Machine Learning Algorithms For Diagnosis Of Leukemia

Italia Joseph Maria, T. Devi, D. Ravi

Abstract: Leukemia is cancer of the blood, which includes the bone marrow and the lymphatic tissues, usually involving white blood cells. Unlike usual cancer, leukemia does not form solid tumours, but form large number of abnormal white blood cells which crowd out the normal blood cells. Machine Learning algorithms are largely employed in the treatment of Leukemia, be it for classification of different leukemia types or for detecting if leukemia is present in a patient. This paper describes Support Vector Machines, k-Nearest Neighbour, Neural Networks, Naïve Bayes and Deep Learning algorithms which are used to classify leukemia into its sub-types and presents a comparative study of these algorithms.

Index Terms: Comparison of Machine Learning Algorithms, Leukemia Diagnosis, Leukemia Classification, Machine Learning

1. INTRODUCTION

EXRACTING information about white blood cells (WBCs) is very important for hematologists because WBCs play a crucial role in detecting many diseases, mainly leukemia [1]. Leukemia is unwanted or abnormal growth of white blood cells. As a primary step, taking blood counts can help hematologists in screening leukemia. Based on cell type and rate of growth, Leukemia is of four main types, Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL) and Chronic Myeloid Leukemia (CML) [2], of which ALL is more common in children and AML is the most common type of acute leukemia in adults. Based on morphology and cytochemical staining of blasts, the French–American–British (FAB) classification systems classify AML into 8 sub-types (M0 – M7) and ALL into 3 sub-types (L1 – L3). The differentiation based on morphology includes cell size, prominence of nuclei, cell colour and amount and appearance of cytoplasm [2]. Table 1 gives the FAB classification. The algorithms taken for study in this paper, takes as input, colour images of stained blood smears, preprocesses them, performs image segmentation, feature extraction and classifies whether the patient is affected by leukemia, whether the patient suffers from AML or ALL, or which subtype of AML or ALL the patient suffers from. Segmentation algorithms, feature extraction techniques and classification algorithms are used for the purpose of preprocessing and this paper is an outcome of a comparative study of the classification algorithms from the literature.Classification algorithms can be compared upon criteria such as whether they are supervised or unsupervised, whether they work on small or big datasets, whether it is a binary classifier or not, whether it can work with large number of dimensions, should the problem space be linearly separable and what be the accuracy obtainable.

Table 1. FAB Classification

Category	Name	General Description
ALL (3)	L1	Small monotonous lymphocytes
	L2	Small monotonous lymphocytes
	L3	Large homogeneous blast cells
AML (8)	M0	Acute myeloblastic leukemia, undifferentiated
	M1	Acute myeloblastic leukemia, without maturation
	M2	Acute Myeloblasts with maturation (best AML prognosis)
	M3	Acute promyelocytic leukemia
	M4	Acute myelomonocytic leukemia
	M5	Acute monocytic leukemia
	M6	Erythroleukemia/DiGuglielmo syndrome
	M7	Acute megakaryoblastic leukemia

2 CLASSIFICATION ALGORITHMS USED IN LEUKEMIA TREATMENT

2.1 Support Vector Machines

SVM is a binary classification algorithm and is used to classify sample blood images to lymphoid stem cells and myeloid stem cells which mark them as Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia. Using SVM, Laosai et al. (2014) achieved an accuracy of 92%. In SVM, the inputspace of the dataset is separated by a separating surface which optimizes the margin between the classes. The support vectors are found out by the classifier training algorithm. For each nucleus, sample shape and texture based features are extracted and recorded. The most relevant features are selected from all the features and are used to train the SVM. The number of nuclei lobes, ratio of nuclei to cell, ratio of perimeter to nuclei and entropy are the relevant features selected. Of these,

- *Italia Joseph Maria is currently pursuing doctoral degree program in Computer Applications in Bharathiar University, India, PH-+919447678070. E-mail: reginasabs19@gmail.com*
- *T. Devi is currently working as Professor and Head in Computer Applications department in Bharathiar University, Coimbatore, India, PH-+91 9790004351. E-mail: tdevi5@gmail.com*
- *D. Ravi is currently working as Associate Professor in the PG and Research Department of Botany in Government Arts College, Coimbatore, India, E-mail: dravi_botany@hotmail.com*

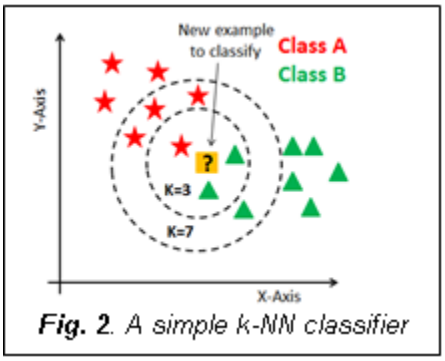

Fig. 2. *A simple k-NN classifier*

number and structure of the nuclei lobe are the prominent feature which are used to identify the class of the WBC [1].

2.2 K-Nearest Neighbour

Subhan et al. (2015) used K-Nearest Neighbour algorithm to classify leukemic cells from normal blood cells and found that k-NN is a ubiquitous classification tool with good scalability [3]. The k-NN algorithm classifies new objects based on similarity measures under the assumption that similar things exist in close proximity. Without taking the burden of building a model tuning several parameters, the KNN algorithm is a lazy learner with no training phase and performing classification or regression by just memorizing its training data set. Supardi et al. (2012) also used k-NN to classify blasts in leukemic cells and classified the cells into Acute Myelogenous Leukemia (AML) and Acute Lymphocytic Leukemia (ALL) with an accuracy of 80%. 12 main features were extracted from leukemic blood images to represent size, colour and shape and k-values and distance metric were tested many times. Using a value of k=4 and cosine distance metric, the k-NN classifier gave good results [4].

2.3 Neural Networks

Neural networks is used by I. Vincent et al. (2014) to classify blood smear images into normal blood cells and leukemic blood cells, which aids as a clinical decision system to diagnose leukemia. The preprocessing steps include image preprocessing, clustering and segmentation. After that Principal Component Analysis (PCA) is performed to extract the principal components which become the input to the Neural Network classifier [5]. The purpose of this classifier is to classify between normal (non-cancerous) and abnormal cells (cancerous). The first two nodes of the input layer are fed with the first two principal components output by PCA. The hidden nodes receive the input load, along with some weight and hidden node bias and get their input as per Eq.1.

$$ni_j = \sum_{k=1}^{2} \sum_{j=1}^{3} xi_k . wi_{kj} + Bi_j \qquad (1)$$

The final output node value will be calculated using Eq. 2.

$$output = \sum_{j=1}^{3} \left(\sum_{k=1}^{2} xi_k * wi_{kj} + Bi_j \right) * Wo_j + B_j \quad (2)$$

The simple structure of this neural network is given in Figure 3.In the work done by I. Vincent et al. (2014), a two-step neural network is used where the second network takes as input, the cancerous cells and classifies them as ALL and AML using the same architecture. The work done by M. Adjouadi et al. (2009) also uses an Artificial Neural network (ANN) to classify blood images as normal, AML or ALL and demonstrates that the classification accuracy can be increased with increased data size [6].

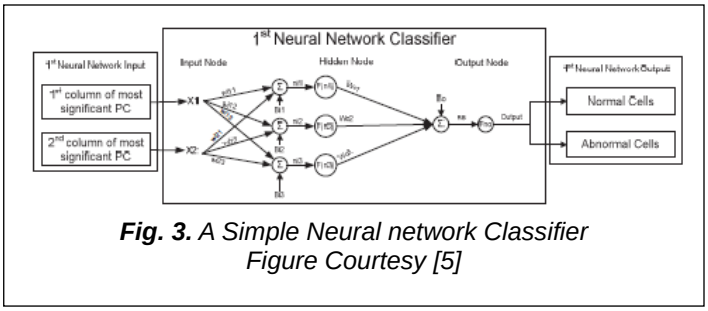

Fig. 3. *A Simple Neural network Classifier*
Figure Courtesy [5]

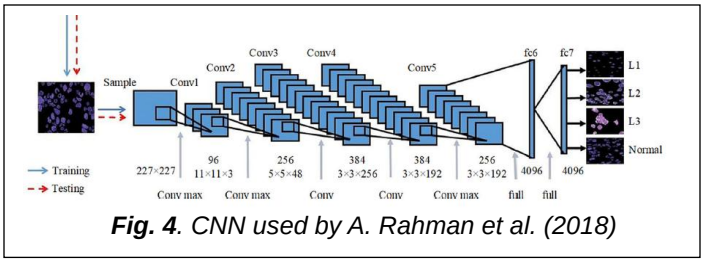

Fig. 4*. CNN used by A. Rahman et al. (2018)*

2.4 Naïve Bayes

Naïve Bayes classifier was used by A. Gautam et al. (2016) to classify leukocytes [7]. Bayes Theorem is a simple probabilistic classifier with independent Naïve assumptions which means that the value of the features is independent of the existence or non-existence of any other features and each of these features contribute independently to the probability. Since it requires only small set of data for training, it can estimate various parameters which are necessary for classification. Being a supervised learning classifier its parameter estimation is based on maximum likelihood scheme. The mean and the variance from every feature of each class c_j is calculated first. They are assumed as independent variables and saved, after that the probability of prior p(c) has been counted as , which is given as

$$P(C_j) = \frac{T_{C_j}}{T_i} \qquad (3)$$

Where T_{C_j} is the trained images of class c, T_i is the total number of trained images, where j=1,2,3,4,5, the leukocyte class.

The posterior probability can be found using Naïve Bayes Classifier using Bayes rule as:

$$P(C_j \mid x_1, x_2, ..., x_n) = P(x_1, x_2, ..., x_n \mid C_j) * P(C_j) \quad (4)$$

Where $P(c_i|x_1, x_2, \ldots, x_n)$ is the posterior probability that x belongs to c_j. Naïve Bayes assumes conditional

independence of attributes the probability of likelihood can be estimated as

$$P(x \mid c_j) = P(x_1 C_j) * P(x_2 C_j) * \ldots * P(x_n C_n) \quad (5)$$

And now the posterior probability can be estimated as:

$$P(C_j \mid x_1, x_2, \ldots, x_n) = P(C_j) * P(x_1 \mid C_j) * P(x_2 \mid C_j) * \ldots * P(x_n \mid C_j) \quad (6)$$

Each white blood cell will be belonging to the class having maximum posterior probability with an accuracy of 80.88% [7].

2.5 Deep Learning

Deep Learning was used for classification of blood images into the subtypes of ALL, that is, L1, L2 and L3 or normal types by A. Rahman et al. (2019). CNNs or Convolutional Neural Network is one of the extensively used deep learning algorithms in biological images data processing. It excludes manual feature extraction and features are learnt directly from the image and thereafter convolves it with input data for the classification to generate an accurate result. The classification process to an extent depends on the discriminant features because, too many features confuse the classifier and too few features are not adequate for effective classification. The blood cell images of both patients having ALL history and normal people were taken as the dataset which was divided as training and test data. It used the Alexnet model with CNN for the classification of ALL into its subtypes or as normal. Configured according to the data, the last three layers of the model are fully connected, softmax and classification layer [8]. As the first step the CNN architecture is defined, and thereafter the input layer and the convolutional layer are defined. The input layer defines the image size to the CNN which corresponds to width, height and number of channels of the image (single channel for grayscale image and 3 for RGB image). The convolutional layer or the second layer consists of the neurons that connect sub region of the image or layers output before it. The convolutional layer learns the features localized by these regions after scanning the image. There is a normalization layer between the Convolutional layer and ReLU layer to speed up the training process and reduce sensitivity [8]. Max pooling layer follows the convolutional layer and is used for downsampling and to reduce overfitting. All features are combined in the last fully connected layer, and is followed by the output layer which consists of softmax function [8]. A. Rahman et al. (2019) used a dataset of 100 images of L1, 100 images of L2, 30 images of L3 and 100 normal images. Taking 80% as training data and 20% as test data, the CNN revealed an accuracy of 97.78% accuracy by taking 1.00e-04 as learning rate and with 20 epochs. The CNN used is shown in the figure 4.

3 DISCUSSION

Each algorithm has its own merits and de-merits. As seen in Table 2, Naïve Bayes and SVM are supervised algorithms whereas ANN and CNN are unsupervised and k-NN is an unsupervised algorithm. SVM is the only binary classifier whereas all other algorithms are capable of classifying into more than two classes. SVM and k-NN are good for small datasets, whereas Neural networks and Naïve Bayes works with both big and small datasets and Deep Learning works well when the dataset is large. Naïve Bayes classifies linear data, whereas k-NN, Neural Networks and Deep learning work with non-linear data and SVM works well for both linear and nonlinear data. Looking at the accuracy values given in Table 2 it cannot be concluded that CNNs are the best, since the comparison is not based on the same dataset. But, the comparison is made between the various algorithms on certain criteria and always remember the Occam's Razor principle: 'use the least complicated algorithm that can address the needs and only go for something more complicated if strictly necessary' [9]. Table 3 provides merits and demerits of the different algorithms discussed.

Table 2. *Comparison of Classification Algorithms*

Algorithm /Features	Supervised/ Unsupervised	No: of classes Supported	Datasets Supported	Works on Linear or Nonlinear Data?	Accuracy as per the samples
SVM	Supervised	2	Small	Both linear and nonlinear	92%
k-NN	Unsupervised	>2	Small	Nonlinear	80%
Neural Networks	Both	>2	Small and Big	Nonlinear	93.7%
Naïve Bayes	Supervised	>2	Small and Big	Linear	80.88%
Deep Learning	Both	>2	Big	Nonlinear	97.78%

Table 3. *Merits and Demerits*

Algorithm	Merits	De-Merits
SVM	• High accuracy • Linearly separable feature space not necessary • Works well with unstructured and semi-structured data • Scales well to high dimensional data	• Takes lot of memory • Does binary classification only • Does not scale to large datasets • Long training time
k-NN	• No training period required • Easy to implement • New data can be added seamlessly	• Cannot handle big datasets • Cannot handle high dimensions • Computation cost is high • Needs feature scaling
Neural Networks	• Works efficiently for both large and small datasets • Needs only less statistical training • Able to detect complex nonlinear relationships between dependent and independent variables • Fault-tolerant	• Great computational burden • Prone to overfitting • Unexplained behaviour of network causes problems • Duration of network unknown • Works with numerical data
Naïve Bayes	• Simple Classifier • Quick convergence • Needs only less training data • Works efficiently for both large and small datasets • Each feature independent of others • Highly scalable	• Assumes all attributes are linearly independent; but in real life it's not so. • Chance of loss of accuracy • Cannot modify dependencies • Assumes numeric attributes are normally

		distributed
Deep Learning	• Reduces the need for feature engineering; features are automatically deduced • Can be applied to many different applications and data types	• Requires expensive GPUs • Extremely expensive to train due to complex data models • What is learned is not easy to understand

4 CONCLUSION

Machine Learning algorithms are gaining popularity and this paper is an attempt to compare five Machine Learning algorithms used for classification: Support Vector Machines, k-Nearest Neighbour, Neural Networks, Naïve Bayes and Deep Learning. Literature on these algorithms to classify and predict leukemia was taken for the study. Further work of this research will lead to identification of the effect of the treatment given to leukemic patients, by the effective use of appropriate Machine Learning algorithms.

5 REFERENCES

[1] J. Laosai and K. Chamnongthai. "Acute leukemia classification by using SVM and K-Means clustering" Proceedings of the International Electrical Engineering Congress pp. 1-4. 2014

[2] “Types of leukemia”15 Nov. 2019 <https://www.cancercenter.com/cancer-types/leukemia/types>

[3] Subhan, Ms. Parminder Kaur. “Significant Analysis of Leukemic Cells Extraction and Detection Using KNN and Hough Transform Algorithm” International Journal of Computer Science Trends and Technology, vol. 3, no.1, pp. 27-33, 2015

[4] Supardi, N. Z., Mashor, M. Y., Harun, N. H., Bakri, F. A., & Hassan, R. “Classification of blasts in acute leukemia blood samples using k-nearest neighbour”. IEEE 8th International Colloquium on Signal Processing and Its Applications, pp. 461-65, 2012

[5] Vincent, I., Kwon, K.-R., Lee, S.-H., & Moon, K.-S. (2015). “Acute lymphoid leukemia classification using two-step neural network classifier” 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV) Jan. 2015

[6] Adjouadi, M., Ayala, M., Cabrerizo, M. et al. “Classification of Leukemia Blood Samples Using Neural Networks” Annals of Biomedical Engineering, vol. 38, no. 4, pp. 1473-82, Apr.2010

[7] Gautam, A., Singh, P., Raman, B., & Bhadauria, H. “Automatic classification of leukocytes using morphological features and Naïve Bayes classifier” IEEE Region 10 Conference (TENCON), pp.1023-27, Nov. 2016

[8] Rehman, A., Abbas, N., Saba, T., Rahman, S. I. ur, Mehmood, Z., & Kolivand, H. “Classification of acute lymphoblastic leukemia using deep learning” Microscopy Research and Technique, vol. 81, no. 11, 23 Oct. 2018

[9] “Occam's razor” 17 Nov. 2019 <https://simple.wikipedia.org/wiki/Occam%27s_razor>