# A Survey on Automated Leukemia Detection and Classification: Exploring Image Processing and Gene Expression Analysis Approaches

Priyush Panwar
Department of Computer Engineering
Shri Govindram Seksaria Institute of Technology and Science
Indore, India

D.A. Mehta
Department of Computer Engineering
Shri Govindram Seksaria Institute of Technology and Science
Indore, India

## ABSTRACT

Leukemia is a haematological disorder which affects blood and bone marrow. Automated leukemia detection is significant for facilitating timely diagnoses and increasing the chances of successful treatment outcomes. Further, the classification of leukemia is crucial because each type demands a distinct treatment strategy. Deep learning and machine learning techniques are being used by computer aided diagnosis systems for their high precision in identifying leukemia from microscopic blood smear images using image processing techniques and gene microarray data using gene expression analysis techniques. However, automation of this task using image processing techniques is challenging due to the uneven structure and overlapping of cells. The large number of genes in microarray data makes the classification challenging, which is accomplished by applying a variety of feature selection techniques. In this study, recent developments in automated leukemia detection and classification employing image processing and gene expression analysis methods are explored. Furthermore, the paper compares and contrasts various techniques to provide a comprehensive overview that can assist in the continuous refinement of the detection and classification process.

### Keywords

Leukemia, Image Segmentation, Convolutional Neural Networks, Feature Selection.

## 1. INTRODUCTION

Blood is an essential element of the human body. White blood cells (WBCs), commonly known as leukocytes, red blood cells (RBCs), and platelets are the three primary components of blood [6]. Leukemia, a life-threatening illness commonly known as blood cancer, results from the rapid proliferation of immature WBCs in the blood and bone marrow. These undeveloped cells are referred to as blast cells. The aberrant production of WBCs interferes with the generation of other key blood components, thereby impacting the blood and immune system.

Depending on how quickly leukemia progresses, leukemia is classified as either acute or chronic. Acute leukemia advances rapidly and is more severe, whereas chronic leukemia develops at a slower pace. Leukemia is further classified as either myeloid or lymphoid based on the originating cell line. These classifications result in four primary types of leukemia: Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL), and Chronic Myeloid Leukemia (CML) [37].

The World Health Organisation (WHO) classifies ALL into B-cell lymphoblastic leukaemia and T-cell lymphoblastic leukaemia based on the distinct collection of proteins called immunophenotypes detected on cells [5]. The most prevalent form of ALL is B-cell lymphoblastic leukemia, while T-cell lymphoblastic leukemia is more common in adults than in children. ALL is also classified on the basis of morphology of cells by French American British (FAB) Classification into L1, L2, L3 categories. AML is characterized by blast cells that are not lymphocytes. AML is classified into eight subtypes, M0-M7, according to the FAB classification [37]. A vast amount of research is being done based on all the above classification to make the detection and diagnosis process automated using machine learning (ML) and deep learning (DL) techniques.

Leukemia is typically diagnosed by a hematologist by manually analyzing blood and bone marrow smears. Researchers have employed a range of methods to identify and categorize leukemia. Image processing techniques include pre-processing to eliminate noise, segmentation to isolate blast cells, feature extraction, and classification. Automated Image analysis method is cheap, fast and reduces dependency on experts. Gene expression data is crucial in analyzing diseases and diagnosing cancer. However, due to the presence of large number of genes in gene expression data, it becomes challenging to classify data accurately [24]. To detect leukemia using gene expression data, the most significant step is feature selection. Feature selection is an essential step for any classification process as it reduces computational time and cost and improves classification accuracy by selecting a relevant gene subset as a classifier. After feature selection, the data is classified using appropriate classification technique.

## 2. AUTOMATED LEUKEMIA DETECTION USING BLOOD SMEAR IMAGE ANALYSIS

Computer vision techniques have been employed to identify hematological abnormalities. The primary methods for analysing medical images to produce more precise disease diagnoses are ML and DL methods. Computer aided diagnosis (CAD) systems utilize image processing techniques to automate the identification of lymphoblasts. The methods utilised can be divided into three major categories: pure DL, manual feature extraction with DL and manual feature extraction with shallow ML classifiers. The process of detecting leukemia using CAD involves five stages: image acquisition, image pre-processing, image segmentation, feature extraction, and classification of leukemia cells. Pre-processing is used to boost the quality of the images. By removing platelets and RBCs, segmentation aims to separate overlapping cells and extract the targeted WBCs [10]. After that, pertinent features are extracted, and a classifier is employed to produce a classification performance that is more effective.

### 2.1 Image Acquisition

There are a variety of datasets available for analysis of correctness of output obtained by various techniques. Public Datasets are available for experimentation with algorithms for diagnosis of leukemia. Private Datasets are those datasets which are constructed by researchers for there specific research and not made available publicly.

*2.1.1 Public Datasets.* ALL-IDB[28] and The American Society of Hematology (ASH) Imagebank [21] are the two most common datasets used for leukemia diagnosis. ALL-IDB contains labelled blood cell images of only ALL subtype. ALL-IDB1 contains around 39000 blood elements, with lymphocytes identified by oncologists. The ALL-IDB2 dataset, which includes cropped regions of interest from blast and normal cells contained in the ALL-IDB1 dataset, was created to evaluate the accuracy of classification methods. ASH imagebank maintains an extensive online image repository with haematological images taken with different microscopes at varying resolutions. The ASH imagebank dataset encompasses images of all forms of leukemia, including ALL, AML, CML and CLL.

*2.1.2 Private Datasets.* Public datasets do not contain a sufficient quantity of blood smear images for all leukemia categories to enable classification. A substantial amount of data is necessary to train machine and deep learning algorithms. As a result, researchers often gather their own private datasets from local hospitals.

By applying random changes to pre-existing training images, the approach of data augmentation is used to create a large training dataset. To improve the amount of data accessible to assist with classification, data augmentation techniques are used [3]. Image transformation methods including rotation, shearing, blurring, mirroring, translation, and grey scale picture transformation are among the different data augmentation that can be applied.

### 2.2 Image Pre-Processing

Image pre-processing enhances low-quality microscopic images by filtering noisy elements. This results in increased performance prior to classification. It is a essential initial step before applying ML and DL algorithms [16]. Noise refers to unwanted variations in the pixel values of an image that can degrade the quality of the image. Noise in a image may be caused by various factors including sensor noise, transmission errors. Noise is a major factor affecting the accuracy and performance of model. Pre-processing includes various techniques like Histogram Equalization, Linear Contrast Stretching, Unsharp Masking, Gaussian filtering, Median filtering [37].

*2.2.1 Median filtering.* Median filtering, a non-linear filtering approach, that employs the exquisite strategy of replacing the value of each pixel with the median value derived from the array of pixels in its immediate proximity. This inspiring technique excels magnificently in its ability to expunge the pernicious salt-and-pepper noise with utmost precision and efficacy.

*2.2.2 Gaussian filtering.* Gaussian filtering is a linear filtering method that smooths an image by replacing each pixel with a weighted average of the pixels in its immediate vicinity. It is effective in removing Gaussian noise, which follows a Gaussian distribution.

*2.2.3 Denoising autoencoders.* Denoising autoencoders are a category of neural network used to reconstruct a clean image from a noisy image. They can be effective at removing various types of noise from images.

Genovese et al. [17] focused on improving pre-processing rather than classification, as accurate pre-processing can improve the overall accuracy and performance of the model. To normalise the cell radius, assess focus quality, adaptively increase image sharpness, and perform classification, Francis et al. [16] applied image processing methods and deep learning. Four steps made up their improved pre-processing: cell radius normalisation using image registration, focus quality estimation and adaptive image unsharpening, shallow convolutional neural networks (CNNs) for fine-tuning adaptive image unsharpening, and final adaptive image unsharpening.

### 2.3 Image Segmentation

Image Segmentation involves dividing image into required sections or extracting the desired part of image [29]. In leukemia diagnosis, WBCs in blood cells are an important component. Segmentation is used to extract immature WBCs from blood cell images. Segmentation highlights the tumor area before classification is performed. There are two types of methods used for segmentation: traditional methods and DL-based methods. The latter, which utilize DL techniques, provide higher accuracy and better performance while requiring less time for computation.

*2.3.1 Thresholding.* Thresholding involves the conversion of a gray scale image into a binary image by segmenting foreground from background regions. The image will have either a zero or one value based on the intensity value relative to the threshold. If the intensity is above the threshold, the image will have a value of one, represented by white color, and if it is below the threshold, it will have a value of zero, represented by black color.

*2.3.2 Region Growing.* Region growing is a image segmentation technique in which an initial seed point is selected and each of the neighbouring seed point is added to same region on the basis of similarity. This process goes on until a stopping criterion is reached. The characteristic of similarity may be selected as colour, texture or intensity value. The major challenge in this technique is to choose the initial seed point.

*2.3.3 Watershed Segmentation.* Watershed segmentation is an image processing technique employed to separate and segment distinct regions within an image by utilizing intensity values or color. It is based on the mathematical idea of watersheds, which is the area of a picture where pixel intensity is constant. The basic idea behind the algorithm is that it transforms an image into a topographic surface, where the intensity of the pixels represents the height of the surface. It then uses markers or seeds, placed at the desired segmentation boundaries, to flood the basins and separate the image into different regions [40]. The watershed segmentation is sensitive to noise and the placement of markers can be critical to the algorithm's performance.

*2.3.4 Morphological Segmentation.* A predetermined structural element is moved over an image during morphological segmentation in a manner analogous to sliding windows [34]. The morphological operations utilized in the detection of leukemia comprise erosion, dilation, opening, closing, and hole filling.

*2.3.5 K-means Clustering.* It is the most common method used for segmentation. Each data point can be part of only one cluster [30][31]. The clusters are made on the basis of similarity between points.

*2.3.6 Fuzzy C-means.* It can be viewed as flexible k-means clustering because it allows points to be part of multiple clusters [8].

## 2.4 Feature Extraction

For medical image analysis using DL, it is essential to extract and reduce features from the input images. This process involves identifying relevant features such as texture, shape, and intensity, and then reducing the dimensionality of the feature space by removing unnecessary or redundant features. Feature reduction is often done using techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [10]. In DL, CNNs are used for feature extraction and methods like Dropout or L1/L2 regularization are used for feature reduction. Additionally, methods based on transfer learning are frequently used to extract more prominent features that are essential for classification.

## 2.5 Detection and Classification

The detection and classification phase through medical imaging involves identifying and categorizing WBCs, specifically those that may be leukemia cells and distinguishing between cancerous and non-cancerous cells. The last step involves classifying WBCs as either healthy or cancerous.

CNNs [25] , a type of deep learning model, have been used effectively in leukemia detection in medical imaging. CNNs excel at processing and analyzing image data, making them suitable for tasks such as identifying patterns in microscopic images of blood cells. In the context of leukemia detection, CNNs can be trained to classify WBCs as healthy or cancerous based on their morphological characteristics. Many research studies have used CNNs for this purpose, with some achieving high accuracy rates in detecting leukemia. Also, many studies have employed trained models like ResNet-50 and VGG-16 [38] [22], fine-tuned to the specific task dataset. Furthermore, CNNs can be used in combination with other techniques, such as traditional ML algorithms, to enhance leukemia detection systems performance.

# 3. AUTOMATED LEUKEMIA DETECTION USING GENE EXPRESSION DATA

Obtaining gene expression data for leukemia patients typically involves microarray analysis, which requires collecting a small amount of blood or bone marrow from the patient and processing it to extract RNA molecules. Gene expression data is critical for understanding the genetic basis of leukemia [24], as distinct gene expression patterns are associated with different types of the disease. In addition, gene expression profiling can be useful for predicting the prognosis of leukemia and guiding treatment decisions. By analyzing hundreds or thousands of genes at once, researchers can identify sets of genes that are co-regulated or dysregulated in leukemia and use this knowledge to develop innovative diagnostic and therapeutic strategies.

Automated leukemia detection using gene expression data involves analyzing the expression levels of numerous genes concurrently to identify patterns that suggest the existence or nonexistence of leukemia. However, due to the vast number of genes implicated, accurately classifying leukemia based on gene expression data is challenging. The initial step in leukemia detection using gene expression data is to perform appropriate pre-processing techniques on the data to eliminate noise and other undesired variations. Once pre-processing is completed, feature selection is carried out to identify a subset of genes that are most informative for classification. This not only decreases computational time and cost but also increases classification accuracy. In the end, the samples can be classified using algorithms based on the chosen features. Numerous datasets are accessible for evaluating the accuracy of gene expression analysis techniques. Some of them are Gene Expression Dataset by Golub et al., Curated Microarray Database (CuMiDa) and GEO platform public database.

## 3.1 Feature Selection

Identifying the most pertinent and important genes for addressing the classification problem is the main goal of feature selection. Filters, wrappers and hybrid methods are used to minimise the feature space and improve the performance of the model in order to achieve this goal. Each method has a particular purpose and approach for dealing with genes.

*3.1.1 Filter Methods.* Filter methods are frequently used as a preliminary stage in feature selection to decrease dimensionality [4]. They generally compute a score indicating the relevance of each feature or gene, order them according to their scores, and exclude those with low scores. This technique offers several benefits, including improved versatility with less computational demands, making it well-suited for high-dimensional datasets. Some of the feature selection methods are Information gain, Mutual information, Random forest ranking, Fast Correlation-Based Filter.

*3.1.2 Wrapper Methods.* Wrapper methods utilize both a learning algorithm and a classifier to identify the optimal subset of features. In this method, a subset of features are chosen by searching through the primary feature space. Wrapper approaches are superior than feature-ranking algorithms because they take the classifier hypothesis into account, despite their high computational requirements and unsuitability for high-dimensional datasets. The Genetic Algorithm, Flower Pollination Algorithm, and Particle Swarm Optimisation are some examples of wrapper feature selection techniques.

### 3.2 Classification

To evaluate the performance of different algorithms, their ability to correctly classify data is compared using predictive modeling. An essential measure for assessing the quality of any model's performance is its classification accuracy, which indicates how accurately it predicts various classes. To perform classification, a ML or DL algorithm is trained on a dataset of gene expression data that includes examples of both leukemia patients and healthy individuals. The algorithm is trained to recognize patterns in the data that indicate leukemia and uses this knowledge to make predictions on new data.

## 4. LITERATURE SURVEY

Researchers have explored various methods for detecting leukemia using image processing and gene expression microarray data analysis. Several effective models have been developed. Table I presents a comparison of various algorithms for segmentation, feature extraction, detection, and classification. Table II lists numerous algorithms for gene expression data analysis, feature selection, and classification.

Ain et al. [3] used median filtering and performed filtering using 3 x 3 size median filters for removal of noise and improvement of image quality. Segmentation was performed using DeepLabv3+. ResNet and MobileNetv2 were used as the pre-trained CNN in DeepLabv3+. Leukemia classification was performed using Alex-Net and ResNet-34. Boreiri et al. [8] implemented a convolutional neuro-fuzzy network for classification. Median Filtering was performed followed by unsharp masking for pre-processing of images. Before performing segmentation microscopic images were transformed to L*a*b colour space from RGB colour space. Colour based clustering is implemented using two stage segmentation, for separation of leukocytes from other blood components. Das et al. [11] used a model based on an Orthogonal SoftMax Layer (OSL) and ResNet 18 for acute leukemia detection. B-Cell and T-Cell ALL have been classified using Alexnet in accordance with the WHO categorization by Anilkumar et al. [5]. Aftab et al. [1] proposed a method using BigDL library in Apache Spark and transfer learning for leukemia diagnosis. They have converted the image to gray scale followed by applying normalization and resizing. Classification is performed using GoogleNet model using softmax activation function.

Rajeswari et al. [33] have proposed method for classification of all the subtypes of leukemia using a hybrid model. A hybrid model uses the ensemble averaging technique and offers results with more precision. They have used a hybrid of Inception V3 and Xception models. Initial pre-processing is applied by performing contrast stretching. Segmentation is done on image using Otsu's thresholding and edge detection. Negm et al. [30] have used two different techniques, Neural Network and Decision Tree. Neural Network model gave better results compared to the decision tree model. Decision tree model worked faster compared to neural network. They have extracted 26 cell features for classification.

Shafique et al. [35] focussed on the subtype classification of ALL and classified ALL according to the FAB classification into L1, L2, L3. They used Alexnet, whose last layers were replaced by new layers capable of classifying the input images into four groups. Shafique et al. [36] have transformed the RGB image into CMYK colour model and applied histogram equalization over the images for overcoming the lightning effects. Zack's Algorithm is applied for segmentation to extract the desired WBCs. Kassani et al. [23] have applied normalization, resizing and data augmentation as pre-processing steps. This pioneering methodology seamlessly integrates low-level features extracted from intermediate layers, culminating in the creation of a discerning high-level feature map. They have harnessed the combined power of two formidable deep learning models, namely MobileNet and VGG16, through ensembling.

The SN-AM dataset was used by D. Kumar et al. [26] to classify Multiple Myeloma (MM) and ALL. The data is normalized and shuffled before division between testing and training data. Classification is performed using proposed DCNN. Support Vector Machine for classification was used by Das et al. [12] to identify leukaemia. The extraction of important elements, such as shape, colour, and texture features, which are crucial in the classification step, was emphasised. The Contrast Limited Adaptive Histogram Equalisation (CLAHE) method is employed to enhance the image quality. PCA is used to maintain more important features when dimensionality is reduced. M. Claro et al. [9] developed Alert Net-RWD, a convolutional neural network model for the automated diagnosis of acute lymphoid and acute myeloid leukaemia. The proposed model's reduced file size makes it more appealing for use in mobile device apps.

Bhadra et al. [7] developed a unique unsupervised feature selection method that incorporates hierarchical feature clustering and singular value decomposition (SVD). Their technique uses hierarchical clustering to divide the features into a number of groups, then applies SVD to each group to determine which feature is the most important one based on SVD-entropy. The resulting subset of chosen features enhances interdependence with nearby unselected features while reducing dependence on itself. A randomised ensemble approach was presented by Koul et al. [24] for choosing characteristics from cancer gene expression data. Recursive feature removal and mutual information are combined in this method. As classifiers, linear SVM and logistic regression were employed.

Based on the discussion above on leukemia detection using image processing and gene expression analysis, image processing based leukemia detection methods need segmentation techniques in order to function as expected. Leukemia diagnosis is now more effective due to recent developments in deep learning, notably transfer learning, which, in contrast to traditional deep learning schemes, can deliver promising results even in small data sets. In the analysis of gene expression microarray data, feature selection methods are commonly employed to identify informative genes.

## 5. DISCUSSION

Table I shows that image processing with a focus on segmentation, preceded by pre-processing, is generally used to detect leukemia. According to literature, thresholding is a common method of segmentation in studies of blood smear analysis, such as for leukemia. Clustering-based segmentation techniques like K-means and Fuzzy C-means are frequently used in leukemia diagnosis and classification studies for precisely segmenting blast cells. Various techniques have been employed during the segmentation stage to improve leukemia cell detection, resulting in better accuracy levels. In leukemia detection through microarray data analysis, the

Table 1. Leukemia detection and classification based on image processing

| Author | Dataset | Area of research | Segmentation | Feature Extraction | Classifier | Accuracy |
|---|---|---|---|---|---|---|
| S. Rajeswari et al., [33] | ALL-IDB, ASH ImageBank | Detection and Classification of Various Types of Leukemia | Otsu's Thresholding | N/A | Inception, Xception convolutional models | 91.71% |
| Mishra et al., [31] | Private Dataset | Automated leukaemia detection using microscopic images | K-means Clustering | Geometrical, Colour, Size, Texture | SVM | 93.57% |
| Negm et al., [30] | ALL-IDB, Private Dataset | Classification of Acute Leukemia | K-means clustering | Geometrical, Colour, Size, Texture | Decision tree, Artificial Neural Network | 99.517% |
| P. Kumar et al., [27] | ASH ImageBank | Detection of Acute Myeloid Leukemia | K-Means clustering | Shape and Texture Features. | SVM | 95% |
| Shafique et al., [35] | ALL-IDB 2 | Detection of Acute Lymphoblastic Leukemia and Subtype Classification | N/A | N/A | AlexNet | ALL - 99.50%,<br>ALL Subtypes - 96.06% |
| Shafique et al., [36] | ALL-IDB | ALL detection | Zack's Algorithm | Shape and colour features | Support Vector Machine (SVM) | 93.70% |
| Kassani et al., [23] | C-NMC [19] | Hybrid Deep Learning Architecture for Leukemic B-lymphoblast Classification | N/A | N/A | Hybrid CNN – MobileNet and VGG16 | 96.17% |
| Ahmed et al., [2] | ALL-IDB and ASH ImageBank | Detection and Subtype Classification of Leukemia | N/A | N/A | CNN | Leukemia - 88.25%,<br>Subtype - 81.74% |
| D. Kumar et al., [26] | SN-AM Dataset [20] | ALL and Multiple Myeloma (MM) Classification | N/A | N/A | DCNN | 97.2% |
| Das et al., [12] | ALL-IDB1 | Detection and Classification of Acute Lymphocytic Leukemia | Color based K-means clustering | Texture and Shape Features. | SVM | 96.00% |
| M. Claro et al., [9] | Hybrid of 16 Datasets | Leukemia Classification using Hybrid Datasets | N/A | N/A | Modified CNN called Alert Net-RWD | 97.18% |
| Aftab et al., [1] | ASH ImageBank | Using Transfer Learning to Execute Spark BigDL for Leukemia Detection | N/A | N/A | GoogleNet deep transfer learning | 97.33% |
| Anilkumar et al., [5] | ASH ImageBank | Automated Detection of Acute Lymphoblastic Leukemia in B Cells and T Cells | N/A | N/A | AlexNet | 94.12% |
| Das et al., [11] | ALL-IDB1, ALL-IDB2, ASH ImageBank, CNMC 2019 Dataset | Acute Leukemia Detection and Classification Using Orthogonal Softmax Layer-based Model | N/A | ResNet 18 | OSL-based classification | ALL-IDB1 - 99.09%,<br>ALL-IDB2 - 98.08%,<br>ASH ImageBank - 97.5%,<br>C-NMC 2019 - 89.67% |
| Boreiri et al., [8] | ALL-IDB1, ALL-IDB2 | Acute Leukemia Classification using a Convolutional Neuro-Fuzzy Network | Fuzzy-based two-stage color segmentation algorithm | N/A | Convolutional neuro-fuzzy network | 97% |
| Ain et al., [3] | ALL-IDB | Diagnosing Leukemia Disease using Deep Learning | N/A | N/A | ResNet, Alex-Net | ResNet – 98.4%,<br>Alex-Net – 96.8% |

high dimensionality of gene expression is reduced using filters and wrappers methods by selecting relevant genes. Filters use each gene's statistical properties to determine its ability to differentiate between classes. While the computation is quick, the predictions may be inaccurate. Wrappers use a specific classifier to select genes that maximize classification accuracy, but the computation can be demanding. Hybrid methods for feature selection are being applied to yield accurate results. Detection algorithms based on gene expression analysis are focused on feature selection. Studies that have skipped the feature selection step have included

Table 2. Leukemia detection and classification based on gene expression analysis

| Author | Dataset | Area of research | Feature Selection | Classifier | Accuracy |
|---|---|---|---|---|---|
| Koul et al., [24] | Gene Expression Dataset [18] | Ensemble Feature Selection from Gene Expression Data | Ensemble of MI and recursive feature elimination | Logistic Regression<br>and Linear SVM | 316 features - 99%,<br>4 features - 95% |
| Ashok et al., [13] | Gene Expression Dataset [18] | Microarray Gene Expression Data for Effective Cancer Classification | N/A | Artificial neural network | 98% |
| Patel et al., [32] | CuMiDa Database [14] | Analyzing Leukemia Gene Expression using Multi-Classifier from Curated Microarray Database (CuMiDa) | PCA | KNN, SVM, Decision<br>Tree, Random Forest, Logistic Regression | Logistic Regression<br>- 92% |
| Rita et al., [15] | GEO platform public database | Leukemia feature selection using a DL and genetic algorithm | Bayesian feature selection, Autoencoders | DNN, SVM | Autoencoder + DNN - 92% |
| Simsek et al., [39] | Gene Expression Dataset [18] | Classifying Leukemia Sub-Types using Machine Learning Techniques | N/A | Linear Discriminant, KNN, SVM and Ensemble Classifiers | SVM - 98.6% |

irrelevant or redundant features in the dataset, increasing the risk of overfitting. In this situation, the model gets excessively complex and performs admirably on training data but poorly on test data.

Many studies have focused on the segmentation step, which involves extracting relevant cells from blood smear images. Segmentation is a crucial step, but the uneven structure of blast cells makes it complex and achieving accurate results becomes challenging. Detection algorithms based on image processing that skipped the segmentation step are affected by the presence of irrelevant or overlapping cells, leading to inaccurate results. Additionally, processing the entire image without segmentation results in a significant increase in computational complexity.

Segmentation and feature selection have been highlighted as essential procedures for precise results in the difficult challenge of detecting leukaemia through image processing and gene expression analysis. Therefore, in order to get optimal results, it's crucial to carefully evaluate the best strategies and methodologies for each stage of the detection process.

## 6. CONCLUSION

In order to detect and classify leukemia, this study explores the use of ML and DL approaches. Researchers are continuously investigating effective segmentation algorithms to precisely segment unevenly shaped blast cells. To achieve better results, pre-trained CNNs and their modified variants are being used. The detection of all subtypes of leukemia, particularly chronic leukemia and its subtypes, has been the subject of relatively few studies. This can be attributed to the insufficient availability of data for classification purposes. Thus, there is a need to collect and analyze data pertaining to these less studied subtypes, in order to facilitate the development of highly accurate and efficient detection methods.

Hybrid feature selection methods are being employed to analyze gene expression data. Additionally, researchers are also exploring ensemble and hybrid methods for feature extraction, combining filter and wrapper methods. Efficient and accurate feature selection methods are critical and there is a need for better hybrid feature selection methods that can be applied to multiple datasets for more generalized results.

## 7. REFERENCES

[1] M.O Aftab, M Javed Awan, S Khalid, R Javed, and H Shabir. Executing spark bigdl for leukemia detection from microscopic images using transfer learning. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 216–220, 2021.

[2] N. Ahmed, A. Yigit, Z. Isik, and A. Alpkocak. Identification of leukemia subtypes from microscopic images using convolutional neural network. *Diagnostics*, 9(3):104, August 2019.

[3] Q. Ul Ain, S. Akbar, S. A. Hassan, and Z. Naaqvi. Diagnosis of leukemia disease through deep learning using microscopic images. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pages 1–6, 2022.

[4] H Almazrua and H Alshamlan. A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data. *IEEE Access*, 10:71427–71449, 2022.

[5] K K Anilkumar, V.J. Manoj, and Sagi M. Automated detection of b cell and t cell acute lymphoblastic leukaemia using deep learning. *IRBM*, 43:405–413, 2021.

[6] R.G Bagasjvara, Ika Candradewi, Sri Hartati, and Agus Harjoko. Automated detection and classification techniques of acute leukemia using image processing: A review. pages 35–43, 10 2016.

[7] T Bhadra, S Mallik, A Sohel, and Z Zhao. Unsupervised feature selection using an integrated strategy of hierarchical clustering with singular value decomposition: An integrative biomarker discovery method with application to acute myeloid leukemia. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3):1354–1364, may-june 2022.

[8] Z Boreiri, A.N Azad, and A Ghodousian. A convolutional neurofuzzy network using fuzzy image segmentation for acute leukemia classification. In *2022 27th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–7, 2022.

[9] Maíla Claro, Luis Vogado, Rodrigo Veras, André Santana, João Tavares, Justino Santos, and Vinicius Machado. Convolution neural network models for acute leukemia diagnosis. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 63–68, 2020.

[10] P. K. Das, D. V. A, S. Meher, R. Panda, and A. Abraham. A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia. *IEEE Access*, 10:81741–81763, 2022.

[11] P.K Das, B Sahoo, and S Meher. An efficient detection and classification of acute leukemia using transfer learning and orthogonal softmax layer-based model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[12] Pradeep Kumar Das, Priyanka Jadoun, and Sukadev Meher. Detection and classification of acute lymphocytic leukemia. In *2020 IEEE-HYDCON*, pages 1–5, 2020.

[13] A.K. Dwivedi. Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Comput Applic*, 29:1545–1554, 2018.

[14] B.C et al Feltes. Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 2019.

[15] R Francese, M Frasca, M Risi, and G Tortora. A deep learning and genetic algorithm based feature selection processes on leukemia data. In *2022 26th International Conference Information Visualisation (IV)*, pages 412–417, 2022.

[16] E. U. Francis, M. Y. Mashor, R. Hassan, and A. A. Abdullah. Screening of bone marrow slide images for leukemia using multilayer perceptron (mlp). In *Proc. IEEE Symp. Ind. Electron. Appl.*, pages 643–648, September 2011.

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

[18] T.R et al Golub. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring". *Science*, pages 531–537, 1999.

[19] A. Gupta and R. Gupta. All challenge dataset of isbi 2019. The Cancer Imaging Archive, 2019.

[20] A. Gupta and R. Gupta. Sn-am dataset: White blood cancer dataset of b-all and mm for stain normalization. The Cancer Imaging Archive, 2019.

[21] American Society Of Haemotology. Ash-dataset. `http://imagebank.hematology.org/`.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] S.H Kassani, P.H Kassani, M.J Wesolowski, K.A Schneider, and R Deters. A hybrid deep learning architecture for leukemic blymphoblast classification. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 271–276, 2019.

[24] N. Koul and S. S. Manvi. Ensemble feature selection from cancer gene expression data using mutual information and recursive feature elimination. In *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–6, Bengaluru, India, 2020.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[26] Deepika Kumar, Nikita Jain, Aayush Khurana, Sweta Mittal, Suresh Chandra Satapathy, Roman Senkerik, and Jude D. Hemanth. Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks. *IEEE Access*, 8:142521–142531, 2020.

[27] P Kumar and S.M Udwadia. Automatic detection of acute myeloid leukemia from microscopic blood smear image. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1803–1807, 2017.

[28] R. D. Labati, V. Piuri, and F. Scotti. All-idb: The acute lymphoblastic leukemia image database for image processing. In *Proc. 18th IEEE Int. Conf. Image Process.*, pages 2045–2048, September 2011.

[29] A.S Abdul Nasir, N Mustafa, and N.M Nasir. Application of thresholding technique in determining ratio of blood cells for leukemia detection. In *Proc Int Conf Man-Mach Syst (ICoMMS)*, pages 1–6, Penang, Malaysia, October 2009.

[30] A.S Negm, O.A Hassan, and A.H Kandil. A decision support system for acute leukaemia classification based on digital microscopic images. *Alexandria Eng J*, 57(4):2319–2332, December 2018.

[31] N Patel and A Mishra. Automated leukaemia detection using microscopic images. *Proc Comput Sci*, 58:635–642, January 2015.

[32] S Patel, H Patel, D Vyas, and S Degadwala. Multi-classifier analysis of leukemia gene expression from curated microarray database (cumida). In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1174–1178, 2021.

[33] S. Rajeswari, Ch. Siva Vasanth, Ch. Bhavana, and K Sethu Sandeep Chowdary. Detection and classification of various types of leukemia using image processing, transfer learning and ensemble averaging techniques. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6, 2022.

[34] A. Rehman, N. Abbas, T. Saba, S. I. U. Rahman, Z. Mehmood, and H. Kolivand. Classification of acute lymphoblastic leukemia using deep learning. *Microsc. Res. Technique*, 81(11):1310–1317, November 2018.

[35] S. Shafique and S. Tehsin. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technol. Cancer Res. Treatment*, 17:1533033818802789, September 2018.

[36] S. Shafique, S. Tehsin, S. Anas, and F. Masud. Computer-assisted acute lymphoblastic leukemia detection and diagnosis. In *2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*, pages 184–189, 2019.

[37] A. Shah, S. S. Naqvi, K. Naveed, N. Salem, M. A. U. Khan, and K. S. Alimgeer. Automated diagnosis of leukemia: A comprehensive review. *IEEE Access*, 9:132097–132124, 2021.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[39] Badem H. Okumus I.T Simsek, E. Leukemia sub-type classification by using machine learning techniques on gene expression. In *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems, vol 217. Springer, Singapore, 2022.

[40] E. Suryani, S. Palgunadi, and T. N. Pradana. Classification of acute myelogenous leukemia (aml m2 and aml m3) using momentum back propagation from watershed distance transform segmented images. *J. Phys., Conf. Ser.*, 801:012044, January 2017.