Random Oversampling Bagging Ensemble Model for Leukemia Classification

Dr.Mariena A A[1],
[1]*Assistant Professor, Department of Computer Science, Little Flower College (Autonomous), Guruvayur*

***Abstract*—In medical research, Digital Image Processing plays a crucial role encompassing image acquisition, contrast enhancement, image segmentation, feature extraction, and classification. During the feature extraction phase, the Gray Level Co-occurrence Matrix (GLCM), along with statistical and geometrical features from blood smear images, are extracted to form a feature vector. These extracted features are then used for leukemia prediction and classification. A Support Vector Machine (SVM) can be employed for prediction, while a Random Oversampling-based Bagged Ensemble method will be utilized for final classification. The performance of the proposed classification model should be evaluated using metrics such as accuracy, precision, and recall.**

***Index Terms*—Classification, Ensemble Model, Random Oversampling.**

I. INTRODUCTION

An ensemble model is a composite one that combines the predictions from other models. It also combines multiple instances of a base learner and the results are used for prediction purposes. Ensemble improves the performance and it analyses the output that is hard to analyze. Mostly, there are four types of ensembles learning methods like bagging, boosting, randomization and stacking. In bagging, the user has to develop different decision structures by having dissimilar training sets of the same size. It can be done by sampling the original dataset. The sampling procedure can be done by replacing the data in the dataset. The model for each decision tree has been built by using the same learning algorithm. Finally, the user has to combine the Predictions from various models using voting procedure. Bagging is a straightforward method for manipulating the training set .It produces bootstrap replicate of the original training set that consists of sub samples taken by sampling with replacement.

A decision tree is one tree in which the nodes represent decisions based on the features and the edges are binary representing possible paths from one node to another. A binary DT separates the data into two subsets by calculating the best feature split determined by a chosen split criterion. The two resulting subsets become the new parent nodes and are divided further into two child nodes until all observations have been classified. The decision tree is composed of multiple judgment nodes, representing a mapping relationship between attributes and values. A decision tree works through several phases. It starts from the root node by testing the corresponding attributes in the items to be classified. Then it selects the output branch according to their value until reaching the leaf nodes. Finally, the classes of these leaf nodes are outputted as the decision results. The core of the decision tree algorithm is the selection the appropriate splitting conditions. Thus, it is very important to choose the appropriate measure metrics of order and then qualify the split by information gain. Decision trees are sensitive to the specific data on which they are trained. If the training data is changed, the resulting decision tree can be quite different and in turn, the predictions can be quite different. The decision tree learns the dataset recursively splitting the dataset from the root onwards according to the splitting metric at each decision node.

II.LITERATURE REVIEW

Leukemia ends up in mortality due to the time lag in its detection and treatment. Therefore, a system for its quick detection and treatment is a medical imperative. The detection of leukemia and its classification have many stages and the final stage is the latter. A classifier is trained to classify unlabeled or unknown type of the disease based on a given

training set. There are two kinds of classification - supervised and unsupervised. Whereas the former provides a collection of pre-classified images and the latter does not have to consider any given prior data to be trained. Among the many classes of leukemia, SVM, KNN and decision tree are the most useful. SVM is a classification algorithm applied in many fields of science, and these areas comprise face detection, leukemia detection and object recognition. SVM training algorithm builds a model that combines new instances into one. The algorithm outputs an optimal hyperplane which categorizes new instances. The distinction between these groups is by drawing a line between the two classes. In case there are more than one line separating them, the best line between subsets is found and the same is called hyperplane. The SVM algorithm based on the hyperplane gives the largest minimum distance to the training classes. The separating hyperplane maximizes the margin of the training data. Decision Tree is a hierarchical method consisting of rules that divide the independent variables into homogeneous areas. Basic idea behind DT is to get a set of rules that can be used in the prediction process through the results that are obtained from the data and the input variables. A Decision Tree is called a regression or a classification tree if the target variables are continuous or discrete, respectively. Cho J H et al. [2] used Decision tree and applied in a large number of areas, including health care and in the prediction and classification process. Ahmed S. Negm et al. [1] suggested Multilayer Perceptron for classification. The cells are classified according to their morphological features. Initially, the WBCs are detected using Kmeans technique and the morphological features are extracted for classification. The proposed work is tested with dataset consisting of 757 images as preferred by a pathologist. There are three types of images including blast, myelocyte, and segmented cells. The proposed work ensures overall accuracy of 99.51%, the sensitivity of 99.348%, and specificity of 99.529%. The neural network is more sensitive than decision tree in classifying the three leukemia cells in the dataset. It lacks in accuracy in the case of huge dataset. Kumar P S et al. [6] designed an automated system for leukemia classification using multi class SVM. Initially, the nucleus of WBC was extracted by KM technique. The blasted color of the nuclei, The GLCM features were extracted and classified as cancerous or non-cancerous cells or its subtypes. The accuracy of the classifier obtained a value of 90%. The experimental results showed that proposed algorithm could attain an improved performance in the diagnosis of AML and ALL. This technique focused only on two types of leukemia. Minal D Joshi et al. [7] designed an automated system for segmentation and classification of acute leukemia from blood smear images. KNN was used for classification of blast cells.108 images were taken for testing, and 93 percent accuracy was obtained. Yet, the system has not classified other types of leukemia. Muhammed Sajjed et al. [8] described an approach namely EMC-SVM for classification of leukocytes. Among the dataset, 70 percent was utilized for training and 30 percent of the data was used for testing purpose. The proposed classifier obtained an accuracy of 98.6% compared to Naive Bayes and linear classifier. The ensemble SVM has had better accuracy compared to traditional SVM. Agaian et al. [3] suggested a classification method consists of various validation techniques such as k-fold validation, holdout validation and leave-one-out validation. In holdout method, the data are split into two non-overlapped parts: one for testing and the other for training. In k-fold cross-validation, the data are partitioned into k equal size sets. ‘Leave one out’ is working by taking all the data except for a single observation, which is used for training in every iteration and the model, is tested on that particular single observation. The proposed approach is very efficient and applicable for acute leukemia diagnosis. However, it is time consuming.

Reta et al. [9] introduced KNN technique with Euclidean distance metric for classification. The edges were converted from Cartesian space to polar space to split the overlapped regions and discontinuous points were interpolated using linear interpolation. Totally 633 images were used for classification and 80 cells for testing. Various features like statistical, geometrical and texture features were used for classification. This approach gives 92.5% accuracy.

III.RANDOM OVERSAMPLING BASED BAGGING ENSEMBLE MODEL

A.Oversampling

Random oversampling tries to balance class distribution by randomly duplicating minority class instances. However, this technique can hike the possibility of overfitting since it makes precise copies of existing instances. Here the user has less number of images for CLL and CML. Hence, there is a need to balance the dataset. Sampling techniques are the most useful for handling the imbalance problem. The oversampling performed on the minority class instances to balance the dataset also increases its size. The main advantage of oversampling is that it does not lose any significant information from the dataset. The most used technique for oversampling is the random oversampling method. Here the minority class features are replicated until the samples of both classes are balanced. Yet, there is a chance for overfitting since same instances occur multiple times. Oversampling can provide a balanced distribution without losing information on majority class that increases the number of the minority class by the duplication of the original data.

B. *Bagging*

Breiman L et al. [5] defines that bagging is the process of creating bags that comprise subsets of original dataset using sampling with replacement method. Bootstrap Aggregation is a procedure that can be used to reduce the variance for those algorithms that have high variance like decision tree. It derives a model for each subset using a learning scheme and the new instance is classified using majority voting from the desired models. All the models have a result that contains class label and the class label receives majority votes assigned to the instance C. *Architecture of Proposed ROBE Model*

Fig:1 Overall Framework

D. *ROBE Algorithm*

Input: Training set S, no. of iterations n.

Output: Ensemble Model, M*

1. Start
2: for i=1 to n do
3: create bootstrap sample S_x by sampling n with replacement.
4: use S_x and learning scheme to derive a model M_x.
5: end for
6: To use M* to classify a new instance, i_{new}: each $M_x \in M^*$ classify i_{new} and return majority vote.
7. Stop

E *Random Forest Model*

Breiman L et al. [5] defines The random forest (RF) Model is an ensemble learning technique consisting of the aggregation of a large number of decision trees, through bootstrap aggregation resulting in a reduction of variance compared to the single decision trees. Machine learning deals with the analysis and creation of algorithms that facilitate the making of data-driven decisions from previously provided data. The algorithms use the provided data, also known as training data, to "learn" by building a model and utilize it to generate predictions. Machine learning algorithms can be broadly classified into two categories- supervised learning and unsupervised learning. Supervised learning is the process of identifying new or unknown instances by using classification algorithms on a group of samples with known class values. This training data is used to facilitate the creation of a predictive model. Random Forest is one of the supervised classification algorithms that works in three phases. In the first phase, it extracts K training sets by using bootstrap random sampling with size of original training set. Secondly, it creates the decision tree for each of the bootstrap training sets to produce k decision trees to form a forest. Finally, voting is made based on output from each decision tree. The training process of each decision tree is independent so that random forest model can work in parallel improving efficiency. A 'Decision Tree' is considered as the base classifier for the random forest model and this model synthesizes the output of multiple decision trees by using majority voting. The majority of the votes decide the overall prediction of Random Forest model. This aggregate vote of several DTs is inherently less noisy and less susceptible to outliers

than a single DT output. The output of random forest is determined by the result of the decision tree with highest vote. The specific working of this model is shown in figure 5.1 RF model has better accuracy compared to other models. Moreover, it deals with large datasets. The number of decision trees in the random forest model is the key to the result of the model. The training time increases along with the number of decision trees. The user has chosen the number of trees by increasing the number of trees on run after run until the accuracy begins to stop showing improvement.

IV.PERFORMANCE ANALYSIS

The accuracy is the most important performance metric. It is the simple ratio of correctly predicted observation to the total observations. If a system gets high-level accuracy, it is the best model. The precision defined as a ratio of correctly predicted positive observations to the totally predicted observation. The recall is also known as sensitivity and it is the ratio of correctly predicted positive observation to all observation in actual class. Finally, f1 score is the weighted average of precision and recall and it is nothing but the harmonic mean between precision and recall. So these parameters are very helpful to in measuring the performance of the proposed model.Accuracy is the proportion of the total number of predictions that are correct, which is determined by the following formula:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Precision is the proportion of the predicted positive cases that are correct, which is calculated using:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall also called Sensitivity or True Positive Rate (TPR), is the proportion of positive cases that are correctly identified, which is calculated using:

$$\text{Recall}= \frac{TP}{TP+FN}$$

Where, TP is the positive tuple correctly classified by the classifier.TN is the negative tuple correctly classified by the classifier. FP is the negative tuple that are incorrectly labelled as positive. FN is the positive tuple mislabeled as negative.

Table 1 Performance analysis on Accuracy

Si. No	Model	Accuracy
1	Stochastic Gradient	48
2	Decision Tree	84
3	Random Forest	83
4	ROBE	89

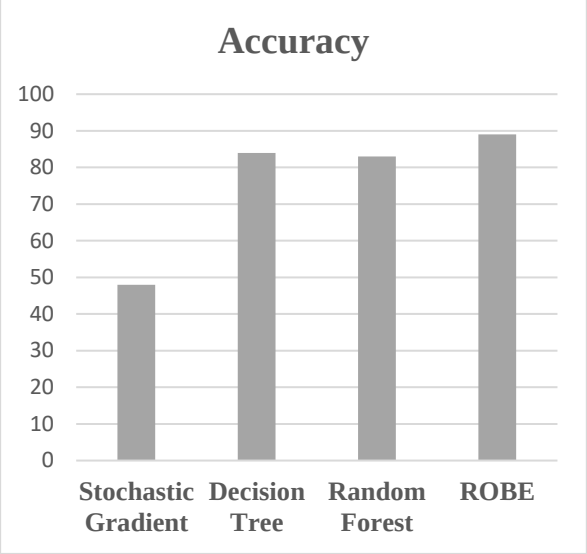

Fig.3 Graphical Representation of Accuracy Results

Table 2 Performance Analysis on Precision

Si.No	Model	Precision
1	Stochastic Gradient	55
2	Decision Tree	85
3	Random Forest	87
4	ROBE	90

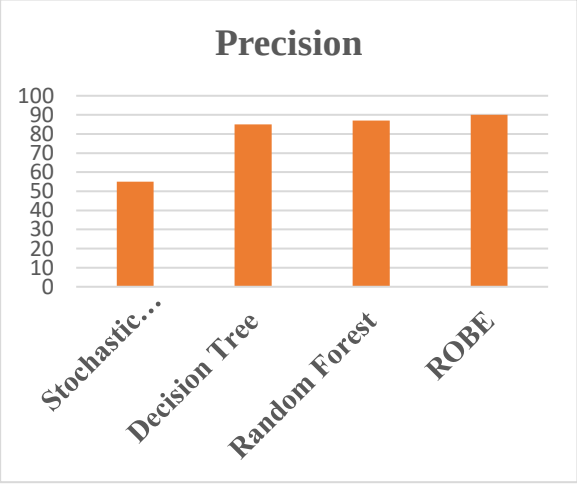

Fig.4 Graphical Representation of Precision Results

Table 3 Performance Analysis on Recall

Si. No	Model	Recall
1	Stochastic Gradient	48
2	Decision Tree	84
3	Random Forest	83
4	ROBE	89

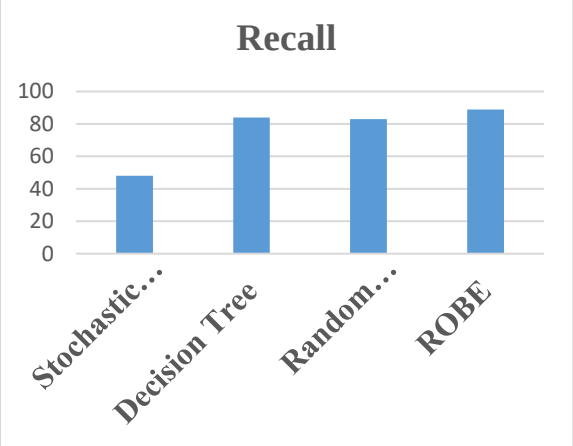

Fig.5 Graphical Representation of Recall

V. EXPERIMENTAL STUDY

Performance of the proposed architecture is evaluated using Accuracy, precision and recall. The ROBE model has improved the classification accuracy from 72% to 89%, Recall improved from 72% to 89%, and Precision improved from 76 % to 90 %. From these results, it is observed that the proposed approach works better for classification.

VI. CONCLUSION

The proposed technique has been tested on its performance based on quality parameters such as accuracy, precision and recall. Totally, there are eight intensity histogram features, which have been extracted from the image. Twenty-one GLCM features are extracted for classification. The ROBE model has been used for final classification task. The classification accuracy of this technique has been compared well with the other approaches like decision tree, logistic regression and random forest. The proposed technique outperforms the existing techniques and provides better results for classification.

REFERENCES

[1] Ahmed S. Negm, Osama A. Hassan, Ahmed H. Kandil. (2017)," A decision support system for Acute Leukaemia classification based on digital microscopic images*", Alexandria Engineering Journal, 57(4), 2319-2332.*

[2] Cho, J. H.,&Kurup, P. U. (2011)," Decision tree approach for classification and dimensionality reduction of electronic nose data", *Sensors and Actuators B: Chemical, 160(1)*,542-548.

[3] Agaian, S,Madhukar, M.,Chronopoulos, New Acute leukaemia-automated classification system*", ComputerMethodsin Biomechanics and Biomedical Engineering: Imaging & Visualization, 6(3),* 303-314.

[4] AshwiniRejintal, &Aswini, N. (2016)," Image processing-based leukemia cancer cell detection", *International Journal of Engineering Research & Technology, (06).*

[5] Breiman, L. (1996)," Bagging predictors. *Machine Learning", 24(2),* 123-140.

[6] Kumar P S Vasuki S. (2017)," Automated diagnosis of acute lymphocytic leukemia and acute myeloid leukemia using multi-svm", *Biomed Imag Bio eng, 1(1).*

[7] Minal D. Joshi, Atul H. Karode, S.R. Suralkar. (2013)", White Blood Cells Segmentation and Classification to Detect Acute Leukemia", *International Journal of Emerging Trends & Technology in Computer Science, 2(3)*

[8] Muhammad Sajjad, Siraj Khan, Zahoor Jan, Khan Muhammad, Hyeonjoon Moon, Jin Tae Kwak, Seungmin Rho, Sung Wook Baik, Irfan Mehmood. (2016), "Leukocytes Classification and Segmentation in Microscopic Blood Smear: A Resource-Aware Healthcare Service in Smart Cities", 2169-3536 (c)

[9] Reta, C., Altamirano, L., Gonzalez, J. A., Diaz-Hernandez, R., Peregrina, H., Olmos, I., Alonso, J. E., & Lobato, R. (2015). Segmentation and classification of bone marrow cell images using contextual information for medical diagnosis of acute leukemia. PLOS ONE, 10(6), e0130805.

[10] Engineering, Management & Applied Science, 6(4).