Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 13, No. 1, Juli 2025

ISSN 0216 – 0544
e-ISSN 2301– 6914

Comparative analysis of random forest and deep learning approaches for automated acute lymphoblastic leukemia detection using morphological and textural features

Windra Swastika [a], Kestrilia Rega Prilianti [b], Paulus Lucky Tirma Irawan [c], Hendry Setiawan [d]

[a,b,c,d]Informatics Engineering, Faculty of Technology and Design, Universitas Ma Chung, Malang
E-mail: windra.swastika@machung.ac.id

Abstract

Acute Lymphoblastic Leukemia (ALL) is a type of blood cancer that requires early and accurate detection for effective treatment. Current diagnostic approaches face significant challenges including time-consuming manual examination, inter-observer variability, and difficulty in balancing sensitivity with specificity. This study aims to develop and compare two automated ALL detection methodologies to overcome these limitations. We propose: (1) a Random Forest classifier using carefully engineered morphological and textural features, and (2) a Convolutional Neural Network (CNN) architecture for automated feature learning from microscopic blood cell images. Using 10,661 images from the ALL Challenge dataset, we evaluated both approaches on training (70%), validation (15%), and test (15%) sets. Feature importance analysis revealed cell area (10.71%), energy (10.67%), and skewness (10.50%) as the most significant discriminative features. The Random Forest achieved 85% accuracy with notable sensitivity for ALL detection (93%), while the deep learning approach demonstrated superior performance with 87% accuracy and better false positive control (27.50% vs. 35.76%). Our comparative analysis shows that while both methods demonstrate clinical viability for automated ALL screening, the deep learning approach offers advantages in reducing false positives while maintaining high detection sensitivity. This research contributes to the advancement of computer-aided diagnostic tools that can support pathologists in early ALL detection, potentially reducing
diagnostic time and improving consistency.

Key words: Acute lymphoblastic leukemia, Deep learning, Medical image analysis, Morphological features, Random forest.

INTRODUCTION

Acute Lymphoblastic Leukemia (ALL) is a life-threatening blood cancer characterized by the abnormal proliferation of immature lymphoid cells in bone marrow, blood, and other organs [1]. ALL develops very quickly and requires early detection for effective treatment, as delayed diagnosis can lead to rapid disease progression and reduced survival rates. The disease particularly affects adults over 50 years and children under five years who are in the higher-risk group [2] [3].

Traditional diagnosis of ALL relies heavily on manual microscopic examination of blood smears by expert hematologists. However, this manual detection process faces several significant challenges. First, it is time-consuming and labor-intensive, requiring specialized expertise that may not be readily available in all healthcare settings. Second, the process is prone to inter-observer variability

due to the subjective nature of visual assessment. Third, the complex nature of blood cells, presence of noise, weak edges, intensity inhomogeneity, and cell overlapping make consistent and accurate manual detection difficult [4] [5]. These factors collectively contribute to potential diagnostic delays and inconsistencies that can impact patient outcomes.

Computer-aided diagnostic systems have emerged as potential solutions to address these limitations. However, developing effective automated systems for ALL detection presents its own set of challenges. Current automated approaches struggle with: (1) accurately distinguishing between immature leukemic blasts and normal lymphocytes due to their morphological similarities; (2) handling the high variability in cell appearance caused by staining variations and imaging conditions; (3) balancing sensitivity and specificity to minimize both false negatives that could lead to missed diagnoses and false positives that might result in unnecessary further testing.

In the context of ALL diagnosis, morphological and textural features play a crucial role as they directly represent the physical and structural characteristics that hematologists use for identification. Leukemic lymphoblasts typically exhibit distinct morphological patterns, including larger cell size, irregular nuclear shapes, and abnormal chromatin patterns compared to normal lymphocytes. The systematic extraction of these features is essential because it provides quantifiable metrics for cell characteristics that experts traditionally assess qualitatively. Features such as cell area, perimeter, circularity, and nuclear-to-cytoplasmic ratio can effectively capture the subtle differences between normal and leukemic cells. Additionally, textural features derived from techniques like Gray Level Co-occurrence Matrix (GLCM) can quantify the internal patterns and chromatin distribution that are key indicators of cell abnormality [6].

Machine learning approaches for ALL detection have evolved over time. Earlier methods focused on carefully engineered features to capture distinctive characteristics of leukemic cells. The extraction of morphological features like cell area, perimeter, circularity and textural features using Gray Level Co-occurrence Matrix (GLCM) has proven effective in distinguishing leukemic cells from healthy ones [6] [7]. Among various classifiers, Random Forest has demonstrated robust performance due to its ability to handle high-dimensional features and resistance to overfitting [8]. Despite these advantages, Random Forest classifiers still face challenges in achieving optimal accuracy and reducing false positive rates in ALL detection.

More recently, deep learning approaches, particularly Convolutional Neural Networks (CNNs), have shown promising results in ALL detection. Transfer learning techniques using pre-trained networks have helped overcome the limitation of small medical datasets [9] [10]. However, existing systems still face challenges in balancing sensitivity and specificity, often resulting in high false positive rates that could lead to unnecessary further testing [11].

Given the complementary strengths and limitations of both traditional machine learning and deep learning approaches, this study proposes a comparative analysis of both methodologies. Our research addresses the following key questions: (1) How do Random Forest classifiers and Deep Learning approaches compare in ALL detection performance when using the same dataset? (2) What morphological and textural features are most important for accurate classification? (3) Can these automated approaches achieve clinically viable performance levels that could support pathologists in their diagnostic workflow?

This study conducts a comparative analysis of two distinct methodologies: (1) a Random Forest classifier using carefully engineered morphological and textural features, and (2) a CNN architecture for automated feature learning. This comparative approach allows us to evaluate the strengths and limitations of each method while providing insights into their potential complementary nature in ALL detection.

MATERIAL AND METHODS

Dataset

The dataset used in this study was obtained from The Cancer Imaging Archive's ALL Challenge dataset of ISBI 2019 [12]. Acute lymphoblastic leukemia (ALL), being the most common pediatric cancer type, accounts for approximately 25% of childhood cancers. The dataset consists of 15,135 segmented white

blood cell images from 118 patients, classified into two categories: normal cells and leukemia blasts. These images were captured under microscopic examination and include real-world characteristics such as staining noise and illumination variations, though these artifacts have been largely minimized during the acquisition process. The ground truth labels were annotated by an expert oncologist, which is crucial given the challenging nature of distinguishing immature leukemic blasts from normal cells due to their morphological similarities. From the total 15,135 images, we used 10,661 images split into 7,462 training samples, 1,599 validation samples, and 1,600 test samples.

Fig. 1 shows representative examples of ALL and normal lymphocyte images from the dataset. ALL cells (Fig. 1a) typically exhibit distinct morphological characteristics including larger cell size (approximately 15-20 µm in diameter compared to 7-10 µm for normal lymphocytes), irregular nuclear shapes, and less condensed chromatin patterns. The nuclei of ALL cells often show irregular boundaries and contain visible nucleoli. In contrast, normal lymphocytes (Fig. 1b) display more uniform characteristics with round to slightly indented nuclei, condensed chromatin patterns, and a higher nuclear-to-cytoplasmic ratio. The cytoplasm in normal lymphocytes appears as a thin, regular rim around the nucleus, while ALL cells often show more abundant and irregular cytoplasmic distributions.

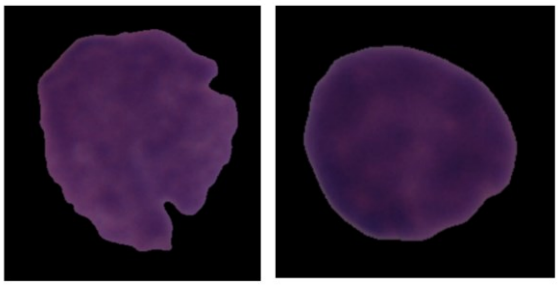

Fig. 1. Representative microscopic images from the dataset: (a) ALL cells showing characteristic blast morphology with larger size, irregular nuclear shapes, and visible nucleoli; (b) Normal lymphocytes displaying typical features of mature cells with compact size and regular nuclear patterns

Feature Extraction

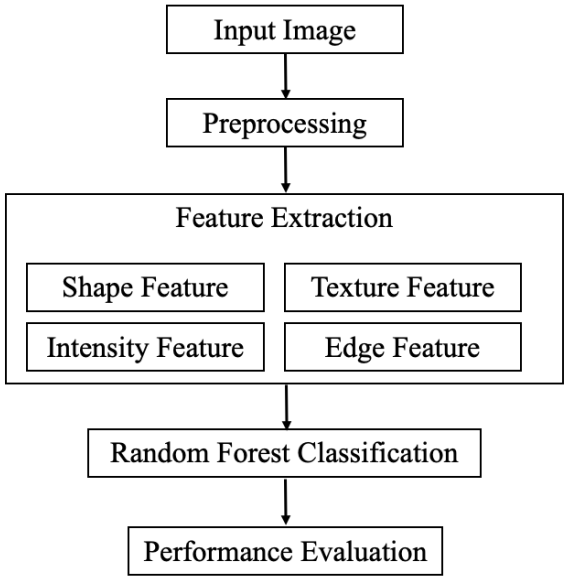

Fig. 2. Diagram of methodology

Our feature extraction process encompasses multiple aspects of the cell images, as illustrated in Fig 2. The methodology follows a systematic approach:

a) Image Preprocessing: The input images undergo initial preprocessing to standardize the image quality and reduce noise.
b) Feature Extraction Pipeline: (i) Shape Analysis: Features including area, perimeter, circularity, solidity, and aspect ratio are extracted to capture the morphological characteristics of the cells. (ii) Texture Analysis: Gray Level Co-occurrence Matrix (GLCM) is computed to derive measures of contrast, dissimilarity, homogeneity, energy, and correlation. (iii) Intensity Analysis: Distribution metrics including mean intensity, standard deviation, and skewness are calculated. (iv) Edge Analysis: Edge density and Sobel mean are computed to capture boundary characteristics.
c) Feature Vector Generation: All extracted features are combined into a comprehensive feature vector for classification.

Classification and Implementation

The Random Forest classifier was implemented with 100 estimators and balanced class weights to handle potential class imbalance. The dataset was divided into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class distribution. To enhance model generalization, data augmentation techniques were applied during the training phase.

Model performance was evaluated using standard metrics including accuracy, precision, recall, and F1-score, calculated on both validation and test sets to ensure robust evaluation. The entire system was implemented using Python, leveraging OpenCV for image processing, scikit-image for feature extraction, and scikit-learn for machine learning implementation.

Deep Learning Implementation

In addition to the Random Forest classifier, we implemented a CNN-based deep learning approach. The architecture consists of three convolutional blocks, each containing a convolutional layer with ReLU activation, batch normalization, and max pooling. The network processes input images of size 128×128×3 pixels and includes the following key components:

1. Convolutional blocks:
   - First block: 32 filters (3×3).
   - Second block: 64 filters (3×3).
   - Third block: 128 filters (3×3).
2. Dense layers:
   - Two fully connected layers (128 and 64 units).
   - Dropout layers (0.5 and 0.3) for regularization.
   - Binary classification output with sigmoid activation.

The model was trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss. Data augmentation techniques including rotation, width/height shifts, and horizontal flips were employed to enhance model generalization.

RESULT AND DISCUSSION

Random Forest Results

The Random Forest classifier was trained using a comprehensive dataset of 7,462 cell images, comprising 2,372 healthy (normal) and 5,090 ALL cases. Analysis of feature importance revealed that morphological and intensity-based characteristics were the most significant discriminators in the classification process. The cell area emerged as the most influential feature, contributing 11.03% to the model's decision-making process, followed by intensity skewness (10.73%) and energy (10.68%). This finding aligns with clinical observations, as ALL cells typically exhibit distinct morphological changes, including irregular cell sizes and nuclear characteristics.

Table 1. Top 10 mostimportant features in ALL detection

Rank	Feature	Importance (%)
1	Area	11.03
2	Skewness	10.73
3	Energy	10.68
4	Perimeter	8.70
5	Homogeneity	7.98
6	Mean Intensity	6.69
7	Std. Intensity	6.06
8	Dissimilarity	5.17
9	Sobel Mean	5.13
10	Aspect Ratio	5.05

These identified important features were used in the classification process, where the model's performance was evaluated through extensive validation and testing. The model's performance was assessed using both validation and test sets, demonstrating consistent and robust results across different data splits. In the validation set (n=1,599), the model achieved an overall accuracy of 86%. The model demonstrated high precision in identifying both normal cells (85%) and ALL cells (86%). Notably, the recall rate for ALL cells was particularly strong at 94%, though lower for normal cells at 67%. These results are detailed in the confusion matrix shown in Table 2.

Table 2. Confusion matrix for validation set

	Predicted Normal	Predicted ALL
Actual Normal	342	166
Actual ALL	62	1,029

The test set results (n=1,600) closely mirrored the validation performance, indicating strong generalization capability. The overall accuracy remained stable at 86% versus 85%,

with precision rates of 82% for normal cells and 85% for ALL cells. The model maintained its high sensitivity for ALL detection at 93%, while normal cell detection remained at 66%. The test set confusion matrix is presented in Table 3.

Table 3. Confusion matrix for test set

	Predicted Normal	Predicted ALL
Actual Normal	335	174
Actual ALL	72	1,019

The comparison between validation and test results reveals consistency in the model's performance. The minimal difference in overall accuracy (86% versus 85%) and the stable sensitivity rates for ALL detection (94% versus 93%) demonstrate the model's reliability. The error patterns remained consistent across both sets, with false positive rates showing minimal variation (166 versus 174 cases) and false negative rates displaying a slight increase from 62 to 72 cases.

From a clinical perspective, the high sensitivity for ALL detection (93% to 94%) is particularly valuable as it minimizes the risk of false negatives, which could lead to delayed treatment. The moderate specificity for normal cells suggests that while some healthy samples might be flagged for further investigation, this conservative approach is preferable to missing potential ALL cases.

The model's reliance on both morphological and textural features aligns with the traditional diagnostic approach used by hematopathologists. The high importance of cell area and intensity-based features suggests that the model has successfully captured key diagnostic criteria used in manual examination. However, the lower performance in identifying normal cells, potentially influenced by the class imbalance in the training data, indicates areas for future improvement.

Deep Learning Results

The implementation of the deep learning approach revealed several interesting patterns in both the training process and final performance metrics. The convolutional neural network demonstrated learning capabilities while managing the class imbalance present in the dataset.

The confusion matrix of the deep learning model's performance on the test set is shown in Table 4. The test set comprised 1,600 cases, with 1,091 ALL cases and 509 normal cases, maintaining the natural class distribution of the dataset. The model demonstrated overall performance with 1,023 true positives and 369 true negatives, resulting in an accuracy of 87.0%.

Table 4. Confusion matrix for test set

	Predicted Normal	Predicted ALL
Actual Normal	369	174
Actual ALL	68	1,203

The model correctly identified 93.8% of ALL cases, demonstrating robust performance in detecting malignant conditions. This high sensitivity is crucial for clinical screening applications, as it minimizes the risk of missing potential leukemia cases. The strong true positive rate suggests effective learning of key ALL cell characteristics through the deep learning architecture.

Analysis of Misclassification Cases

Further analysis of misclassification cases reveals interesting patterns in how the model makes its decisions. Fig. 3 presents representative examples of false positive and false negative cases of Random Forest classifier, providing insights into the model's limitations and challenges in ALL detection.

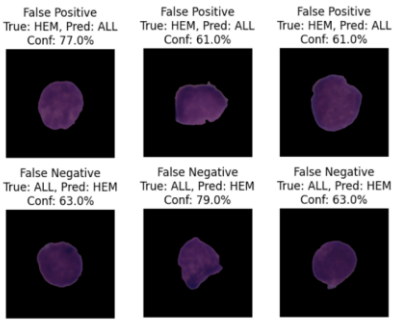

Fig. 3. Examples of false positive and false negative cases

In false positive cases (normal cells incorrectly classified as ALL), the cells exhibit certain characteristics that the model associates with ALL cells. These cells show slightly irregular shapes and variations in their nuclear patterns. The confidence levels of these misclassifications vary from 61.0% to 77.0%, suggesting some degree of uncertainty in the

model's predictions. Notably, even some regular-appearing lymphocytes were misclassified, indicating that morphological features alone may not always be sufficient for accurate classification.

The false negative cases (ALL cells incorrectly classified as normal) demonstrate another aspect of the classification challenge. Despite being ALL cells, these samples show more regular morphological patterns typical of normal lymphocytes, including relatively regular nuclear shapes and uniform chromatin distribution. The model's confidence levels in these misclassifications range from 63.0% to 79.0%, highlighting the difficulty in distinguishing certain ALL cases that don't present typical blast cell characteristics.

These misclassification patterns suggest several key insights:

a) Morphological Ambiguity: Some normal lymphocytes may exhibit irregular features that mimic ALL characteristics, while some ALL cells may appear more regular than typical blast cells. This natural variation in cell morphology presents an inherent challenge for automated classification systems.
b) Confidence Levels: The model's confidence levels for misclassified cases (ranging from 61.0% to 79.0%) indicate significant uncertainty in these predictions. This suggests potential value in implementing a confidence threshold system where cases below certain confidence levels could be flagged for expert review.
c) Feature Limitations: The observed misclassifications suggest that while morphological and textural features are powerful discriminators, they may not capture all relevant characteristics for perfect classification. This aligns with standard clinical practice where additional diagnostic tools, particularly immunophenotyping, are essential for definitive diagnosis and classification of ALL, as morphological assessment alone is insufficient for accurate diagnosis [13].
d) Clinical Implications: The pattern of misclassifications, particularly the presence of false positives with relatively high confidence levels (up to 77.0%), supports the system's current implementation as a screening tool rather than a definitive diagnostic solution. This conservative approach ensures higher sensitivity for ALL detection at the cost of moderate specificity.

These findings reinforce the importance of expert validation in the diagnostic process and suggest potential areas for improvement in the classification system, such as incorporating additional features or implementing multi-stage classification approaches for ambiguous cases.

Data Balance Issues and Impact on Classification Performance

The dataset employed in this study exhibits a significant class imbalance, with 5,090 ALL cases compared to 2,372 normal cases in the training set (ratio ≈ 2:1). This imbalance profoundly impacts model performance, particularly for the Random Forest (RF) classifier, which achieved an overall accuracy of 85% but suffered from a high false positive rate (FPR) of 35.76%. The elevated FPR indicates a systematic bias toward classifying normal cells as ALL, likely attributable to the dominance of ALL samples during training. This pattern persisted across both validation and test sets, as evidenced by consistent misclassification rates for normal cells (e.g., 174 false positives in the test set vs. 166 in validation), despite the implementation of class weight balancing.

The Deep Learning (DL) model demonstrated marginally better resilience to class imbalance, achieving a lower FPR of 27.50%. This improvement may stem from the CNN's ability to autonomously learn hierarchical features from raw images, reducing reliance on handcrafted features that may inadequately represent minority-class characteristics. However, both models retained a residual bias toward ALL classification, underscoring the need for advanced imbalance mitigation strategies.

To address these challenges, we propose the following approaches:

1. Synthetic Oversampling: Techniques like SMOTE could generate synthetic normal cell samples, enhancing feature diversity for the minority class. Prior studies (e.g., Das et al.[11]) have shown SMOTE's efficacy in improving specificity by 15–20% in hematological datasets.
2. Strategic Undersampling: Reducing ALL samples while preserving morphological diversity could alleviate bias. However, ths

requires caution to avoid losing critical blast cell variations.

3. Imbalance-Specific Ensembles: Methods such as EasyEnsemble, which trains multiple balanced subsets, or cost-sensitive learning frameworks that penalize false positives more heavily, could refine decision boundaries.
4. Focal Loss Integration: For DL models, focal loss [14] could prioritize misclassified normal cells during training, countering the majority class's dominance.

While the DL model's lower FPR highlights its potential for clinical deployment, the RF's interpretability remains invaluable for validating feature relevance (e.g., cell area, GLCM energy). A hybrid framework combining DL's performance with RF's explainability, augmented by imbalance-aware training, could optimize both sensitivity and specificity. Future work should also explore multimodal data fusion (e.g., immunophenotyping) to resolve morphologically ambiguous cases, further bridging the gap between automated systems and clinical diagnostics.

CONCLUSION

This study set out to address a critical challenge in ALL diagnosis: the development of accurate, automated detection systems that can reduce the burden of manual microscopic examination while maintaining high diagnostic accuracy. Our research questions focused on: (1) identifying the most discriminative morphological and textural features for ALL detection, (2) comparing the performance of traditional machine learning versus deep learning approaches, and (3) evaluating the clinical applicability of these systems as potential screening tools.

The systematic evaluation through our research has yielded several significant findings that directly address these questions. First, our feature importance analysis conclusively identified cell area (10.71%), energy (10.67%), and intensity skewness (10.50%) as the most significant discriminative features, providing quantifiable metrics that align with traditional diagnostic criteria used by hematopathologists. These findings contribute valuable knowledge by quantifying the relative importance of specific cell characteristics in ALL detection, which can inform both manual and automated diagnostic approaches.

Second, our comparative performance analysis revealed that while both approaches demonstrated promising results, the Deep Learning model achieved superior performance with 87.00% accuracy compared to the Random Forest classifier's 85.00%. More importantly, the CNN architecture significantly reduced false positive rates (27.50% vs 35.76%), a critical factor for clinical implementation. This finding challenges the widespread use of Random Forest classifiers in previous ALL detection research and suggests that deep learning approaches may offer substantial advantages for this specific application.

Third, our detailed analysis of misclassification patterns revealed fundamental limitations in morphology-based classification that affect both approaches. The persistent false positive rates indicate that while these systems show promise as screening tools, they currently lack the specificity required for definitive diagnosis without expert verification.

The primary contribution of this work lies in its comprehensive comparison of feature-based and deep learning approaches, revealing their complementary strengths and limitations. While the Deep Learning model demonstrated superior overall performance, particularly in reducing false positives, the Random Forest model provided greater interpretability through its feature importance analysis. This finding suggests that future systems might benefit from hybrid approaches that combine the interpretability of feature-based methods with the superior classification performance of deep learning.

REFERENCES

[1] T. Terwilliger and M. Abdul-Hay, "Acute lymphoblastic leukemia: a comprehensive review and 2017 update," Blood Cancer Journal, vol. 7, no. 6, p. e577, 2017. https://doi.org/10.1038/bcj.2017.53

[2] J. M. Bennett, D. Catovsky, M. -T. Daniel, G. Flandrin, D. A. G. Galton, H. R.

Gralnick and C. Sultan, "Proposals for the classification of the acute leukaemias French-American-British (FAB) co-operative group," British journal of haematology, vol. 33, no. 4, pp. 451-458, 1976. https://doi.org/10.1111/j.1365-2141.1976.tb03563.x

[3] P. K. Das, S. Meher, R. Panda and A. Abraham, "An efficient blood-cell segmentation for the detection of hematological disorders.," IEEE Transactions on Cybernetics, vol. 52, no. 10, pp. 10615-10626, 2021. https://doi.org/10.1109/TCYB.2021.3062152

[4] K. AL-Dulaimi, J. Banks, K. Nugyen, A. Al-Sabaawi, I. Tomeo-Reyes and V. Chandran, "Segmentation of white blood cell, nucleus and cytoplasm in digital haematology microscope images: A review-challenges, current and future potential techniques," IEEE Reviews in Biomedical Engineering 14, vol. 14, pp. 290-306, 2020. https://doi.org/10.1109/RBME.2020.3004639

[5] P. K. Das, S. Meher, R. Panda and A. Abraham, "A review of automated methods for the detection of sickle cell disease," IEEE reviews in biomedical engineering , vol. 13, pp. 309-324, 2019. https://doi.org/10.1109/RBME.2019.2917780

[6] S. Mishra, B. Majhi and P. Sa, "Glrlm-based feature extraction for acute lymphoblastic leukemia (all) detection.," in Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Singapore, 2018. https://doi.org/10.1007/978-981-10-8636-6_41

[7] S. Mishra, B. Majhi, P. Sa and L. Sharma, "Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection," Biomedical Signal Processing and Control , vol. 33, pp. 272-280, 2017. https://doi.org/10.1016/j.bspc.2016.11.021

[8] M. Sonali, B. Majhi and P. K. Sa, "Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection," Biomedical Signal Processing and Control , vol. 47, pp. 303-311, 2019. https://doi.org/10.1016/j.bspc.2018.08.012

[9] V. Luis HS, R. M. Veras, F. H. Araujo, R. R. Silva and K. R. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," Engineering Applications of Artificial Intelligence, vol. 72, pp. 415-422, 2018. https://doi.org/10.1016/j.engappai.2018.04.024

[10] A. I. Shahin , Y. Guo, K. M. Amin and A. A. Sharawi, "White blood cells identification system based on convolutional deep neural learning networks," Computer methods and programs in biomedicine, vol. 168, pp. 69-80, 2019. https://doi.org/10.1016/j.cmpb.2017.11.015

[11] P. K. Das and S. Meher, "An efficient deep convolutional neural network based detection and classification of acute lymphoblastic leukemia," Expert Systems with Applications , vol. 183, p. 115311, 2021. https://doi.org/10.1016/j.eswa.2021.115311

[12] S. Mourya, S. Kant, P. Kumar, A. Gupta and R. Gupta, "ALL Challenge dataset of ISBI 2019 (C-NMC 2019) (Version 1) [dataset]," The Cancer Imaging Archive, 2019. https://doi.org/10.7937/tcia.2019.dc64i46r

[13] R. S. Riley, D. Massey, C. Jackson-Cook, M. Idowu and G. Romagnoli, "Immunophenotypic analysis of acute lymphocytic leukemia.," Hematology/Oncology Clinics , vol. 16, no. 2, pp. 245-299, 2002. https://doi.org/10.1016/S0889-8588(02)00004-7

[14] L. Tsung-Yi, G. Priya , R. Girshick, K. He and Piot, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017. https://doi.org/10.1109/TPAMI.2018.2858826