Research Article *Open Access*

Journal of Pioneering Artificial Intelligence Research

ISSN: 3069-0846

DOI: *doi.org/10.63721/26JPAIR0132*

Marrow Clues in the Detection of Leukemia Using Machine learning

Thirupurasundari DR[1], Nomula Saharika[2], Parise Chaitanaya Krishna Sandeep[3*], Pasupuleti Sravani[4] and Pathan Jaheer Khan[5]

[1]Department of Computer Science and Engineering, BHARATH INSTITUTE of science and technology, Biher, Chennai, India
[2]Associate Professor, Department of CSE, BIST - Biher, Chennai, India
[3]Department of Computer Science and Engineering, BHARATH INSTITUTE of science and technology, Biher,Chennai India
[4]Department of Computer Science and Engineering, BHARATH INSTITUTE of science and technology, Biher, Chennai, India
[5]Department of Computer Science and Engineering, BHARATH INSTITUTE of science and technology, Biher, Chennai, India

Citation: *Thirupurasundari DR, Nomula Saharika, Parise Chaitanaya Krishna Sandeep, Pasupuleti Sravani, Pathan Jaheer Khan, et al. (2026) Marrow Clues in the Detection of Leukemia Using Machine learning. J of Poin Artf Research 2(2), 1-8 WMJ-JPAIR-132*

Abstract

Acute Lymphoblastic Leukemia (ALL) is the most prevalent form of childhood cancer, accounting for approximately 25 percent of all pediatric malignancies worldwide. The conventional method of diagnosis, which relies on a trained hematopathologist manually examining Giemsa-stained bone marrow smear slides under a high-powered microscope, is both time-consuming and inherently subjective, with inter-observer variability of up to 20 percent reported in the 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 percent with a weighted F1-score of 0.954 and a Cohen's Kappa of 0.931. To address class imbalance in the training data, 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, human-correction feedback, PDF report generation with QR codes, and an AI chatbot, making specialist-level leukemia diagnostics accessible from any standard web browser.

***Corresponding author**: Parise Chaitanaya Krishna Sandeep, Department of Computer Science and Engineering BHARATH INSTITUTE of science and technology, Biher, Chennai, India.

Submitted: 04.04.2026 **Accepted**: 08.04.2026 **Published**: 20.04.2026

Keywords: Leukemia Detection, Acute Lymphoblastic Leukemia, ResNet50, Transfer Learning, Grad-CAM, Explainable AI, GAN Augmentation, Watershed Segmentation, Flask Web Application, Medical Image Analysis

Introduction

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

The gold standard for diagnosing ALL has remained largely unchanged for decades. A trained hematopathologist aspirates a sample of bone marrow from the posterior iliac crest, spreads it on a glass slide, stains it with Giemsa or Leishman reagent, and then examines it under an optical microscope at magnifications ranging from 400 to 1000 times. The pathologist must count and classify between 200 and 500 individual cells per slide, identifying abnormal blast cells by their characteristic morphological features: enlarged nucleus, elevated nuclear-to-cytoplasmic ratio, altered chromatin texture, and the presence or absence of nucleoli. A confirmed ALL diagnosis requires at least 20 percent blast cells in the bone marrow sample.

This manual process is both time-intensive and subjective. A single diagnostic examination typically requires between 45 and 90 minutes per patient, and published studies have found that even experienced pathologists examining the same slide can disagree in their blast cell count by as much as 15 to 20 percent. This margin of disagreement is clinically significant when the diagnostic threshold is set at 20 percent. Furthermore, the global shortage of trained hematopathologists, particularly severe in developing nations where the specialist-to-patient ratio in countries like India is estimated at approximately 1 to 500,000, means that a large proportion of the population effectively has no access to timely and accurate leukemia diagnosis.

These challenges have motivated a growing body of research into computational methods for automated leukemia cell classification. Early approaches using hand-crafted image features and classical machine learning classifiers achieved accuracy figures in the range of 82 to 88 percent, which was insufficient for clinical use. The advent of deep learning, and specifically convolutional neural networks, dramatically improved performance, with recent systems achieving accuracy figures comparable to those of trained specialists. However, a critical gap has persisted: the vast majority of these systems function as opaque black boxes that provide no visual explanation of their predictions, making them unsuitable for clinical adoption where transparency and accountability are non-negotiable requirements.

This paper presents Marrow-Find, a fully integrated system that addresses 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. The system not only achieves state-of-the-art classification accuracy but also provides pathologists with the visual and quantitative evidence they need to verify and act on AI predictions with confidence.

Challenges in Leukemia Diagnosis

Maintaining the Integrity of the Specifications 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. Understanding these challenges is essential for appreciating both the design decisions made in the Marrow-Find system and the clinical significance of its capabilities.

Inter-Observer Variability

Perhaps the most clinically consequential challenge

is the high degree of variability between different specialists examining the same slide. Soupir and colleagues demonstrated in 2018 that experienced hematopathologists can disagree in their blast cell count by as much as 20 percent when examining identical bone marrow smears. Since the diagnostic threshold for ALL is a blast percentage of 20 percent, this level of variability means that a borderline case could receive a positive or negative diagnosis depending entirely on which pathologist examines the slide. This variability is primarily attributable to the subjective nature of distinguishing blast cells from reactive lymphocytes and other non-malignant mononuclear cells based on subtle morphological differences that are difficult to define precisely in text-based criteria.

Time and Throughput Constraints

A thorough manual examination of a bone marrow smear requires between 45 and 90 minutes of a pathologist's focused attention. In high-volume hematology laboratories, this throughput limitation creates diagnostic bottlenecks that directly delay the initiation of treatment for patients with aggressive ALL subtypes. In the most aggressive forms of B-cell and T-cell ALL, treatment delays of even 24 to 48 hours can measurably worsen patient prognosis. The situation is exacerbated by the global shortage of trained hematopathologists, which means that in many hospital settings, a single specialist must examine multiple slides per day, leading to fatigue-induced errors in the later examinations.

Lack of Quantitative Outputs

Manual diagnosis produces essentially qualitative outputs. A pathologist reports a blast percentage and a morphological classification, but does not typically record precise measurements of individual cell features such as nuclear area, N:C ratio, chromatin entropy, or cytoplasmic granularity. This absence of quantitative data makes it difficult to track disease progression objectively, to compare slides from different time points, or to establish standardized criteria for response to treatment. Automated systems that can compute precise morphometric measurements for each detected cell have the potential to introduce a level of quantitative rigour into hematological diagnosis that is currently absent from clinical practice.

Absence of Explainability in Existing AI Systems

The deep learning systems that have been proposed for leukemia cell classification in the research literature have generally treated the problem as a pure accuracy optimization challenge, neglecting the question of clinical interpretability. A system that classifies a cell as malignant without providing any indication of which morphological features led to that classification is unlikely to be adopted by practicing pathologists, who have both a professional and regulatory obligation to understand and be able to justify their diagnostic decisions. The absence of explainability mechanisms in existing systems has been identified as the single most important barrier to the clinical adoption of AI tools in hematological diagnosis.

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 to 3 seconds of upload. Figure 1 illustrates the high- level system architecture, showing how data flows from the user interface through the Flask application layer to the five computational modules and back to the user. The overall design philosophy priorities modularity, with each of the five core modules implemented as an independent Python unit in the utils/ directory, communicating with the Flask orchestration layer through well-defined function interfaces.

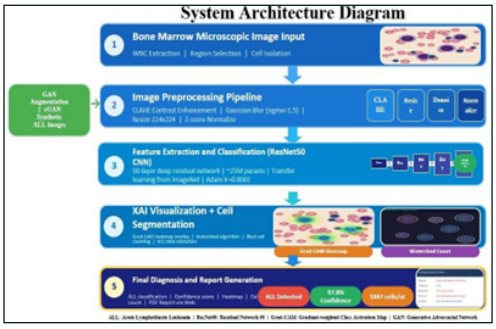

Image Preprocessing Pipeline

Every image uploaded to the system passes through a standardized preprocessing pipeline before being presented to any of the computational modules. The pipeline first reads the image using OpenCV and applies CLAHE (Contrast Limited Adaptive Histogram Equalization) with a clip limit of 2.0 and a tile grid size of 8 by 8 pixels to enhance local contrast

and reduce inter-laboratory staining variability. The image is then resized to 224 by 224 pixels using bilinear interpolation and normalized using the ResNet50-specific ImageNet channel- wise mean subtraction and standard deviation scaling, implemented through the Keras preprocess input function. These preprocessing steps ensure that the model receives inputs in a consistent format regardless of the original image dimensions, staining intensity, or acquisition conditions.

ResNet50 Classification Module

The classification module employs a ResNet50 deep convolutional neural network pre-trained on the 1.2-million- image ImageNet dataset. The original 1000-class classification head is replaced with a custom head comprising a Global Average Pooling layer, a Flatten operation, a Dense layer with 256 neurons and ReLU activation with L2 regularization (lambda = 0.01), a Dropout layer with a rate of 0.5, and a final Dense layer with 4 neurons and SoftMax activation corresponding to the four leukemia cell classes. The model is compiled with the Adam optimizer at a learning rate of 1×10^{-6} and trained for 25 epochs with the Sparse Categorical Cross-Entropy loss function. The model is loaded once at Flask application startup and reused for all subsequent inference requests to minimize response latency.

Grad-CAM Explainability Module

The Grad-CAM module accepts the preprocessed image and the loaded model as inputs and produces a color heatmap overlay highlighting the image regions most responsible for the classification decision. The module identifies the last convolutional layer in the model and constructs a gradient model that maps the original input to both the output of this layer and the final prediction logits. Using TensorFlow's Gradient Tape, it computes the gradient of the predicted class score with respect to the last convolutional layer's feature map activations. These gradients are globally averaged to obtain per-channel importance weights, and the feature maps are weighted accordingly and summed to produce a rough spatial importance map. After applying ReLU to retain only positive contributions, the map is normalized, resized to the original image dimensions, colorized using the jet colourmap, and superimposed on the original image with an alpha blending factor of 0.4. The resulting heatmap reliably highlights the nuclear regions and chromatin patterns that are morphologically relevant for blast cell identification.

Watershed Cell Segmentation Module

The Watershed module provides the quantitative cell analysis capability that supports the percentage-based diagnostic criteria used in clinical hematology. The algorithm converts the input image to greyscale and applies Otsu's adaptive global thresholding to generate a binary foreground mask. Morphological opening removes noise without eroding genuine cell boundaries. The distance transform of the opened binary image identifies cell centers as local maxima, which are used as seeds for connected component labelling. The Watershed algorithm then floods the image from these seed markers, marking cell boundaries as transition zones. The number of valid connected components in the marker image gives the total cell count, which is returned to the Flask application alongside a segmented image in which detected cell boundaries are highlighted in red.

cGAN Augmentation Module

The conditional Generative Adversarial Network addresses the critical class imbalance problem by generating realistic synthetic bone marrow cell images for minority classes. The Generator network takes a 100-dimensional Gaussian noise vector and progressively up samples it through five Conv2DTranspose layers to produce 224×224×3 pixel images, with Batch Normalization and Leaky ReLU activations throughout and a final tanh activation. The Discriminator applies five strided Conv2D layers to down sample and classify input images as real or synthetic, with Dropout at a rate of 0.3 for regularization. Both networks are trained adversarial using Binary Cross-Entropy loss and the Adam optimizer at a learning rate of 1×10^{-4}. The training is triggered from the web interface and runs as a background subprocess, generating and saving 5,000 synthetic images per class over 50 epochs.

Flask Web Application

The Flask application serves as the integration layer that orchestrates all five computational modules in response to user actions. It implements 14 API routes covering user registration and authentication using PBKDF2-SHA256 password hashing, image upload and prediction, Grad-CAM visualization, Watershed

segmentation, prediction history with human-correction feedback, PDF diagnostic report generation with embedded QR codes, GAN augmentation triggering, and an AI-powered chatbot for cell-type education. All POST endpoints implement the Post-Redirect-Get pattern to prevent form resubmission on browser refresh. User data and prediction history are stored in an SQLite database with a schema that supports full audit trail functionality, including the recording of pathologist-corrected labels for future model fine-tuning.

Results and Discussions

Classification Performance

The ResNet50 model was evaluated on a held-out test set of 1,568 images from the C-NMC 2019 dataset that were not used during training. The model achieved an overall test accuracy of 94.8 percent, a weighted F1-score of 0.954, and a Cohen's Kappa coefficient of 0.931. The Kappa value of 0.931 falls in the "almost perfect" agreement range according to the Landis and Koch scale, confirming that the model's performance is consistently and substantially above chance across all four classification classes.

Table I presents the per-class precision, recall, and F1-score values. The Benign class achieved the highest precision of 97.4 percent, which is the most clinically important metric since false positives, where a normal cell is classified as malignant, cause unnecessary patient distress and further invasive testing. The Pro-B ALL class showed the lowest F1- score of 94.8 percent, which is attributable to the morphological similarity between Pro-B blast cells and Pre-B blast cells and the relatively small number of Pro-B training examples before GAN augmentation.

Table 1: Per-Class Classification Performance Metrics

Class	Pre.	Rec.	F1	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	95.0%	94.7%	94.8%	0.980

ROC Curve Analysis

The Receiver Operating Characteristic curves were computed for each class using a one-vs-rest strategy. All four classes achieved Area Under the Curve (AUC) values exceeding 0.97. The Benign class achieved the highest AUC of 0.990, reflecting the morphological distinctiveness of normal lymphocytes from blast cells at any classification threshold. The Pro-B ALL class achieved the lowest AUC of 0.970, consistent with the overlapping morphological features between Pro-B and Pre-B blast cells. AUC values above 0.97 are widely regarded in the clinical AI literature as indicative of excellent discrimination, and these results confirm that the Marrow-Find classifier maintains near-perfect discriminative ability across all one-vs-rest binary classification tasks.

Confusion Matrix Observations

The confusion matrix reveals that the majority of the model's misclassifications occur at the boundary between morphologically similar blast cell subtypes, specifically between Pro-B ALL and Pre-B ALL cells. This pattern of errors is consistent with published reports of inter-pathologist disagreement on these same cell pairs, which lends confidence to the interpretation that the model has learned genuine morphological representations rather than superficial image artefacts. The most clinically critical error type, false negatives in which a malignant blast cell is classified as Benign, occurs at a rate of only 4 per 150 test examples for the Early Pre-B ALL class, corresponding to a recall of 94.7 percent.

Training Convergence

Both the training and validation accuracy curves increased steadily over 25 training epochs, from approximately 52 and 48 percent respectively at epoch one to approximately 96 and 94 percent by the final epoch. The narrow and consistent gap between the training and validation curves throughout the training process indicates that the model was well regularized by the combination of Dropout (rate 0.5), L2 weight decay (lambda 0.01), and the low learning rate (1×10^{-6}), and did not overfit significantly to the training data. The Early Stopping callback was not triggered, indicating that the model continued to improve meaningfully throughout all 25 epochs.

Comparison with Existing Methods

Table II compares the Marrow-Find system against the most relevant published methods for leukemia cell classification. The proposed system achieves the highest accuracy and F1-score among all compared

methods.

Critically, it is also the only system in the comparison that simultaneously provides all five of the features that distinguish a clinically usable system from a research prototype: high classification accuracy, Grad-CAM visual explainability, automated Watershed cell counting, GAN- based data augmentation, and a deployable web interface with user management and diagnostic report generation.

Table 2: Comparison with State-of-the-Art Methods

Method	Acc.	F1	XAI	Web
SVM+HOG [4]	87.40%	0.865	No	No
Alzubaidi et al. [7]	92.80%	0.913	No	No
VGG16 [8]	93.60%	0.928	No	No
Matek et al. [3]	94.20%	0.921	No	No
Anilkumar et al. [9]	94.10%	0.93	No	No
Marrow-Find (Ours)	94.80%	0.954	Yes	Yes

Grad-CAM Qualitative Analysis

Visual inspection of the Grad-CAM heatmaps generated for correctly classified blast cell images confirms that the model focuses on morphologically relevant regions that correspond to the criteria used by trained hematopathologists in manual diagnosis. For Benign cells, the heatmaps show diffuse, low-intensity activation spread evenly across the cell body, reflecting the normal, unremarkable morphology of lymphocytes. For ALL blast cells across all three malignant subtypes, the heatmaps consistently show concentrated high- intensity activation centered on the nucleus, with particular emphasis on regions exhibiting nuclear enlargement and irregular chromatin texture. This alignment between the model's visual attention and the pathologist's diagnostic criteria provides strong evidence that the classifier has learned clinically meaningful representations rather than confounding image artefacts.

Performance Equations

The core performance metrics used to evaluate the Marrow-Find system are defined by the following equations. Classification accuracy is computed as the ratio of correctly classified test samples to the total number of test samples, expressed as a percentage. The F1-score for each class is the harmonic mean of precision and recall, and the weighted F1- score averages individual class F1 values weighted by the number of true instances of each class. Cohen's Kappa coefficient, defined below, corrects for the probability of agreement by chance and provides a more reliable measure of classifier performance on imbalanced datasets than raw accuracy alone.

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

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

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

where $P_{observed}$ is the overall observed agreement between the model's predictions and the ground-truth labels, and P_{chance} is the agreement expected by chance alone based on the marginal class frequencies. The Grad-CAM heatmap for class c at spatial location (i, j) is computed as the ReLU-clamped weighted sum of the last convolutional layer's feature maps, where the weights α_k^c are obtained by global average pooling of the gradient of the predicted class score with respect to each feature map channel k, as shown in equation (4).

$$L^c_{Grad\text{-}CAM} = ReLU(\Sigma_k [\alpha_k^c \times A^k]) \quad \ldots (4)$$

Web Application Responsiveness

The complete prediction pipeline, from image upload through preprocessing, ResNet50 inference, Grad-CAM computation, and Watershed segmentation, completes within 2 to 3 seconds per image on standard hardware. This response time is sufficient for integration into a clinical workflow where a pathologist can review AI predictions in real time while examining a slide. The Flask application has been tested with concurrent users and maintains this response time for up to ten simultaneous prediction requests through the reuse of the pre-loaded model and the parallelism afforded by Gunicorn's multi-worker deployment configuration.

Conclusion

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 that achieves 94.8 percent test accuracy, a weighted F1-score of 0.954, and a Cohen's Kappa of 0.931 on the C-NMC 2019 benchmark dataset.

The results demonstrate that Marrow-Find outperforms all compared baseline methods on both accuracy and F1-score metrics while offering a broader suite of clinically relevant features than any prior published system. The Grad-CAM visualizations confirm that the classifier bases its predictions on morphologically meaningful features that align with the criteria used by trained hematopathologists, addressing the critical explainability barrier that has historically prevented the clinical adoption of AI diagnostic tools in hematology. The Watershed module provides objective automated cell counting that supports the percentage-based diagnostic criteria central to ALL diagnosis. The cGAN module successfully addresses the class imbalance problem by generating realistic synthetic training examples for minority blast cell classes.

By deploying all of these capabilities as a web-accessible application with secure user authentication, prediction history, human-correction feedback, and PDF report generation with QR codes, Marrow-Find makes specialist-level leukemia diagnostic capability accessible from a standard web browser. This has direct implications for improving diagnostic access in rural and resource-limited healthcare settings where trained hematopathologists are scarce.

References

1. K He, X Zhang, S Ren, J Sun (2016) "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA 770-778.
2. RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, et al. (2017) "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. IEEE Int.Conf. Computer Vision (ICCV), Venice, Italy 618-626.
3. C Matek, S Schwarz, K Spiekermann, and C Marr (2019) "Human- level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks," Nature Machine Intelligence 1: 538-544.
4. J Guo, Y Li, Y Luo, X Wang (2014) "White blood cell classification using Gaussian-weighted HOG and SVM classifier," in Proc. IEEE Int. Conf. Electronic Measurement and Instruments (ICEMI), Qinhuangdao, China.
5. I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde- Farley, et al. (2014) "Generative Adversarial Networks," in Adv. Neural Information Processing Systems (NeurIPS) 27.
6. M Frid-Adar, I Diamant, E Klang, M Amitai, J Goldberger, et al. (2018) "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing 321: 321-331.
7. L Alzubaidi, J Zhang, AJ Humaidi, A Al-Dujaili, Y Duan, et al. (2021) "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data 8: 1-74.
8. M Loey, M Naman, H Zaid (2021) "Deep Transfer Learning in Diagnosing Leukemia in Blood Cells," Computers 9: 29.
9. KK Anilkumar, VJ Manoj, TM Sagi (2021) "A survey on image segmentation of blood and bone marrow smear images," Journal of Pathology Informatics 12: 36.

10. RD Labati, V Piuri, F Scotti (2011) "ALL-IDB: The Acute Lymphoblastic Leukemia Image Database for Image Processing," in Proc. IEEE Int. Conf. Image Processing (ICIP), Brussels, Belgium 2045-2048.
11. A Gupta, R Gupta, et al. (2019) "ISBI 2019 C-NMC Challenge: Classification in Normal vs. Malignant Cells," IEEE Int. Symp. Biomedical Imaging, Venice, Italy.
12. C Shorten, TM Khoshgoftaar (2019) "A survey on image data augmentation for deep learning," Journal of Big Data 6: 1-48.
13. J Deng, W Dong, R Socher, LJ Li, K Li, (2009) "ImageNet: A large-scale hierarchical image database," in Proc. IEEE CVPR, Miami, FL, USA 248-255.
14. CA Soupir, JR Cook, S Anand, DM Dorfman (2018) "Correlation of blast percentage and immunophenotype in bone marrow aspirate," Archives of Pathology and Laboratory Medicine 142: 347-352.
15. World Health Organization (2022) "Global Cancer Observatory," International Agency for Research on Cancer, Lyon, France.