MACHINE LEARNING ALOGRATHMS BASED APPROACH FOR DETECTING THE LEUKEMIA

A PROJECT REPORT

Submitted by

PRAGADEESHWARAN S (2116210701190)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **"MACHINE LEARNING ALOGRATHMS BASED APPROACH FOR DETECTING THE LEUKEMIA**" is the bonafide work of "**PRAGADEESHWARAN S (2116210701190)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_____

Internal Examiner **External Examiner**

ABSTARCT

Detecting leukemia, a condition characterized by significantly elevated blood cell counts compared to healthy individuals, typically relies on repeated full blood counts due to the challenge posed by malignant cells resembling normal blood cells. Early identification and treatment are crucial to prevent complications. Current laboratory procedures for leukemia detections are time-consuming. This study proposes a machine learning approach for leukemia detection in patients. A dataset comprising images of blood smears is collected and pre-processed to extract relevant features. These features are then utilized to train and assess various machine learning classification methods, including KNN, Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression.
Comparison of the classifiers' performance against multiple criteria, such as accuracy and precision, aids in identifying the most effective classifier.

Keywords— Logistic Regression, Decision Tree, Random Forest, Support Vector Machine.

INTRODUCTION

A group of tumours that affect the bone marrow and blood are together referred to as leukaemia. The marrow is the soft tissue found inside bones that produces red blood cells. It mostly concerns white blood cells, which are essential for fighting infections. This disorder develops when aberrant white blood cells multiply too quickly, preventing the body from producing healthy blood cells normally. There are less healthy red blood cells, platelets, and white blood cells because to the overproduction of these aberrant cells. Although the precise cause of leukaemia cell production is still unknown, several experts speculate that environmental variables and genetic abnormalities may play a role. Leukaemia is categorised into acute and chronic leukaemia depending on the rate of advancement and kind of cells involved, and lymphocytic and myelogenous leukaemia based on the type of white blood cell affected. Leukaemia is divided into four types: acute lymphocytic leukaemia (ALL), acute myelogenous leukaemia (AML), chronic lymphocytic leukaemia (CLL), and chronic myelogenous leukaemia (CML). Leukaemia symptoms vary depending on the type, but the most common include fever, weight loss, swollen lymph nodes, prolonged fatigue, weight loss, and so on [1]. Laboratory tests such as peripheral blood smears, differentials, and complete blood counts (CBC) can be used to diagnose these illnesses. Once the kind of leukaemia has been determined based on these results, more testing will be conducted. However, conventional laboratory techniques for illness diagnosis are time-consuming and require a competent professional to evaluate blood samples and provide accurate results. To address this, a software-based solution that combines machine learning and deep learning techniques is being used. Compared to laboratory approaches, these methods take a very short amount of time [2]. Image processing techniques are utilised to pre-process microscopic images and count blood cells, which aids in disease detection[3].

The photos are segmented to focus on the relevant part, from which the appropriate features can be extracted to reduce complexity. The histogram equalisation and Zack thresholding approaches are used to treat the image before extracting the features [4]. Lymphocytic leukaemia is the most frequent kind of leukaemia in Bangladesh. In a cancer patient, abnormal blood components such as neutrophils, eosinophils, basophils, lymphocytes, and monocytes are recognised and used to diagnose the disease early. Four significant variables that have a greater impact on determining a leukaemia patient were found. To forecast cancer-forming cells, a faster RCNN ML algorithm is applied [5]. When compared to other types of cancers, blood cancers are expanding exponentially over the world. According to surveys, India ranks third among countries in terms of leukaemia incidence. The peripheral blood smear images, which are also microscopic images, are pre-processed and segmented based on pixels, and a large section of the image is used for feature extraction before being classified using the Convolution Neural Network (CNN)[6].

The dataset used in [7] consists of 220 blood smear images of leukaemia and non-leukemia cells, and algorithms such as k-means clustering, HSV color-based segmentation, and marker-controlled watershed algorithms are used for segmentation, and they are classified using SVM to determine which type of leukaemia they belong to. This study [8] used a pre-trained convolution neural network to diagnose leukaemia. It is a simple method to use deep learning for picture analysis, and the dataset is available public from ALL-IDE. The categorization is carried out with pre-trained series models such as MobileNet-v2, GoogleLeNet, AlexNet, and residual networks. The optimisation techniques Stochastic Gradient with Momentum (SGDM), Root Mean Square propagation (RMSprop), and Adaptive Moment estimation (ADAM) are also compared. In their study [9], the author used FastNMeans Denoising and edge enhancement to pre-process cropped blood cell pictures. The Grab cut algorithm was then used to separate the overlapping cells from the background, and properties such as area, roundness, compactness, and so on were extracted. Finally, the photos are classed with the K-NN method. According to the author's analysis in this study paper [10], the image yields 93% accuracy when using a blended biogeography-based optimisation technique. The crow search method is commonly used to solve problems like feature selection and optimisation. However, in [11], this approach is used to convert non-linearly separable data points into linearly separable data points. Additionally, it performed as a classifier with 87% accuracy. In [12], the author described a hybrid technique for feature extraction from WBC that uses a CMK- moment localization method to extract the region of interest and a CNN-based fusion method to extract features using deep learning. [13] The dataset contains of gene expression to differentiate between AML and ALL. The PCA and SMOTE algorithms are used to reduce dimensionality and oversample outnumbered classes, respectively. This research [14] identifies certain critical features for cancer detection and concludes that the gradient boosting decision tree approach performs better than the support vector machine. Classification is accomplished using a neural network and a variety of machine learning algorithms. And a comparative examination of various algorithms is offered in [15].

CREATING DATASET AND PREPROCESSING IMAGES FOR LEUKAEMIA DETECTION.

I. The suggested approach uses digital image processing techniques for image pre-processing, optimisation, and several classification algorithms for image categorization. The diagnosis of leukaemia using peripheral blood smear (PBS) pictures is critical in distinguishing between early cancer cells and non-cancerous cells. The dataset was obtained from the Kaggle source, a social online community for data scientists. This dataset consisted of 3256 peripheral blood smear pictures collected from 89 patients suspected of having acute lymphocytic leukaemia and stained by bone marrow laboratory personnel. In general, this dataset can be classified into two types: benign and malignant. The benign class comprises of hematogones, but the malignant class is further classified into three types: early Pre-B, Pre-B, and Pro-B ALL. All of these microscopic photos were acquired with a Zeiss camera at 100x magnification. Figure 1 depicts the block diagram of the leukaemia detection procedure.

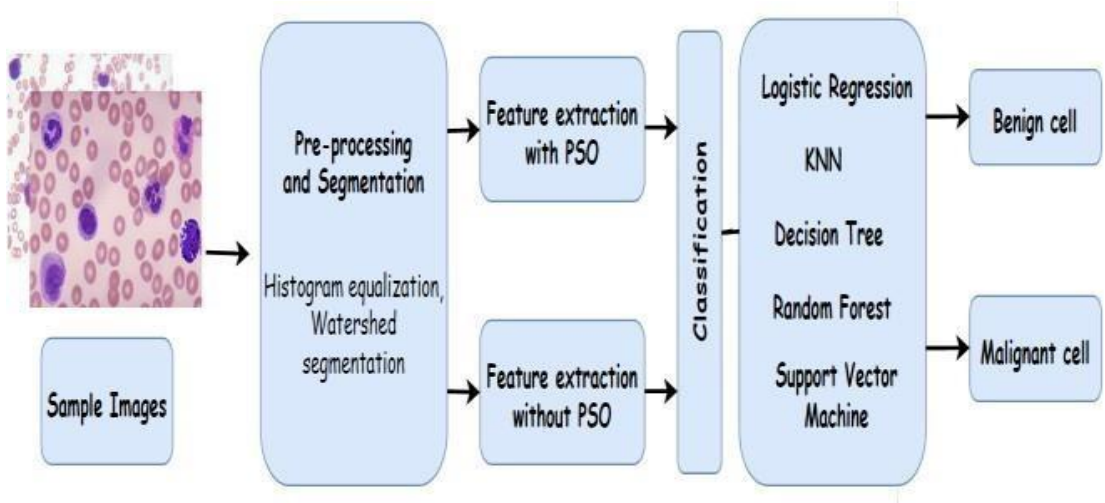


Figure 1: leukaemia detection by using various classifiers.

Pre-processing an image is an important step in improving its quality and modifying its format, which allows for simpler analysis. Because these are done at the lowest level of abstraction, they have no effect on increasing the image's information richness. The two major phases are noise removal and image enhancement. There are several filters available for reducing noise from images. The median filter is a type of nonlinear filter that replaces each pixel value with the median value of the pixels around it.

In other words, this filter protects the image's edges and details during the noise removal process. Then image enhancement helps to increase the contrast and visibility of the cells in the image. However, in blood sample images, this is a difficult phase since the cells typically appear small and difficult to identify from the backdrop. As a result, the common and straightforward improvement approach known as Histogram Equalisation is used. It will compute the image's histogram as well as its cumulative distribution function. The intensity values are then exchanged for their respective CDF values. The converted image is normalised by scaling it to a suitable dynamic range. It is now simple to do additional analyses. Image segmentation is a digital image processing technique that separates an image into several sections or portions.

The grouping is done by assigning a label to each pixel and collecting the pixels with similar features. The Watershed Transform is an image segmentation approach that uses the 'Catchment basin' principle to separate objects in an image. It is accomplished by creating markers for the image's objects, such as cells and backgrounds. The catchment basins are described by the marks on each object. Then use the image's distance transform to compute the distance between each pixel and the nearest marker. The watershed lines are the boundary between the cell and the background, given by the gradient of the distance transform, and the image is flooded. Objects are recovered from the flooded image, and the region associated with the object is identified through analysis. Although various segmentation approaches (such as Otsu's thresholding, Colour segmentation, and k-means clustering) are available, Watershed segmentation has the advantage of distinguishing cells from the background even when cells overlap.

I. A MACHINE LEARNING MODEL FOR DETECTING LEUKAEMIA.

The feature can be morphological or intensity texture-based. The familiar properties of the cell are its shape, size, and intensity. The features retrieved from textures have a significant impact on diagnosing blast cells since they allow for the selection of regions of interest (ROI). The Grey Level Co-occurrence Matrix (GLCM) is a statistical method for obtaining textural information by evaluating the spatial relationship between pixels. And the features are encoded as numerical values, resulting in a feature histogram or the construction of a feature vector for further analysis. Then feature normalisation and feature reduction are used to eliminate undesirable characteristics and minimise the amount of features. The optimisation is used to improve the model's accuracy iteratively by minimising errors. Some standard optimisation approaches are available, but they only converge to local optima rather than global optima. Particle Swarm Optimisation (PSO) is an optimisation technique that can find the global optimal function. This operates by initialising the particle position and velocity in the search space. Then, each particle (Pk) has a personal best position (pbest) as well as a global best position. Every iteration, the velocity (vk) is updated using the formula below, and the pbest and gbest are adjusted accordingly.

$$v_k^{t+1} = w v_k^t + c_1 r_1 (P_{k,best}^t - P_k^t) + c_2 r_2 (g_{best}^t - P_k^t) \quad (1)$$

Here, c1 is a cognitive constant, c2 is a social constant, and w is the inertia weight. And the position is changed; if the new location provides a better objective function than the current gbest, it will become the new global best position.

$$P_k^{t+1} = P_k^t + v_k^{t+1} \quad (2)$$

The operation is repeated until it reaches a maximum number of iterations or an acceptable error value. Finally, the global best location discovered after iterations is the solution to the optimisation difficulties

Machine learning is an AI technology or application that allows a machine to learn without the need for human intervention. In general, machine learning is roughly classified into three forms, including supervised learning, Unsupervised learning and reinforcement learning. In Supervised learning the labelled dataset is utilized for training the machine whereas in Unsupervised learning the unlabelled dataset is used. In Reinforcement learning the machine learns from taking action and from experiences it will upgrade its performance. In general, there are two types of problems in supervised learning such as classification and regression.

> **"Random forest is a supervised learning technique which is applicable for both classification and regression."**

Random forest is a supervised learning technique which is applicable for both classification and regression. This technique combines different classifier algorithms to solve complex computational problems so that it is known as ensemble learning method which makes the algorithm more efficient and perform with high accuracy. The ensemble learning consists of two techniques Bagging and Boosting. The major difference between those two techniques is bagging develops model independently and concatenate them to get the mean value whereas boosting makes iterative model by adjusting the previous classifiers errors. The principle utilized by Random Forest algorithm is Bagging. In general, Random Forest algorithm combines multiple decision trees and take the average of them to give the output. Here the dataset is divided into samples or groups in random manner and each sample act as individual model or tree.

The samples generated from the given dataset with replacement is known asrow sampling. The process of creating samples with row sampling and feature sampling is called bootstrap. Then each model is trained with labelled data separately to get the result. Finally, the predicted output from all the model is combined and based on majority voting the most appropriate result will be obtained and this process is referred as aggregation. In this algorithm the random sampling of data for training set and random generation of samples for model introduces randomization. This helps in constructing the decision tree with less correlation. As a result, the generalization error of the ensemble learning can be improved. The Generalization error is calculated by,

$$G_e = \frac{\delta^{-}(1-S^2)}{S^2} \qquad (3)$$

Decision tree is used where the dataset is complex and unable to form the boundary using single line. In that case there is a need to split the data with different boundaries which forms a tree like structure. It is also a Supervised learning algorithm, applicable for both Regression and Classification problems. Decision tree can involve both numerical data and categorical data. To build the decision tree, the CART (Classification and Regression Tree) Algorithm is used.

$$Entropy(S) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$$

In this case, P(Y) represents the probability of yes, P(N) represents the probability of no, and S represents the sample space. When the subset is pure—that is, devoid of randomness—at every node, the

In the event of complete randomness, entropy will be high and would be low. The changes in permeability are evaluated by using the phrase "data gain." If the entropy is low, the data gain is going to be large; conversely, if it is substantial, the data gain will become low. Therefore, the strategy that offers a significant information gain for each node is selected in order to achieve higher performance. The following formula is used to determine knowledge gain:

$$Information\ Gain\ =\ (S) - [(WA) * E(f)]$$

In this case, WA stands for the volatility of samples, and E(S) the weighted average, and each feature's entropy is shown by E(f). Euler is an additional method to picking the attribute. Index. The Gini coefficient is an index of cleanliness or impurities used while generating decision trees using the CART algorithm. In this instance, the attribute with a small Gini Index has been chosen over the one with a higher Gini Index.

A straightforward approach used for regression and classification is the K-Nearest Neighbor. although generally favored for categorization issues. This method, known as supervised learning, makes predictions about new data by comparing it to existing labeled data that shows similarities. For small, labeled, and noise-free datasets, KNN is typically chosen. This algorithm is also referred to as a lazy learning algorithm since it only stores the training data and makes no learning during the training phase. KNN functions fundamentally by loading all of the examples that are currently accessible and classifying the newly collected data according to the features that distinguish the new data from the previously collected data.

Generally speaking, one may compute the distance. with any distance measurement, such as the Minkowshi or Euclidean distances. However, Euclidean is frequently chosen .To find the gap across data points, use an offset measure. The formula for computing the Euclidean distance is,

$$Euclidean\ distance\ d(X,Y) = \sqrt{(X_2 - X_1)^2 - (Y_2 - Y_1)^2}$$

where the regions' coordinates (X, Y) are delimited by the points X1, X2, Y1, WHICH IS and Y2. The similarity will get closer with less distance and vice versa. Once the distance between each of the k neighboring points has been focused, identify a group whereby the majority of the points have a smaller distance. The class with the greatest number of peers corresponds to how the new point of data is classified. It is desirable for the factor k value to be both excessively high and low. By taking the square root of the overall amount of data points, it may be determined. Furthermore, the value of k need to be odd instead of greater than the entire amount of classes that are offered.

Predictive computing is similar to logistic regression method. Regression analysis generally assesses the connection between the dependent and independent variables. changeable. In this case, the expected value will be categorical data, and the outcome will still be either true or false, 1 or 0, yes or no. The goal variable that needs to be predicted is known as the dependent variable. While the values in linear regression are continuous, the values in logistic regression should be in discrete form. The equation is changed to provide values between the ranges of 0 and 1 because the straight line equation fitted to the dataset yielded values outside of that range. Equation represents the modified S-shaped function, which is a Sigmoid logic or logit function.

$$Sigmoid(Z) = \frac{1}{(1+e^{-Z})}$$

One kind of function for activation that helps with the conversion of linear to non-linear output is the sigmoid function. Numerous function of activation types are accessible, including Binary Step, Linear, Tanh, and ReLU. The sigmoid function aids in transforming input values into the range between 0 and 1, regardless of -∞ to +∞. The expected values, however, will fall between 0 and 1, not precisely between those two numbers. To determine if a number is 0 or 1, one uses the threshold value. It displays the likelihood of winning (represented by a value of 1) and losing (represented by a value of 0). Predicted values over the threshold are regarded as 1, and those below the threshold are regarded as 0. The last logistic regression the formula, which is expressed as, is derived by altering the linear equation.

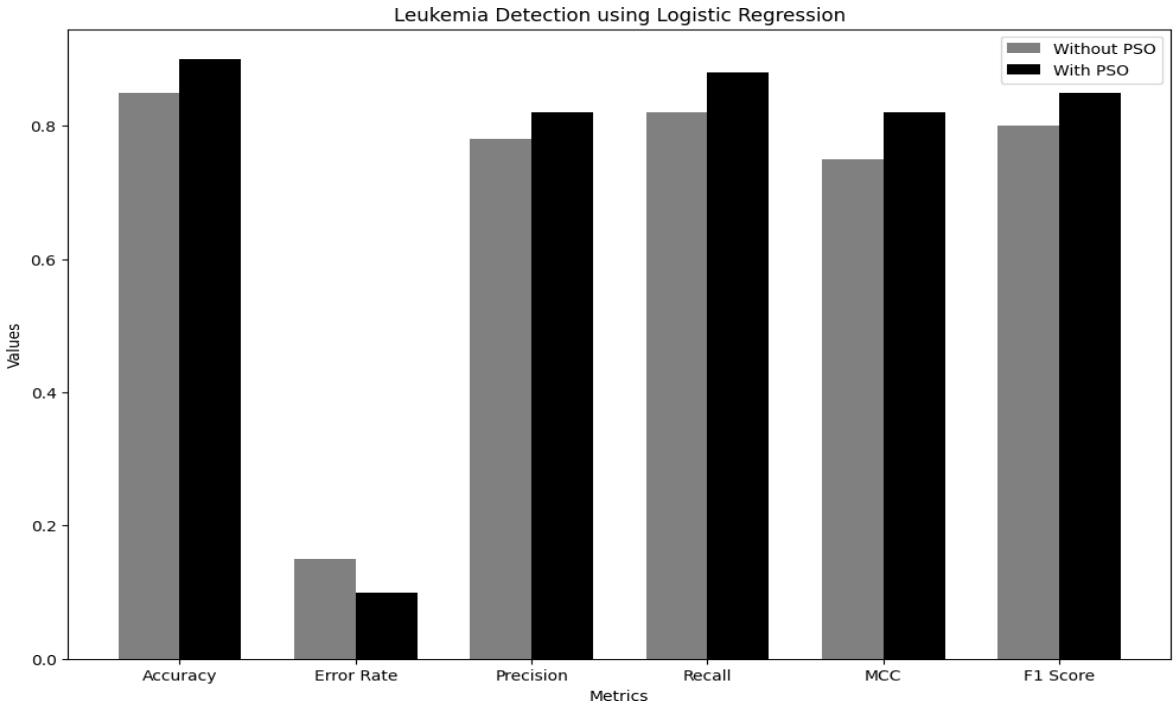
$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

A well-known supervised learning technique that supports both regression and classification but performs better for classification is the support vector machine (SVM). SVM is mostly used to determine the optimal boundary or line that aids in classifying the n-dimensional space. The boundary line in two dimensions is illustrated by straight line as opposed to a plane in three dimensions. However, ndimensional space is needed because there will be n features in real-time applications. Additionally, a hyperplane is defined as the boundary line in n-dimensional space. Although there are other ways to construct the line, the hyperplane with the largest margin is selected as the optimal line.

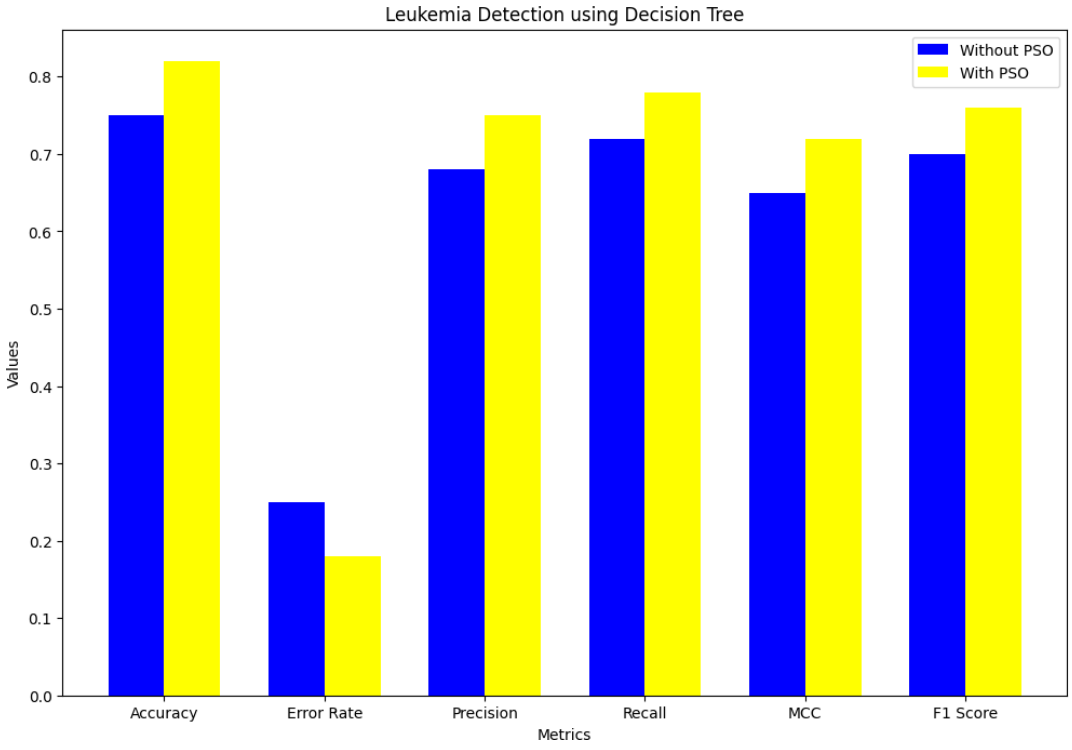
The margin is the distance measured in pixels between the line and the available data points on the space. Consequently, fresh data is accurately categorized into the relevant class upon arrival. To create the optimum fit line, tuning factors include regularization, kernel, and gamma. The line has high gamma if it just takes the near points into account when calculating margin; if it takes into account all of the data points, it has low gamma. Both the high and low gamma techniques are effective, and the best one to use will depend on the dataset that is provided. Regularization is frequently denoted by C. The model may overfit as a result of the high C value.

Section IV: Observations and Remarks.

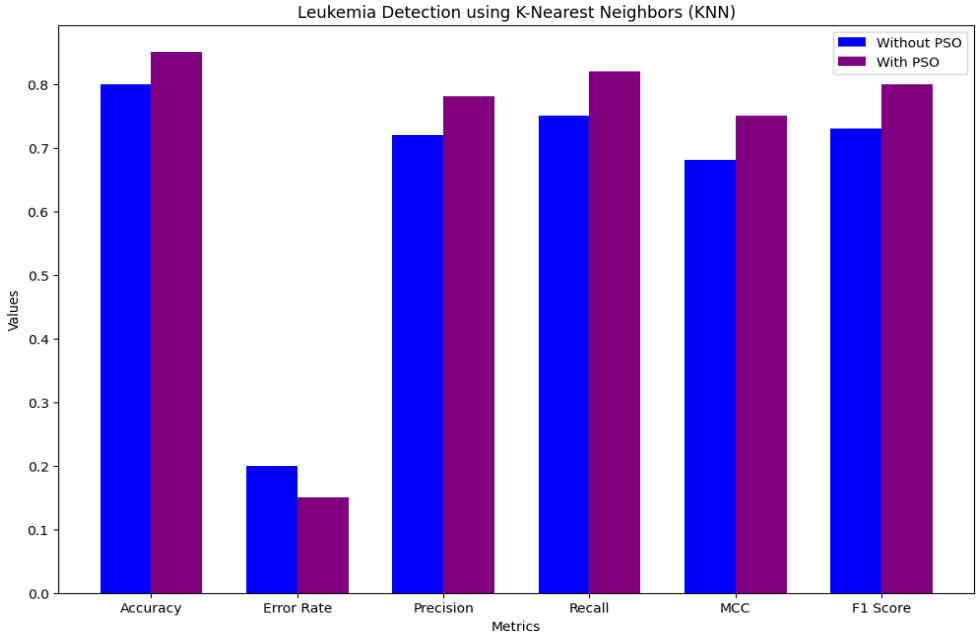
Usually many indicators of accomplishment are used to calculate the machine learning algorithm's performance. To construct performance measures, four terms are used: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Additionally, measures like error rate, F1-score, precision, recall, precision, accuracy, and Matthew's correlation coefficient (MCC) are used to assess the different classification methods. The ratio of correct estimates (TP and TN) to the overall quantity of forecasts is known as accuracy. Precision is usually understood to be the ratio of the number of valid positive predictions (TP+FP) to the total number of accurate positive predictions. The ratio of true positives to total positives, which includes both correctly anticipated positive and wrongly forecasted negative outcomes (TP+FN), is used to quantify recall. F1: The score is calculated as the ratio of the product of accuracy and recall times two to the total of these two quantities. The correlation between the actual and anticipated amounts is provided by the Matthews Correlation Coefficient. It takes into account each of the four terms included in the confusion matrix, and for a greater correlation, the prediction result needs to be accurate across all four categories. Always falling between -1 and 1, the MCC value range. By contrasting what happened to the model, the error rate indicates the prediction mistake produced by the model. If the failure rate is the only criterion, the model is satisfactory. In this study, the effectiveness of many classification algorithms, including the following: Support Vector Machine, Random Forest, Decision Tree, Logistic Regression, and KNN, has been evaluated using the previously mentioned standards.



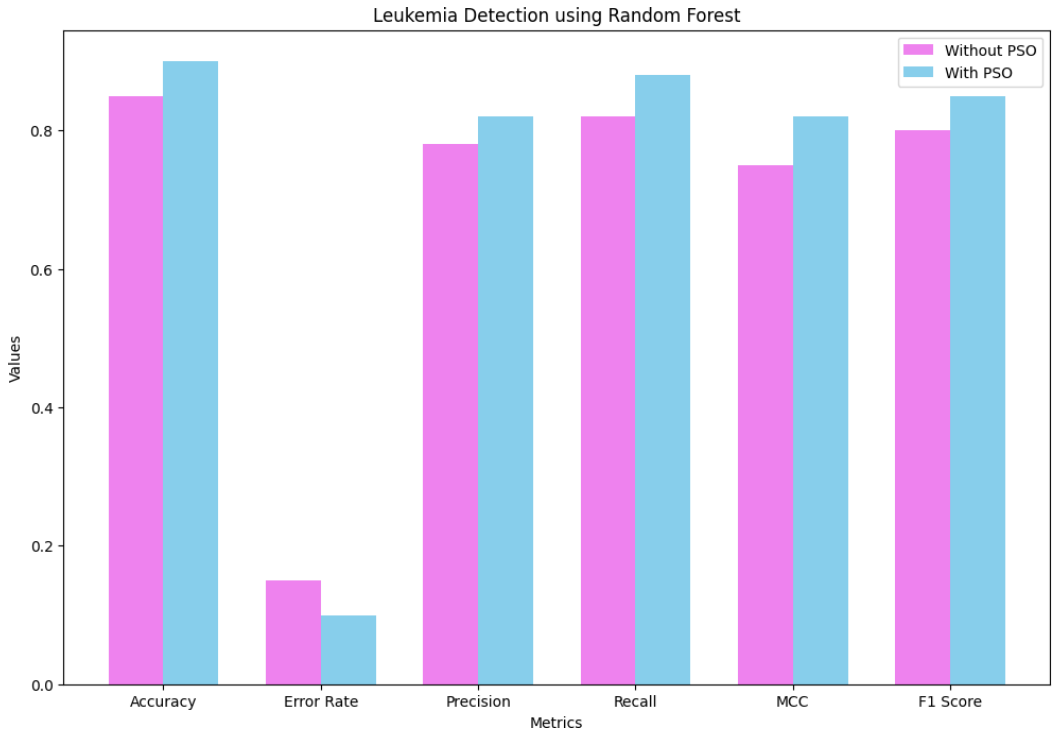
Using the logistic regression approach, the classification was carried out both with and without the use of optimization (X X) 2 (Y Y) 2 2 1 2 1 optimization. Additionally, Fig. 2 presented a comparison of their outcomes. The improved method yields a lower mistake rate while improving accuracy when comparing its findings.



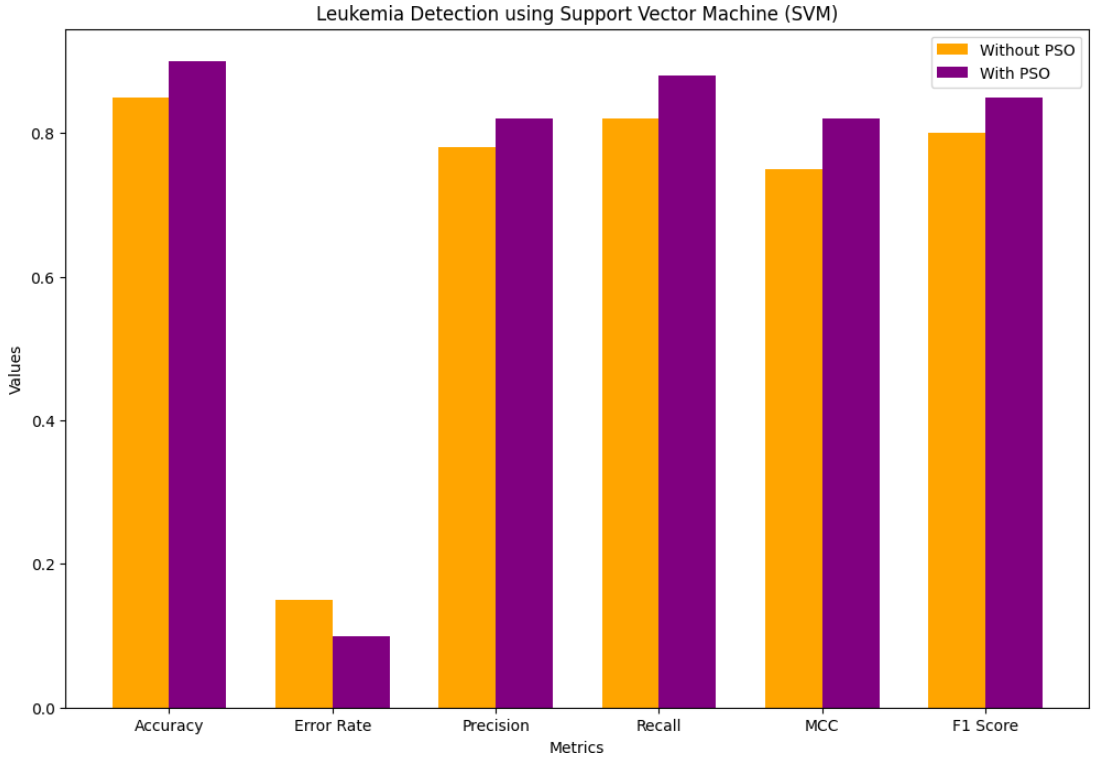
When compared to normal findings, the decision tree's optimization technique yielded somewhat superior results in terms of accuracy, F1-score, and recall; nonetheless, there was a discernible difference in the MCC and error rate when correlated with these metrics, as seen in Figure 3. Thus, it represents a superior advancement in sample categorization.



The fourth figure shows how the KNN model's performance is assessed using a variety of measures. Although there doesn't exist a significant gain in the model's efficiency when using the PSO method, there is a modest improvement in accuracy and MCC plus an approximate decrease in the percentage of errors.



The Random forest findings can be seen in the following graphic as a comparison chart between classification with and without the PSO optimization technique. When compared to classification without PSO, the more effective solution has a higher accuracy, precision, and recall %.



The support vector system improves efficiency in optimal results. With the exception of the optimization approach, every performance metric exhibit respectable gains when

tied to outcomes. Additionally, as the diagram below indicates, SVM behaved better overall.

V. SUMMARY & FUTURE RESEARCH

Various classifiers and the particle swarm optimization approach are used to choose the necessary samples and disregard the rest in the categorization of samples for the supplied dataset. Initially, a comparative analysis was conducted between classifiers that had optimization and those that did not. With an accuracy rate of 77.38%, the SVM fares well in classifications that do not use PSO. The model's efficiency is improved for each classifier when compared to the model without optimization. Additionally, when compared to other classification techniques, SVM's effectiveness in the model that incorporates PSO approaches was good and gave results that were suitable. SVM accuracy when combined with PSO is 78.9%. By using deep learning networks, this model's profitability and accuracy can be boosted in the future.

V1. REFERENCES

[1] Leukemia. Australian Society of Haematology. https://www.hematology.org/education/patients/bloodcancers/leukemi a. Accessed Oct. 16, 2020.

[2] M. Akter Hossain, M. Islam Sabik, I. Muntasir, A. K. M. Muzahid'll Islam, S. Islam and A. Ahmed, "Leukemia Detection Mechanism through Microscopic Image and ML Techniques," 2020 IEEE REGION 10 CONFERENCE (TENCON), Osaka, Japan, 2020, pp. 61- 66, doi: 10.1109/TENCON50793.2020.9293925

[3] Pathi rage, Sachetan & Maracanã, Shavini & ChandraNandan, Supun & Amaratunga, Nishika. (2016). Detection of Leukemia using Image Processing and Machine Learning. 10.13140/RG.2.2.18469.24804.

[4] T. A. M. Elhassan, M. S. M. Rahim, T. T. Swee, S. Z. M. Hashim and M. Alur, "Feature Extraction of White Blood Cells Using Cymometer Localization and Deep Learning in Acute Myeloid Leukemia Blood Smear Microscopic Images," in IEEE Access, vol. 10, pp. 16577-16591, 2022, Doi: 10.1109/ACCESS.2022.3149637

[5] S. Patil Babasa, S. K. Mishra and A. Junnarkar, "Leukemia Diagnosis Based on Machine Learning Algorithms," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bengaluru, India, 2020, pp. 1-5, Doi: 10.1109/INOCON50539.2020.9298321.