International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A SMOTE Boosting Based Blood Cancer Prediction Using Leukemia Microarray Dataset

M.V. Phanindra[1], Dr. G. Thippanna[2]

[1]Research Scholar, Dept. of Computer Science and Engineering, NIILM University, Kaithal, Haryana, India
[2]Supervisor, Dept. of Computer Science and Engineering, NIILM University, Kaithal, Haryana, India

Abstract:
Leukemia is a type of blood cancer primarily involving abnormal white blood cell production. This condition leads to an irregular increase in white blood cells compared to normal levels. Despite advancements, accurately classifying cancers using microarray data remains challenging. Many data mining techniques have struggled due to limited sample sizes, posing significant challenges for organizations. While frequently used in cancer diagnosis, these methods often fall short in delivering improved results. This research introduces an innovative approach utilizing ensemble learning algorithm to analyze microarray data from leukemia cells, aiming to predict early-stage leukemia. SMOTE (Synthetic Minority Over-sampling Technique) boosting is an advanced technique used to address class imbalance in predictive modeling, particularly in the context of medical data. SMOTE works by generating synthetic samples of the minority class by interpolating between existing instances. The integration of AdaBoost, SMOTE enhanced the model's ability to focus on misclassified instances, thus improving the overall prediction accuracy.

Keywords: SMOTE. Machine Learning, Blood Cancer, Leukemia Microarray

1 Introduction

Cancer is a broad term referring to a group of diseases characterized by the uncontrolled growth and division of abnormal cells. In some types of cancer, cells grow excessively, while in others, cells divide more rapidly than normal[1]. Some cancers lead to the formation of visible masses known as tumors, while others, like leukemia, do not. Hematologic cancers, which affect the blood and bone marrow, typically originate in the bone marrow, where blood cells are produced. Blood cancers occur when abnormal blood cells proliferate uncontrollably, disrupting the normal function of blood cells, impeding infection defense, and affecting the production of new blood cells. There are approximately a hundred different types of cancer, with the three primary categories of blood cancers being leukemia, myeloma, and lymphoma[2]. Leukemia, a type of blood and bone marrow cancer, originates in the tissue responsible for blood formation. Unlike solid tumors, leukemias involve an overproduction of dysfunctional white blood cells (leukemia cells and blasts) that overcrowd the blood and bone marrow, leaving insufficient space for healthy cells. This can lead to a shortage of red blood cells, making it harder for the body to deliver oxygen to tissues, control blood loss, and fight infections[3].

Machine learning (ML) algorithms can often struggle in the prediction of blood cancers data due to several challenges inherent in the data and disease characteristics. One of the primary issues is the imbalanced

nature of the dataset, where the number of samples for one class may be much smaller compared to the other (e.g., healthy individuals)[4]. This imbalance can lead to biased models that fail to detect the minority class, resulting in poor performance. Additionally, microarray data is high-dimensional, meaning that the number of features (gene expressions) can be significantly larger than the number of samples, which can cause overfitting and hinder the model's generalization ability. Moreover, the complexity of leukemia's genetic landscape, with numerous factors influencing its development, makes it difficult for traditional ML algorithms to accurately capture the patterns required for prediction. Boosting with SMOTE can significantly improve the prediction of leukemia by addressing these challenges. Boosting techniques like AdaBoost enhance weak models by focusing on difficult-to-predict instances, iteratively refining the model's predictions. When combined with SMOTE, which generates synthetic examples of the minority class to balance the dataset, these techniques become much more effective. SMOTE helps mitigate the class imbalance problem by generating artificial instances of the underrepresented class, allowing the model to learn more robust patterns for prediction. As a result, the combined approach improves model performance, particularly in detecting rare occurrences of leukemia, leading to more accurate and reliable predictions.
The structure of this article is organized as follows: Section II reviews the work of various researchers who have utilized the IoMT dataset. Section III provides an overview of the proposed AdaBoost classifier and SMOTE algorithm. Section IV details the dataset and experimental setup used in this study. Section V presents the results of the experimental validation for the Leukemia Microarray dataset. Finally, Section VI concludes the article.

2 Literature review

[5] presents an innovative approach that utilizes machine learning algorithms applied to leukemia microarray data from GSE9476 cells. The primary goal is to predict the onset of leukemia. Various machine learning techniques, including decision tree (DT), naive Bayes (NB), random forest (RF), gradient boosting machine (GBM), linear regression (LinR), and support vector machine (SVM), are employed. Additionally, a novel ensemble model combining Logistic Regression (LR), DT, and SVM, referred to as the ensemble LDSVM model, is proposed. The study employs k-fold cross-validation and grid search optimization to classify leukemia in patients with an accuracy of 99%.
[6] introduces a supervised ML approach for predicting blood cancer, utilizing a leukemia microarray dataset containing 22,283 gene expressions. To address challenges posed by imbalanced and high-dimensional datasets, the study employs Chi-squared (Chi^2) feature selection and the SMOTE-Tomek resampling technique. SMOTE-Tomek generates synthetic data to balance the dataset across target classes, while Chi^2 identifies the most relevant features for training the models from the 22,283 genes. Additionally, a novel weighted convolutional neural network (CNN) model is proposed for classification, leveraging the combined power of three distinct CNN models and derived an accuracy of 99.9%.
[7] presents a method for predicting blood cancer using a supervised ML approach. The research utilizes a leukemia microarray dataset consisting of 22,283 genes. To address issues related to imbalanced and high-dimensional datasets, the study employs ADASYN resampling and Chi^2 feature selection techniques. ADASYN generates synthetic data to balance the dataset for each target class, while Chi^2 selects the most relevant features from the 22,283 genes to train the models. For classification, a hybrid Logistic Vector Trees (LVTrees) classifier is proposed, combining LR, SVM, ETC. Extensive experiments were conducted on the dataset, and the proposed approach was compared with state-of-the-art methods. The LVTrees

model, using ADASYN and Chi^2 techniques, achieved a remarkable 100% accuracy, outperforming all other models.

[8] introduces a multi-population particle swarm optimization (MPSO) approach for feature selection to identify the most important gene subsets for classifying ML algorithms. In this study, MPSO is applied to enhance the search diversity of the traditional particle swarm optimization (PSO) method. It is integrated with the SVM classifier to create a wrapper-based feature selection model that captures the interactions between the features and the classifier. The proposed model is assessed using 10-fold cross-validation, and the accuracy 80.6% demonstrate that MPSO provides more consistent classification performance compared to conventional PSO for all ML algorithms.

3 Methodology

AdaBoost (Adaptive Boosting) is an ensemble learning technique that aims to improve the accuracy of weak classifiers by combining them into a strong classifier. The working principle of AdaBoost involves training a series of weak models, typically decision trees, in a sequential manner. Initially, each training instance is given equal weight, but as each model is trained, AdaBoost adjusts the weights of the misclassified instances, making them more important for the next model. This means that subsequent classifiers focus more on the instances that were misclassified by previous ones. After each model is trained, AdaBoost assigns a weight to it based on its accuracy, with more accurate models receiving higher weights[9]. The final prediction is made by combining the predictions of all the weak classifiers, with each classifier's prediction weighted according to its performance. This iterative process helps to reduce bias and variance, resulting in a highly accurate and robust model that performs well on complex datasets, even when individual models might not.

The model weighting of each weak classifier is assigned a weight based on its accuracy. Classifiers that perform better (i.e., make fewer errors) are given higher weights, whereas classifiers with worse performance receive lower weights. The weight of a classifier is computed based on its error rate Eq. (1). Where α_t is the weight of classifier t, and ϵ_t is the error rate of classifier t. This means that a stronger model will have more influence in the final prediction. Once a specified number of classifiers have been trained, AdaBoost makes its final prediction by combining the predictions of all the weak classifiers in the ensemble. Each classifier's vote is weighted according to its accuracy, meaning more accurate classifiers have a larger influence on the final decision. The final prediction is typically determined by a weighted majority vote Eq. (2). Where $h_t(x)$ is the prediction of classifier t for input x, and T is the total number of classifiers in the ensemble. This process is repeated for a predefined number of iterations, each time refining the model to focus on the errors made by previous classifiers. By the end of the boosting process, AdaBoost creates a strong classifier that performs better than any individual weak learner could.

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \qquad \text{Eq. (1)}$$

$$\text{Final Prediction} = \text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right) \qquad \text{Eq. (2)}$$

SMOTE is a popular method for addressing the problem of imbalanced datasets, particularly in classification tasks where one class significantly outnumbers another. In such cases, standard ML models may become biased toward the majority class, leading to poor predictive performance for the minority class[10]. SMOTE works by generating synthetic samples for the minority class rather than duplicating

existing ones. It creates new instances by interpolating between a minority class instance and its nearest neighbours in feature space. This approach maintains the diversity of the dataset while effectively balancing the class distribution. By incorporating these synthetic samples into the training set, SMOTE helps machine learning algorithms better learn the characteristics of the minority class, leading to improved generalization and enhanced model performance for imbalanced datasets. The algorithm 1 gives a brief working principle of SMOTE[11].

Algorithm 1: SMOTE
Input : Minority data $(D) = \{ x_i \in X\}$ *where* $i = 1,2, \dots, T.$
T *number of minority instances,* N *SMOTE percentage*
k *is number of nearest neighbours*
Output: Return synthetic data ξ
For $i = 1,2, \dots, T$ do
Find the k nearest(minority class) neighbors of x_i
$\widehat{N} = N/100$
While $\widehat{N} \neq 0$ do
Select one of the k nearest neighbor $\bar{x}$
Select a rando number $\alpha \in [0,1]$
$\hat{x} = x_i + \alpha\ (\bar{x} - x_i)$
Append $\hat{x}$ to ξ
$\widehat{N} = \widehat{N} - 1$

4 Datasets Description and Experiment Setup

4.1 Dataset

The characteristics of the microarray dataset of leukemia cancer target class is given in Figure 1. The datasets is in csv format with 3573 coloums, which is a pre-processed and derived through 3051 gene expression with 72 samples. The leukemia cancer dataset also known colon cancer microarray dataset was originally analysed by. Figure 1 depicts the distribution of the target classes in the dataset. The data is split into 80% for training and 20% for testing, with 50 samples allocated to training and 22 samples set aside for testing. Figures 2 and 3 represent the training and validation samples, visualized using kernel PCA and their respective class distributions.

4.2 Experiment Setup

The evaluation was conducted using the Python Google Colab environment, utilizing the Scikit-learn library as the framework for implementation and graph plotting.

5 Results and Discussion

In this study, the SMOTE with Boosting approach, using a decision tree as the base estimator, showed outstanding performance. As illustrated in Figures 4-5, the model achieved impressive results in accuracy and F1-score during training with base estimators and samples. The confusion matrix in Figure 6 shows flawless performance, with no misclassifications. The absence of false positives and false negatives suggests that the model is successfully differentiating between classes, with near-perfect classification of

normal and tumor cases.

The Table 1 shows the performance of the classification model for detecting blood cancer is outstanding, with perfect results across all key evaluation metrics. The model demonstrates flawless prediction capabilities, achieving a Recall of 1.0 for both normal and tumor categories. The model achieves **perfect Precision** of 1.0 for both categories. This means that every time the model predicts an instance to be a tumor or normal, it is correct. High precision is critical in medical diagnostics, as it ensures that the model is not wrongly flagging healthy individuals as having cancer, preventing unnecessary interventions. The F1 Score of 1.0 for both classes reflect a perfect balance between precision and recall. The AUC of 1.0 further underscores the model's exceptional ability to discriminate between normal and tumor instances. This suggests that the model's performance would remain robust across different thresholds for decision-making, making it adaptable for various clinical scenarios.

Overall, the model demonstrates unparalleled reliability and accuracy, making it highly suitable for use in critical medical applications in detection of blood cancer. With no false positives or negatives, perfect precision and recall, and an AUC of 1.0, the model's performance is ideal for real-world clinical settings, ensuring that blood cancer detection is both precise and reliable. This kind of model could significantly enhance diagnostic accuracy and support healthcare professionals in making timely and informed decisions for patient care.

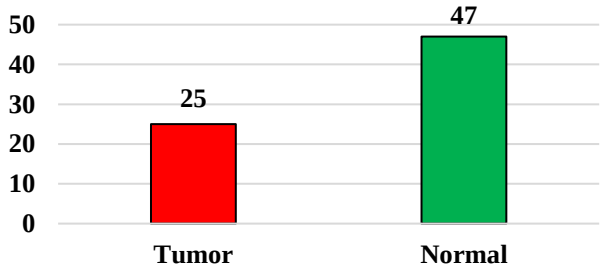

Figure 1: Class label distribution Microarray dataset

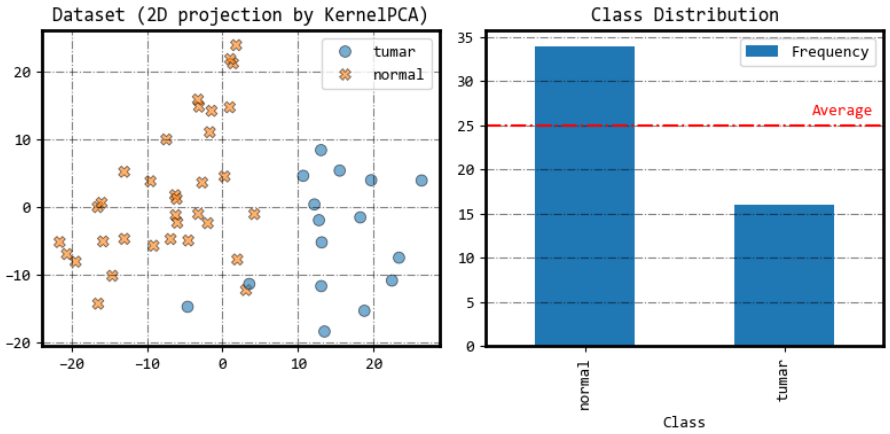

Figure 2: Projection of training dataset

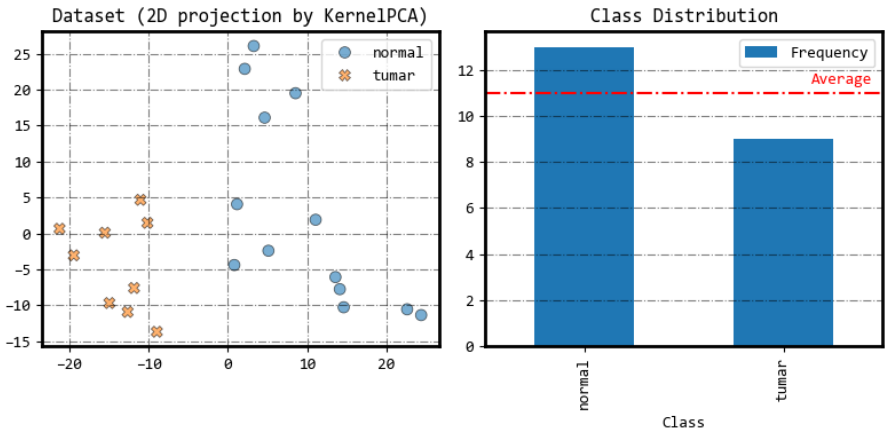

Figure 3: Projection of validation dataset

Table 1: Performance metrics (%)

Labels	Precision	Recall	F1-score	AUC	Accuracy
Normal	1.00	1.00	1.00	1.00	100
Tumor	1.00	1.00	1.00	1.00	

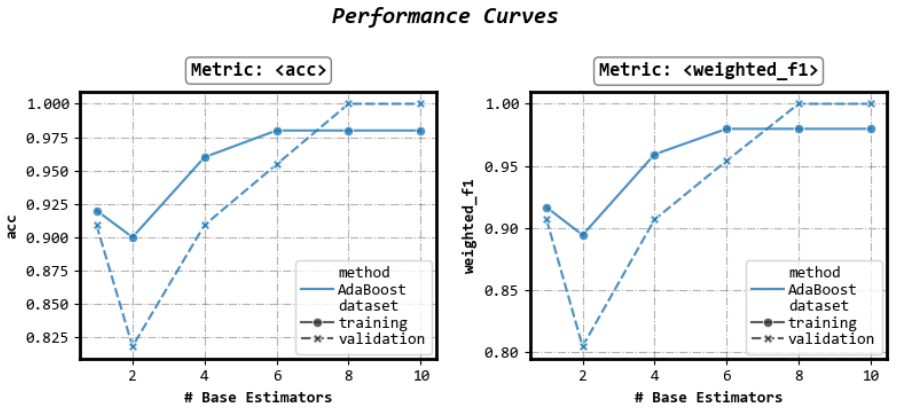

Figure 4: Performance curves with base estimators

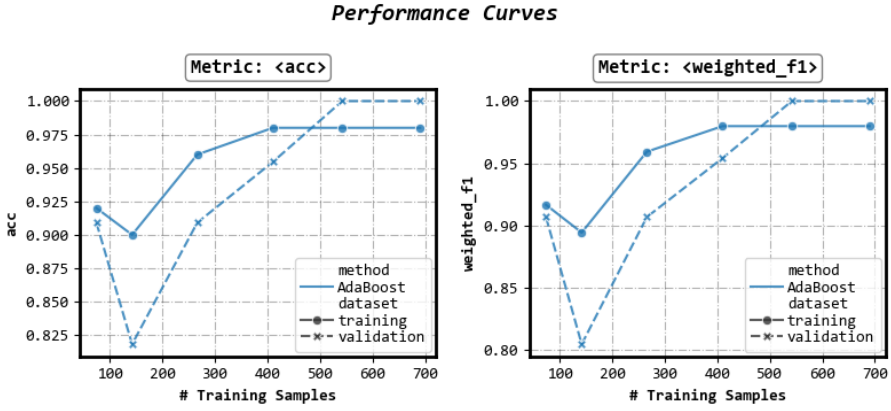

Figure 5: Performance curves with training samples

6 Conclusion

This study explores and introduces the combination of SMOTE with the AdaBoost classifier. The blood cancer prediction model has demonstrated outstanding performance, with perfect results across all key evaluation metrics, including precision, recall, and F1-score, all scoring 1.00 for both normal and tumor classes. The confusion matrix reinforces these results, showing that all 22 samples (13 normal and 9 tumor) were correctly identified, with no errors in classification. The overall accuracy of 1.0 further highlights the model's flawless predictive capability. These results suggest that the model is highly reliable in distinguishing between normal and tumor cases, making it a robust tool for blood cancer detection. This makes the model an ideal candidate for clinical applications, where accurate and reliable detection of blood cancer is crucial.

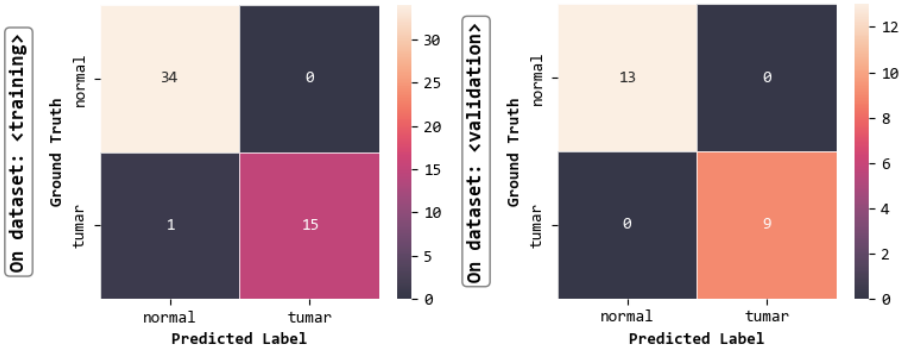

Figure 6: Confusion matric

References

1. "International Agency for Research on Cancer, World Health Organization." [Online]. Available: https://gco.iarc.fr/today/home
2. M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A. Roshanpour, "Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images: A Systematic Review," 2021, *Hindawi Limited*. doi: 10.1155/2021/9933481.
3. A. H. Chen, Y.-W. Tsau, and C.-H. Lin, "Novel methods to identify biologically relevant genes for leukemia and prostate cancer from gene expression profiles," 2010. [Online]. Available: http://www.biomedcentral.com/1471-2164/11/274
4. D. K. K. Reddy, H. Swapnarekha, H. S. Behera, S. Vimal, A. K. Das, and D. Pelusi, "Issues and future challenges in cancer prognosis: (Prostate cancer: A case study)," in *Computational Intelligence in Cancer Diagnosis: Progress and Challenges*, Elsevier, 2022, pp. 337–358. doi: 10.1016/B978-0-323-85240-1.00001-8.
5. A. Karim, A. Azhari, M. Shahroz, S. B. Belhaouri, and K. Mustofa, "LDSVM: Leukemia cancer classification using machine learning," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3887–3903, 2022, doi: 10.32604/cmc.2022.021218.
6. E. A. Alabdulqader *et al.*, "Improving prediction of blood cancer using leukemia microarray gene data and Chi2 features with weighted convolutional neural network," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-65315-7.
7. V. Rupapara, F. Rustam, W. Aljedaani, H. F. Shahzad, E. Lee, and I. Ashraf, "Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-04835-6.

8. W. Son Ng, S. Chin Neoh, K. Kyaw Htike, and S. Li Wang, "Progress in Energy and Environment Particle Swarm Feature Selection for Microarray Leukemia Classification," *Progress in Energy and Environment*, vol. 2, pp. 1–8, 2017.
9. D. K. K. Reddy, H. S. Behera, G. M. S. Pratyusha, and R. Karri, "Ensemble Bagging Approach for IoT Sensor Based Anomaly Detection," in *Lecture Notes in Electrical Engineering*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 647–665. doi: 10.1007/978-981-15-8439-8_52.
10. S. K. Pemmada, K. S. Naidu, and D. K. K. Reddy, "SMOTE Integrated Adaptive Boosting Framework for Network Intrusion Detection," in *Intelligent Systems Reference Library*, vol. 60, Springer Science and Business Media Deutschland GmbH, 2024, pp. 1–25. doi: 10.1007/978-3-031-54038-7_1.
11. D. K. K. Reddy, B. K. Rao, and T. A. Rashid, "Intelligent Under Sampling Based Ensemble Techniques for Cyber-Physical Systems in Smart Cities," 2024, pp. 219–244. doi: 10.1007/978-3-031-54038-7_8.