A Coarse-to-Fine Pathology Patch Selection for Improving Gene Mutation Prediction in Acute Myeloid Leukemia

Chun-Chia Chiu[1], Jeng-Lin Li[1], Yu-Fen Wang[2], Bor-Sheng Ko[3], Chi-Chun Lee[1]

Abstract— **Identifying gene mutation is essential to prognosis and therapeutic decisions for acute myeloid leukemia (AML) but the current gene analysis is inefficient and non-scalable. Pathological images are readily accessible and can be effectively modeled using deep learning. This work aims at predicting gene mutation directly by modeling bone marrow smear images. Traditionally, bone marrow smear slides are cropped into patches with manual segmentation for patch-level modeling. Slide-level modeling, such as multi-instance learning, could aggregate patches for holistic modeling, though suffer from excessive redundancy. In this study, we propose a discriminative multi-instance approach to select useful patches in a coarse-to-fine process. Specifically, we preprocess a slide into patches by using a trained pre-selector network. Then, we rule out low quality patches in the coarse selection with known prior knowledge, and refine the model using gene-discriminative patches in the fine selection. We evaluate the framework for CEBPA, FLT3, and NPM1 gene mutation prediction and obtain 71.67%, 56.26%, and 56.34% F1-score. Further analysis show the effect of different selection criteria on prediction gene mutations using pathological images.**

Clinical relevance— **This study makes the gene mutation predictable (better than hematologists) from pathological images for AML to improve clinical availability of gene information.**

I. INTRODUCTION

Acute myeloid leukemia (AML) is a genetically heterogeneous clonal malignancy which stops bone marrow cells from maturing and leads to serious hematopoietic failure. The current clinical treatment guideline still results in higher than 50 percent of unsatisfactory relapse and mortality rate. Researchers globally are continuously seeking factors to help determine effective therapeutic strategies. Genetic analysis has been suggested in the leukemia diagnosis and risk stratification guideline from the World Health Organization (WHO) [1]. For example, FMS-like tyrosine kinase 3 (FLT3) gene is enlisted as a factor for poor prognosis while CCAAT/enhancer binding protein alpha (CEBPA) would suggest a favorable prognosis outcome [2]. However, current gene mutation analyses instrument while exists, it remains costly and rare that hinders scalable and efficient use of gene information for clinical decision-making.

The advancement of deep learning for digital pathology has significantly reduced interpretation efforts for gigabyte-sized biopsy slides. Notably, leukemia classification using blood smear images improved clinical efficiency for diagnosis by accelerating cell counts and interpretation procedure [3]. A recent study has found that using a deep learning approach can discriminate specific gene mutations using bone marrow smear pathological images [4]. This work is an exciting advancement since even experienced pathologists are not able to identify the morphological and cytogenetic relationship, and it also opens up opportunity for leveraging gene information with more readily accessible clinical data (pathological images).

When designing a deep learning algorithm for digital pathology, irrelevant regions leading to unreliable prediction and the overwhelming data size causing computational resource issues are two critical challenges. The learning strategies for large pathological images are thus divided into two categories based on the processing granularity: patch-level and slide-level approaches. Patch-level algorithms often rely on manually labeled patches from region of interests (ROIs) and segmented the targeted patches for classifiers. Several algorithms, including support vector machine and convolutional neural network (CNN), have been applied on segmented patches with only one or a few leukocytes [5]. Eckardt et al. has also aggregated the predicted values from patch-level models in an ensemble approach for the prediction on a whole slide [4].

The direct slide-level algorithms attempt to automatically perform prediction and identify important patches from entire scan to avoid the manual segmentation at the same time. Several multi-instance learning (MIL) based methods developed for breast, lung, and kidney images regarded patches as instances which were aggregated in a bag (whole slide). The aggregation was usually implemented as a differentiable pooling layer. Attention mechanism has also been deployed to improve MIL performance using the learned patch weights for survival prediction [6]. Chen et al. have developed a unified memory mechanism to reduce the memory usage in whole-slide training [7]. Shao et al. have proposed a modified Transformer to reduce model complexity and leverage patch relationship for better MIL performance [8]. Although these prior studies have attempted to model the whole slide by reducing computational requirement for deep network training. There has not been a domain knowledge sensitive patch selection with discrminative encoding approach.

In this study, we propose a process of selecting the important patches from AML bone marrow images and encodes with representation used for slide-level gene mutation prediction. Specifically, we develop a coarse selection selector

[1] CC Chiu, JL Li and CC Lee are with Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan and Joint Research Center for AI Technology and All Vista Healthcare, Ministry of Science and Technology, Taiwan (phone: +88635162439. e-mail: joy19980824@gmail.com , cllee@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw).

[2] YF Wang is with AHEAD Medicine, Taipei, Taiwan.

[3] BS Ko is with Department of Hematological Oncology, National Taiwan University Cancer Center, Taipei, Taiwan, and Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan.

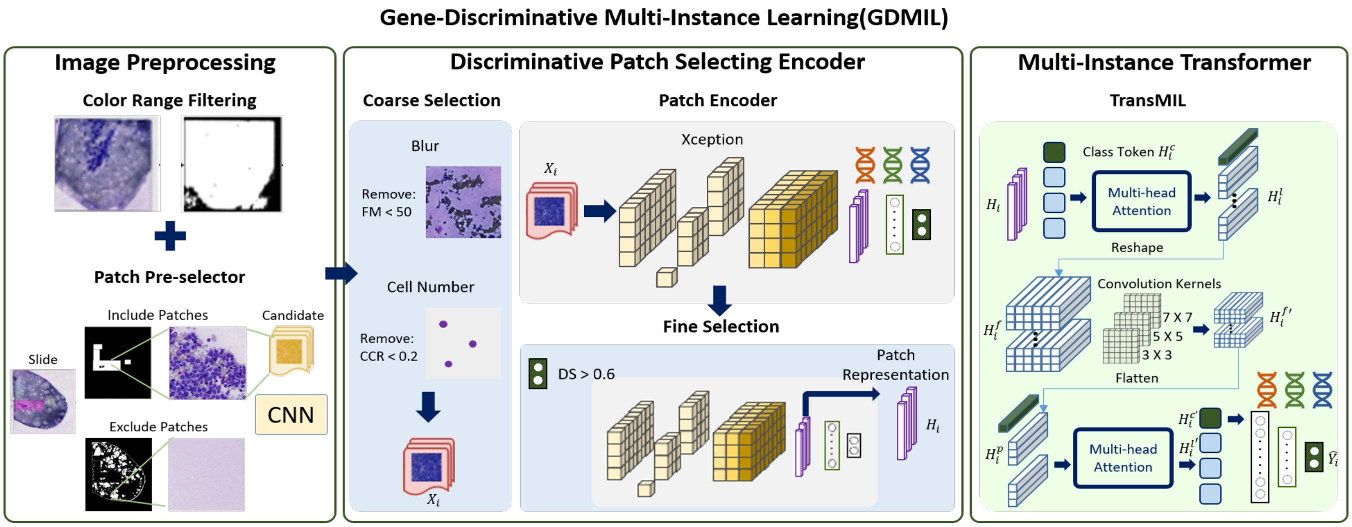

Fig. 1. The overall framework consists of three parts, image preprocessing, discriminative patch selecting encoder and multi-instance transformer. We denote the focus measure, cell coverage ratio, discriminative score of a patch as FM, CCR, and DS, respectively.

based on prior pathological knowledge. The patches are embedded as representations using a patch encoder network that is discriminatively trained to predict gene mutations using each single patch. Then, we perform a fine-grained selection based on the single-patch prediction output probability from the patch encoder. The preserved gene discriminative patches are finally aggregated by a TransMIL model for patient-level gene prediction. With this coarse-to-fine selecting design, we achieve 71.67%, 56.26%, and 56.34% F1-score and 81.97%, 66.54%, and 61.91% AUC for CEBPA, FLT3 and NPM1 gene prediction on an AML bone marrow image dataset.

II. Methods

A. Database and Study Population

The data cohort in this study was collected from the National Taiwan University Hospital[1] and contained 386 bone marrow smear slides from different patients. All patients had gene testing for both FLT3 and NPM1 and 308 of them had tests for CEBPA. The slide and patch distribution are shown in Table I where the genes are labeled as 1 if mutated and otherwise 0. A hematologist annotated 76 slides in a 3000x3000 downsampled scale using the Computer Vision Annotation Tool (CVAT) for the ROIs. The CVAT allowed the hematologist labeling a region as an included or excluded region based on a general morphological pattern instead of a cell-level delicate segmentation. These labeled ROIs were used to train a patch pre-selector described in Section II-B.1 which avoided labeling the whole dataset.

B. Bone Marrow Smear Image Classification

Fig. 1 shows the overall framework which consists of image preprocessing, discriminative patch selecting encoder and multi-instance transformer.

TABLE I
Gene mutation distribution in the dataset. The genes are labeled as 1 if mutated and otherwise 0.

	patient (slide)			patch		
Gene	0	1	total	0	1	total
CEBPA	257	51	308	25108	5056	30164
FLT3	304	82	386	29850	8064	37914
NPM1	315	71	386	30855	7059	37914

1) Image Preprocessing: In the preprocessing stage, we segment the 20X magnified images by defining a valid color range, (107, 12, 0) to (136, 255, 255) in hue, saturation, lightness (HSV) dimensions. The color range specifies the smear boundary from background. Then, we crop each image into patches of size 512x512x3. The choice of high resolution helps preserve details and sufficient number of candidates for further gene-discriminative information extraction. Additionally, we used patches from the 76 slides with labeled ROI regions to train a CNN model and predicted whether to include each patch or not. This model is named as “patch pre-selector” that is used to rank the most appropriate patches to be includes by sorting the prediction score (here, we take the top 100 patches from each bone smear slide).

2) Discriminative Patch Selecting Encoder: In this section, we introduce a coarse-to-fine selecting mechanism along with a patch encoder to discriminatively preserve the gene-related information in representations. The coarse-to-fine selecting mechanism includes coarse selection rules and a fine selection rule applied before and after the patch encoder network. The coarse selection rules remove low quality patches which are constructed by hematologists’ prior knowledge. We detected blur patches using a focus measure (FM) value that computes variations of the Laplacian; it measures the second derivatives for intense change of pixels. If the focus measure was lower than a threshold, we removed the patch for blurriness. We defined another rule, cell coverage ratio (CCR), which removed the patches with very few

[1] IRB: 201906018RINB

TABLE II

RESULTS OF GENE PREDICTION USING DIFFERENT ALGORITHMS WITH BEST PERFORMED TWO VALUES ON EACH METRIC HIGHLIGHTED IN BOLD.

Gene	CEBPA				FLT3				NPM1			
Metric(%)	UAR	F1	ACC	AUC	UAR	F1	ACC	AUC	UAR	F1	ACC	AUC
InceptionV3+AvgPool	71.65	66.09	76.30	79.67	57.80	54.03	60.88	61.49	55.44	51.01	59.33	58.81
Xception+AvgPool	**72.24**	66.14	75.97	80.63	56.30	53.17	60.62	61.01	**57.93**	53.45	66.06	**63.94**
Xception+AMIL	64.98	61.70	74.35	69.87	58.30	**56.61**	66.58	**62.67**	56.36	**55.82**	**71.50**	60.74
Xception(Fixed)+TransMIL	53.86	50.14	61.04	58.19	56.22	51.49	56.99	60.32	51.02	50.83	68.13	53.27
Xception+TransMIL	71.63	**68.18**	**80.19**	**80.95**	**59.42**	**56.88**	**65.54**	61.07	57.68	54.41	64.77	59.10
GDMIL	**73.97**	**71.67**	**82.79**	**81.97**	**61.29**	56.26	62.18	**66.54**	**58.43**	**56.34**	**68.65**	**61.91**

cells by computing the area with stained color. The low CCR indicates that rare nucleated cells are in the patch.

After the coarse selection, we built a "patch encoder" to embed gene mutation related information into a latent space by a deep Xception network. The Xception model was pretrained on ImageNet dataset and we performed transfer learning using patch data downsampled to a class-balanced condition. The learned patch encoder could estimate patch-level gene mutation prediction scores, and we preserved the patches in the training set which were correctly predicted with the prediction scores higher than a discriminative score (DS) threshold. Therefore, the patch encoder was further retrained by the patches that could better describe the pathological characteristics for gene mutations in a high confidence. We extracted the latent representation of each patch for the following multi-instance learning.

3) Multi-Instance Transformer: For a set of N patients passed through the selection described in Section II-B.2, the whole slide images X_i where $i = 1, ..., N$ correspond to the binary gene mutation labels Y_i. With the patch encoder, a bag of patches $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,T}\} \in R^{T \times h \times w \times c}$ from a patient were transformed as hidden representations $H_i = \{h_{i,1}, h_{i,2}, ..., h_{i,T}\} \in R^{T \times d}$, where T is the number of patches for the patient, h, w and c are the height, width and channel number of a patch, d is the dimension of a patch representation. We proposed to use a transformer based correlated multiple instance learning (TransMIL) approach [8] to aggregate varied number of patches in a bag and therefore perform gene mutation prediction on a patient level. The TransMIL firstly transformed a representation concatenating H_i and a class token H_i^c to $H_i^l \in R^{(T+1) \times d}$ using a multi-head attention layer and reshaped the representation of H_i^l except the class token dimension as a matrix $H_i^f \in R^{\sqrt{T} \times \sqrt{T} \times d}$. Multiple convolutional layers with different kernel sizes were applied to extract the relations between the patches and resulted in a matrix $H_i^{f'} \in R^{\sqrt{T} \times \sqrt{T} \times d}$. We then obtained relationship embedded representations $H_i^p \in R^{(T+1) \times d}$ by flattening $H_i^{f'}$ and concatenated H_i^p with the class token representation for another multi-head attention layer transformation. Finally, we used the class token $H_i^{c'} \in R^{1 \times d}$ extracted from the derived $H_i^{l'} \in R^{(T+1) \times d}$ in a fully-connection network to predict gene mutation $\hat{Y}_i$.

C. Experimental Analysis

To evaluate our pre-processing stage (pre-selector), we sampled 882 gathered patch images evenly distributed from 96 patients for hematologists to rate. The hematologists manually annotated theses sampled patches as abnormal, normal, and unknown based on their experiences. Our of these 882 sample patches (output from our pre-selector), there are 695 patches labeled as abnormal (78.80%), 187 as unknown patches (21.20%), and none normal patches (0%). This analysis demonstrates out of all the automatically selected samples to be included in our model learning, about 80% of them are deemed important, i.e., with pathological forms of being abnormal.

The aim of our proposed framework is to perform a binary gene mutation prediction task. The targeted gene mutation labels were CEBPA, FLT3, Nucleophosmin (NPM1). We carried out a 5-fold cross-validation experiments (60% for training, 20% for validation, and 20% for testing) in each fold. The model hyper-parameters were determined by grid-search for the best validation performance. The patch encoder was trained using weighted cross-entropy loss, Adam optimizer, a learning rate as 1e-3 and a batch size as 16. For transMIL, cross-entropy loss is optimized by Lookahead optimizer with a learning rate of 2e-4 and a batch size of 1. The focus measure (FM) threshold, cell coverage ratio (CCR) threshold, and discriminative score (DS) threshold described in Section II-B.2 are set to 50, 0.2, and 0.6, respectively.

1) Exp I: Comparison of patch encoder and MIL approaches: Four metrics including UAR, unweighted f1 score (F1), accuracy (ACC), and area under the ROC curve (AUC), are used to evaluate the following models.

- InceptionV3+AvgPool: using the InceptionV3 [9] as the patch encoder and aggregating patches by average pooling
- Xception+AvgPool: replacing *InceptionV3+AvgPool* with Xception [10] for the patch encoder
- Xception+AMIL: using an Attention-based MIL approach [11] for patch aggregation
- Xception(Fixed)+TransMIL: using fixed pre-trained Xception model as the patch encoder and a TransMIL approach [8] for patch aggregation
- Xception+TransMIL: Our proposed framework without discriminative patch selection mechanism
- GDMIL: Our proposed framework denoted as Gene-Discriminative MIL

2) Exp II: Analysis of different selecting factors: In this section, we aim to investigate the effects of different factors to the gene mutation prediction task. The factors contain parameters in the coarse selection, fine selection, and bag

TABLE III
RESULT ANALYSES ON SELECTING FACTORS IN TERMS OF UAR.

Gene		CEPBA	FLT3	NPM1
GDMIL		73.97	61.29	58.43
Coarse Selection	FM $>$ 60	71.44	**62.66**	**60.23**
	CCR $>$ 0.3	72.06	57.59	57.40
	Non-Boundary	**76.17**	60.10	56.93
Fine Selection	DS $>$ 0.7	**74.77**	58.33	**58.48**
Bag Order	Spatial	73.58	60.06	**60.38**

order of TransMIL. For the coarse selection, prior knowledge based factors include different thresholds of blurriness (FM), thresholds of cell coverage (CCR), and an additional removal of patches close to the smear boundary. For the fine selection, we increase discriminative scores (DS) to examine the results with higher gene-related confidence but fewer patches. We also reshape patches by spatial order for TransMIL.

III. RESULTS

In this study, we compare to different slide learning algorithms and the results are shown in Table II. We observe that our proposed *GDMIL* consistently outperforms the other approaches across all the metrics (73.97% UAR, 71.67% F1, 82.79% ACC, and 81.97% AUC) for CEBPA gene mutation prediction. The 61.29% UAR and 66.54% AUC for FLT3 and the 58.43% UAR and 56.34% F1 for NPM1 using *GDMIL* are also the highest results. The two strong baselines fine-tuning from the InceptionV3 and Xception pretrained networks for the patch encoder obtain 71.65%, 57.80%, and 55.44% UAR (*InceptionV3+AvgPool*) and 72.24%, 56.30%, and 57.93% UAR (*Xception+AvgPool*) for CEBPA, FLT3, and NPM1. These two approaches can effectively capture characteristics in each patch. The state-of-the-art *Xception+TransMIL* modifying the attention mechanism in *AMIL*, the derived UAR are 71.63%, 59.42%, and 57.68% which result in an improvement of 6.65%, 1.12%, 1.32% for CEBPA, FLT3, and NPM1 compared to *Xception+AMIL*. The original paper used a pretrained patch encoder without fine-tuning (*Xception(Fixed)+TransMIL*) which could deteriorate the performance due to the domain gap. Our proposed *GDMIL* using additional selection mechanism attains further improvement compared to *Xception+TransMIL* with 2.34%, 1.87%, and 0.75% UAR.

The analysis results in terms of UAR are shown in Table III. We obtain improved UAR (62.66% and 60.23%) for FLT3 and NPM1 by increasing the FM threshold from 50 to 60. Removing 15% patches close to the boundary can significantly improve the performance to 76.17% UAR for CEBPA while stricter CCR threshold does not add improvement. Increasing DS threshold from 0.6 to 0.7 in the fine selection benefits the prediction of CEBPA. Organizing patches based on spatial order in TransMIL helps the performance achieve 60.38% UAR for NPM1.

IV. DISCUSSIONS

Our experiment results show an encouraging coarse-to-fine patch selection scheme to incorporate both human knowledge and discriminative learning in the gene prediction task. Our proposed *GDMIL* outperforms the other algorithms with the advantage of selected patches. We find that transfer learning is essential in the comparison to the state-of-the-art approach, *Xception(Fixed)+TransMIL*, using a fixed patch encoder. Intriguing observations are revealed in Exp II that the mutation of CEBPA is sensitive to redundant patches, such as locating at smear boundary or less discriminative in the patch encoder. Multiple factors affect the performance of NPM1 mutation prediction while most do not affect FLT3.

V. CONCLUSIONS

In this study, we propose a framework to select discriminative patches for multi-instance gene prediction. The more gene-related patches and the better patch representation learning improve the performance on three different gene mutations. We observe that different gene mutation would be sensitive to different factors in the selection process. In the future work, we will expand the dataset and examine the algorithms to contribute for different diseases and genes.

REFERENCES

[1] Hartmut Döhner, Elihu Estey, David Grimwade, Sergio Amadori, Frederick R Appelbaum, Thomas Büchner, Hervé Dombret, Benjamin L Ebert, Pierre Fenaux, Richard A Larson, et al., "Diagnosis and management of aml in adults: 2017 eln recommendations from an international expert panel," *Blood, The Journal of the American Society of Hematology*, vol. 129, no. 4, pp. 424–447, 2017.

[2] Jifeng Yu, Yingmei Li, Danfeng Zhang, Dingming Wan, and Zhongxing Jiang, "Clinical implications of recurrent gene mutations in acute myeloid leukemia," *Experimental hematology & oncology*, vol. 9, no. 1, pp. 1–11, 2020.

[3] Mustafa Ghaderzadeh, Farkhondeh Asadi, Azamossadat Hosseini, Davood Bashash, Hassan Abolghasemi, and Arash Roshanpour, "Machine learning in detection and classification of leukemia using smear blood images: a systematic review," *Scientific Programming*, vol. 2021, 2021.

[4] Jan-Niklas Eckardt, Jan Moritz Middeke, Sebastian Riechert, Tim Schmittmann, Anas Shekh Sulaiman, Michael Kramer, Katja Sockel, Frank Kroschinsky, Ulrich Schuler, Johannes Schetelig, et al., "Deep learning detects acute myeloid leukemia and predicts npm1 mutation status from bone marrow smears," *Leukemia*, pp. 1–8, 2021.

[5] Luis HS Vogado, Rodrigo MS Veras, Flavio HD Araujo, Romuere RV Silva, and Kelson RT Aires, "Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 415–422, 2018.

[6] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, pp. 101789, 2020.

[7] Chi-Long Chen, Chi-Chung Chen, Wei-Hsiang Yu, Szu-Hua Chen, Yu-Chan Chang, Tai-I Hsu, Michael Hsiao, Chao-Yuan Yeh, and Cheng-Yu Chen, "An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning," *Nature communications*, vol. 12, no. 1, pp. 1–13, 2021.

[8] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang, "Transmil: Transformer based correlated multiple instance learning for whole slide image classication," *arXiv preprint arXiv:2106.00908*, 2021.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[10] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[11] Maximilian Ilse, Jakub Tomczak, and Max Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.