International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Blood Cancer Detection and Classification using ML Algorithms

[#1]M.V. PHANINDRA, Research Scholar,

[#2]Dr. G. THIPPANNA, Supervisor,

[#1,#2]*Dept. of Computer Science and Engineering*

[#1,#2]NIILM University, Kaithal, Haryana, India

Abstract: The automatic detection of blood cancer, is a challenging yet crucial task for ensuring timely diagnosis and treatment. Leukemia affects the blood and bone marrow, and its symptoms often overlap with those of other conditions, making it difficult to detect. While traditional diagnostic methods, such as blood tests, bone marrow biopsies, and imaging techniques like CT scans and MRIs, are commonly used, they rely heavily on the expertise of trained medical professionals for accurate interpretation. However, these methods may not always provide reliable or consistent results. Machine learning algorithms offer a promising solution by automating the detection process. By analysing large datasets of patient information and medical images, these algorithms can identify subtle patterns and indicators of leukaemia that might be challenging for human experts to spot, improving both the speed and accuracy of detection.

1 Introduction

Blood cancer, also known as hematologic cancer, encompasses a group of malignancies that affect the blood, bone marrow, lymphatic system, and spleen. This category includes leukaemia, lymphoma, and myeloma, each of which originates from different types of blood cells and presents distinct clinical challenges. Leukemia primarily impacts white blood cells, leading to their uncontrolled proliferation, while lymphoma involves the malignant growth of lymphocytes, a key component of the immune system. Myeloma, on the other hand, results from the abnormal growth of plasma cells in the bone marrow [1]. These cancers disrupt the normal functioning of the hematologic system, leading to a range of symptoms such as fatigue, immune system compromise, and abnormal bleeding. The complexity of blood cancers lies not only in their diverse molecular mechanisms but also in their varied response to treatment, which may include chemotherapy, stem cell transplantation, and immunotherapy. Despite advances in treatment, blood cancers continue to pose significant challenges in terms of early diagnosis, personalized therapy, and long-term outcomes, making them a critical area of ongoing research.

Machine learning (ML) algorithms have emerged as powerful tools for predicting and diagnosing blood cancers, particularly in enhancing early detection, improving treatment strategies, and personalizing patient care [2]. Blood cancers are complex and multifaceted diseases that involve intricate genetic, molecular, and clinical factors. Traditional diagnostic methods can be time-consuming and may not always provide the granularity needed for early intervention. ML algorithms, by contrast, can process vast amounts of data and identify hidden patterns that can significantly improve prediction accuracy and treatment outcomes. One key area where ML is beneficial is in the early detection and classification of blood cancers [3]. By analysing clinical and laboratory data such as blood counts, imaging results, genetic mutations, and patient demographics, ML algorithms can predict the likelihood of a patient developing a blood cancer or help classify subtypes of the disease. Supervised learning algorithms can be trained on datasets that contain labelled instances of cancerous and non-cancerous cases [4]. These models can then

predict the presence or absence of blood cancer based on a patient's specific features, often with high accuracy, enabling faster diagnosis and reducing human error.

By integrating multi-omics data (including genomics, proteomics, and metabolomics) and treatment histories, machine learning models can be used to predict how individual patients will respond to various chemotherapy drugs, targeted therapies, or immunotherapy regimens. This personalized approach, often referred to as precision medicine, can significantly improve patient outcomes by tailoring treatments to an individual's specific genetic makeup, potentially leading to more effective and less toxic therapies. Furthermore, ML algorithms can be employed in monitoring and detecting minimal residual disease, which refers to the small number of cancer cells that remain in the body after treatment and can cause relapse. ML models can analyse post-treatment data, such as gene expression or liquid biopsy results, to identify traces of minimal residual disease that might be missed by traditional diagnostic methods, allowing for timely interventions before the disease reoccurs [5]. In summary, ML has the potential to revolutionize blood cancer diagnosis, prognosis, and treatment by enabling early detection, personalized treatment plans, and better disease monitoring. As research advances and datasets grow, these technologies will likely become integral tools in improving patient outcomes and guiding clinical decision-making in hematologic oncology [6].

This article is structured as follows: Section II discusses and outlines of various researchers employed work on blood cancer detection. Section III describes the proposed ML algorithms. Section IV discusses dataset description and experiment setup employed in this research. Section V presents the results of the experimental validation for the BCCD dataset. The article concludes with Section VI.

2 Literature review

This chapter reviews the literature on the use of image recognition and computer vision in connection to the medical system, blood cell identification. This paper covers the background and present state of haematological systems before reviewing studies done on both normal and pathological blood samples. Understanding the present state of the work and identifying its flaws are the goals of this examination.

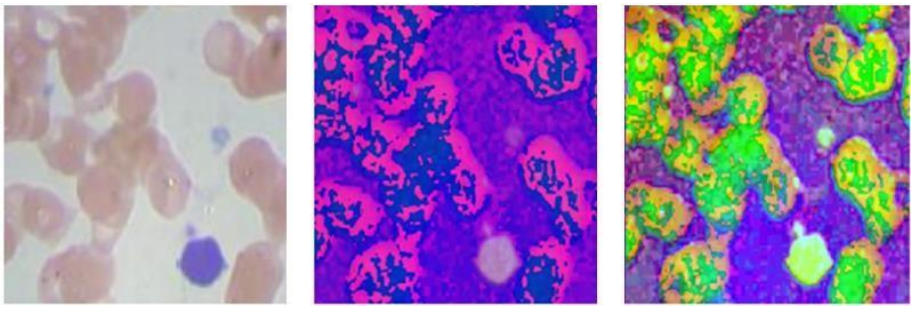
Researchers have conducted some previous studies in this section. [7] Reviewing machine learning algorithms applied to leukaemia detection, this paper provides a comprehensive overview. With its emphasis on standardized datasets and proper evaluation metrics, the review serves as a valuable resource for researchers and 91% in KNN, 91% in Naive Bayes, and 75% in SVM because of the small size of the dataset.
[8] Vogado et al. achieved 100% accuracy in the performance, challenges, and challenges associated with Multilayer Perceptron's, SVM, and RFs. Considering the results of the tests, it can be concluded that segmentation is not necessary to locate specific cells with this method, as it reduces the processing time for creating the blood smear image. Balakumar K *et al.* [9] This paper presents exhibits that combine machine learning algorithms to ensure accurate and early detection of the disease, feature selection techniques and model optimization are employed, found that the accuracy of the classification was 92% in comparison to other standard classifiers.

Khaled A. S. Abu Daqqa *et. al.* [10] proposed a prediction and diagnosis system for leukemia using Classification algorithms. They applied several classification algorithms such as DT, SVM, and KNN. In their approach, decision trees obtained higher accuracy than others. The accuracy that they have found using a DT, SVM, and KNN is 77.30%, 76.82%, and 70.15%, respectively. They show the prediction accuracy of different classification model but did not work on any image processing technique.

3 Methodology

3.1 preprocessing of image

Any categorization requires preprocessing a picture since it sets the image up for further processing. The RGB colour pictures from the CT scan of the red blood sample are first transformed to HSV images for the operation. Colour conversion is the process of combining RBC subtractive coloured layers for a certain application and defined colour combination. Once this is done, the picture is subjected to histogram equalization. Brightness dependent contrast is a phase image-based application, where histogram equalized pictures are used to increase pixel quality for pre-processed images by calculating pixel intensity. The global contrast of the cancerous RBC picture increases with intensity value.



(A) (B) (C)

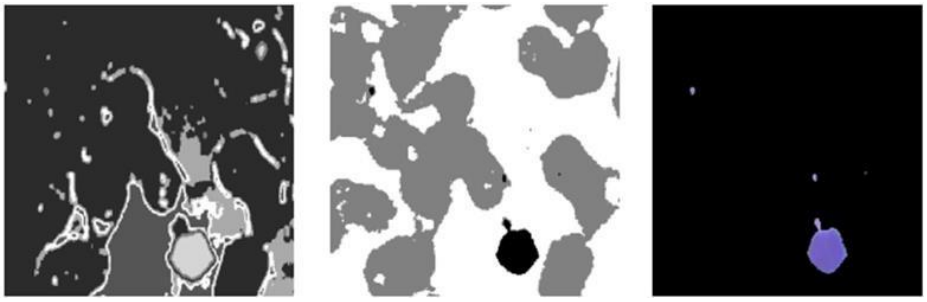
Figure 1: (A) Benign RBC Tumour sample 1 pre-processed picture (B) Histogram-based input colour-converted image and (C) sample image equalization

The colour scale histogram offers the equalization step for detection at the threshold's limited values. For human viewers, the sense of visual borders is growing at a significant pace. The result of histogram-specified equalization has a uniform distribution of intensity and provides intensity values in the input picture. For uniformly based histogram values, it raises the objective of contrast with nominal histogram equalization. Histogram equalization, which modifies image intensities to boost background and foreground contrast, is used to improve performance. When data for constant contrast values at a global level of contrast are available, this approach is used. The resulting improved picture for the segmentation stage is more aesthetically pleasing and features distinct nodules.

3.2 Image segmentation

Following the preprocessing stage, the output from this phase is used for segmentation in order to remove malignant cells from CT RBC pictures. The disorders that are of important structures with categorization, the segments of different classes of tissue organs, and the imaging at medical condition. The RBC segmentation state is now emerging in the low contrast imaging ambiguity range. Functionally, the suggested map-based segmentation job uses a tagged reference picture and an MRF system to identify feature values.

To effectively recognize cluster pixels, this clustering is identified. K-means centroid begins using the suggested method, effectively assigning from the reference cluster with a decrease in mean aligned between the allocated pixels within the cluster centre. K-means clustering is a technique that effectively analyses online pixel grouping with comparable intensity levels in non-cancerous zones. Radiologists may diagnose patients more accurately by using the suggested approach, which precisely separates malignant cells from non-cancerous cells in fewer affected areas. By determining the minimal area threshold values based on the previous information of the image collected series, the online region-based segmentation is applied. Effective segmentation is provided by the suggested ORBS approach, which processes data live. It makes the foreground pixels' border larger. Segmentation of the picture is applied according to its form and features.



(A) (B) (C)

Figure 2: Benign Red Blood Cell Tumour Sample 1's Segmented Image (A) Segmentation Map (B) K-means Cluster and (C) Suggested ORBS

Even the regular tissue portions are taken into consideration together with all the foreground elements. Segmentation mapping expands the area according to foreground pixel boundaries. A picture may be retrieved based on its form and properties. A preliminary extraction involving minor components exists. The K stands for clustering, which uses processing methods to distinguish between healthy and malignant cells. Using new segmentation, a powerfully simple method splits lighter items at background cells. In smaller blood cells, this divides cancer cells. For use in medical analysis, the suggested online region-based segmentation is carried out automatically. This technique eliminates every flaw in the previous approaches and produces accurate visual results when separating the afflicted RBC cells from the unaffected RBC cells.

Table 1: CT imaging of Benign Red Blood Cell Tumours: Features and Values

Image	Contrast	Clustering Shade values	Clustering prominence values	Correlation density	Dissimilarity coefficient	Probability Values
1	0.453402	-24.35896	324.884	0.928544	0.364405	0.1952
2	0.480787	-23.97541	369.751	0.748975	0.201361	0.1544
3	0.278814	-22.28947	270.458	0.926434	0.160338	0.1613
4	0.425737	-23.97845	296.354	0.954037	0.193531	0.1643
5	0.482967	-24.82447	332.154	0.849975	0.241061	0.1944
6	0.581298	-18.94672	271.394	0.827944	0.173282	0.1552
7	0.491035	-19.35874	264.897	0.754218	0.343656	0.1653
8	0.407573	-17.94246	279.365	0.955328	0.475228	0.1797
9	0.496626	-14.75612	256.684	0.751768	0.203421	0.1861
10	0.595226	-16.38726	222.333	0.959457	0.348617	0.1883
11	0.241714	-21.36726	278.394	0.824956	0.244935	0.1651
12	0.597532	-18.97521	284.654	0.913457	0.396749	0.1766
13	0.523191	-22.97564	345.897	0.745301	0.427157	0.1842
14	0.579228	-24.36994	336.655	0.915996	0.271558	0.1518
15	0.496519	-19.06789	397.254	0.810656	0.378492	0.1371
16	0.395769	-19.87624	378.245	0.708442	0.417742	0.1758
17	0.439558	-17.59251	363.252	0.951269	0.272483	0.1445
18	0.397581	-16.54212	397.845	0.752689	0.274834	0.1650
19	0.394613	-18.42351	387.112	0.870231	0.324856	0.1848
20	0.377559	-23.97511	278.951	0.867869	0.494871	0.1920
21	0.486621	-21.34562	202.646	0.844381	0.374931	0.1990

3.3 Feature extraction

For feature extraction, histogram-based features are often taken into consideration, with feature values determined by the threshold values of the histogram. The classifier uses these properties to classify the illness in the red blood cell image. These database photos are used to generate features based on histograms.

4 Datasets Description and Experiment Setup

4.1 Dataset

The BCCD (Blood Cell Count and Detection) dataset consists of 200 benign and 200 malignant blood cell images (JPEG format) from 534 patients, along with corresponding cell type labels. The dataset includes four types of blood cells: neutrophils, eosinophils, lymphocytes, and monocytes. Figure 1 displays the dataset samples of BCCD. Initially, the benign cell images are processed to identify potential cancer cells. The data is split into a training set (80%) and a test set (20%) using random stratified sampling. However, due to the limited number of training samples, deep learning models may face challenges in learning effectively. To address this, we applied image augmentation techniques such as rotating and flipping the white blood cell (WBC) images to artificially expand the training set. This augmentation resulted in 400 images for each type of WBC, which were used to create a more robust training set. For model training, we leveraged pre-trained models, initially trained on the ImageNet dataset, and fine-tuned them by adjusting their weights based on our augmented dataset. To optimize the models, we performed five-fold cross-validation, selecting the best model based on validation accuracy. The performance of the model was then evaluated using the test set.

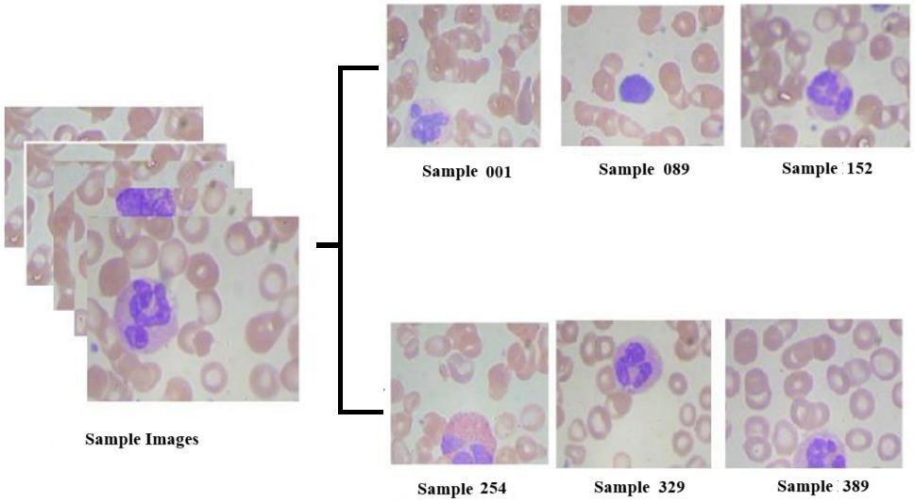


Figure 3: BCCD Dataset Samples

4.2 Experiment Setup

The evaluation was performed by using the statistical tool-rich MATLAB 2018a program for designing framework and to plot the graphs.

5 Results and Discussion

The example pictures in this section are classified using ANN, KNN, SVM, and BRT classifiers. Efficiency is influenced by sensitivity, specificity, accuracy, precision, and error value, among other things.

The SVM classifier is given characteristics that are based on histograms. Class 0 represents a benign tumour picture and Class 1 represents a malignant red blood cell image in Figure 4(a). Eight of the fifty benign tumours were wrongly identified as malignant, whereas 42 benign tumours were accurately diagnosed as benign thanks to the use of an SVM classifier and feature values based on histograms. Seven instances are mistakenly labelled as benign, whereas 43 cases accurately diagnosed as malignant are reviewed out of every 50 malignant cases. In the SVM uncertainty matrix, TP is 41.9 percent, FP is 58.1 percent, FN is 57.7 percent, and TN is 42.3 percent. The accuracy score of the SVM classifier is 94.7 percent. The rate of misclassification is 5.3 percent.

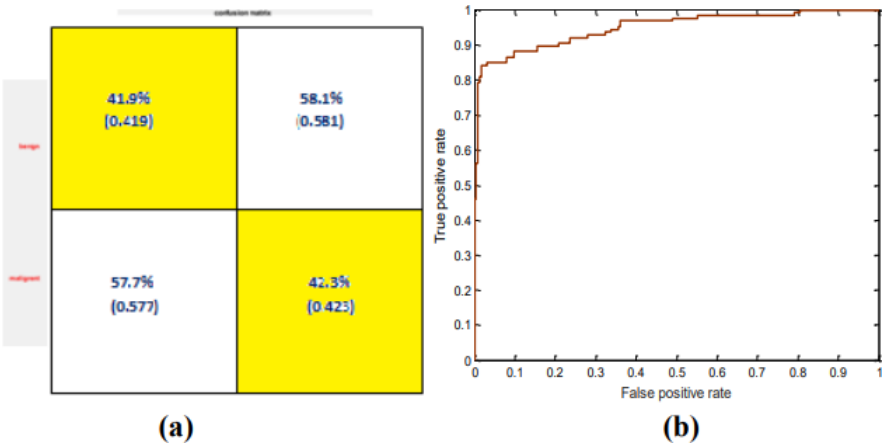
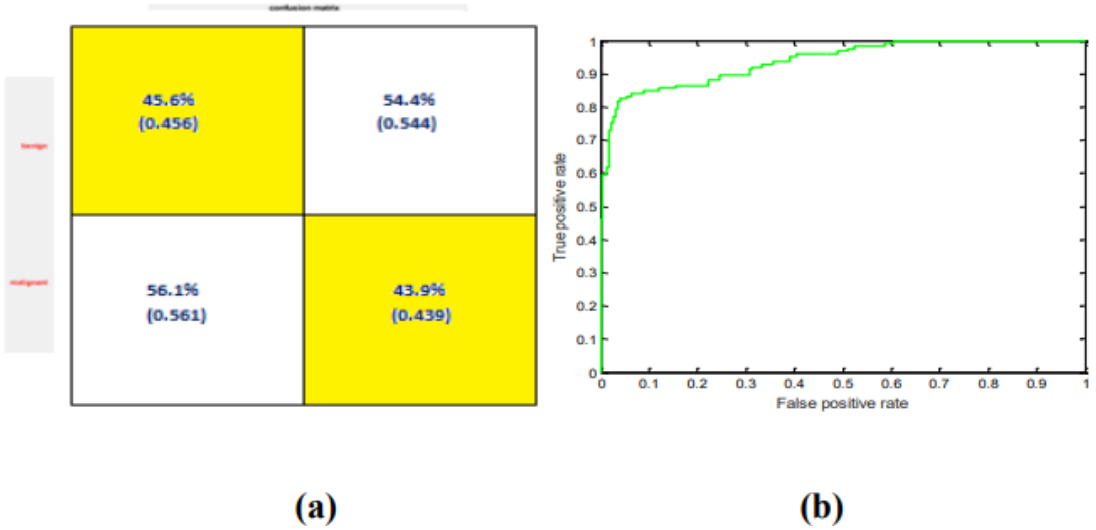


Figure 4: a) Confusion matrix, b) ROC Curve- SVM Classifier

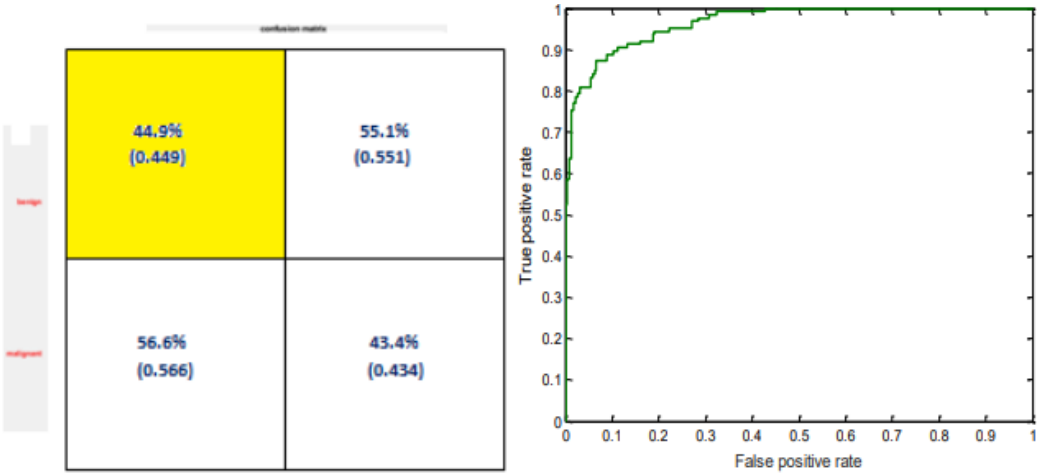
KNN provides a benign or malignant classification to the images in the Red Blood Cell Tumour Database. Class 0 represents a benign tumour picture and Class 1 represents a malignant red blood cell image in Figure 5(a). Four of the 50 benign are mistakenly categorized as malignant, while 46 of the 50 benign are accurately identified as benign using a KNN classifier and feature values based on histograms. Six instances are mistakenly labelled as benign, and 44 cases accurately identified as malignant out of every 50 malignant cases.

ANN provides fresh characteristics and outputs for back propagation training weights are first set at random. Errors occur at both the intended and actual manufacturing phases because of weight changes at each era. The confusion matrix with ROC for the ANN classifier that was produced using features based on histograms is shown in Figure 6. Class 1 represents a benign tumour picture and Class 2 represents a malignant red blood cell image in Figure 6(a). Using an ANN classifier using Histogram-based feature values, 45 of the 50 benign cases are categorized as benign, whereas 5 are incorrectly labelled as malignant. Six instances are mistakenly labelled as benign, and 44 cases accurately identified as malignant out of every 50 malignant cases.



(a) (b)

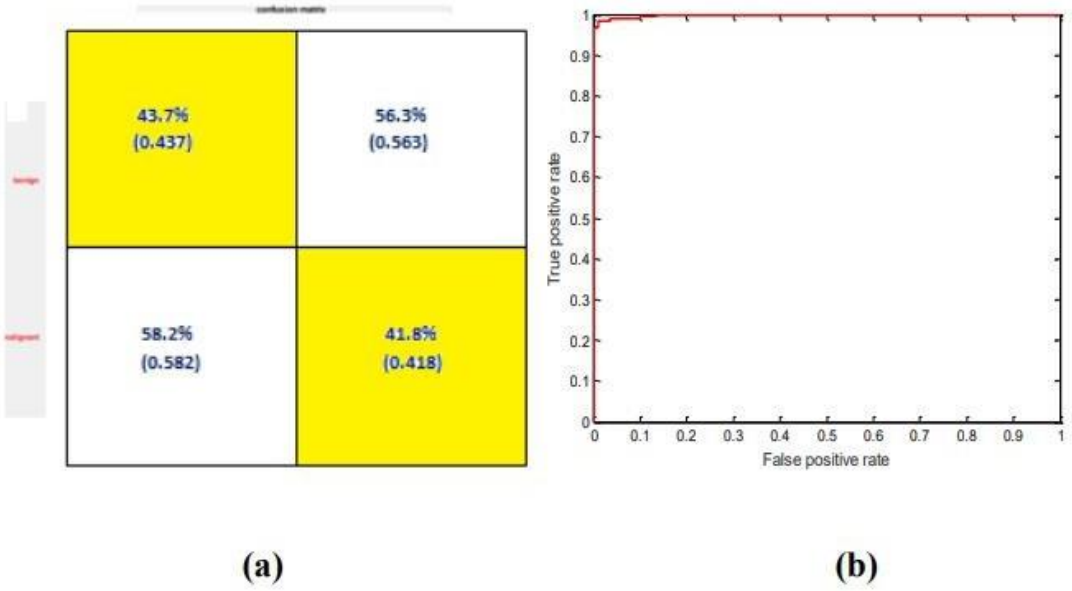
Figure 5: a) Confusion matrix, b) ROC Curve - KNN Classifier



(a) (b)

Figure 6: a) Confusion matrix, b) ROC Curve - ANN Classifier

A Bagged Random Tree (BRT) classifier is a bootstrap aggregation tree. A BRT classifier based on features derived from histograms is produced by combining the confusion matrix with the ROC curve, as seen in Figure 7. Class 0 represents a benign tumour picture and Class 1 represents a malignant red blood cell image in Figure 7(a). 44 of the 50 benign were rated as benign by the BRT classifier using Histogram-based function values, whereas 6 were incorrectly labelled as malignant. Eight instances are mistakenly labelled as benign, and 42 cases accurately identified as malignant out of every 50 malignant cases evaluated.



(a) (b)

Figure 7: a) Confusion matrix, b) ROC Curve - BRT Classifier

Table 2: Performance values of different classifier

Types	BRT	SVM	KNN	ANN
Sensitivity(%)	91.3	94.6	93.1	92.3
Specificity(%)	93.6	96.7	83.6	93.5
Precision(%)	95.5	94.3	84.3	83.7
Accuracy(%)	92.7	96.1	92	93.7
Time (second)	43	30	27	35

In the context of blood cancer detection, the performance of different machine learning models can be assessed through metrics such as sensitivity, specificity, precision, and accuracy. Among the models tested SVM, K-NN, ANN, and BRT. SVM stands out as the most effective model for this task.

The SVM model achieves the highest sensitivity 94.6%, indicating it excels at identifying true positive cases of blood cancer. It also has the highest specificity 96.7%, meaning it is effective at minimizing false positives and correctly identifying healthy individuals. This balance between sensitivity and specificity makes SVM particularly reliable for detecting blood cancer with minimal misclassification. K-NN offers a strong sensitivity 93.1% but falls behind in specificity 83.6% and precision 84.3%, suggesting it may have a higher rate of false positives. This is a critical factor in medical diagnoses, where false positives can lead to unnecessary treatments or anxiety. The ANN model, with a sensitivity of 92.3% and accuracy of 93.7%, provides decent overall performance but suffers from lower precision 83.7%, indicating it may misclassify a considerable number of negative cases as positive. BRT model delivers the highest precision 95.5% and solid specificity 93.6%, which suggests it is good at correctly identifying non-cancer cases. However, its lower sensitivity 91.3% indicates it might miss some true positives, which is a concern in critical medical applications where catching all positive cases is crucial. Overall, SVM is the most balanced model for blood cancer detection, demonstrating superior performance in both sensitivity and specificity. While BRT offers strong precision, it is outperformed by SVM in terms of sensitivity, making SVM the more reliable choice for ensuring accurate and comprehensive detection of blood cancer.

6 Conclusion

In conclusion, ML models, particularly the SVM, have demonstrated significant effectiveness in predicting blood cancer detection. Using key performance metrics like sensitivity, specificity, precision, and accuracy, these models show promise in improving diagnostic accuracy and reliability. SVM, with its high sensitivity and specificity, stands out as a robust tool for detecting blood cancer, ensuring both accurate identification of positive cases and minimizing false positives. Other models like K-NN, ANN, and BRT also offer valuable insights, with each excelling in specific areas such as precision or accuracy. However, the overall comparison highlights the potential of ML in enhancing medical diagnostics by providing precise, reliable, and timely predictions. With further advancements in these techniques, ML can play an integral role in improving early detection and treatment outcomes for blood cancer patients, making it a powerful tool in the healthcare domain.

References

[1] Bukhari, Maryam, Sadaf Yasmin, Saima Sammad, and Ahmed A. Abd El-Latif. "A deep learning framework for leukemia cancer detection in microscopic blood samples using squeeze and excitation learning." *Mathematical problems in engineering* 2022, no. 1 (2022): 2801227.

[2] D. K. K. Reddy, H. S. Behera, J. Nayak, A. R. Routray, P. S. Kumar, and U. Ghosh, “A Fog-Based Intelligent Secured IoMT Framework for Early Diabetes Prediction,” 2022, pp. 199–218. doi: 10.1007/978-3-030-81473-1_10.

[3] Ghaderzadeh, Mustafa, Farkhondeh Asadi, Azamossadat Hosseini, Davood Bashash, Hassan Abolghasemi, and Arash Roshanpour. "Machine learning in detection and classification of leukemia using smear blood images: a systematic review." *Scientific Programming* 2021, no. 1 (2021): 9933481.

[4] D. K. K. Reddy, J. Nayak, H. S. Behera, V. Shanmuganathan, W. Viriyasitavat, and G. Dhiman, “A Systematic Literature Review on Swarm Intelligence Based Intrusion Detection System: Past, Present and Future,” Archives of Computational Methods in Engineering, Mar. 2024, doi: 10.1007/s11831-023-10059-2.

[5] J. N. ; P. S. K. ; D. K. R. ; B. Naik, “Identification and classification of hepatitis C virus: an advance machine-learning-based approach,” in Blockchain and Machine Learning for e-Healthcare Systems, 2020, ch. 16, pp. 10–11. doi: 10.1049/PBHE029E_ch.

[6] 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.

[7] A Comprehensive Study of Machine Learning Algorithms for Predicting Leukemia Based on Biomedical Data, author: Anamika Das Mou and Pratap Kumar Saha, 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), the year 2019.

[8] L. H. S. Vogado, R. D. M. S. Veras, A. R. Andrade, F. H. D. de Araujo, R. R. V. Silva, and K. R. T. Aires, "Diagnosing Leukemia in Blood Smear Images Using an Ensemble of Classifiers and Pre-Trained Convolutional Neural Networks," 2017 30th SIBGRAPI Conference on Graphics, Patterns, and Images (SIBGRAPI), Niteroi, Brazil, 2017, pp. 367-373.

[9] B. K, G. A. T, N. G, and S. Umamaheswari, "Improving the Performance of Leukemia Detection using Machine Learning Techniques," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 867-872.

[10] Daqqa KA, Maghari AY, Al Sarraj WF," Prediction and diagnosis of leukaemia using classification algorithms",8thInternational Conference on Information Technology (ICIT), 2017 May 17 (pp. 638-643).