An Explainable Computer Vision in Histopathology: Techniques for Interpreting Black Box Model

Subrata Bhattacharjee
Dept. of Computer Engineering
Inje University
Gimhae, Republic of Korea
subrata_bhattacharjee@outlook.com

Yeong-Byn Hwang
Dept. of Digital Anti-Aging Healthcare
Inje University
Gimhae, Republic of Korea
hyb1345679@gmail.com

Kobiljon Ikromjanov
Dept. of Digital Anti-Aging Healthcare
Inje University
Gimhae, Republic of Korea
kobiljonikromjanov@gmail.com

Rashadul Islam Sumon
Dept. of Digital Anti-Aging Healthcare
Inje University
Gimhae, Republic of Korea
Sumon39.cst@gmail.com

Hee-Cheol Kim
Dept. of Digital Anti-Aging Healthcare
Inje University
Gimhae, Republic of Korea
heeki@inje.ac.kr

Heung-Kook Choi
Department of Computer Engineering
Inje University
Gimhae, Republic of Korea
cschk@inje.ac.kr

***Abstract*—Computer vision is a field of artificial intelligence (AI) that is being used increasingly in histopathology to identify pathologies in slide images with a high degree of accuracy. In this paper, we focus on the different interpreting techniques of explainable computer vision (XCV). Analysis of histopathology images is a challenging task, and specialized knowledge is mandatory to make AI decisions. To carry out this analysis, a deep learning model has been used to classify and differentiate the scoring (i.e., benign and malignant) of Prostate cancer (PCa). However, the AI models are complex and opaque, and it is important to understand model decision-making. Therefore, to address this problem, we present three techniques for accountability and transparency of the model, namely Activation Layer Visualization (ALV), Local Interpretable Model-Agnostic Explanation (LIME), SHapley Additive exPlanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM). XCV is AI in which the results of the black-box model can be understood by humans. The robustness of our model has been confirmed by using an external test dataset including 100 histopathology images. The model performance has been evaluated using the receiver operating characteristic (ROC) curve.**

Keywords— explainable computer vision, histopathology, artificial intelligence, black box, prostate cancer

I. INTRODUCTION

The analysis of histopathology images is a gold standard for the detection of different types of cancer regions and performs diagnosis using AI algorithms [1, 2]. Histopathology study is carried out under the microscope for disease diagnosis. Hematoxylin and Eosin (H&E) staining is used routinely in histopathology laboratories to analyze different types of cells and tissue and provides important information about the pattern, shape, cell structure in a tissue sample [3, 4]. Also, H&E dyes make it easier for pathologists to see different parts of the cell under a microscope Hematoxylin has a deep blue-purple color which shows the ribosomes, chromatin within the nucleus. In contrast, Eosin has an orange-pink-red color which shows the cytoplasm, cell wall, collagen, connective tissue, and other structures that surround and support the cell. Image classification and segmentation are two basic tasks in digital histopathology. Image classification is carried out by categorizing and labeling groups of pixels within an image [5]. In this study, the image classification task was carried out using PCa tissue samples, and it is a type of cancer that has always been an important challenge for pathologists. For manual diagnosis of PCa, expert pathologists need more attention to analyze the tissue pattern, structure of cells, and glands under a microscope, which is time-consuming. However, to make the work easier for pathologists, many researchers are developing different types of computer-aided diagnosis (CAD) systems that can make decisions automatically.

In recent years, AI algorithms have shown tremendous performance in different kinds of applications, especially in medical health. It has been used in many fields as exemplified by computer vision and is well-recognized for image classification [6]. Recently, the activation features of convolution neural networks (CNN) have achieved splendid triumphs in computer vision [7-9]. XCV is AI in which the results of the black-box model can be understood by humans. Nowadays, AI systems and machine learning (ML) algorithms are widespread in many areas. Data is used almost everywhere to solve problems and help humans, a large factor for this success is the progress in the DL area, but also generally the development of new and creative ways how we can use data.

As a consequence, the complexity of these systems becomes incomprehensible even for AI experts. Therefore, the models are usually also referred to as black boxes. The meaning of "black-box" is that it is generally difficult to clearly explain the decisions made by the models [10]. Explainability and interpretability are very important in medical areas because the CAD system needs to be transparent and understandable to gain the trust of doctors and patients. The procedures and methods that allow human users to understand and trust the results and output created by the models are called XCV [11]. Fig. 1 shows a schematic representation of XCV.

In this paper, we introduce a Light-Dense CNN (LDCNN) model for histopathology image classification. The model consists of multiple layers, including the input layer, convolutional layers, concatenation layers, dropout layers, and classification layer. This model has been modified from the light-weight CNN (LWCNN) architecture which was introduced in our previous study [12]. Also, we have explained the processes and outputs of the supervised model so that it is understandable for other readers. The ALV, LIME [13], SHAP [14, 15], and Grad-CAM [16] techniques were used to generate the activated feature maps and visualize the model's decisions which produce a coarse localization map of the important region in the image, thus interpreting the decision of the neural network.

978-1-6654-5818-4/22/$31.00 ©2022 IEEE

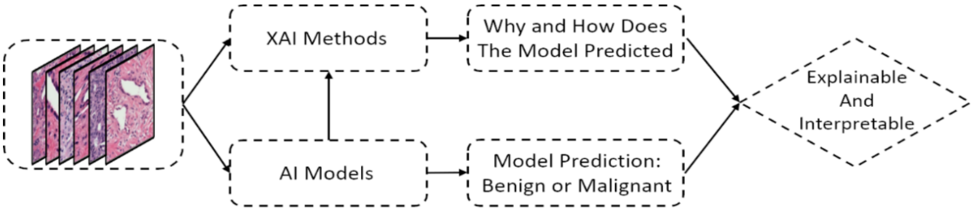
XAI Methods
Why and How Does The Model Predicted
Explainable And Interpretable
AI Models
Model Prediction: Benign or Malignant

Fig. 1. A brief schematic representation of explainable computer vision

The contributions of this paper are summarized as follows:

- The binary classification was performed successfully using the LDCNN model.
- Experiments were conducted using PCa histopathology images with different magnifying factors (i.e., 20× and 40×).
- Two types of datasets were used to perform the experiments: a public dataset (i.e., PANDA Challenge) and a private dataset.
- The XCV methods are used to visualize the results of the black-box model, which include ALV, SHAP, and Grad-CAM.

The remainder of this article is structured as follows. In Section II, we described the related work and recent studies about images classification and XAI. Section III illustrates the complete methodology of this study, which includes data acquisition, model development, and XCV methods. Results of the AI models are presented in Section IV. Section V discusses the paper and lastly, the paper is concluded in Section VI.

II. Related Work

Research on computer vision for histopathology image analysis provided valuable findings regarding the problems of automatic detection and classifying PCa tissue images. In [17], they developed a patch-based classifier using CNN for the automated classification of histopathology images. Their proposed method achieved promising results for both binary and multiclass classification. In [18], they developed a dual-channel residual convolution neural network to classify the histopathology images of the lymph node section. They performed binary classification to discriminate between cancerous from noncancerous tumors. In [19], a novel method was proposed for histopathological image classification of colorectal cancer. They developed a novel bilinear convolution neural network (BCNN) model that consists of two CNNs, and the outputs of the CNN layers are multiplied with the outer product at each spatial domain. This proposed model of this paper performed better than the traditional CNN by classifying colorectal cancer images into eight different classes. In [20], they developed an Inception Recurrent Residual Convolution Neural Network (IRRCNN) model for the histopathology image classification of Breast Cancer. They developed their model based on three powerful DL architectures, namely Inception, Residual, and Recurrent Network. In [12], the author proposed a lightweight CNN model to classify the histopathology images of prostate cancer. The model achieved promising accuracy of 94.0% for binary classification. The comparative analysis was performed with other state-of-the-art pre-trained models. In [21], the author proposed a fully automatic method that detects prostatectomy WSIs with a high-grade Gleason score. The model achieved an accuracy of 78% in a balanced set of 46 unseen test images.

In recent years, researchers are focusing on the XAI because the decision-making process of deep neural networks is largely unclear, and it is difficult to understand for humans. In [22], the author used different methods to generate the importance map from the black-box model indicating how salient each pixel importance using gradients or other internal network states. Also, they address the problem of XAI for deep neural networks that take images as input and output a class probability. In [13], the author proposed a novel explanation technique (i.e., LIME) to explain the predictions of nay classifier. Also, they demostrated the flexibility of this method by explaining ML models (e.g. random forests) for text and DL models (e.g. neural networks) for image classification. In [14], the author present a unified framework for interpreting model prediction, SHAP. It assigns feature importance for a particular prediction. Also, they proposed new methods that show better consistency with human intuition than previous approaches. In [15], the author surveyed the current progress of XAI and in particular its advances in healthcare applications. They discussed different approaches (i.e., Grad-CAM, LIME, and SHAP) to unbox the black-box for medical explainable AI via multi-modal and multi-center data fusion. In [23], the author evaluated k-means clustering and random forest algorithms using two very popular xEplainable techniques (i.e., LIME and SHAP) to see and understand the output of the black-box model.

In previous research works, the authors developed different kind of CNN models for histopathology image classification and achieved promising results. Also, few explainable techniques were proposed for interpreting model prediction. However, in the present study we developed a LDCNN model which is a modified version of LWCNN [12], and it is not complicated like other CNN models discussed in this section. The proposed model performs better than LWCNN in terms of accuracy, overfitting issue, and computational cost.

III. Materials and Methods

A. Dataset

We have used two different datasets from two different centers. Out of which, one is public and the other one is private.

Public Dataset: It was collected online from the Kaggle repository [24]. The whole slide images (WSIs) were prepared at Radboud University Medical Center, Netherland. The slides were scanned using 3Dhistech Pannoramic Flash II 250 scanner at 20× magnification. Sample patch images used for model testing are shown in Fig. 2 The size of the patch images extracted from WSIs is 512 × 512 pixels.

Private Dataset: This dataset is not publicly available online. It was collected from the Severance Hospital of Yonsei

University, South Korea. The slides were scanned at 40× optical magnification with 0.3 NA objective using a digital camera (Olympus C-3000) attached to a microscope (Olympus BX-51). The sizes of patches extracted from WSIs are 256 × 256 and 512 × 512 pixels. Fig. 3 shows the sample images of PCa used for model training and validation.

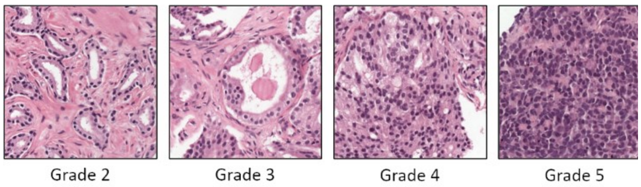

Fig. 2. The four types of PCa histopathology images from the PANDA challenge dataset. Grade 2 is considered a benign tumor and Grade 3, Grade 4, and Grade 5 is considered a malignant tumor

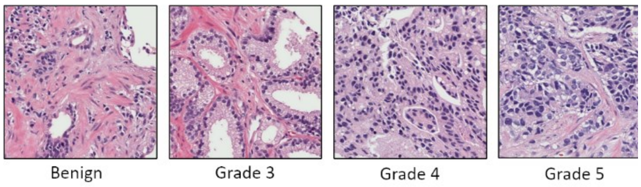

Fig. 3. The four types of histopathology images of PCa (benign, grade 3, grade 4, and grade 5) from a private dataset

Image resizing is a crucial step in computer vision. DL models train faster on smaller images and require the same input image dimensions (i.e., height × width) for all the input samples. Also, the model produces an error if the image is too small or too big. So according to the rule of thumb method, we decided to use 256 × 256 pixel images for the DL model.

Moreover, we performed data augmentation to increase the number of samples in the dataset because the huge data increases the likelihood that it contains useful information, which is advantageous for the DL model. Also, by adding more data in the training set the chances of overfitting decrease rather than increase. Therefore, we generated 2 samples from each input sample with rotation (i.e., 10° and 180°) augmentation technique. Table I shows the statistics for the PCa classification datasets.

TABLE I. STATISTICS FOR PRIVATE AND PUBLIC DATASET

Training and Validation	**Private Dataset**		
	Benign	***Malignant***	***Total***
Total Number of Samples	900	900	1800
Number of Training Samples Before Augmentation	810	810	1620
Number of Training Samples After Augmentation	1620	1620	3240
Number of Validation Samples	90	90	180
Testing	**Public Dataset**		
Number of Test Sample	50	50	100

B. LDCNN Model for Prostate Cancer Recognition

To perform supervised learning, we introduce an LDCNN model for PCa classification. We have used two different datasets from two different centers. Out of which, one is public and the other one is private. Although the model was not trained with a sufficient amount of data, the proposed model has shown state-of-the-art performance. The model provides better recognition performance using fewer network parameters.

To construct the model, we utilized a concatenation operation between the CNN layers to build the dense connections in the network. Here, the output feature maps of the layer are concatenated with the incoming feature maps. The model has several advantages: it strengthens feature propagation, encourages feature reuse, and substantially reduces the number of parameters. Nonetheless, the model may require high graphics processing unit (GPU) due to concatenation operation. The model included CNN layers, such as those for input, convolution, rectified linear unit (ReLU), concatenation, dropout, global average pooling (GAP), and classification. In this model, 'Stride=2' was utilized in the convolution layer instead of the 'Maxpooling (2 × 2)' to down-sample an input representation (image, hidden-layer output matrix, etc.), reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned. The entire model is shown in Fig. 4.

For classification, we set the input shape to 256 × 256 × 1 while building the model. The model contains 9 convolutional layers, 3 concatenation layers, a GAP layer, and a classification layer. Softmax activation function was utilized for the binary classification. An Adam [25] optimizer was used during training and the ReduceLROnPlateau function was used to control the learning rate (LR) of the model. To avoid model overfitting, we used the early stopping function which is a form of regularization. A total of 100 epochs were set for training the model and the learning stopped at 45 because there was no progress on the validation set for consecutive 10 epochs. All the experiments were conducted on a workstation with an NVIDIA GeForce RTX 3060 GPU, 32 GB of RAM using Tensorflow and Keras libraries.

C. Activation Layer Visualization

ALV is the technique for visualizing the feature maps by digging into neural networks. In the CNN model, activation layers are a crucial part of the design, and each layer produces a different number of feature maps that are the result of applying the filters to an input image. Therefore, visualizing the activation layer of the black-box model is important because it shows the output (i.e., the activated feature maps) of specific activation layers and this is done by looking at each specific layer.

D. Gradient Weighted Class Activation Map

Grad-CAM is another popular and effective technique for interpreting black-box models. It is a simple method compared to SHAP but we can see which regions in the image were relevant in a specific class. To visualize the attention regions in the image, a GAP layer was used instead of fully connected (FC) layers at the end before the final classification layer, and the network takes the convolutional feature maps as input through the GAP layer to produce the outputs for each class. Therefore, the class-discriminative localization map is obtained by computing the gradient of class C with respect to feature maps F of a convolutional layer. The global average pooled gradients flow back to obtain the importance weights α_k^c. The computation for Grad-CAM can be expressed as:

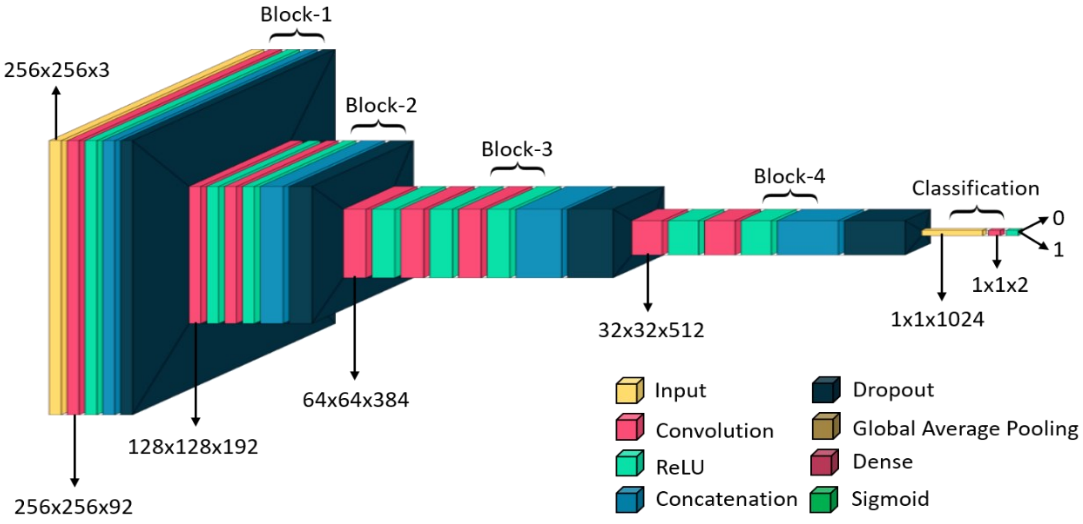

Fig. 4. Light Dense Convolutional Neural Network Architecture

$$\alpha_k^c = \frac{1}{uv}\sum_i^u \sum_j^v \frac{\partial y^c}{\partial F_{ij}^k} \quad (1)$$

$$GradCAM_c = ReLU\left(\sum_{k=1}^{k} \alpha_k^c F^k\right) \quad (2)$$

where α_k^c is the average of all gradient score of class C for the kth feature map, u and v are the length and width of the image, $\frac{1}{uv}\Sigma_i^u\Sigma_j^v$ is the global average pooling, $\frac{\partial y^c}{\partial F_{ij}^k}$ are the gradients via backpropagation, $ReLU$ is the activation function of a convolutional layer, and $GradCAM_c$ is the final attention result for the predicted class.

E. Local Interpretable Model-Agnostic Explanation

LIME is one of the novel explanation methods that explains the model predictions form each data sample in a faithful way by approximating the local interpretable models [23]. The implementation strategies of LIME are different for different data format (i.e., tabular, text data, and image data) [26].

To explain the decision of the black-box model for image data, LIME technique was used to visualize the important regions that contributed to the prediction results. In LIME algorithm, first an image segmentation method (i.e., Quickshift) is utilized to separate the original data into multiple pixel blocks. Then the pixel block is used as the original data set and perturbs it to achieve model interpretation. The equation for LIME explainer can be expressed as:

$$\xi(x) = \frac{argmin}{g\epsilon G} Loss(f,g,\pi_x) + \Omega(g) \quad (3)$$

where f is an original predictor (i.e., the CNN model), x is the original features, g is a local model, $Loss(f,g)$ signifies the local approximation degree of the target model f and the proxy model g [26], π_x is the measure of proximity of an instance y from x, $\Omega(g)$ is measure of complexity of $g\epsilon G$ [23], and $\xi(x)$ is an interpreter.

F. SHAP Gradient Explainer

SHAP is a very popular AI technique and game theory-based approach used for explaining the output of any black-box model (e.g., DL or ML). SHAP technique is used in this paper to measure feature importance and explain model decisions using expected gradients (i.e., an extension of integrated gradients). Generally, the feature attribution method is called integrated gradient which is used for deep neural networks. Therefore, the SHAP value indicates the importance of each feature in the model and how much it is contributed to the predictions for each given instance.

To explain and interpret the decisions of the black-box model, we used the SHAP algorithm to visualize the attention regions by plotting the SHAP values of every important feature for every predicted tissue sample. The equation for Shapely value estimation can be expressed as:

$$\emptyset_i(f,x) = \sum_{S\subseteq F} \frac{|S|!\,(F-|S|-1)!}{F!} \times \left[f_x(S) - f_x(S\backslash i)\right] \quad (4)$$

where $\emptyset_i$ Shapley value for feature i, f is the black-box model, x is the input dataset, $S \subseteq F$ is the feature subsets, and F is the set of all features. The SHAP value is computed based on the model prediction with the training dataset $f_x(S)$ and testing dataset $f_x(S\backslash i)$.

IV. Results and Discussion

We trained and tested the LDCNN model on high-resolution H&E stained image datasets collected from two different centers. A total of 900 images were used from private dataset and 100 from public dataset for training and testing, respectively. We performed data augmentation to avoid overfitting issues and improve the performance and outcomes of the CNN model by creating new and different samples to train the dataset. Further, to learn our CNN model, we divided private dataset into training and validation datasets according to an 9:1 ratio. Both LWCNN and LDCNN models were trained and tested on private and public datasets, respectively. From the comparative analysis (Table II), it can be observed that our proposed model achieved the best performance. In particular, LDCNN obtained a better performance by 6.0% on accuracy compared to LWCNN. Moreover, at testing phase, the public dataset was further separated into five-split for determining the generalizability of the learned model (i.e., LDCNN). Therefore, the model showed promising results and achieved an accuracy of 100%, 100%, 95.0%, 90.0%, and 85.0%, and area under the curve (AUC) of 1.00, 1.00, 1.00, 0.90, and 0.99 at test split 1, 2, 3, 4, and 5, respectively. In addition, we also provided the ROC curves to evaluate and compare the CNN models which illustrates the diagnostic ability of a binary classifier system, shown in Fig. 5. From the figure, we can observe that both the models performed well at training phase and achieved an overall AUC of 98.0%. In contrast, at testing phase, LWCNN model did not achieve better results compared to LDCNN.

TABLE II. COMPARISON RESULTS OF LWCNN AND LDCNN

Model Performance at Testing Phase	Public Dataset	
	LWCNN	*LDCNN*
Accuracy (%)	87.0	93.0
Precision (%)	96.0	92.0
Recall (%)	81.4	93.8
F1-Score (%)	88.1	92.9
AUC (%)	98.0	98.0

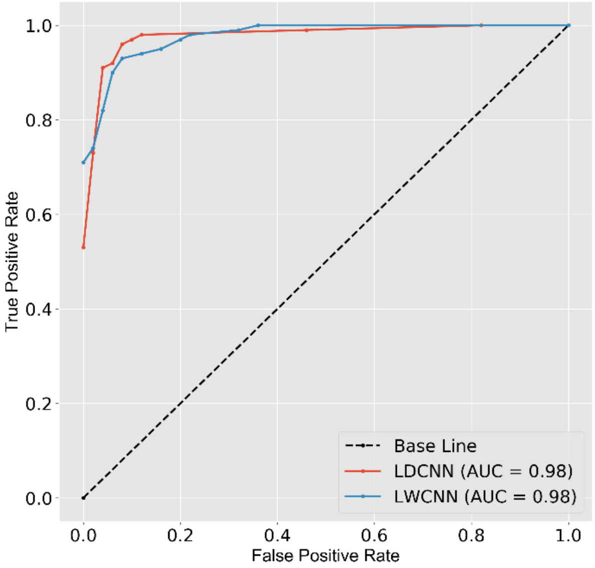

Fig. 5. ROC curve for analyzing the model performance on each test split generated by plotting the model's confidence scores

Unboxing the black-box model is very important for the medical image analysis. Many AI algorithms cannot provide any evident how and why a decision has been cast. Therefore, in this paper, we adopted few techniques for interpreting the black-box model. Fig. 6 shows the visualization results of four different activation layers extracted from our proposed CNN model. From the figure, we can observe low- and high-level feature maps obtained by CNN to identify cancer types (i.e., benign and malignant).

We also examined to interpret the decisions of the CNN black-box model using the SHAP technique (Fig. 7). This shows how each feature is significant in determining the final prediction of the model outputs. This technique could recognize the cell nuclei surrounding circular regions and highly scattered in other sections in benign and malignant tissue samples. Fig. 8 shows another popular effective method (Grad-CAM) for interpreting the CNN model. This technique is utilized to see which regions in the image are relevant to the particular class. Fig. 9 shows the model explanation via LIME to visualize the super-pixels that contributed to the benign and malignant prediction results. To visualize the interpretable results, we used the predicted benign and malignant samples, shown in Fig. 7a and b, respectively.

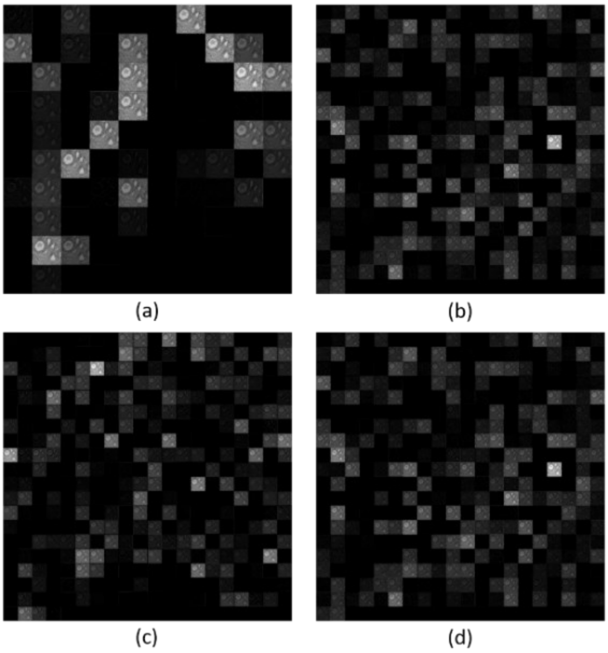

Fig. 6. Feature maps are generated from the activation layers of the CNN model. (a) Block 1 activation layer. (b) Block 2 activation layer. (c) Block 3 activation layer. (d) Block 4 activation layer. The bright pixels represents the activated and significant features

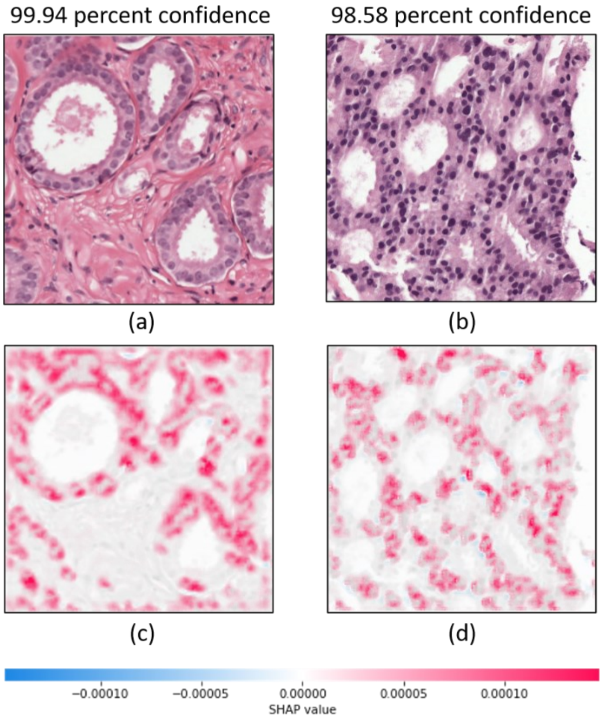

Fig. 7. Visualization of the attention regions that are positively contributed to the prediction via the SHAP method. (a) and (b) Predicted benign and

malignant tissue samples. (c) and (d) Interpretable results of (a) and (b), respectively. The color bar signifies the SHAP value. The red and blue color represents the positive and negative value that increases and decreases the model's output, respectively

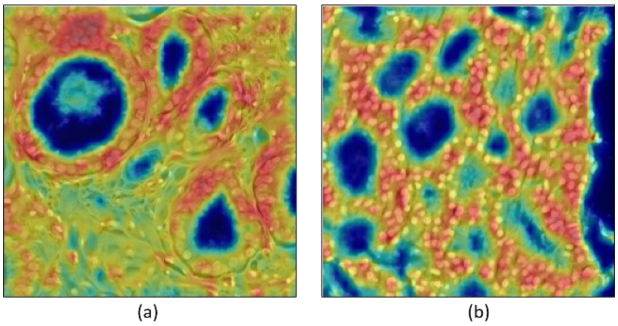

Fig. 8. Visualization of the attention regions using the Grad-CAM method. (a) Benign tissue sample. (b) Malignant tissue sample. The red color signifies the most class-specific discriminative parts of the image

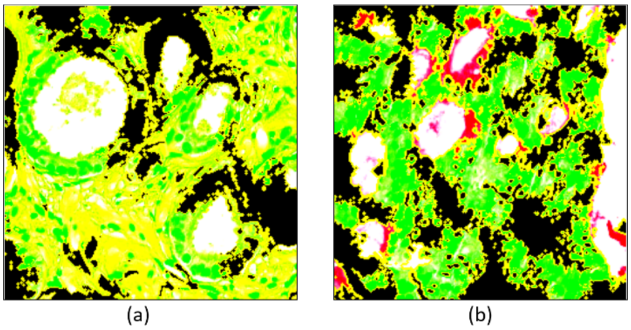

Fig. 9. Visualization of the attention resions and super-pixels that are positively contributed to the prediction via the LIME method. (a) Benign tissue sample. (b) Malignant tissue sample. The green color signifies the most class-specific discriminative parts of the image

In this study, we have first discussed the methods utilized for image classification and XCV. Binary classification (i.e., benign vs. malignant) was performed using LDCNN and LWCNN models, and comparative analysis was carried out to analyze their performance on a public dataset used for testing. Then, we adopted different explainable and interpretable techniques to visualize the attention regions and contribution of each pixels for the prediction. The main difference between the two models is that the LDCNN consists of the concatenation layers that create dense connections in the network and each convolutional block is constructed using the combinations of activation functions (i.e., Tanh and ReLu).

Diagnosis of PCa using histopathology images has been one of the key topics in recent oncology. In this study, we focused on interpretability and explainability in DL. Our proposed CNN model classified the H&E stained images of PCa into benign and malignant and achieved the best result at test split 1 and 2, obtaining an AUC of 1.00. The interpretable and explainable techniques (i.e., ALV, Grad-CAM, LIME, and SHAP) are very beneficial for individual diagnosis by analyzing each super-pixel in the predicted sample. Also, these methods explain how local explanations affect the final prediction. It is of note that PCa detection and classification is a widely investigated problem in medical data analysis. Therefore, meta-explanation is important to describe the behavior of the black-box model at a more human-understandable level [14]. Research in XCV should be more precise and meaningful because human users are the viewers of XCV results.

Nevertheless, the visualization results extracted from our CNN model are interpretable and explainable which shows the activated and significant feature maps for classification and attention regions in the predicted outputs. However, from the existing research works, it has been analyzed that there are no standardized metrics to evaluate the explainability techniques of CNN.

V. Conclusion

In this paper, an LDCNN model was developed for the classification of H&E stained images of PCa (i.e., benign and malignant). This model was modified from the LWCNN model introduced in our previous study. The modified model plotted some astounding results by prompting an accuracy of 100% at test split 1 and 2. The approaches we used in this study are significantly superior for tissue image classification and detection of the cancer regions. Therefore, the quantitative and qualitative results are promising to rationalize further research of our approach for other domains. In the future, the deliberated approach will be applied to other cancers. However, the present work motivated and encouraged us by providing an excellent output that could be useful in real-life scenarios for the healthcare industry.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2008576).

References

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009, doi: 10.1109/RBME.2009.2034865.

[2] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever, "Breast Cancer Histopathology Image Analysis: A Review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400–1411, May 2014, doi: 10.1109/TBME.2014.2303852.

[3] J. P. Hinton *et al.*, "A Method to Reuse Archived H&amp;E Stained Histology Slides for a Multiplex Protein Biomarker Analysis," *Methods Protoc.*, vol. 2, no. 4, p. 86, Nov. 2019, doi: 10.3390/mps2040086.

[4] J. P. Hinton *et al.*, "A Method to Reuse Archived H&amp;E Stained Histology Slides for a Multiplex Protein Biomarker Analysis," *Methods Protoc.*, vol. 2, no. 4, p. 86, Nov. 2019, doi: 10.3390/mps2040086.

[5] Y. Xu *et al.*, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinformatics*, 2017, doi: 10.1186/s12859-017-1685-x.

[6] A.-S. Metwalli, W. Shen, and C. Q. Wu, "Food Image Recognition Based on Densely Connected Convolutional Neural Networks," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Feb. 2020, pp. 027–032, doi: 10.1109/ICAIIC48513.2020.9065281.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[8] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale Orderless Pooling of Deep Convolutional Activation Features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, pp. 392–407.

[9] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

[10] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*. 2020, doi: 10.3390/JIMAGING6060052.

[11] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[12] S. Bhattacharjee, C.-H. Kim, D. Prakash, H.-G. Park, N.-H. Cho, and H.-K. Choi, “An Efficient Lightweight CNN and Ensemble Machine Learning Classification of Prostate Tissue Using Multilevel Feature Analysis,” *Appl. Sci.*, vol. 10, no. 22, pp. 71-93, Nov. 2020, doi: 10.3390/app10228013.

[13] M. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Feb. 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.

[14] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Adv. Neural Inf. Process. Syst.*, May 2017, [Online].

[15] G. Yang, Q. Ye, and J. Xia, “Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond,” *Inf. Fusion*, 2022, doi: 10.1016/j.inffus.2021.07.016.

[16] Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham, and G. Slaney, “Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging.,” *J. Neurosci. Methods*, vol. 353, p. 109098, Apr. 2021, doi: 10.1016/j.jneumeth.2021.109098.

[17] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri, “Patch-based system for Classification of Breast Histology images using deep learning,” *Comput. Med. Imaging Graph.*, vol. 71, pp. 90–103, Jan. 2019, doi: 10.1016/j.compmedimag.2018.11.003.

[18] S. Chakraborty, S. Aich, A. Kumar, S. Sarkar, J.-S. Sim, and H.-C. Kim, “Detection of cancerous tissue in histopathological images using Dual-Channel Residual Convolutional Neural Networks (DCRCNN),” in *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, Feb. 2020, pp. 197–202, doi: 10.23919/ICACT48636.2020.9061289.

[19] C. Wang, J. Shi, Q. Zhang, and S. Ying, “Histopathological image classification with bilinear convolutional neural networks,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, vol. 2017, pp. 4050–4053, doi: 10.1109/EMBC.2017.8037745.

[20] M. Z. Alom, C. Yakopcic, M. S. Nasrin, T. M. Taha, and V. K. Asari, “Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network,” *J. Digit. Imaging*, 2019, doi: 10.1007/s10278-019-00182-7.

[21] Del Toro, O. J., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., ... & Müller, H. (2017, March). Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. In *Medical Imaging 2017: Digital Pathology*, vol. 10140, p. 101400. International Society for Optics and Photonics.

[22] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized Input Sampling for Explanation of Black-box Models,” *Br. Mach. Vis. Conf. 2018, BMVC 2018*, Jun. 2018.

[23] A. Gramegna and P. Giudici, “SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk,” *Front. Artif. Intell.*, vol. 4, pp. 1-6, Sep. 2021, doi: 10.3389/frai.2021.752558.

[24] “Prostate cANcer graDe Assessment (PANDA) Challenge”, Accessed on: Oct. 10, 2021. [Online]. Available: https://www.kaggle.com/c/prostate-cancer-grade-assessment/overview/description

[25] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2015.

[26] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, “Explaining the black-box model: A survey of local interpretation methods for deep neural networks,” *Neurocomputing*, vol. 419, pp. 168–182, Jan. 2021, doi: 10.1016/j.neucom.2020.08.011.