Afro-Asian Journal of Scientific Research (AAJSR)
المجلة الأفروآسيوية للبحث العلمي
E-ISSN: 2959-6505
Volume 4, Issue 3, 2026
Page No: 66-72
Website: https://aajsr.com/index.php/aajsr/index
SJIFactor 2024: 5.028 ISI 2025: 0.915 معامل التأثير العربي (AIF) 2025: 0.76

Deep Feature Selection and AutoML Ensembles for Acute Lymphoblastic Leukemia Cell Classification

Akram Khalil Ghaidan[1*], Mustafa Mohammed Alkharsh[2], Nourah Mohammed Abdulmjid[3], Rafea M. Almejrab[4], Ahmad Allafee Moftah[5]
[1]Computer Science Department, Faculty of Education, University of Derna, Al Qubbah, Libya
[2,3,4,5]Computer Science Department, College of Computer Technology, Benghazi, Libya

انتقاء السمات العميقة وتجميعات التعلم الآلي المؤتمت لتصنيف خلايا ابيضاض الدم اللمفاوي الحاد

أكرم خليل غيضان[1*]، مصطفى محمد الخرش[2]، نورة محمد عبد المجيد[3]، رافع محمد المجراب[4]، أحمد اللافي مفتاح[5]
[1]قسم الحاسوب، كلية التربية، جامعة درنة، القبة، ليبيا
[2,3,4,5] قسم علوم الحاسوب، كلية تقنيات الحاسوب، بنغازي، ليبيا

*Corresponding author: akram_kalil2010@yahoo.com

Received: April 24, 2026	Accepted: July 04, 2026	Published: July 18, 2026

Abstract:
Acute lymphoblastic leukemia (ALL) classification from microscopic peripheral blood smear images remains challenging because of staining variability, inter-class morphological similarity, and class imbalance. This paper presents a four-class classification framework for distinguishing benign hematogones, Early Pre-B ALL, Pre-B ALL, and Pro-B ALL, rather than the easier binary benign-versus-malignant screening task. The framework uses VGG19 as a fixed deep feature extractor, Mutual Information (MI) for label-aware feature selection, and H2O AutoML stacked ensembles for automated classifier construction. On the held-out 80:20 test partition, the MI-based configuration achieved 94.14% overall accuracy with a Wilson 95% confidence interval of 92.02%-95.72% and a macro-F1 of approximately 0.931. Class-level analysis showed strong performance for Pre-B and Pro-B cells, while benign hematogones remained the most difficult class because of visual similarity with early malignant cells. The results support the value of combining deep feature embeddings with automated model selection for ALL image classification, but the system is not yet a substitute for clinical diagnosis without patient-level and multi-center external validation.

Keywords: Acute Lymphoblastic Leukemia, AutoML, VGG19, Mutual Information, H2O, Blood Smear Images, Medical Image Classification, Feature Selection.

الملخص:
يظل تصنيف ابيضاض الدم اللمفاوي الحاد (ALL) اعتمادًا على الصور المجهرية لمسحات الدم المحيطي مهمةً صعبة بسبب تباين الصبغ، والتشابه المورفولوجي بين الفئات، وعدم توازنها. تقدم هذه الورقة إطارًا رباعي الفئات للتمييز بين الخلايا السلفية اللمفاوية الحميدة (الهيماتوغونات)، وابيضاض الدم اللمفاوي الحاد من النمط البائي المبكر (Pre-B Early ALL)، والنمط Pre-B ALL، والنمط Pro-B ALL، بدلًا من مهمة الفرز الثنائية الأسهل بين الحالات الحميدة والخبيثة. ويستخدم الإطار نموذج VGG19 مستخرجًا ثابتًا للسمات العميقة، والمعلومات المتبادلة (MI) لانتقاء السمات المراعية للفئات، وتجميعات H2O AutoML المكدسة لبناء المصنفات تلقائيًا. وفي مجموعة الاختبار المحجوزة وفق تقسيم 80:20، حقق الإعداد القائم على المعلومات المتبادلة دقة كلية بلغت 94.14%، بفاصل ثقة ويلسون قدره 95% يتراوح

بين 92.02% و95.72%، وبقيمة Macro-F1 تقارب 0.931. وأظهر التحليل على مستوى الفئات أداءً قويًا لخلايا Pre-B وPro-B، في حين ظلت الخلايا السلفية اللمفاوية الحميدة الفئة الأصعب بسبب تشابهها البصري مع الخلايا الخبيثة المبكرة. وتدعم النتائج قيمة الجمع بين تمثيلات السمات العميقة والاختيار الآلي للنماذج في تصنيف صور ابيضاض الدم اللمفاوي الحاد؛ غير أن النظام لا يُعد بديلًا عن التشخيص السريري دون تحقق خارجي على مستوى المرضى وفي مراكز متعددة.

الكلمات المفتاحية: ابيضاض الدم اللمفاوي الحاد، التعلم الآلي المؤتمت، VGG19، المعلومات المتبادلة، H2O، صور مسحات الدم، تصنيف الصور الطبية، انتقاء السمات.

Introduction:

Acute lymphoblastic leukemia (ALL) is an aggressive hematological malignancy characterized by abnormal proliferation of immature lymphoid precursors and substantial biological and morphological heterogeneity [1]. Microscopic examination of blood or bone-marrow smears remains an important diagnostic-support activity, but manual visual assessment is time-consuming and may be affected by observer variability, staining quality, overlapping cells, and limited specialist availability [2,3].

Deep learning has improved medical image analysis by learning hierarchical visual representations directly from images, and recent ALL studies have shown the value of transfer learning and deep feature extraction on blood-smear data [4,5]. However, end-to-end convolutional neural networks often require substantial data, careful tuning, and specialized expertise. Transfer learning can reduce these requirements by reusing feature representations learned from large-scale image collections, while Automated Machine Learning (AutoML) can reduce the manual burden of classifier selection, hyperparameter tuning, and ensemble construction [6-9].

This paper focuses on a narrow and defensible contribution: evaluating whether deep features extracted from VGG19, reduced through feature-selection strategies, and classified using H2O AutoML can improve multi-class ALL cell classification. The emphasis is on reproducible class-level evaluation, transparent limitations, and avoiding unsupported claims of clinical deployment.

A key design decision in this manuscript is to treat the problem as a four-class ALL classification task rather than as a binary screening task. Binary benign-versus-malignant experiments can naturally report higher accuracy because the model only separates two broad categories. The four-class setting used here is more challenging and clinically more informative because it evaluates whether the model can distinguish benign hematogones from Early Pre-B, Pre-B, and Pro-B ALL subtypes.

Contributions:

The main contributions are: (1) a four-class ALL cell classification pipeline based on VGG19 feature embeddings and H2O AutoML ensembles; (2) a comparative evaluation of Mutual Information and PCA as feature-selection or dimensionality-reduction strategies; (3) reporting of class-level precision, recall, and F1-score rather than accuracy alone; and (4) an explicit discussion of methodological validity threats, including class imbalance, single-center data, and the need for patient-level and external validation.

Related Work:

Recent leukemia-image research has increasingly moved from handcrafted descriptors toward transfer learning, deep feature extraction, segmentation-assisted pipelines, and ensemble classification. Ghaderzadeh et al. evaluated multiple CNN backbones for B-ALL diagnosis and subtype classification [2], while Rastogi et al. introduced a deep feature extractor for acute-leukemia microscopy [4]. Das et al. combined transfer learning with a discriminative classification layer for small leukemia datasets [5], and later work explored neural feature ensembles, optimized feature selection, and joint classification-segmentation pipelines [7], [8]. Park et al. further demonstrated deep-learning differentiation of myeloblasts and lymphoblasts using peripheral blood-cell images [9].

Although many recent studies still emphasize binary malignant-versus-normal detection, current research increasingly recognizes the difficulty of subtype discrimination, data imbalance, morphology overlap, and cross-dataset generalization [7], [10,11]. Unlike binary screening studies, this manuscript evaluates four classes: benign hematogones and three malignant ALL subtypes. This is a stricter and clinically more informative task, although it is also more sensitive to class imbalance and subtype similarity.

Materials and Methods:

1. Dataset:

The experiments used the Acute Lymphoblastic Leukemia image dataset prepared at Taleqani Hospital, Tehran and described in the recent B-ALL subtype-classification literature [2]. The dataset contains 3,256 peripheral blood smear images from 89 suspected ALL patients. The four classes used

in this paper are benign hematogones and three malignant subtypes: Early Pre-B ALL, Pre-B ALL, and Pro-B ALL. The class distribution is summarized in Table 1.

Table (1): Dataset distribution used for multi-class ALL classification

Class group	Subtype	Patients	Images	Image size
Benign	Hematogones	25	504	1024 x 768
Malignant	Early Pre-B ALL	20	985	1024 x 768
Malignant	Pre-B ALL	21	963	1024 x 768
Malignant	Pro-B ALL	23	804	1024 x 768

2. Preprocessing:

Images were preprocessed through segmentation and resizing. K-Means clustering was used to separate relevant cellular regions from the background, and HSV-based masking helped emphasize stained nuclear regions. Resizing ensured consistent input dimensions for VGG19 feature extraction. These steps were designed to reduce background noise and improve representation of cell morphology. Systematic data augmentation (such as rotation, flipping, brightness jittering, or stain perturbation) was not evaluated in the reported experiments and is therefore treated as a limitation rather than an implemented step.

3. Deep Feature Extraction:

VGG19 pre-trained on ImageNet was used as a fixed deep feature extractor rather than an end-to-end fine-tuned model, consistent with recent leukemia studies that benchmark pre-trained CNNs as feature extractors [2], [4], [8]. Fully connected classification layers were replaced by global average pooling, producing a 512-dimensional embedding for each image. This fixed-feature design was selected to reduce trainable parameters, limit overfitting risk on a moderate-size single-center dataset, and focus the contribution on whether AutoML ensembles can exploit compact deep representations. Fine-tuning VGG19 or lighter modern architectures is therefore treated as a natural extension rather than as a claim of the present experiment.

4. Feature Selection and Reduction:

Two alternatives were evaluated. Mutual Information selected the most discriminative features with respect to the class labels, while PCA reduced dimensionality by retaining variance without using labels. Following the experimental protocol, the MI configuration retained the top 200 features from the 512-dimensional VGG19 embedding. This value was used as a fixed empirical setting to balance compactness and class discrimination; future extended experiments will examine a systematic sensitivity curve over different numbers of selected features.

5. AutoML Classification:

H2O AutoML was used to train and rank multiple models, including gradient boosting machines, random forests, generalized linear models, deep neural networks, XGBoost where available, and stacked ensembles. Recent healthcare AutoML reviews show that such systems can streamline algorithm selection, hyperparameter optimization, and ensemble construction, while still requiring careful interpretation and transparent reporting [6]. The AutoML search was configured as a time-bounded model-selection process with cross-validation used internally for candidate ranking. In the completed run, the best-performing learner was a stacked ensemble, which combined complementary base learners through a meta-learner. The implementation record available from the source thesis documents the environment only at a high level: Python on a personal computer. Accordingly, this manuscript does not infer exact CPU, GPU, RAM, package-version, random-seed, or runtime configurations that were not recorded. To improve reproducibility, the documented workflow, dataset counts, feature-extraction setting, feature-selection alternatives, split setting, and reported metrics are stated explicitly; future replications should archive code, package versions, hardware specifications, random seeds, and split files.

6. Evaluation Protocol:

The main reported experiment used an image-level stratified 80:20 train-test split, producing 631 test images for metric reporting. The split preserved the four-class distribution as closely as possible. Current medical-imaging AI reporting guidance recommends explicit disclosure of partitioning level, model-selection procedures, software details, and statistical uncertainty [12,13]. For leakage-controlled reimplementation, feature selection should be fitted using the training partition only before applying the learned feature subset to the held-out test set. Because the original thesis report does not provide enough implementation detail to independently audit this point, feature-selection leakage is retained as a threat to validity. Furthermore, although the public dataset contains images from 89 suspected patients, patient identifiers were not used to construct a strict patient-level split in the reported experiment. Therefore, potential patient-level leakage remains a major threat to validity, and the 94.14%

result is interpreted as an image-level estimate that requires confirmation under patient-level and external validation protocols. Metrics include overall accuracy, per-class precision, recall, F1-score, macro-F1, and a Wilson 95% confidence interval for accuracy.

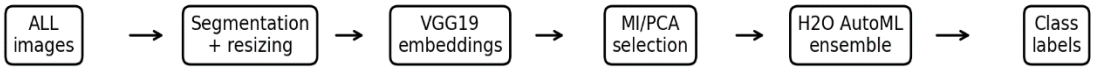

Fig. (1): Proposed VGG19 deep-feature and AutoML ensemble pipeline for four-class ALL cell classification.

Results:

1. Multi-Class Confusion Matrix:

Table II reports the held-out test confusion matrix for the MI-based VGG19-AutoML configuration under the 80:20 split. The model correctly classified 594 out of 631 test images, corresponding to 94.14% overall accuracy. Most errors occurred in the benign and Early Pre-B classes, which are morphologically closer and more sensitive to staining and segmentation variability.

Table (2): Confusion matrix for MI + VGG19 + H2O AutoML under the 80:20 split.

Actual / Predicted	Benign	Early	Pre	Pro	Error rate
Benign	77	13	2	0	15 / 92
Early	13	181	0	1	14 / 195
Pre	0	5	188	1	6 / 194
Pro	1	0	1	148	2 / 150
Total predicted	91	199	191	150	37 / 631

2. Class-Level Metrics:

The class-level metrics in Table III confirm that the model performs strongly for Pre-B and Pro-B classes, while benign-cell discrimination remains the main source of error. The 95% Wilson confidence interval for overall accuracy is 92.02% to 95.72%, calculated from 594 correct predictions out of 631 test samples.

Table (3): Class-level performance of the proposed MI-AutoML configuration

Metric	Benign	Early Pre-B	Pre-B	Pro-B
Precision	0.846	0.909	0.984	0.993
Recall	0.837	0.928	0.969	0.986
F1-score	0.841	0.918	0.976	0.990

3. Comparison of Feature Selection Strategies:

Table IV summarizes the main internal comparisons. Mutual Information outperformed PCA under both 70:30 and 80:20 split settings. This suggests that label-aware feature selection preserved discriminative information more effectively than variance-driven dimensionality reduction for this dataset.

Table (4): Internal comparison of feature-selection strategies

Configuration	Split	Overall accuracy	Main observation
VGG19 + MI + H2O AutoML	70:30	93.10%	Strong overall performance; benign recall lower than malignant subtypes
VGG19 + PCA + H2O AutoML	70:30	86.20%	Lower class discrimination, especially benign
VGG19 + MI + H2O AutoML	80:20	94.14%	Best reported configuration
VGG19 + PCA + H2O AutoML	80:20	~89.80%	Below MI; above PCA 70:30

4. Baseline Comparison:

Compared with conventional machine-learning baselines reported on the same feature space, the MI-H2O AutoML configuration achieved higher accuracy than Lazy Predict models such as

LGBMClassifier and SVC, which both achieved approximately 92% accuracy. This improvement is meaningful because the proposed configuration also reports class-level stability rather than relying on a single aggregate metric.

Table (5): Comparison with selected conventional baselines

Model	Accuracy	F1-score	Time taken
MI + H2O AutoML	94.14%	Macro-F1 approx. 0.931	AutoML run
LGBMClassifier	0.92	0.92	9.74
SVC	0.92	0.92	0.54
SGDClassifier	0.90	0.90	0.33
Logistic Regression	0.90	0.90	0.26
Random Forest	0.89	0.89	4.18

Discussion:

The results support the value of combining VGG19 deep embeddings with MI-based feature selection and AutoML ensembles in a four-class ALL setting. MI performed better than PCA because it explicitly uses class-label information and can retain discriminative features even when their variance is not dominant. Although the 94.14% accuracy is lower than what may be observed in binary benign-versus-malignant screening, it is more defensible for this manuscript because the reported task separates four visually similar categories rather than two broad groups.

The benign class is the most challenging category. In the 80:20 MI experiment, 13 benign cases were classified as Early Pre-B and two as Pre-B. This indicates that the model may confuse benign hematogones with early malignant morphology. From a diagnostic-support perspective, such confusion deserves careful attention because false alarms and missed benign cases can affect downstream clinical workflow. Future refinement should investigate class weighting, balanced sampling, additional benign examples, and augmentation strategies to improve benign-cell discrimination.

The results do not support readiness for direct clinical deployment. They are promising for decision-support research and controlled prototyping, but external testing on independent centers, scanners, and staining conditions is required. This cautious interpretation is consistent with recent CLAIM and TRIPOD+AI guidance, which emphasizes patient-disjoint partitioning, uncertainty reporting, reproducibility, and external evaluation before clinical translation [12,13]. A future version can be strengthened substantially by adding Grad-CAM or another explainability analysis to show whether the model focuses on diagnostically meaningful cellular regions; recent augmentation research has also demonstrated the value of combining attention and Grad-CAM in ALL classification [10].

Threats to Validity and Limitations:

- Single-center dataset: all images come from one source, which limits generalization to other laboratories, microscopes, staining protocols, and patient populations.
- Class imbalance: malignant classes outnumber benign samples, so accuracy may overstate performance unless class-level metrics are reported.
- No systematic data augmentation: the reported experiments did not evaluate augmentation strategies such as rotation, flipping, brightness or stain variation. This may limit robustness to acquisition and staining variability.
- Patient-level leakage risk: the reported 80:20 split was image-level rather than a confirmed patient-level split. If multiple images from the same patient appear in both training and testing partitions, performance may be optimistic. A strict patient-level split is required for stronger clinical evidence.
- Feature-selection leakage risk: the original report does not fully document whether MI was fitted strictly within the training partition. Applying feature selection before splitting can inflate test performance, so future reproducible experiments should fit MI inside each training fold or training split only.
- Limited explainability: the present manuscript reports classification metrics but does not yet include visual explanation such as Grad-CAM. Such visual evidence would help verify whether the model attends to diagnostically relevant cellular regions rather than background or staining artifacts, as recommended by recent medical-imaging AI reporting guidance [12].
- External testing: no independent dataset such as C-NMC was used for final external testing [3].
- Limited implementation metadata: the source thesis documents only Python and a personal-computer environment. Exact software versions, hardware specifications, random seeds, and split files were not available for independent verification, limiting full reproducibility.

Data Availability and Ethical Considerations:

This secondary analysis used a publicly available benchmark image dataset with expert labels [2]. No new patient samples were collected in this study. Any reuse must follow the data-use terms of the

dataset provider. The original experimental record did not include publicly archived code, exact split files, or complete environment logs; therefore, future replications should release these materials to strengthen reproducibility in line with current reporting recommendations [12,13].

Manuscript Origin and Author Consent:

This manuscript is derived from the master's thesis work of Ahmad Allafee Moftah, entitled "Using Image Processing Methods to Diagnose Leukemia Cells from Microscopic Images," conducted at the Libyan Academy - Benghazi Branch [14] under the supervision of Akram Khalil Ghaidan. Ahmad Allafee Moftah has approved the reuse of the thesis methodology, experimental results, tables, and findings for publication and is listed as the first author because the core methodology and results originate from his master's thesis. Akram Khalil Ghaidan is included as a co-author in recognition of supervisory and scientific contribution to the original thesis work. The manuscript has been substantially reformulated as a focused research article, with revised framing, limitations, reporting, and discussion for scholarly submission.

Author Contributions:

Ahmad Allafee Moftah conducted the original thesis study and approved the publication use of the thesis-derived methodology and results. Akram Khalil Ghaidan supervised the original thesis work and contributed scientific oversight. Rafea M. Almejrab contributed to manuscript reformulation, academic revision, critical review, and preparation of the article for submission. All authors reviewed and approved the final manuscript.

Conclusion:

This paper presented a focused multi-class ALL cell classification framework combining fixed VGG19 deep features, Mutual Information feature selection, and H2O AutoML stacked ensembles. The best configuration achieved 94.14% held-out image-level accuracy with a 95% confidence interval of 92.02%-95.72%. The study is stronger than a binary screening report because it evaluates benign, Early Pre-B, Pre-B, and Pro-B classes separately. Nevertheless, the result is best viewed as a controlled image-level experiment rather than clinical validation, especially because patient-level splitting, complete implementation metadata, and feature-selection leakage auditing were not fully available from the original record. Future work will prioritize patient-level splitting, multi-center testing, VGG19 fine-tuning and lightweight alternatives, feature-count sensitivity analysis, systematic data augmentation, class-balancing strategies, full code/environment archival, and Grad-CAM-based explainability.

References:

1. R. Alaggio et al., “The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms,” Leukemia, vol. 36, no. 7, pp. 1720–1748, 2022, doi: 10.1038/s41375-022-01620-2.
2. M. Ghaderzadeh, M. Aria, A. Hosseini, F. Asadi, D. Bashash, and H. Abolghasemi, “A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images,” International Journal of Intelligent Systems, vol. 37, pp. 5113–5133, 2022, doi: 10.1002/int.22753.
3. R. Gupta, S. Gehlot, and A. Gupta, “C-NMC: B-lineage acute lymphoblastic leukaemia: A blood cancer dataset,” Medical Engineering & Physics, vol. 103, Art. no. 103793, 2022, doi: 10.1016/j.medengphy.2022.103793.
4. P. Rastogi, K. Khanna, and V. Singh, “LeuFeatx: Deep learning-based feature extractor for the diagnosis of acute leukemia from microscopic images of peripheral blood smear,” Computers in Biology and Medicine, vol. 142, Art. no. 105236, 2022, doi: 10.1016/j.compbiomed.2022.105236.
5. P. K. Das, B. Sahoo, and S. Meher, “An efficient detection and classification of acute leukemia using transfer learning and orthogonal softmax layer-based model,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 3, pp. 1817–1828, 2023, doi: 10.1109/TCBB.2022.3218590.
6. H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, “Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare,” Medicine Advances, vol. 2, no. 3, pp. 205–237, 2024, doi: 10.1002/med4.75.
7. M. Awais, R. Ahmad, N. Kausar, A. I. Alzahrani, N. Alalwan, and A. Masood, “ALL classification using neural ensemble and memetic deep feature optimization,” Frontiers in Artificial Intelligence, vol. 7, Art. no. 1351942, 2024, doi: 10.3389/frai.2024.1351942.
8. D. S. Ametefe et al., “Automatic classification and segmentation of blast cells using deep transfer learning and active contours,” International Journal of Laboratory Hematology, vol. 46, no. 5, pp. 837–849, 2024, doi: 10.1111/ijlh.14305.
9. [9] S. Park, Y. H. Park, J. Huh, S. M. Baik, and D. J. Park, “Deep learning model for differentiating acute myeloid and lymphoblastic leukemia in peripheral blood cell images via myeloblast and

lymphoblast classification," Digital Health, vol. 10, Art. no. 20552076241258079, 2024, doi: 10.1177/20552076241258079.
10. T. Mustaqim, C. Fatichah, N. Suciati, T. Obi, and J.-S. Lee, "FocusAugMix: A data augmentation method for enhancing acute lymphoblastic leukemia classification," Intelligent Systems with Applications, vol. 26, Art. no. 200512, 2025, doi: 10.1016/j.iswa.2025.200512.
11. W. Hussain Shah, S. Rafia Fatima, R. Jaimes-Reátegui, D. E. Arévalo-Simental, P. T. Villalobos-Gutiérrez, and A. N. Pisarchik, "A systematic review of machine and deep learning techniques for acute lymphoblastic leukemia diagnosis," Artificial Intelligence in Medicine, vol. 176, Art. no. 103393, 2026, doi: 10.1016/j.artmed.2026.103393.
12. A. S. Tejani et al., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," Radiology: Artificial Intelligence, vol. 6, no. 4, Art. no. e240300, 2024, doi: 10.1148/ryai.240300.
13. G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.
14. A. A. Moftah, "Using Image Processing Methods to Diagnose Leukemia Cells from Microscopic Images," M.S. thesis, Department of Computer Science, School of Basic Sciences, Libyan Academy—Benghazi Branch, Benghazi, Libya, 2025.