Towards Interpretable Attention Networks for Cervical Cancer Analysis

Ruiqi Wang[1], Mohammad Ali Armin[1], Simon Denman[2], Lars Petersson [1], David Ahmedt-Aristizabal[1,2]

Abstract— Recent advances in deep learning have enabled the development of automated frameworks for analysing medical images and signals, including analysis of cervical cancer. Many previous works focus on the analysis of isolated cervical cells, or do not offer explainable methods to explore and understand how the proposed models reach their classification decisions on multi-cell images which contain multiple cells. Here, we evaluate various state-of-the-art deep learning models and attention-based frameworks to classify multiple cervical cells. Our aim is to provide interpretable deep learning models by comparing their explainability through the gradients visualization. We demonstrate the importance of using images that contain multiple cells over using isolated single-cell images. We show the effectiveness of the residual channel attention model for extracting important features from a group of cells, and demonstrate this model's efficiency for multiple cervical cells classification. This work highlights the benefits of attention networks to exploit relations and distributions within multi-cell images for cervical cancer analysis. Such an approach can assist clinicians in understanding a model's prediction by providing interpretable results.

I. INTRODUCTION

Cervical cancer is a serious health problem and it is one of the most common types of cancer in women worldwide [1]. With the development of promising computer vision techniques, more and more practical and efficient image analysis models exist to provide reliable auxiliary diagnosis results based on cell images. In cervical cell image classification tasks, the input can be an image showing a single isolated cell, or an image showing multiple cells. A model must classify the type of cell in the image (*e.g* koilocytotic, metaplastic).

Various works have performed cervical cell classification using neural networks. Plissiti et al. [2] applied VGG-16 [3], a deep convolutional neural network, to classify isolated cervical cell images. Talo et al. [4] proposed a DenseNet-161 [5] model which improved upon the results of [2]. To further improve performance, Haryanto et al. [6] introduced a padding scheme to AlexNet [7]. Apart from deep neural networks, Win et al. [8] combined various traditional machine learning methods such as random forests, support vector machines and k-nearest neighbors for segmentation and classification. GV et al. [9] proposed a segmentation-free PCA based approach combined with a deep convolutional neural network to achieve the state-of-the-art result. However, these promising results are based on isolated cell images, and thus consider only a single isolated cell. Un-cropped cervical cell images contain multiple cells in different regions, and we refer to these as multi-cell images. Focusing only on a single cell discards vital information in the multi-cell image. For example, different types of cells in a multi-cell image have different distributions, and the relation between them varies. In addition, although these deep learning models achieve high performance in classification of isolated cells, they do not provide explainability and interpretabilty information for their model, which makes it difficult to understand the rationale behind their decisions. To enable adoption of these methods, we need improved explanations and interpretabilty to build user confidence and acceptance.

In this paper, we aim to develop interpretable deep learning models for the classification of multi-cell cervical cell images. Particularly, our work focuses on exploring the feasibility of adapting attention-based frameworks. Several prominent explainability methods for CNN-based models have been introduced including class activation mapping [10] and guided backpropagation [11]. While there is much interesting research within this field, it is immature and there are only a few works that investigate explanation methods for cervical cancer classification [12]. We verify our models using an attribution prediction technique and compare the interpretable learning results offered by traditional CNN models and the proposed attention-based model.

Our main contributions are summarized as follows:
1) We compare and introduce multiple deep learning models including a residual network, dense network, classic residual attention model and residual channel attention mechanisms to classify multi-cell cervical cell images.
2) We demonstrate the effectiveness of interpretable attention networks for cervical cancer analysis.

II. METHODOLOGY

In this paper, we analyze different deep learning models for cervical cell classification and compare their performance. We introduce attention-based frameworks to extract important features from multi-cell cervical cell images. We aim to demonstrate the effect of the attention mechanism on the classification of multi-cell cervical images, which contain blank backgrounds with multiple cells of the same type. We also provide reliable explanations to demonstrate how the attention mechanisms work, and show its interpretability through gradient visualization.

A. Deep convolutional neural networks

With the development of deep learning, various proposed deep convolutional neural networks have achieved success

[1] CSIRO, DATA61, Canberra, Australia. Corresponding author: david.ahmedtaristizabal@data61.csiro.au
[2] Image and Video Research Laboratory, SAIVT, Queensland University of Technology, Brisbane, Australia.

in classification, detection and segmentation tasks. We select the following two prominent deep convolutional neural networks, which recently have shown to be successful for cervical image classification [4], [13].

1) Residual Convolutional Networks (ResNet): Recent studies have shown the high performance of ResNet [14], which leverages the residual block structure, for image recognition and classification tasks. It uses residual (or skip) connections to allow information to more readily propagate through the network. ResNet models have many variants with different number of layers and different residual block structures. The baseline model in our experiments is based on ResNet-50, which has 48 Convolutional layers, along with 1 MaxPool and 1 Average Pool layer.

2) Dense Convolutional Networks (DenseNet): DenseNet [5] simplifies the connectivity pattern between layers in a deep neural network, which significantly reduces the number of parameters and prevents learning from redundant feature maps. The structure of DenseNet is different from a traditional deep neural network, which concatenates the output feature maps of each layer with incoming feature maps, instead of summing them together. Similar to how ResNet is divided into residual blocks, DenseNet is divided into Dense Blocks with the same dimension as the feature maps, but with a different number of filters. Therefore, each layer in DenseNet has direct access to its preceding feature maps, and collect the new information it learned from the input. We chose one variant from this light and effective model, DenseNet101, for use in our experiments.

B. Attention-based frameworks

Recently, attention mechanisms have shown an ability to achieve substantial improvements when added to deep learning models for various applications. Attention mechanisms helps a network to focus on important features in the data, and lead to more accurate decisions. Hence, we introduce two different attention mechanisms in our model for further experiments.

1) Residual attention networks: The residual attention network [15] is a convolutional neural network based on the attention mechanism of [16]. A naive attention learning creates a soft mask on the input to generate attention-aware features. However, it may lead to a performance drop as its dot product operation with a mask will degrade the value of features in deep layers. Residual attention networks solve this problem by using a trunk branch, a pre-activation ResNet unit [17] for feature extraction; and a mask branch, which uses a bottom-up top-down structure for learning the mask. In the residual attention network, it adds the attention module after each residual layer, making features clearer as depth increases. We use this residual attention network based on the ResNet-50 framework as our first attention-based model.

2) Residual channel attention: For image super-resolution, it is necessary to avoid learning abundant low-frequency information from low-resolution inputs and features. These unnecessary features are treated equally across channels. The residual channel attention network

(RCAN) [18] is made up of residual-in-residual (RIR) blocks with channel attention, which ignores abundant low-frequency information through skip connections within blocks, and adaptively rescales features between channels. Channel attention combines channel-wise features with different weights by exploiting the interdependencies among them. Therefore, for highly accurate image super-resolution, it is more flexible and more powerful for feature extraction between channels. We aim to use this RCAN model to verify the effectiveness of attention mechanisms for multi-cell images classification, and compare it with the residual attention networks (see Section II-B.1).

C. Explainability and interpretability

Lack of transparency is identified as one of the main barriers for AI adoption in clinical practice. A step towards making AI trustworthy is the development of explainable AI methods. In this analysis, we adopted the integrated gradient model [19] to illustrate the advantages of these methods for helping clinicians to identify the parts of the images that are critical in the decision.

The *integrated gradient model* computes the attribution of the prediction of the deep neural network by using the gradient operation. Different to previous attribution techniques, it is characterized by two axioms: sensitivity and implementation invariance; which makes it more flexible and easier to apply to a variety of deep networks.

III. EVALUATION

A. Datasets and experimental setup

The performance of each proposed method is evaluated on the SIPaKMeD dataset [2], which is an open-source cervical cell image database. It consists of two different types of data: 966 multi-cell images and 4049 isolated cell images. There are five different categories of cervical cells in this dataset: superficial-intermediate, parabasal, koilocytotic, dysketarotic and metaplastic. Sample images from each class are shown in Fig. 1. Class distribution details are presented in Table I. These cell images were acquired using a high resolution CCD camera connected to an optical microscope. From the data distribution in Table I it can be noted that the classification task for multi-cell images is more challenging as, apart from the imbalanced class distribution, the volume of data is smaller than for the isolated cell images. In our experiments, we focus on multi-cell images only, as isolated cell images contains features of a single cell only and thus do not contain information on cell distributions and relationships between cells. We also aim to demonstrate that some of these isolated cell images may suffer from the lack of cell information. To

TABLE I: Data distribution of the SIPaKMeD dataset

Category	Multi-cell Images	Isolated Images
Dyskeratotic	223	813
Koilocytotic	238	825
Metaplastic	271	793
Parabasal	108	787
Superficial	126	813

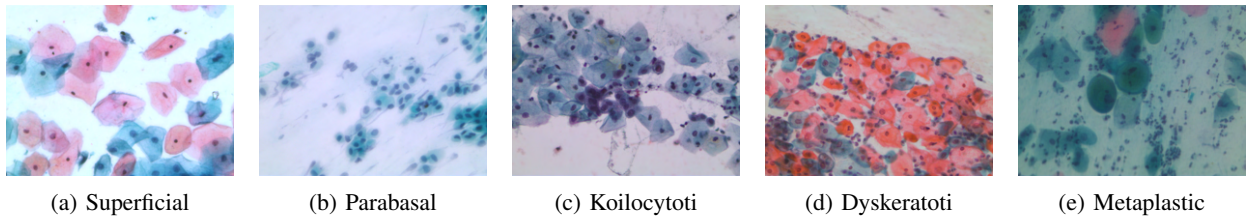| (a) Superficial | (b) Parabasal | (c) Koilocytoti | (d) Dyskeratoti | (e) Metaplastic |

Fig. 1: Multi-cell images of the five categories considered.

verify the capability and effectiveness of each model, we also applied them to the isolated cell images. Both types of data (multi-cell and single-cell images) are split such that 70% of the data samples are allocated for training, 20% for validation, and 10% for testing.

B. Evaluation metric and implementation

As this is a multi-class classification task with an uneven class distribution, we use a weighted-F1 score to measure performance. This is calculated as follows,

$$\text{Weighted } F_1 = \sum_{i=1}^{5} \frac{2 \times \text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i} \times w_i, \quad (1)$$

where w_i is the weight of the $i-th$ class and depends on the number of positive examples in that class.

The overall accuracy on the test set is also an evaluation metric, which simply shows the overall performance of different models. Categorical cross-entropy loss and the Adam optimizer [20] (learning rate=0.001, other parameters are default in Pytorch) are used to train the models. Models are trained for 50 epochs with a mini-batch size of 16, as we found more epochs lead to overfitting. All models were implemented in Pytorch 1.6.0 [21] with same parameter setting. Results are reported based on 5-fold cross-validation.

C. Experimental results and discussion

An evaluation of all proposed models and baseline methods for each dataset are shown in Table II.

From the accuracy results on both isolated and multi-cell images, it is obvious that the DenseNet-121 model with the residual channel attention mechanism has a significant advantage compared to other methods commonly used for medical image classification [4]. The original DenseNet-121 model achieves better results than the ResNet-50, as it is deeper and able to extract more hidden features. The attention-based model of DenseNet-121 also has the highest accuracy on the test set.

However, it is worth noting that the residual attention model based on ResNet-50 decreases the accuracy of ResNet-50. To explore its prediction results in more details, we compute the F1-score for each model on each class, shown in the lower section of Table II. The baseline ResNet-50 model has difficulty classifying koilocytotic and metaplastic cells correctly since both are in large size and some koilocytotic cells are a type of metaplastic cell; although there are still slight differences between them in their color, contour, size and shape. Therefore, the introduction of attention in ResNet-50 targets these differences between

TABLE II: Results

	ResNet-50	DenseNet-121 [4]	RAN-ResNet-50	RCAN-DenseNet-121
Accuracy	95.11%	95.84%	94.13%	**96.33%**

(a) Overall accuracy on isolated cervical cell images

	ResNet-50	DenseNet-121 [4]	RAN-ResNet-50	RCAN-DenseNet-121
Accuracy	85.15%	89.11%	84.16%	**91.09%**

(b) Overall accuracy on multi-cell cervical cell images

|	ResNet-50	DenseNet-121	RAN-ResNet-50	RCAN-DenseNet-121
Dyskeratotic	0.869	0.898	0.851	**0.978**
Koilocytotic	0.744	0.809	0.783	**0.869**
Metaplastic	0.839	**0.929**	0.877	0.896
Parabasal	0.96	0.957	0.88	**1.0**
Superficial	0.846	**0.889**	0.815	**0.889**

(c) F1 score of each class on multi-cell cervical cell images

these two classes and improves the precision. However, as in the residual attention model, the attention layer is added after each residual layer and it learns similar noise for other classes. Therefore, although it improves upon the weakness of the baseline model, it decreases the precision of other well-performing classes and thus lowers the overall accuracy.

The performance of residual channel attention DenseNet-121 is outstanding in four of the five classes. Comparing with the original DenseNet-121 model, the introduction of residual channel attention significantly improves the precision on the dyskeratotic class and koilocytotic class with a slight loss in precision for metaplastic cells. This result also gives strong proof of the effectiveness of the channel attention mechanism in this cervical cell classification task.

In order to explore more details of the attention learning progress, we visualize the gradient attribution prediction of DenseNet-121 and the residual channel attention based model in Fig. 2.

From the gradient visualization results of a random test image from the dyskeratotic category, we can observe that DenseNet-121 considers features from the input image from a large area of cells, including a part of the background. Conversely, the model with residual channel attention is more focused on a small specific region of the multi-cell image, which means it makes the decision from a few cells only, ignoring the background and other noisy information. From this interpretable visualization result of the deep neural network, we also understand how the attention mechanism works in classifying different cervical cells. It gives more weight in parts of the cell groups for classification, focusing on parts which contains useful features and relations. This visualization results also provide information about the specific regions in the multi-cell images, which helps to assist
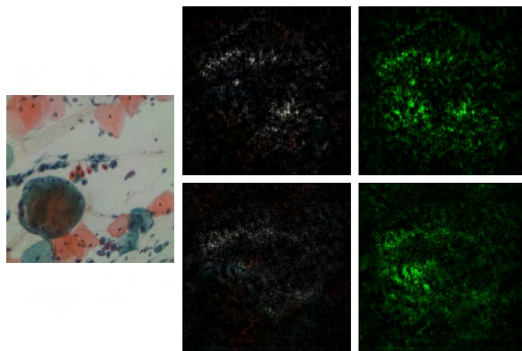
(a) DenseNet-121 without attention
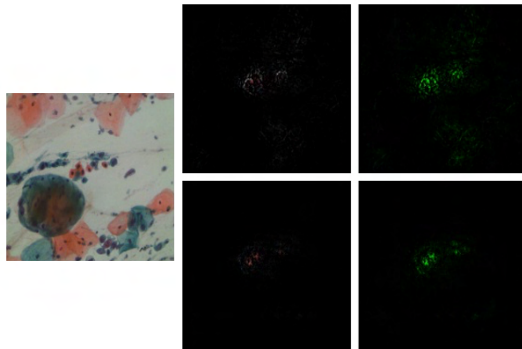
(b) DenseNet-121 with residual channel attention

Fig. 2: Attribution Prediction Gradient Visualization. Columns from left to right: the original image, gradient and integrated gradient overlay, gradient and integrated gradient.

the identification of cervical cells of particular interest for experts in a real world application.

Our current evaluation results show that the residual channel attention mechanism is efficient for analyzing the multi-cell cervical cell images and classifies them more precisely. It also offers information regarding hidden relations between cell groups in the multi-cell image, and it is worth noting that the attention often falls on a specific group of cells. There may also be other factors such as the distribution of different classes of cervical cells which can be analyzed. CNNs have been commonly used in the digital pathology domain for the classification of fixed-sized biopsy image patches. However, learning over patch-wise features limits the model to capturing global contextual information. Recently, graph data representations have attracted significant attention in the analysis of histological images [22] due to their ability to represent the tissue architecture by modeling a tissue section as a multi-attributed spatial graph of its constituent cells. Graph-based representations can encode the spatial relationships across the patches for fine-grained classification [22]. In future work, relations between cells in multi-cell images can be analyzed through this technique.

IV. CONCLUSIONS

We introduce and compare deep convolutional neural networks with different attention mechanisms for cervical cell classification. We also provide a detailed explanation on how interpretation methods can be applied to interpret classification results for cervical cell images. Our experiments and analysis show that the residual channel attention framework is effective in distinguishing between features for different classes and isolating a specific region of interest for multi-cell cervical cell images.

REFERENCES

[1] T. P. Canavan and N. R. Doshi, "Cervical cancer," *American family physician*, vol. 61, no. 5, pp. 1369–1376, 2000.

[2] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images," in *ICIP*, 2018, pp. 3144–3148.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[4] M. Talo, "Diagnostic classification of cervical cell images from pap smear slides," *Academic Perspective Procedia*, vol. 2, no. 3, pp. 1043–1050, 2019.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[6] T. Haryanto, I. S. Sitanggang, M. A. Agmalaro, and R. Rulaningtyas, "The utilization of padding scheme on convolutional neural network for cervical cell images classification," in *CENIM*, 2020, pp. 34–38.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[8] K. P. Win, Y. Kitjaidure, K. Hamamoto, and T. Myo Aung, "Computer-assisted screening for cervical cancer using digital image processing of pap smear images," *Applied Sciences*, vol. 10, no. 5, p. 1800, 2020.

[9] K. K. GV and G. M. Reddy, "Automatic classification of whole slide pap smear images using cnn with pca based feature interpretation." in *CVPR Workshops*, 2019, pp. 1074–1079.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[12] T. Conceição, C. Braga, L. Rosado, and M. J. M. Vasconcelos, "A review of computational methods for cervical cells segmentation and abnormality classification," *International journal of molecular sciences*, vol. 20, no. 20, p. 5114, 2019.

[13] E. Hussain, L. B. Mahanta, C. R. Das, and R. K. Talukdar, "A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network," *Tissue and Cell*, vol. 65, p. 101347, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[15] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017, pp. 3156–3164.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016, pp. 630–645.

[18] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018, pp. 286–301.

[19] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017, pp. 3319–3328.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8026–8037.

[22] B. Aygüneş, S. Aksoy, R. G. Cinbiş, K. Kösemehmetoğlu, S. Önder, and A. Üner, "Graph convolutional networks for region of interest classification in breast histopathology," in *Med. Imaging 2020: Digit. Pathol.*, vol. 11320, 2020, p. 113200K.