International Journal of Artificial Intelligence, Internet of Things and Cloud Computing
Volume: 1 (2022), pp. 58-63

Leukemia Detection employing Machine Learning: A Review andTaxonomy

Tejal Nemade[1]

[1]Research Scholar, Department of Computer Science and Engineering Sushila Devi Bansal College of Technology, Indore, India

Abstract

Blood Leukemia is one of the most deadly diseases in the world with one of the most deadly mortality rates. The detection of blood leukemia at early stages is extremely difficult owing to the fact that leukemia's symptoms do not manifest themselves completely early. Off late, artificial intelligence is being used in several applications of healthcare which are complex to be handled by conventional or traditional techniques. One such domain is the automated classification of leukemia using artificial intelligence based techniques. The study of previous work in the domain shows the fact that the classification accuracy is an extremely important parameter related to automated leukemia classification and attaining high accuracy is a difficult task. Several approaches have their pros and cons in this regard. 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. Theincidence 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 natureand 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[5],[6].

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.

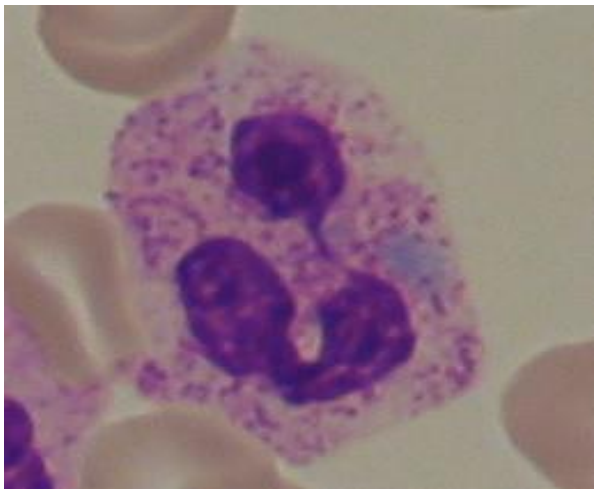

Fig. 1 A typical microscopic image [4].

2. Automated Detection Of Leukemia

The automated detection of leukemia is challenging due to the following reasons:

1) Unsupervised Learning.
2) Supervised Learning.
3) Semi-supervised Learning.

2.1 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:

$$y = \theta\theta_1 + \theta\theta_2 x$$

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:

$$y = \theta\theta_0 + \theta\theta_1 x^3 + \theta\theta_2 x^2 + \theta\theta_3 x$$

Here,

x is the independent variabley is the dependent variable

$\theta\theta_1, \theta\theta_2 \ldots\ldots$ 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

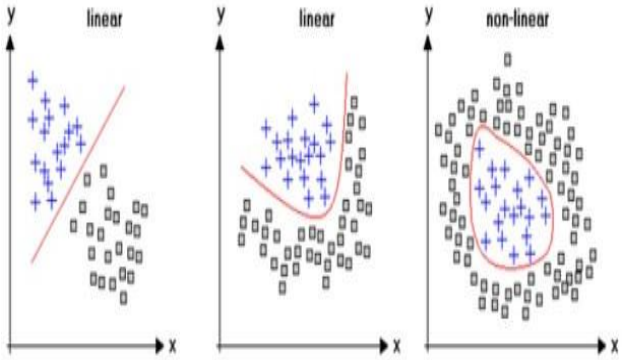

Fig. 2 Linear and Non-Linear Regression fitting

2.2 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.

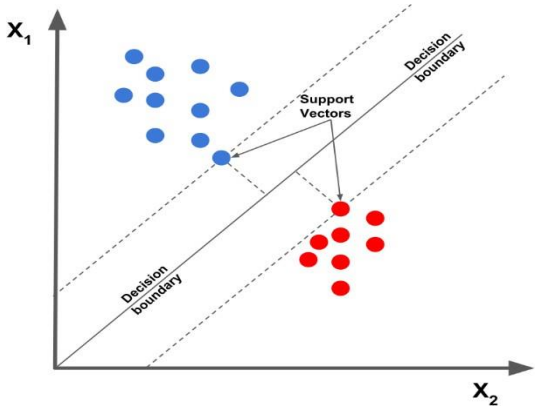

Fig. 3 Separation of data classes using SVM

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

$$d = \sqrt{x_1^2 + \cdots \ldots \ldots . x_n^2}$$

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:

$$d^z_x = MMMn(d^z_{C1}, d^z_{C2} \ldots d^z_{C2=m})$$

Here,

d_x^z is the minimum separation of a new data sample from **'m'** separate categories

$C1$, $C2$ … d^z d^z $C2=m$ are the Euclidean distances of the new data sample **'z'** from m separate data categories.

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 of91.8% achieved.	The Support Vector Machine (SVM) suffersfrom performance saturation.
S.Kumar et al. [3]	K-means clustering followed by K nearest neighbor (KNN) employed.	Accuracy of 92.8% achieved for useddataset.	Feature optimization notperformed.
J. Rawat etal. [4]	Multi Layer Perceptron (MLP)kernel based SVM	Accuracy of 91.4%	No feature optimization and noise removal adopted.
Y.Mao etal. [5]	A Deep Convolutional Neural Network (DCNN) with and without HOG features was tested.	F-1 score of 78% and91% obtained for the two methodologies.	Convolutional Neural Networks are prone to overfitting thereby negatively impacting transfer learning models.Feature optimization nit done.
V. Shankaret al. [6]	Zack algorithm for image enhancement and EuclideanDistance 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.

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

4. Conclusion

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.

References

[1] J Denny, MM Rubeena, JK Denny, "Cloud based Acute Lymphoblastic Leukemia Detection Using Deep Convolutional Neural Networks", 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE (2020), 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.