Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Machine learning based analytical system for predictive detection of Leukemia using WEKA

Ms. Harsh Paliwal[1]* and Dr. Kuntal Barua[2]

[1] Ph.D Research Scholar, Institute of Advance Computing, SAGE University, Indore, Madhya Pradesh, India

harsh.paliwal113@gmail.com

[2] Associate Professor, Institute of Advance Computing, SAGE University, Indore, Madhya Pradesh, India

kuntal.barua@gmail.com

Article Info
Page Number: 8880-8894
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 15 September 2022
Revised: 25 October 2022
Accepted: 14 November 2022
Publication: 21 December 2022

Abstract—
In recent times, classification of leukemic blood cell by using machine learning techniques has gained the attention of many researchers for developing an automated model which can assists doctors in detection of leukemia. Also, it is quite challenging to accurately predict the blood cancer as symptoms are very general in initial stages. In this manuscript, we have presented an approach for predictive detection of leukemia by observing the important features from the blood test and using various classifiers. We have observed that AdaBoostM1 classification algorithms gives better result than Bayes Net classifier. We have also derived some most important features age, infected ("Yes", "No"), white blood cell count, red blood cell count, platelet count, leucocytes count, mch, hemoglobin, hematocrit, neutrophils, eosinophils, lymphocytes, monocytes, basophils, mpv, nrbc hash, diastolic blood pressure, total cholesterol, triglycerides, hdl cholesterol, which has significant impact on leukemia detection. We have achieved 98.50% accuracy, 96.99% sensitivity, 98.7% specificity and 98.30% precision values for detection of leukemia by using Random Forest classifier.

Keywords— Classification, Bayes Net, blood cancer, Leukemia,.

I. INTRODUCTION

The Leukemia is a type of cancer, in which a large amount of immature blood cells get developed in bone marrow and also get spread in the body. Blood cancer or Leukemia is developed due to abnormality in production of blood cells. Leukemia initiates from the bone marrow and results in the manufacturing of large amounts of abnormal cells. By the time, these cells enter into the body tissues and cause fatal diseases. WHO has declared cancer as 2nd among deadliest diseases[1]. This disease affects the immune system of the human body. Depending on the growth rate leukemia can be classified as acute or chronic leukemia. The subclasses of leukemia further vary depending on the fact that which type of leukocytes is being affected. If the affected leukocytes are lymphocytes then leukemia is identified as Lymphocytic leukemia, and if monocytes and granulocytes are found to be abnormal then it is named as myeloid Leukemia. Main cause of leukemia is production of immature WBC in bone marrow. In the year 2015 it is observed that around 8, 76,000 people were diagnosed

with ALL globally and out of them 111,000 people died of this disease [2] [3]. Leukemia can occur in any person of any age group from children of two year to older people of 60+ years of age. It is observed that early detection of leukemia is crucial in saving lives of patients. The common symptoms that can be seen in a leukemia affected patient are the pale color of skin, tiredness in a patient, enlargement of lymph node, fever, pain in joint etc. Various ways are available for detection of blood cancer including complete blood count test, blood protein test, biopsy test of bone marrow, and analyzing microscopic images of cell.

In automatic detection of leukemia from a given dataset of patient's blood parameters machine learning algorithms play a vital. Haneen T. Salah et. Al.[4] discussed about applications of Machine learning in the diagnosis of leukemia. Deep Convolutional neural networks were used to diagnose leukemia and its various classes identified by French American British (FAB) classifications with higher accuracy. Metrics which are used widely for evaluating the performance f a model are sensitivity, specificity, accuracy, precision, and rarely AUC(Area Under Curve). Segmentation are performed on nucleus or cytoplasm. Due to lack of availability of data, data augmentation techniques and Generative Adversarial Network (GAN) are used in order to increase data set size by some researchers. Anamika Das Mouet. Al.[1]considered thirteen attributes, i.e. Gender, Age, Height(cm), Weight(Kg), Body Mass Index(BMI), Diastolic Blood Pressure, Pulse, S.Total Cholesterol(TC),S.Triglycerides, HDL Cholesterol, LDL Cholesterol and class(Yes,No) and found them very informative. The accuracy is measured by using two methods- splitting data and k-fold cross validation. ShakirMahmood Abbas et. Al.[5] proposed a model which they named as COMPUTER-AIDED DETECTION SYSTEMS (CAD3). In proposed model YOLO v2, CNN and Visualization methods are applied for detection, classification and visualization respectively of WBC in input image and provide complete details about number and size of WBC in image. The overall accuracy of the system is 94.3% in detecting and classifying the leukocytes in leukemia. The proposed system can operate on images brought from the laboratories directly without the need of preprocessing. It is observed that type of Leukemia depends on the type of leukocyte which is being affected hence identification of abnormal leukocyte is most important stage in leukemia detection process. A.M.Patilet. Al.[6] proposed a model to perform classification of blood cells from images into four types i.e. Eosinophil, Lymphocyte, Monocyte, Neutrophil. The system is consist of a CNN segment, which uses the Xception model, another stage that uses the two directional long-short term memory model and third stage is the Canonical Correlation Analysis which is used for feature extraction to improve accuracy. Overall accuracy obtained is 95.89%. Due to overlapping of blood cells in images, classification time got reduced, resulting in compressed dimension of input images and faster convergence of networks with more accurate weight parameters. For detection of leukemia WBC, RBC and platelets count plays the most important role, hence Mohammad MahmudulAlamet. Al.[7] proposed a deep learning based blood cell counting method where YOLO is used for automatic identification and counting of these blood cells. YOLO(You Only Look Once) threshold is used for identification of cells. Accuracy of the system for RBC 96.09%, WBC 86.89%, and Platelet 96.36% is achieved. Limitation of the system is that it sometimes double counts the same platelets from the neighboring grid, k-nearest neighbors and intersection over union is used in each platelet to overcome this issue. **Anita**

et. Al.[8] proposed a method in which they used an artificial electric field algorithm (AEFA) and Velocity and Position Clamping Based AEFA (AEFA-C). After preprocessing the image an edge detection (Morphological edge detection) algorithm, for segmentation diffused expectation-maximization is used which gives WBC, RBC and background pixels as output. This edge map image is provided as input to the AEFA-C based ellipse detection scheme. Detection rate of WBC achieved 96.90 % and 3.09% false alarm rate. **Muhammad Shahzadet. Al.[9]** suggested a robust method for semantic segmentation of microscopic images of blood cell which points out the what (semantics) and where (location) about the image which is under observation. The information regarding semantics and location is encoded in a nonlinear local-to-global pyramid fashion by using deep feature extraction. For preprocessing pixel-level labeling is used and acquired masked images are then converted from RGB to grayscale, then pixel fusing, and unity masks are generated. VGG16 is used for feature extraction. System classified RBCs with 97.45%,WBCs with 93.34%, and platelets with 85.11% accuracy and global and mean accuracies were 97.18% and 91.96% respectively. Ahmed T. Sahlolet. Al.[10] suggested an efficient approach for classification of WBC Leukemia. They used VGGNet for feature extraction, statistically enhanced Salp Swarm Algorithm (SESSA) for feature filtration and removing noise and Chi-square is used to remove highly correlated. 83.2% accuracy is achieved in classification of WBC Leukemia with Improved Swarm Optimization of Deep Features.

II. METHODS

The whole method of predictive detection of leukemia has been divided into six sections such as data collection, feature reduction , Pre-processing filtering, building classification model and evaluating classifier as discussed in below subsections.

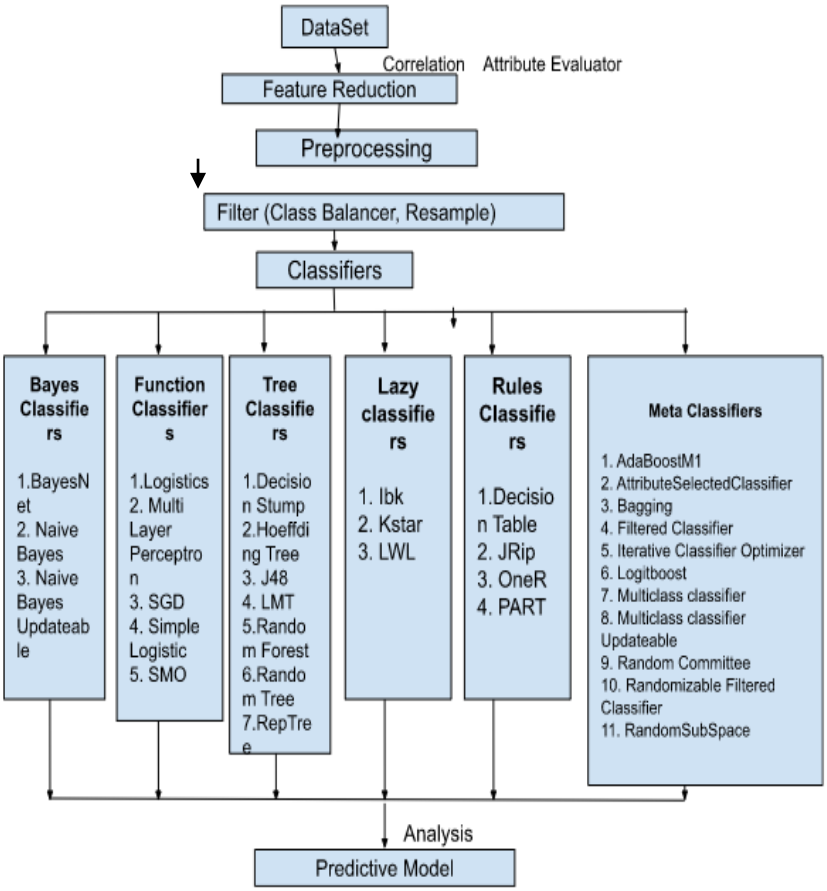


Figure 1: Proposed model

A. Data Collection

For training and testing the proposed model large amount of real time dataset is required. Some local hospitals were visited for this purpose and data of around a thousand patients has been collected, which was consisting of both infected and healthy people.

B. Feature Reduction

It is observed from previous studies also and found correct on reducing the parameters of the dataset accuracy of the model can be improved. Hence some feature reduction algorithms have been used and features are reduced to 20 attributes. These 20 attributes are collection of those attributes which have ranked higher in feature selection process and identified as those having significant impact on detection process.

C. Pre-Processing filtering

In preprocessing stage the dataset collected has been normalized first. For pre-processing Class balancer and resample filters were used. Class balancer selected 200 instances from a dataset of 1000 instances. Filters played an important role in balancing the dataset.

D. Building Classification Model

We have considered total 32 features, which are further analyzed further for checking their importance in finding out the disease. On the basis of feature selection using select attribute evaluator of WEKA 20 features were selected to train the classifiers. Ten different classifiers have been used Bayes net, naïve Bayes, naïve Bayes updateable from Bayes classifiers and decision stump, hoeffding tree, j48, LMT, random forest, random tree, RepTree from Tree classifiers were employed for the detection of leukemia. We have used WEKA tool framework for building the proposed model. WEKA provides various classification algorithms and various testing techniques for evaluating the model.

E. Evaluation of Classification Model

The classification models were evaluated on the basis of various performance metrices like Accuracy, Sensitivity, Specificity and precision. Values of True positive (TP), false positive (FP) and false negative (FN) are used for calculating these performance metrices.

True positive in confusion matrix represents the number of positive instances in a dataset which are predicted as positive i.e. correctly identified positive records. False positive in confusion matrix represents the number of negative instances in a dataset which are predicted as positive i.e. incorrectly identified as positive while they were not positive. Such a scenario is known as Type 1 error. False negative in confusion matrix represents the number of positive instances in a dataset which are predicted as negative i.e. incorrectly identified as negative while they were positive. Such a scenario is termed as Type2 error. True negative in confusion matrix represents the number of negative instances in a dataset which are predicted as negative i.e. correctly identified as negative when they are actually negative.

Accuracy

Accuracy represents the state of being correct, in other words accuracy technically denotes the degree up to which the outcome of a calculation conforms to the standard or correct value. Accuracy can be calculated as either the sum of two correct predictions (TP + TN) divided by the total number of instances in datasets (Positive + Negative) or by dividing the total number of correctly identified instances by the total number of instances. In weka, accuracy of the model can be observed by % of correctly classified instances. The accuracy value considered best is 1.0 and the worst is 0.0.

Recall/sensitivity (%)

Sensitivity is also termed as Recall (REC) or True Positive Rate. It is computed as the number of positive predictions (TP) which are correctly identified by the model divided by the total number of positive (P) instances in the dataset.The best or most desired sensitivity for any model is 1.0 and the worst is 0.0.

specificity (%)

Specificity can be computed by dividing the number of correctly identified negative predictions (TN) by the total number of negatives (N) instances in the dataset. The best and most desired specificity is 1.0 and the worst is 0.0.

Precision (%)

Precision is measured as the number of correct positive predictions (TP) divided by the total number of positive predictions (TP + FP) i.e. the correctly identified negative and positive predictions. The best precision is 1.0 and the worst is 0.0.

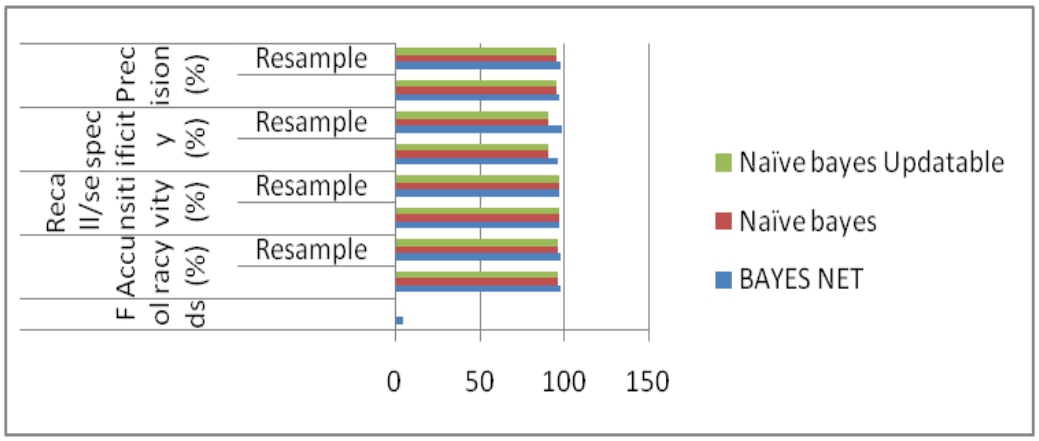
III. RESULTS

We have calculated accuracy, sensitivity, specificity, and precision for each classification model over the dataset by using 5 and 10 folds of cross validation, for evaluating the performance of implemented ten classifiers, which is described in the table I. Also, we have separately evaluated the performance of the classifiers after reducing the parameters in the dataset. Table II shows the performance of classifier by on reduced features. Also we have applied two filters on the dataset which are class balancer and resample.

Performance of Bayesian classifiers for 5 folds of cross validation over complete dataset-

Classifier	**Folds**	**Accuracy (%)**		**Recall/sensitivity (%)**		**Specificity (%)**		**Precision (%)**	
		Class Balancer	**Resample**	**Class Balancer**	**Resample**	**Class Balancer**	**Resample**	**Class Balancer**	**Resample**
BAYES NET	5	98.00	98.50	96.99	96.99	96.81	98.70	97.63	98.30
Naïve bayes		96.40	96.40	96.99	96.99	91.16	90.86	95.46	95.51
Naïve bayes Updateable		96.40	96.40	96.99	96.99	91.16	90.86	95.46	95.51

TABLE I. PERFORMANCE OF CLASSIFIERS for 5 folds

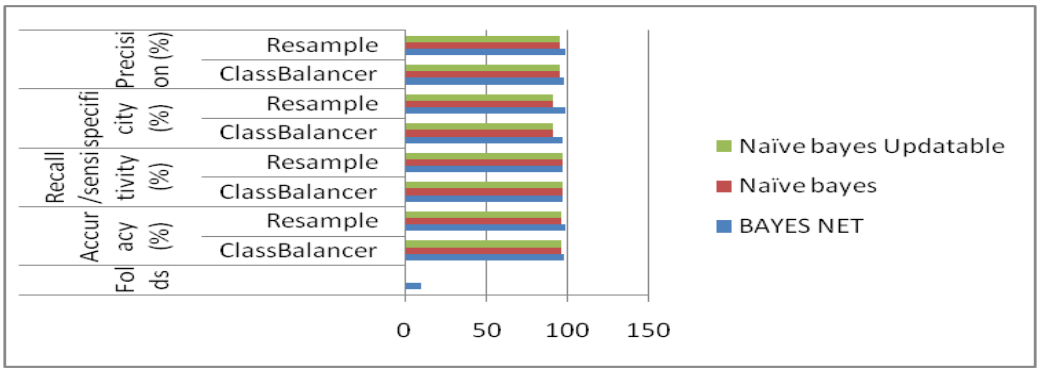


Graph 1: Bayesian classifiers performance over complete dataset for 5 folds

Performance of Bayesian classifiers for 10 folds of cross validation-

Classifier	Folds	Accuracy (%)		Recall/sensitivity (%)		specificity (%)		Precision (%)	
		Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample
BAYES NET	10	98.00	98.50	96.99	96.99	96.81	98.70	97.63	98.30
Naïve bayes		96.40	96.40	96.99	96.99	91.16	90.86	95.46	95.51
Naïve bayes Updateable		96.40	96.40	96.99	96.99	91.16	90.86	95.46	95.51

TABLE II. PERFORMANCE OF CLASSIFIERS for 10 folds

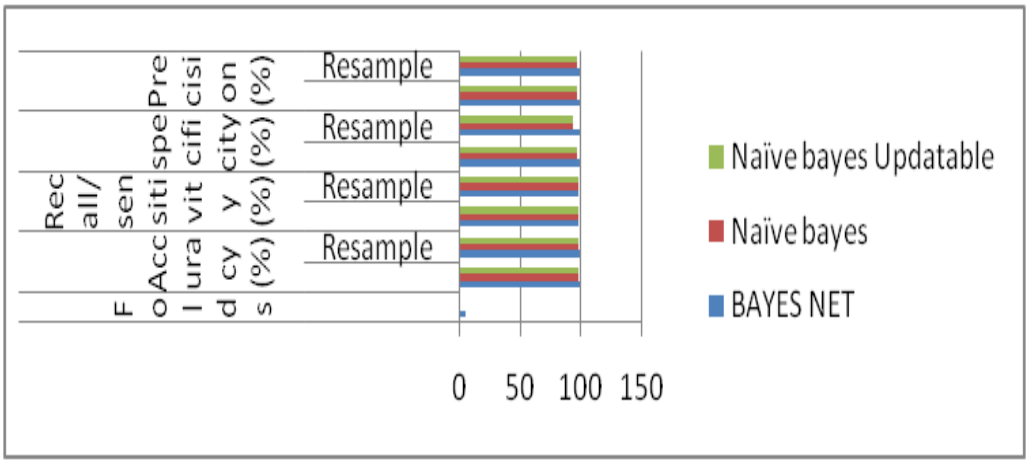


Graph 2: Bayesian classifiers performance over complete dataset for 10 folds

Performance on 20 reduced dataset for 5 folds:

Classifier	Folds	Accuracy (%)		Recall/sensitivity (%)		specificity (%)		Precision (%)	
		Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample
BAYES NET	5	98.50	98.50	96.99	96.99	98.70	98.70	98.30	98.30
Naïve bayes		97.31	96.90	96.99	96.99	96.27	92.62	96.00887	96.18
Naïve Bayes Updateable		97.31	96.90	96.99	96.99	96.27961	92.62	96.00887	96.18

TABLE III. PERFORMANCE OF CLASSIFIERS for 5 folds

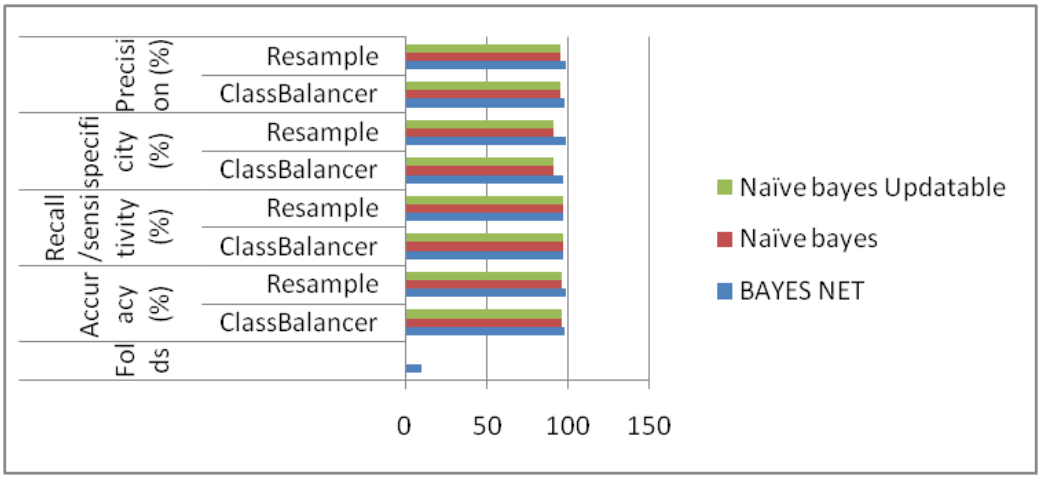


Graph 3: Bayesian classifiers performance over 20 reduced dataset for 5 folds

Performance on 20 reduced dataset for 10 folds:

Classifier	Folds	Accuracy (%)		Recall/sensitivity (%)		specificity (%)		Precision (%)	
		Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample
BAYES NET	10	98.50	98.50	96.99	96.99	98.7	98.70	98.3	98.30
Naïve bayes		97.22	96.80	96.99	96.99	96.105	92.26	95.839	96.04
Naïve bayes Updateable		97.22	96.80	96.99	96.99	96.105	92.26	95.839	96.04

TABLE IV. PERFORMANCE OF CLASSIFIERS for 10 folds



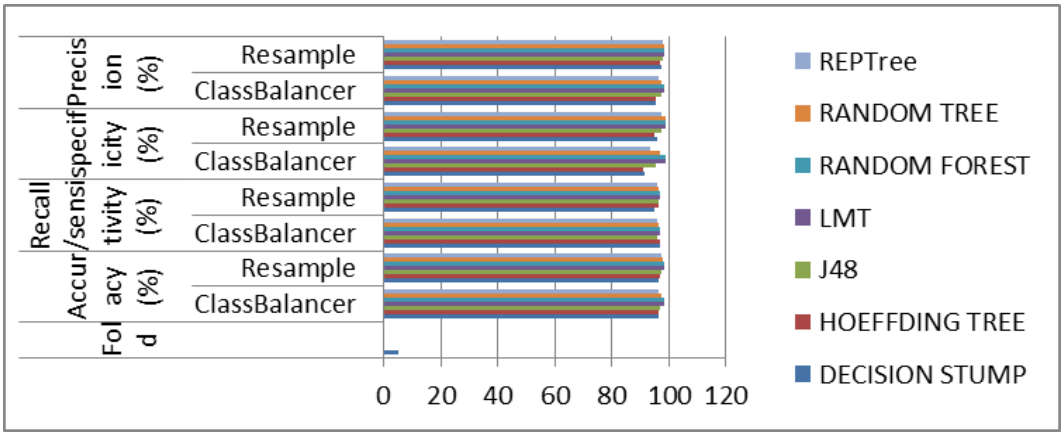
Graph 4: Bayesian classifiers performance over 20 reduced dataset for 10 folds

From the graph 1, 2, 3 and 4 two things can be observed one is Bayes Net is showing better results regarding the overall performance and second is the performance has been improved on reducing the parameters.

Performance of Tree classifiers for 5 folds of cross validation-

Classifier	Folds	Accuracy (%)		Recall/sensitivity (%)		specificity (%)		Precision (%)	
		Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample
DECISION STUMP	5	96.50	96.40	96.99	95.1478	91.51	95.78	95.60	97.37
HOEFFDING TREE		96.20	97.10	96.85127	96.3224	90.80	95.11	95.33	97.10
J48		96.90	97.60	95.91154	96.19	95.60	97.47	97.22	97.90
LMT		98.40	98.40	96.85537	96.85	98.70	98.70	98.30	98.30
RANDOM FOREST		98.40	98.50	96.86	96.99	98.70	98.70	98.30	98.30
RANDOM TREE		97.50	98.10	96.31886	96.4616	96.78	98.70	97.63	98.30
REPTree		96.30	97.30	95.90275	95.80	93.40	97.46	96.41	97.90

TABLE V. PERFORMANCE OF TREE CLASSIFIERS for 5 folds



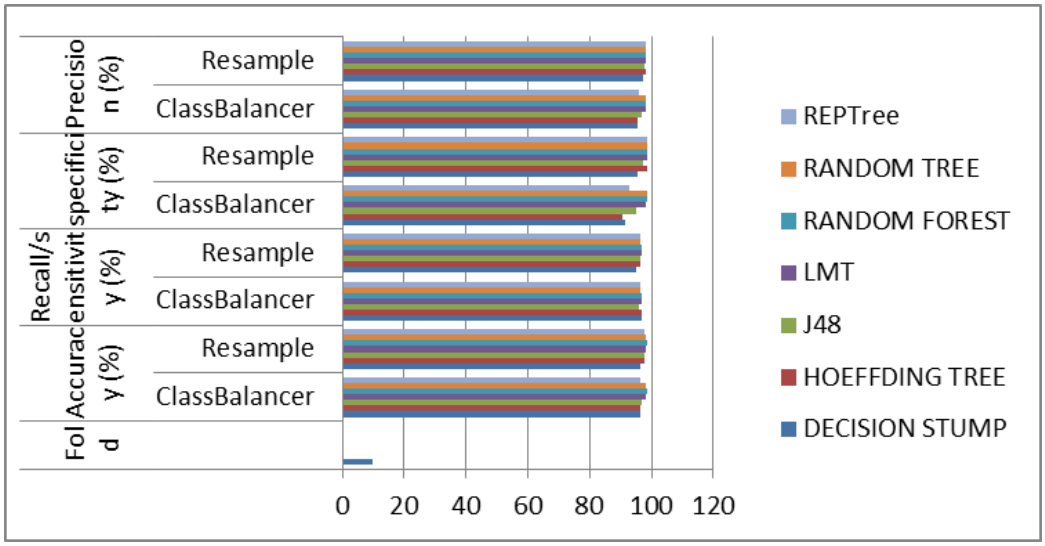
Graph 5: Tree classifiers performance over complete dataset for 5 folds

Performance of Tree classifiers for 10 folds of cross validation-

Classifier	Folds	Accuracy (%)		Recall/sensitivity (%)		specificity (%)		Precision (%)	
		Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample	Class Balancer	Resample
DECISION STUMP	10	96.50	96.30	96.99	95.1454	91.51	95.38	95.60	97.24
HOEFFDING TREE		96.20	97.90	96.85127	96.1994	90.80	98.70	95.33	98.30
J48		96.70	97.90	95.90863	96.5915	94.86	97.49	96.95	97.90
LMT		98.40	98.40	96.99	96.8573	98.31	98.70	98.17	98.30
RANDOM FOREST		98.50	98.50	96.99	96.99	98.70	98.70	98.30	98.30
RANDOM TREE		98.00	98.20	96.32048	96.59317	98.70	98.70	98.30	98.30
REPTree		96.50	97.90	96.44	96.1	92.7	98.7	96.1	98.3

				19	994	8	0	4	0

TABLE VI. PERFORMANCE OF TREE CLASSIFIERS for 10 folds



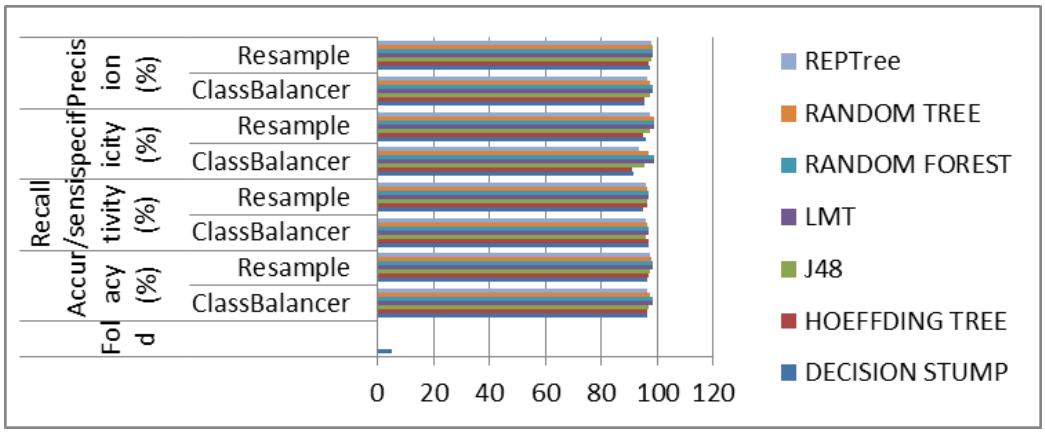
Graph 6: Tree classifiers performance over complete dataset for 10 folds

Performance on 20 reduced dataset for 5 folds:

Classifier	**Folds**	**Accuracy (%)**		**Recall/sensitivity (%)**		**specificity (%)**		**Precision (%)**	
		Class Balancer	**Resample**	**Class Balancer**	**Resample**	**Class Balancer**	**Resample**	**Class Balancer**	**Resample**
DECISION STUMP	5	95.97	96.40	96.99	95.1478	93.69	95.78	93.43	97.37
HOEFFDING TREE		96.27	97.10	96.99	96.3224	94.26	95.11	94.01	97.10
J48		98.16	97.80	96.62669	96.4595	98.38	97.48	98.01	97.90
LMT		98.35	98.30	96.99	96.8572	98.39	98.30	98.01	98.17
RANDOM FOREST		98.50	98.50	96.99	96.99	98.70	98.70	98.30	98.30

RANDOM TREE		96.59	98.40	94.9332	96.85737	96.94	98.70	96.68	98.30
REPTree		97.20	97.30	96.99	95.8042	96.07	97.46	95.80	97.90

TABLE VII. PERFORMANCE OF TREE CLASSIFIERS for 5 folds



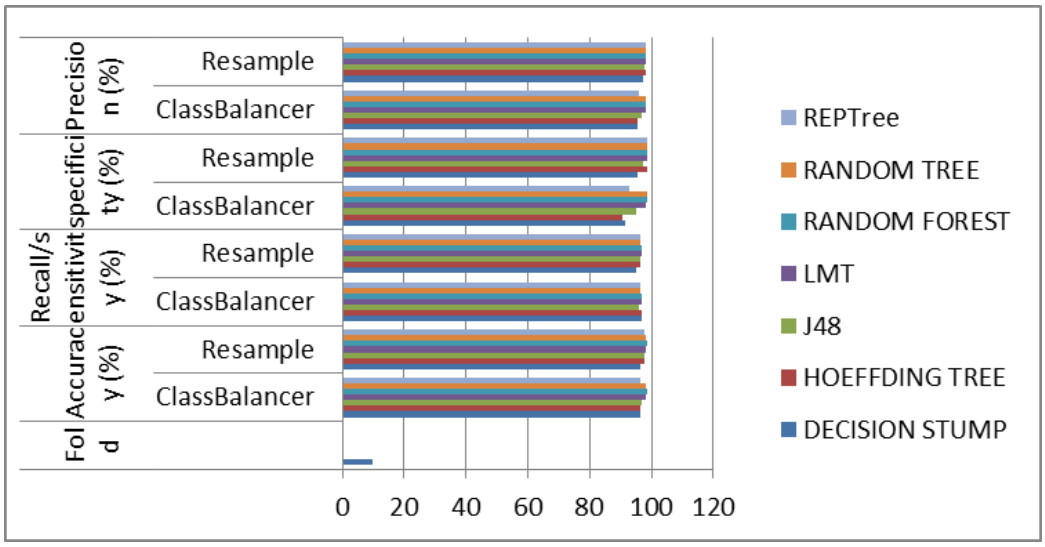
Graph 7: Tree classifiers performance over 20 reduced dataset for 5 folds

Performance of tree classifiers for 10 folds of cross validation-

Classifier	**Folds**	**Accuracy (%)**		**Recall/sensitivity (%)**		**specificity (%)**		**Precision (%)**	
		ClassBalancer	**Resample**	**ClassBalancer**	**Resample**	**ClassBalancer**	**Resample**	**ClassBalancer**	**Resample**
DECISION STUMP	10	95.97	96.30	96.99	95.14	93.69	95.38	93.43	97.24
HOEFFDING TREE		96.75	98.00	96.37454	96.3303	95.77	98.70	95.53	98.30
J48		98.23	97.90	96.57387	96.7236	98.59	97.09	98.20	97.77
LMT		98.35	98.50	96.99	96.99	98.40	98.70	98.02	98.30
RANDOM		98.50	98.50	96.99	96.9	98.7	98.7	98.3	98.3

FOREST					9	0	0	0	0
RANDOM TREE		97.69	98.40	96.80 117	96.8 573	97.2 4	98.7 0	96.9 3	98.3 0
REPTree		97.00	97.90	96.99	96.4 602	95.6 7	97.8 8	95.4 2	98.0 3

TABLE VIII. PERFORMANCE OF TREE CLASSIFIERS for 10 folds



Graph 8: Tree classifiers performance over 20 reduced dataset for 10 folds

From the graph 5, 6, 7 and 8 two things can be observed one is Random Forest clssifier is showing better results regarding the overall performance and second is the performance is consistent irrespective of number of parameters used in dataset.

IV. CONCLUSION

We have selected features from the dataset and built classifier using Bayes net, naïve Bayes, naïve Bayes updateable from Bayes classifiers and decision stump, hoeffding tree, j48, LMT, random forest, random tree, RepTree classifiers using WEKA. Our proposed model in which Bayes Net classifier is used achieved highest accuracy in both 5 and 10 folds of cross validations among Bayesian classifiers and Random Forest performed better than other tree classifiers employed in the model in detection of leukemia. On comparing Bayes and tree classifiers as whole Random Forest outperformed remaining nine classifiers by achieving 98.50% accuracy, 96.99% sensitivity, 98.7% specificity and 98.30% precision values.

Also, the feature reduction showed improvement in the performance of the detection results. In future also we will work to enhance the performance of the model by using different classifiers and datasets. Also we will attempt to reduce the features to a minimum in order to focus on the most important features.

References:

1. A. Das Mou and P. Kumar Saha, "A Comprehensive Study of Machine Learning algorithms for Predicting Leukemia Based on Biomedical Data," 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), 2019, pp. 1-5, doi: 10.1109/ICIET48527.2019.9290544.
2. Vos, Theo, C. Allen, M. Arora, R.M. Barber,Z. M. Brown,A. Carter, A.Casey et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990– 2015: a systematic analysis for the Global Burden of Disease Study 2015." The Lancet 388, no. 10053.2016, pp. 1545-1602.
3. Wang, Haidong, M. Naghavi, A. Carter, R. M. Barber , Z. M. Brown, A. Casey et al. "Global, regional, and national life expectancy, allcause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015." The lancet 388, no. 10053. 2016, pp. 1459-1544.
4. Salah HT, Muhsen IN, Salama ME, Owaidah T, Hashmi SK. Machine learning applications in the diagnosis of leukemia: Current trends and future directions. Int J Lab Hematol. 2019 Dec;41(6):717-725. doi: 10.1111/ijlh.13089. Epub 2019 Sep 9. PMID: 31498973.
5. Abass, Shakir & Mohsin Abdulazeez, Adnan & Zeebaree, Diyar. (2021). A YOLO and convolutional neural network for the detection and classification of leukocytes in leukemia. Indonesian Journal of Electrical Engineering and Computer Science. 25. 10.11591/ijeecs.v25.i1.pp200-213.
6. "A.M. Patil, M.D. Patil, G.K. Birajdar, White Blood Cells Image Classification Using Deep Learning with Canonical Correlation Analysis, IRBM, Volume 42, Issue 5, 2021, Pages 378-389, ISSN 1959-0318, https://doi.org/10.1016/j.irbm.2020.08.005. (https://www.sciencedirect.com/science/article/pii/S195903182030141X)"
7. Alam MM, Islam MT. Machine learning approach of automatic identification and counting of blood cells. Healthc Technol Lett. 2019 Jul 17;6(4):103-108. doi: 10.1049/htl.2018.5098. PMID: 31531224; PMCID: PMC6718065.
8. "Anita, Anupam Yadav, An Intelligent Model for the Detection of White Blood Cells using Artificial Intelligence, Computer Methods and Programs in Biomedicine, Volume 199, 2021, 105893, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2020.105893. (https://www.sciencedirect.com/science/article/pii/S0169260720317260)"
9. Muhammad Shahzad, Arif Iqbal Umar, Muazzam A. Khan, Syed Hamad Shirazi, Zakir Khan, Waqas Yousaf, "Robust Method for Semantic Segmentation of Whole-Slide Blood Cell Microscopic Images", Computational and Mathematical Methods in Medicine, vol. 2020, Article ID 4015323, 13 pages, 2020. https://doi.org/10.1155/2020/4015323
10. A. Ratley, J. Minj and P. Patre, "Leukemia Disease Detection and Classification Using Machine Learning Approaches: A Review," 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), 2020, pp. 161-165, doi: 10.1109/ICPC2T48082.2020.9071471.