

An Automated Supervised Machine Learning Models For (ALL) Cancer Classification

Deepali Soni¹, Prof. Priyanshu Dhameniya²
^{1,2} AITR Indore

Abstract- In this paper, we have present a machine learning based approach for the classification of blood leukaemia using microscopic images employing a Probabilistic Neural Network. Probabilistic Neural Network is based on Bayes's theorem of Conditional Probability and is a famed paradigm for data classification for systems employing artificial intelligence. Pre-processing has been achieved using gray scaling and thresholding. Discrete Wavelet Transform (DWT) has been used as a tool to remove the abrupt variations in the calculated feature values. Principal component analysis (PCA) has been used to find particular trends in the computed feature data and minimize the redundancy. We have shown that the proposed technique achieves 98% percent classification accuracy which can be attributed to the highly rigorous pre-processing and feature extraction mechanisms which culminates to training a Probabilistic Neural Network which is used for the final classification of the data.

Keywords- Artificial Intelligence (AI), Probabilistic Neural Network (PNN), Discrete Wavelet Transform (DWT), Principal Component Analysis (PCA), Accuracy, Acute lymphocytic leukemia (ALL).

I. INTRODUCTION

Artificial Neural Network (ANN) has recently emerged as one of the most powerful tools for contemporary computation. Its design is based on the fact that the human brain is a:

- 1) Highly Non Linear Structure
- 2) Highly Parallel Structure

On extremely important attribute of the neural model is its ability in following trends in the input fed to it. No matter how complex or abruptly the output for a corresponding input may change, the network maps the input and output in the form of experiences called weights. The parallel structure enables data or inputs X from various paths design the weights W. The design of the network culminates in the decision making according to some function θ called the bias. The structure can be mathematically modelled as:

$$Y = \sum_{i=1}^n X_i W_i + \theta$$

Here X represents the signal
 W represents the weight
 θ represents the bias.

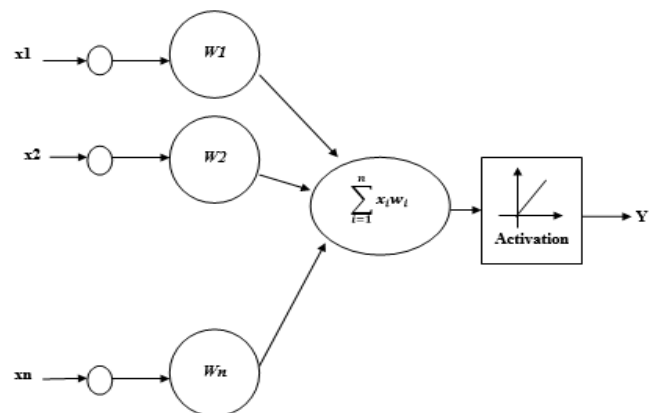


Fig.1 Mathematical Formulation of Neural Network

The above figure is pretty generic in nature. The task of following a trend in the given input data can be accomplished accurately using a particular architecture of a neural network. [3] The network continuously leans and adapts according to the provided input data and its corresponding data. Thus it can be thought of working as the human brain that also adapts itself according to experiences. The experiences themselves have counterparts in the mathematical ANN structure called weights. If the correspondance between the input X and a target Y is given to the network to design and adapt its weights, then the deviation of the expected output and the actual output is given by an error E which can be mathematically defined as:

$$E = Y_a - Y_p$$

Y_a is the actual output
 Y_p is the predicted output
 E is the error.

II. PROBABILISTIC NEURAL NETWORK (PNN)

PNN is built on the theory of Bayesian network and the estimation of probability density function. This theory allows for cost function to represent the fact that it may be worse to misclassify a vector that is actually a member of class A but it's classified as a vector that belongs to class B. [1] PNN is extensively used for classification problem. It works as a supervised classifier. We give a sample of our data set to it as an input and the data is then handled through numerous layers. The Bayes rule such that the input vector belonging to class A is classified as

$$P_A C_A f_A(x) > P_B C_B f_B(x)$$

Where P_A –Priori probability of occurrence of pattern in class A. C_A - Cost associated with classifying vectors.

$f_A(x)$ - Probability density functions of class A.

PNN can take any number of input and can map them to any number of output so for that reason PNN works more efficiently than other Neural Networks. Neural Networks in common have a back propagation which this neural network doesn't have making it faster than others.

The different layers of Probabilistic Neural Network are:

- Input Layer: It consists of the input to the neural network the training sample whose class has to be found out.
- Pattern Layer: It consists of the Gaussian function with the training samples are the center to it. We then find the Gaussian distance for all of them using the formula.
- Summation Layer: Here all the output of the pattern layer is received and then they are added and send to the output layer for class determination. It is also the hidden layer in the Neural Network.
- Output Layer: Here all the outputs from the summation layers are taken and a selection of the largest value is done and that particular class of the selection determines the class of the sample.

The PNN works by creating a set of multivariate probability densities that are derived from the training vectors presented to the network. The input instance with unknown category is propagated to the pattern layer. Once each node in the pattern layer receives the input, the output of the node will be computed. The summation layer neurons compute the maximum likelihood of pattern x being classified into c by

summarizing and averaging the output of all neurons that belong to the same class

$$P_i^c = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left[\frac{-(x - x_{ij})^T (x - x_{ij})}{2\sigma^2} \right]$$

where N_i denotes the total number of samples in class c . If the a priori probabilities for each class are the same, and the losses associated with making an incorrect decision for each class are the same, the decision layer unit classifies the pattern x in accordance with the Bayes's decision rule based on the output of all the summation layer neurons

$$C(x) = \text{argmax} \{P_i(X)\}, i = 1, 2, \dots, c$$

where $C(x)$ denotes the estimated class of the pattern x and m is the total number of classes in the training samples. If the a priori probabilities for each class are not the same, and the losses associated with making an incorrect decision for each class are different, the output of all the summation layer neurons will be

$$C(x) = \text{argmax} \{P_i(X) \text{cost}_i(x) \text{apro}_i(x)\}, i = 1, 2, \dots, c$$

Where $\text{cost}_i(x)$ is, the cost associated with misclassifying the input vector and $\text{apro}_i(x)$. Thus PNN acts as an effective classifier to data classification into categories.

There are quite a lot of advantages of using PNN instead of Back Propagation (BP) or multilayer perceptron. PNNs are much faster than BP multilayer networks. It provides better accuracy.

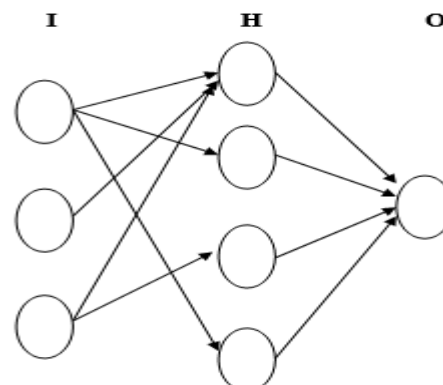


Fig.2 Model of PNN with Input Feature Values

This network is relatively insensitive to outliers and generates accurate predicted target probability scores.

Some of the advantages of PNN are:

- It can map any number of input to any number of classification.
- Unlike other Neural Network it doesn't have a back propagation which makes it a fast learner.
- Once trained removing or adding a data sample doesn't require retraining of the sample.
- It saves a lot of computational space.

III. DATA PRE-PROCESSING

Prior to feature extraction and classification, data pre-processing is done to achieve train the PNN accurately and obtain high sensitivity and accuracy of classification.[5]-[6]

(i) Segmentation:

The division of an image into meaningful structures, image segmentation, is often an essential step in image analysis, object representation, visualization, and many other image processing tasks. A disjunct categorization does not seem to be possible though, because even two very different segmentation approaches may share properties that defy singular categorization. [5] The categorization presented in this chapter is therefore rather a categorization regarding the emphasis of an approach than a strict division. The following categories are used:

Threshold based segmentation: Histogram thresholding and slicing techniques are used to segment the image. They may be applied directly to an image, but can also be combined with pre- and post-processing techniques.

Edge based segmentation: With this technique, detected edges in an image are assumed to represent object boundaries, and used to identify these objects.

Region based segmentation: Where an edge based technique may attempt to find the object boundaries and then locate the object itself by filling them in, a region based technique takes the opposite approach, by (e.g.) starting in the middle of an object and then "growing" outward until it meets the object boundaries

Segmentation plays a crucial role in the feature extraction and classification.

(ii) Principal Component Analysis (PCA)

Principal Component Analysis is a method by which we can find the attributes which contribute the most to the final classification of the data set. Principal component analysis (PCA) is a mathematical procedure that converts a set

of correlated variables into a set of values of uncorrelated variables called principal components.

PCA is a mathematical process that does a linear transformation of the dataset to a new system such that the element which influences to the class greatest is set as the first coordinate, that is the first principal component followed by the next great and so on. But when the data is non-linear in nature, PCA is unable to find the principal components.

(iii) Discrete Wavelet Transform (DWT)

The Wavelet Transform is rather a recent tool for the analysis of randomly fluctuating non-smooth signals.

The mathematical description of the wavelet transform can be given by:

$$C(S, P) = \int_{-\infty}^{\infty} f(t) ((S, P, t)) dt$$

Here S stands for scaling

P stands for position

t stands for time shifts.

C is the Continuous Wavelet Transform (CWT)

The main disadvantage of the CWT is the fact that it contains an enormous amount of data. The sampled version of the CWT is the Discrete Wavelet Transform (DWT). The DWT is a down sampled version of the CWT and its characteristic nature is to smoothen out abrupt fluctuations which are possible due to both abruptly changing base functions and down sampling.

The scaling function can be defined as:

$$W\Phi(J_0, k) = \frac{1}{\sqrt{M}} \sum_n S(n) \cdot \Phi(n)_{j_0 k}$$

The Wavelet function can be defined as:

$$W\Psi(j, k) = \frac{1}{\sqrt{M}} \sum_n S(n) \cdot \psi(n)_{j, k}$$

Where $\frac{1}{\sqrt{M}}$ is Normalizing term

IV. PROPOSED SYSTEM

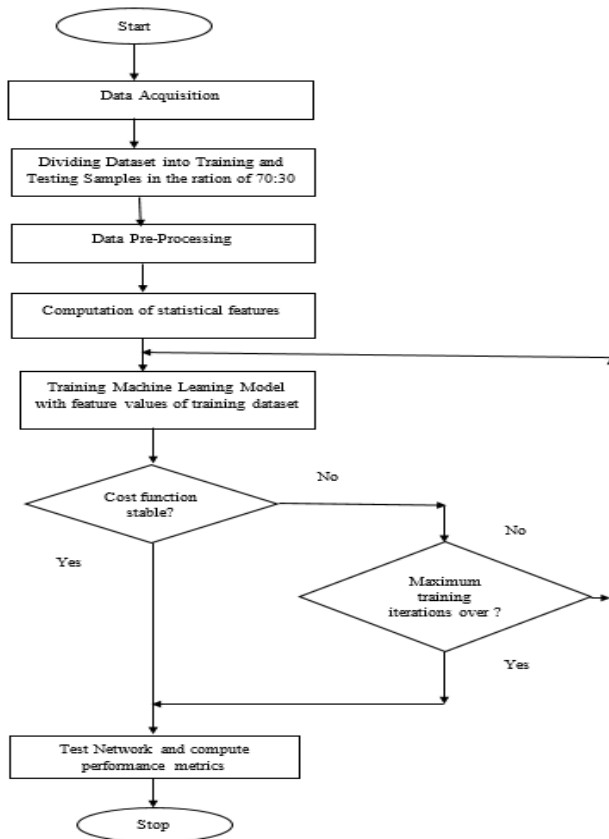


Fig.3 Proposed Flow Chart

The proposed technique can be understood using the following steps of the algorithm:

- 1) One image is loaded at a time to the MATLAB workspace.
- 2) RGB to Gray scale Conversion and Binary Thresholding to convert the image into a Binary image have been employed as a pre-processing tool.
- 3) Segmentation has been applied to locate affected cellular area..
- 4) Subsequently Discrete Wavelet Transform (DWT) has been applied smoothen out the non-linearities in the image to facilitate feature extraction.
- 5) Principal Component Analysis (PCA) has been employed to find regular patterns or trends in the DWT coefficients computed above.
- 6) Twelve feature values have been calculated for the image under consideration.
- 7) The above process is employed for all the training data samples in the data set.
- 8) The feature values obtained for all the training data samples are used to train a Probabilistic Neural Network (PNN).

- 9) The Neural Network is tested for the remaining images of the data set.
- 10) Sensitivity and Accuracy have been evaluated for the Proposed System Design.

V. RESULTS

Evaluation Parameters:

The various parameters for the classification are:

1. **True Positive (TP):** It is the case when a sample belongs to category and the test also predicts its belongingness.
2. **True Negative (TN):** It is the case when a sample does not belong to category and the test also predicts its non-belongingness.
3. **False Positive (FP):** It is the case when a sample does not belong to category and the test predicts its belongingness.
4. **False Negative (FN):** It is the case when a sample belongs to category and the test predicts its non-belongingness.

Accuracy (Ac): It is mathematically defined as:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

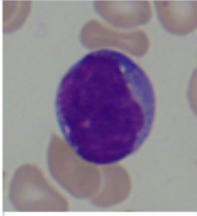
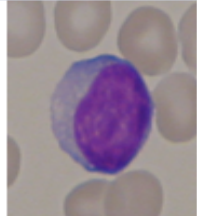
	HEALTHY	LEUKEMIA
		
	0.566477273	0.403409091
	0.134990779	0.079914238
	0.805264075	0.689903151
	0.938386769	0.910672348
	0.004660435	0.009083011
	0.106527261	0.106241438
	2.208658726	3.764806736
	0.106600358	0.106600358
	0.011328815	0.011261729
	0.895968612	0.943774119
	20.05759459	7.15389322
	2.159108057	0.788548077

Fig.4 :Tabulation of featue values for Normal and Leukaemia Blood samples

A total of 200 images have been used for training and 60 images have been used for testing. It has been found that the accuracy is **98%**. It is significantly higher than the accuracy, **91.84%** obtained by previously existing system for the same dataset [1].

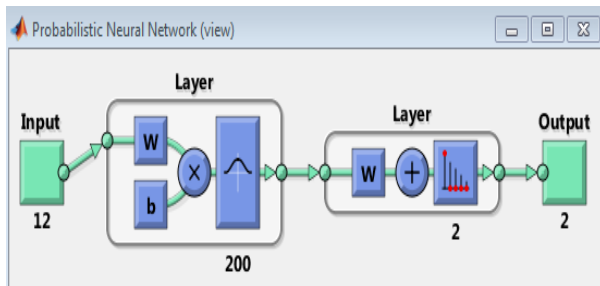


Fig.5 :Designed Neural Network

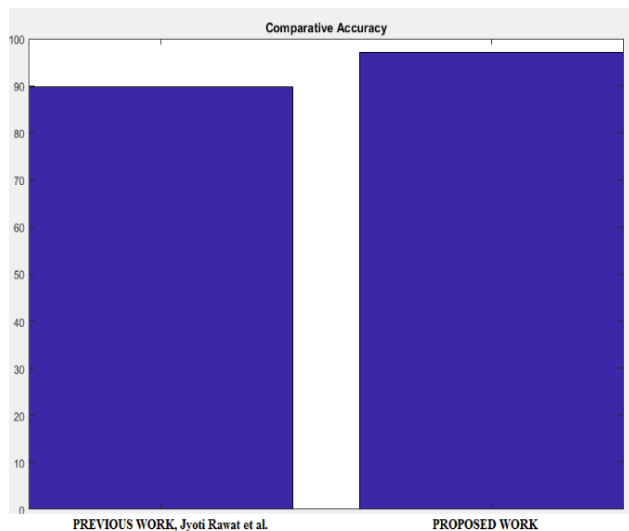


Fig.6 : Compared Accuracy Analysis

VI. CONCLUSION

In this paper blood leukaemia image are automatically classified into normal or malignant (with leukaemia). The classifier used is the Probabilistic Neural Network (PNN) classifier. Accuracy of 98% is obtained. This work will act as supportive tool for medical practitioners and will help doctors for fast diagnosis based on which the treatment plan can be decided. We have found that the proposed technique achieves 98% classification accuracy which can be attributed to the efficacy of the proposed method.

REFERENCES

[1] E Tuba, I Strumberger, N Bacanin, D Zivkovic, “Acute Lymphoblastic Leukemia Cell Detection in Microscopic

Digital Images Based on Shape and Texture Features”, Springer 2020

- [2] Sachin Kumar ,Sumita Mishra, Pallavi Asthana, Pragya, “Automated Detection of Acute Leukemia Using K-mean Clustering Algorithm, Springer 2018
- [3] JyotiRawat et.al , “Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia”, Elsevier 2017
- [4] Sonali Mishra , Lokesh Sharma et.al, “Microscopic Image Classification Using DCT for the Detection of Acute Lymphoblastic Leukemia (ALL)”, Springer 2017
- [5] Yunxiang Mao , Zhaozheng Yin , Joseph Schober , “A deep convolutional neural network trained on representative samples for circulating tumor cell detection”, IEEE 2016
- [6] Vasuki Shankar , Murali Mohan Deshpande et.al, “ Large Scale Image Feature Extraction from Medical Image Analysis”, IEEE 2016
- [7] Jyoti Rawat , H.S. Bhadauria, “(Computer Aided Diagnostic System for Detection of Leukemia using Microscopic Images”, Elsevier 2015
- [8] D. Goutam , S. Sailaja , “Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier, IEEE 2015
- [9] Sos Agaian , Monica Madhukar et.al., “Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images”, IEEE 2014
- [10] Nur Alom Talukda et.al, “Automated Blood Cancer Detection Using Image Processing Based on Fuzzy System”, IJARCSSE 2014
- [11] Aditya Sharma et.al , “Quantum Dots Self Assembly Based Interface for Blood Cancer DetectionBased on Fuzzy System”, Langmuir 2013
- [12] MarcVendrell et.al , “Surface-enhanced Raman scattering in cancer detection and imaging, “ Elsevier 2013
- [13] Devin C. Koestler et.al , “Peripheral Blood Immune Cell Methylation Profiles Are Associated with Nonhematopoietic Cancers”, AACR 2012
- [14] Hayan T. Madhloom et al., “An Image Processing Application for the Localization and Segmentation of Lymphoblast Cell Using Peripheral Blood Images”, Springer 2012
- [15] Subrajeet Mohapatra et al., “Fuzzy Based Blood Image Segmentation for Automated Leukemia Detection”, IEEE 2011
- [16] Yujie LI et al., “An Improved Detection Algorithm Based on Morphology Methods for Blood Cancer Cells Detection”, JOFCIS 2011
- [17] Siyang Zheng et al., “3D microfilter device for viable circulating tumor cell (CTC) enrichment from blood”, Springer 2011

- [18] Subrajeet Mohapatra , Dipti Patra et al., “Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images”, IEEE 2010
- [19] Subrajeet Mohapatra et al., “Image analysis of blood microscopic images for acute leukemia detection”, IEEE 2010
- [20] Waidah Ismail et al., “Detecting Leukaemia (AML) Blood Cells Using Cellular Automata and Heuristic Search”, Springer 2010
- [21] JM Corchado, JF De Paz, S Rodríguez, “Model of experts for decision support in the diagnosis of leukemia patients”, Elsevier 2009
- [22] CE Pedreira, L Macrini, MG Land, “New decision support tool for treatment intensity choice in childhood acute lymphoblastic leukemia”, IEEE 2009