Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identification of Mutations in Giemsa-Stained Bone Marrow Images of Acute Myeloid Leukemia

Garima Jain

garima.j@icmr.gov.in

Indian Council of Medical Research

Ravi Kant Gupta

Indian Institute of Technology Bombay

Priyansh Jain

Indian Institute of Technology Bombay

Ardhendu Sekhar

Indian Institute of Technology Bombay

Abhijeet Patil

Indian Institute of Technology Bombay

Amit Sethi

Indian Institute of Technology Bombay

Article

Keywords: Acute Myeloid Leukemia, Molecular prognostication, Bone marrow smears, Blasts, Heamtopathol- ogy, Giemsa stain

Posted Date: March 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8624068/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

IDENTIFICATION OF MUTATIONS IN GIEMSA-STAINED BONE MARROW IMAGES OF ACUTE MYELOID LEUKEMIA

Garima Jain, Ravi Kant Gupta*, Priyansh Jain*, Ardhendu Sekhar, Abhijeet Patil, Amit Sethi*
Corresponding author: Garima Jain (garima.j@icmr.gov.in)

Indian Institute of Technology Bombay, Mumbai, India

ABSTRACT

In this study, we propose a robust methodology for identification of myeloid blasts followed by prediction of genetic mutation in single-cell images of blasts, tackling challenges associated with label accuracy and data noise. We trained an initial binary classifier to distinguish between leukemic (blasts) and non-leukemic cells images, achieving 90% accuracy. To test the model's generalization, we applied this model to a separate large unlabeled dataset and validated the predictions with two hemato-pathologists, finding an approximate error rate of 20% in the leukemic and non-leukemic labels. Assuming this level of label noise, we further trained a four-class model on images predicted as blasts to classify specific mutations. The mutation labels were known for only a bag of cell im- ages extracted from a single slide. Despite the tumor label noise, our mutation classification model achieved 85% accu- racy across four mutation classes, demonstrating resilience to label inconsistencies. This study highlights the capability of machine learning models to work with noisy labels effectively while providing accurate, clinically relevant mutation predictions, which is promising for diagnostic applications in areas such as hemato-pathology.

Index Terms— Acute Myeloid Leukemia, Molecular prognostication, Bone marrow smears, Blasts, Heamtopathology, Giemsa stain

1. INTRODUCTION

Acute myeloid leukaemia (AML) is a highly aggressive cancer affecting the hematopoietic system. The understanding of the mutational landscape of AML has expanded significantly in recent years, leading to updates in the classification of AML and revolutionizing its prognostic stratification, treat- ment approaches, and response evaluation [1]. The primary curative approach involves intensive induction chemotherapy, followed by consolidation therapy or allogeneic hematopoi- etic stem cell transplantation. Accurate risk assessment, which is crucial for selecting the most appropriate treatment, is currently based on genetic profiling of mutations and chro- mosomal abnormalities [2], which also have proven to be the most predictive of overall survival and relapse-free sur- vival [3] . However, obtaining genetic information can take several days to weeks after diagnosis, incur substantial costs for healthcare as they are not universally available and require expert domain knowledge [4].

AML classification systems incorporate recurrent cytogenetic abnormalities, such as RUNX1::RUNX1T1, PML::RARA, MYH11::CBFB, and mutations in key genes like NPM1, FLT3, and CEBPA into the classification and prognostic models of AML. In recent years, new therapeutic agents have also been approved to complement intensive first-line treatments for specific genetic subgroups. [4].

Cytomorphology may, in some cases, also lead to the sus- picion of possible underlying genetic mutations. However, to the best of our knowledge, while machine learning has been leveraged to identify blasts from single cell images or to dif- ferentiate between blasts of acute myeloid leukemia and acute lymphoid leukemia, only a couple of studies [5, 6] have lever- aged deep learning to identify underlying genetic mutations on haemato-pathology smears.

In this study, we employ a multi-step deep learning method for automated segmentation and identification of myeloid blasts using the Giemsa stained images of human bone marrow samples. The downstream pipeline then identifies status of NPM1, RUNX1:RUNX1T1, and CBFB: MYH11. mutations between samples of Acute Myeloid Leukemia (AML) and healthy controls and predict the muta- tion.

2. METHODOLOGY

Our methodology comprises two key phases: (1) cancer de- tection through a threshold-based binary classification model for distinguishing malignant from benign cells shown in Fig- ure 1, and (2) mutation classification via a multi-class deep learning framework for subtype identification of specific can- cer mutations using high-resolution single-cell images shown in Figure 2. Advanced preprocessing, data augmentation, and class balancing techniques were implemented to mitigate issues with noisy labels, data sparsity, and class imbalances, thereby enhancing model robustness, sensitivity, and reliabil- ity. This integrated, multi-stage approach provides a compre- hensive framework for cellular-level diagnostics, enabling ac-

*Equal Contribution by Garima Jain, Ravi Kant Gupta, Priyansh Jain

curate cancer detection and mutation-specific profiling within complex biological datasets.

Fig. 1. Labeled single cell images from Dataset 1 were grouped into two bags i.e. cancer vs non-cancer, which is then used to train a tumor detector model

2.1. Dataset

The experiments utilized two distinct datasets of acute myeloid leukemia (AML).

Dataset 1 (D1): This dataset, AML-Cytomorphology MLL · Helmholtz, contains over 170,000 de-identified, expert-annotated cells from bone marrow smears of 945 patients. The dataset comprises a total of 21 classes, including 2 cancerous, 16 non-cancerous, and 3 irrelevant classes as determined by domain experts. The cells were stained using the May-Gru¨nwald-Giemsa/Pappenheim stain and acquired via brightfield microscopy at 40x magnification. This dataset serves as a vital resource for developing and validating computational methods in diagnostic medicine, particularly for hematologic malignancies.

Dataset 2 (D2): The second dataset, AML-Cytomorphology· MLL Helmholtz, includes data on four common subtypes of acute myeloid leukemia (AML) characterized by specific genetic anomalies and typical morphological features as classified by the WHO in 2022. The subtypes are: (i) APL characterized by the PML: RARA fusion, (ii) AML with a mutation in NPM1, (iii) AML marked by the CBFB: MYH11 fusion (excluding NPM1 mutation), and (iv) AML with the RUNX1::RUNX1T1 fusion. Addition- ally, the dataset includes a control group composed of healthy stem cell donors. A total of 189 peripheral blood smears were digitized, each containing between 99 and 500 white blood cells, scanned at 40x magnification. This dataset captures the morphological complexity of acute myeloid leukemia and provides a substantial basis for mutation classification.

2.2. Cancer Detection: Data Preprocessing and Model Development

A notable challenge in processing Dataset 1 (D1) was the pronounced class imbalance between cancerous and non-cancerous cell types. Simple random sampling would have resulted in significant under-representation of cancerous cells, which could bias the model toward the prevalent non- cancerous class. To address this, we implemented a stratified proportional sampling algorithm, which ensured balanced representation across both training and validation sets, thus enhancing model generalization and reducing bias by provid- ing equivalent exposure to both classes.

The cancer detection model was based on ResNet-18 [7] architecture that was pretrained on ImageNet [8], to classify individual single-cell images as cancerous or non-cancerous. To reduce the impact of classification errors, a 5% error threshold was established; if fewer than 5% of the cells for a given patient were classified as cancerous, the patient was labeled as non-cancerous. This thresholding approach min- imized the likelihood of misclassifying patients with sparse cancerous cells and improved classification reliability.

2.3. Mutation Detection: Addressing Noisy Labels and Classification

Given the possibility of miss-classification where some non- cancerous cells might be mislabeled as cancerous, we treated this as a noisy label problem. Smooth cross-entropy [9] was employed to mitigate the impact of noisy labels, using a noisy label percentage of 20%. This method allowed for more flex- ibility in learning from potentially mislabeled data, reduc- ing the impact of erroneous labels on model performance. A multi-class classification model was developed to classify cancerous cells into one of four mutation types.

3. EXPERIMENTAL DETAILS AND RESULTS

3.1. Experimentation

We utilized a pretrained ResNet-34 as our primary feature ex- tractor, as it demonstrated superior performance compared to other architectures. For the ablation study, we experimented with various feature extractors, including ResNet-18, ResNet- 50 [7], VGG13, VGG16, VGG19 [10], ConvNeXt Tiny [11],
Swin Transformer Tiny [12], and EfficientNet [13]. Our findings indicate that ResNet-34 provided the most effective feature extraction capabilities, justifying its selection for our approach. Dataset 1 was partitioned into an 80-20 split for training and validation, with optimization performed us- ing the Adam optimizer (learning rate 0.001, weight decay 0.001, batch size 16). Regularization was reinforced through a dropout layer (0.3) following

feature extraction, enhancing model robustness and mitigating overfitting. Early stopping, with a patience level of 50 epochs and an error tolerance

Fig. 2. Unlabeled images of Dataset 2 were labeled with the help of algorithm 1. Labels cross-verified by two subject experts. Single cell images labeled as cancer were fed to mutation prediction network.

(delta) of 0.02, ensured efficient convergence. Specifically, the ResNet-18 model was trained in two configurations: first with a frozen encoder, training only the classification head, achieving maximum training and validation accuracies of 0.789 and 0.837, respectively, and then with full encoder fine-tuning, resulting in improved maximum accuracies of 0.918 for training and 0.892 for validation. All encoders were initialized with ImageNet pre-trained weights, facilitating effective transfer learning for cancer detection.

For mutation detection, the same encoder architectures were applied to Dataset 2, which was split into 72%, 18%, and 10% for training, validation, and testing, respectively. Optimization was conducted using the Stochastic Gradient Descent (SGD) optimizer, configured with a learning rate of 0.001, weight decay of 0.001, momentum of 0.9, and a batch size of 64 to support stable convergence. A dropout layer (0.3) was incorporated post-feature extraction and preceding the classification head to enhance generalization and mini- mize overfitting risks. Early stopping, with a patience of 50 epochs and an error delta of 0.02, further improved training efficiency by halting training upon stabilized convergence, ensuring computational efficiency and model performance re- liability.

3.2. Results

The performance evaluation of the cancer detection and mutation classification models described in Table 1, was conducted using several rigorous metrics, including accuracy, precision, recall, and F1-score. The model achieved an accuracy of approximately 90% in distinguishing cancerous from non-cancerous cells, demonstrating its robustness in high-stakes diagnostic tasks. The cancer detection model achieved an accuracy of approximately 90% in classifying single cell images of cells from Bone marrow into cancerous and non-cancerous categories for D1. This performance metric indicates a high reliability in differentiating malignant from benign cells, further underscored by precision and re- call values that capture the model's capability to correctly identify positive cases while minimizing false positives and negatives. The F1-score offers a balanced measure, provid- ing insights into the model's robustness in practical, clinical scenarios. By applying a stringent 5% error threshold, the model effectively reduced false positives in the classifica- tion of non-cancerous samples. This threshold ensured that the model maintained a high specificity, thereby minimizing the likelihood of incorrectly identifying healthy cells as can- cerous. This approach enhances the model's reliability for clinical applications, where reducing false positives is critical to prevent unnecessary diagnostic follow-ups and treatments.

The mutation classification model achieved an 85% accu- racy across the four classes using the smooth cross entropy loss that is robust to data mislabeling. This result was ob- tained by employing a majority voting mechanism based on aggregated predictions from individual single-cell images to enhance classification robustness across specific cancer mu-

Dataset	Task	Accuracy	AUC	Precision	Recall	F1-score
D1	Tumor vs Non-Tumor	0.90	0.93	0.88	0.92	0.90
D2	Mutation Prediction	0.85	0.70	0.82	0.85	0.84

Table 1. Results of tumor vs benign classification on D1 and mutation prediction (4-classes) on D2 using the proposed method

tation categories. This approach allowed for a more reliable consensus on mutation classification, reducing variability and aligning closely with expert-annotated benchmarks. The high concordance between the model's predictions and ex- pert analyses highlights the model's potential in supporting accurate and scalable mutation classification.

Collectively, these results underscore the efficacy of the proposed methodologies in advancing diagnostic precision in cancer detection and mutation typing, establishing a robust framework for further exploration and deployment within clinical diagnostics.

4. CONCLUSION AND DISCUSSION

Our study demonstrates robust performance in identifying clinically relevant mutations in AML through cytomorphology images, aligning with advancements in the field while offering distinct improvements in data handling and accuracy. For instance, [2] developed a pipeline using Pappenheim-stained bone marrow smears to predict genetic mutations such as NPM1, achieving an AUROC of 0.86 for NPM1 mutations. Similarly employing deep learning on cytomor- phology, our approach utilizes a multi-step process with CNNs and noise-resistant loss functions, achieving com- parable accuracy (85%)even under noisy label conditions. Other models, such as the Morphogo system [14] trained on a massive dataset of bone marrow cells (2.8 million images), report high overall accuracy in morphology-based classifi- cation, though they often emphasize cell-type identification over mutation prediction. Meanwhile, [15] developed AL- NET, which focused on lineage-specific classification among leukemia subtypes, with high sensitivity and specificity in differentiating acute promyelocytic leukemia and myeloid leukemia lineages. While these models show valuable results, our approach uniquely combines mutation-specific classification with label noise handling, enhancing model robustness in noisy, real-world clinical data contexts and underscoring its potential for broader clinical application.

Our study aligns with and extends prior work on AML subtype identification using deep learning while offering unique insights into handling label noise and robust muta- tion detection. SCEMILA, [5] an explainable AI model, employed attention mechanisms to identify diagnostically relevant cells in AML, achieving F1 scores as high as 0.86 for PML::RARA subtypes with agreement from human ex- pert annotations. Similarly, hierarchical CNN approach for APL detection gave high precision and recall in a previous

study [6], demonstrating the utility of CNNs in rare AML subtype detection, especially in low-resource settings without immediate genetic testing. Our study achieved comparable performance metrics, with an 85% accuracy for mutation classification, addressing similar diagnostic challenges but with a focus on resilience to label noise. By employing noise- handling techniques and smooth cross-entropy loss function, our model demonstrated stable performance under approxi- mately 20% noisy labels. Unlike SCEMILA and Eckardt's models, which focused on attention and cell-specific fea- ture extraction, our approach provides broader applicability across multiple mutations (e.g., NPM1, RUNX1::RUNX1T1) in noise-prone environments. This focus on robust classi- fication amidst label inconsistencies enhances the model's potential for clinical integration, allowing for reliable muta- tion prediction without extensive label refinement, a common barrier in AML cytomorphology applications.

Label noise is a prevalent challenge in medical imaging, especially in tasks like cancer detection, where data labeling requires domain-specific expertise and is often prone to er- rors. To mitigate the effects of noisy labels, various strategies and models have been implemented to ensure robustness and improve diagnostic reliability in deep learning applications. Regularization methods, such as robust loss functions (e.g., cross-entropy with noise handling adjustments) and dropout layers, prevent overfitting to noisy data by focusing the model on reliable signal patterns. This method is particularly valu-able in image classification where features vary significantly, and the goal is to avoid learning from anomalous data due to noise. Hybrid methods like semi-supervised learning use a mix of labeled and unlabeled data, where the latter often pro- vides cleaner representations for model training. Contrastive learning, by focusing on distinguishing between similar and dissimilar pairs, can help the model ignore noisy labels by emphasizing the underlying structure of images rather than la- bels, which is useful in cytomorphology and medical imaging with inherent noise. In complex medical datasets where cer- tain image features may inherently lead to more frequent mis- labeling (e.g., in AML cell variations), instance-dependent noise modeling has proven effective. This approach tailors noise handling to specific data characteristics, thus provid- ing an advantage in nuanced datasets like medical images of bone marrow samples, where noise distribution varies at the instance level.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made publicaly available . Ethical approval was not required as confirmed by the license attached with the open access data.

6. FUNDING

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. DATA AVAILABILITY STATEMENT

No datasets have been generated. However the datasets analysed during the current study are publicaly availaible datsets available in The Cancer Imagimng Archive repository (https://www.cancerimagingarchive.net/collection/aml-cytomorphology_mll_helmholtz/) and AML-Cytomorphology and (https://www.cancerimagingarchive.net/collection/bone-marrow-cytomorphology_mll_helmholtz_fraunhofer/).

8. REFERENCES

1. Garima Jain, Deepshi Thakral, Ranjit Kumar Sahoo, Indresh Kumar, Sangeeta Vashishtha, Pramod Verma, and Ritu Gupta, "Next generation sequencing guided treatment modulation and prognosis in acute myeloid leukemia: Case vignettes," *American Journal of Blood Research*, vol. 10, no. 4, pp. 134, 2020.

2. Jacqueline Kockwelp, Sebastian Thiele, Jannis Bartsch, Lars Haalck, Jo¨rg Gromoll, Stefan Schlatt, Rita Ex- eler, Annalen Bleckmann, Georg Lenz, Sebastian Wolf, et al., "Deep learning predicts therapy-relevant genet- ics in acute myeloid leukemia from pappenheim-stained bone marrow smears," *Blood advances*, vol. 8, no. 1, pp. 70–79, 2024.

3. Jason H Kurzer and Olga K Weinberg, "Updates in molecular genetics of acute myeloid leukemia," in *Seminars in Diagnostic Pathology*. Elsevier, 2023, vol. 40, pp. 140–151.

4. Hartmut Do¨hner, Andrew H Wei, Frederick R Appelbaum, Charles Craddock, Courtney D DiNardo, Herve´ Dombret, Benjamin L Ebert, Pierre Fenaux, Lucy A Godley, Robert P Hasserjian, et al., "Diagnosis and management of aml in adults: 2022 recommendations from an international expert panel on behalf of the eln," *Blood, The Journal of the American Society of Hematology*, vol. 140, no. 12, pp. 1345–1377, 2022.

5. Matthias Hehr, Ario Sadafi, Christian Matek, Peter Lienemann, Christian Pohlkamp, Torsten Haferlach, Karsten Spiekermann, and Carsten Marr, "Explainable ai identifies diagnostic cells of genetic aml subtypes," *PLOS Digital Health*, vol. 2, no. 3, pp. e0000187, 2023.

6. Jan-Niklas Eckardt, Tim Schmittmann, Sebastian Riechert, Michael Kramer, Anas Shekh Sulaiman, Katja Sockel, Frank Kroschinsky, Johannes Schetelig, Lisa Wagenfu¨hr, Ulrich Schuler, et al., "Deep learning identifies acute promyelocytic leukemia in bone marrow smears," *BMC cancer*, vol. 22, no. 1, pp. 201, 2022.

7. S Jian, H Kaiming, R Shaoqing, and Z Xiangyu, “Deep residual learning for image recognition,” in *IEEE Con- ference on Computer Vision & Pattern Recognition*, 2016, pp. 770–778.

8. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vi- sion and Pattern Recognition*, 2009, pp. 248–255.

9. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, “Rethinking the in- ception architecture for computer vision,” in *Proceed- ings of the IEEE conference on computer vision and pat- tern recognition*, 2016, pp. 2818–2826.

10. Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recogni- tion,” in *International Conference on Learning Repre- sentations*, 2015.

11. Zhuang Liu, Jiangfeng Wang, Xie Si, and et al., “Con- vnext: Revisiting convolutional neural networks for vi- sual recognition,” *arXiv preprint arXiv:2201.03545*, 2022.

12. Ze Liu, Yutong Lin, Yuankai Qi, and et al., “Swin trans- former: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF Interna- tional Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

13. Mingxing Tan and Quoc Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

14. Zhanwu Lv, Xinyi Cao, Xinyi Jin, Shuangqing Xu, and Huangling Deng, “High-accuracy morphological identi- fication of bone marrow cells using deep learning-based morphogo system,” *Scientific Reports*, vol. 13, no. 1, pp. 13364, 2023.

15. Laura Boldu´, Anna Merino, Andrea Acevedo, Angel Molina, and Jose´ Rodellar, “A deep learning model (al- net) for the diagnosis of acute leukaemia lineage using peripheral blood cell images,” *Computer Methods and Programs in Biomedicine*, vol. 202, pp. 105999, 2021.

Figures

Image not available with this version

Figure 1

Labeled single cell images from Dataset 1 were grouped into two bags i.e. cancer vs non-cancer, which is then used to train a tumor detector model

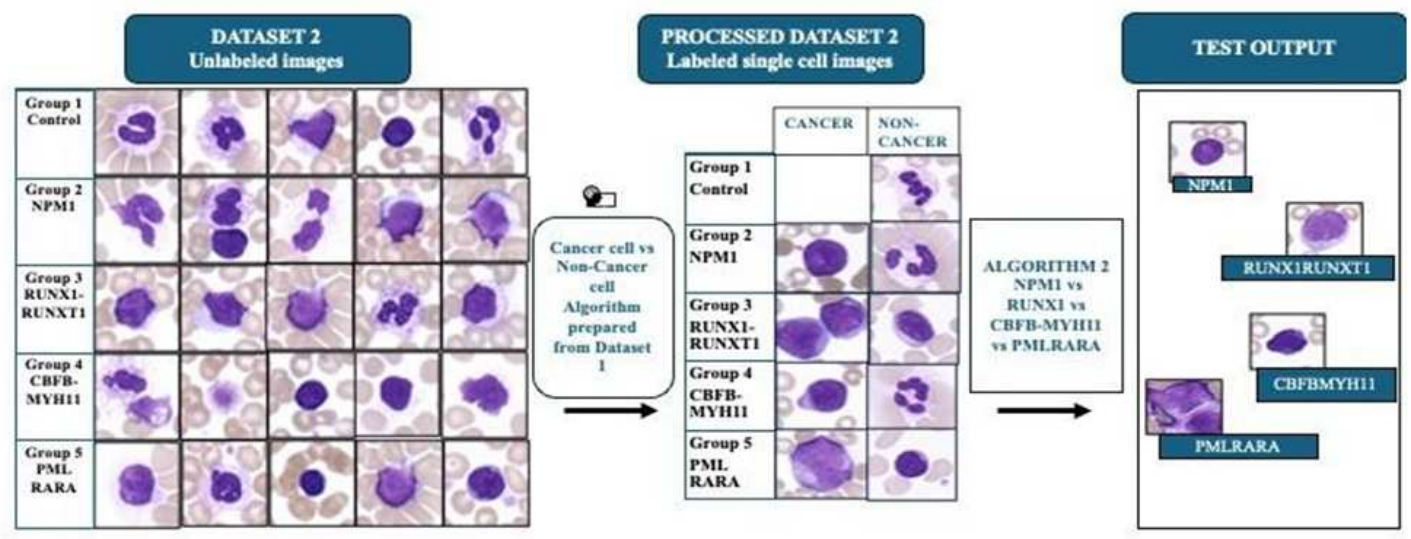


Figure 2

Unlabeled images of Dataset 2 were labeled with the help of algorithm 1. Labels cross-verified by two subject experts. Single cell images labeled as cancer were fed to mutation prediction network.