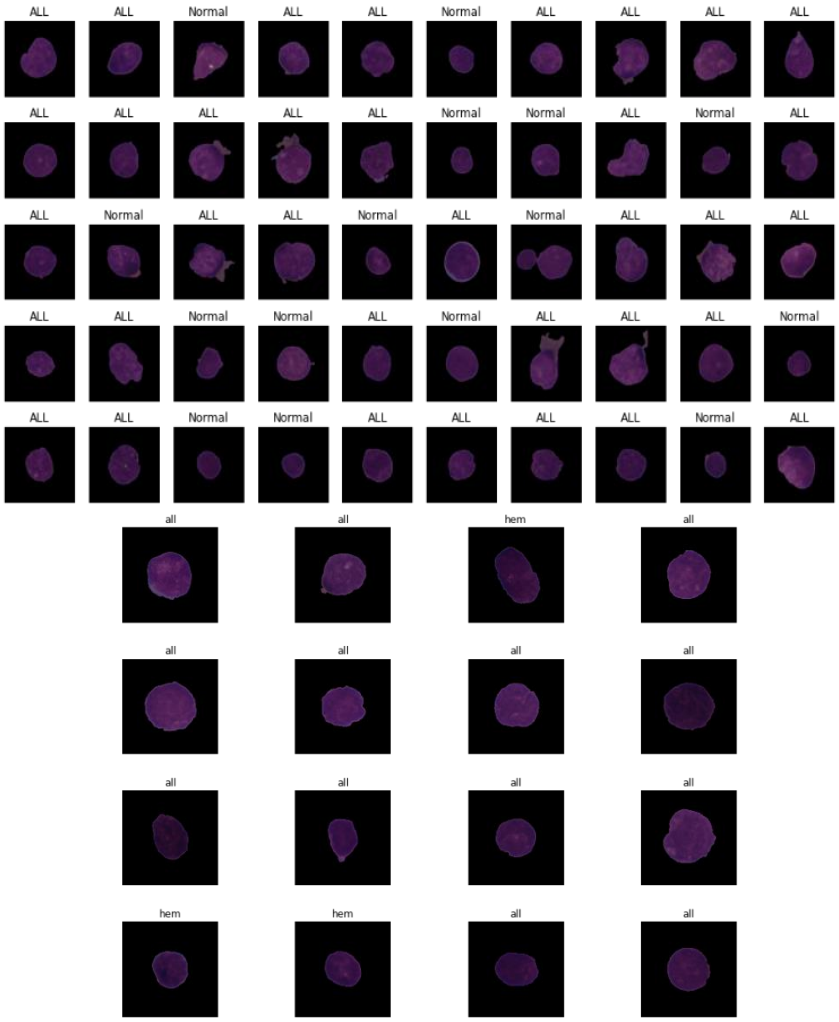


Fig. 21. Classification of Data into different Clusters

Step20: *Prediction*

Data analytics may be used for predictive purposes, thus the term "predictive analytics." This strategy builds a prediction model for foreseeing the future by utilising data, analysis, statistics, and machine learning approaches. Predictive modelling use mathematics to seek for patterns in input data in order to forecast future events or outcomes. Predictive analytics, a subfield of data analytics that makes use of both new and historical data to foresee patterns of behaviour and industry development, relies heavily on this technique.

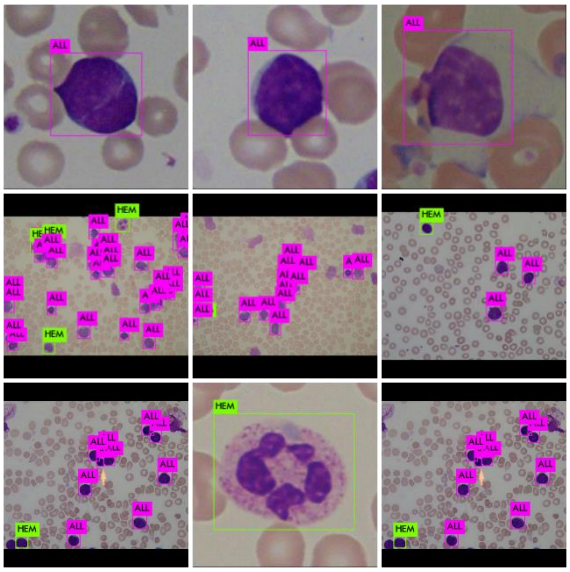
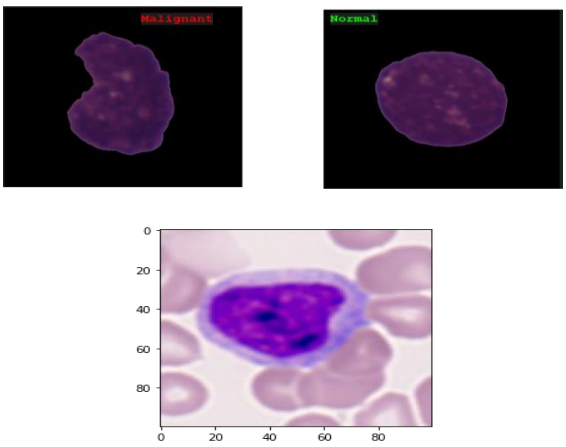
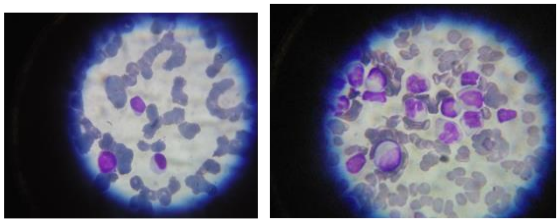


Fig. 22. Prediction of Results

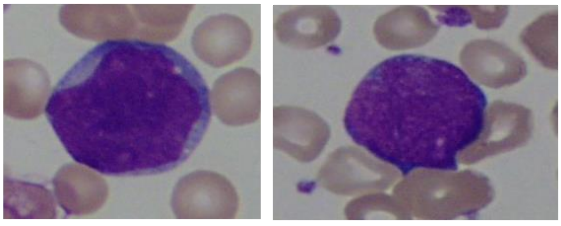
Step21: *Detection*
With the help of an object detection model, we can find out if and where a particular set of objects are in a given image or video.



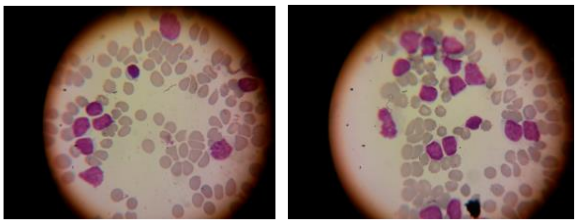
Normal Cell



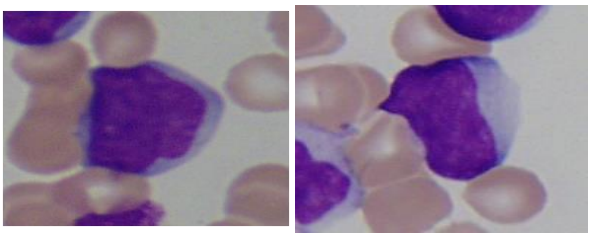
AML (Acute myeloid leukemia)



CML (Chronic myeloid leukemia)



ALL (Acute lymphocytic leukemia)



CLL **(**Chronic lymphocytic leukemia)

Fig. 23.

Fig. 24. Detection of Different types of Cancerous cells

Step22: *Evaluation Model*
It is via the use of several assessment measures that a machine learning model's efficacy, strengths, and weaknesses may be understood. Evaluation of a model's performance is essential in the preliminary phases of a study. Evaluation of the model also helps with keeping an eye on it.

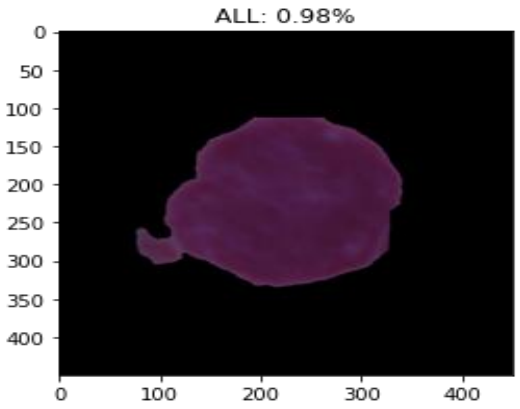


Fig. 25. Evaluation of Loss and Accuracy in the Model

IV. COMPARATIVE STUDY OF VARIOUS ALGORITHMS

K-Neighbors: K-nearest neighbor's approach (KNN) is a supervised learning classifier that employs geographical closeness to make predictions or classifications about where to place a single data point. The supervised machine learning method K-nearest neighbours (KNN) is effective in resolving both classification and regression issues. While it's simple to pick up and use, it becomes painfully sluggish as more data is incorporated.

```
*********  K- Nearest Neighbor  *********

Score train : 1.0
-----------
Score test : 0.91
-----------
AUC: 0.95 (std:0.0199), (splits = 5)

--------
Classification reporting
              precision    recall  f1-score   support

         0.0       0.96      0.90      0.93        30
         1.0       0.82      0.93      0.87        15

    accuracy                           0.91        45
   macro avg       0.89      0.92      0.90        45
weighted avg       0.92      0.91      0.91        45
```

Decision Trees: Trees are used in the decision-making tool known as a "decision tree" to depict several paths and their associated probabilities, costs, and benefits. This is one example of a conditional control statement-based algorithm.

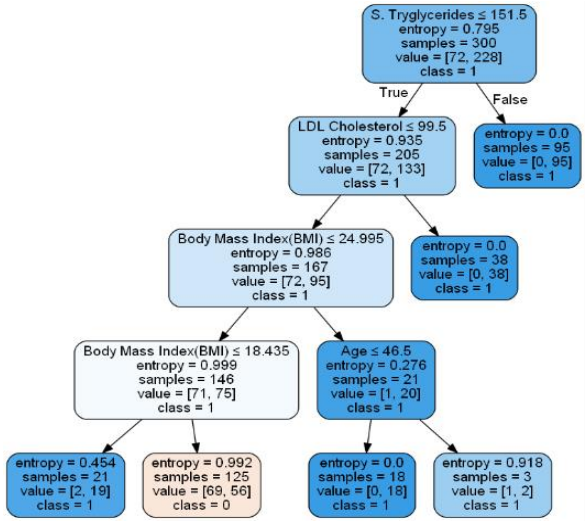


Fig. 26. Decision Tree

```
*********  Decision Tree  *********

Score train : 1.0
-----------
Score test : 0.82
-----------
AUC: 0.76 (std:0.0156), (splits = 5)

--------
Classification reporting
              precision    recall  f1-score   support

         0.0       0.87      0.87      0.87        30
         1.0       0.73      0.73      0.73        15

    accuracy                           0.82        45
   macro avg       0.80      0.80      0.80        45
weighted avg       0.82      0.82      0.82        45
```

Support Vector Machines: The popular supervised learning technique known as Support Vector Machine (SVM) may be applied to both classification and regression issues. However, it is most frequently used to Classification issues in Machine Learning. The SVM approach seeks to rapidly classify future data points by identifying the optimal line (or decision boundary) for splitting an n-dimensional space into classes. The optimal decision boundary is referred to as a hyperplane. Selective Feature Selection (SVM) selects the most out-there vectors and points from which to construct the hyperplane. The SVM method uses support vectors to represent these exceptional instances.

```
*********  SVM  *********

Score train : 0.88
-----------
Score test : 0.84
-----------
-----------
No prob methods for this model :  predict_proba is not available when  probability=False
-----------

--------
Classification reporting
              precision    recall  f1-score   support

         0.0       0.83      0.97      0.89        30
         1.0       0.90      0.60      0.72        15

    accuracy                           0.84        45
   macro avg       0.86      0.78      0.81        45
weighted avg       0.85      0.84      0.83        45
```

Logistic Regression Algorithm: Logistic regression is a technique in statistical analysis for predicting a yes/no outcome based on a collection of prior observations. A dependent data variable may be predicted by using a logistic regression model, which does so by analyzing the relationship between one or more independent variables that are previously known to exist.

```
*********  LR  *********

Score train : 0.91
-----------
Score test : 0.89
-----------
AUC: 0.94 (std:0.034), (splits = 5)

--------
Classification reporting
              precision    recall  f1-score   support

         0.0       0.90      0.93      0.92        30
         1.0       0.86      0.80      0.83        15

    accuracy                           0.89        45
   macro avg       0.88      0.87      0.87        45
weighted avg       0.89      0.89      0.89        45
```

Naïve Bayes: One branch of classification algorithms motivated by Bayes' Theorem is the Naive Bayes classifier. It's not just one technique; rather, it's a group of algorithms predicated on the assumption that no two compared qualities are inherently equivalent. Naive Bayes is a classifier that does well with both binary and multiclass classification. Naive Bayes excels in cases with categorical input variables, but it struggles when dealing with numerical input variables. Useful for seeing into the future and making educated guesses based on previous results.

```
*********  Naive Bayes  *********

Score train : 1.0
-----------
Score test : 0.89
-----------
AUC: 0.92 (std:0.0366), (splits = 5)

--------
Classification reporting
              precision    recall  f1-score   support

         0.0       0.93      0.90      0.92        30
         1.0       0.81      0.87      0.84        15

    accuracy                           0.89        45
   macro avg       0.87      0.88      0.88        45
weighted avg       0.89      0.89      0.89        45
```

Random Forest: By combining the results of many decision trees, we get the random forest classification technique. Through the use of bagging and feature randomization, it creates a set of trees whose aggregate prediction is superior than that of any individual tree. technique. It may be utilized for

classification and regression problems in ML. Ensemble learning, the practice of combining several classifiers to solve complex problems and improve model efficiency, is the foundation on which it is constructed. Random Forests are a type of classifier that uses several decision trees to make more accurate predictions by applying those trees to different parts of the provided dataset and then averaging the results. The random forest takes forecasts from all of the trees and makes a final prediction based on the majority of those projections, rather than relying on just one.

```
*********  Random Forest   *********

Score train : 1.0
-----------
Score test : 0.93
-----------
AUC: 0.95 (std:0.0146), (splits = 5)


--------
Classification reporting
              precision    recall  f1-score   support

         0.0       0.97      0.93      0.95        30
         1.0       0.88      0.93      0.90        15

    accuracy                           0.93        45
   macro avg       0.92      0.93      0.93        45
weighted avg       0.94      0.93      0.93        45
```

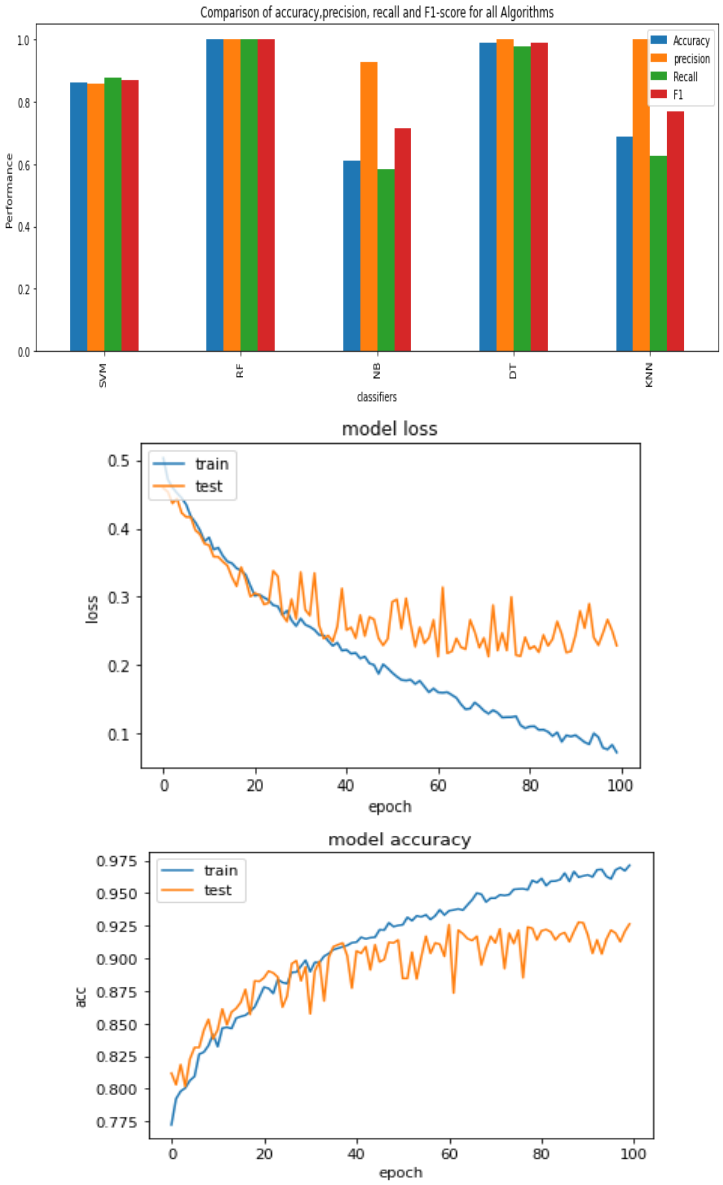


Fig. 27. Performance Comparison of Various Algorithms

Model	Accuracy	Precision	Recall	F1 score
KNN	0.87	0.91	0.88	0.87
SVC	0.98	0.98	0.98	0.98
RF	0.99	0.99	0.99	0.99
NB	0.94	0.94	0.94	0.94
DT	0.87	0.87	0.88	0.87
LR	0.97	0.97	0.97	0.97

TABLE I. COMPARATIVE EVALUATION

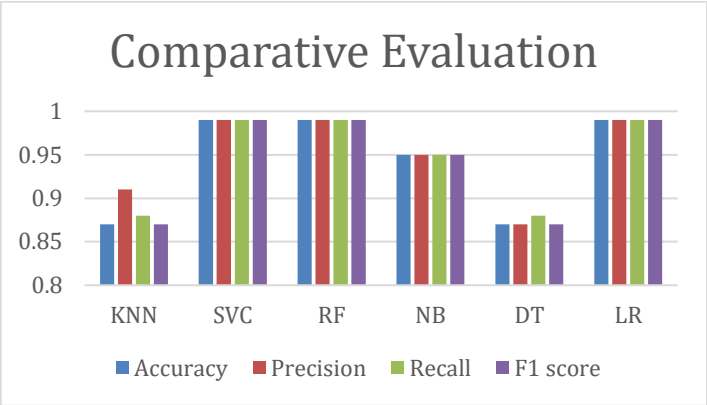


Fig. 28. Comparisons Using Accuracy, Precision, Recall, and the F-1 Score

V. CONCLUSION

For human blood cells, only the leukaemia detection technique described above is employed. This technique uses image processing to segment, eliminate, and fill gaps in blood cell pictures in order to obtain the boundaries of the malignant blood cell. Infrequently utilized on a monthly or temporary basis due to the high cost and lengthy turnaround time of pathological testing. If suspicious cells are obtained through such testing, you should move on with the diagnosis under medical supervision. This effort will help us build more precise techniques for diagnosing diseases by allowing us to swiftly and reliably detect additional types of leukaemia. In light of the increasing popularity of machine learning algorithms, this research examines and contrasts five of the most popular ones: Methods such as Deep Learning, Naive Bayes, k-Nearest Neighbor, Logistic Regression, and Support Vector Machines are used. This study built on prior literature classifying and making predictions about leukaemia. More investigation into this field will pave the way for effective application of relevant Machine Learning algorithms to determine the effect of the treatment provided to patients with leukaemia.

References

[1] G. 1. Eid, M. M., Rashed, A. N. Z., Bulbul, A.A.-M. & Podder, E. Mono-rectangular core photonic crystal fiber (MRC-PCF) for skin and blood cancer detection. Plasmonics 16, 717–727 (2021).

[2] Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71, 209–249 (2021).

[3] T. L. L. Society. Blood cancer facts 2016–2017. https:// www. kaggle. com/ uciml/ sms- spam-colle ction- datas et/ (2017).

[4] Goutam, D. & Sailaja, S. Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier. In 2015 IEEE International Conference on Engineering and Technology (ICETECH), 1–5 (IEEE, 2015).

[5] El-Halees, A. M. & Shurrab, A. H. Blood tumor prediction using data mining techniques. Health Inform. 6, 23–30 (2017).

[6] Vijayarani, S. & Sudha, S. An efficient clustering algorithm for predicting diseases from hemogram blood test samples. Indian J. Sci. Technol. 8, 1 (2015).

[7] Xiao, Y., Wu, J., Lin, Z. & Zhao, X. A deep learning-based multi-model ensemble method for cancer prediction. Comput. Methods Programs Biomed. 153, 1–9 (2018).

[8] Subhan, M. & Kaur, M. Significant analysis of leukemic cells extraction and detection using KNN and Hough transform algorithm. Int. J. Comput. Sci. Trends Technol. (IJCST) 3 (2015).

[9] Gal, O., Auslander, N., Fan, Y. & Meerzaman, D. Predicting complete remission of acute myeloid leukemia: Machine learning applied to gene expression. Cancer Inform. 18, 11769351198 35544 (2019).

[10] Rustam, F. et al. Wireless capsule endoscopy bleeding images classification using CNN based model. IEEE Access 9, 33675–33688 (2021).

[11] Reshi, A. A. et al. An efficient CNN model for COVID-19 disease detection based on x-ray image classification. Complexity 2021 (2021).

[12] Shafique, S. & Tehsin, S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convo-lutional neural networks. Technol. Cancer Res. Treat. 17, 1533033818802789 (2018).

[13] Mohd, F., Noor, N. M. M., Bakar, Z. A. & Rajion, Z. A. Analysis of oral cancer prediction using features selection with machine learning. In The 7th International Conference on Information Technology (ICIT) (2015).

[14] Loey, M., Naman, M. & Zayed, H. Deep transfer learning in diagnosing leukemia in blood cells. Computers 9, 29 (2020).

[15] Abd El-Nasser, A., Shaheen, M. & El-Deeb, H. Enhanced leukemia cancer classifier algorithm.

In 2014 Science and Information Conference, 422–429 (IEEE, 2014).

[16] MoradiAmin, M., Samadzadehaghdam, N., Kermani, S. & Talebi, A. Enhanced recognition of acute lymphoblastic leukemia cells in microscopic images based on feature reduction using principle component analysis. Front. Biomed. Technol. 2, 128–136 (2015).

[17] Kandil, A. & Hassan, O. Automatic segmentation of acute leukemia cells. Int. J. Comput. Appl. 133, 1–8 (2016).

[18] Claro, M. et al. Convolution neural network models for acute leukemia diagnosis. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 63–68 (IEEE, 2020).

[19] Castillo, D. et al. Leukemia multiclass assessment and classification from microarray and RNA-Seq technologies integration at gene expression level. PLoS One 14, e0212127 (2019).

[20] Nazari, E. et al. Deep learning for acute myeloid leukemia diagnosis. J. Med. Life 13, 382 (2020).

[21] Stirewalt, D. Abnormal expression changes in aml. https:// www. ncbi. nlm. nih. gov/ geo/ query/ acc. cgi? acc= GSE94 76 (2018).

[22] Song, G. New markers for minimal residual disease detection in acute lymphoblastic leukemia. https:// www. ncbi. nlm. nih. gov/ geo/ query/ acc. cgi? acc= GSE28 497 (2018).

[23] He, H., Bai, Y., Garcia, E. A. & Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328 (IEEE, 2008)

[24] Mujahid, M. et al. Sentiment analysis and topic modeling on tweets about online education during COVID-19. Appl. Sci. 11, 8438 (2021).

[25] Rustam, F. et al. Classification of Shopify app user reviews using novel multi text features. IEEE Access 8, 30234–30244 (2020).

[26] Smialowski, P., Frishman, D. & Kramer, S. Pitfalls of supervised feature selection. Bioinformatics 26, 440–443 (2009).

[27] Saultz, J.N.; Garzon, R. Acute Myeloid Leukemia: A Concise Review. J. Clin. Med. 2016, 5, 33.

[28] American Society of Hematology. Available online: https://www.hematology.org:443/ (accessed on 29 July 2020). 4. Kumar, C.C. Genetic Abnormalities and Challenges in the Treatment of Acute Myeloid Leukemia. Genes Cancer 2011, 2, 95–107.

[29] Ahmed, N.; Yigit, A.; Isik, Z.; Alpkocak, A. Identification of Leukemia Subtypes from Microscopic Images Using Convolutional Neural Network. Diagnostics 2019, 9, 104.

[30] Prinyakupt, J.; Pluempitiwiriyawej, C. Segmentation of white blood cells and comparison of cell morphology by linear and naïve Bayes classifiers. Biomed. Eng. Online 2015, 14, 63.

[31] Sasada, K.; Yamamoto, N.; Masuda, H.; Tanaka, Y.; Ishihara, A.; Takamatsu, Y.; Yatomi, Y.; Katsuda, W.; Sato, I.; Matsui, H. Inter-observer variance and the need for standardization in the morphological classification of myelodysplastic syndrome. Leuk. Res. 2018, 69, 54–59.

[32] Amin, M.M.; Kermani, S.; Talebi, A.; Oghli, M.G. Recognition of Acute Lymphoblastic Leukemia Cells in Microscopic Images Using K-Means Clustering and Support Vector Machine Classifier. J. Med. Signals Sens. 2015, 5, 49–58.

[33] De Angelis, C.; Pacheco, C.; Lucchini, G.; Arguello, M.; Conter, V.; Flores, A.; Biondi, A.; Masera, G.; Baez, F. The Experience in Nicaragua: Childhood Leukemia in Low Income Countries—The Main Cause of Late Diagnosis May Be 'Medical Delay. Int. J. Pediatr. 2012, 2012, 1–5.

[34] Salah, H.T.; Muhsen, I.N.; Salama, M.E.; Owaidah, T.; Hashmi, S.K. Machine learning applications in the diagnosis of leukemia: Current trends and future directions. Int. J. Lab. Hematol. 2019, 41, 717–725.

[35] Howell, D.A.; Smith, A.G.; Jack, A.; Patmore, R.; Macleod, U.; Mironska, E.; Roman, E. Time-to-diagnosis and symptoms of myeloma, lymphomas and leukaemias: A report from the Haematological Malignancy Research Network. BMC Blood Disord. 2013, 13, 9.

[36] Matek, C.; Schwarz, S.; Spiekermann, K.; Marr, C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat. Mach. Intell. 2019, 1, 538–544.

[37] Abdeldaim, A.M.; Sahlol, A.T.; Elhoseny, M.; Hassanien, A.E. Computer-Aided Acute Lymphoblastic Leukemia Diagnosis System Based on Image Analysis. In Advances in Soft Computing and Machine Learning in Image Processing; Springer: Cham, Switzerland, 2018; Volume 2018, pp. 131–147.

[38] Labati, R.D.; Piuri, V.; Scotti, F. All-IDB: The acute lymphoblastic leukemia image database for image processing. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2045–2048.

[39] Shafique, S.; Tehsin, S. Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Networks. Technol. Cancer Res. Treat. 2018, 17, 1533033818802789.

[40] Kazemi, F.; Najafabadi, T.A.; Araabi, B.N. Automatic Recognition of Acute Myelogenous Leukemia in Blood Microscopic Images Using K-means Clustering and Support Vector Machine. J. Med. Signals Sens. 2016, 6, 183–193.