# Lukemia Classification Using Machine Learning Model

Om Patil
*KIT's College of Engineering (Autonomous)*
Kolhapur, Maharashtra, India
pom72103@gmail.com

Sujit Gade
*KIT's College of Engineering (Autonomous)*
Kolhapur, Maharashtra, India
gadesujit10@gmail.com

Suyog Mali
*KIT's College of Engineering (Autonomous)*
Kolhapur, Maharashtra, India
suyogmali321@gmail.com

Chaitanya Uthale
*KIT's College of Engineering (Autonomous)*
Kolhapur, Maharashtra, India
chaitanyauthale5@gmail.com

Uma Gurav
*KIT's College of Engineering (Autonomous)*
Kolhapur, Maharashtra, India
gurav.uma@kitcoek.in

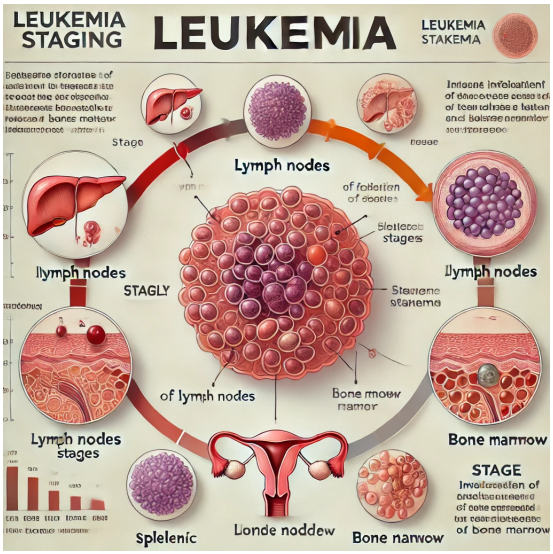
***Abstract*—Leukemia is a type of cancer that affects bone marrow and blood cells; it can be either acute or chronic. Reducing death rates requires early detection. Rapid progression of acute leukemia results in anemia, infections, and bleeding problems because white blood cells build up blood vessels in the bone marrow too soon. An accumulation of aberrant cells results from the delayed progression of chronic leukemia. Algorithms for automated and machine learning are being developed to detect leukemia more reliably and efficiently. These techniques distinguish between normal and aberrant cells by training algorithms on large datasets of blood smear pictures or through the information of required parameters through spectroscopy. Medical pathologists can make better decisions if leukemia is detected more quickly and consistently.**

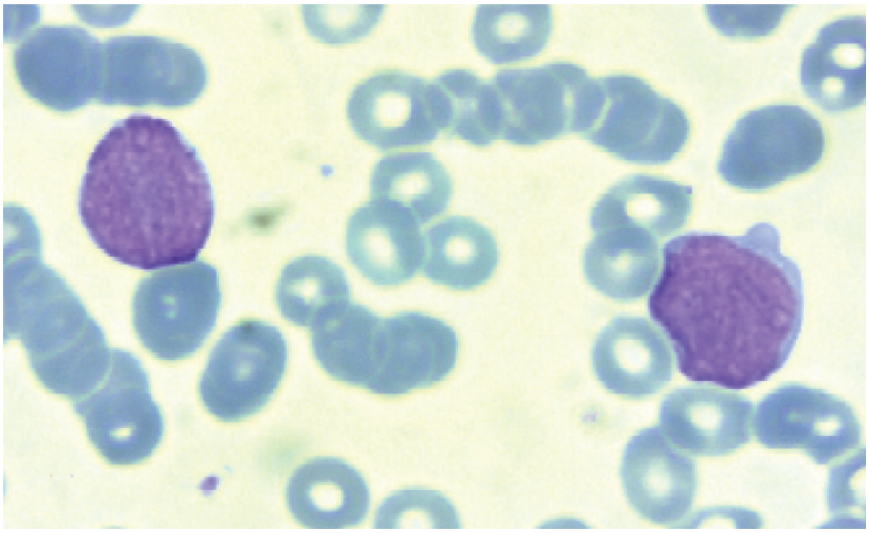
**Even though machine learning algorithms might make detection better, qualified medical personnel are still necessary for result interpretation and the best possible patient care. The suggested model predicts leukemia cells from healthy blood samples by demonstrating a convolutional neural network (CNN) and TensorFlow framework. The accuracy of this method is 97.07, making it a popular tool for both diagnosing and treating leukemia. The second method uses a random forest classifier,scikit learn,opencv and keras library of python to identify huge dataset images of malignant cells from normal cells. Batch normalization of the images further improves accuracy and F1-score.**

## I. Introduction

Leukemia is a disease due to the uncontrolled division of abnormal cells in the body and it develops first in the bone marrow cells of the blood. In the immune system, there is a predominant role played by white blood cells that play an irregular and unrestrained progression. Approximately 3,03,006 people have died of leukemia, out of 4,37,033 cases from the disease were reported The agency will be part of the World Health Organization's International Agency for Cancer Research in the year of 2022 [1]. To manage and improve outcomes for individuals affected by this condition, early detection, and accurate classification are crucial [2]. Although it takes time, the manual method can reach a 100 recognition rate. Medical researchers have become more interested in using machine learning (ML) techniques in recent years. because of their capacity to foresee intricate relationships and patterns in complex datasets. By applying machine learning to the medical field, researchers might obtain important insights into cancer research, potentially leading to better diagnosis. Due to the deep learning method's sophisticated feature extraction capabilities and processing efficiency, picture identification and medical diagnosis can be improved while processing massive volumes of image datasets [3]. The Convolutional Neural Network (CNN) technique is used in leukemia detection to identify patterns and malformations in blood cell pictures and extract important information for an effective leukemia diagnosis.The C-NMC dataset is frequently utilized in leukemia detection [4]. Blood smear slide photos are labeled to show whether they show malignant (leukemia-affected) or normal cells. Leukemia can be detected by machine learning approaches; the same dataset was used for both training and evaluation [4]

979-8-3503-1755-8/24/$31.00 ©2024 IEEE



## II. Related Work

Researchers have previously conducted several studies in this area. This paper offers a thorough review of machine learning algorithms utilized for leukemia detection. By focusing on standardized datasets and appropriate evaluation metrics, the review serves as a significant resource for researchers, noting a 100

## III. Proposed Method

The suggested model illustrated that a comprehensive explanation of the workflow necessitates an examination of the input dataset, the processing methods applied to the input images, the data augmentation strategies employed, and the formulation of a network architecture. This approach has enabled the detection of leukemia through a network-based methodology, as depicted in Figure 1. The proposed system is represented in the subsequent block diagram.

## IV. INPUT DATA

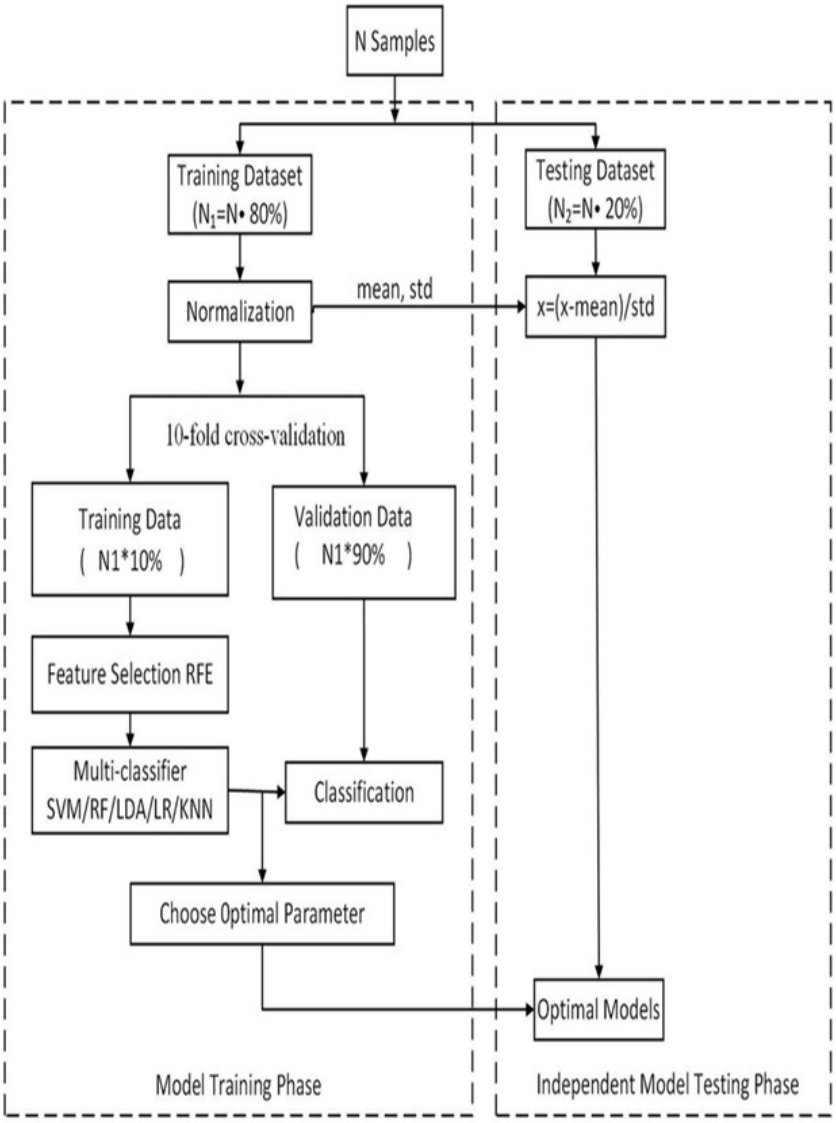
The datasets employed in this study are classified within a particular domain [5]. The C-NMC ALL Challenge dataset from ISBI 2019 encompasses a total of 10,661 cells [6-8]. This dataset contains images of cells that have undergone microscopic segmentation. The images are standardized to dimensions of 200 x 200 x 3 and have been pre-processed to remove all background pixels, concentrating exclusively on the object of interest.

The dataset comprises various image types: (a) ALL-Leukemia, (b) HEM-Normal, (c) segmented image of ALL-Leukemia, and (d) segmented image of HEM-Normal. A HIS (Hue, Saturation, Intensity) image is generated by transforming the original image, which consists of three components of color information: Hue represents the predominant color, Saturation evaluates its purity, and Intensity indicates the brightness of the pixel.

Convolutional Neural Networks (CNNs) employ data augmentation techniques to tackle the challenge of limited labeled data and to enhance generalization capabilities [20]. Various data enhancement methods in CNN algorithms include flipping, rotating, scaling, translating, adjusting brightness and contrast, adding Gaussian noise, cropping, padding, and shearing [9-11]. The implementation of data augmentation improves the accuracy of CNN algorithms, reduces overfitting, and increases robustness, thereby facilitating their application across a variety of computer vision tasks.

A vital component of the analysis involves feature extraction and diagnosis from blood sample images. Relevant features are extracted from these images using various techniques to investigate complex patterns and structures within the cells. Texture analysis methods, such as local binary patterns or Gabor filters, are frequently utilized. To characterize cell shapes, shape features are defined by parameters such as area, perimeter, and eccentricity [3,5]. By counting the cells, it is possible to determine the number of irregular and healthy cells, which is essential for disease diagnosis. Additionally, a color image can be transformed. Inverted images exhibit color reversals compared to their original counterparts, where dark regions appear light and light regions appear dark. Common techniques for blurring include Gaussian blur and median blur.

Convolutional Layer: A convolutional layer receives an image characterized by its dimensions (height, width, channels) and employs learnable filters, also known as kernels. The filter traverses the input image, performing element-wise multiplication of the filter weights with the corresponding pixels within the receptive field. This process, known as convolution, involves sliding the input data over the filters, which then multiply and sum the values to generate the output feature map. Typically, these filters are structured as 3x3 or

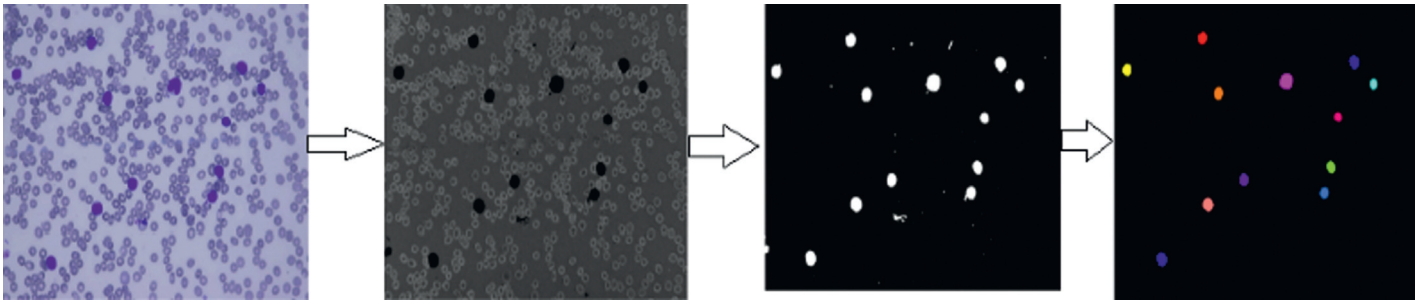



5x5 grids of weight values. By increasing the depth of the convolutional neural network (CNN) through the addition of more layers, it becomes possible to extract a greater variety of features and enhance the understanding of representations. While deeper networks can capture more abstract and intricate features, there is a risk of overfitting, particularly when the available data is limited. Overfitting occurs when the CNN model memorizes the training data, resulting in high accuracy during training but poor performance on unseen data. To address overfitting, techniques such as dropout, regularization, and data augmentation should be employed. The size of the dataset and the careful selection of the model are also vital in mitigating the risk of overfitting.

3.4.2 ReLU Layer: Following the convolution operation, each value in the feature maps is activated on an element-wise basis. The Rectified Linear Unit (ReLU) is essential in CNNs as it introduces non-linearity, promotes sparsity, and facilitates efficient gradient flow. Variants such as Leaky ReLUs and Parametric ReLUs (PReLUs) are frequently utilized as activation functions within convolutional layers. Mathematically, ReLU is defined as $\mathrm{ReLU}(x) = \max(0, x)$. Beyond introducing non-linearity, the activation function enhances the network's expressiveness, enabling it to learn complex patterns. However, the sigmoid activation function has drawbacks, including issues with vanishing gradients and outputs that are not zero-centered. ReLU and its variants are frequently favored in convocational neural networks (CNNs) due to their rapid convergence and enhanced gradient flow.

3.4.3 Padding: There are two primary types of padding: Valid Padding (No Padding), where neither the input image nor the feature maps receive any padding, and Same Padding, which involves padding the input data to ensure that the resulting feature maps maintain the same spatial dimensions as the input data. The calculation of padding is contingent upon the filter size and stride.

3.4.4 Pooling Layer: A pooling layer reduces the spatial dimensions of feature maps through down-sampling. There are two principal types of pooling: maximum pooling and



average pooling. Max Pooling operates by sliding a window across the feature maps, with the maximum values within each window determining the output for that region. Conversely, Average Pooling computes the average values of the elements within the pooling window. The most commonly used sizes for pooling windows are 2x2 or 3x3.

3.4.5 Flatten Layer: The process of flattening in CNNs facilitates the transition from convolutional layers to fully connected layers by transforming all values from the 2D feature maps into 1D vectors. For example, a Flatten layer will convert feature maps of size (224 X 224) into vectors of size (224 X 224) = 50176.

3.4.6 Fully Connected Layer: Fully Connected layers are responsible for mapping the learned features to the final output classes or predictions, utilizing learnable weights and biases. In this layer, neurons apply activation functions to the weighted sum of their inputs, introducing nonlinearity into the network, which is essential for making accurate predictions for various tasks.

3.4.7 Dropout Regularizer: During training, neurons in one or more layers are randomly dropped out at a rate typically ranging from 0.2 to 0.5.

An equal likelihood is assigned to the random dropout of neurons, with no dropout implemented during inference or testing phases. During prediction, the weights of the neurons are adjusted in accordance with the dropout rate.

3.4.8 Batch Normalization: The application of batch normalization enhances the stability of training, accelerates the convergence process, and mitigates the risk of overfitting in Convolutional Neural Networks (CNNs). By normalizing the activations of each layer within a mini-batch during training, the network becomes more resilient and simpler to train. This normalization process involves subtracting the mean of each mini-batch and dividing by its standard deviation.

3.4.9 Loss Function: The Binary Cross Entropy (BCE) loss function is employed for binary classification tasks, comparing predicted probabilities against actual labels. During the training of the classifier, it evaluates the discrepancies between the probability distributions. Since mains speed up and automate the forecasting process, they have become essential. To predict leukemia, a number of researchers have used ML and DL in different ways. For predicting, ML models use either individual or group learners. In certain cases, it has been found that negligible attributes, a subpar pre-processing step, and incorrect categorization system tuning hinder the effectiveness of a truly positive forecasting rate. To solve these issues, this study develops an integrated model that blends ML and DL. In this case, a DL model called Resnet50 is used to collect significant deep features that help build an appropriate features model for classification. Relative to the particular features, the residual characteristics are gathered using Resnet50, also known as residual learning.

Additionally, it was able to extract more residual information from the source photos because it includes 50 layers. This work employs SVM, an ML technique, to classify the deep features. The primary advantages of using SVM are that it operates effectively in high-dimensional spaces, is generally memory-efficient, and clearly distinguishes the various classes with different margins. Given these variables, leukemia with deep characteristics is more accurately predicted by the SVM model. A comparative method that emphasizes the benefits of applying the recommended strategy in terms of efficacy is shown in Table 5.

## V. RESULT

A study looked at 116 publications on machine learning (ML) methods for diagnosing leukemia by analyzing blood smear images. Seventeen articles were selected for examination following a rigorous review process. According to the study, throughout the previous five years, there has been a rise in the application of machine learning techniques for blood smear image analysis.

The majority of research made use of publicly accessible blood smear image databases, particularly the ALL-IDB dataset, which is frequently used to diagnose acute lymphoblastic leukemia (ALL). But since most research used private or homogeneous information, it was challenging to create reliable models that could be applied to other datasets.

According to the study, machine learning (ML) and deep learning (DL) are the two primary types of machine vision algorithms employed for blood smear picture processing.

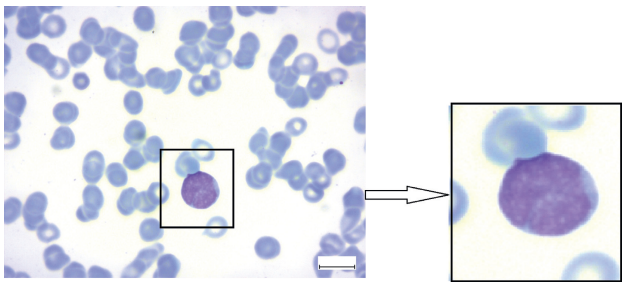
While DL approaches use convolutional neural networks

to automatically extract features, ML methods depend on extracting specific characteristics from images. The study found several characteristics, including chromatin, nucleus structure, and cell shape, that are frequently employed for leukemia detection.

When analyzing blood smear images, segmentation—a method for locating and separating particular cells from an image—is essential. The study investigated a number of segmentation strategies, including as object identification, boundary-based segmentation, and thresholding techniques. It emphasized how crucial segmentation is to accurately extracting information and categorizing leukemia.

The study examined a number of machine learning (ML) segmentation techniques, such as support vector machines (SVM), watershed algorithms, and clustering algorithms. It was discovered that while the watershed technique is thought to be more successful at separating components based on morphological features, clustering algorithms are commonly utilized for blood cell segmentation.

All things considered, the study offers a thorough review of machine learning methods for examining blood smear images in order to identify leukemia. It talks on the difficulties and possibilities involved in creating strong and trustworthy models for leukemia diagnosis and classification, as well as the growing application of machine learning in this area.



## VI. DISCUSSION

The early detection of leukemia is one of the key challenges in aiding in and prolonging life. Consequently, developing a trustworthy detection technique is among the

top priorities. Based on conventional therapy procedures, leukemia assessment and forecasting have proven to be difficult and time-consuming undertakings. Forecast

Because AI and ML are so widely used in the medical field, applying ML algorithms to a digital blood smear image alone has proven very effective and efficient. Developing a paradigm that accurately predicts acute lymphoblastic leukemia (ALL) is the main objective of this effort. To stop ALL from spreading too much, it is essential to predict it in its early phases. A combination of DL, ML, and AI

## VII. CONCLUSION

Leukemia is a type of blood cancer that usually begins inside the bone marrow and creates a large number of aberrant blood cells. There are now four distinct types of leukemia recognized. ALL is one of the common forms of leukemia. In ALL, malignant cells develop and accumulate quickly, necessitating prompt medical intervention.

The conventional process, which entails blood test analysis, genealogical research, and regular medication, takes a long time and may not always produce satisfactory results. If a disease is not identified appropriately, it may spread so fast that it becomes extremely deadly. Machine learning and artificial intelligence are especially useful for overcoming these obstacles. Both recognizing non-blood cells and forecasting the type of sickness are made possible by this automated process, which is extremely successful and efficient. In this work, non-blood cells are identified from digital blood pictures using a color grouping technique.

This technique helps count the non-blood cells by separating the darker non-blood regions from the bloodstream. Additionally, the categorization of non-blood cells is estimated using the three most popular machine learning models. The model with the highest accuracy level is the SVM. A combination strategy is then created to enhance prediction performance. This design achieves accuracy values of over 99 by combining the SVM with the Resnet50 deep neural network framework. In the near future, understanding ALL will require the development of a personalized treatment plan. To generate a prognosis of treatment results that is unique to each patient, the evaluation can take into account a variety of private information, including genomic traits, clinical criteria, and response to treatment statistics. Machine learning (ML) and artificial intelligence (AI) systems work together to forecast the best treatments for each unique condition.

Real-time system tracking is one of the best methods to manage ALL since AI and ML systems are the most advantageous in this prediction aspect.Regularly checking on a patient's condition and response to treatment is crucial. The creation of automated monitoring systems that give medical personnel real-time information on patients' conditions could be the subject of future research. Since the sickness has been known to worsen quickly, this would enable the medical staff to treat it quickly before it worsens.

**References**

1. LDSVM: Leukemia Cancer Classification Using Machine Learning. (2022). [Article]. Computers, Materials Continua, 2, 3888. https://doi.org/10.32604/cmc.2022.021218.

2. Abhishek, A., Jha, R. K., Sinha, R., Jha, K. (2021). Automated classification of acute leukemia on a heteroge-

neous dataset using machine learning and deep learning techniques. Biomedical Signal Processing and Control, 72, 103341. https://doi.org/10.1016/j.bspc.2021.103341

3. Aby, A. E., Salaji, S., Anilkumar, K. K., Rajan, T. (2024). A review on leukemia detection and classification using Artificial Intelligence-based techniques. In Elsevier Ltd., Cochin University College of Engineering Kuttanad, Cochin University of Science And Technology, Believers Church Medical College Hospital, Computers and Electrical Engineering (Vol. 118, p. 109446) [Journal-article]. https://doi.org/10.1016/j.compeleceng.2024.109446

4. Bose, P., Bandyopadhyay, S. (2024). A Comprehensive Assessment and Classification of Acute Lymphocytic Leukemia. Mathematical and Computational Applications, 29–29, 45–45. https://doi.org/10.3390/mca29030045

5. Faiz, M., Mounika, B. G., Akbar, M., Srivastava, S. (2024). Deep and Machine Learning for Acute Lymphoblastic Leukemia Diagnosis: A Comprehensive Review. ADCAIJ ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL, 13, e31420. https://doi.org/10.14201/adcaij.31420

6. Ghaderzadeh, M., Asadi, F., Hosseini, A., Bashash, D., Abolghasemi, H., Roshanpour, A. (2021). Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images: A Systematic review. Scientific Programming, 2021, 1–14. https://doi.org/10.1155/2021/9933481

7. Jawahar, M., Anbarasi, L. J., Narayanan, S., Gandomi, A. H., Leather Process Technology Division, School of Computer Science and Engineering, School of Electronics Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, University Research and Innovation Center (EKIK), Óbuda University. (2024). An attention-based deep learning for acute lymphoblastic leukemia classification. In Scientific Reports (Vol. 14, p. 17447). https://doi.org/10.1038/s41598-024-67826-9

8. Liu, K., Hu, J. (2022). Classification of acute myeloid leukemia M1 and M2 subtypes using machine learning. Computers in Biology and Medicine, 147, 105741. https://doi.org/10.1016/j.compbiomed.2022.105741

9. Mallick, P. K., Mohapatra, S. K., Chae, G., Mohanty, M. N. (2020a). Convergent learning–based model for leukemia classification from gene expression. Personal and Ubiquitous Computing, 27(3), 1103–1110. https://doi.org/10.1007/s00779-020-01467-3

10. Sulaiman, A., Kaur, S., Gupta, S., Alshahrani, H., Reshan, M. S. A., Alyami, S., Shaikh, A. (2023). ResRandSVM: Hybrid Approach for Acute Lymphocytic Leukemia Classification in Blood Smear Images. Diagnostics, 13(12), 2121. https://doi.org/10.3390/diagnostics13122121