Artificial intelligence-based quantitative bone marrow pathology analysis for myeloproliferative neoplasms

Dandan Yu,[1-3*] Hongju Zhang,[1,2*] Yanyan Song,[1,2] Yuan Tao,[1,2] Fengyuan Zhou,[4] Ziyi Wang,[4] Rongfeng Fu,[1-3] Ting Sun,[1-3] Huan Dong,[1-3] Wenjing Gu,[1-3] Renchi Yang,[1-3] Zhijian Xiao,[1,2] Qi Sun[1,2] and Lei Zhang[1-3]

[1]State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Tianjin Key Laboratory of Gene Therapy for Blood Diseases, CAMS Key Laboratory of Gene Therapy for Blood Diseases, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin; [2]Tianjin Institutes of Health Science, Tianjin; [3]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing and [4]XY AI Technologies (Su Zhou) Limited, Jiangsu, China

**DY and HZ contributed equally as first authors.*

Correspondence: L. Zhang
zhanglei1@ihcams.ac.cn

Q. Sun
sunqi@ihcams.ac.cn

Z. Xiao
zjxiao@ihcams.ac.cn

Received: June 19, 2024.
Accepted: May 23, 2025.
Early view: June 12, 2025.

https://doi.org/10.3324/haematol.2024.286123

Methods

To mitigate the computational load caused by the large digital data size of WSIs, the regions of interest (ROIs) were first extracted using a defined workflow (Figure S2). The extracted ROIs were then subjected to preprocessing steps, including image enhancement, to improve the model's generalization ability (Figure S3). Subsequently, the ROIs were divided into small patches (512×512 pixels) for input into segmentation models (Table S2). Finally, a misaligned cutting-overlapping stitching strategy was applied to integrate and quantitatively analyze the patches across the whole slide (Figure S4).

Detection and delineation of cells and tissues

All the target identification algorithm models were trained using a supervised learning approach. Based on the characteristics of cells and tissues in specific stained sections, we employed the U^2-Net and UNeXt to segment cells and tissues in the hematoxylin-eosin (H&E)-stained section and Gomori-stained section, respectively, and the ResNet-18 to classify granulocytes and erythropoiesis. All models were trained in a PyTorch 1.8.0 cuda11.1 environment, using the default initialization methods of PyTorch (Table S3). The performance of the segmentation models is demonstrated through loss and Intersection over Union (IoU) metrics, while the performance of the classification models is represented by the area under the curve (AUC), the confusion matrix, Accuracy, Precision, Recall, and F1 score. Next, we will elaborate on the parameters optimized through multiple rounds of training, performance metric evaluations, and the visual presentation of segmentation and classification results.

Segmentation model training for targets in H&E-stained sections

In H&E-stained sections, it's essential to evaluate marrow cellularity, myeloid-to-erythroid (M: E) ratio, megakaryocyte morphology and distribution. Therefore, the targets for segmentation in H&E-stained sections include hematopoietic tissues, fat cells, bone trabecula, granulocytes, erythropoiesis, and megakaryocytes (Table S2).

For the segmentation task in H&E-stained sections, U^2-Net was used to process 512*512 pixels patches at various image levels. [1] U^2-Net consists of an encoding-decoding phase, which is constituted by ReSidual Ublock (RSU) modules. Each RSU comprises an input convolution layer (with a filter size of 3, stride of 1, and padding of 1) along with batch normalization layers and rectified linear unit (ReLU), a U-Net-like symmetric encoding-decoding structure, and residual connections. The encoder stage of U^2-Net employs RSU structures with depths of 7, 6, 5, 4, 4, and 4, while the decoder mirrors the encoder's structure. The output convolutional layer has a filter size of 1 and is followed by the Sigmoid, outputting the probability of each pixel belonging to a specific target category. This allows for clear differentiation between the target and background, achieving precise boundary delineation and segmentation of targets.

The hyperparameter settings are shown in Table S4. The training and validation datasets of various targets for the U^2-Net segmentation model are shown in Table S7. After three training rounds, the U^2-Net model achieved a high-accuracy segmentation of tissues with an IoU of 0.8 $\pm$0.07 (Table S10). The visualization of segmentation results is presented in Figure S5A-E.

Classification model training for granulocytes and erythropoiesis

Immature myeloid and erythroid cells are challenging to distinguish visually based on an experienced assessment. Three hematopathologists meticulously annotated granulocytes and erythropoiesis independently in H&E-stained sections, and those consistent annotations were taken as the standard training and validation sets. To evaluate the inter-observer consistency, a subset of consistent granulocytes and erythropoiesis annotations (intersection) labeled by three pathologists was selected as the test set and served as the ground truth (intersection). Each pathologist's annotations were compared against the ground truth (Table S12). Results indicate high inter-observer consistency and establish a reliable baseline.

After accurately segmenting granulocytes and erythropoiesis by U^2-Net, we employ the ResNet-18 classification model to differentiate between the two cell types. The image patches of granulocyte and erythropoiesis were extracted from the original segmentation images based on the size of the bounding rectangle of the segmentation mask (Figure S6). The ResNet-18 is a variant of the ResNet architecture that is 18 layers deep. [2] The input convolutional layer has a filter size of 7, followed immediately by a max pooling layer with a filter size of 2. The core section of ResNet-18 consists of four residual blocks (each of which contains two convolutional layers with a filter size of 3), with skip connections at the beginning and end of each residual block to promote the flow of information. A global average pooling layer then simplifies the feature map into a feature vector, which is ultimately processed by a fully connected layer and a SoftMax activation function to output the probabilities of each category.

The hyperparameter settings of ResNet-18 are presented in Table S5. We performed one training round on the model, which achieved a high accuracy in identifying granulocytes and erythropoiesis, with an average AUC of 0.958 in fivefold cross-validation (Table S8 and S11), and achieved an average detection rate of 0.89 (Figure S7). The visualization of the segmentation results is shown in Figure S5D.

Segmentation model training for targets in Gomori-stained sections

The severity of marrow fibrosis (MF) in the Gomori-stained sections is another crucial bone marrow feature of myeloproliferative neoplasm (MPN). This analysis necessitates precisely segmenting reticular fibers and identifying bone trabecula and fat cells (Table S2).

The segmentation model architecture for Gomori-stained sections is based on UNeXt, processing 512x512 pixels patches at various image levels. [3] This network employs an encoder-decoder architecture, where the encoder consists of three convolution blocks and two tokenized multilayer perceptron (MLP) stages, while the decoder comprises two tokenized MLP stages and three convolution blocks. Each convolutional block consists of one convolutional layer, batch normalization, and a ReLU. The convolutional layer has a filter size of three, a stride of one, and padding of one. In the encoder, a max pooling with a filter size of 2 follows the convolution blocks for down-sampling, whereas the decoder utilizes bilinear interpolation layers for up-sampling the feature maps. Features are processed through a Shift MLP before the Tokenized MLP. The Tokenized MLP block consists of a projection layer, a width-shifted MLP layer, a depth-wise convolution layer, a Gaussian Error Linear Unit (GeLU), and a height-shifted MLP layer, in sequence, with output features passed through layer normalization to the next block. The output is a segmentation map defined by the predictions for each pixel.

Hyperparameter settings for the UNeXt model are presented in Table S6. The training and validation datasets of various targets for the UNeXt segmentation model are shown in Table S9. After three training rounds, the UNeXt model achieved a relatively high accuracy segmentation of tissues with an IoU of approximately 0.7 (Table S13). The visualization of segmentation results is presented in Figure S5F-H.

Quantitative analysis of various metrics in the morphological model

Bone marrow cellularity

Bone marrow cellularity is determined by the average ratio of the area occupied by hematopoietic cells to the combined area of hematopoietic cells and fat cells within all valid hematopoietic regions. [4] Therefore, the calculation of bone marrow cellularity is based on the accurate segmentation of hematopoietic areas, fat, and bone trabecula. The segmentation of related tissues with a hematopoietic area can be visualized as shown in Figure 2A and Figure S5A-C. The identification of bone trabecula was used to correct the segmentation of the hematopoietic areas. The actual area occupied by hematopoietic cells was determined by the identified hematopoietic tissue region subtracted from the blank area.

The formula for the bone marrow cellularity is as follows:

$$\text{Bone marrow cellularity} = \frac{\text{Actual hematopoietic area}}{\text{Actual hematopoietic area} + \text{Fat cell area}}$$

Since the bone marrow cellularity involves the intact hematopoietic areas, therefore, fragmented tissue areas—those with smaller areas and incomplete fat cells should be excluded. Fragmented tissues are quantified by the area less than 40000 μm² and a local hematopoietic tissue proportion higher than the average proportion of hematopoietic tissue in the whole sections. The overall bone marrow cellularity is then determined by the average ratio of the area occupied by hematopoietic cells to the combined area of hematopoietic cells and fat cells in the remaining areas.

Megakaryocyte morphology and distribution

Metrics related to megakaryocytes include the megakaryocyte size, nucleus, and distribution. Regarding the megakaryocyte nucleus, we incorporated two indicators, the nuclear-cytoplasmic ratio and the presence of naked nuclei. The literature did not provide a clear numerical definition of megakaryocyte morphology. We quantitatively defined the two characteristics based on the megakaryocytes in H&E-stained trephine sections from healthy donors.

We took megakaryocyte cell size from healthy donors (n=8, total 380 cells) as the reference megakaryocyte cell size range (Figure S8). Based on the 2.5th to 97.5th percentiles to define the normal reference range, cells smaller than the 2.5th percentile were classified as small megakaryocytes, while those larger than the 97.5th percentile were defined as large megakaryocytes.

Image patches for each megakaryocyte can be extracted based on boundaries after segmentation. Subsequently, each megakaryocyte image patch was processed based on the

color threshold to extract the cytoplasm and nucleus, enabling the calculation of the nucleocytoplasmic ratio and the proportion of the nucleus for each megakaryocyte (Figure S9). Similarly, we have taken the megakaryocyte nuclear-cytoplasmic ratio range among healthy donors as the reference megakaryocyte nuclear-cytoplasmic ratio range. Larger than the 97.5th percentile of the reference nuclear-cytoplasmic ratio range was determined as megakaryocytes with a high nuclear-cytoplasmic ratio (Figure S10).

Three hematopathologists manually categorized megakaryocyte clusters into dense and loose clusters in six H&E-stained section digital images and reviewed each other's work to ensure annotation consistency. The distance between megakaryocytes was defined as the distance between the centroids of two megakaryocytes minus the intracellular distance between the megakaryocytes (Figure S11). The numerical characteristics of megakaryocyte spacing were determined based on the manual annotations and clustering identified using the density-based spatial clustering of applications with noise (DBSCAN) algorithm. [5]

Fibrosis grading

In this phase, the slide images were segmented into larger patches (1536*1536 pixels, 0.263 μm per pixel) at 40X magnification to simulate the microscope insight for assessing fibrosis severity. Patches with more than 60% blank area were excluded from the analysis. Three hematopathologists independently classified the fibrosis grading of each patch (from 6 pre-PMF and 2 PMF) into MF-0 to MF-3 based on the 2022 WHO diagnostic criteria and other official consensus on grading bone marrow fibrosis. [6] Patches labeled with consistent fibrosis grading across three hematopathologists (547 patches) were taken as the standard sets for the following further analysis.

First, we extracted the effective fibrotic regions within each consistent annotated patch (Figure S12) and then calculated the density of reticular fibers within these areas for each patch. Based on the 2022 WHO diagnostic criteria, fibers in the MF-0 region predominantly appear as single, non-crosslinked short fibers. [6] Therefore, fibers in the MF-1, MF-2, and MF-3 regions can be considered a dense accumulation of single short fibers. Therefore, the relative quantity of fibers within the effective fibrotic areas, or the fiber density, can, to some extent, represent the severity of fibrosis. After identifying the fibers by UNeXt, we can determine the number and area of fibers in each patch. In 94 patches marked as MF-0 among standard sets, the total number of fibers is 3818, with the area of a single fiber ranging from 3 to 50 (median, 8) μm^2. Next, in the patches of grades 1, 2, and 3 among standard sets, we calculate the fiber density by dividing the total fiber area in each patch by the median area of a single fiber (8 μm^2), obtaining the relative fiber quantity. Then, the fiber density in each patch is determined by dividing the relative fiber quantity by the effective fibrotic area of the patch.

Next, we correlated these densities with the annotated fibrosis grading of each patch (MF-0 to MF-3), determining the fibrous density corresponding to different fibrosis gradings of each patch (Figure S13). This allows us to predict the fibrosis grading of unannotated patches. Determine the fiber density for the effective fibrosis area within each patch to derive the fibrosis grade for each patch of a biopsy section. Then, calculate the total effective fibrosis area for patches with corresponding fibrosis grade (MF-0 to MF-3) and the overall effective fibrosis area of the entire biopsy section. According to the 2022 WHO diagnostic criteria, using an area ratio of 30% as the baseline, the highest fibrosis grade with an effective fibrosis area ratio exceeding

30% of the total effective fibrosis area will be the final fibrosis grade for the entire biopsy section (Figure S14).

Classification

Fourteen pathological indicators were included in constructing the bone marrow classification model (Figure 4A), and six clinical indicators contributed to a clinical classification model (Figure 4B). On this basis, all fourteen pathological indicators and six clinical indicators were incorporated to build the comprehensive classification model (Figure 4C). The definition of clinical indicators implemented in classification models is as follows: hemoglobin (Hb), white blood cell count (WBC), platelet count (PLT), and lactate dehydrogenase (LDH) level were analyzed as continuous variables using their raw quantitative measurements. While gene mutation status and spleen sizes were analyzed as categorical variables. Gene mutation status was categorized into “mutated” (presence of $\geq$1 driver mutation in JAK2, CALR, or MPL genes) versus "Wild-type" (no detectable mutations). Splenomegaly grading was stratified into four categories: absent, mild, moderate, or severe.

All classification models were random forest classifiers. The composition ratios of disease samples in the training and validation sets are presented in Table S14.

The model parameters for those three classification models are identical. Each model consists of 100 decision trees. For each tree, 80% of the original dataset samples are randomly selected with replacements to form the training dataset. A certain number of features, determined to be one-third of the total number of features, are randomly chosen to create a feature subset. The classification and regression tree (CART) algorithm is then applied to construct the decision tree based on the selected dataset and feature subset. [7] No maximum depth limit is set for the decision trees, and their growth stops when the number of samples at a node decreases to less than two. The final prediction result is obtained by averaging the prediction results of each tree.

Initially, 45 samples (45/342, 13%) were randomly selected from all 342 samples from our center as the internal test set (Table S14). The remaining samples were randomly divided into training and validation sets in an 80%/20% ratio and underwent 500 independent trials. In each trial, three classification models—bone marrow, clinical, and comprehensive models—were independently trained. Figure S15 presents the predicted performance of the three models in the 500 independent trials. The classification performances of the final three classification models are shown in Table S15.

Supplementary Tables

Supplementary Table 1. Clinical Characteristics of All Cohort from Our Center.

	Blood counts (value)			LDH (U/L) (Median, range)	Mutation status (number)				Splenomegaly (number)			
	Platelet count (10^9/L) (Median, range)	White cell count (10^{12}/L) (Median, range)	Hemoglobin (g/L) (Median, range)		TN	JAK2 (V617F)	CALR	MPL	no	mild	moderate	severe
MPN												
ET (n=78)	771 (342-1680)	8.44 (3.12-23.62)	138.5 (59-182)	226.5 (142-536.9)	10	48	19	1	56	10	8	4
Pre-PMF (n=37)	918 (119-2562)	8.78 (3.06-36.12)	131 (66-159)	269.3 (159.7-482.7)	2	23	11	1	16	8	9	4
PMF (n=167)	218 (7-1541)	7.82 (0.67-196.23)	102 (37-190)	486 (141-2073.2)	16	108	39	4	14	9	54	90
PV (n=27)	576 (217-1387)	11.91 (4.32-25.44)	191 (124-235)	316.6 (142.4-533.4)	2	25	0	0	5	9	11	2
Nonneoplastic												
Normal (n=8)	241 (166-291)	7 (4.66-10.98)	140 (128-184)	165.1 (113-260)	8	0	0	0	8	0	0	0
IDA (n=25)	264 (31-1191)	5.9 (1.27-16.78)	77 (35-113)	182.9 (107.2-1549)	25	0	0	0	15	6	2	2

Abbreviations: MPN, myeloproliferative neoplasm; ET, essential thrombocythemia; PMF, primary myelofibrosis; pre-PMF, prefibrotic PMF; PV, polycythemia vera; IDA, iron-deficiency anemia; LDH, lactate dehydrogenase; TN, triple-negative.

Supplementary Table 2. Hierarchical Processing of Various Tissues in Bone Marrow Trephine Section.

Section	Research Metric	Tissues	Segmentation Model	Magnification [a]	Hierarchies	Pixel Size (μm)	Image Patch Size (pixel) [b]	Number of Annotated Slides (Total Annotated Cells/Tissue)
H&E-stained section	Marrow cellularity	Fat	U^2-Net	10X	Level 1	1.052	512*512	32 (53559)
		Bone		10X	Level 1	1.052	512*512	50 (7240)
		Hemopoietic tissue		10X	Level 1	1.052	512*512	50 (1430)

	Myeloid-to-erythroid ratio	Granulocyte & Erythropoiesis	U^2-Net	80X	NA [c]	0.132	512*512	87 (89449) [d]
	Megakaryocyte	Megakaryocyte	U^2-Net	40X	Level 0	0.263	512*512	141 (33217)
Gomori-stained section	Fibrosis severity	Fibrosis	UNeXt	20X	Level 0.5	0.526	512*512	52 (98464)
		Fat		10X	Level 1	1.052	512*512	37 (63354)
		Bone		10X	Level 1	1.052	512*512	50 (11870)

[a] The specific magnifications were determined by the visual observation of hematopathologists and the requirement of the model development.

[b] The actual size of each pixel depended on the specific magnification corresponding to specific stained sections, aimed at accurately matching the characteristics of various cells and tissues.

[c] Pixel scaling has no corresponding hierarchy.

[d] Granulocytes and erythropoiesis were annotated jointly, with the reported value representing the sum of both. Given the significant quantity of erythropoiesis and granulocytes present in a single section image, and to account for variability across different samples, annotations of these cells are performed within selected regions on each image rather than comprehensive annotation across the entire slide, which was the approach used for the annotations of other tissues.

Abbreviations: H&E-stained section, hematoxylin-eosin-stained section.

Supplementary Table 3. Model Training Environment Configurations.

	Configurations
CPU	Intel(R) Xeon(R) Gold 6226R @2.90GHz
GPU	Tesla A100 40G
System Platform	CentOS 7.3
Deep Learning Framework	PyTorch 1.8.0 cuda11.1

Abbreviations: CPU, Central Processing Unit.

Supplementary Table 4. Hyperparameters of the U^2-Net Model for Segmentation Applied in H&E-stained Section.

Tissue / Parameter	Fat	Bone	Hemopoietic tissue	Granulocyte & Erythropoiesis	Megakaryocyte
Batch Size	16	16	16	16	16
Learning Rate	0.001	0.001	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam	Adam	Adam
Epochs	250	200	200	200	200
Loss function	Cross Entropy Loss				

Abbreviations: H&E-stained section, hematoxylin-eosin-stained section; Adam, Adaptive Moment estimation.

Supplementary Table 5. Hyperparameters of the ResNet-18 Model for Identifying Granulocytes and Erythropoiesis.

Tissue / Parameter	Granulocyte & Erythropoiesis
Batch Size	32
Learning Rate	0.001
Optimizer	SGD
Epochs	80
Loss function	Cross Entropy Loss

Abbreviations: SGD, Stochastic Gradient Descent.

Supplementary Table 6. Hyperparameters of the UNeXt Model for Segmentation Applied in Gomori-stained Section.

Tissue / Parameter	Bone	Fat	Fibrosis
Batch Size	16	16	16

Learning Rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	200	250	250
Loss function	Cross Entropy Loss		

Abbreviations: Adam, Adaptive Moment estimation.

Supplementary Table 7. Training and Validation Image Patches and Samples for the U^2-Net Model in H&E-stained Section.

H&E-stained section	Initial round		1st round		2nd round		3rd round		Samples
	Training	Validation	Training	Validation	Training	Validation	Training	Validation	
Hemopoietic tissue	1796	316	2874	507	7185	1264	8980	1585	165
Fat	2507	443	3449	608	8622	1521	14750	2594	
Bone trabecula	1734	301	2737	483	6842	1206	13435	2061	
Granulocyte & Erythropoiesis	3163	558	5435	959	7165	1264	42150	7435	249
Megakaryocyte	3094	547	5396	923	7901	1394	63034	11123	

Abbreviations: H&E-stained section, hematoxylin-eosin-stained section.

Supplementary Table 8. Training and Validation Image Patches and Samples for the ResNet-18 model in identifying granulocytes and Erythropoiesis.

H&E-stained section	1 st round		Samples
	Training	Validation	
Granulocyte	37816	8703	249
Erythropoiesis	20113	3549	
Total	57929	12252	

Abbreviations: H&E-stained section, hematoxylin-eosin-stained section.

Supplementary Table 9. Training and Validation Image Patches and Samples for the UNeXt Model in Gomori-stained Section.

Gomori-stained section	Initial round		1st round		2nd round		3rd round		Samples
	Training	Validation	Training	Validation	Training	Validation	Training	Validation	
Fibrosis	1808	375	2823	423	4454	658	12220	2155	
Fat	2175	384	3245	572	4300	602	12795	2255	132
Bone trabecula	1856	328	2284	296	4051	576	11420	2015	

Supplementary Table 10. Segmentation Performance of the U^2-Net Model Applied in H&E-Stained Section.

	Initial round	1st round	2nd round	3rd round
Hemopoietic tissue				
Loss	1.6045	0.5902	0.4373	0.4570
IoU	0.57	0.62	0.70	0.73
Fat				
Loss	0.8655	0.6580	0.3459	0.2586
IoU	0.45	0.60	0.71	0.76
Bone trabecula				
Loss	0.2473	0.1182	0.1301	0.1364
IoU	0.76	0.76	0.85	0.87
Megakaryocyte				
Loss	1.2949	0.8426	0.6950	0.5209
IoU	0.62	0.75	0.74	0.82
Erythropoiesis & Granulocyte				
Loss	1.2714	0.7977	0.4874	0.4541
IoU	0.73	0.79	0.78	0.81

Abbreviations: H&E-stained section, hematoxylin-eosin-stained section; IoU, Intersection over Union.

Supplementary Table 11. Fivefold Cross-validation of the ResNet-18 Model Applied in Identifying Granulocyte & Erythropoiesis.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
AUC	0.957	0.958	0.959	0.959	0.958	0.958
Accuracy	0.961	0.962	0.967	0.962	0.963	0.963
Precision	0.955	0.956	0.958	0.955	0.957	0.957
Recall	0.961	0.961	0.962	0.961	0.961	0.961
F1 score	0.958	0.959	0.960	0.959	0.960	0.958

Abbreviations: AUC, area under the curve.

Supplementary Table 12. Segmentation Performance of the UNeXt Model Applied in Gomori-Stained Section.

	Initial round	1st round	2nd round	3rd round
Fibrosis				
Loss	0.6522	0.1625	0.1315	0.0973
IoU	0.27	0.41	0.53	0.57
Fat				
Loss	0.7027	0.2621	0.0793	0.0672
IoU	0.43	0.51	0.50	0.63
Bone trabecula				
Loss	0.5571	0.0564	0.0515	0.0499
IoU	0.63	0.75	0.74	0.78

Abbreviations: IoU, Intersection over Union.

Supplementary Table 13. Inter-observer Consistency and Model Overfitting Evaluation.

	Granulocyte identification				Erythropoiesis identification			
	FNR	FPR	Precision	Recall	FNR	FPR	Precision	Recall
Intersection VS User1	0.000	0.053	0.94	1	0.000	0.215	0.78	1
Intersection VS User2	0.000	0.025	0.97	1	0.006	0.081	0.91	1
Intersection VS User3	0.000	0.050	0.94	1	0.000	0.081	0.91	1
AI VS Intersection	0.065	0.194	0.78	0.93	0.180	0.186	0.82	0.82
AI VS User1	0.065	0.257	0.7	0.93	0.109	0.323	0.69	0.89
AI VS User2	0.069	0.218	0.76	0.93	0.146	0.222	0.78	0.85
AI VS User3	0.057	0.248	0.72	0.94	0.176	0.240	0.76	0.82

Evaluation of inter-observer baseline, segmentation performance, and overfitting in the ResNet-18 classification model. The False Negative Rate (FNR) was used to quantify the missed detection rate, defined as the proportion of actual positive samples incorrectly classified as negative. The False Positive Rate (FPR) was employed to describe the false detection rate, representing the proportion of actual negative samples erroneously classified as positive. The intersection refers to a subset of consistent annotations provided by three pathologists (User1, User2, and User3), serving as an uncontested diagnostic consensus. ”AI” denotes the annotations predicted by the ResNet-18 classification model. Inter-observer consistency was evaluated by comparing each pathologist’s annotations with the intersection, while model performance was assessed by comparing its predictions with both the intersection and individual pathologist annotations. These comparisons demonstrate the model’s reliability and confirm that it is not overfitting to any specific pathologist.

Abbreviations: FNR, false negative rate; FPR, false positive rate; AI, artificial intelligence.

Supplementary Table 14. Datasets for Random Forest Models Applied in Differentiating Nonneoplastic and MPN Subtypes.

	Internal				External
	Total (samples, n)	Training (samples, n)	Validation (samples, n)	Test (samples, n)	Test (samples, n)
ET	78	55	13	10	19
Pre-PMF	37	26	6	5	20
PV	27	18	4	5	38
PMF	167	117	30	20	19
Nonneoplastic	33	22	6	5	10
Total	342	238	59	45	106

The number of samples utilized for the training and validation sets in the table corresponds to the quantity extracted for each random experiment.

Abbreviations: MPN, myeloproliferative neoplasm; ET, essential thrombocythemia; PMF, primary myelofibrosis; pre-PMF, prefibrotic PMF; PV, polycythemia vera.

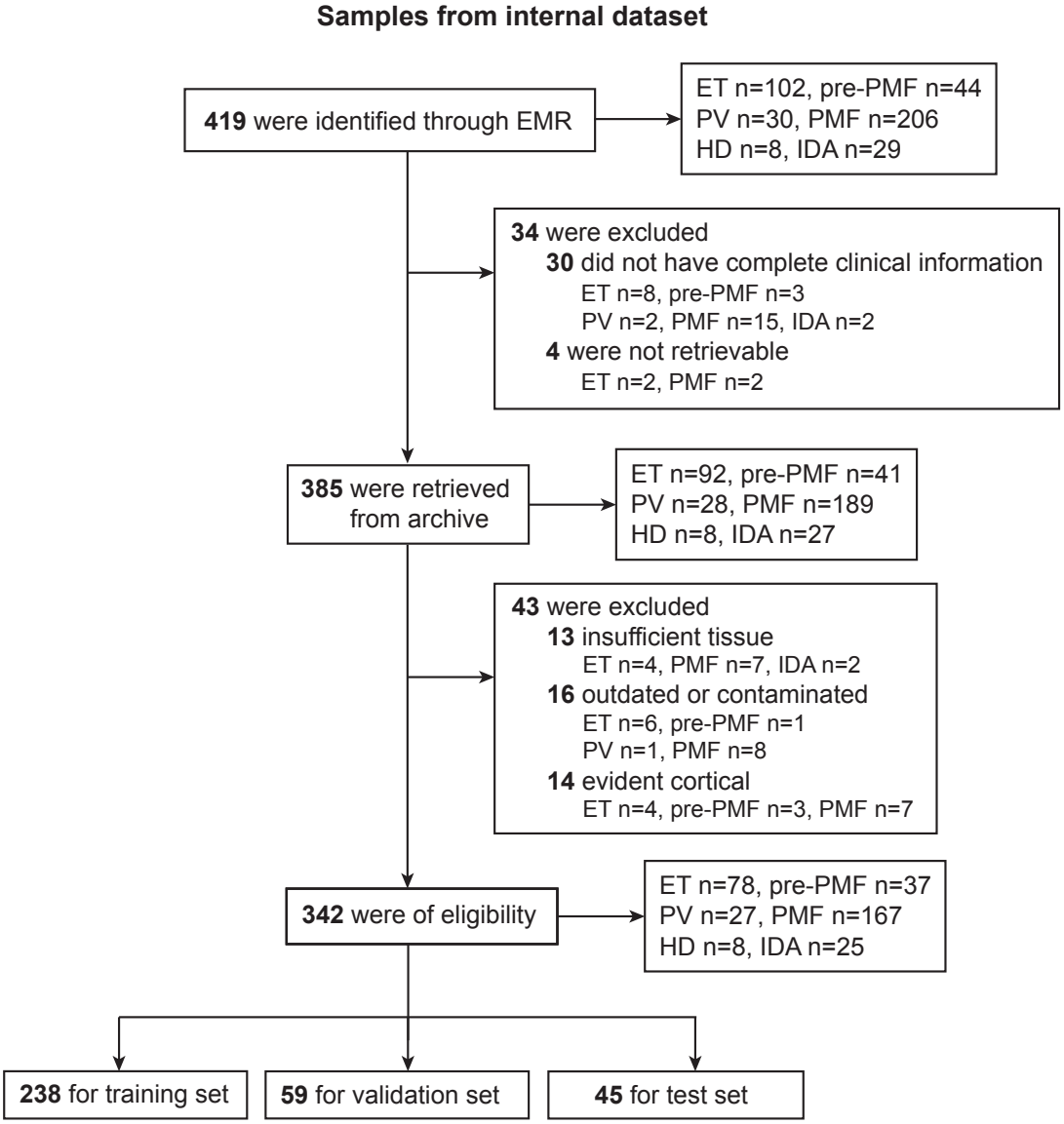
Supplementary Table 15. Classification Performances of Classification Models Applied in Differentiating Nonneoplastic and MPN Subtypes.

		ET	Pre-PMF	PMF	PV	Nonneoplastic	Macro	Micro
Internal test set	**Bone marrow model**							
	AUC	0.91	0.92	1	0.98	1	0.96	0.8
	Accuracy	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	Precision	0.67	0.4	1	0.67	0.83	0.71	0.8
	Recall	0.6	0.4	0.95	0.8	1	0.75	0.8
	F1 score	0.63	0.4	0.97	0.73	0.91	0.73	0.8
	Clinical model							
	AUC	0.9	0.84	0.95	0.88	1	0.92	0.78
	Accuray	0.73	0.73	0.73	0.73	0.73	0.73	0.73
	Precision	0.7	0.67	0.74	0.67	0.83	0.72	0.78
	Recall	0.7	0.4	0.85	0.4	1	0.67	0.73
	F1 score	0.7	0.5	0.79	0.5	0.91	0.68	0.73
	Comprehensive model							
	AUC	0.94	0.94	0.96	0.97	1	0.96	0.8
	Accuracy	0.85	0.85	0.85	0.84	0.85	0.85	0.85
	Precision	0.64	0.5	1	1	1	0.83	0.84
	Recall	0.7	0.6	0.95	0.8	1	0.81	0.84
	F1 score	0.67	0.55	0.97	0.89	1	0.82	0.84
External test set	**Bone marrow model**							
	AUC	0.95	0.84	1.0	0.92	0.99	0.94	0.58
	Accuracy	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	Precision	0.76	0.73	0.95	0.86	0.82	0.82	0.84
	Recall	0.84	0.53	1.0	0.84	1.0	0.84	0.84
	F1 score	0.8	0.62	0.97	0.85	0.9	0.83	0.84

Abbreviations: MPN, myeloproliferative neoplasm; ET, essential thrombocythemia; PMF, primary myelofibrosis; pre-PMF, prefibrotic PMF; PV, polycythemia vera; AUC, area under the curve.

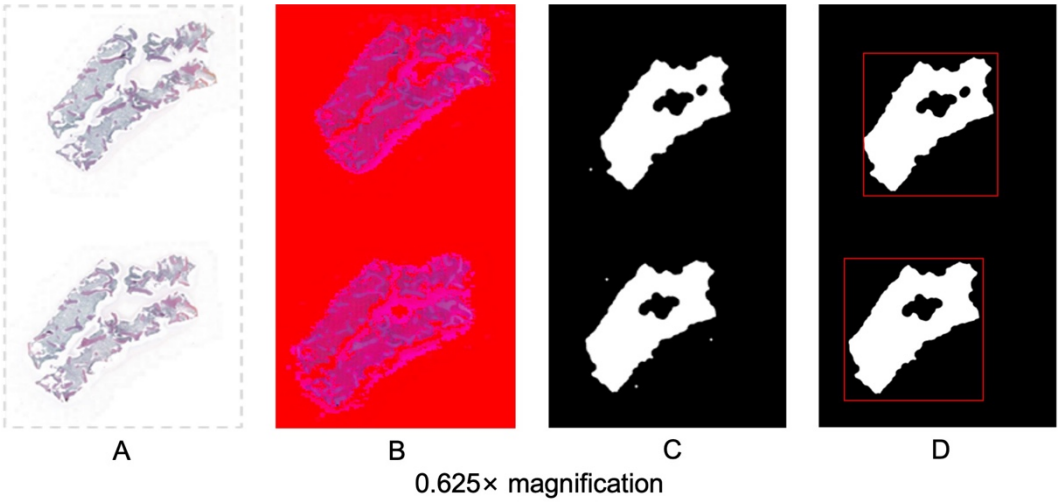
Supplementary Figures

Supplementary Figure 1. Flowchart of MPN (n=309) and Nonneoplastic (n=33) Case Selection Process.



The plot illustrates the inclusion and exclusion flowchart for samples in the internal dataset, with samples from the external dataset enrolled based on the same criteria. For inclusion, each slide must encompass sufficient tissue ($\geq 0.5 \times 0.2$ cm^2 for each). Outdated or contaminated slides were excluded. Additionally, biopsies displaying evident cortical bone or severely fragmented tissue, thereby hindering the assessment of hematopoietic status, were also excluded. Abbreviations: MPN, myeloproliferative neoplasm; EMR, electronic medical record; ET, essential thrombocythemia; PMF, primary myelofibrosis; pre-PMF, prefibrotic PMF; PV, polycythemia vera; HD, healthy donor; IDA, iron-deficiency anemia.

Supplementary Figure 2. The Extraction of the Region of Interest (ROI).



A&B. The original image was transformed into an image within the HSV color space, a visible light subset in a three-dimensional color space composed of Hue (H), Saturation (S), and Value (V). C. By conducting color localization tracking on the bone marrow tissue within the HSV color space and calculating the HSV thresholds, a color mask can be constructed based on these threshold values. Applying bitwise operations to the original image and this mask yields the effective area of the bone marrow slide as illustrated in the image. D. Further morphological transformations, including erosion, dilation, and contour area threshold filtering, were applied for refinement. And the area within the bounding rectangle was the Region of Interest (ROI).

Supplementary Figure 3. The Pre-processing of Whole Slide Images (WSIs).

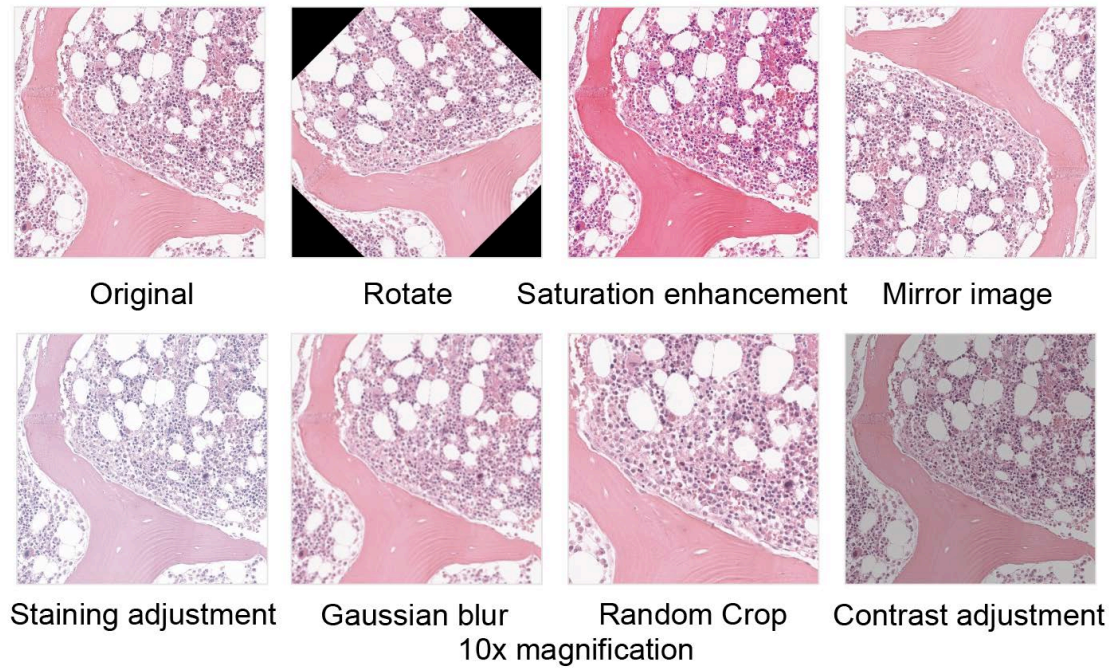
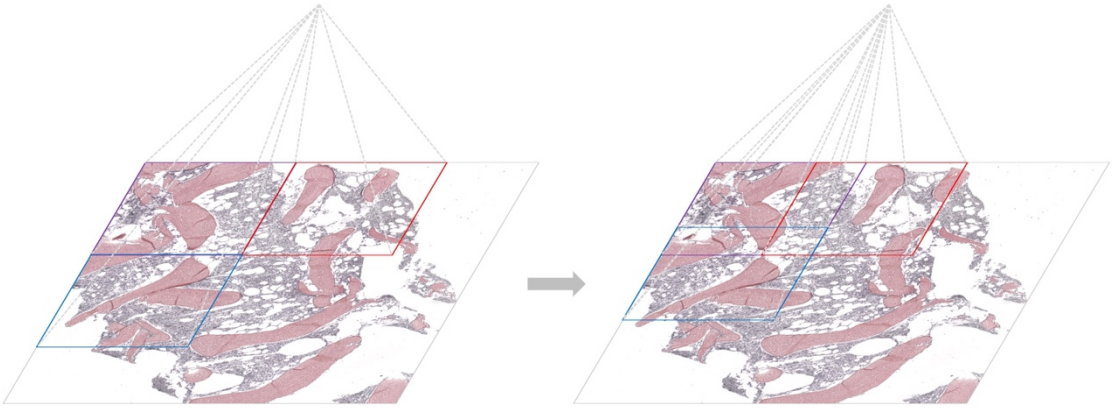


Image Augmentation techniques were applied to the Whole Slide Images (WSIs) to enhance the generalizability of the image recognition model. These included symmetrical transformations, rotation, HSV color space conversion, histogram equalization, blurring, motion blur, distortion, elastic deformation, inversion, channel scraping, Gaussian noise, salt-and-pepper noise, random brightness changes, random contrast adjustments, grayscale modifications, enhanced elastic deformations, and staining enhancement methods specific to H&E-stained sections targeting eight

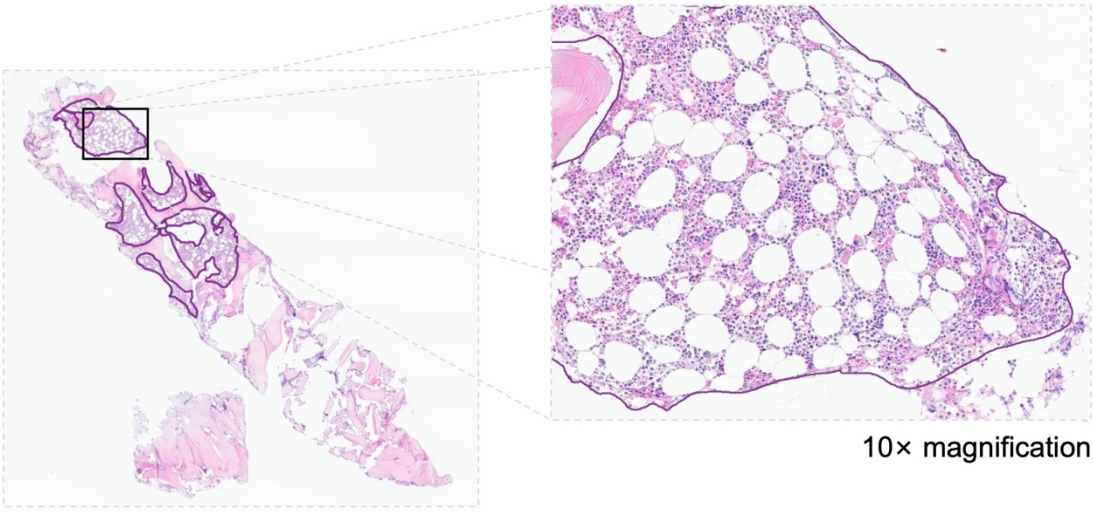
staining gradients.

Supplementary Figure 4. Misaligned Cutting-Overlapping Stitching Strategy.

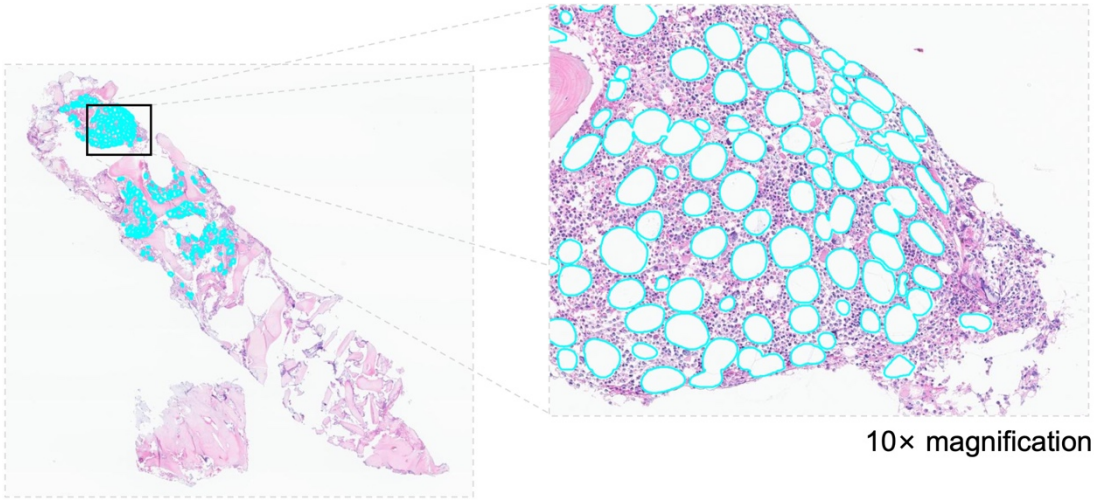


For structural segmentation of images, initial misaligned splitting into small image patches (512*512 pixels) is performed, followed by inference on each patch and subsequent reassembly into the original image size. This misaligned splitting and overlapping reassembly strategy enhances image segmentation accuracy while eliminating seam marks from image assembly, improving overall visualization.

Supplementary Figure 5. Visualization of the Segmentation of Cells and Tissues.

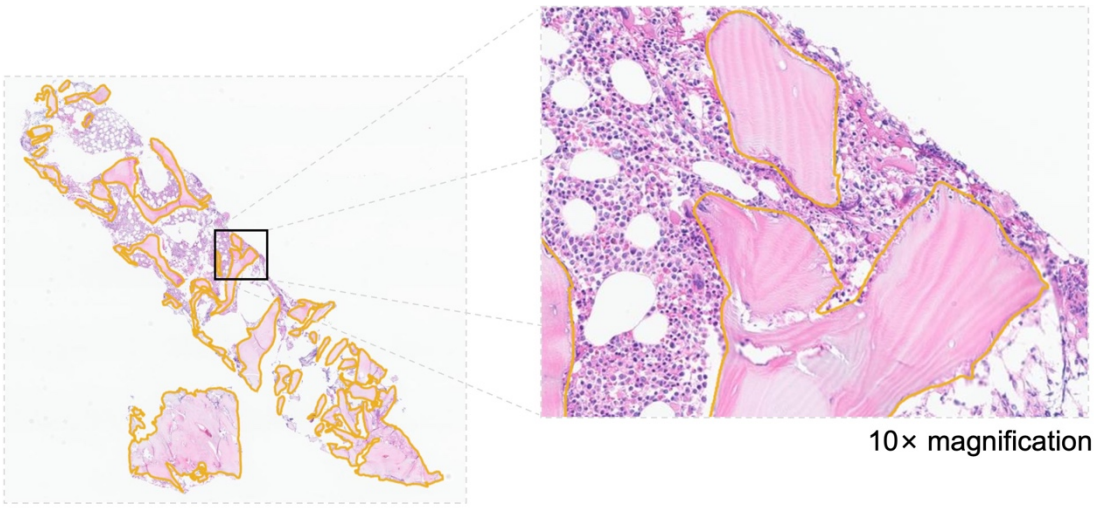


A. Hemopoietic area



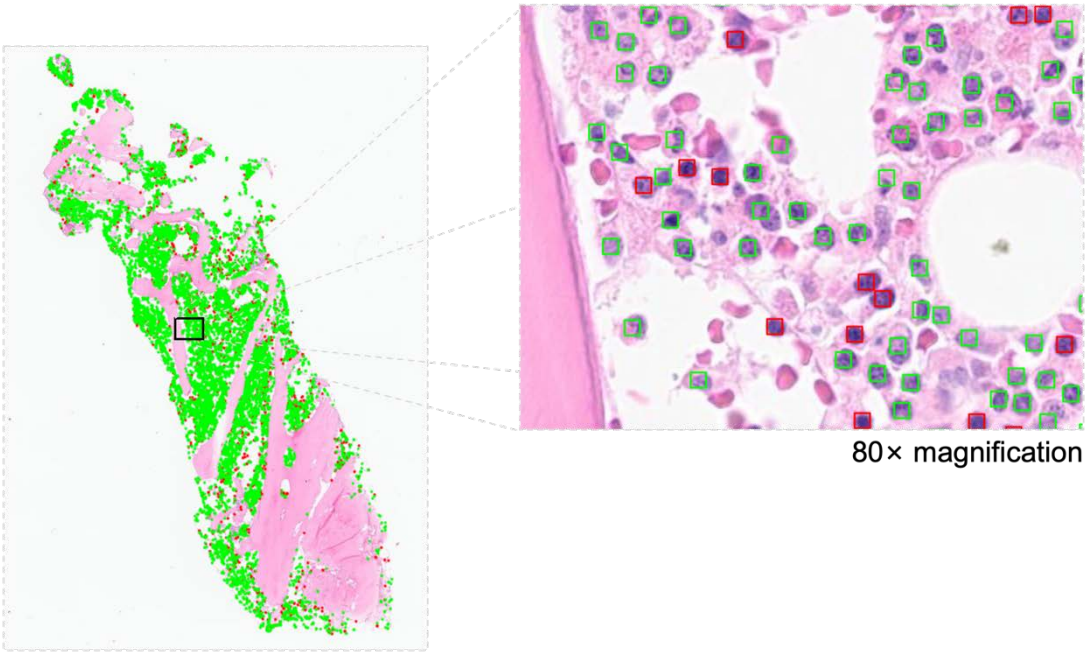
10× magnification

B. Fat



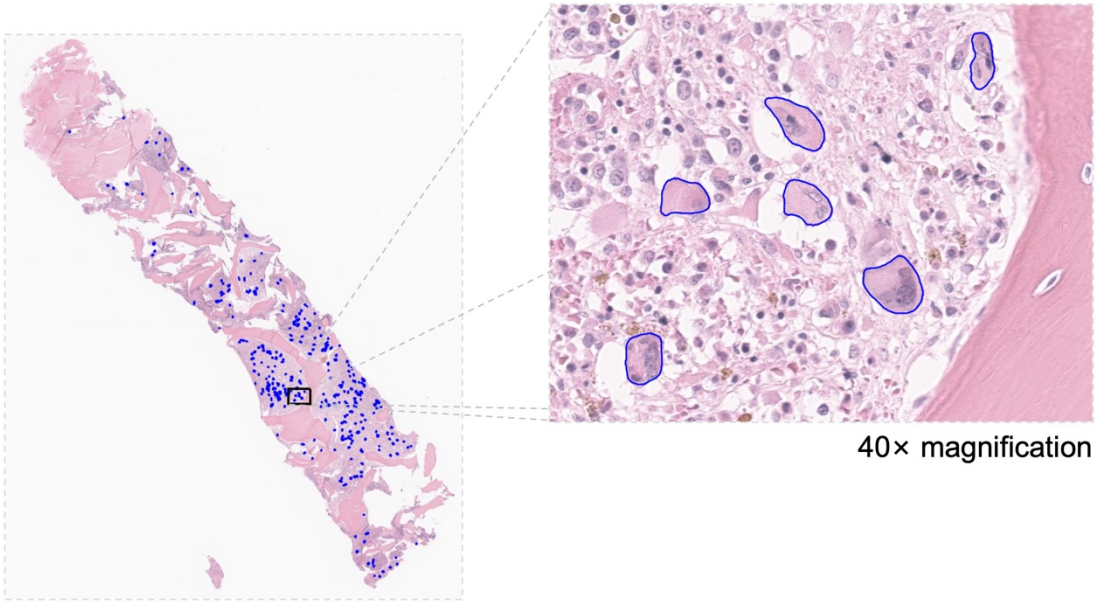
10× magnification

C. Bone trabecular



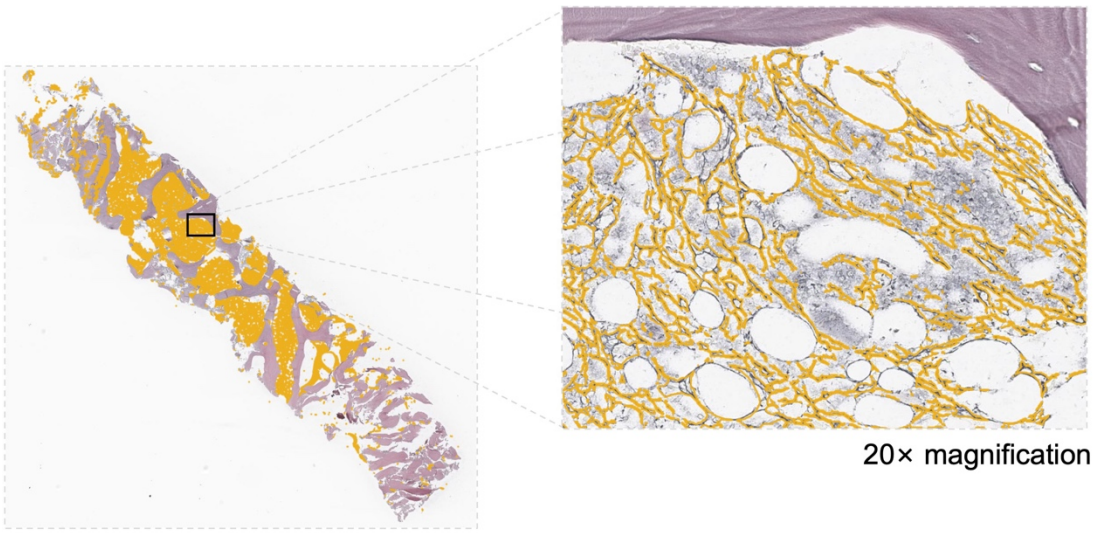
80× magnification

D. Granulocyte & Nucleated erythroid cells



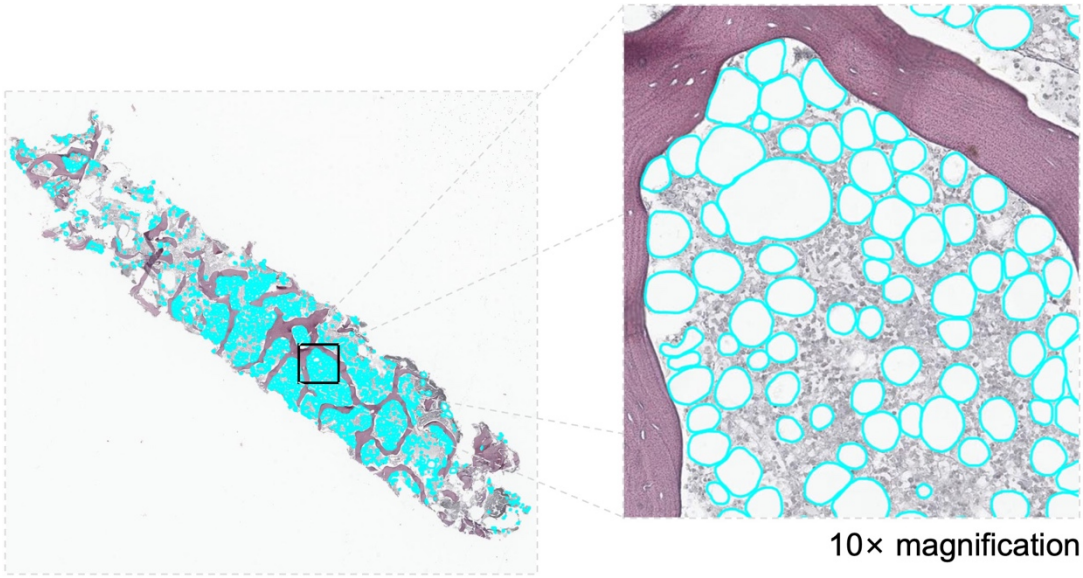
40× magnification

E. Megakaryocyte



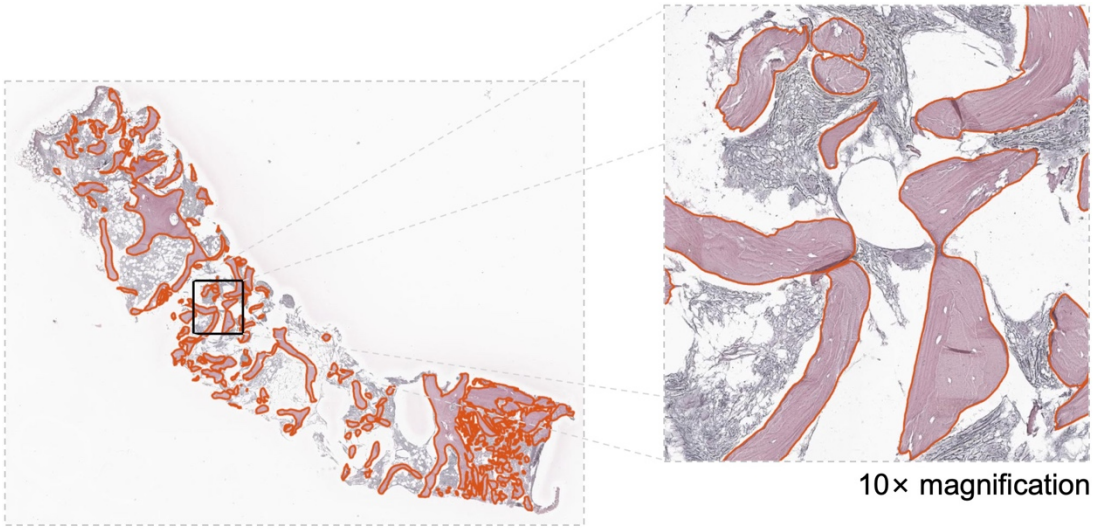
20× magnification

F. Reticulin fibrosis



10× magnification

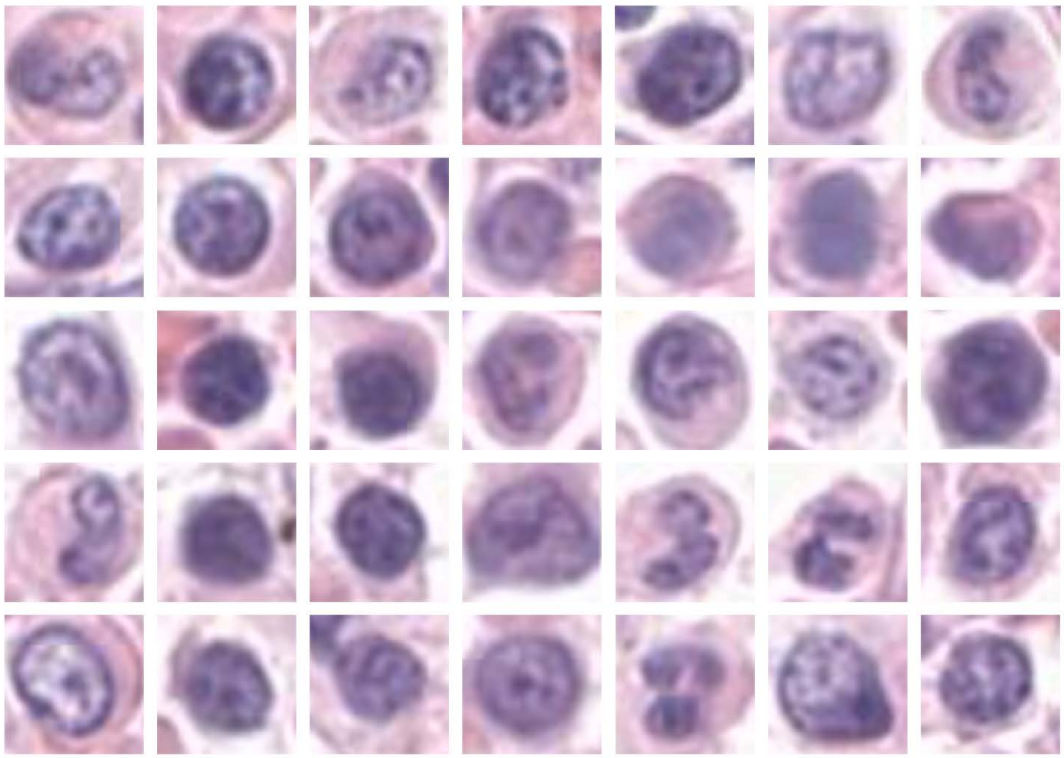
G. Fat



H. Bone trabecular

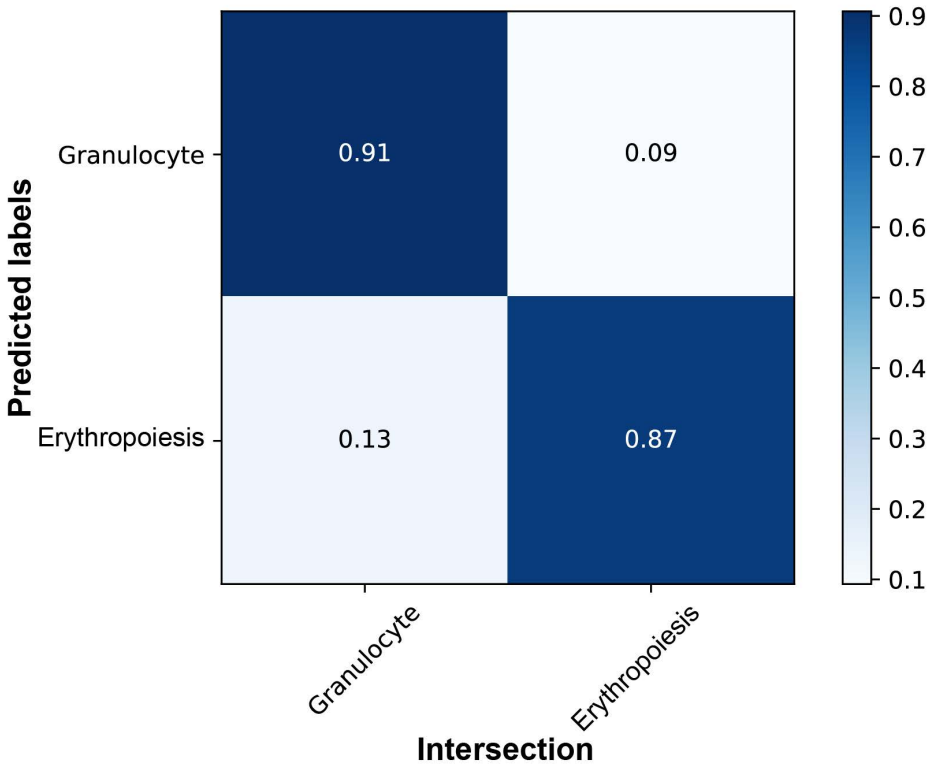
A-E. The segmentation of targets in hematoxylin and eosin-stained section by U^2-Net model and the identification of granulocyte and erythropoiesis by RseNet-18 model. F-H. The segmentation of targets in Gomori-stained section by UNeXt model.

Supplementary Figure 6. Input Image Patches Sizes of Granulocytes and Erythropoiesis for ResNet-18 Model.



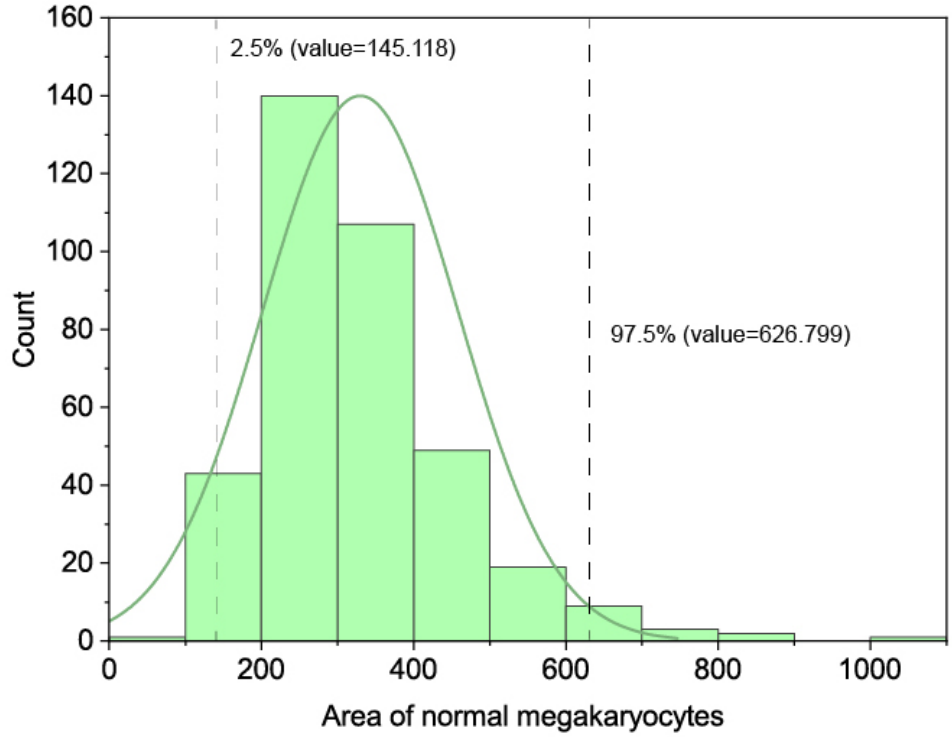
Due to the small size of granulocytic and erythroid cells, which also vary in size, each image patch was resized by pixel scaling to 64*64 pixels (16.768 μm per pixel) to facilitate model training. Each image patch contains only one cell, thus the number of patches used corresponds to the number of cells.

Supplementary Figure 7. The Confusion Matrix of the ResNet-18 Model in Recognizing Granulocyte & Erythropoiesis.



The confusion matrix was generated to evaluate the performance of the ResNet-18 model in identifying granulocytes and erythropoiesis in the test set. The test set was defined as a subset of consistent annotations (intersection) provided by three pathologists, serving as the "gold standard". Predicted labels were derived from the ResNet-18 model. The model achieved an average detection rate of 0.89.

Supplementary Figure 8. Area Range of Megakaryocytes from Healthy Donors.



Supplementary Figure 9. Analysis for Nuclear-cytoplasmic Ratio and Nuclear Proportion of Megakaryocyte.

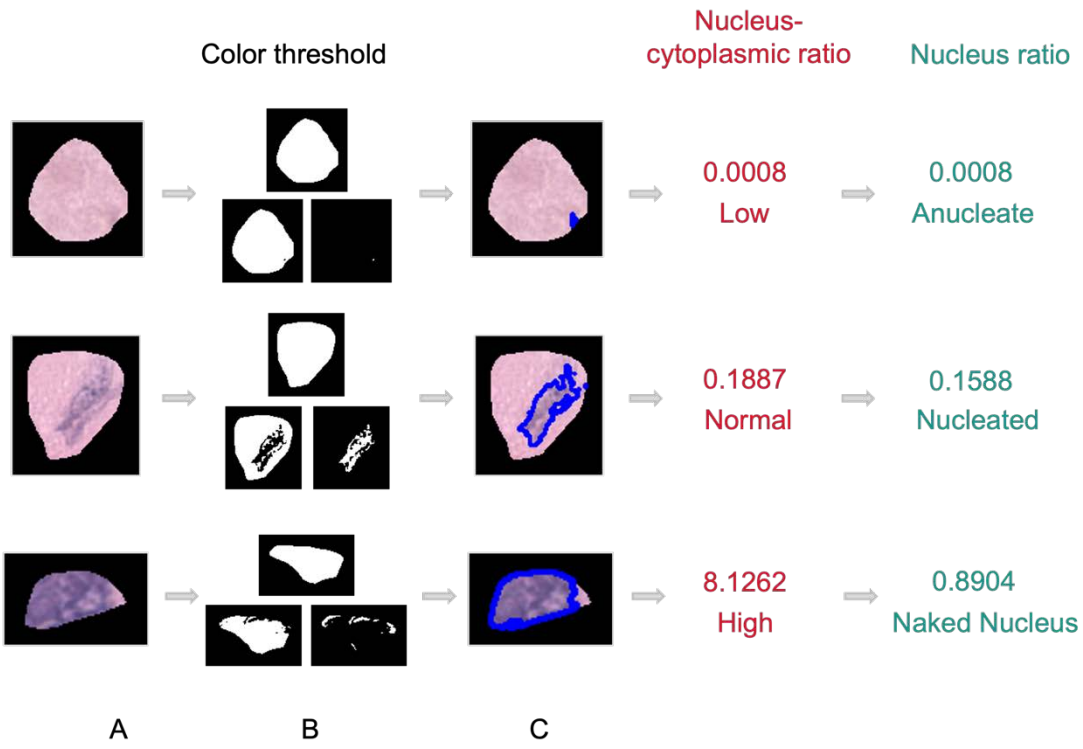
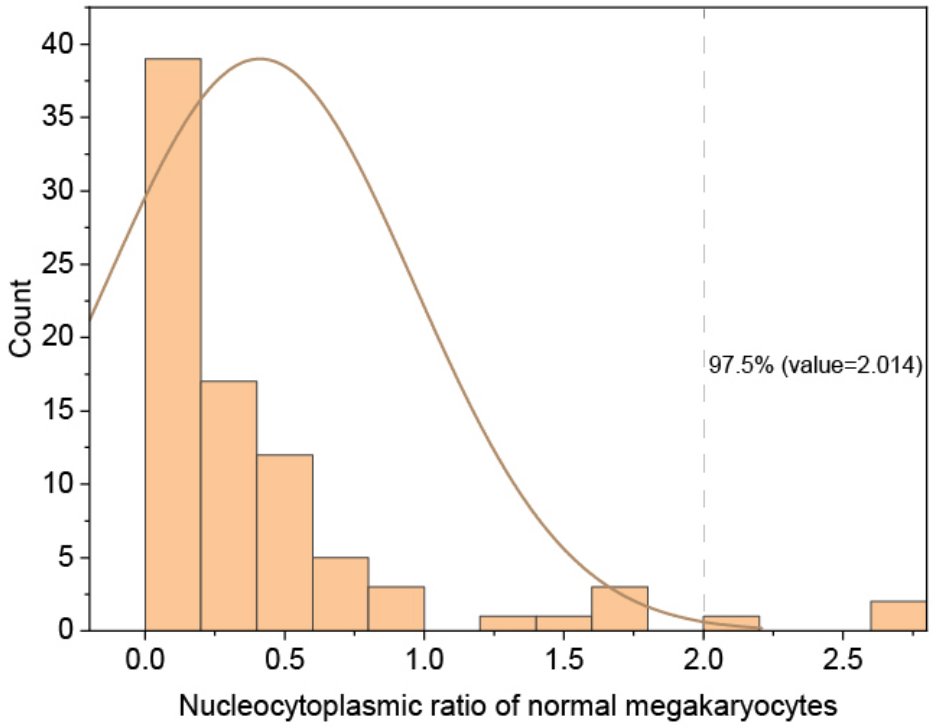


Image patches for each megakaryocyte could be extracted based on delineated boundary after segmentation (A). Then, cytoplasm and nuclei were extracted from each megakaryocyte image patch using color thresholding (B), enabling the

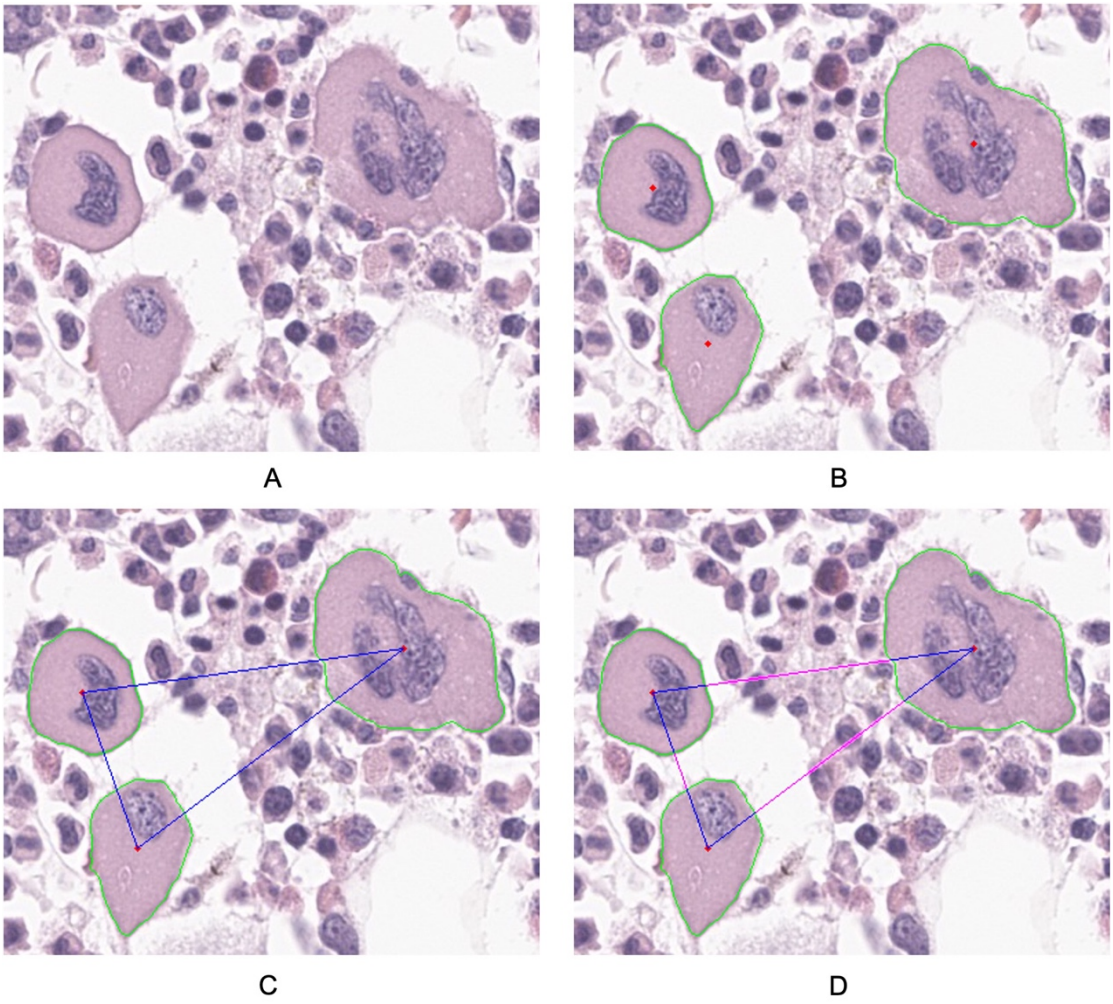
calculation of individual nucleo-cytoplasmic ratio and nuclear proportion (nuclear ratio) (C). The nuclear-cytoplasmic ratio is calculated by nuclear area / (total cell area − nuclear area). The nuclear proportion (nuclear ratio) is calculated by nuclear area / total cell area.

Supplementary Figure 10. Nuclear-cytoplasmic Ratio Range of Megakaryocytes from Healthy Donors.



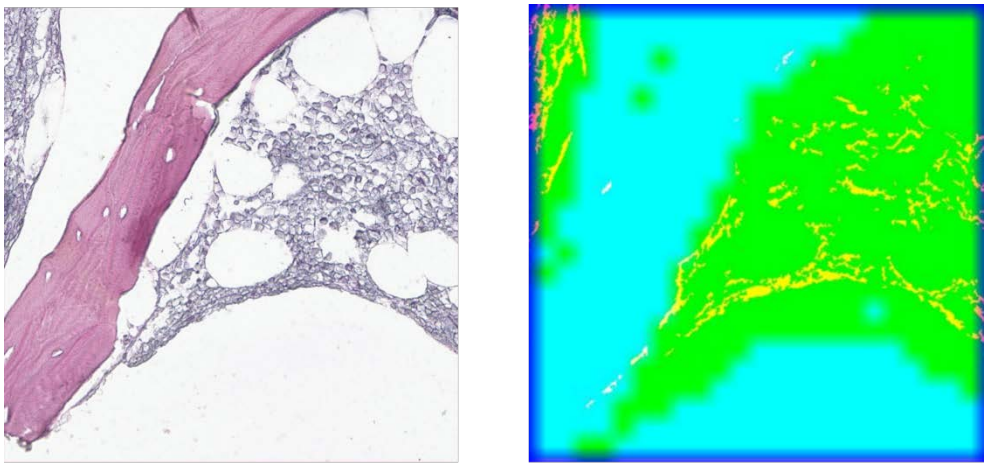
The nuclear-cytoplasmic ratio is calculated by nuclear area / (total cell area − nuclear area), as shown in Supplementary Figure 9.

Supplementary Figure 11. Visualization of Megakaryocyte Centroids and Intercellular Spacing.



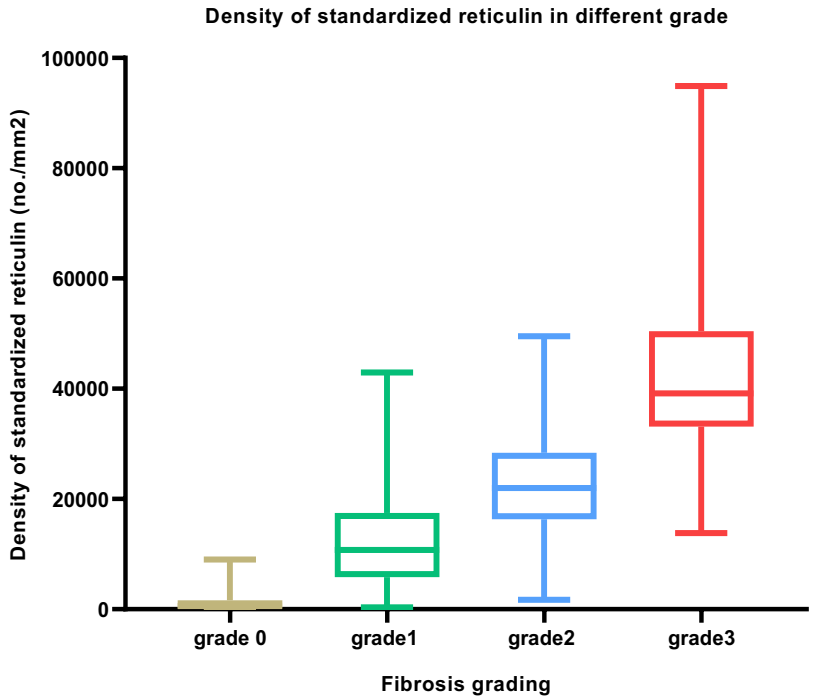
A. Original image at 40X magnification. B. Megakaryocytes (green) and their centroids (red); the centroid of a megakaryocyte represents its geometric center, calculated by averaging the positions of all pixels within the megakaryocyte. C. Distances between megakaryocyte centroids (purple lines). D. Actual distances between megakaryocytes (pink lines), which were determined by subtracting the internal distance of a megakaryocyte from the centroid distance.

Supplementary Figure 12. The Extraction of Effective Fibrosis Area.



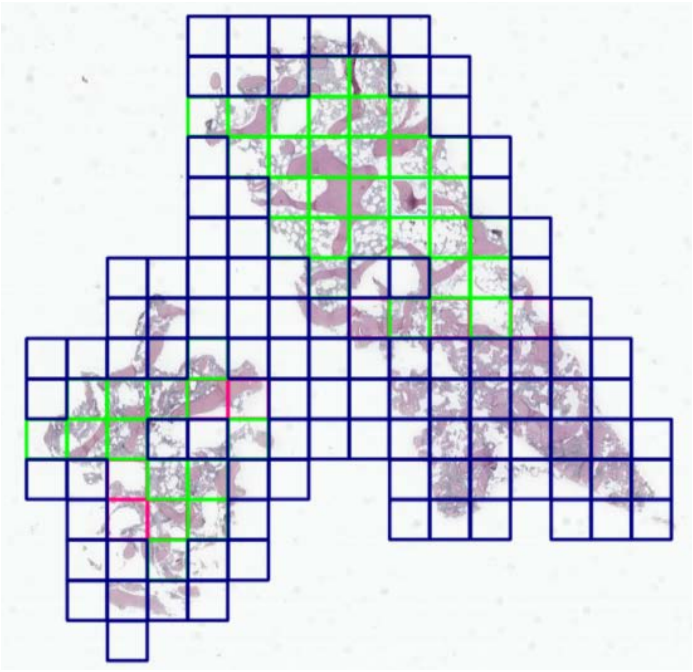
The effective fibrosis region within each patch was calculated as the total area of the patch minus the regions of bone trabecula and blank area.

Supplementary Figure 13. Analysis of Fiber Density for Each Image Patch.



Fiber density is the ratio of the number of fibers and the effective fibrosis region within each image patch. The box plot demonstrates the clear separation in fiber density among patches of different levels, allowing for the grading calculation of individual patches based on fiber density within the patch. MF-0: fiber density ~1550.863; MF-1: 1550.863 ~16384.87; MF-2: 16384.87 ~28336.58; MF-3: 28336.58~ .

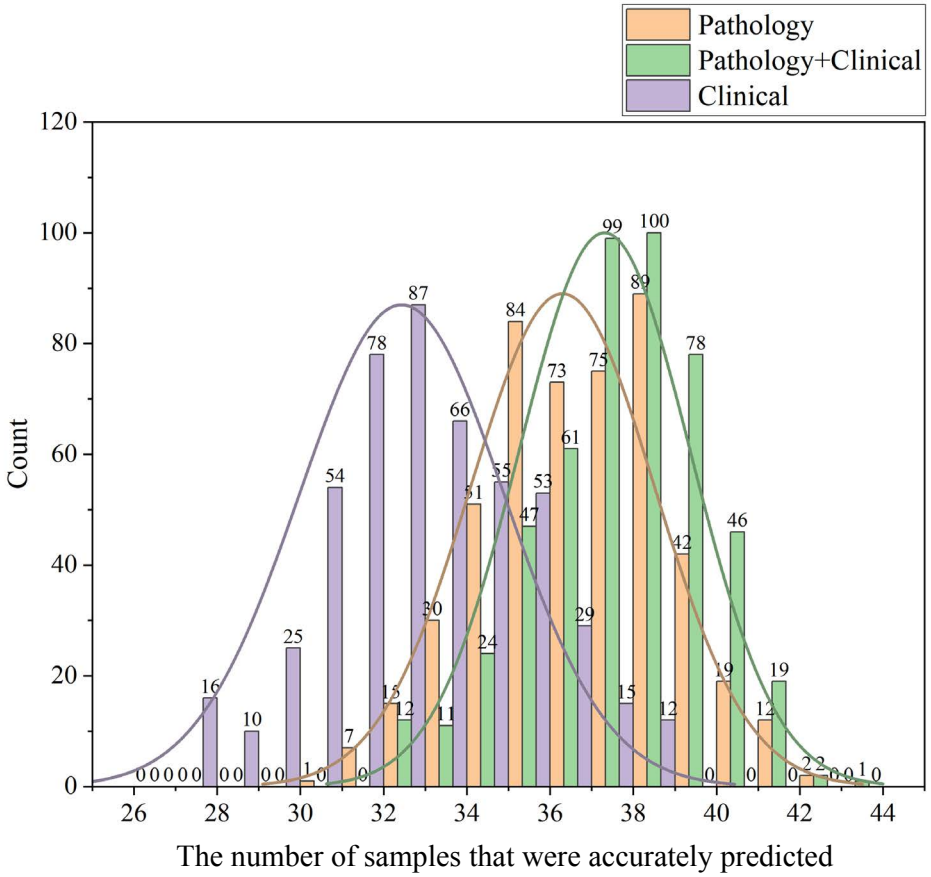
Supplementary Figure 14. Visualization of the severity of fibrosis.



An image of a section sample leveled as MF-1 at 1X magnification. Grid colors of small image patches correspond

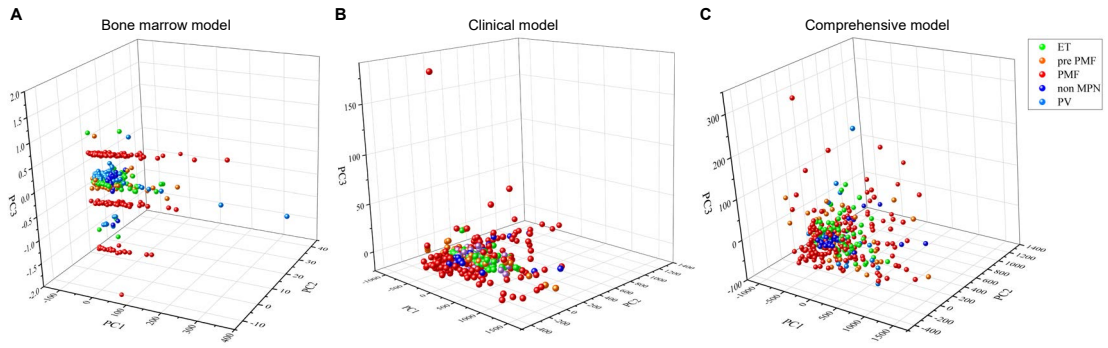
to fibrosis grades green represents MF-0, yellow represents MF-1, pink represents MF-2, red represents MF-3, and blue represents the original patches (pre-grading or excluded from the assessment of fibrosis grade). The fibrosis level was calculated based on the regions with the highest level of effective fibrosis exceeding 30%. The prediction of the fibrosis level of individual samples (MF-0~1 and MF-2~3) achieved an accuracy of 0.916.

Supplementary Figure 15. Performances of The Three Classification Models in the Internal test set (n=45).



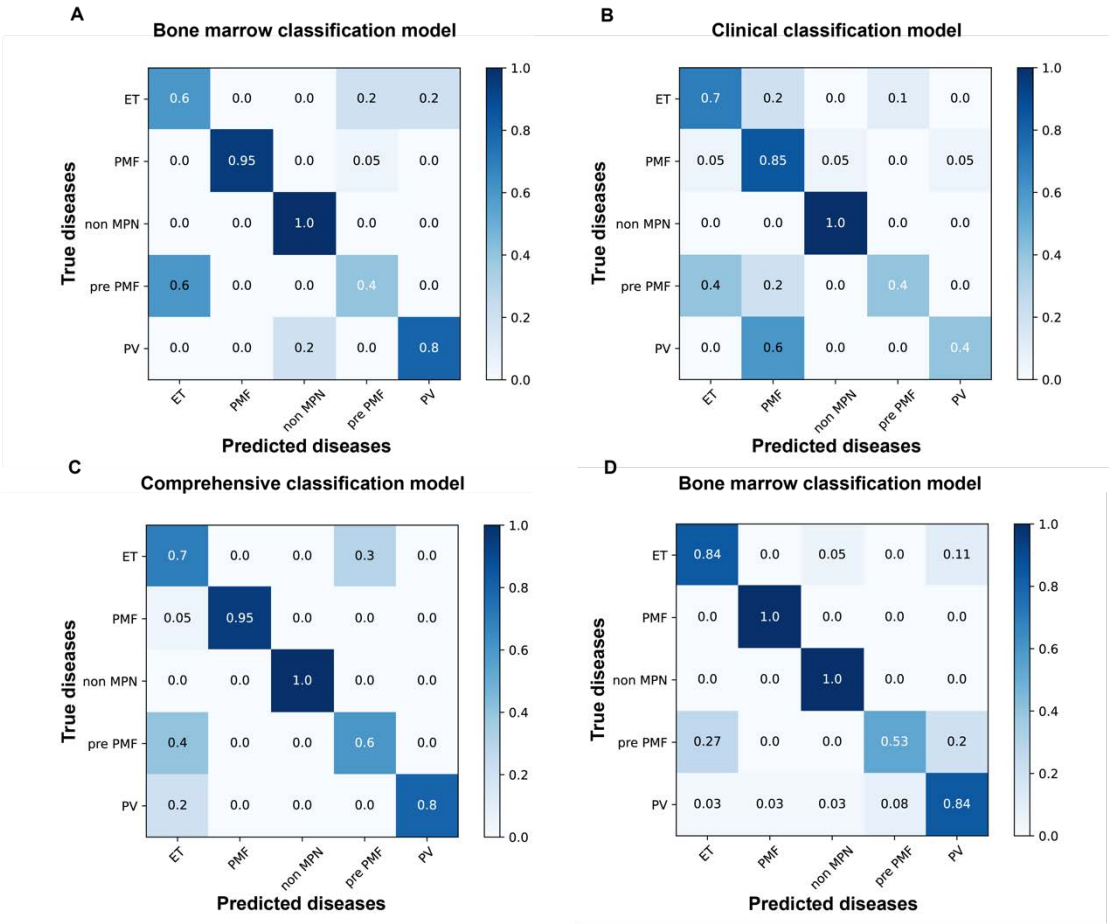
The distribution of correct predictions made by the bone marrow, clinical, and comprehensive models on the internal test set across 500 independent trials. The x-axis represents the number of correctly predicted samples by the three classification models in each independent trial. The y-axis represents the total count of occurrences for each value on the x-axis across 500 trials. The plot showed that the three classification models' highest frequency of correct predictions occurs at 32, 36 and 38 samples, respectively.

Supplementary Figure 16. Visualization of Categorizing Nonneoplastic and MPN subtypes by Classification models.



PCA plots illustrate the separation among non-MPN and MPN subtypes through the bone marrow, clinical, and comprehensive classification models.

Supplementary Figure 17. Classification Performances of Classification Models Applied in Differentiating Nonneoplastic and MPN Subtypes.



The classification performance of models on the internal and external test sets are shown in the confusion matrices. A-C. The classification performance of the internal test set is shown for the bone marrow, clinical, and comprehensive classification model, with average precisions of 0.75, 0.67, and 0.81, respectively. D. The classification performance of the external test set is presented for the bone marrow classification model, achieving an average precision of 0.842. When misclassifications occurred, ET and pre-PMF were most frequently mistaken for one another.

Abbreviations: MPN, myeloproliferative neoplasm; ET, essential thrombocythemia; pre-PMF, prefibrotic PMF; PMF, primary myelofibrosis; PV, polycythemia vera.

1. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*. 2020;106doi:10.1016/j.patcog.2020.107404
2. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016:770-778.
3. Valanarasu JMJ, Patel VM. UNeXt: MLP-Based Rapid Medical Image Segmentation Network. Springer Nature Switzerland; 2022:23-33.
4. Arber DA, Orazi A, Hasserjian RP, et al. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. *Blood*. Sep 15 2022;140(11):1200-1228. doi:10.1182/blood.2022015850
5. Saha PK, Logofatu D. Efficient Approaches for Density-Based Spatial Clustering of Applications with Noise. Springer International Publishing; 2021:184-195.
6. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*. Jul 2022;36(7):1703-1719. doi:10.1038/s41375-022-01613-1
7. Krzywinski M, Altman N. Classification and regression trees. *Nature Methods*. 2017/08/01 2017;14(8):757-758. doi:10.1038/nmeth.4370