International Research Journal of Education and Technology
Peer Reviewed Journal, **ISSN 2581-7795**

Exploring Machine Learning-Based Automation in Leukemia Diagnosis: A Critical Review

Harmeet Singh Lubana[1], Prof. Virendra Verma[2]
[1]M.tech Scholar, [2]Assistant Professor, [1,2]Department of Computer science Engineering
[1,2]LNCT(Bhopal) Indore Campus, Indore

Abstract: **Blood Leukemia remains among the most lethal diseases globally, marked by a significantly high mortality rate. Early diagnosis poses a major challenge due to the subtle and incomplete manifestation of symptoms in initial stages. Recently, artificial intelligence has emerged as a transformative force in healthcare, offering solutions to problems that traditional methods struggle to address. A key area of advancement is the AI-driven classification of leukemia. Prior research underscores the critical role of classification accuracy in this domain, though achieving consistently high precision remains complex. Various methodologies offer distinct advantages and limitations. This study delivers an in-depth evaluation of machine learning techniques applied to automated leukemia detection, emphasizing the unique strengths of each approach. This paper presents a comprehensive analysis of the various machine learning based approaches employed for automated blood leukemia detection, highlighting the salient features of each approach.**

Keywords—**Blood Leukemia, Microscopic images, machine learning, automated classification, classification accuracy.**

1. INTRODUCTION

Blood Cancer or Leukemia is one of the most dreaded diseases in the world with a high mortality rate. The incidence has been prevalent to a great extent and comes up with very common symptoms at the initial stage of the illness. It is seen that quicker detection of the disease leads to better treatment possibilities and success of the treatment [1]. There has been rampant technological advancement in the field of image processing and

allied technologies that have led to improved medical image clarity and has aided in better diagnosis. Hence this domain of medical technology and classification of the cancer, the type i.e. benign or malignant etc have seen increase in in-depth research and study. The improved and efficient image services increases the accuracy and efficacy of diagnosis. Effective treatment can happen when the disease is detected quickly and accurately. This is possible with high end medical diagnosis and accuracy of diagnosis in less time. Consequently, modern techniques have become the go-to option for the evaluation and detection of this serious blood cancer cases that can predict it accurately and faster than other conventional methods [2].

Blood leukemia is form of cancer that exhibits uncontrolled growth of the white blood cells and is potentially very serious and life threatening form of blood cancer. For detection of the disease, images of the blood samples have to be evaluated by a hematologist for any other than normal feature. The images of blood samples are microscopic in nature and therefore correct diagnosis and identification is dependent on the accuracy and clarity of the images long with the expertise of the hematologist. Serious problem can arise if the images are erroneous because it can lead to incorrect line of diagnosis and incorrect treatments [3]-[4].. Hence this a major concern for accurate diagnosis and detection and proper identification of microscopic images of the blood samples. Time factor is also a major concern and the quicker is the diagnosis done, there is better treatment probability for the illness. Computer

aided diagnosis system is an active tool used for early detection but 10%-30% of patients who have the disease and undergo diagnosis have negative classification. Two-third of these false negative cases was evident retrospectively. These mistakes in the visual interpretation are due to poor image quality, eye fatigue of the radiologist, subtle nature of the findings, or lack experienced radiologists especially in third-world regions. Nowadays the computer-aid systems play the main role in early detection and diagnosis of blood leukemia. Increasing confidence in the diagnosis based on computer-aid systems would, in turn decrease the number of patients with suspected blood cancer who have to undergo surgical blood biopsy, with its associated complications.

Fig. 1 A typical microscopic image [4].

Since detection of blood leukemia in extremely challenging, and yet critically important, hence the motivation of the proposed work is to detect blood leukemia cased with high accuracy. Since manual inspection and detection is prone to errors, automated detection is a strong alternative or at least can cast a strong second opinion. For this purpose, use of Artificial Intelligence and Machine Learning is proposed with an aim to detect blood leukemia cases successfully. The paper is divided into the following parts discussing different aspects of the automated detection mechanism.

2. Automated Detection of Leukemia

The automated detection of leukemia is challenging due to the following reasons:
The disease is a very fatal one if it is not detected and treated in time. This form of cancer has many variants based on the type of cells showing improper behavior and function. Basically the lymphoid cells and the myeloid cells become affected in this illness resulting in the respective types of the disease [5].
Also based on how rapidly is the growth of cells, it can be classified into grades and also into chronic or acute. The chronic form of the illness, it progresses slowly and not very aggressively. But with the case of the acute version of the disease, it aggravates very fast and needs immediate treatment [6]. Hence a major aspect of this illness and solution is very fast line of action in terms of diagnosis as well as treatment.
The symptoms as discussed earlier are of very common types that cannot be guessed immediately. And many a times the illness is asymptomatic and does not show any symptom which is when the detection becomes even more difficult. Also as there are major symptoms that is observed, so the classification of the images of the blood samples also becomes improper and non conclusive.
The microscopic image of the blood has to have a lot of clarity and precision so that the medical expert can detect any tiny variation in the sample of the blood. This detection and evaluation phase is immensely important and has to be performed without any dint of error otherwise the entire thing could go wrong including the treatment. Therefore a mechanism is needed that can detect the images clearly and also the clarity of medical images has to be maintained so that the entire process is accurate and on point [7]-[8].

Based on the image processing and feature extraction, the classification is done. Automated classification requires training a classifier with the pre-defined and labelled data set and subsequently classifying the new data samples. Off late machine learning based classifiers are being used for the classification problems. Machine learning can be crudely understood as the design of automated computational systems which mimic the human

behaviour and can be trained in the sense that they can learn from data fed to the system. Primarily machine learning is categorized into three major categories which are [13]-[15]:

1) Unsupervised Learning: In this approach, the data set is not labelled or categorized prior to training a model. This typically is the most crude form of training wherein the least amount of apriori information is available regarding the data sets.

2) Supervised Learning: In this approach, the data is labelled or categorized or clustered prior to the training process. This is typically possible in case the apriori information us available regarding the data set under consideration.

3) Semi-Supervised Learning: This approach is a combination of the above mentioned supervised and unsupervised approaches. The data is demarcated in two categories. In one category, some amount of the data is labelled or categorized. This is generally not the larger chunk of the data. In the other category, a larger chunk of data is un-labeled and hence the data is a mixture of both labelled and un-labeled data groups.

Some other allied categories of machine learning are:

4) Reinforcement Learning
5) Transfer Learning
6) Adversarial Learning
7) Self-Supervised learning etc.

While these learning algorithms can be studied separately, however they are essentially the modified versions of unsupervised, supervised and semi-supervised learning architectures. A more advanced and useful category of machine learning is deep learning which is the design of deep neural nets with multiple hidden layers.

Machine learning based classifiers are typically much more accurate and faster compared to the conventional classifiers. They render more robustness to the system as they are adaptive and can change their characteristics based on the updates in the dataset [16]. The common classifiers which have been used for the classification of pests are:

Regression Models: In this approach, the relationship between the independent and dependent variable is found utilizing the values of the independent and dependent variables. The most common type of regression model can be thought of as the linear regression model which is mathematically expressed as [15]:

$$y = \theta_1 + \theta_2 x \quad (1)$$

Here,
x represents the state vector of input variables
y represents the state vector of output variable or variables.
Θ1 and Θ2 are the co-efficients which try to fit the regression learning models output vector to the input vector.
Often when the data vector has large number of features with complex dependencies, linear regression models fail to fit the input and output mapping. In such cases, non-linear regression models, often termed as polynomial regression is used. Mathematically, a non-linear or higher order polynomial regression models is described as:

$$y = \theta_0 + \theta_1 x^3 + \theta_2 x^2 + \theta_3 x \quad (2)$$

Here,
x is the independent variable
y is the dependent variable
$\theta_1, \theta_2 \ldots . \theta_n$ are the co-efficients of the regression model.
Typically, as the number of features keep increasing, higher order regression models tend to fit the inputs and targets better. A typical example is depicted in figure 2

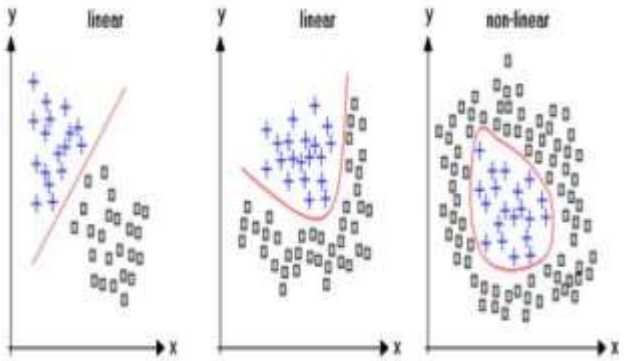

Fig. 2 Linear and Non-Linear Regression fitting [14]

Support Vector Machine (SVM): This technique works on the principle of the hyper-plane which

tries to separate the data in terms of 'n' dimensions where the order of the hyperplane is (n-1). Mathematically, if the data points or the data vector **'X'** is m dimensional and there is a possibility to split the data into categories based on **'n'** features, then a hyperplane of the order **'n-1'** is employed as the separating plane. The name plane is a misnomer since planes corresponds to 2 dimensions only but in this case the hyper-plane can be of higher dimensions and is not necessarily a 2-dimensional plane. A typical illustration of the hyperplane used for SVM based classification is depicted in figure 3.

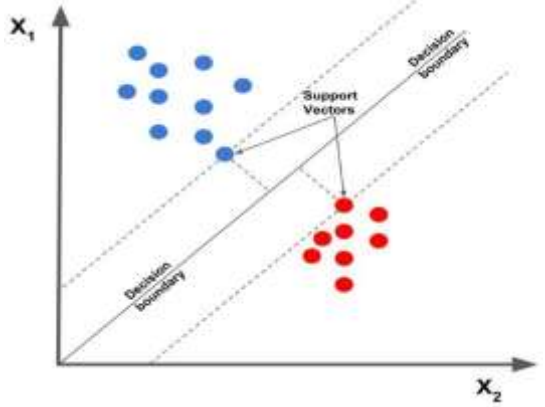

Fig. 3 Separation of data classes using SVM [15]

The selection of the hyperplane H is done on the basis of the maximum value or separation in the Euclidean distance d given by:

$$d = \sqrt{x_1^2 + \cdots \ldots \ldots x_n^2} \quad (3)$$

Here,
x represents the separation of a sample space variables or features of the data vector,
n is the total number of such variables
d is the Euclidean distance

The (n-1) dimensional hyperplane classifies the data into categories based on the maximum separation. For a classification into one of **'m'** categories, the hyperplane lies at the maximum separation of the data vector **'X'.** The categorization of a new sample **'z'** is done based on the inequality:

$$d_x^z = Min(d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z) \quad (4)$$

Here,
d_x^z is the minimum separation of a new data sample from **'m'** separate categories
$d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z$ are the Euclidean distances of the new data sample **'z'** from m separate data categories.

Neural Networks: Owing to the need of non-linearity in the separation of data classes, one of the most powerful classifiers which have become popular is the artificial neural network (ANN). The neural networks are capable to implement non-linear classification along with steep learning rates. The neural network tries to emulate the human brain's functioning based on the fact that it can process parallel data streams and can learn and adapt as the data changes. This is done through the updates in the weights and activation functions. The mathematical model of the neural network is depicted in figure 4.

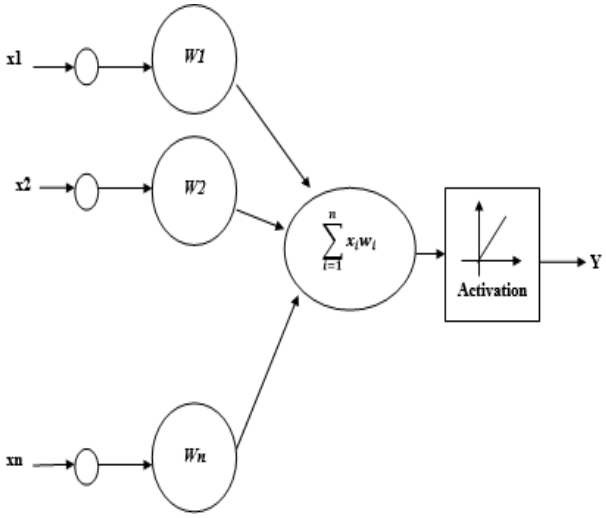

Fig. 4 Mathematical Model of Single Neuron

The mathematical equivalent of an artificial neuron is depicted in figure 4 where the output can be given by:

$$y = f(\sum_{i=1}^{n} x_i w_i + b) \quad (5)$$

Here,
x denote the parallel inputs
y represents the output
w represents the bias
f represents the activation function

The neural network is a connection of such artificial neurons which are connected or stacked with each other as layers. The neural networks can be used for both regression and classification problems based on the type of data that is fed to them. Typically the neural networks have 3 major conceptual layers which are the input layer, hidden layer and output layer. The parallel inputs are fed to the input layer whose output is fed to the hidden layer. The hidden layer is responsible for analysing the data, and the output of the hidden layer goes to the output layer. The number of hidden layers depends on the nature of the dataset and problem under consideration. If the neural network has multiple hidden layers, then such a neural network is termed as a deep neural network. The training algorithm for such a deep neural network is often termed as deep learning which is a subset of machine learning. Typically, the multiple hidden layers are responsible for computation of different levels of features of the data. Several categories of neural networks such as convolutional neural networks (CNNs), Recurrent Neural Network (RNNs) etc. have been used as effective classifiers [17] .

3. PREVIOUS WORK

This section cites the various contemporary approaches employed for automated pest and weed detection in plants. The salient features of each approach in terms of the technique adopted, performance metrics obtained and detected research gaps or limitations are also mentioned for a quick analysis of the contemporary techniques employed in the domain.

Table I. Previous Work.

Authors	Approach Used	Performance	Limitations
J. Denny et al. [1]	Deep Learning based on Convolutional Neural Networks (CNN)	Classification accuracy of 80 achieved	Separate image enhancement not employed.
E. Tuba et al. [2]	Support Vector Machine (SVM) used for classification.	Highest accuracy of 91.8% achieved.	The Support Vector Machine (SVM) suffers from performance saturation.
S.Kumar et al. [3]	K-means clustering followed by K nearest neighbor (KNN) employed.	Accuracy of 92.8% achieved for used dataset.	Feature optimization not performed.
J. Rawat et al. [4]	Multi Layer Perceptron (MLP) kernel based SVM	Accuracy of 91.4%	No feature optimization and noise removal adopted.
Y.Mao et al. [5]	A Deep Convolutional Neural Network (DCNN) with and without HOG features was tested.	F-1 score of 78% and 91% obtained for the two methodologies.	Convolutional Neural Networks are prone to overfitting thereby negatively impacting transfer learning models. Feature optimization nit done.
V. Shankar et al. [6]	Zack algorithm for image enhancement and Euclidean Distance classifier used for classification.	Accuracy of 90% achieved.	No probabilistic classifier used for classification. The Euclidean classifier renders erroneous results for overlapping data samples.

J. Rawat et al. [7]	GLCM features extracted from the images and SVM used for classification.	Classification Accuracy of 87.6% achieved.	Statistical features not computed. SVM inherently suffers from performance saturation.
D. Goutam et al. [8]	K-mean clustering, Local Directional path (LDP), and support vector machine (SVM) used.	F-Measure of 93.4% achived.	Very small testing dataset of 90 images chosen. No feature optimization analysis done.
S. Agaian et al. [9]	LBP and GLCM features computed and SVM used for classification.	Accuracy of 98% obtained for the dataset used.	Image denoising not done. Relatively small testing sample set. SVM prone to performance saturation.
AR Ali et al. [10]	Fuzzy C means approach used for classification of malenoma.	Classification accuracy.	Accuracy not clearly specified. Feature optimization not done.
C. Moschopoulos et al. [11]	A Genetic Algorithm for Pancreatic Cancer Diagnosis	Classification accuracy of 88%.	Image pre-processing and feature optimization not done. GA is a relatively old technique in today's context.
L Putzu et al. [12]	Decision Tress and Support Vector Machine employed.	Accuracy of 76% and 84% obtained respectively.	Both models are prone to overfitting and performance saturation.
S. Mohapatra et al. [13]	Rough-fuzzy hybrid-clustering technique for leukocyte image segmentation	global silhouette index (SL) [40] and partition index (SC) of 0.3624 and 0.0018 achieved	No classification done. Separate noise removal not undertaken.
T Madhloom et al. [14]	An Image Processing Application for the Localization and Segmentation of Lymphoblast Cell Using Peripheral Blood Images	Accuracy of 90-96% achieved	No classification done.
A Perez et al. [15]	Fuzzy Logic controller for classification of Leukemia types.	Accuracy of 93.52% in classification of acute leukemia, 87.36% in lymphoblastic subtypes and 94.42% in myeloid subtypes	Image pre-processing and feature optimization not undertaken.

The performance metrics of the classifiers are generally computed based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values which are used to compute the accuracy and sensitivity of the classifier, mathematically expressed as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Sensitivity: It is mathematically defined as:

$$Se = \frac{TP}{TP+FN} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$Precisiosn = \frac{TP}{TP+FP} \quad (9)$$

$$F - Measure = \frac{2.Precision.Recall}{Precision+Recall} \quad (10)$$

The aim of any designed approach is to attain high values of accuracy of classification along with other associated parameters. The computation complexity of the system often evaluated in terms of the number of training iterations and execution time is also a critically important metric which decides the practical utility of any algorithm on hardware constrained devices.

4. Conclusions

It can be concluded that It can be concluded that AI based techniques can prove to be a strong supporting tool to medical practitioners aiming to detect blood leukemia. Development of such techniques are not aimed at replacing doctors, rather supporting and augmenting them. Several AI and ML based techniques have been proposed with their own strengths and limitations. Different stages of the data processing and segmentation have been enlisted. The significance of different image features and extraction techniques have been clearly mentioned with their utility and physical significance. Various machine learning based classifiers and their pros and cons have been highlighted. The mathematical formulations for the feature extraction and classification gave been furnished. A comparative analysis of the work and results obtained has been cited in this paper. It can be concluded that image enhancement and feature extraction are as important as the effectiveness of the automated classifier, hence appropriate data processing should be applied to attain high accuracy of classification.

Some of the future directions of work can be separate image enhancement and data optimization to avoid both over fitting and under-fitting, moreover, employing separate image denoising to extract features more accurately. Solely computing feature based on deep learning architecture can be compared with statistical feature extraction. This would make the system application to a large variety of datasets. Moreover, classifiers which do not saturate in terms of performance with increasing sat size can be employed.

References

[1] J Denny, MM Rubeena, JK Denny, “Cloud based Acute Lymphoblastic Leukemia Detection Using Deep Convolutional Neural Networks”, Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE (2019), pp. 530-536.

[2] E Tuba, I Strumberger, N Bacanin, D Zivkovic, “Acute Lymphoblastic Leukemia Cell Detection in Microscopic Digital Images Based on Shape and Texture Features”, Advances in Swarm Intelligence. ICSI 2019. Lecture Notes in Computer Science, Springer, vol 11656 (2019), pp: 142-151.

[3] Sachin Kumar ,Sumita Mishra, Pallavi Asthana, Pragya, “Automated Detection of Acute Leukemia Using K-mean Clustering Algorithm, Advances in Computer and Computational Sciences. Advances in Intelligent Systems and Computing, Springer vol 554, (2018) pp: 655-670.

[4] J Rawat, A Singh, HS Bhadauria, J Virmani, “Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia”, Biocybernetics and Biomedical Engineering, Elsevier, vol. 37, no. 4, (2017), pp: 637-654.

[5] Yunxiang Mao , Zhaozheng Yin , Joseph Schober , “A deep convolutional neural network trained on representative samples for circulating tumor cell detection”, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, (2016), pp. 1-6,

[6] Vasuki Shankar , Murali Mohan Deshpande, Automatic detection of acute lymphoblasitc leukemia using image processing", 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, (2016), pp. 186-189.

[7] Jyoti Rawat , H.S. Bhadauria, "Computer Aided Diagnostic System for Detection of Leukemia using Microscopic Images", Procedia Computer Science, Elsevier (2015) vol- 70, pp: 748-756.

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

[9] Sos Agaian , Monica Madhukar., "Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images", IEEE Systems Journal, vol. 8, no. 3, (2014) pp. 995-1004.

[10] AR Ali, MS Couceiro, A E Hassenian, "Melanoma detection using fuzzy C-means clustering coupled with mathematical morphology", 2014 14th International Conference on Hybrid Intelligent Systems, (2014), pp. 73-78.

[11] C. Moschopoulos., D. Popovic, A. Sifrim, G. Beligiannis, B. De Moor, Y. Moreau. "A Genetic Algorithm for Pancreatic Cancer Diagnosis" Engineering Applications of Neural Networks. EANN Communications in Computer and Information Science, Springer, (2013). vol 384, pp.222-230.

[12] L Putzu, C Di Ruberto, "Investigation of different classification models to determine the presence of leukemia in peripheral blood image", International Conference on Image Analysis and Processing (2013) Springer, pp: 612-621

[13] S Mohapatra, D Patra, K Kumar, "Unsupervised leukocyte image segmentation using rough fuzzy clustering", vol-2012,Hindawi Publications (2012), pp: 1-12.

[14] HT Madhloom, SA Kareem, H Ariffin., "An Image Processing Application for the Localization and Segmentation of Lymphoblast Cell Using Peripheral Blood Images", Journal of Medical Systems, Springer (2012), vol-36, pp: 2149–2158.

[15] A Rosales-Pérez, CA Reyes-García "An Improved Detection Algorithm Based on Morphology Methods for Blood Cancer Cells Detection", International Conference on Advances in Artificial Intelligence, (2011) pp:537-548.