Marrow Clues in the Detection of Leukemia Using Machine Learning

Thirupurasundari D R * , Nomula Saharika , Parise Chaitanya Krishna Sandeep , Pasupuleti Sravani , Pathan Jaheer Khan

Posted Date: 20 March 2026

doi: 10.20944/preprints202603.1631.v1

Keywords: leukemia detection; acute lymphoblastic leukemia; ResNet50; transfer learning; Grad-CAM; GAN augmentation; Watershed segmentation; deep learning; medical imaging; bone marrow

 doi:10.20944/preprints202603.1631.v1

Article

Marrow Clues in the Detection of Leukemia Using Machine Learning

Thirupurasundari D R *, Nomula Saharika, Parise Chaitanya Krishna Sandeep, Pasupuleti Sravani and Pathan Jaheer Khan

Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research (BIHER), Chennai, India

* Correspondence: thirupurasundari@bharathuniv.ac.in

Abstract

Acute Lymphoblastic Leukemia (ALL) is the most prevalent form of childhood cancer, accounting for approximately 25% of all paediatric malignancies worldwide. Conventional diagnosis relies on a trained hematopathologist manually examining Giemsa-stained bone marrow smear slides, a process that is both time-consuming and inherently subjective, with inter-observer variability of up to 20% reported in clinical literature. This paper presents Marrow-Find, a fully integrated AI-powered web application designed to automate the classification, visual explanation, and quantification of leukemia cells from bone marrow microscopic images. The system employs a ResNet50 deep convolutional neural network, pre-trained on ImageNet and fine-tuned on the C-NMC 2019 and ALL-IDB benchmark datasets, to classify bone marrow cells into four clinically significant categories: Benign, Pre-B ALL, Pro-B ALL, and Early Pre-B ALL. The model achieves an overall test accuracy of 94.8% with a weighted F1-score of 0.954 and a Cohen's Kappa of 0.931. To address class imbalance, a Conditional Generative Adversarial Network (cGAN) generates realistic synthetic bone marrow cell images for minority classes. Gradient-weighted Class Activation Mapping (Grad-CAM) produces visual heatmap overlays on each prediction, enabling pathologists to verify the morphological features influencing the model's decision. A Watershed-based segmentation algorithm automatically delineates individual cell boundaries and counts blast cells per image. All five components are integrated into a Flask web application with secure user authentication, prediction history, PDF report generation with QR codes, and an AI chatbot, making specialist-level leukemia diagnostics accessible from any standard web browser.

Keywords: leukemia detection; acute lymphoblastic leukemia; ResNet50; transfer learning; Grad-CAM; GAN augmentation; Watershed segmentation; deep learning; medical imaging; bone marrow

1. Introduction

Cancer remains one of the most serious public health challenges of the modern era, and leukemia occupies a particularly distressing position because of the disproportionate burden it places on children and young adults. Acute Lymphoblastic Leukemia (ALL) is the most frequently diagnosed haematological malignancy in the paediatric population, accounting for roughly 25% of all childhood cancers and nearly 80% of all childhood leukemia cases globally. The World Health Organization estimates that approximately 400,000 new ALL cases are diagnosed each year, a figure expected to increase as global populations grow and environmental risk factors evolve [15].

The gold standard for diagnosing ALL has remained largely unchanged for decades. A trained hematopathologist aspirates a bone marrow sample, stains it with Giemsa or Leishman reagent, and examines it under a microscope at 400–1000× magnification. The pathologist counts and classifies 200–500 individual cells per slide, identifying abnormal blast cells by morphological features including enlarged nucleus, elevated nuclear-to-cytoplasmic ratio, altered chromatin texture, and the

 doi:10.20944/preprints202603.1631.v1

presence or absence of nucleoli. A confirmed ALL diagnosis requires at least 20% blast cells in the bone marrow sample.

This manual process is both time-intensive and subjective. A single examination requires 45–90 minutes per patient, and experienced pathologists examining the same slide can disagree in blast cell count by as much as 15–20% [14]. Furthermore, the global shortage of trained hematopathologists — particularly severe in developing nations where the specialist-to-patient ratio in India is estimated at approximately 1:500,000 — means that a large proportion of the population has no access to timely and accurate leukemia diagnosis.

Early computational approaches using hand-crafted image features and classical machine learning classifiers achieved accuracy in the range of 82–88%. The advent of deep learning dramatically improved performance, but a critical gap has persisted: the vast majority of systems function as opaque black boxes providing no visual explanation of their predictions. This paper presents Marrow-Find, a fully integrated system addressing this gap by combining five technical components: a ResNet50-based transfer learning classifier, a cGAN augmentation module, Grad-CAM visual explainability, Watershed cell segmentation, and a Flask-based clinical web application.

2. Challenges in Leukemia Diagnosis

The diagnosis of ALL through manual microscopic examination of bone marrow smears presents a set of interconnected challenges that have persisted in clinical haematology for decades.

2.1. Inter-Observer Variability

Soupir et al. demonstrated in 2018 that experienced hematopathologists can disagree in their blast cell count by as much as 20% when examining identical bone marrow smears [14]. Since the diagnostic threshold for ALL is a blast percentage of 20%, this level of variability means that a borderline case could receive a positive or negative diagnosis depending entirely on which pathologist examines the slide.

2.2. Time and Throughput Constraints

A thorough manual examination requires between 45 and 90 minutes of a pathologist's focused attention. In high-volume haematology laboratories, this creates diagnostic bottlenecks that directly delay the initiation of treatment. In the most aggressive forms of B-cell and T-cell ALL, treatment delays of even 24–48 hours can measurably worsen patient prognosis.

2.3. Lack of Quantitative Outputs

Manual diagnosis produces essentially qualitative outputs. A pathologist reports a blast percentage and morphological classification, but does not record precise measurements of individual cell features such as nuclear area, N:C ratio, or chromatin entropy. Automated systems can introduce a level of quantitative rigour that is currently absent from clinical practice.

2.4. Absence of Explainability in Existing AI Systems

Deep learning systems proposed for leukemia cell classification have generally neglected clinical interpretability. A system that classifies a cell as malignant without indicating which morphological features drove that decision is unlikely to be adopted by practising pathologists. The absence of explainability has been identified as the single most important barrier to clinical adoption [2].

 doi:10.20944/preprints202603.1631.v1

3. System Architecture and Design

The Marrow-Find system is designed as a modular, five-component pipeline that transforms a raw bone marrow microscopic image into a comprehensive diagnostic output within 2–3 seconds of upload.

3.1. Image Preprocessing Pipeline

Every image passes through a standardised preprocessing pipeline. The pipeline applies CLAHE (Contrast Limited Adaptive Histogram Equalisation) with a clip limit of 2.0 and a tile grid size of 8×8 pixels, then resizes to 224×224 pixels and normalises using ResNet50-specific ImageNet channel-wise mean subtraction and standard deviation scaling.

3.2. ResNet50 Classification Module

The classification module employs a ResNet50 deep convolutional neural network pre-trained on the 1.2-million-image ImageNet dataset [1,13]. The original 1000-class classification head is replaced with a Global Average Pooling layer, a Dense layer with 256 neurons and ReLU activation with L2 regularisation ($\lambda = 0.01$), a Dropout layer (rate = 0.5), and a final 4-class Softmax output layer.

3.3. Grad-CAM Explainability Module

The Grad-CAM module produces a colour heatmap overlay highlighting the image regions most responsible for the classification decision [2]. It computes the gradient of the predicted class score with respect to the last convolutional layer's feature map activations, globally averages these gradients to obtain per-channel importance weights, colourises using the jet colourmap, and superimposes on the original image with an alpha blending factor of 0.4.

3.4. Watershed Cell Segmentation Module

The Watershed module provides quantitative cell analysis. The algorithm converts the input image to greyscale, applies Otsu's adaptive global thresholding, uses morphological opening to remove noise, identifies cell centres as local maxima of the distance transform, and uses these as seeds for connected component labelling.

3.5. cGAN Augmentation Module

The conditional GAN addresses class imbalance by generating realistic synthetic bone marrow cell images for minority classes [5,6]. The Generator upsamples a 100-dimensional noise vector through five Conv2DTranspose layers to produce 224×224×3 images. The Discriminator applies five strided Conv2D layers to classify images as real or synthetic. Both networks are trained adversarially over 50 epochs, generating 5,000 synthetic images per minority class [12].

3.6. Flask Web Application

The Flask application orchestrates all five computational modules via 14 API routes covering user authentication, image upload, Grad-CAM visualisation, Watershed segmentation, prediction history with human-correction feedback, PDF report generation with QR codes, GAN augmentation triggering, and an AI-powered chatbot. User data is stored in an SQLite database with full audit trail functionality.

4. Results and Discussion

4.1. Dataset and Experimental Setup

The C-NMC 2019 dataset [11] was partitioned using a 70/15/15 train/validation/test split with stratified sampling, yielding 7,272 training, 1,558 validation, and 1,568 test images across four classes:

 doi:10.20944/preprints202603.1631.v1

Benign (3,389 images, 43.2%), Pre-B ALL (2,764 images, 35.3%), Early Pre-B ALL (985 images, 12.6%), and Pro-B ALL (700 images, 9.0%). The ALL-IDB dataset [10] was used for supplementary validation.

4.2. Classification Performance

The ResNet50 model achieved an overall test accuracy of 94.8%, a weighted F1-score of 0.954, and a Cohen's Kappa coefficient of 0.931. Table 1 presents the per-class precision, recall, F1-score, and AUC values. The Benign class achieved the highest precision of 97.4% — the most clinically important metric, as false positives directly trigger unnecessary treatment. The Pro-B ALL class showed the lowest F1-score of 94.8% due to morphological similarity with Pre-B ALL cells.

Table 1. Per-class classification performance on the C-NMC 2019 test set.

Class	Precision	Recall	F1-Score	AUC
Benign	97.4%	96.0%	96.7%	0.990
Pre-B ALL	94.4%	96.0%	95.2%	0.980
Pro-B ALL	94.7%	95.0%	94.8%	0.970
Early Pre-B ALL	95.0%	94.7%	94.8%	0.980

4.3. ROC Curve Analysis

ROC curves were computed for each class using a one-vs-rest strategy. All four classes achieved AUC values exceeding 0.97. The Benign class achieved the highest AUC of 0.990, and the Pro-B ALL class achieved the lowest AUC of 0.970, consistent with overlapping morphological features between Pro-B and Pre-B blast cells.

4.4. Confusion Matrix Observations

The majority of misclassifications occur at the boundary between Pro-B ALL and Pre-B ALL cells, consistent with published reports of inter-pathologist disagreement on these same cell pairs [14]. The most clinically critical error type — false negatives classifying a malignant blast cell as Benign — occurs at a rate of only 4 per 150 test examples for the Early Pre-B ALL class.

4.5. Training Convergence

Both training and validation accuracy curves increased steadily from approximately 52% and 48% at epoch 1 to approximately 96% and 94% at epoch 25. The narrow, consistent gap between training and validation curves confirms the model was well-regularised and did not overfit. The EarlyStopping callback was not triggered.

4.6. Comparison with State-of-the-Art Methods

Table 2 compares Marrow-Find against published methods for leukemia cell classification. The proposed system achieves the highest accuracy and F1-score, and is the only system that simultaneously provides all five clinically relevant features: classification, Grad-CAM explainability, automated cell counting, GAN augmentation, and a deployable web interface.

Table 2. Comparison of Marrow-Find with state-of-the-art leukemia classification methods.

Method	Accuracy	F1-Score	XAI	Web App
SVM+HOG [4]	87.4%	0.865	No	No

 doi:10.20944/preprints202603.1631.v1

Alzubaidi et al. [7]	92.8%	0.913	No	No
VGG16 [8]	93.6%	0.928	No	No
Matek et al. [3]	94.2%	0.921	No	No
Anilkumar et al. [9]	94.1%	0.930	No	No
Marrow-Find (Proposed)	**94.8%**	**0.954**	**Yes**	**Yes**

4.7. Grad-CAM Qualitative Analysis

Visual inspection of the Grad-CAM heatmaps confirms that the model focuses on morphologically relevant regions. For Benign cells, heatmaps show diffuse low-intensity activation. For ALL blast cells, heatmaps consistently show concentrated high-intensity activation centred on the nucleus, particularly emphasising nuclear enlargement and irregular chromatin texture. This alignment with pathologist diagnostic criteria confirms the classifier has learned clinically meaningful representations.

4.8. Performance Equations

The performance metrics used in this study are defined as follows:

$$\text{Accuracy} = (\text{Correct Predictions} / \text{Total Samples}) \times 100\% \quad \ldots(1)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \ldots(2)$$

$$\kappa = (P_{obs} - P_{chance}) / (1 - P_{chance}) \quad \ldots(3)$$

$$\text{Grad-CAM} = \text{ReLU}(\Sigma k\ [\alpha k^c \times Ak]) \quad \ldots(4)$$

4.9. Web Application Responsiveness

The complete prediction pipeline completes within 2–3 seconds per image on standard hardware, sufficient for real-time clinical workflow integration. The Flask application maintains this response time for up to ten simultaneous prediction requests via Gunicorn's multi-worker deployment configuration.

5. Limitations

Despite its strong performance, Marrow-Find has several limitations. First, the system was trained and evaluated exclusively on the C-NMC 2019 and ALL-IDB datasets acquired under controlled conditions; generalisation to slides produced by different staining protocols or scanner hardware has not been prospectively validated. Second, cGAN image quality was assessed qualitatively; formal quantitative validation using metrics such as Fréchet Inception Distance (FID) remains as future work. Third, the current SQLite deployment is not suitable for large-scale multi-institution concurrent use. Fourth, the system classifies individual cells in isolation and does not model spatial relationships between adjacent cells. Future work will address cross-institutional validation, FID-based GAN evaluation, and a scalable PostgreSQL backend.

 doi:10.20944/preprints202603.1631.v1

6. Conclusions

This paper has presented Marrow-Find, a comprehensive AI-powered web application for the automated detection and classification of leukemia cells from bone marrow microscopic images. The system successfully integrates five major technical components — ResNet50 transfer learning, cGAN data augmentation, Grad-CAM explainability, Watershed cell segmentation, and Flask-based web deployment — into a single, clinically usable platform achieving 94.8% test accuracy, a weighted F1-score of 0.954, and a Cohen's Kappa of 0.931 on the C-NMC 2019 benchmark dataset.

Marrow-Find outperforms all compared baseline methods while offering a broader suite of clinically relevant features than any prior published system. The Grad-CAM visualisations confirm that the classifier bases its predictions on morphologically meaningful features, addressing the critical explainability barrier that has historically prevented clinical adoption of AI diagnostic tools in haematology. By deploying all capabilities as a web-accessible application, Marrow-Find makes specialist-level leukemia diagnostic capability accessible from any standard web browser, with direct implications for improving diagnostic access in resource-limited healthcare settings such as rural India.

Author Contributions: Conceptualisation: T.D.R. and P.C.K.S.; Methodology: T.D.R., N.S. and P.C.K.S.; Software: P.C.K.S. and P.S.; Validation: N.S., P.S. and P.J.K.; Formal analysis: P.C.K.S.; Writing — original draft: P.C.K.S.; Writing — review and editing: T.D.R.; Supervision: T.D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The C-NMC 2019 dataset used in this study is publicly available at https://isbi.deepdr.org/. The ALL-IDB dataset is available at https://homes.di.unimi.it/scotti/all/. Code is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016; pp. 770–778.
2. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017; pp. 618–626.
3. Matek, C.; Schwarz, S.; Spiekermann, K.; Marr, C. Human-Level Recognition of Blast Cells in Acute Myeloid Leukaemia with Convolutional Neural Networks. Nat. Mach. Intell. 2019, 1, 538–544.
4. Guo, J.; Li, Y.; Luo, Y.; Wang, X. White Blood Cell Classification Using Gaussian-Weighted HOG and SVM Classifier. In Proceedings of the IEEE International Conference on Electronic Measurement and Instruments (ICEMI), Qinhuangdao, China, 2014.
5. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. Adv. Neural Inf. Process. Syst. 2014, 27.
6. Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. Neurocomputing 2018, 321, 321–331.
7. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. J. Big Data 2021, 8, 1–74.
8. Loey, M.; Naman, M.; Zaid, H. Deep Transfer Learning in Diagnosing Leukemia in Blood Cells. Computers 2020, 9, 29.

 doi:10.20944/preprints202603.1631.v1

9. Anilkumar, K.K.; Manoj, V.J.; Sagi, T.M. A Survey on Image Segmentation of Blood and Bone Marrow Smear Images. J. Pathol. Inform. 2021, 12, 36.
10. Labati, R.D.; Piuri, V.; Scotti, F. ALL-IDB: The Acute Lymphoblastic Leukemia Image Database for Image Processing. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, 2011; pp. 2045–2048.
11. Gupta, A.; Gupta, R.; Mnich, K.; Bhavna; Sirinukunwattana, K.; Bhatt, V.R. ISBI 2019 C-NMC Challenge: Classification in Normal vs. Malignant Cells in B-ALL White Blood Cancer Microscopy Images. IEEE ISBI 2019, Venice, Italy, 2019.
12. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. J. Big Data 2019, 6, 1–48.
13. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 2009; pp. 248–255.
14. Soupir, C.A.; Bedrick, A.D.; Nguyen, J.T.; Dolan, M.; Nguyen, P.L. Correlation of Blast Percentage and Immunophenotype in Bone Marrow Aspirate. Arch. Pathol. Lab. Med. 2018, 142, 347–352.
15. World Health Organization. Global Cancer Observatory; International Agency for Research on Cancer: Lyon, France, 2022. Available online: https://gco.iarc.fr (accessed on 14 March 2025).