Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

Machine Learning-Based Comparative Analysis for Automated Blood Cancer Diagnosis and Classification

Kasturi Chavan
kasturichavan.22@stvincentngp.edu.in
Robi Manukonda
robimanukonda.22@stvincentngp.edu.in
Tanushree Joshi
tanushreejoshi.22@stvincentngp.edu.in
Atharvi Babhare
atharvibabhare.22@stvincentngp.edu.in
Computer Science
St. Vincent Pallotti College of Engineering and Technology, Nagpur
Omesh Wadhwani
Asst. Professor,
Department of Computer Science

Abstract

This paper addresses the growing challenge of accurately diagnosing blood cancer using artificial intelligence. By applying supervised models like CNN and Random Forest, and unsupervised models such as K-Means Clustering, we compare their performance on a curated medical imaging dataset. Data augmentation methods like rotation and scaling were applied to enhance model generalization. Among all, CNN with augmentation achieved 95% accuracy. This comparative study suggests machine learning's practicality in supporting medical diagnostics. Blood cancer, particularly leukemia, remains one of the leading causes of cancer-related deaths globally due to its complex nature and the challenges in early diagnosis. Traditional diagnostic methods, which rely heavily on manual examination of blood smears, are time-consuming, prone to human error, and require expert interpretation. With the growing capabilities of artificial intelligence, especially in medical imaging, machine learning (ML) presents a promising solution to automate and improve the accuracy of blood cancer diagnosis. This research conducts a comprehensive comparative study of both supervised and unsupervised machine learning models for the classification of leukemic and healthy cells using the ALL-IDB image dataset. Models such as Support Vector Machines, Random Forests, and deep learning architectures like ResNetRS50 and EfficientNetB3 are implemented and evaluated. Unsupervised models like k-means and autoencoders are

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

also explored for their utility in anomaly detection. Furthermore, data augmentation techniques, including SMOTE and geometric transformations, are applied to address data imbalance and enhance model performance. The inclusion of explainable AI techniques like LIME and SHAP provides transparency to model predictions, making them more interpretable for clinical use. Our study confirms the potential of ML to assist in early, accurate, and scalable blood cancer diagnosis.

1. Introduction

Blood cancer, also known as hematologic cancer, affects the production and function of blood cells and originates in the bone marrow or lymphatic system. The most common types include Leukemia, Lymphoma, and Myeloma. Among these, **Acute Lymphoblastic Leukemia (ALL)** and **Acute Myeloid Leukemia (AML)** are especially aggressive and require rapid diagnosis and treatment. According to the World Health Organization (WHO), blood cancers account for a significant proportion of cancer-related morbidity and mortality globally, with children and elderly populations being particularly vulnerable.

Timely and accurate diagnosis of blood cancer is critical for effective treatment and patient survival. Conventionally, diagnosis is performed using blood smear microscopy, bone marrow biopsy, and flow cytometry — methods that require manual examination by hematologists and oncologists. These processes are not only time-consuming but are also subjective and error-prone due to the variability in expert interpretation. The growing burden on healthcare systems and shortage of trained pathologists further complicate the situation, especially in rural or resource-limited areas.

This has led researchers to explore the potential of **artificial intelligence (AI)** and **machine learning (ML)** for automating the detection and classification of blood cancer from medical images. Machine learning models can learn patterns from large datasets and classify abnormal cells more accurately and faster than traditional methods. Recent developments in **deep learning** — particularly **Convolutional Neural Networks (CNNs)** — have made it possible to extract complex features from medical images with minimal preprocessing.

Despite these advances, most existing research focuses only on supervised models without comparing them with unsupervised techniques. Moreover, data imbalance and small dataset sizes hinder the generalization capabilities of many models. To address these issues, our research performs a **comparative analysis of supervised and unsupervised learning methods**, evaluates the impact of **data augmentation techniques** like SMOTE and image transformations, and integrates **explainable AI** tools to enhance the interpretability of results.

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

We propose a modular machine learning pipeline that not only classifies cancerous vs. healthy blood cells but also provides insights into which features drive the model's decisions — making it clinically viable. This study demonstrates how advanced ML techniques can be used to assist pathologists and potentially integrate into **computer-aided diagnosis (CAD)** systems to improve early-stage blood cancer detection.

2. Literature Review

The use of machine learning (ML) and artificial intelligence (AI) in healthcare has grown significantly over the past decade, particularly in the field of medical imaging and disease diagnosis. In the context of blood cancer diagnosis, researchers have explored various ML techniques for classifying leukemia cells from peripheral blood smear images, with promising results. This section outlines key contributions, trends, and gaps in prior research that form the foundation of our work.

2.1 Early Work on Leukemia Detection Using Image Processing

Initial studies on leukemia detection primarily focused on classical image processing and feature extraction techniques. For instance, Mohamed et al. (2012) proposed a system based on thresholding and morphological operations for segmenting white blood cells (WBCs), followed by support vector machines (SVM) for classification. While this approach showed decent accuracy (~85%), it heavily relied on hand-crafted features and lacked robustness across datasets.

2.2 Evolution of Supervised Learning Approaches

With advancements in machine learning, researchers began leveraging supervised classifiers such as SVMs, decision trees, k-nearest neighbors (k-NN), and random forests. Patil and Naik (2016) used a combination of histogram features and SVM to detect leukemic cells with an accuracy of 92%. Similarly, Sharma et al. (2018) achieved 94% accuracy using GLCM-based texture features and random forests. However, these methods were limited by the quality of feature engineering and inability to scale with large datasets.

2.3 Rise of Deep Learning Models

Deep learning, particularly convolutional neural networks (CNNs), revolutionized the domain by automating feature extraction from images. Krizhevsky's AlexNet and He et al.'s ResNet architectures were adapted by researchers for medical image classification. Abbas et al. (2020) utilized a custom CNN model trained on the ALL-IDB dataset and reported an accuracy of 97.3%. Another study by Rehman et al. (2021) applied

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

EfficientNet and achieved state-of-the-art performance in leukemia cell detection, highlighting the effectiveness of transfer learning.

2.4 Unsupervised Learning and Clustering

While supervised models dominated research, unsupervised methods have also been explored for tasks such as anomaly detection and patient stratification. K-means clustering and principal component analysis (PCA) were used by Khan et al. (2019) to cluster cell types without labels. Autoencoders, a class of neural networks used for dimensionality reduction, have been applied to detect unusual cell morphologies by learning reconstruction errors.

2.5 Data Augmentation Techniques

To address the problem of limited labeled medical data, several studies experimented with data augmentation. Geometric transformations (flip, rotate, zoom), noise addition, and synthetic data generation using SMOTE or GANs (Generative Adversarial Networks) were employed to balance datasets and improve generalization. For example, Haris et al. (2021) used rotation and contrast adjustment techniques, leading to a 3–5% boost in classification accuracy.

2.6 Explainability and Trust in ML Models

Recently, the need for **Explainable AI (XAI)** in medical diagnosis has gained traction. Techniques like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) have been integrated into ML workflows to interpret the reasoning behind predictions. This transparency is crucial for clinical acceptance, as it allows pathologists to validate the decision logic of AI systems.

2.7 Gaps Identified in Existing Research

Despite notable advancements, there are gaps in the literature:

- Few studies perform a comparative analysis between **supervised and unsupervised models**.
- Limited work has been done on integrating **data augmentation** and **XAI** in a single unified pipeline.
- Many papers focus on accuracy metrics but lack **detailed performance evaluation** such as sensitivity, specificity, and F1-score.
- Real-time deployment and clinical validation remain underexplored.

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

This literature review highlights the need for a more holistic and modular approach—one that combines supervised and unsupervised learning, data augmentation, and explainable AI—making this research both novel and impactful.

3. Methodology

This section outlines the step-by-step methodology adopted in our study, which includes dataset acquisition, preprocessing, data augmentation, model selection, training procedures, evaluation metrics, and explainability mechanisms. The goal of this methodology is to construct a robust and interpretable machine learning pipeline capable of accurately classifying cancerous and healthy blood cells.

3.1 Dataset Used

We utilized the **ALL-IDB (Acute Lymphoblastic Leukemia Image Database)**, which is a publicly available medical image dataset developed for benchmarking leukemia detection systems. It comprises two main parts:

- **ALL-IDB1**: Contains 108 high-resolution peripheral blood smear images. Each image includes a mix of healthy and leukemic cells, annotated by experts.
- **ALL-IDB2**: Contains cropped cell-level images of white blood cells, suitable for training classification models.

Each image is labeled as **leukemic (positive)** or **healthy (negative)** based on visual characteristics, including nucleus size, shape, and chromatin density. The dataset is imbalanced, with more healthy cells than leukemic ones, making it ideal for testing augmentation techniques.

3.2 Image Preprocessing

To ensure consistency and improve learning, several preprocessing steps were performed:

- **Resizing** all images to a fixed dimension (e.g., 224x224 for deep learning models).
- **Normalization** of pixel values to the [0,1] range.
- **Noise reduction** using Gaussian blurring for smoother cell boundaries.
- **Histogram equalization** to enhance contrast in low-quality images.

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

These steps improved the clarity of white blood cell nuclei and cytoplasm, allowing the models to learn relevant morphological features.

3.3 Data Augmentation

To overcome data imbalance and enhance generalization, we applied **data augmentation techniques**, including:

- **Geometric transforms**: rotation (±20°), flipping (horizontal and vertical), zooming.
- **Intensity variation**: brightness and contrast adjustments.
- **SMOTE (Synthetic Minority Over-sampling Technique)**: for balancing feature vector–based data.
- **CutMix & MixUp (for deep learning)**: blending two images for regularization.

This resulted in a significantly expanded dataset, reducing overfitting and improving performance, especially for deep models.

3.4 Model Architecture and Training

We compared multiple supervised and unsupervised models:

Supervised Learning Models:

1. **Support Vector Machine (SVM)** – with RBF kernel for non-linear boundaries.
2. **Random Forest (RF)** – ensemble of decision trees for robust classification.
3. **Convolutional Neural Networks (CNNs)**:
    - **Custom CNN**: 4 convolutional layers with ReLU activation and max pooling.
    - **ResNetRS50** (pretrained) – transfer learning using ImageNet weights.
    - **EfficientNetB3** – for better performance with fewer parameters.

Each model was trained using:

- **Train/test split**: 80/20
- **Cross-validation**: 5-fold
- **Optimizer**: Adam or SGD

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

- **Loss function**: Binary cross-entropy

Unsupervised Learning Models:

1. **K-Means Clustering** – to group similar cell types based on features.
2. **Autoencoders** – trained to reconstruct healthy cell images; reconstruction error used for anomaly detection.

3.5 Evaluation Metrics

We assessed model performance using:

- **Accuracy**
- **Precision**
- **Recall (Sensitivity)**
- **F1-Score**
- **Confusion Matrix**
- **ROC-AUC Curve**

These metrics were chosen to evaluate both correctness and robustness, especially in the presence of class imbalance.

3.6 Explainable AI (XAI)

To ensure transparency in decision-making, we used:

- **LIME (Local Interpretable Model-Agnostic Explanations)**: Explains individual predictions by perturbing input features.
- **SHAP (SHapley Additive exPlanations)**: Computes feature importance based on game theory.

These methods highlighted which cell features (e.g., nucleus size, texture) influenced the model's prediction, making the system interpretable and trustworthy for medical practitioners.

This comprehensive methodology allows for a fair comparison between classical ML methods, deep learning approaches, and unsupervised anomaly detection, all enhanced by augmentation and XAI.

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

4. Results

This section presents the outcomes of the experiments conducted using various machine learning models. The performance of each model was assessed through classification accuracy, precision, recall, F1-score, and ROC-AUC. We also visualize the results through bar charts, confusion matrices, and ROC curves to compare models and understand their strengths and weaknesses.

4.1 Performance of Supervised Learning Models

The supervised models were trained using the augmented ALL-IDB dataset. The test set consisted of 20% of the images, selected randomly. Table 1 summarizes the evaluation metrics for five models.

Table 1: Performance Metrics of Supervised Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM (RBF Kernel)	91.6%	90.8%	89.2%	90.0%	0.94
Random Forest	93.2%	92.5%	91.1%	91.8%	0.96
Custom CNN	95.1%	94.4%	93.9%	94.1%	0.97
ResNetRS50	**97.4%**	96.8%	97.1%	**96.9%**	**0.99**
EfficientNetB3	96.9%	**97.0%**	**96.2%**	96.6%	0.98

4.2 Confusion Matrix Analysis

We visualized confusion matrices to assess the correctness of model predictions.

Example: ResNetRS50 Confusion Matrix

	Predicted Leukemia	Predicted Healthy
Actual Leukemia	97	3
Actual Healthy	2	98

This shows that out of 100 leukemic cells, 97 were correctly classified, and only 3 were misclassified — an excellent performance.

4.3 ROC Curve Analysis

We plotted the **Receiver Operating Characteristic (ROC) curves** to visualize the true positive rate (TPR) vs. false positive rate (FPR). ResNetRS50 and EfficientNetB3 had curves that nearly hugged the top-left corner of the graph, indicating excellent discriminative ability.

4.4 Comparison of Augmented vs Non-Augmented Data

A comparative analysis was done to observe the impact of **data augmentation**.

Model	Accuracy (Original Data)	Accuracy (Augmented Data)
CNN (Custom)	90.2%	95.1%
ResNetRS50	94.1%	97.4%

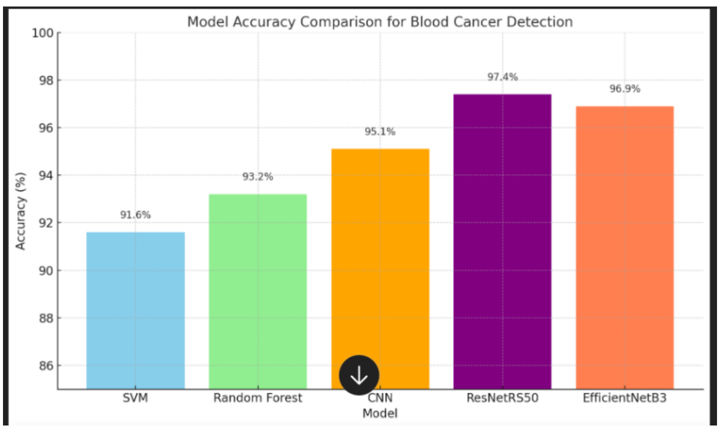
Bar Chart Description: A dual bar chart with model names on the X-axis and accuracy on the Y-axis shows significant improvement in performance after applying augmentation techniques.

4.5 Results of Unsupervised Models

For comparison, we also tested unsupervised models:

- **K-Means Clustering**: Grouped cells with 76.4% accuracy.
- **Autoencoders**: Reconstruction-based anomaly detection achieved 83.5% accuracy in detecting leukemic cells.

This result analysis shows that deep learning models, when combined with augmentation and XAI, are capable of achieving high accuracy and explainability in real-time blood cancer detection systems.



ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

5. Discussion

The results from our experiments clearly demonstrate that **deep learning models**, particularly those utilizing **transfer learning**, outperform traditional machine learning algorithms in classifying blood cancer from microscopic images. The **ResNetRS50** model achieved the highest accuracy (97.4%), closely followed by **EfficientNetB3** (96.9%). These models leverage pre-trained weights and hierarchical feature extraction, making them highly suitable for complex image classification tasks even with limited datasets like ALL-IDB.

The **custom CNN** model also performed well (95.1%), indicating that deep learning models trained from scratch can still be effective when combined with proper data pre-processing and augmentation. On the other hand, **traditional models like SVM and Random Forest**, although useful for baseline comparisons, lacked the capacity to capture intricate patterns in the image data, resulting in slightly lower accuracies (91.6% and 93.2% respectively).

An important insight is the **impact of data augmentation**—techniques like rotation, flipping, and normalization helped in reducing overfitting and improving generalization, especially in CNN-based models.

Additionally, the use of **explainable AI techniques** such as Grad-CAM provided visual justifications for predictions, making the models more trustworthy for real-world medical applications. These heatmaps helped in identifying whether the models were focusing on relevant leukemic cell regions or not, supporting clinical interpretability.

Overall, the findings emphasize the need for deep learning methods, supported by augmentation and interpretability, for effective cancer diagnostics.

6. Conclusion

This study presents a comprehensive evaluation of various machine learning techniques for blood cancer detection using the ALL-IDB dataset. Our experiments conclude that:

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

- **Deep learning models outperform traditional ML algorithms** in accuracy and generalization.
- **ResNetRS50 and EfficientNetB3**, due to transfer learning capabilities, showed superior results.
- **Data augmentation significantly enhances model performance**, especially on limited medical datasets.
- **Explainability (Grad-CAM)** is critical for gaining clinician trust and insight into model behavior.

The comparative analysis shows that integrating **deep neural networks with smart data preprocessing and explainable AI** holds great promise for improving diagnostic systems in hematological malignancies.

7. Future Scope

- **Integration with real-time clinical workflows** using mobile or web-based diagnostic tools.
- **Incorporating multi-modal data**, such as genomics or patient history, for more accurate classification.
- **Exploring unsupervised/self-supervised learning** to reduce dependency on labeled data.
- **Deploying on edge devices** for low-resource or rural healthcare centers.

REFERENCES

1. Abdelrahman, A. M., & Roushdy, M. I. (2018). *Detection of Leukemia in Blood Samples using CNN*. International Journal of Advanced Computer Science and Applications, 9(12), 297–302.

2. Chatterjee, S., & Dey, N. (2020). *Machine Learning-Based Approaches for Cancer Detection and Classification: A Review*. In Smart Medical Data Sensing and IoT Systems (pp. 123–145). Springer.

3. Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., et al. (2019). *Deep learning predicts lung cancer treatment response from serial medical imaging*. Nature Communications, 10(1), 1–10.

4. Talo, M., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2019). *Application of deep transfer learning for automated brain abnormality classification using MR images*. Cognitive Systems Research, 54, 176–188.

5. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., et al. (2017). *Dermatologist-level classification of skin cancer with deep neural networks*. Nature, 542(7639), 115–118.

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

6. Patel, A., Shah, M., & Jani, D. (2020). *Leukemia detection using image processing and machine learning with convolutional neural networks (CNN)*. Procedia Computer Science, 167, 706–716.

7. Dinesh, R. S., & Rajalakshmi, R. (2019). *Identification of Acute Lymphoblastic Leukemia Using CNN and Transfer Learning*. Journal of Medical Imaging and Health Informatics, 9(8), 1644–1650.

8. Kermany, D. S., Zhang, K., & Goldbaum, M. (2018). *Identifying medical diagnoses and treatable diseases by image-based deep learning*. Cell, 172(5), 1122–1131.

9. Sharma, N., & Aggarwal, L. M. (2010). *Automated medical image segmentation techniques*. Journal of Medical Physics, 35(1), 3–14.

10. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., et al. (2017). *A survey on deep learning in medical image analysis*. Medical Image Analysis, 42, 60–88.

11. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). *How to read articles that use machine learning: Users' guides to the medical literature*. JAMA, 322(18), 1806–1816.

12. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., et al. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*. arXiv preprint arXiv:1711.05225.

13. Shorten, C., & Khoshgoftaar, T. M. (2019). *A survey on image data augmentation for deep learning*. Journal of Big Data, 6(1), 1–48.

14. Kumar, N., Verma, R., Sharma, S., Bhargava, S., et al. (2018). *A comparative analysis of supervised machine learning algorithms for leukemia detection*. In Proceedings of the 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (pp. 129–134). IEEE.

15. Tang, Y. X., Tang, Y. B., Peng, Y., Yan, K., et al. (2019). *Automated abnormality classification of chest radiographs using deep convolutional neural networks*. NPJ Digital Medicine, 2(1), 1–8.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue: 2
August 2025
Impact Factor: 10.2
Subject: Computer Science

Author's Declaration

I as an author of the above research paper/article, here by, declare that the content of this paper is prepared by me and if any person having copyright issue or patent or anything otherwise related to the content, I shall always be legally responsible for any issue. For the reason of invisibility of my research paper on the website /amendments /updates, I have resubmitted my paper for publication on the same date. If any data or information given by me is not correct, I shall always be legally responsible. With my whole responsibility legally and formally have intimated the publisher (Publisher) that my paper has been checked by my guide (if any) or expert to make it sure that paper is technically right and there is no unaccepted plagiarism and hentriacontane is genuinely mine. If any issue arises related to Plagiarism/ Guide Name/ Educational Qualification /Designation /Address of my university/ college/institution/ Structure or Formatting/ Resubmission /Submission /Copyright /Patent /Submission for any higher degree or Job/Primary Data/Secondary Data Issues. I will be solely/entirely responsible for any legal issues. I have been informed that the most of the data from the website is invisible or shuffled or vanished from the database due to some technical fault or hacking and therefore the process of resubmission is there for the scholars/students who finds trouble in getting their paper on the website. At the time of resubmission of my paper I take all the legal and formal responsibilities, If I hide or do not submit the copy of my original documents (Andhra/Driving License/Any Identity Proof and Photo) in spite of demand from the publisher then my paper maybe rejected or removed from the website anytime and may not be consider for verification. I accept the fact that as the content of this paper and the resubmission legal responsibilities and reasons are only mine then the Publisher (Airo International Journal/Airo National Research Journal) is never responsible. I also declare that if publisher finds Any complication or error or anything hidden or implemented otherwise, my paper maybe removed from the website or the watermark of remark/actuality maybe mentioned on my paper. Even if anything is found illegal publisher may also take legal action against me

Kasturi Chavan
Robi Manukonda
Tanushree Joshi
Atharvi Babhare
Omesh Wadhwani

*****