Annotation-Free Deep Learning for Predicting Gene Mutations from Whole Slide Images of Acute Myeloid Leukemia

Bo-Han Wei[1,2], **Xavier Cheng-Hong Tsai**[3,4,5,6], **Min-Yen Lo**[7], **Sheng-Yu Hung**[8], **Wen-Chien Chou**[3,8], **Hwei-Fang Tien**[3,9], **Hsin-An Hou**[3], and **Chien-Yu Chen**[1,2,10*]

[1] *Center for Advanced Computing and Imaging in Biomedicine, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City 106319, Taiwan (R.O.C.)*
[2] *Department of Biomechatronics Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City 106319, Taiwan (R.O.C.)*
[3] *Division of Hematology, Department of Internal Medicine, National Taiwan University Hospital, No.1, Changde St., Zhongzheng Dist., Taipei City 100229, Taiwan (R.O.C.)*
[4] *Department of Medical Education and Research, National Taiwan University Hospital Yunlin Branch, No. 579, Sec. 2, Yunlin Rd., Douliu City, Yunlin County 640203, Taiwan (R.O.C.)*
[5] *Department of Medical Education and Research, National Taiwan University Hospital Yunlin Branch, No. 579, Sec. 2, Yunlin Rd., Douliu City, Yunlin County 640203, Taiwan (R.O.C.)*
[6] *Department of Hematological Oncology, National Taiwan University Cancer Center, National Taiwan University Hospital, No.57, Ln. 155, Sec. 3, Keelung Rd., Da'an Dist., Taipei City 106, Taiwan (R.O.C.)*
[7] *Division of Hematology, Department of Internal Medicine, National Taiwan University Hospital Yunlin Branch, No. 579, Sec. 2, Yunlin Rd., Douliu City, Yunlin County 640203, Taiwan (R.O.C.)*
[8] *Department of Laboratory Medicine, National Taiwan University Hospital, No.1, Changde St., Zhongzheng Dist., Taipei 100229, Taiwan (R.O.C.)*
[9] *Department of Internal Medicine, Far-Eastern Memorial Hospital, New Taipei City, Taiwan, No. 21, Section 2, Nanya S. Road, Banqiao District, New Taipei City, 220, Taiwan (R.O.C.)*
[10] *Smart Medicine and Health Informatics Program, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City 106319, Taiwan (R.O.C.)*

Abstract

The rapid development of deep learning in recent years has revolutionized the field of medical image processing, including the applications of using high-resolution whole slide images (WSIs) in predicting gene mutations in acute myeloid leukemia (AML). Although the potential of characterizing gene mutations directly from WSIs has been demonstrated in some studies, it still faces challenges due to memory limitations and manual annotation requirements. To address this, we propose a deep learning model based on multiple instance learning (MIL) to predict gene mutations from AML WSIs with no patch-level or cell-level annotations. The proposed MIL-based deep learning model offers a promising solution for gene mutation prediction on *NPM1* mutations and *FLT3*-ITD. With the property of annotation-free, the proposed method eliminates the need for manual annotations, reducing the manpower and time costs associated with traditional patch-based or cell-based approaches. We assessed our MIL models using a dataset of 572 WSIs from AML patients. By exclusively utilizing annotation-free WSIs for cell-level training, we achieved an AUC of 0.75 for predicting *NPM1* mutations and 0.68 for *FLT3*-ITD. Furthermore, upon applying upsampling and ensemble techniques to address the data imbalance issue, the AUC improved from 0.75 to 0.86 for *NPM1* mutations and from 0.68 to 0.82 for *FLT3*-ITD. These enhancements, leading to more precise predictions, have brought AML WSI analysis one step closer to being utilized in clinical practice.

Keywords: *Multiple instance learning (MIL), Acute myeloid leukemia (AML), whole slide images (WSIs), NPM1 mutations, FLT3-ITD*

*Corresponding author. *E-mail address:* chienyuchen@ntu.edu.tw

1. Introduction

Bone marrow evaluation, including morphology, immunophenotyping, molecular studies, and flow cytometry, is very critical for patients with hematologic malignancies. The integration of this information, combined with clinical presentation, enables physicians to arrive at a definitive diagnosis. Gene mutation analysis, as one of the fundamental diagnostic tools, serves as the cornerstone for diagnosis, prognosis assessment, and the identification of potential therapeutic targets. Nevertheless, the identification of gene mutations often relies on time-consuming methods with limited accessibility, such as next-generation sequencing (NGS).

Whole slide images (WSIs) employ digital imaging technology to transform pathological specimens into high-resolution and exceptionally precise digital images, providing intricate details of cellular and histological structures [1]. These attributes embody the tissue's morphological characteristics, which hold significant importance in various critical clinical decision-making points. Moreover, the ability to discern subtle morphological alterations in bone marrow cells is essential for diagnosing and assessing the risk of individuals with myelodysplastic syndrome (MDS).

In recent years, deep learning has been demonstrated to perform tasks such as binary morphological classification, distinguishing between tumor and normal tissues, and histological grading in numerous studies by utilizing WSIs [2–4]. However, challenges abound in the realm of bone marrow aspirates, given the complexity of cytological specimens. Aspirates contain only small regions suitable for cytology, with significant non-cellular debris and numerous distinct cell types often clustered or overlapping. The inherent complexity of bone marrow cytology poses a notably intricate challenge in the field of computational pathology. An essential hurdle involves identifying areas of interest (ROIs), which entails selecting a few cytology-appropriate regions or patches from extensive WSIs. Furthermore, the task encompasses distinguishing individual bone marrow cells or non-cellular objects as distinct entities separated from the background. Both of these issues necessitate multi-step image preprocessing, including segmentation and denoising. Although deep learning has been employed for object detection [5,6], there remains a requirement for experts to manually annotate a considerable number of segmented single cells, which is a labor-intensive and time-consuming endeavor. Hence, it is imperative to develop a more efficient and accurate method for the analysis of bone marrow aspirate WSIs.

In contemporary histopathology research, a common approach involves a two-stage workflow, emphasizing patch-level and slide-level training [7]. Initially, a Convolutional Neural Network (CNN) is trained on patches extracted from WSIs with patch-level annotations. This CNN learns the complex patterns within these patches. In the second stage, the features learned at the patch level are utilized to train a slide-level model. This slide-level model plays a pivotal role in discerning the diagnosis from the WSIs, capitalizing on the insights gleaned from the patch-level analysis. These two-stage methods have been widely used for tasks such as cancer identification [8–10], classification [11,12], and metastasis detection [8]. Nevertheless, these approaches require substantial manual annotations from pathologists. To directly harness slide-level labels, Multiple Instance Learning (MIL) has been employed [13,14]. n slide-level cancer classification using MIL, if the patch with the highest score is cancerous, the entire slide is classified as cancer. Conversely, if the highest-scoring patch is normal, the slide is classified as benign. By using slide-level ground truth as weak supervision, MIL effectively reduces the annotation burden.

We hypothesized that deep learning could be employed to predict gene mutations based on cellular morphology. In this study, we present an end-to-end artificial intelligence framework for bone marrow cytology. Unlike the conventional two-stage approach, this model is uniquely trained using WSIs with slide-level annotations. We aim to introduce a training approach that leverages annotation-free WSIs for gene mutation predictions, demonstrating the capability of deep learning in predicting gene mutations based on cellular morphology. We aim to highlight that, in the realm of AML WSIs, models trained at the cell level outperform those trained at the patch level. Additionally, we aim to illustrate how the techniques, such as upsampling and ensemble learning, can enhance the predictive performance of the model, especially in scenarios with limited training data.

2. Methods

2.1. Dataset

Between 1994 and 2015, 572 patients who were diagnosed with *de novo* AML at the National Taiwan University Hospital (NTUH) were enrolled in this study. Bone marrow smears, as well as peripheral blood smears, were scanned as WSIs after staining with hematoxylin and eosin (H&E) (Fig. 1). The gene mutation status was determined using NGS [15], which were annotated as 1 (indicating the presence of pathogenic mutations) or 0 (indicating the absence of pathogenic mutations). In our cohort, a total of 34 genes were frequently found to have mutations, and in this study, we listed the mutation frequencies of more than 10% and selected the top two highest frequent mutations as the targets (Table 1). This retrospective study was approved by the NTUH Research Ethics Committee, and written informed consent was obtained from all participants in accordance with the Declaration of Helsinki (Approval number: 201802021RINC).

Table 1. Number of WSIs with gene mutations in the our cohort

Gene mutations	WSI Number	Percentage
FLT3-ITD	103	18.0%
NPM1	101	17.7%
DNMT3A	99	17.3%
NRAS	92	16.1%
ASXL1	86	15.0%
RUNX1	84	14.7%
CEBPA	83	14.6%
TET2	78	14.5%
IDH2	70	12.2%
TP53	60	10.5%

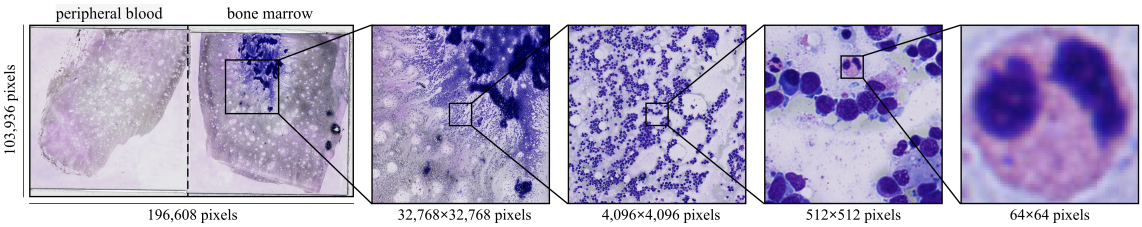


Figure 1. Example of whole slide image. The WSI has an original resolution of approximately 200,000 × 100,000 pixels.

2.2. Cell Image Generation

The bone marrow smears of WSIs underwent a three-step filtering process to identify the cells for subsequent model training. First, we used the PyHIST tool to generate patches [16]. This tool was applied to filter out background regions and non-smear areas in the WSIs (Fig. 2a). The patches (512×512 pixels) were extracted from the highest resolution (40X magnification), and the generation method was based on the graph method, with a content threshold of 0.05. The parameters 'tilecross-downsample' and 'mask-downsample' adopted the default values, and the parameter 'output-downsample' was set to 1 in order to obtain patches at the original resolution. Secondly, the patches generated from the PyHIST tool were further classified into ROI and non-ROI patches. We utilized an ROI detection model from a previous study [17], which was based on DenseNet121 [18] and was pre-trained and fine-tuned on labeled patches (ROI/non-ROI). In this step (Fig. 2b), patches containing densely packed leucocytes or areas without any leucocytes were removed, significantly reducing the number of input data for the subsequent step: cell detection modeling.

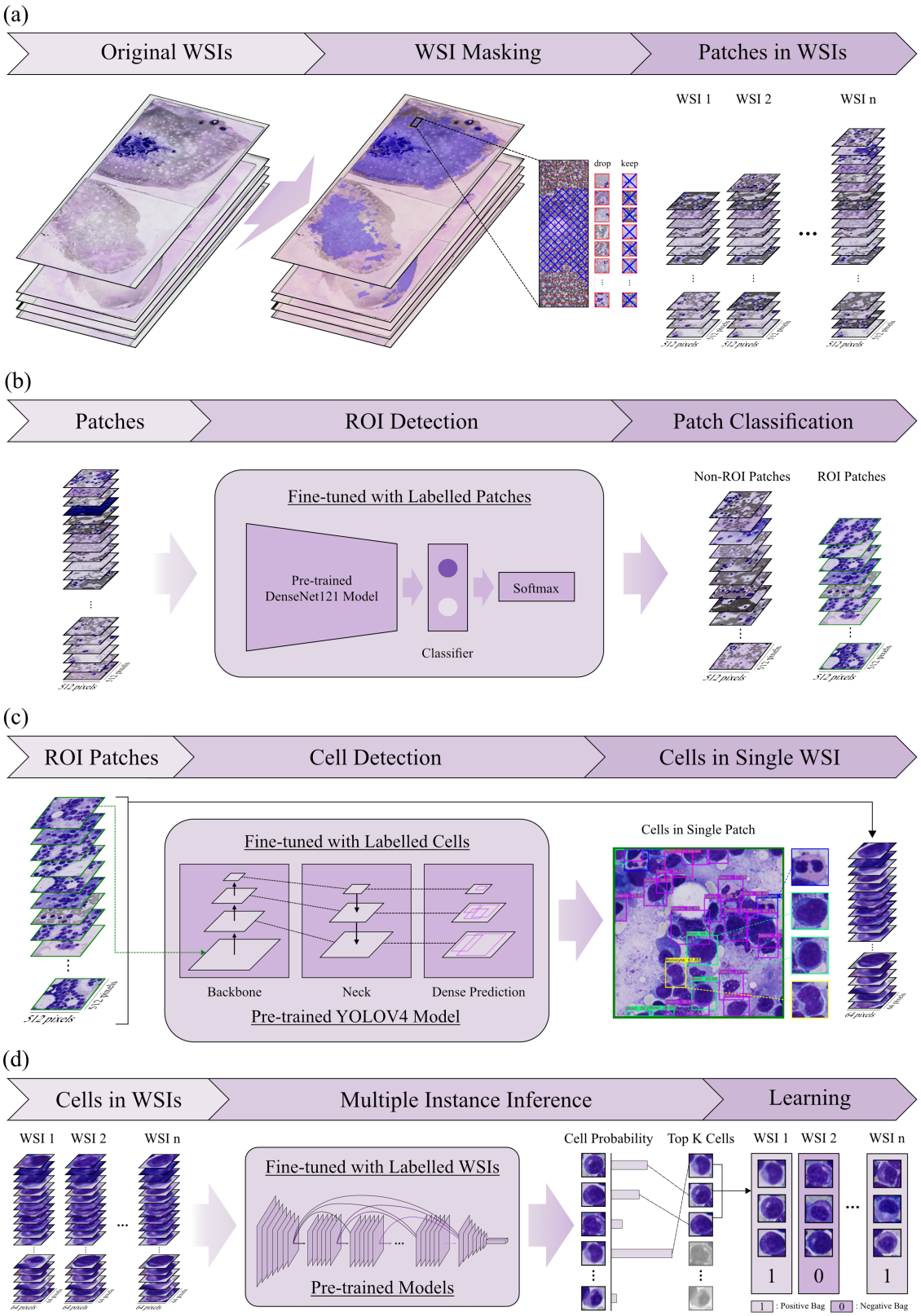


Figure 2. Overview of the proposed method. (a). generation of patch images; (b). selection of ROI patches; (c). leucocyte (cell) detection; and (d). MIL training

After ROI selection, the ROI patches in each WSI would be used to get the cells in the third step. We applied the cell detection model from the same research of ROI detection [17]. This model was based on YOLOv4 [19] and trained to predict the bounding boxes of leucocyte objects. In this step, all leucocytes in each ROI patch in bone marrow aspirates were detected (Fig. 2c).

2.3. Multiple Instance Learning

Fully supervised approaches for histopathology image analysis require detailed manual annotations, which are time-consuming and intrinsically ambiguous, even for well-trained experts. Standard unsupervised approaches usually fail due to their complicated patterns.

MIL is a weakly supervised learning approach used in tasks where the training data consists of labeled groups of instances, known as bags, rather than individually labeled instances. It works well for the current study because it takes advantage of both supervised and unsupervised approaches. The main idea of MIL is to learn local patterns using global annotations. In MIL, each bag contains multiple instances, but only the bag is labeled with a class label, and the instances within the bag are unlabeled. In previous research, they transformed binary classification tasks into MIL problems by dividing WSIs into multiple instances. Here, the instances could be patches [13] or could be cells as proposed in this study. Positive WSIs contain at least one positive instance, while negative WSIs are without any positive instances. With the above two criteria, we expect positive WSIs to contain only a few key instances, while negative WSIs can have all the patches or cells as instances.

In this study, the term "bags" refers to WSIs, while "instances" denotes distinguishable cells. We categorize WSIs with and without mutations as mutated and standard bags, respectively. Each WSI can be seen as a bag containing multiple cells (Fig. 2d). To predict a bag as mutated, at least one cell should be classified as mutated. On the other hand, all cells should be classified as standard when predicting a bag as standard. In a bag, all cells will be classified, and they are ranked based on their probability of being classified as mutated. For mutated bags, the probability of the highest-ranked cell should be higher than a threshold, while for standard bags, it should be lower than that threshold. The MIL task involves learning a cell-level representation that effectively discriminates distinguishing cells within mutated WSIs from all other cells. Based on the cell detection process described above, bags $B = \{B_i : i = 1, 2, \ldots, n\}$ are generated, where $B_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m_i}\}$ is a bag containing m_i cells from WSI s_i.

During training, the embedding models AlexNet, DenseNet, EfficientNet, ResNet34, ResNet101, ResNet152, ResNext, and VGG16 were employed and compared. These models are initialized with pre-trained weights from ImageNet. The model is represented as a function f_θ, where the current parameters θ map the input cells $c_{i,j}$ to probabilities of "standard" and "mutated" classes. For a bag B_i, a vector list $P = \{p_{i,1}, p_{i,2}, \ldots, p_{i,m_i}\}$ is obtained, with each vector corresponding to the probability of the "mutated" class for each cell $c_{i,j}$ ($j = 1, 2, ..., m_i$, where m_i is the number of cells in B_i) in WSI s_i. Once we have obtained the index k_i of the highest-ranked "mutated" cell for each WSI, denoted as $k_i = argmax_j(p_{i,j}), j = 1, 2, ..., m_i$, we proceed to label this specific cell as "mutated." This represents the strictest version of MIL, but the standard MIL assumption can be relaxed by introducing a hyperparameter K, assigning the top K cells as mutated in mutated WSIs. For $K = 1$, the highest-ranked cell in bag B_i is c_{i,k_i}. Then, the network's output $\widetilde{y}_i = f_\theta(b_{i,k_i})$ is compared to the label y_i of WSI s_i using the cross-entropy loss function l as in Eq. 1:

$$l = -w_1[y_i log(\widetilde{y}_i)] - w_0[(1 - y_i)log(1 - \widetilde{y}_i)], \tag{1}$$

where w_0 and w_1 are the weights for standard and mutated classes, the sum of these two weights must equal 1. Similarly, if $K > 1$, all the top K cells from a WSI share the same label y_i, and the loss can be computed with Eq. 1 for each one of the K cells.

2.4. Data Upsampling

The class imbalance in the WSI data was indeed substantial, with a limited number of samples exhibiting gene mutations compared to standard samples, which could have a significant impact

on training outcomes. To address this imbalance, we introduced class weights, denoted as w_0 and w_1 for the standard and mutated classes, respectively. These weights allowed us to assign greater importance to the under-represented mutated examples.

Furthermore, prior to MIL training, we faced the challenge of working with WSIs, each containing tens of thousands of cells. To expedite the training process, we randomly selected a fixed number of cells to represent each WSI, reducing the input quantity. Since the number of mutated WSIs was considerably smaller compared to the negative class, we adopted a strategy of dividing mutated WSIs into multiple bags, each containing the same fixed number of cells. These bags were considered independent training data, effectively increasing the training data for the mutated class and resulting in a more balanced training dataset.

2.5. Ensemble Learning

Ensemble methods represent a potent approach that encompasses the training and amalgamation of multiple models to tackle intricate problems. At its heart, ensemble methods rely on the fundamental concept that a gathering of individual "weak learners" can synergistically give rise to a "strong learner." In this strategy, each model casts its vote or contributes valuable insights, with the ensemble method harmonizing these inputs to formulate a definitive prediction. The overarching objective of ensembles is to mitigate bias and variance in predictions. By harnessing the combined capabilities of multiple models instead of relying solely on a single one, these methods enhance accuracy and robustness significantly.

In this study, we utilized an ensemble method based on a loss function approach. This technique aimed to optimize the model combination of each base model $M_K (K = 1, 3, 5, 10, 20$, where K is the top K cells) by employing a loss function l_{M_K} as in Eq. 1 to quantify the difference between predicted and actual values. Each base model within the ensemble was assigned a weight W_{M_K} determined by its performance as in Eq. 2:

$$W_{M_K} = \frac{exp(-l_{M_K})}{\sum_{i \in \{M_K, K=1,3,5,10,20\}} exp(-l_i)} \tag{2}$$

The ensemble model then combined the probabilities P_{M_K} of these models according to their respective weights, as illustrated in Eq. 3:

$$P_{ensemble} = \sum_{i \in \{M_K, K=1,3,5,10,20\}} W_i \times P_i \tag{3}$$

This approach ultimately led to improved prediction accuracy.

3. Results

3.1. Automatic Selection of ROI Patches

In digital pathology, glass slides of bone marrow aspirate smears were scanned using a digital slide scanner to generate high-resolution WSIs for hematopathologist analysis. To initiate this process, a dataset was sampled from 572 bone marrow aspirate WSIs obtained within two years at the National Taiwan University Hospital. To address the issue of detecting ROI patches, we developed a pipeline to select ROI patches in stages, automatically identifying areas within the bone marrow aspirate WSIs that are suitable for cytological analysis.

A single WSI might contain only a few areas suitable for cytology. These areas are sparsely distributed, with minimal cell overlap and staining artifacts, and exhibit subtle and intricate cytological features required for cell classification. To acquire such areas efficiently, the PyHIST tool [16] was first used with the default graph method to remove the blank background regions,

retaining only the stained portions. Subsequently, we employed a fine-tuned DenseNet121 architecture, utilizing pre-trained weights from previous research to classify individual patches as ROI patches (appropriate patches) or non-ROI patches (inappropriate patches) [17].

We observed that the classification results aligned with expectation, i.e., a real-world scenario where usually only 10-20% of a WSI might be the ROI regions for cytology. The results obtained after applying DenseNet121 reveal that for most WSIs, the ROI patches were reduced to around 10-25% of their original counts after undergoing this selection process, indicating a substantial reduction in patch counts (Fig. 3). This screening process significantly reduced the time required for subsequent cell detection processes and exhibited outstanding filtering effectiveness in removing areas with excessive cell overlap, excessive tissue artifacts, or patches devoid of blood cell presence within the designated range. These problematic patches were effectively eliminated, while patches with a minimal number of cells were not erroneously dropped (Fig. 4).

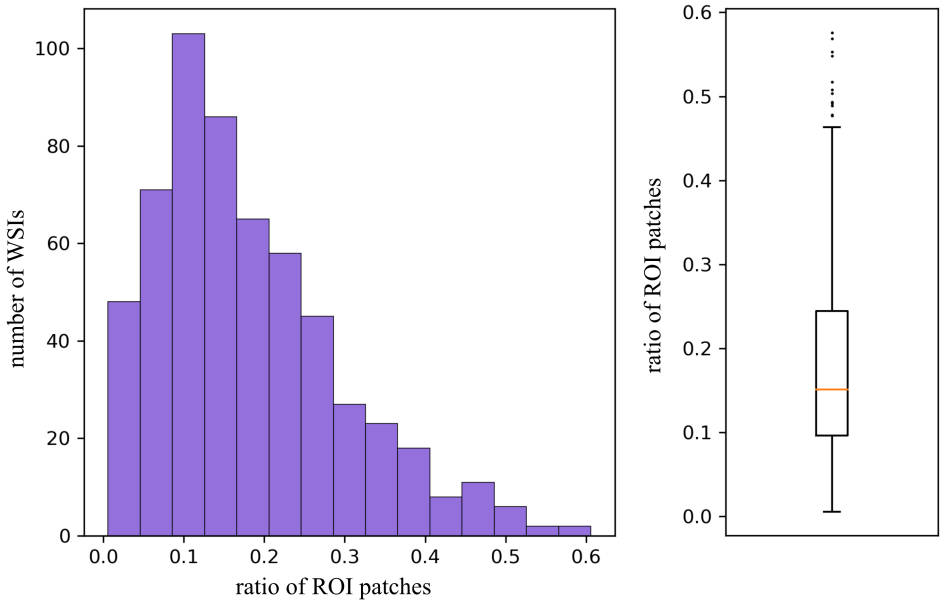


Figure 3. Distribution (histogram and boxplot) of the selected ROI patches.

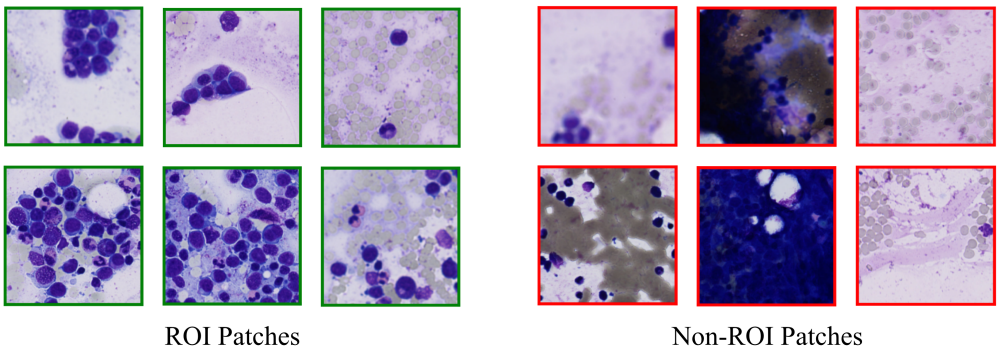


Figure 4. Examples of the selected ROI patches versus non-ROI patches. ROI patches cover areas with single or multiple leucocytes that are evenly distributed, clear, and not blurry. Non-ROI patches contain only red blood cells but no leucocytes, regions with cell overlaps making identification challenging, a considerable mix of heterogeneous tissues, areas devoid of blood cells within the designated range, or blurry patches.

3.2. Leukocyte Detection

After completing the ROI selection process, we employed a YOLOv4 model to automatically detect and classify cells and non-cellular objects within the selected ROI patches. The training weights in the same study of selecting ROI patches were used to fine-tune this YOLOv4 model [17], with DenseNet model-identified ROI patches used as input. The YOLOv4 model aimed to detect various leucocytes, including neutrophils, metamyelocytes, myelocytes, promyelocytes, blast, lymphocytes, monocytes, eosinophils, and basophils. For object predictions, a confidence threshold of 0.5 was set. Cells with confidence scores below this threshold were not captured, ensuring high-quality cell selection for subsequent MIL model training. In addition to managing confidence levels, we factored in cell size during the process. Cells were chosen if their sizes were within the range of 51×51 pixels to 80×80 pixels, where the patch size is 512×512. The chosen cells were then uniformly resized to 64×64 pixels, ensuring a consistent input data size for convenient execution of subsequent augmentation steps (fig. 5).

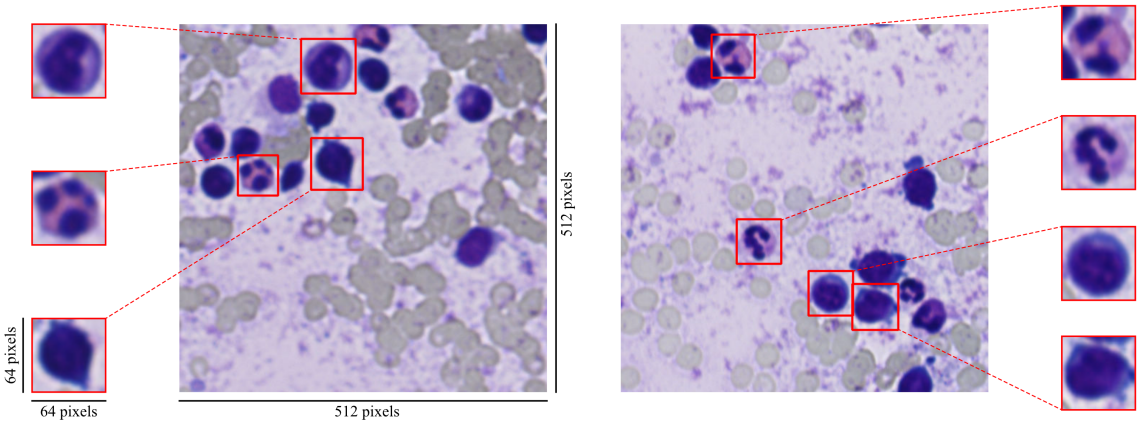


Figure 5. Examples of cell detection. The resolution of the patches is 512×512 pixels. The size of filtered cells is within the range of 51×51 pixels to 80×80 pixels and then resized to a uniform size of 64×64 pixels.

Following the YOLOv4 model detection and simple threshold-based filtering, the retained cells were selected to represent the entire WSI characteristics comprehensively while maintaining the best possible quality. The number of cells left per WSI varied between 100 to 100,000, with the majority of cell counts being below 20,000 and an average of 11,273 (Fig. 6). This outcome effectively reduced the input quantity for the MIL model, accelerating training time without compromising the accuracy. For WSIs with fewer than 2,000 cells, we applied augmentation to the selected cells to maintain the minimum number of cells for each WSI. The augmentation processes included flipping, translation, rotation, and color adjustments such as random contrast (scaling by 0.5–1.5), brightness (scaling by 0.65–1.35), hue (shifting by -32–32), and value (shifting by -32–32). The augmented data helped to enhance the robustness of the model.

3.3. Patch-Level Multiple Instance Learning

In order to demonstrate the contribution of cell-level MIL, this study included a patch-level MIL implementation as a comparison. The total dataset of 572 WSIs was split into training, validation, and test sets at slide level, with a ratio of around 6:2:2 (343:113:116 WSIs). We compared up-sampling and different embedding models separately to validate the effectiveness of our approach. For comparison, we used the ResNet34 with mini-batches of size 512, initialized with pre-trained weights from ImageNet. To emulate a real-world scenario, we evaluated the models on imbalanced data, specifically, more standard WSIs than mutation WSIs. Each epoch was evaluated using a validation set of 113 WSIs without applying upsampling, with a total of 100 epochs for training. Early stopping was implemented to prevent overfitting. The separate testing set included 96 standard WSIs and 20 *NPM1* mutations WSIs, along with 95 standard WSIs and 21 *FLT3*-ITD WSIs.

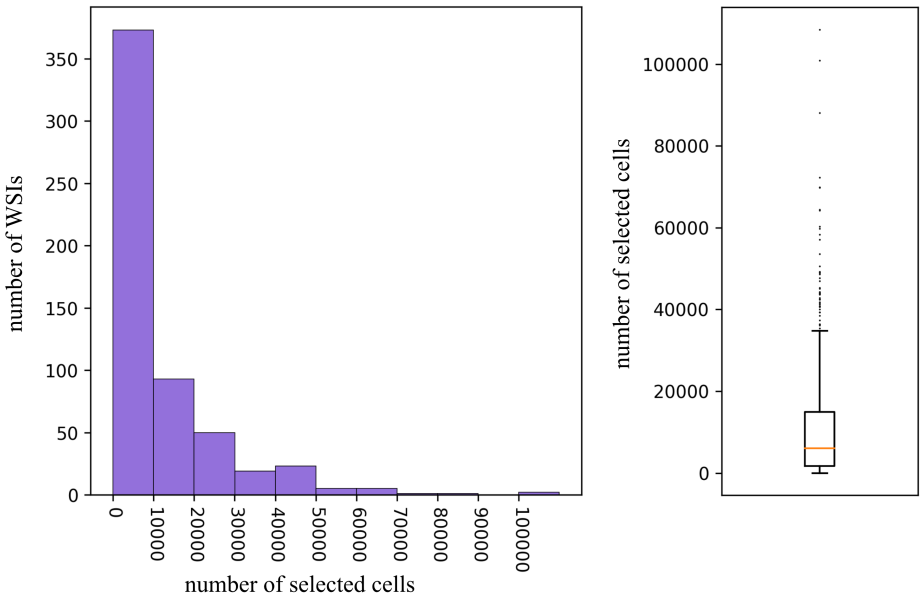


Figure 6. Distribution (histogram and boxplot) of the selected cells

3.3.1. Upsampling Performance

We maintained the use of ResNet34 with the same parameters but modified the training data to include the upsampled dataset. The upsampling method was applied by configuring each bag to comprise 500 patches, meaning that a WSI can generate multiple bags, each containing exactly 500 exclusive patches. This method effectively increased the number of mutated bags in the *NPM1* mutations training dataset from the original 60 to 344, resulting in the total number of training bags increasing from 343 to 627. For *FLT3*-ITD, a similar approach effectively raised the number of mutated bags in the training dataset from 61 to 469, and the total training bags increased from 343 to 751. Since both the non-upsampling and upsampling data still exhibited class imbalance, we fine-tuned the training parameters using weights $(w0, w1)$ in Eq. 1. Without upsampling, the weights were set to $(0.17, 0.83)$ for *NPM1* mutations and $(0.18, 0.82)$ for *FLT3*-ITD. With upsampling, the weights changed to $(0.55, 0.45)$ for *NPM1* mutations and $(0.62, 0.38)$ for *FLT3*-ITD.

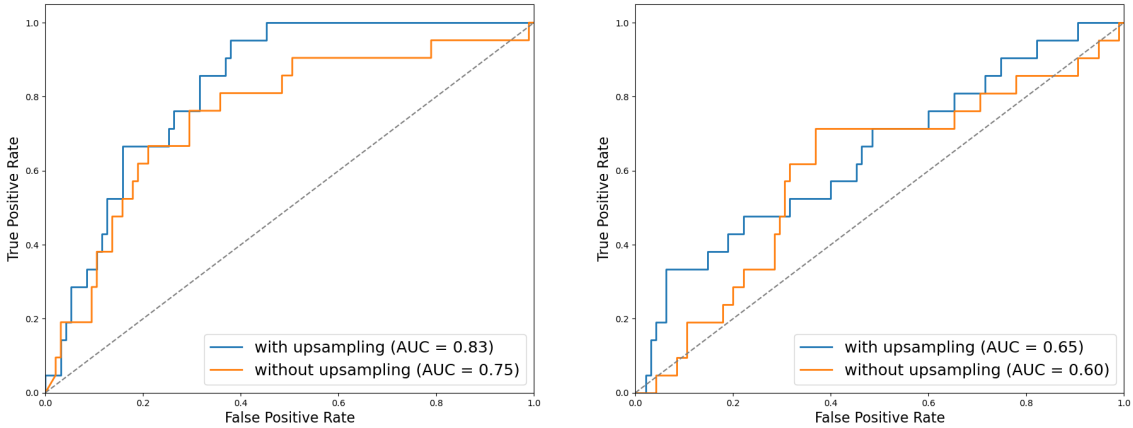
Here, we first demonstrated that the proposed upsampling method improved the model's performance at the patch-level MIL. Specifically, the AUC for *NPM1* mutations increased from 0.75 to 0.83, while for *FLT3*-ITD, it increased from 0.60 to 0.65. By comparing the performance of the two models, we assessed the impact of data balance on training results (Fig. 7).

3.3.2. Different Embedding Models

Furthermore, we employed various embedding models trained with the data of upsampling *NPM1* mutations. These models included AlexNet, DenseNet121, EfficientNetV2, ResNet34, ResNet50, ResNet101, ResNet152, ResNext, and VGG19, all initialized with pre-trained weights from ImageNet. The AUC values for these models, when tested with the testing set, are as follows: 0.67, 0.78, 0.73, 0.83, 0.79, 0.75, 0.71, 0.63, and 0.74, respectively (Fig. 8a). When comparing the performance of different models, it became evident that the ResNet and DenseNet models achieved better performance. Specifically, ResNet34 outperformed the others with an AUC of 0.83, which aligns with previous research findings [13].

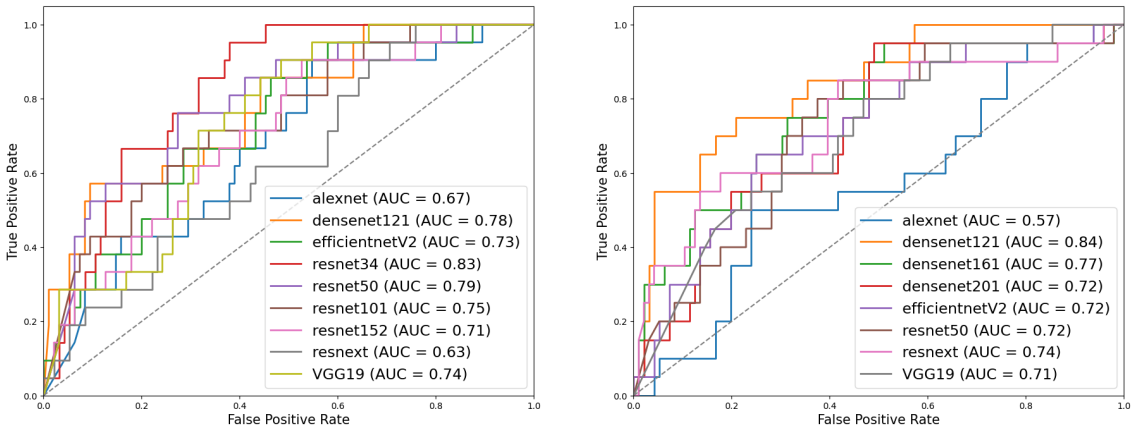
3.4. Cell-Level Multiple Instance Learning

Alongside the patch-level MIL model discussed earlier, we incorporated the cell-level models in this subsection. This was motivated by the observation that in the diagnosis of AML, pathologists



(a) *NPM1* mutations **(b)** *FLT3*-ITD

Figure 7. Performance of MIL models with upsampling versus without upsampling on ROC curves at patch level. a. *NPM1* mutations; b. *FLT3*-ITD.



(a) patch-level **(b)** cell-level

Figure 8. Comparison of using different pre-trained embedding MIL models. a. patch-level; b. cell-level.

routinely evaluate individual cells when analyzing bone marrow smears. To increase the size of the training dataset, we divided the entire dataset of 572 WSIs into training, validation, and test sets at the slide level, following a distribution ratio of around 7:1:2 (400:56:116 WSIs).

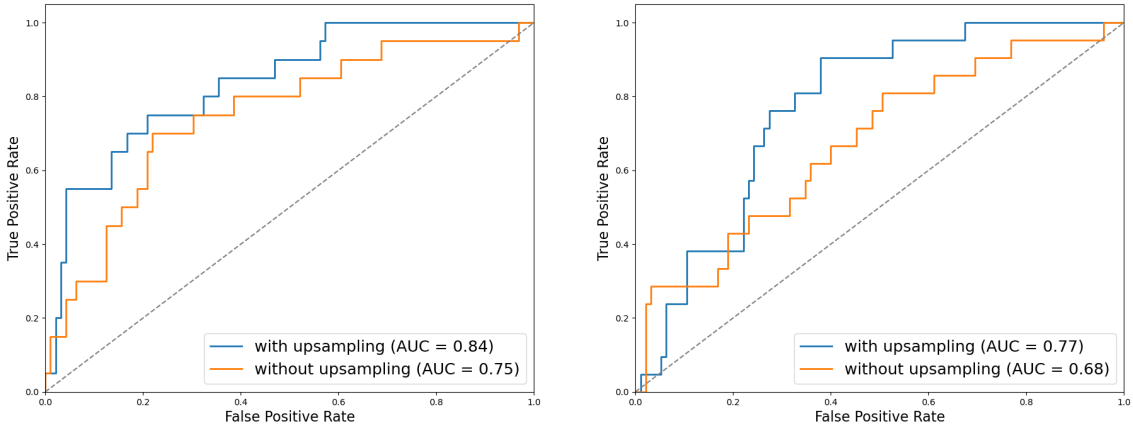
Due to the smaller image size of cells (64×64 pixels) at the cell-level MIL, we switched to utilizing the DenseNet121 as the first embedding model, with mini-batches of size 1024. This model was also initialized with pre-trained weights from ImageNet. Each epoch was evaluated using a validation set of 56 WSIs without applying upsampling, with 100 epochs for training. Early stopping was implemented to prevent overfitting. We evaluated these models on the same imbalanced testing dataset as the patch-level MIL.

3.4.1. Upsampling Performance

We next implemented the upsampling method to address the class imbalance issue and expand the training data. Given that the number of cells in a WSI is approximately 4-5 times that number of patches, we set the number of cells in a bag to 2000 and applied upsampling only to the mutation WSIs. This resulted in an increase in the number of mutated bags in the *NPM1* mutations training dataset from the original 70 to 323 and total training bags from 400 to 653. For *FLT3*-ITD, the mutated bags increased from 69 to 568, with the total bags increasing from 400 to 899. We

continued to use the training parameters weights $(w0, w1)$ to address the class imbalance, with values of $(0.17, 0.83)$ for *NPM1* mutations and $(0.17, 0.83)$ for *FLT3*-ITD without upsampling, while with upsampling, the values became $(0.49, 0.51)$ for *NPM1* mutations and $(0.63, 0.37)$ for *FLT3*-ITD.

In the cell-level experiment, we once again demonstrated that the proposed upsampling method can improve the model's performance. The AUC for *NPM1* mutations increased from 0.75 to 0.84, while for *FLT3*-ITD, it increased from 0.68 to 0.77. By comparing the performance of the two models, we assessed the impact of data balance on training results (Fig. 9).



(a) *NPM1* mutations **(b)** *FLT3*-ITD

Figure 9. Performance of MIL models with upsampling versus without upsampling on ROC curves at the cell level. a. *NPM1* mutations; b. *FLT3*-ITD.

3.4.2. Different Embedding Models

We then employed the upsampled *NPM1* data to train various embedding models with the goal of identifying the best pre-trained embedding model. These models included AlexNet, DenseNet121, DenseNet161, DenseNet201, EfficientNetV2, ResNet50, ResNext, and VGG19, all initialized with pre-trained weights from ImageNet. The AUC values for these models on the testing set are as follows: 0.57, 0.84, 0.77, 0.72, 0.72, 0.72, 0.74, and 0.71, respectively (Fig. 8b).

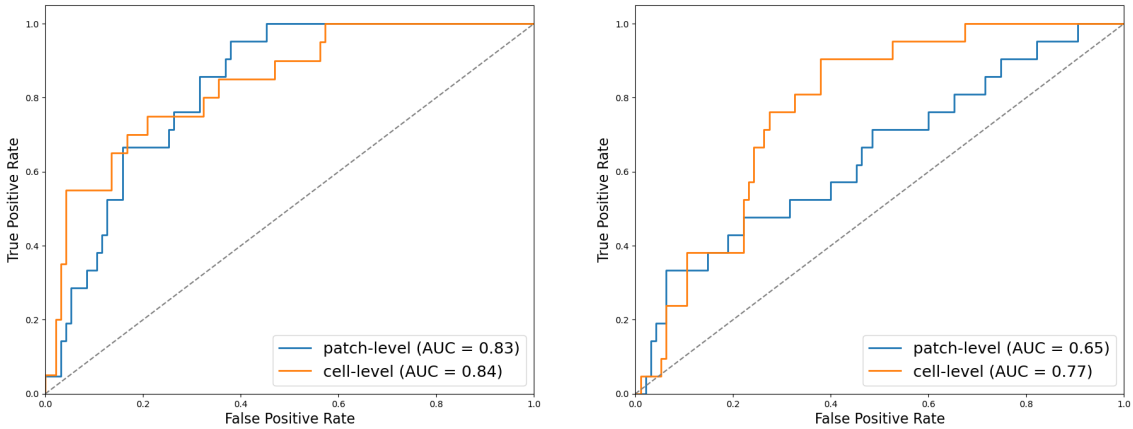
Given the 64×64 pixels size of the input images, some models, like the ResNet series and ResNext, needed to resize the images to 224×224 pixels in the first layer. This resizing could lead to inaccuracies in feature selection. Consequently, the DenseNet series of models, which could directly process 64×64 pixels input data, outperformed others. Among them, DenseNet121 exhibited the best performance with an AUC of 0.84.

3.4.3. Comparison with patch-level MIL

Moreover, we revealed that the cell-level training results were superior to the patch-level results. Without upsampling, *NPM1* mutations did not exhibit a significant improvement, but *FLT3*-ITD increased from 0.60 to 0.68. With upsampling, the differences were even more pronounced. *NPM1* mutations' AUC improved from 0.83 to 0.84, while *FLT3*-ITD increased from 0.65 to 0.77 (Fig. 10).

In clinical diagnosis, sensitivity represents the ability of a diagnostic tool to detect a positive result from individuals who are actually affected by the condition. We aim to maintain the same sensitivity while minimizing the probability of false positives. In this experiment, we set the sensitivity to 0.7 and compared the false positive rates between patch-level and cell-level MIL models. For the prediction of the *NPM1* mutations, the false positive rate decreased from 0.25 to 0.17. This result was even more significant in the case of the *FLT3*-ITD gene, where the false

positive rate decreased from 0.48 to 0.26. This outcome further validates the greater effectiveness of the cell-level MIL model in aiding clinical diagnosis, resulting in a lower probability of false positives in cases diagnosed as mutated.



(a) *NPM1* mutations **(b)** *FLT3*-ITD

Figure 10. Comparison of patch-level and cell-level MIL model performances on ROC curves. a. *NPM1* mutations: Patch level using the ResNet34 embedding model, and at the cell level using the DenseNet121 embedding model; b. *FLT3*-ITD: Patch level using the ResNet34 embedding model, and at the cell level using the DenseNet121 embedding model .

3.4.4. Ensemble of Different K Values in MIL

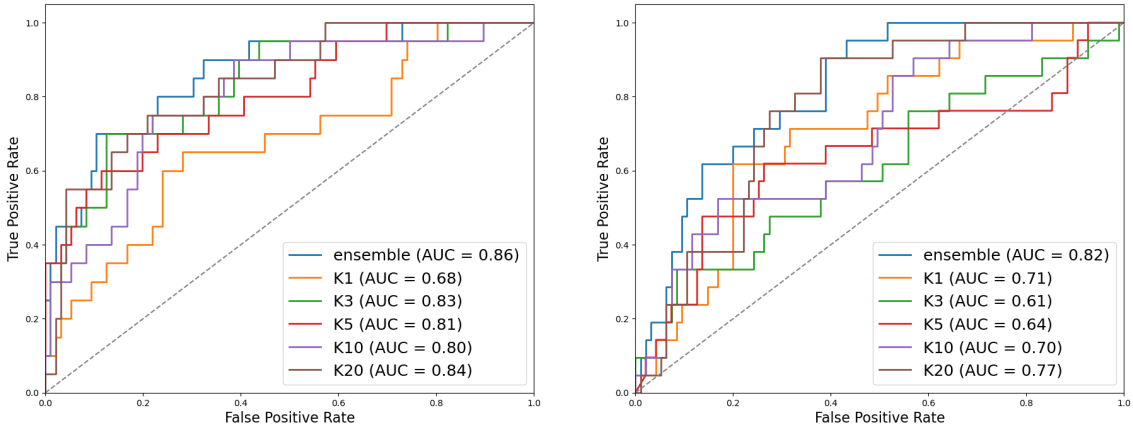
In ensemble learning, employing different values of K in MIL serves to diversify the models' perspectives on the data. Each K represents the number of positive instances considered within a bag, leading to the creation of models with varying focuses on distinct data subsets. When these models are combined through ensemble techniques, their diverse viewpoints contribute to a more comprehensive understanding of underlying data patterns.

In this study, we established minimum AUC standards for models to be included in the ensemble learning process, which were set at 0.80 or higher for *NPM1* mutations and 0.70 for *FLT3*-ITD. Models surpassing these predefined thresholds were selected to contribute to the ensemble based on their respective weights W_{M_K} calculated using Eq. 2. Ultimately, through Eq. 3, we achieved predictive results with an AUC of 0.86 for *NPM1* mutations and 0.82 for *FLT3*-ITD (Fig. 11). This approach ensured that only high-performing models were part of the ensemble, enhancing its overall robustness and accuracy. By adhering to these AUC criteria, we minimized the risk of overfitting and effectively harnessed the collective insights from diverse data facets, reaffirming the efficacy of ensemble learning techniques.

We also used sensitivity as the standard for the reliability analysis of clinical diagnosis. The results of the ensemble learning model were particularly remarkable in the case of *NPM1*, where the false positive rate decreased from 0.17 to 0.10 when setting sensitivity to 0.7. For *FLT3*-ITD, there was a slight reduction from 0.26 to 0.24. This outcome once again confirms that ensemble learning can enhance the effectiveness of our model.

4. Conclusion

In our research, we extensively explored Multiple Instance Learning (MIL) at both patch and cell levels to address the challenges in diagnosing acute myeloid leukemia (AML). Utilizing cell-level MIL with DenseNet121, we achieved impressive AUC values for *NPM1* mutations, reaching 0.84, and for *FLT3*-ITD, which achieved 0.77. To reduce false positive rates and maintain high sensitivity levels, we also incorporated upsampling and ensemble learning techniques, ensuring that only high-performing models contributed to the ensemble. These strategies yielded remarkable outcomes, with an impressive AUC of 0.86 for *NPM1* mutations and 0.82 for *FLT3*-ITD.



(a) *NPM1* mutations **(b)** *FLT3*-ITD

Figure 11. Performance of different top-K and ensemble learning. a. *NPM1* mutations ensemble model $= K3 + K5 + K10 + K20$; b. *FLT3*-ITD ensemble model $= K1 + K10 + K20$.

Furthermore, the proposed method aims at maintaining the same sensitivity while reducing false positive occurrences. When expecting to achieve a sensitivity rate of 0.7, the false positive rates of cell-level MIL models led to significant improvements. In *NPM1* mutation prediction, we reduced the false positive rate from 0.25 to 0.10, while for *FLT3*-ITD, the false positive rate decreased from 0.48 to 0.24. These results strongly emphasize the superiority of the cell-level MIL model in clinical diagnosis, as it substantially reduces the likelihood of false positives in cases identified as mutated. This underscores the model's effectiveness in providing reliable diagnoses while maintaining high sensitivity levels.

In conclusion, our research underscores the immense potential of ensemble learning, upsampling techniques, and MIL in predicting gene mutations in AML patients. Importantly, we accomplished this using training data labeled exclusively at the slide level, eliminating the labor-intensive manual annotation required at the patch and cell levels. This approach streamlined the end-to-end training and prediction process, emphasizing the value of integrating advanced machine-learning approaches to effectively address complex real-world challenges in the field of medical image analysis.

References

[1] [5] Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., & Madabhushi, A. (2019). Artificial intelligence in digital pathology — New tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology, 16(11), 703-715. https://doi.org/10.1038/s41571-019-0252-y

[2] Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D. R., & Rajpoot, N. M. (2016). Gland segmentation in colon histology images: The glas challenge contest. Medical Image Analysis, 35, 489-502. https://doi.org/10.1016/j.media.2016.08.008

[3] Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D. R., & Rajpoot, N. M. (2016). Gland segmentation in colon histology images: The glas challenge contest. Medical Image Analysis, 35, 489-502. https://doi.org/10.1016/j.media.2016.08.008

[4] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Gilbert, F. J., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., . . . Shetty, S. (2019). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6

[5] Chandradevan, R., Aljudi, A. A., Drumheller, B. R., Kunananthaseelan, N., Amgad, M., Gutman, D. A., Cooper, L. A., & Jaye, D. L. (2019). Machine-based detection and classification for bone marrow aspirate differential counts: Initial development focusing on nonneoplastic cells. Laboratory Investigation, 100(1), 98-109. https://doi.org/10.1038/s41374-019-0325-7

[6] Reta, C., Altamirano, L., Gonzalez, J. A., Diaz-Hernandez, R., Peregrina, H., Olmos, I., Alonso, J. E., & Lobato, R. Segmentation and Classification of Bone Marrow Cells Images Using Contextual Information for Medical Diagnosis of Acute Leukemias. PLOS ONE, 10(6), e0130805. https://doi.org/10.1371/journal.pone.0130805

[7] Dimitriou N, Arandjelović O, Caie PD. Deep Learning for Whole Slide Image Analysis: An Overview. Front Med (Lausanne). 2019 Nov 22;6:264. doi: 10.3389/fmed.2019.00264. Erratum in: Front Med (Lausanne). 2020 Aug 19;7:419. PMID: 31824952; PMCID: PMC6882930.

[8] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318(22):2199–2210. doi:10.1001/jama.2017.14585

[9] Burlutskiy, N., Gu, F., Wilen, L. K., Backman, M., & Micke, P. (2018). A Deep Learning Framework for Automatic Diagnosis in Lung Cancer. ArXiv. /abs/1807.10466

[10] Chuang, W., Chang, S., Yu, W., Yang, C., Yeh, C., Ueng, S., Liu, Y., Chen, T., Chen, K., Hsieh, Y., Hsia, Y., Wang, T., Hsueh, C., Kuo, C., & Yeh, C. (2020). Successful Identification of Nasopharyngeal Carcinoma in Nasopharyngeal Biopsies Using Deep Learning. Cancers, 12(2), 507. https://doi.org/10.3390/cancers12020507

[11] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine, 24(10), 1559-1567. https://doi.org/10.1038/s41591-018-0177-5

[12] Santos, A. G., Da Rocha, G. O., & De Andrade, J. B. (2019). Occurrence of the potent mutagens 2- nitrobenzanthrone and 3-nitrobenzanthrone in fine airborne particles. Scientific Reports, 9(1), 1-13. https://doi.org/10.1038/s41598-018-37186-2

[13] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25(8), 1301-1309. https://doi.org/10.1038/s41591-019-0508-1

[14] X. Wang et al., "Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis," in IEEE Transactions on Cybernetics, vol. 50, no. 9, pp. 3950-3962, Sept. 2020, doi: 10.1109/TCYB.2019.2935141.

[15] Tsai, X. C., Sun, K. J., Lo, M. Y., Tien, F. M., Kuo, Y. Y., Tseng, M. H., Peng, Y. L., Chuang, Y. K., Ko, B. S., Tang, J. L., Sun, H. I., Liu, M. C., Liu, C. W., Lin, C. C., Yao, M., Chou, W. C., Hou, H. A., & Tien, H. F. (2023). Poor prognostic implications of myelodysplasia-related mutations in both older and younger patients with de novo AML. Blood cancer journal, 13(1), 4. https://doi.org/10.1038/s41408-022-00774-7

[16] Muñoz-Aguirre, Manuel, et al. "PyHIST: a histological image segmentation tool." PLoS computational biology 16.10 (2020): e1008349.

[17] Tayebi, Rohollah Moosavi, et al. "Automated bone marrow cytology using deep learning to generate a histogram of cell types." Communications medicine 2.1 (2022): 45.

[18] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[19] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).