Automated Report Generation for Lung Cytological Images Using a CNN Vision Classifier and Multiple-Transformer Text Decoders: Preliminary Study

Atsushi Teramoto, Ayano Michiba, Yuka Kiriyama, Tetsuya Tsukamoto,
Kazuyoshi Imaizumi, and Hiroshi Fujita

Abstract—Cytology plays a crucial role in lung cancer diagnosis. Pulmonary cytology involves cell morphological characterization in the specimen and reporting the corresponding findings, which are extremely burdensome tasks. In this study, we propose a report-generation technique for lung cytology images. In total, 71 benign and 135 malignant pulmonary cytology specimens were collected. Patch images were extracted from the captured specimen images, and the findings were assigned to each image as a dataset for report generation. The proposed method consists of a vision model and a text decoder. In the former, a convolutional neural network (CNN) is used to classify a given image as benign or malignant, and the features related to the image are extracted from the intermediate layer. Independent text decoders for benign and malignant cells are prepared for text generation, and the text decoder switches according to the CNN classification results. The text decoder is configured using a Transformer that uses the features obtained from the CNN for report generation. Based on the evaluation results, the sensitivity and specificity were 100% and 96.4%, respectively, for automated benign and malignant case classification, and the saliency map indicated characteristic benign and malignant areas. The grammar and style of the generated texts were confirmed as correct and in better agreement with gold standard compared to existing LLM-based image-captioning methods and single-text-decoder ablation model. These results indicate that the proposed method is useful for pulmonary cytology classification and reporting.

Index Terms—Convolutional neural network, cytology, deep learning, image captioning, report generation, Transformer

I. Introduction

CANCER statistics show that lung cancer has the highest incidence and death rate among all cancers [1]. In addition to surgery, effective methods for treating lung cancer, including radiotherapy and chemotherapy, have been developed. Early and accurate diagnosis is necessary to obtain good therapeutic outcomes with these treatment modalities.

Pathological diagnosis, which involves cytological and histological analyses, plays an important role in detailed lung cancer diagnosis. In cytological diagnosis, a specimen is prepared from cells collected via bronchoscopy or other procedures, and a cytotechnologists or cytopathologists uses a microscope to observe the cell nucleus, cytoplasm, and cell arrangement to distinguish benign from malignant cases and determine the histological type; the results are then described in a report. However, cytological diagnosis is time-consuming and burdensome because it is extremely complicated and requires the observation of a large number of cells under a microscope while writing reports. If the image analysis and report generation processes could be automated, the diagnosis efficiency would improve, and the burden associated with diagnosis could be greatly reduced.

Various classification methods have been developed for cytology. Zhang et al. [2] proposed a convolutional neural network (CNN)-based cervical cell classification method. Fine-tuning of the CNN pre-trained on ImageNet using two datasets resulted in a 98.3% classification accuracy. Gelardi et al. [3] used a three-block CNN to classify nasal cytology images and obtained a 99.0% classification accuracy. Bal et al. [4] used a CNN to differentiate ductal carcinoma of the breast using hematoxylin and eosin (HE)- and Giemsa-stained breast specimens; the classification accuracy for the HE- and Giemsa-stained specimens was 96.5% and 97.5%, respectively. In contrast, we focused on lung cytology and conducted several studies. First, a simple five-layer CNN was used to classify the lung cancer tissue types in Papanicolaou-stained lung specimens [5]; obtaining a classification accuracy of 71.1%. Subsequently, by integrating the results of fine-tuning multiple pre-trained CNNs, the classification accuracy improved to

78.9% [6]. In addition, the pre-trained CNNs were used for benign and malignant lung cell classification, and a classification accuracy of 79.2% was obtained [7]. Moreover, as the first cytology study using generative artificial intelligence (AI), we proposed a method using images generated by generative adversarial networks and classification accuracy was improved to 85.3% [8].

Several studies have been conducted to generate diagnostic reports using radiological images. For example, Wang et al. [9] proposed a report-generation method with a network combining a CNN and a recurrent neural network (RNN), which was evaluated using the ChestXRay-14 dataset with a BLEU-1 of 0.286 and ROUGE-L of 0.226 [9]. Hou et al. [10] proposed a report-generation method that combined a CNN and Transformer. DenseNet121 was used as the CNN and its features were used in the Transformer to generate text sentences, with evaluation results of 0.232 for BLEU-1 and 0.240 for ROUGE-L. Studies on report generation for pathological images have been conducted on histological images. Zhou et al. [11] used a graph-neural network and Transformer to analyze histological images of urothelial papillary carcinoma and the structural cell information extracted from them to generate a report. The accuracy of this report was 0.608 for BLEU-4 and 0.409 for SPICE, with a classification accuracy of 79.0%.

To the best of our knowledge, no report-generation method for cytological images has been reported to date. In this paper, we propose a hybrid scheme for diagnosis and report generation for lung cytological images that combines our previously developed classification method [7] with a text-generation method. The main contributions of this study can be summarized as follows.

1) A report-generation method for cytology specimens.
2) We propose an end-to-end model consisting of a cell-classification model and a Transformer-based text decoder that is selected based on the results.
3) A highly accurate classification model improves the diagnostic accuracy and generates higher-quality reports than existing image captioning methods.
4) The proposed model provides classification results, reports, and saliency maps, and can be used as a support tool for cytological diagnosis in terms of understanding, transparency, and accuracy.

II. METHODS

A. Outline

An overview of the proposed scheme is shown in Fig. 1. The vision model uses a CNN fine-tuned using cytological images to distinguish malignant from benign cells and extracts features for text decoders. Based on the CNN classification results, text decoders (Text Decoders #1 and #2) are invoked for benign and malignant cells, and a report is generated based on the CNN features. The saliency map generated by the Grad-CAM is also exported to the report. To confirm the effectiveness of the proposed method, we compared it with a model with a single-text decoder and existing state-of-the-art (SOTA)

image-captioning models. Cytological images are provided to the CNN, which classifies them as benign or malignant. Based on the results, the two text decoders are switched and a report is generated based on the CNN-provided image features.

Fig. 1. Outline of the proposed report-generation scheme.

B. Dataset

This study was conducted as a retrospective study with approval from the Institutional Review Board of Fujita Health University (IRB No. HM23-390). Informed consent was obtained from all patients under data anonymization. All experimental protocols were performed in compliance with the Declaration of Helsinki and in accordance with relevant guidelines and regulations. In this study, lung cells were collected from 206 patients via interventional cytology, using either bronchoscopy or computed-tomography-guided fine-needle aspiration cytology; the collected cells consisted of 71 benign and 135 malignant cases. The malignant cases included 83 adenocarcinomas and 52 squamous cell carcinomas. These diagnoses were combined with the histological analysis of the biopsy specimens for the final determination. Biopsy tissues were collected at the same time as the cytology specimens, fixed in 10% formalin, dehydrated, and embedded in paraffin. The cytological specimens were prepared via liquid-based cytology using the BD SurePathTM liquid-based Pap test (Beckton Dickinson, Franklin Lakes, NJ, USA) and stained using the Papanicolaou method. A microscope (BX53, Olympus Corporation, Tokyo, Japan) attached to a digital camera (DP20, Olympus Corporation) was used to acquire 219 microscopic benign cell images and 460 malignant cell images in JPEG format with a size of 1280×960 pixels.

To construct the image dataset for report generation, a cytotechnologist and cytopathologist extracted a 296×296-pixel patch image of the area where cells were present from the original microscopic image, as shown in Fig. 2. The final dataset consisted of 797 patch images, with 325 benign and 472 malignant cell images. Microscopic findings were then prepared for these images to describe the cell type, shape of the cell nucleus, cell arrangement, and background, which represented conditions other than the target cells. The method used to describe the findings was based on the World Health Organization (WHO) reporting system for lung cytopathology [12]. A sample of the dataset is shown in Fig. 3.

Finally, the datasets are randomly divided into training and evaluation datasets. The training dataset consisted of 270 images from 57 benign cases and 377 images from 108 malignant cases, while the evaluation dataset consisted of 55 images from 14 benign cases and 95 images from 27 malignant cases. Images of the same patient were not mixed in the training and evaluation datasets.

Fig. 2. Preparation of patch images and the report. Cytology specimens were taken under a microscope, and experts selected the areas of the specimens that should be output as a report and converted into a patch image. A report corresponding to each image was then prepared.

	The background is relatively clean. There is the clusters of columnar epithelial cells
	There are columnar epithelial cells in the inflammatory background.
	There are adenocarcinoma cells with hyperchromatic nucleus, nucleus with irregular shape, pale cytoplasm, promient nucleoli, filled with foamy mucus.
	There are squamous cell carcinoma cells with keratinized squamous cell, hyperchromatic nucleus, nucleus with irregular shape.

Fig. 3. Four examples of patch images and the corresponding report description in our dataset.

C. Vison Model

Vision models have multiple roles. First, a given patch image is classified as benign or malignant. Fig. 4 shows the structure of the CNN-based visual model used in this study. For the CNN architecture, we introduced VGG16 [13], InceptionV3 [14], ResNet50 [15], and DenseNet121 [16], which are existing SOTA models. The parameters of these CNNs were pre-trained using the ImageNet dataset. The fully connected layers of the original CNNs were replaced with ones comprising 1024 units and 3units; and the entire network was fine-tuned using actual patch images. The images were rotated 90° for data augmentation during training.

For classification using CNNs, to see the regions of interest along with the output report, we used Grad-CAM to obtain a saliency map [17]. When a CNN is used for classification into two classes, the background area, where no object exists, may have a high value on the saliency map. To focus on the target benign and malignant cells for analysis, we added one category for no cells, similar to object detection models [18]. In addition to benign and malignant images, 300 background images without lung cells were prepared, and a CNN was trained to classify these three categories. The benign and malignant cell probabilities were compared and the category with the highest probability was used as the classification result.

In the training of each CNN model used in this study, we used the Adam optimization algorithm, a learning coefficient of $1.0×10^{-5}$, a batch size of 16, and a fixed number of 50 training epochs, after which the training was terminated automatically if the validation error did not improve. A CNN classifies images into benign, malignant, and background. The features obtained from the intermediate layer are used in the text decoder.

Fig. 4. Structure of the vision model.

D. Text Decoder

The text decoder outputs the text based on the image features provided by the vision model. Before being input into the text decoder, the text provided as training data was divided into tokens by separating them with spaces and punctuation marks. Then, each token was converted into an integer value using a vectorizer, which also provided positional information to vectorize the text. Next, the text was decoded using a Transformer with multiple Transformer layers, as shown in Fig. 5. Each layer consisted of a multi-headed causal self-attention layer, a cross-attention layer, and a fully connected layer with 512-256 units. Finally, the output-vectorized text was converted into words using the reverse tokenization procedure to obtain the final text.

In this study, two text decoders specializing in benign and malignant cells were implemented, and the decoder was switched according to the classification results of the vision model. Each decoder was trained to generate reports only for benign and malignant cells, and the learning parameters for the text decoder were the Adam optimization algorithm and a learning coefficient of $1.0×10^{-4}$. Regarding the number of training epochs, training was terminated if the validation error did not improve after 100 epochs.

Fig. 5. Text decoder structure. The Transformer layer consists of a multi-head causal self-attention layer, a cross-attention layer, and a feed forward network. The output is the vectorized text.

E. Evaluation Metrics

We evaluated the classification performance of the CNNs used in the proposed method and the quality of the report-generation model. First, to evaluate the CNN performance for classifying benign and malignant cases, we performed five iterations with multiple CNN-based SOTA models and calculated their detection sensitivity, specificity, and average (balanced accuracy). Based on the results, we selected the CNN-based model to be used in the vision model of the proposed method.

Next, we evaluated the image-captioning techniques by assessing the agreement between the ideal and generated

reports using five metrics: BLEU [19], METEOR [20], ROUGE [21], CIDEr [22], and SPICE [23].

The number of layers and heads of the Transformer used for text decoding can vary, and the performance of the proposed method changes accordingly. Therefore, we evaluated the performance of the proposed method by varying the number of layers between one, two, and three, and the number of heads between two, four, and six.

As an ablation study, we also built a single-text decoder model of the proposed method that was shared by both the benign and malignant cells and compared the performance of the proposed method's Transformer. The aforementioned number of heads and layers of the Transformer were also optimized in the same manner as in the proposed method.

In addition, we fine-tuned the major captioning models, namely, GIT [24], BLIP [25], and BLIP2 [26], using large natural language models (LLMs) and compared them with the proposed method. For GIT and BLIP, we used the Base and Large models, while for BLIP2, we used a model incorporating OPT [27] with parameters of 2.7B and 6.7B as the LLM. For these processes, a PC with an NVIDIA RTX 6000Ada GPU and an AMD Ryzen9 CPU was used.

III. RESULTS

Table I presents the classification performance evaluation results for the CNN used as the vision model. In this evaluation, fivefold cross-validation was repeated five times using the training data, and the average and standard deviations of the sensitivity, specificity, and balanced accuracy were obtained. By comparing the classification performance of the multiple CNN models, it was found that the balanced accuracy did not differ significantly among the models. Considering the importance of sensitivity in classifying benign and malignant cases, it was determined that ResNet50 was the best CNN model for use as a vision encoder.

Subsequently, ResNet50 was trained using all the training data prepared in this study, and the confusion matrix of the classification procedure was calculated using the test data, as presented in Table II. The sensitivity and specificity for

detecting malignant cells were 100% and 98.1%, respectively.

TABLE I
CLASSIFICATION PERFORMANCE OF THE CNNS FOR THE VISION MODEL

CNN MODEL	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY
RESNET50	0.981±0.019	0.938±0.046	0.958±0.024
VGG16	0.979±0.015	0.989±0.010	0.962±0.027
INCEPTIONV3	0.924±0.029	0.924±0.047	0.953±0.025
DENSENET121	0.966±0.040	0.924±0.076	0.945±0.035

TABLE II
CLASSIFICATION CONFUSION MATRIX (RESNET50)

		PREDICTED	
		BENIGN	MALIGNANCY
ACTUAL	BENIGN	54	1
	MALIGNANCY	0	95

Next, we processed and evaluated the entire report-generation method, using ResNet50 as the CNN for the vision model and changing the number of layers and heads used in the Transformer. Table III presents the evaluation accuracy results for the reports generated using BLEU, ROUGE, METEOR, CIDEr, and SPICE. As the proposed method uses different text decoders for benign and malignant cells, the evaluation was divided into benign and malignant cell cases. For benign cells, the smallest model with one layer and two heads exhibited the best generated text output, whereas for malignant cells, the model with two layers and four heads exhibited the best performance.

Table IV presents the evaluation results for the proposed method, the single-text decoder prepared as an ablation study, and the existing SOTA image-captioning models, including GIT, BLIP, and BLIP2. The Hugging-Face library was used to implement the captioning process using GIT, BLIP, and BLIP2.

TABLE III
NETWORK STRUCTURE OPTIMIZATION IN THE TEXT DECODER

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
LAYER 1 HEAD 2	BENIGN	**0.959**	**0.950**	**0.941**	**0.932**	**0.683**	**0.966**	**8.509**	**0.946**
	MALIGNANT	0.850	0.798	0.749	0.714	0.487	0.841	4.598	0.715
LAYER 1 HEAD 4	BENIGN	0.961	0.950	0.939	0.928	0.671	0.965	8.170	0.942
	MALIGNANT	0.849	0.796	0.746	0.712	0.485	0.835	4.419	0.684
LAYER 1 HEAD 6	BENIGN	0.950	0.938	0.926	0.914	0.653	0.950	7.976	0.920
	MALIGNANT	0.852	0.797	0.744	0.709	0.486	0.837	4.651	0.723
LAYER 2 HEAD 2	BENIGN	0.959	0.949	0.939	0.928	0.675	0.966	8.340	0.943
	MALIGNANT	0.847	0.791	0.736	0.699	0.481	0.835	4.459	0.717
LAYER 2 HEAD 4	BENIGN	0.944	0.926	0.908	0.888	0.623	0.943	7.103	0.906
	MALIGNANT	**0.853**	**0.804**	**0.756**	**0.724**	**0.494**	**0.844**	**4.883**	0.718
LAYER 2 HEAD 6	BENIGN	0.950	0.933	0.917	0.901	0.637	0.946	7.539	0.925
	MALIGNANT	0.853	0.800	0.749	0.714	0.489	0.842	4.735	0.710
LAYER 3 HEAD 2	BENIGN	0.954	0.939	0.923	0.905	0.639	0.956	7.297	0.922
	MALIGNANT	0.824	0.767	0.710	0.672	0.462	0.814	4.025	0.674
LAYER 3 HEAD 4	BENIGN	0.949	0.938	0.927	0.916	0.668	0.957	8.230	0.938
	MALIGNANT	0.852	0.799	0.748	0.713	0.488	0.842	4.704	**0.721**
LAYER 3 HEAD 6	BENIGN	0.940	0.924	0.908	0.891	0.627	0.941	7.273	0.902
	MALIGNANT	0.841	0.787	0.736	0.700	0.480	0.833	4.563	0.702

In the Transformer for benign and malignant tumors, as shown in Fig. 5, the numbers of layers and heads were changed to assess the accuracy of the report. The indexes other than CIDEr have higher performance as they approach 1, and the larger the CIDEr, the higher the performance.

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND SOTA MODELS

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Proposed Method	BENIGN	0.959	0.950	0.941	0.932	0.683	0.966	8.509	0.946
	MALIGNANT	**0.853**	**0.804**	**0.756**	**0.724**	**0.494**	**0.844**	**4.883**	0.718
	BALANCED	**0.906**	**0.877**	**0.849**	**0.828**	**0.589**	**0.905**	**6.696**	0.832
CNN + single text decoder	BENIGN	0.959	0.948	0.937	0.927	0.670	0.960	8.339	0.945
	MALIGNANT	0.835	0.777	0.722	0.685	0.471	0.816	4.370	0.689
	BALANCED	0.897	0.863	0.830	0.806	0.571	0.888	6.355	0.817
GIT Base	BENIGN	0.940	0.924	0.911	0.899	0.636	0.935	7.976	0.921
	MALIGNANT	0.810	0.745	0.681	0.633	0.446	0.797	3.139	0.676
	BALANCED	0.875	0.835	0.796	0.766	0.541	0.866	5.558	0.799
GIT Large	BENIGN	0.941	0.926	0.911	0.897	0.633	0.939	7.554	0.913
	MALIGNANT	0.821	0.752	0.685	0.636	0.448	0.803	3.073	0.691
	BALANCED	0.881	0.839	0.798	0.767	0.541	0.871	5.314	0.802
BLIP Base	BENIGN	**0.971**	**0.962**	**0.952**	0.942	0.688	**0.975**	8.339	**0.952**
	MALIGNANT	0.835	0.777	0.719	0.672	0.467	0.828	3.010	**0.717**
	BALANCED	0.903	0.870	0.836	0.807	0.578	0.902	5.675	**0.835**
BLIP Large	BENIGN	0.933	0.919	0.905	0.890	0.635	0.939	7.529	0.899
	MALIGNANT	0.831	0.766	0.701	0.653	0.457	0.812	2.939	0.695
	BALANCED	0.882	0.843	0.803	0.772	0.546	0.876	5.234	0.797
BLIP2 2.7B	BENIGN	0.405	0.393	0.380	0.366	0.517	0.642	0.856	0.909
	MALIGNANT	0.844	0.783	0.721	0.671	0.469	0.833	2.879	**0.739**
	BALANCED	0.625	0.588	0.551	0.519	0.493	0.738	1.868	0.824
BLIP2 6.7B	BENIGN	0.962	0.955	0.949	**0.944**	**0.697**	0.961	**8.994**	0.942
	MALIGNANT	0.824	0.753	0.683	0.631	0.448	0.807	2.791	0.688
	BALANCED	0.893	0.854	0.816	0.788	0.573	0.884	5.893	0.815

The results of the independent evaluation of benign and malignant cells and the average of both are labeled as Balanced.

<Benign cells>

	Gold Standard	The background is relatively clean. There are small clusters of columnar epithelial cells.
	Proposed	the background is relatively clean there are small clusters of columnar epithelial cells
	Single decoder	the background is relatively clean there are small clusters of columnar epithelial cells
	BLIP Base	the background is relatively clean there are small clusters of columnar epithelial cells

	Gold Standard	The background is relatively clean. There are small clusters of columnar epithelial cells.
	Proposed	the background is relatively clean there are small clusters of columnar epithelial cells
	Single decoder	there are adenocarcinoma cells with hyperchromatic nucleus nucleus with irregular shape pale cytoplasm nuclear enlargement
	BLIP Base	the background is relatively clean. there are small clusters of columnar epithelial cells.

	Gold Standard	The background is relatively clean. There are small clusters of columnar epithelial cells.
	Proposed	the background is relatively clean there are small clusters of columnar epithelial cells
	Single decoder	there are adenocarcinoma cells with hyperchromatic nucleus nucleus with irregular shape pale cytoplasm nuclear enlargement
	BLIP Base	the background is relatively clean. there are small clusters of columnar epithelial cells.

<Malignant cells>

	Gold Standard	There are adenocarcinoma cells with hyperchromatic nucleus, nucleus with irregular shape, pale cytoplasm, promient nucleoli, filled with foamy mucus.
	Proposed	there are squamous cell carcinoma cells with thick cytoplasm hyperchromatic nucleus nucleus with irregular shape
	Single decoder	there are squamous cell carcinoma cells with thick cytoplasm hyperchromatic nucleus nucleus with irregular shape
	BLIP Base	there are squamous cell carcinoma cells with thick cytoplasm, hyperchromatic nucleus, nucleus with irregular shape.

	Gold Standard	There are squamous cell carcinoma cells with thick cytoplasm, dark nucleus, nucleus with irregular shape.
	Proposed	there are adenocarcinoma cells with hyperchromatic nucleus nucleus with irregular shape pale cytoplasm nuclear enlargement
	Single decoder	there are adenocarcinoma cells with hyperchromatic nucleus nucleus with irregular shape pale cytoplasm nuclear enlargement
	BLIP Base	the background is relatively clean. there are clusters of columnar epithelial cells.

Fig. 6. Images with saliency maps and corresponding generated reports with gold standard. The saliency map was obtained from ResNet50 in the proposed method. The red font in the report output indicates an incorrect part.

Finally, Fig. 6 shows the proposed method output, including the input images, saliency maps using Grad-CAM, image classification results, and report output. For the output text, in addition to the proposed method, the results of the single-text decoder from the ablation study and the BLIP Base, which exhibited the best performance among the existing captioning models, are also shown.

IV. DISCUSSION

In this study, we developed an image-classification and report-generation model for lung cytological images using feature extraction by a CNN and a text decoder using a Transformer. The fine-tuned CNN classifier exhibited acceptable performance with a 95.8% correct classification rate for benign and malignant lung cells. Additionally, the saliency map extracted by Grad-CAM also exhibited high values in typical benign and malignant cell areas. These results confirm that feature extraction was properly performed.

Based on the CNN classification results, two text decoders were used to generate reports on benign and malignant cells. The text decoders consist of multiple Transformer layers with causal self-attention and cross-attention, and each attention layer has a multi-head attention structure. By optimizing the text decoder network structure by changing the number of layers and heads, it was found that the optimal decoder model that generated reports for benign cells was one with one layer and two heads, while for the text decoder for malignant cells, it

was one with two layers and four heads. For benign cells, there were only two cell types, columnar and squamous epithelial cells; therefore, a simple network was sufficient. By contrast, malignant cells with a wide variety of cell types require large networks. The optimal network was selected based on the complexity of the target, yielding reasonable results.

Regarding the report output accuracy, the proposed method correctly provided a malignant description for all malignant cells, and incorrectly provided malignant cells for only one case of benign cells. In contrast, the single-text decoder model evaluated in an ablation study and general captioning methods incorrectly output benign and malignant descriptions for some of the evaluated images. This difference in performance indicates that the proposed method has a CNN that classifies images into benign and malignant in the first stage and has a high classification accuracy. The quantitative text evaluation metrics also show that the proposed method outperforms the SOTA LLM-based captioning models, indicating the effectiveness of the CNN for image classification and feature extraction.

Despite its many advantages, there is room for improvement in the proposed method in terms of the histological lung cancer cell description in the generative text for malignant cell images for any method. Regarding malignant cells in the lungs, adenocarcinoma and squamous cell carcinoma cells were included; however, it is often difficult, even for pathologists, to distinguish between the two in Papanicolaou-stained specimens. For this reason, the two are often treated together as "non-small cell carcinoma." The report outputs of all models evaluated in this study showed that many of them did not correctly discriminate between adenocarcinoma and squamous cell carcinoma. To improve the malignant cell classification, it is necessary to train the CNN not only to classify benign and malignant cells but also to classify the cell tissue type to evaluate whether the report-generation performance can be improved. In addition, although small-cell carcinoma, which is epidemiologically less numerous, was excluded from this study, we will add small-cell carcinoma as a tissue-type classification category.

V. Conclusion

In this paper, we propose an image-classification and report-generation model for lung cytological images. The proposed method, which has a CNN that combines image classification and feature extraction and two Transformer-based text encoders, outperforms existing image capturing models and single-text-encoder models in terms of image identification and report-generation quality. These results indicate that the proposed method may be useful for generating cytological image reports.

References

[1] American Cancer Society, "Cancer facts and figures 2023". Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf.

[2] L. Zhang *et al.*, "DeepPap: Deep convolutional networks for cervical cell classification," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1633–1643, 2017, doi: 10.1109/JBHI.2017.2705583.

[3] G. Dimauro *et al.*, "Nasal cytology with deep learning techniques," *Int. J. Med. Inform.*, vol. 122, pp. 13–19, 2019, doi: 10.1016/j.ijmedinf.2018.11.010.

[4] A. Bal, *et al.*, "BFCNet: a CNN for diagnosis of ductal carcinoma in breast from cytology images," Pattern Anal. Appl., vol. 24, no. 3, pp. 967–980, 2021, doi: 10.1007/s10044-021-00962-4.

[5] A. Teramoto *et al.*, "Automated classification of lung cancer types from cytological images using deep convolutional neural networks," *BioMed Res. Int.*, vol. 2017, pp. 4067832, 2017, doi: 10.1155/2017/4067832.

[6] T. Tsukamoto *et al.*, "Comparison of fine-tuned deep convolutional neural networks for the automated classification of lung cancer cytology images with integration of additional classifiers," *Asian Pac. J. Cancer Prev.*, vol. 23, no. 4, pp. 1315–1324, 2022, doi: 10.31557/APJCP.2022.23.4.1315.

[7] A. Teramoto *et al.*, "Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network," *Inform. Med. Unlocked*, vol. 16, p. 100205, 2019, doi: 10.1016/j.imu.2019.100205.

[8] A. Teramoto *et al.*, "Deep learning approach to classification of lung cytological images: Two-step training using actual and synthesized images by progressive growing of generative adversarial networks," *PLOS ONE*, vol. 15, no. 3, p. e0229951, 2020, doi: 10.1371/journal.pone.0229951.

[9] X. Wang *et al.*, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays" in, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, Jun. 2018, 2018, pp. 9049–9058, doi: 10.1109/CVPR.2018.00943.

[10] B. Hou *et al.*, "RATCHET: Medical Transformer for chest X-ray diagnosis and reporting" in *Med. Image Comput. Comput. Assist. Interv. MICCAI*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng and C. Essert, Eds. Cham: Springer International Publishing, vol. 2021, pp. 293–303, 2021.

[11] Y. F. Zhou *et al.*, "GNNFormer: A Graph-based Framework for Cytopathology Report Generation," arXiv, Available at: arXiv:2303.09956, 2023.

[12] F. C. Schmitt, et al., "The World Health Organization reporting system for lung cytopathology," *Acta Cytol.*, vol. 67, no. 1, pp. 80–91, 2023, doi: 10.1159/000527580.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition" in International Conference on Learning Representations, vol. 2015, 2015.

[14] C. Szegedy, et al., "Going deeper with convolutions" in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2015, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[15] K. He *et al.*, "Deep residual learning for image recognition" in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2016, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[16] G. Huang *et al.*, "Densely connected convolutional networks" in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2017, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[17] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization" in IEEE International Conference on Computer Vision (ICCV), vol. 2017, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[18] R. Girshick *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation" in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2015, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[19] K. Papineni *et al.*, "Bleu: A method for automatic evaluation of machine translation" in Annual Meeting of the Association for Computational Linguistics, pp. 1106–1114, 2001, doi: 10.3115/1073083.1073135.

[20] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language" in EACL Workshop on Statistical Machine Translation, 2014, pp. 376–380, doi: 10.3115/v1/W14-3348.

[21] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries" in *Text Summarization Branches Out*, 2004, pp. 74–81.

[22] R. Vedantam *et al.*, "Cider: Consensus-based image description evaluation" in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.

[23] P. Anderson *et al.*, "Spice: Semantic propositional image caption evaluation" in European Conference on Computer Vision, 2016, pp. 382–398.

[24] J. Wang *et al.*, *"GIT: A Generative Image-to-text Transformer for Vision and Language,"* arXiv, Available at: arXiv:2205.14100v5, 2022.

[25] J. Li *et al.*, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *Proc. 39th International Conference on Machine Learning, PMLR*, vol. 162, pp. 12888-12900, 2022.

[26] J. Li *et al.*, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv., Available at: arXiv:2301.12597v3, 2023.

[27] S. Zhang, et al., *"OPT: Open Pre-trained Transformer Language Models," arXiv*, Available at: arXiv:2205.01068v4, 2022.