Improving Leukemia Diagnosis: A Feature Selection-Driven BPNN Approach

Abinash Rath
Department of Computer Science
C. V. Raman Global University
Bhubaneswar, India
abinashrath610@gmail.com

Debadatta Rout
Department of Computer Science
C. V. Raman Global University
Bhubaneswar, India
routdebadatta22@gmail.com

Abinash Baliarsingh
Department of Computer Science
C. V. Raman Global University
Bhubaneswar, India
abinash7735251881@gmail.com

Amit Ranjan Bastia
Department of Computer Science
C. V. Raman Global University
Bhubaneswar, India
125amitbastia@gmail.com

Ghansyama Mohanty
Department of Computer Science
C. V. Raman Global University
Bhubaneswar, India
mohantyghansyama2005@gmail.com

Debendra Muduli
Department of Computer Science
C. V. Raman Global University
Bhubaneswar, India
debendra.muduli@cgu-odisha.ac.in

***Abstract*—Leukemia is a malignant disorder of the blood and bone marrow, posing a significant global health challenge. Early and precise classification of leukemia is essential for timely intervention and personalized treatment planning. This study introduces an advanced leukemia classification framework that systematically incorporates dataset balancing, feature selection, feature ranking, and machine learning to enhance predictive accuracy. During preprocessing, the Synthetic Minority Oversampling Technique (SMOTE) is employed to address class imbalance and improve model performance. Subsequently, feature selection techniques, including ANOVA, Information Gain, and Correlation Analysis, are applied to identify the most relevant biomarkers. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is then utilized to rank these selected features, ensuring that the most significant attributes contribute to the classification process. During classification, uses the back propagation neural network (BPNN) as the primary clssifier. Performance evaluation is conducted using various activation functions, where ReLU achieves a remarkable classification accuracy of 95.45% on the leukemia dataset. The results underscore the effectiveness of a structured machine learning pipeline, demonstrating the potential of feature selection-driven approaches in improving leukemia diagnosis and optimizing treatment strategies.**

***Index Terms*—Leukemia Prediction, Feature Selection, BPNN, TOPSIS**

I. Introduction

Leukemia, a malignant disease affecting blood-forming tissues, remains a significant global health challenge due to its diverse subtypes and complex progression patterns [1].It begins in the bone marrow, causing an unchecked growth of abnormal white blood cells, which disrupts normal blood cell production and weakens the immune system. The disease is categorized into acute and chronic forms, with acute leukemia progressing rapidly and requiring immediate intervention, while chronic leukemia develops more slowly. Despite advancements in treatment, including chemotherapy, targeted therapies, and immunotherapy, leukemia remains difficult to manage due to its heterogeneity and the unpredictable nature of treatment responses [2]. Given these challenges, early detection and precise prognosis are critical to improving patient outcomes.

Machine learning technique facilitate the early detection of diseases such as breast cancer [15]–[18], brain tumor detection, leukemia [3]–[14], glaucoma detection [20]–[22], and diabetes [19] by analyzing medical imaging and patient records with high accuracy.

Recent advances in computational technologies, particularly machine learning (ML), have shown great potential in enhancing leukemia diagnosis, prognosis, and treatment optimization. ML models can process extensive datasets, including clinical [5], genomic, and proteomic information, to uncover complex patterns beyond human analytical capabilities. These predictive models assist in stratifying patients into risk groups, optimizing treatment strategies, and forecasting survival rates and recurrence risks. Techniques such as decision trees, Support vector machines (SVMs), Convolutional neural networks (CNNs), and deep learning have demonstrated notable success in leukemia classification, treatment outcome prediction, and biomarker discovery.

Several studies have explored the application of ML in leukemia prognosis and treatment response prediction. Nguyen et al. [6] effectively applied CNNs to classify acute lymphoblastic leukemia (ALL) subtypes using histopathological images, while Mukherjee et al. [7] reported novel biomarkers and therapies for leukemia. Ahmed et al. [9] developed a deep learning-based leukemia detection model using microscopic blood smear images. These findings underscore the transformative role of ML in hematological malignancies, offering non-invasive, rapid, and reliable diagnostic solutions.

Recent studies have explored various methodologies for improving classification accuracy in complex datasets. Maiapalli & Athavale (2021) a hybrid approach integrating Convolutional Neural Networks (CNN) with VGG, ResNet, and Inception models, yielding competitive results [11]. Manescu et al. (2022) achieved notable performance using the MILLI

method [12], while Raina et al. (2022) demonstrated the efficacy of a deep learning model, performance of the proposed approaches in this research was evaluated and compared to ensure alignment with expected outcomes [13]. Ahmed et al. (2022) utilized CNNs to achieve strong results [9], whereas Mondal et al. (2021) employed a weighted ensemble of CNNs [14], reporting comparatively lower performance. The proposed method in this research combines ANOVA, Information Gain (IG), Correlation, and TOPSIS, achieving results on par with the highest-performing existing methods, underscoring its effectiveness.

The literature review indicates that the workflow of the proposed method begins with feature selection, where the dataset undergoes preprocessing, including the application of SMOTE to address class imbalance. Feature selection is conducted using ANOVA, Information Gain, and Correlation to identify the most relevant features . These features are then ranked and prioritized using the TOPSIS method, which evaluates their contribution to model performance. Finally, a Backpropagation Neural Network (BPNN) is employed for classification, specifically targeting the differentiation of cell types such as PCELL, TCELL, and BCELL. This integrated approach, combining robust feature selection techniques with TOPSIS-based ranking, enhances the model's ability to capture discriminative features, resulting in improved classification performance. The method's comprehensive design ensures effective feature extraction and classification, demonstrating its superiority over existing approaches.

This subsequent part of document has structured the following manner. Section II outlines the proposed methodology describes the dataset and feature selection methods. Section III presents results and discussions. Finally, Section IV summarizes key findings, emphasizes on future work.

II. MATERIAL AND METHODS

A. Proposed Methodlogy

The proposed methodology in this study focuses on developing an optimized feature selection-driven machine learning model for leukemia outcome prediction. Figure 1 presents the architectural diagram of the implemented scheme. The implemented model integrates multiple techniques to enhance classification accuracy and robustness.In step 1, the leukemia dataset is preprocessed to handle missing values and class imbalances using imputation methods and the Synthetic Minority Oversampling Technique (SMOTE). Feature normalization is applied to standardize numerical variables.In step 2, feature selection techniques, including ANOVA, Information Gain, and Correlation Analysis, are applied to extract the most relevant biomarkers. The selected features are ranked using the TOPSIS technique to ensure optimal input selection for classification.In step 3, the refined dataset is used to train various classifiers, including Support Vector Machine, Naïve Bayes, Random Forest, and the Back Propagation Neural Network, with activation functions such as Rectified Linear Unit (ReLU) to enhance learning efficiency.In step 4, the trained models are evaluated using test data, and performance

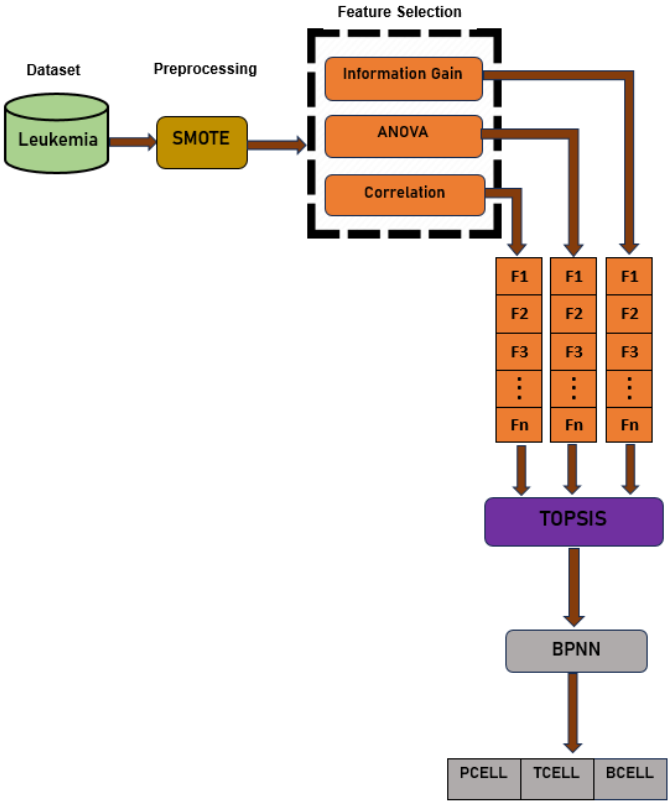

Fig. 1. Block diagram of our purposed model

metrics such as accuracy, recall and precision are analyzed .Finally, in step 5, the classification results are compared with previous research to assess improvements in leukemia prediction, demonstrating the effectiveness of the proposed method.

B. Dataset

The study utilizes the leukemia.csv dataset [6] , containing 7,130 genomic features from microarray experiments on 72 leukemia patients, with labeled subtypes in the "CLASS" column for supervised learning. Given its high dimensionality and small sample size, feature selection techniques were applied to identify key biomarkers for leukemia classification [9]. Data preprocessing addressed missing values and class imbalances, ensuring a high-quality dataset for effective model training and analysis.

C. Data Pre-Processing

The data pre-processing phase was essential to address issues inherent in healthcare datasets, such as missing values and imbalances.Missing data were addressed using imputation methods, where numerical variables were filled with their mean or median values, while categorical variables were replaced with their mode. Continuous variables like age and WBC count were normalized to standardize their scales. To address class imbalances, particularly in rare genetic mutations, the Synthetic Minority Oversampling Technique (SMOTE)

was employed. This ensured equitable representation of all classes in the training process. Finally, the dataset was divided into training, testing subsets in a 70:30 ratio, enabling robust model evaluation.

D. Feature Selection

The feature selection phase was conducted to identify and prioritize the most relevant predictors for leukemia outcomes [11]. The following methods were employed and shown in Fig. 2:

1) Anova: Statistical significance of features was evaluated to understand their impact on the target variable using Eq. (1).

The Analysis of Variance (ANOVA) coefficient f is given by:

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} \tag{1}$$

where:

- F represents the ANOVA coefficient, which is used to test the hypothesis in analysis of variance.
- $\text{MS}_{\text{between}}$ denotes the Mean square between groups, measuring the variance between different groups.
- $\text{MS}_{\text{within}}$ signifies the Mean square within groups, quantifying the variance within each group.

2) Information Gain: Each feature's contribution to reducing uncertainty was assessed to determine its relevance using Eq. (2).

The Information Gain for a feature A with respect to the target variable Y is given by:

$$\text{Information Gain}(Y, A) = Entropy(Y) - Entropy(Y|A) \tag{2}$$

where:

- $Entropy(Y)$ represents the entropy of the target variable Y, which measures the uncertainty or randomness in Y.
- $Entropy(Y|A)$ denotes the conditional entropy of Y given the feature A, quantifying the remaining uncertainty in Y after knowing A.

3) Correlation: Features with low correlation to the target variable or high multicollinearity were excluded using Eq.(3).

The Pearson Correlation Coefficient (r) between two variables X and Y is defined as:

$$r_{XY} = \frac{\text{Covariance}(X, Y)}{\sigma_X \sigma_Y} \tag{3}$$

where:

- $\text{Covariance}(X, Y)$ represents the covariance between X and Y, which is computed as:

$$\text{Covariance}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right) \tag{4}$$

- σ_X and σ_Y denote the standard deviations of X and Y, respectively, calculated using:

$$\sigma_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \tag{5}$$

- $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y, respectively, given by:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{6}$$

The results of these methods were consolidated using the TOPSIS technique, enabling a comprehensive ranking of features shown in Fig. 2. This hybrid approach combined the strengths of all individual feature selection methods, ensuring that only the most significant features were retained for classification.

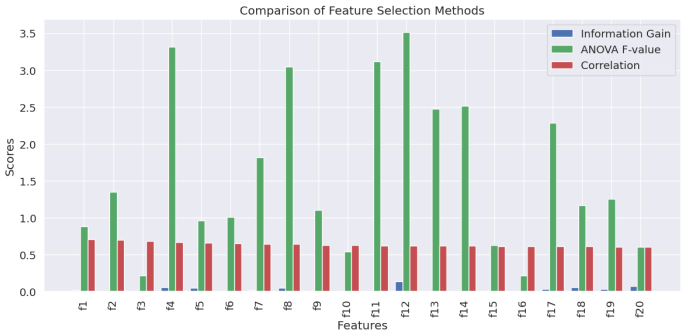

Fig. 2. Comparison of feature selection methods

E. Classification

The selected features were used to train various classifiers, as depicted in Table 1.The models evaluated in this study encompassed Logistic Regression, Naïve Bayes, K-Nearest Neighbors , Support Vector Machine , Decision Tree, Random Forest, AdaBoost combined with Random Forest, AdaBoost paired with SVM, XGBoost, Backpropagation Neural Networks, Artificial Neural Networks, and a Voting classifier.. The accuracy and performance of each classifier were evaluated to identify the optimal predictive model for leukemia outcomes. .

The performace Evaluation metrics for all the classifiers is given by using accuracy,precision,recall,specificity,f1-score and MCC.

III. RESULT AND DISCUSSION

The performance evaluation of various classifiers after applying the TOPSIS method highlights their effectiveness in predicting leukemia outcomes using high-dimensional genomic data. The Back Propagation Neural Network (BPNN) [9] achieved the highest accuracy of 95.45%, demonstrating the superior ability of neural network-based models to capture intricate patterns within the data. Similarly, the Naïve Bayes classifier also attained an accuracy of 95.45%, showcasing

its strength in probabilistic classification tasks, particularly when handling imbalanced datasets. Other classifiers, including Artificial Neural Network (ANN), Logistic Regression, K-Nearest Neighbor (KNN), Random Forest, SVM, and the Voting Classifier, consistently achieved an accuracy of 90.90%, indicating their robustness and reliability in handling genomic datasets. However, the Decision Tree classifier recorded a lower accuracy of 86.36%, which could be attributed to its limited capability to manage high-dimensional and complex feature spaces effectively. The XGBoost model, despite its popularity for achieving high performance in diverse applications, obtained the lowest accuracy of 81.81% is shown in Table.1, likely due to challenges in hyperparameter tuning or overfitting in this specific context.

TABLE I
PERFORMANCE EVALUATION METRICS WITH TOPSIS

Classifiers	Accuracy (in %)
Logistic Regression	86.36
SVM	90.90
Naïve Bayes	95.45
Random Forest	90.90
Adaboost + SVM	90.90
K-Nearest Neighbour	90.90
XGBoost	81.81
ANN	90.90
Decision Tree	86.36
BPNN	**95.45**
Voting Classifier	90.90

Comparative analysis of leukemia detection models highlights the superior performance of the proposed methodology, which achieved a baseline precision of 95.45%, surpassing several state-of-the-art studies, including Matapalli & Athavale [11] and Mondal et al. [14], which reported significantly lower accuracies. While some studies, such as Raina et al. [13] and Ahmed et al. [9]and Manescu et al. [12] demonstrated performance closer to this baseline, the consistent underperformance in other works underscores the challenges of high-dimensional genomic datasets and model optimization Table. 2. By integrating preprocessing, feature selection (IG+ANOVA+Correlation) through the TOPSIS method, and robust classifiers like BPNN .This method successfully tackles these issues, establishing a new standard for leukemia prediction and highlighting the critical role of sophisticated data preprocessing and feature selection in enhancing accuracy and generalizability.

TABLE II
PERFORMANCE COMPARSION WITH EXISTING MODELS

Author	Methods	Accuracy
Matapalli & Athavale (2021)	Hybrid CNN +VGG + ResNet+Inception models	92.10%
Manescu et al. (2022)	MILLI	94.00%
Raina et al. (2022)	Deep Learning Model	95.45%
Ahmed et al. (2022)	CNN	94.94%
Mondal et al. (2021)	Weighted Ensemble of CNNs	86.20%
Proposed Method	**IG+ ANOVA + Correlation +TOPSIS+ BPNN**	**95.45%**

We presented the confusion matrix for the best-performing model, the Back Propagation Neural Network (BPNN), using the test dataset, as illustrated in Fig. 3. The outcomes reveal that our proposed BPNN model achieves high accuracy in classifying leukemia cases.

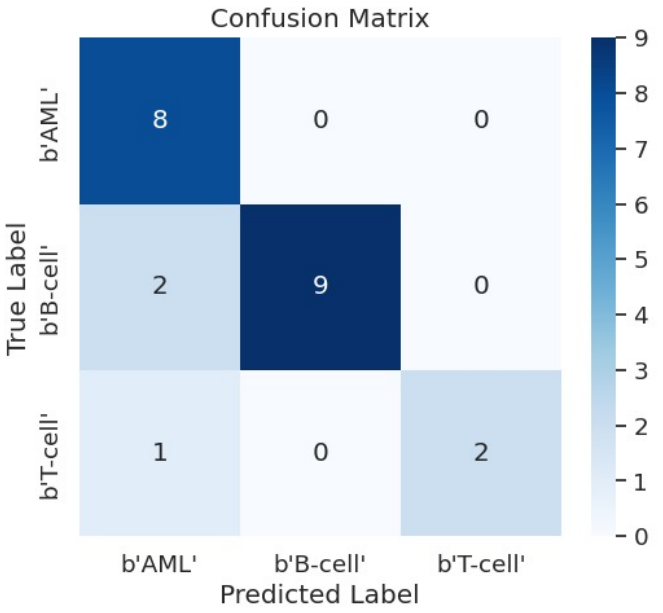

Fig. 3. Confusion matrix obtained for BPNN model

The ROC curves for the BPNN model are depicted in Figure 4, providing a visual representation of the binary classifier's performance. The curves are generated by plotting the true positive rate (TPR) against the false positive rate (FPR) to illustrate performance comparisons.

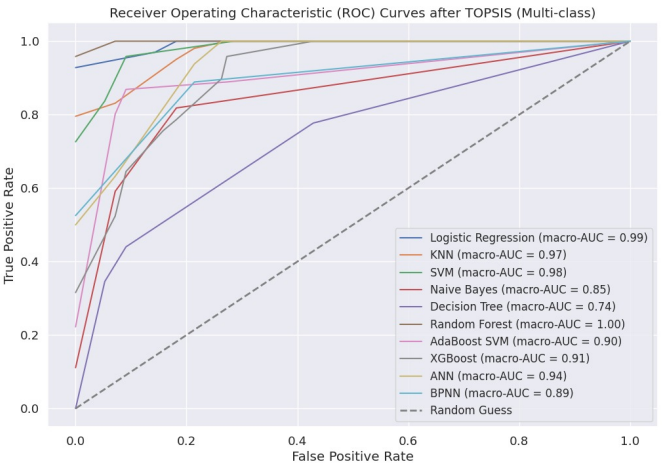

Fig. 4. ROC curve for all purposed model

IV. CONCLUSION

his study presents an optimized leukemia classification framework integrating SMOTE for dataset balancing, advanced feature selection methods, TOPSIS for feature ranking, and BPNN for classification. With ReLU activation, the

proposed model achieved a high classification accuracy of 95.45%, demonstrating its effectiveness in leukemia diagnosis. Future work will explore deep learning models, alternative feature selection techniques, and multi-modal data integration to enhance classification performance and clinical applicability. Expanding the dataset and developing an automated decision-support system will further improve robustness and real-world implementation, contributing to more accurate and timely leukemia diagnosis.

REFERENCES

[1] Ghiuzeli, C., Mawad, R., & Percival, M.-E. (2024). "Genetic testing helps predict leukemia prognosis." *Department of Medicine News, University of Washington*.

[2] Ugandhar Chapalamadugu et al. / *Asian Journal of Research in Pharmaceutical Sciences and Biotechnology*. 3(1), 2015, 12-26.

[3] Pokharel et al. "Leukemia - A review article," *International Journal of Advanced Research in Pharmaceutical and Biosciences*, 2(3), 2012, pp. 397-407.

[4] Lubomir Sokol et al. "Large Granular Lymphocyte Leukemia," *The Oncologist*, Jan 3, 2000.

[5] Goldstone AH et al. "In adults with standard-risk acute lymphoblastic leukemia, the greatest benefit is achieved from a matched sibling allogeneic transplantation in first complete remission," *Blood*, 111(4), 2008, pp. 1827-33.

[6] Nguyen, T. T. P., et al. (2019). "Detection of acute lymphoblastic leukemia and its subtypes using deep convolutional neural networks." *Scientific Reports*, 9, 4927. Demonstrates a successful application of CNNs for ALL subtype classification using histopathological images.

[7] Mukherjee, S., et al. (2021). "Novel biomarkers and therapies in leukemia: Progress and challenges." *Trends in Molecular Medicine*, 27(5), 383-400.

[8] He, K., et al. (2021). "Explainable artificial intelligence for survival prediction of AML patients treated with hypomethylating agents." *Frontiers in Oncology*, 11, 639660. Focuses on the use of explainable AI techniques to predict treatment outcomes for AML patients.

[9] Abdelmageed Ahmed, Alaa Nagy, Ahmed Kamal, Daila Farghl (2022). "Leukemia detection based on microscopic blood smear images using deep learning."

[10] Wahidul Hasan Abir, Md. Fahim Uddin, Faria Rahman Khanam, Mohammad Monirujjaman Khan (2023). "Explainable AI in Diagnosing and Anticipating Leukemia Using Transfer Learning Method."

[11] Sai Mattapalli, Rishi Athavale (2021). "ALLNet A Hybrid Convolutional Neural Network to Improve Diagnosis of Acute Lymphocytic Leukemia ALL in White Blood Cells."

[12] Petru Manescu et al (2022). "Automated Detection of Acute Promyelocytic Leukemia in Blood Films and Bone Marrow Aspirates with Annotation-free Deep Learning."

[13] Rohini Raina et al (2022). "A Systematic Review on Acute Leukemia Detection Using Deep Learning Techniques."

[14] Mondal et al. (2021). "Acute Lymphoblastic Leukemia Detection from Microscopic Images Using Weighted Ensemble of Convolutional Neural Networks."

[15] Muduli, Debendra, Ratnakar Dash, and Banshidhar Majhi. "Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach." *Biomedical Signal Processing and Control*, 71 (2022): 102825.

[16] Muduli, Debendra, Ratnakar Dash, and Banshidhar Majhi. "Automated breast cancer detection in digital mammograms: A moth flame optimization-based ELM approach." *Biomedical Signal Processing and Control*, 59 (2020): 101912.

[17] Muduli, Debendra, Ratnakar Dash, and Banshidhar Majhi. "Fast discrete curvelet transform and modified PSO based improved evolutionary extreme learning machine for breast cancer detection." *Biomedical Signal Processing and Control*, 70 (2021): 102919.

[18] Muduli, Debendra, et al. "An empirical evaluation of extreme learning machine uncertainty quantification for automated breast cancer detection." *Neural Computing and Applications* (2023): 1-16.

[19] Sharma, Santosh Kumar, et al. "A Diabetes Monitoring System and Health-Medical Service Composition Model in Cloud Environment." *IEEE Access*, 11 (2023): 32804-32819.

[20] Sharma, Santosh Kumar, et al. "An evolutionary supply chain management service model based on deep learning features for automated glaucoma detection using fundus images." *Engineering Applications of Artificial Intelligence*, 128 (2024): 107449.

[21] Muduli, Debendra, et al. "Retinal imaging based glaucoma detection using modified pelican optimization based extreme learning machine." *Scientific Reports*, 14.1 (2024): 29660.

[22] Sharma, Santosh Kumar, et al. "Discrete ripplet-II transform feature extraction and metaheuristic-optimized feature selection for enhanced glaucoma detection in fundus images using least square-support vector machine." *Multimedia Tools and Applications* (2024): 1-33.