Classification of Leukemia Image Using Genetic Based K-Nearest Neighbor (G-KNN)

M. Bennet Rajesh¹ and S. Sathiamoorthy²*

¹ & ² Assistant Professor, Division of Computer and Information Science
Annamalai University, Annamalai Nagar, Tamil Nadu, India
E-Mail: ks_sathia@yahoo.com
*Corresponding Author

Abstract - In the medical diagnostic system, categorization of blood cell is more vital to analyze and detect the disease. The diseases that are associated with blood can be divided only after the categorization of blood cell. Leukemia is a blood cancer that starts with bone marrow. Therefore, it must be treated at the initial stage and makes to death if left unprocessed. This paper introduces a new Genetic based KNN pre-processing approach for removing the noise in Leukemia image without affecting the accuracy of an image. This paper integrates the Genetic algorithm and KNN for noise removal and pre-processing of Leukemia image datasets.

Keywords: Leukemia, K-Nearest Neighbor, Genetic Algorithm (GA), Pre-Processing, Noise Removal, Median Filter approach

1. INTRODUCTION

Leukemia is a kind of cancer [1] that marks the bone marrow causing the better fabrication of white blood cells which inflow the bloodstream which is more essential, infection-fighting part of your immune system, made in bone marrow. Hence, diagnose myelogenous leukemia has perfectly analyzed a physician will edict a comprehensive blood count bone mellow biopsy and cytophenetic tests.

Leukemia is a malignancy (cancer) of blood cells. In leukemia, atypical blood cells have formed in the bone marrow. White Blood Cells (WBC) are the cells accountable for fighting infections. Leukaemia comprises the fabrication of the abnormal WBC.

II. RELATED WORKS

The following table 1 represents the literature survey on the Image Processing techniques used for the detection of the Leukemia.

Niranjan Chatap, Sini Shibu in [2], analyzed that Classical watershed-based algorithm overcomes the problem of Cell extraction and segmentation from massive, noisy images.

A. Arputha Regina in [3], suggested that K-Means Clustering, Hausdorff Dimension, Local Binary Pattern for detecting the leukemia using segmentation, Feature Extraction, and Classification.

Tejashri G. Patil, V. B. Raskar in [4] used Contour Signature, Otsu’s Segmentation for detecting the acute leukemia using feature extraction.

C.Vidhya, P. Saravana Kumar, K. Keerthika, C.Nagalakshmi, B. Medona devi in [5] used K-Means Clustering, SVM Classifiers for input image is resized and CIELAB color conversion. Extraction of nuclei from the leukocytes.

Mrs. Trupti A. Kulkarni-Joshi, Prof. Dilip S. Bhosale in [6] suggested Otsu’s Thresholding Algorithm used for segmentation of Acute Lymphoblastic Leukemia.


Sonal G. Deore Prof. Neeta Nemade in [8] Adaptive Pre-Filtering and Segmentation Algorithms for extracting morphological indexes from those cells, and finally, it classifies the presence of leukemia.


Ms. Minal D. Joshi, Prof. Atul H. Karode, Prof. S.R. Suralkar in [12] suggested Histogram, Global Thresholding Otsu Method, KNN for feature extracted, classifier classifies the affected cells or healthy cells and with the accuracy of 93%, kNN classification is used.

Khot S.T, Sneha Bhalekar, Divya Jaggi and Dolly Rani in [13] used Support Vector Machine for the features extracted from the images and applied to classifier to detect the leukemia cell.

Monika Mogra, Vivek Srivastava in [15] used Watershed Transform and Morphological Image Processing Technique to identify the affected white blood cells.

Ms. Sneha Dhakne, Ms. Kumudini K. Borkute, Ms. Priyanka Ikhar in [16] utilized Watershed Algorithm and Clustering Algorithm for the determination of the cells.

III. PROPOSED GENETIC BASED KNN PRE-PROCESSING APPROACH FOR LEUKEMIA IMAGE DATASET

A. Filter based Noise Removal

The best-known order-statistic filter in digital image processing is the median filter. It is a useful tool for reducing salt-and-pepper noise in an image. The median filter plays a key role in image processing and vision. In median filter, the pixel value of a point p is replaced by the median of pixel value of 8-neighborhood of a point, p. The operation of this filter can be expressed as:

$$g(p) = \text{median}\{f(p), \text{where } p \in N_8(p)\}$$

The median filter is popular because of its demonstrated ability to reduce random impulsive noise without blurring edges as much as a comparable linear low pass filter. However, it often fails to perform well as linear filters in providing sufficient smoothing of no impulsive noise components such as additive Gaussian noise. One of the main disadvantages of the basic median filter is that it is location-invariant in nature, and thus also tends to alter the pixels not disturbed by noise.

B. K-Nearest Neighbor Approach

In an image processing, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning.

In K-NN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance. The limitation of the K-NN algorithm is it’s sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction. In Feature space, extraction is taken place on raw data before applying K-NN algorithm. The Figure 1 narrates the steps involved in a K-NN algorithm.

C. Genetic Algorithm

In the computer science field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. GAs is inspired by Darwin’s Theory about Evolution “Survival of Fittest”. GAs is adaptive heuristic search based on the evolutionary ideas of natural selection and genetics. GAs simulate the survival of the fittest among individuals over consecutive generation for solving a problem. Each generation consists of a population of character strings that are analogous to the chromosome that we see in our DNA. Each individual represents a point in a search space and a possible solution. The individuals in the population are then made to go through a process of evolution.
Individuals in the population die and are replaced by the new solutions, eventually creating a new generation once all mating opportunities in the old population have been exhausted. In this way it is hoped that over successive generations better solutions will thrive while the least fit solutions die out.

New generations of solutions are produced containing, on average, better genes than a typical solution in a previous generation. Eventually, once the population has converged and is not producing offspring noticeably different from those in previous generations, the algorithm itself is said to have converged to a set of solutions to the problem at hand.

**D. Proposed Genetic based KNN Pre-Processing Framework for Leukemia Image Dataset**

This algorithm is proposed to select the best k-value with the minimum misclassification rate. The K-NN algorithm is one of the best and commonly used for clustering and classification. In this method, each sample value fit to the test dataset and it is classified according to the nearest k sample based on the training data. The class numbers values obtained from k sample values, the maximum number is determined from class samples. The distance measurement calculated by using Euclidean distance measure. The distance of nearest neighbors is calculated based on the test image using distance weighted formula as:

\[
\text{distance (x)} = \sqrt{\sum_{i=1}^{n} \text{weight}_i(x_i - y_i)^2}
\]

where x and y are two images, n is the number of features. Subsequently, the unidentified sample value is selected most relevant to the class from k nearest neighbor algorithm and it is used to find the real value from unidentified sample values. In this proposed method, k value is obtained by Genetic Algorithm.

Pseudo code for Proposed Genetic based KNN pre-processing approach

**Step 1:** Choose k number of samples from the training set to generate the initial population (p1).
**Step 2:** Calculate the distance between the training samples in each chromosome and the testing samples, as fitness value.
**Step 3:** Choose the chromosome with highest fitness value and store it as global maximum (Gmax).

- **Step 3.1:** For i = 1 to L do
- **Step 3.2:** Perform Reproduction
- **Step 3.3:** Apply the Crossover operator
- **Step 3.4:** Perform mutation and get the new population. (p2)

- **Step 3.5:** Calculate the local maximum (Lmax).
- **Step 3.6:** If Gmax < Lmax then
  - **Step 3.6.1:** Gmax = Lmax;
  - **Step 3.6.2:** p1 = p2;
  - **Step 3.6.3:** Repeat

**Step 4:** Output – The chromosome which obtains Gmax and has the optimum k-neighbors.

![Diagram of Proposed Genetic based KNN Pre-processing framework](image)

**IV. RESULT AND DISCUSSION**

The performance of Genetic based KNN pre-processing method is calculated based on the classification accuracy, sensitivity, specificity, finding minimum distance and processing time. The classification accuracy, specificity and sensitivity calculated by using accuracy formula:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Specificity} = \frac{TN + FP}{TN} \\
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

Here:

- True Positive (TP) = accurately classified positive values
- True Negative (TN) = accurately classified negative cases
- False Positive (FP) = inaccurately classified negative values
- False Negative (FN) = inaccurately classified negative values

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Classification Accuracy of Finding Leukemia</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLPNN</td>
<td>90%</td>
</tr>
<tr>
<td>SVM</td>
<td>87%</td>
</tr>
<tr>
<td>Proposed G-KNN pre-processing Framework</td>
<td>96.57%</td>
</tr>
</tbody>
</table>
Fig. 3 Graphical Representation of the performance analysis of the proposed Genetic based K-Nearest Neighbor pre-processing framework with existing classification Multi-Layer Perceptron Neural Network (MLPNN) and Support Vector Machine (SVM).

TABLE II PERFORMANCE ANALYSIS OF THE PROPOSED GENETIC BASED K-NEAREST NEIGHBOR PRE-PROCESSING FRAMEWORK AGAINST CLASSIFICATION ACCURACY, PROCESSING TIME AND MISCLASSIFICATION RATE

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Classification accuracy</th>
<th>Processing Time</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLPNN</td>
<td>90%</td>
<td>3 sec</td>
<td>3%</td>
</tr>
<tr>
<td>SVM</td>
<td>87%</td>
<td>3.5 sec</td>
<td>1.73%</td>
</tr>
<tr>
<td>Proposed G-KNN pre-processing Framework</td>
<td>96.57%</td>
<td>2.9 sec</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

Fig. 4 Graphical Representation of the performance analysis of the proposed Genetic based K-Nearest Neighbor pre-processing framework with existing classification Multi-Layer Perceptron Neural Network (MLPNN) and Support Vector Machine (SVM) against Sensitivity.

TABLE III PERFORMANCE ANALYSIS OF THE PROPOSED GENETIC BASED K-NEAREST NEIGHBOR PRE-PROCESSING FRAMEWORK AGAINST ACCURACY, SENSITIVITY, SPECIFICITY, PRECISION AND ERROR RATE

<table>
<thead>
<tr>
<th>Classification Methods</th>
<th>Performance Analysis in %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>MLPNN</td>
<td>90%</td>
</tr>
<tr>
<td>SVM</td>
<td>87%</td>
</tr>
<tr>
<td>Proposed G-KNN pre-processing Framework</td>
<td>96.57%</td>
</tr>
</tbody>
</table>

Fig. 5 Graphical Representation of the performance analysis of the proposed Genetic based K-Nearest Neighbor pre-processing framework with existing classification Multi-Layer Perceptron Neural Network (MLPNN) and Support Vector Machine (SVM) against Specificity.

From the above figures 3-7, it is clear that the proposed Genetic based K-Nearest Neighbor Pre-processing framework gives better result than the existing methods like Support Vector Machine (SVM) and Multi-Layer Perceptron Neural Network (MLPNN). The proposed G-KNN pre-processing framework composed two major process, Image denoising by median filter approach and...

Fig. 6 Graphical Representation of the performance analysis of the proposed Genetic based K-Nearest Neighbor pre-processing framework with existing classification Multi-Layer Perceptron Neural Network (MLPNN) and Support Vector Machine (SVM) against Precision.

Fig. 7 Graphical Representation of the performance analysis of the proposed Genetic based K-Nearest Neighbor pre-processing framework with existing classification Multi-Layer Perceptron Neural Network (MLPNN) and Support Vector Machine (SVM) against Error rate.

From the above figures 3-7, it is clear that the proposed Genetic based K-Nearest Neighbor Pre-processing framework gives better result than the existing methods like Support Vector Machine (SVM) and Multi-Layer Perceptron Neural Network (MLPNN). The proposed G-KNN pre-processing framework composed two major process, Image denoising by median filter approach and...
enhancing the image by using G-KNN method in the pre-processing stage for the better classification accuracy of Leukemia image dataset.

V. CONCLUSION

In this paper, the proposed G-KNN pre-processing framework composed two major processes, Image denoising by median filter approach and enhancing the image by using G-KNN method in the pre-processing stage of Leukemia image dataset. In this proposed G-KNN pre-processing method, K-NN is combined with Genetic Algorithm to enhance the classification accuracy of the lung cancer identification.

REFERENCES