Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multinomial Classification of Patterns in Lung Cancer Biopsy Slides Using Customized Convolutional Neural Network

Jung Wook Yang
Gyeongsang National University Hospital
Dae Hyun Song
Gyeongsang National University
Hyo Jung An
Gyeongsang National University
Sat Byul Seo (sbseo@kyungnam.ac.kr)
Kyungnam University

Research Article

Keywords: Lung Cancer, Convolutional Neural Network, AUC

Posted Date: June 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-608551/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Identifying the lung carcinoma subtype in small biopsy specimens is an important part of determining a suitable treatment plan but is often challenging without the help of special and/or immunohistochemical stains. Pathology image analysis that tackles this issue would be helpful for diagnoses and subtyping of lung carcinoma. In this study, we developed AI models to classify multinomial patterns of lung carcinoma (adenocarcinoma, squamous cell carcinoma, small cell carcinoma, large cell neuroendocrine carcinoma) and non-neoplastic lung tissue based on convolutional neural networks (CNN or ConvNet). Four CNNs that were pre-trained using transfer learning and one CNN built from scratch were used to classify patch images from pathology whole-slide images (WSIs). We evaluated the diagnostic performance of each model in the test sets. The Xception model achieved the highest performance among pre-trained CNNs with an accuracy of 0.86 and an area under the curve (AUC) of 0.97. The built from scratch CNN model obtained an accuracy of 0.92 and an AUC ranging from 0.99 to 1.00 for subtyping lung carcinoma tasks. These results demonstrate how promising CNN models are for developing improved diagnostic workflow systems for diagnosis and subtyping of lung carcinoma. Of particular note is the fact that the built from scratch CNN described in this paper achieves prompt and consistent results so has the potential to be applied in working hospitals for pathological diagnoses.

Introduction

Lung cancer is most common cause of cancer death (18.0%) while it was the second most commonly diagnosed cancer (2,206,771; 11.4%) worldwide in 2020 behind female breast cancer (2,261,419; 11.7%).[1] The operability of lung cancer is determined by the cancer stage. At the time of diagnosis, approximately 70% of lung cancer patients have advanced stage and are inoperable. In many cases, a final histological diagnosis is made from a small biopsy tissue. After a small biopsy, lung carcinomas are classified as adenocarcinoma (ADC), squamous cell carcinoma (SCC), small cell carcinoma, large cell neuroendocrine carcinoma (LCNEC), or “non-small cell carcinoma (NSCC), not otherwise specified (NOS)”, etc. by additional special and/or immunohistochemical stains, if necessary. The classification “NSCC, NOS” is the diagnostic term used when a lung carcinoma cannot be classified into a specific type from a small biopsy specimen despite the use of the additional stains or when no stains are available.[2] Subtyping for lung carcinoma in small biopsy specimens is often challenging without these special and/or immunohistochemical stains.

Histological subtyping of lung carcinoma is a vital part of determining a suitable treatment plan. In cases of small cell carcinoma, chemotherapy is usually used alone without any surgical treatment because of the advanced cancer stage at the time of diagnosis; it is hard to meet surgical specimens of small cell carcinoma in practice. In the patients with NSCC, surgery is determined by the cancer stage, and the chemotherapy regimen depends on the histological subtype.[3–4]

With the emergence of digital pathology, there have been studies that attempt to analyze digital slide images of lung cancers using deep learning.[5–9] Deep learning-based pathology image analysis that can

accurately diagnose and subtype lung carcinoma would be extremely useful in daily practice as an auxiliary means to help quickly arrive at the most suitable treatment plan. However, there have been no studies that propose technology that is capable of distinguishing non-neoplastic lung tissue and lung carcinoma subtype including adenocarcinoma, squamous cell carcinoma, small cell carcinoma, and LCNEC, which are frequently seen in lung biopsy. In this study, we aimed to develop deep learning-based AI models for multinomial classification of patterns in lung cancer biopsy slides.

Results

Overview of AI models for multi-class classification

We developed and evaluated AI models with a convolutional neural network structure for multi-classification of lung cancer subtypes (Table 2, Figure 1c-1) either built from scratch on the Keras Sequential API (https://keras.io/) or based on four well-known, pre-trained CNNs using transfer learning (Figure 1c-2). Pathology slides from 190 patients of lung or bronchus biopsies were collected at the Gyeongsang National University Hospital. Cancer regions on the WSIs were annotated by pathologists, and the tumor areas in these images were extracted and used to generate nonoverlapping patches 256×256 pixels in size at a magnification of 20× using DeepPATH[7], as shown in Figure 1a. Figure 1b shows the dataset containing a total of 5 classes with 18 ADC, 12 LCNEC, 19 SCC, 18 small cell carcinoma whole slide images (WSIs), and 123 non-tumor WSIs. 10235 sample images were generated from 190 WSIs to represent the four lung cancer subtypes as well as the negative case. Out of the total of 10235 patches, 7366 patches were used to construct a training set (72%), 816 patches were selected for a validation set(8%) and 2053 patches were assigned to the test set (20%), as shown in Table 1. The architecture of the built from scratch CNN is described in Table 2. Four pre-trained convolution neural networks, Xception[10], ResNet152[11], VGG19[12], and NASNETLarge[13], were also used in this study. To compile the model for each CNN, Nadam and categorial cross entropy were chosen as the respective optimizer and loss function. We generated three independent datasets containing training, validation, and test sets from the original dataset using the split-folders python library (https://pypi.org/project/split-folders/). Each AI model diagnosed the images and output either one of the four types of lung cancer or the negative case using the three test sets, as shown in Figure 1d.

Predictive analysis of multi-classification results using AI models

To evaluate the performance of each multi-classification AI model, confusion matrices and normalized confusion matrices were generated. Each row of one of the matrices represents the number of patches in each predicted class according to the AI models, while each column represents the actual number of instances in each class according to the pathologists. The test set consisted of 2053 patches with 353 of ADC, 214 of LCNEC, 263 of SCC, 344 of small cell carcinoma, and 344 of non-tumors. Figure 2 shows the confusion matrix and normalized confusion matrix for each AI model. Table 3 shows the accuracy and loss of each AI model for the three different test sets used. The prediction rate range of the VGG19 model for test set A was 0.54-0.86, as shown in Figure 2a. Figure 2b shows the accuracy of VGG19 was 0.7053

and its loss was 0.8191. Figure 2c shows the prediction rate range of the Xception model was 0.68-0.90, Figure 2d shows its accuracy and loss were 0.8378 and 0.9609, respectively. For the NASNetLarge model, the prediction rate for each class ranged from 0.59 to 0.86 as shown in Figure 2e, while Figure 2f shows its accuracy and loss were 0.7781 and 0.6035, respectively. Figure 2g shows the built from scratch CNN had prediction rates from 0.90 to 0.94 over the various classes, while Figure 2h shows its accuracy and loss were 0.9196 and 0.3056, respectively. The ResNet152 model had the lowest performance across the board so its results are excluded. Table 4 displays the classification results, i.e. the precision, recall, f1 score, and accuracy, of the built from scratch CNN model for each class.

High AUC performance of AI models for discriminating patterns

We evaluated the diagnostic performance of the AI models for multinomial classification of lung cancer subtypes in the test set. The area under the curve (AUC) of the receiver operation characteristic curve (ROC) were evaluated for each model over the three different test sets. VGG19 scored 0.92 for the AUC of the micro-average ROC curve and 0.89 for the AUC of the macro-average ROC curve as shown in Fig 3 a and b. The AUC range of VGG19 for discriminating each lung cancer subtype was 0.86-0.95. The pretrained Xception model achieved AUC of 0.97 for the micro-average ROC curve and the macro-average ROC curve, its AUCs for classifying subtypes of lung cancer ranged from 0.96 to 0.98, as shown in Fig 3 c,d. Fig 3 e,f showed results for the NASNetLarge model with AUCs of 0.95 and 0.94 for the micro-average ROC curve and the macro-average ROC curve, respectively. The AUC range of NASNetLarge for discriminating each subtype of lung cancer was 0.93-0.97. The built from scratch CNN model provided achieved an AUC of 0.99 for both the micro-average ROC curve and the macro-average ROC curve, its AUCs for classifying subtypes of lung cancer ranged from 0.99 to 1.00, as shown in Fig 3 g,h.

Discussion

In this work, we generated various deep learning-based AI models for multinomial classification of patterns in lung cancer biopsy slides. The AI models for the classification of subtype lung carcinoma proceeded with the flow of digital WSI data preparation, convolutional neural networks, loss functions, and diagnostic performance evaluation. Previous research involving lung cancer image analysis with deep learning has focused on differentiating between non-cancer tissue and lung cancer,[5,6] to distinguish adenocarcinoma and squamous cell carcinoma,[7] or looked to identify small cell carcinoma in addition to the cancers previously mentioned.[8] Gonzalez et al. attempted to distinguish LCNEC from small cell carcinoma using deep learning, but this is based on cytology images as the input information,[9] we could not find any studies that have attempted to classify the major carcinoma subtypes, including LCNEC, using lung biopsy images as the input. To the best of our knowledge, our study is the first to identify LCNEC in lung biopsies.

In this study, we took two approaches to developing AI models that distinguish multinomial patterns of lung carcinoma: transfer learning with pre-trained CNNs and a built from scratch CNN. First, we used transfer learning with pre-trained convolutional neural networks based on Xception,[10] ResNet152,[11]

VGG19,[12] and NASNETLarge.[13] Among those models, Xception achieved the highest performance (0.86 accuracy, 0.97 AUC), however, it was difficult for any of these four models to exceed an accuracy of 0.86 with the three test sets used in our experiments, as shown in Table 3. These results demonstrate that transfer learning with pre-trained CNNs on ImageNet is relatively easy to access and ensures obtaining a certain level of verified accuracy. However, these results also imply, even among the pre-trained CNN models, the performances of each model may vary in pathology image analysis, as shown in Table 3. In other words, it may be crucial to choose an appropriate model for the specific task, such as analyzing pathological images.

The second approach was to customize a new convolutional neural network model. The built from scratch CNN model produced outstanding results (0.92 accuracy and AUC ranging from 0.99 to 1.00); this means that the AI model could discriminate which subtype of cancer was present in the test images with great reliability. This level of diagnostic performance has not been reported for the pathological diagnosis of lung cancer subtypes, including the LCNEC subtype. In addition, these results imply that the built from scratch CNN model fitted to the specific task at hand, such as pathological diagnosis of lung cancer subtypes, can be expected to produce better performances than that of transfer learning using pre-trained CNNs on a large benchmark dataset.

There are some limitations to our study. The first limitation of our study is the small number of samples in our dataset. Due to the relatively small number of LCNEC biopsy cases, we decided not to add a great number of samples of the other types of carcinomas to ensure we had a balanced dataset. As such, the total number of carcinomas in our dataset was intentionally kept relatively small to prevent bias. Secondly, our models were built using an intra-hospital dataset. In other studies, public data such as the TCGA dataset was used to classify lung cancer versus normal tissue or the multinomial classification of the various cancer subtypes,[4,5] however, at the time of writing, there are no public databases that include LCNEC and small cell carcinoma slides available. To compensate for these limitations, we generated three independent test sets randomly from within our original dataset. We were then able to obtain consistent cross-validation results, as shown in Table 3.

In conclusion, we generated two types of deep learning-AI models for multinomial classification of patterns in lung cancer biopsy slides. The first type of model was a CNN pre-trained using transfer learning, and our experiments showed that this model was able to consistently classify the various classes of lung cancer a certain level of verified accuracy using actual pathological data as the input. The performance achieved in the experiments conducted demonstrates that a CNN model has the potential to be the basis for developing diagnostic workflow systems for the diagnosis and subtyping of lung cancers. However, the pre-trained CNNs on ImageNet are generally complicated and required a time-consuming process to run, thus expensive equipment is required for experiments. This seems to be challenging for AI models to apply and practical use to all hospitals immediately. The information gathered in this study suggests that another approach to pathology image analysis is to use a relatively simple and built from scratch model fitted to the specific task, like the model demonstrated in this paper.

Especially, the prompt and consistent results of the built from scratch CNN mean that it could be applied in working hospitals for pathological diagnoses.

Material & Methods

Patients

We collected hepatoxylin-eosine stained pathology slides from the pathology reports of 178 patients that underwent lung or bronchus biopsies at Gyeongsang National University Hospital, Jinju, Korea, in 2012 and the pathology slides of 12 patients diagnosed with large cell neuroendocrine carcinoma in their biopsy at Gyeongsang National University Hospital from 2012 to 2018. Out of the patients the pathology slides were taken from, 18, 19, and 18 patients were respectively diagnosed with adenocarcinoma, squamous cell carcinoma, and small cell carcinoma while the others slides (n=123) all came from non-tumor cases. Each diagnosis was histopathologically confirmed by two experienced pathologists. This study was approved by the Institutional Review Board of Gyeongsang National University Hospital with a waiver for informed consent (2021-04-016), and all methods were performed in accordance with the relevant guidelines and regulations.

Whole slide image (WSI) dataset

A total of 190 WSIs were acquired from 190 pathology slides with an Aperio AT2 slide scanner (Leica Biosystems Division of Leica Microsystems Inc., IL, USA) and 400x. Two experienced pathologists annotated the cancer regions on the WSIs with Aperio ImageScope v12.4.3 (Leica Biosystems Division of Leica Microsystems Inc., IL, USA). Tumor areas were extracted from the annotated whole slide images (WSIs), the extracted areas were used to generate nonoverlapping patches 256 256 pixels in size at a magnification of 20 using DeepPATH based on the OpenSlide library in Python (Fig. 1a).[7] In this process, 10235 patches were generated from the original 190 WSIs containing either one of the four lung cancer subtypes or a negative case, as shown in Table 1.

Patch Dataset

The dataset consisted of slides from patients in one of 5 classes with 18 ADC, 12 LCNEC, 19 SCC, and 18 small cell carcinoma whole slide images (WSIs) of lung cancer subtypes as well as 123 non-tumor WSIs. The whole slide images were used to generate 1759 ADC patches, 1061 LCNEC patches, 1314 SCC patches, 1711 small cell carcinoma patches, and 4390 patches of non-tumor whole slide images, as shown in Fig 1b and in Table 1. The patches were removed if the percentage of background in the patch was above 25% according to the DeepPATH program.[7] The input data were randomly divided into a training set, a validation set, and a testing set using the split-folders library in the Python programming language version 3.8.3 (https://pypi.org/project/split-folders/), we then generated three different datasets, A, B, and C, using the split-folders library from the original dataset. Out of 10235 patches, 7366 patches were used to construct the training sets (72%), 816 patches were used for the validation sets (8%) and 2053 patches were used in the test sets (20%).

Convolutional Neural Networks

A deep neural network (DNN) is a supervised classifier that contains multiple layers between input and output layers. [16] A convolutional neural network (CNN, or ConvNet) is a specialized kind of a DNN, CNNs are known to perform particularly when analyzing images.[17] We constructed a convolutional neural network model for the multinomial classification of lung cancer biopsies with the possible outputs being the four lung cancer types or the negative case. Our CNN was built on the Keras Sequential API (https://keras.io/), written in Python and running on TensorFlow (https://www.tensorflow.org/).[18] CNN models take tensors of a certain shape as input, for image analysis CNNs the shape of these tensors are dictated by the height of input images, width, and color channels. Our model takes inputs with dimensions of 244 x 244 x 3 and consists of four convolution blocks with a max pool layer in each. The 1st and 2nd hidden layers of the model have 16 and 32 filters, respectively, with a kernel size of (2, 2) and use a rectified linear unit (ReLU) as their activation functions. The 3rd and 4th hidden layers have 64 filters with a kernel size of (2, 2) and use a rectified linear unit (ReLU) as their activation functions, as shown in Table 2. The fully connected dense layer of the model has 5 units and uses a softmax activation function. Batch size of 200 and 100 epochs were determined as the optimum values for the model when considering both time and computational costs. When compiling the model, Nadam was chosen as the optimizer and categorial cross entropy was selected for the loss function.

Transfer Learning with Pre-Trained ConvNets

We evaluated four AI models that used transfer learning to implement state-of-the-art pre-trained convolutional neural networks. Transfer learning is a subfield of machine learning and artificial intelligence which uses the learned weights of an already trained model to solve a different problem instead of starting the training process of a model over from scratch, this approach saves time and computational costs.[19] Transfer learning for computer vision problems is normally executed by applying pre-trained ConvNet architectures (e.g. VGG, ResNet, Xception etc) that were trained on large benchmark datasets (e.g. ImageNet[14]) to solve a particular problem. The pre-trained convolutional neural networks (ConvNets) allow us to build AI models for image classification with relatively high accuracy and diagnostic performance even if the target dataset is small or if the people tackling the problem do not have the required expertise to train a CNN from scratch. Pre-trained image classification networks are trained on a subset of the ImageNet database used in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC).[20] Four pre-trained convolution neural networks; Xception,[10] ResNet152,[11] VGG19,[12] and NASNETLarge[13] were used in this study. Xception which was build and trained by Google, is a novel deep convolutional neural network inspired by Inception.[10] It slightly outperformed InceptionV3(GoogLeNet)[21] on the ImageNet database. ResNet152 is a newer version of ResNet (Residual Network), which is a convolutional neural network built and trained by Microsoft.[11] VGG19, which has a depth of 19 layers, was established by the University of Oxford in 2014.[12] Lastly, NASNETLarge is a state-of-art neural image classification model built and trained by Google in 2018.[13] The pre-trained models are freely accessible through the Keras Application (https://keras.io/api/applications/), which is a deep

learning library. After the pre-trained models were chosen, we repurposed the knowledge that had already being learned; the layers, features, weights, and biases by fine-tuning to generate the correct outputs for our problem. Batch sizes of 20 and 10 epochs were determined as the optimum values for the pre-trained CNNs in consideration of time and computational costs. When compiling each model, Nadam was chosen as the optimizer and categorial cross entropy was selected for the loss function.

Statistical analysis

To evaluate the classification performance of the AI models, area under the curve (AUC) of the receiver operating characteristic curve (ROC), precision, recall with accuracy, and f1-score were utilized.

True positive (TP): the number of cases where the class was correctly identified versus the rest of classes

False positive (FP): the number of cases where the class was incorrectly identified versus the rest of classes

True negative (FP): the number of cases correctly identified as healthy or other cancer type

False negative (FN): the number of cases incorrectly identified as healthy or other cancer type

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TN}{TN + FP}$$

$$F_1\ \ score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$precision_{micro} = \frac{\sum_{k=1}^{n} TP_k}{\sum_{k=1}^{n} TP_k + \sum_{k=1}^{n} FP_k}, \qquad n = 5$$

$$precision_{macro} = \frac{1}{n} \sum_{k=1}^{n} precision_k, \qquad n = 5$$

All statistical analyses were performed using the scikit-learn library (https://scikit-learn.org/) from Python version 3.8.3 (https://www.python.org/).

Declarations

Data availability

The dataset used in this study might be shared upon reasonable request to Jung Wook Yang, MD, PhD.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government, MSIP; Ministry of Science, ICT & Future Planning (2020R1G1A1A0100746911) to S.S. We are also grateful Prof. Sang Soo Kang for scientific advice and discussion.

Author Contributions: J.Y. and S.S. conceived and designed this study; J.Y., D.H. and H.J. prepared patient samples and pathologic images; J.Y. and D.H. provided annotation and review images; S.S. developed neural network architectures and performed experiment; J.Y. and S.S. discussed the experimental results and wrote the manuscript.

Conflicts of Interest: The authors have no conflicts of interest to disclose.

References

1. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* **0**, 1-41 (2021).
2. Travis, W.D. et al. (Eds.) WHO Classification of Tumours of the Lung, Pleura, Thymus, and Heart, 16-21 (IARC, 2015)
3. Johnson DH, Fehrenbacher L, Novotny WF, Herbst RS, Nemunaitis JJ, Jablons DM, Langer CJ, DeVore RF 3rd, Gaudreault J, Damico LA, Holmgren E, Kabbinavar F. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. *J Clin Oncol*. 22(11), 2184-91 (2004).
4. Selvaggi G, Scagliotti GV. Histologic subtype in NSCLC: does it matter? Oncology (Williston Park). 23(13), 1133-40 (2009).
5. Li, A. et al. Computer-aided diagnosis of lung carcinoma using deep learning – a pilot study. arXiv:1803.05471 (2018).
6. Kanavati, F. et al. Weakly-supervised learning for lung carcinoma classification using deep learning. *Sci Rep*. **10**, 9297 (2020)
7. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med*. **24**,1559-1567 (2018).
8. Kriegsmann, M. et al. Deep Learning for the Classification of Small-Cell and Non-Small-Cell Lung Cancer. *Cancers*. **12**(6), 1604 (2020).
9. Gonzalez, D. et al. Feasibility of a deep learning algorithm to distinguish large cell neuroendocrine from small cell lung carcinoma in cytology specimens. *Cytopathology*. **31**(5), 426-431(2020).
10. Chollet, Francois. Xception: Deep learning with depthwise separable convolutions. arXiv:1610.02357(2016).

11. He, K. et al. Deep residual learning for image recognition. arXiv:1512.03385(2015).
12. Simonyan, K & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR).* (2015).
13. Zoph, B. et al. Learning Transferable Architectures for Scalable Image Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8697-8710 (2018).
14. *ImageNet.* http://www.image-net.org/
15. Li, D. et al. A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. *Nat Commun* **11,** 6004 (2020). https://doi.org/10.1038/s41467-020-19817-3
16. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks.* 61, 85–117 (2015).
17. Goodfellow, I. et al. Deep Learning. *MIT Press.* (2016).
18. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. Available at: tensorflow.org (2015).
19. Torrey, L. et al. Transfer Learning. *Handbook of Research on Machine Learning Applications* (2009).
20. Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*. 115(3), 211-252(2015).
21. Szegedy, C. et al. Rethinking the inception architecture for computer vision. *arXiv* preprint arXiv:1512.00567 (2015).

Tables

Table 1
Summary of number of whole slide images (WSIs) and patches in the dataset: training, validation, and test set among four lung cancer subtypes and a negative case.

Subtypes		Total(n)		Training Set	Validation Set	Test Set
				72%	8%	20%
		WSIs	Patches	Patches	Patches	Patches
Tumor	ADC	18	1759	1266	140	353
	LCNEC	12	1061	763	84	214
	SCC	19	1314	946	105	263
	Small cell carcinoma	18	1711	1231	136	344
Non-tumor		123	4390	3160	351	879
Total		190	10235	7366	816	2053

ADC, adenocarcinoma; **LCNEC**, large cell neuroendocrine carcinoma; **SCC**, squamous cell carcinoma;

Table 2
Summary of 2D CNN model architecture

Layer (options)	Output Shape	Number of Parameters
Input	(None, 224, 224, 3)	-
Conv2D(filters = 16, kernel_size=(2,2), activation='relu')	(None, 223, 223, 16)	208
MaxPooling2D	(None, 111, 111, 16)	0
Conv2D(filters = 32, kernel_size=(2,2), activation='relu')	(None, 110, 110, 32)	2080
MaxPooling2D	(None, 55, 55, 32)	0
Conv2D(filters = 64, kernel_size=(2,2), activation='relu')	(None, 54, 54, 64)	8256
MaxPooling2D	(None, 27, 27, 64)	0
Conv2D(filters = 64, kernel_size=(2,2), activation='relu')	(None, 26, 26, 64)	16448
MaxPooling2D	(None, 13, 13, 64)	0
Global Average Pooling2D	(None, 64)	0
Dense(unit = 5, activation='softmax')	(None, 5)	325

Table 3
Accuracy and Loss of AI models with Test sets

	VGG19		Xception		NASNetLarge		New CNN	
	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
Test set A	0.8191	0.7053	0.9609	0.8378	0.6035	0.7781	0.3056	0.9196
Test set B	0.8419	0.6912	0.5027	0.8266	0.6065	0.7810	0.3028	0.9264
Test set C	0.8448	0.6961	0.5656	0.8563	0.6185	0.7622	0.3637	0.9070

Table 4
Classification Report for a new CNN model

	Precision	Recall	f1-score	Support
ADC	0.8840	0.9065	0.8951	353
LCNEC	0.9571	0.9393	0.9481	214
SCC	0.8655	0.9049	0.8848	263
Small cell carcinoma	0.9456	0.9099	0.9274	344
non-tumor	0.9326	0.9283	0.9304	879
accuracy			0.9196	2053
micro average	0.9127	0.9178	0.9172	2053
weighted average	0.9204	0.9204	0.9198	2053

ADC, adenocarcinoma; **LCNEC**, large cell neuroendocrine carcinoma; **SCC**, squamous cell carcinoma; **Precision**, the fraction of relevant instances among the retrieved instances; **Recall**, the fraction of relevant instances that were retrieved; **f1-score**, a measure of test set's accuracy and the harmonic mean of the precision and recall; **Support**, number of test set for each label and total; **weighted average**, averaging the support-weighted mean per label, **micro average**, averaging the total true positives, false negatives and false positives.

Figures

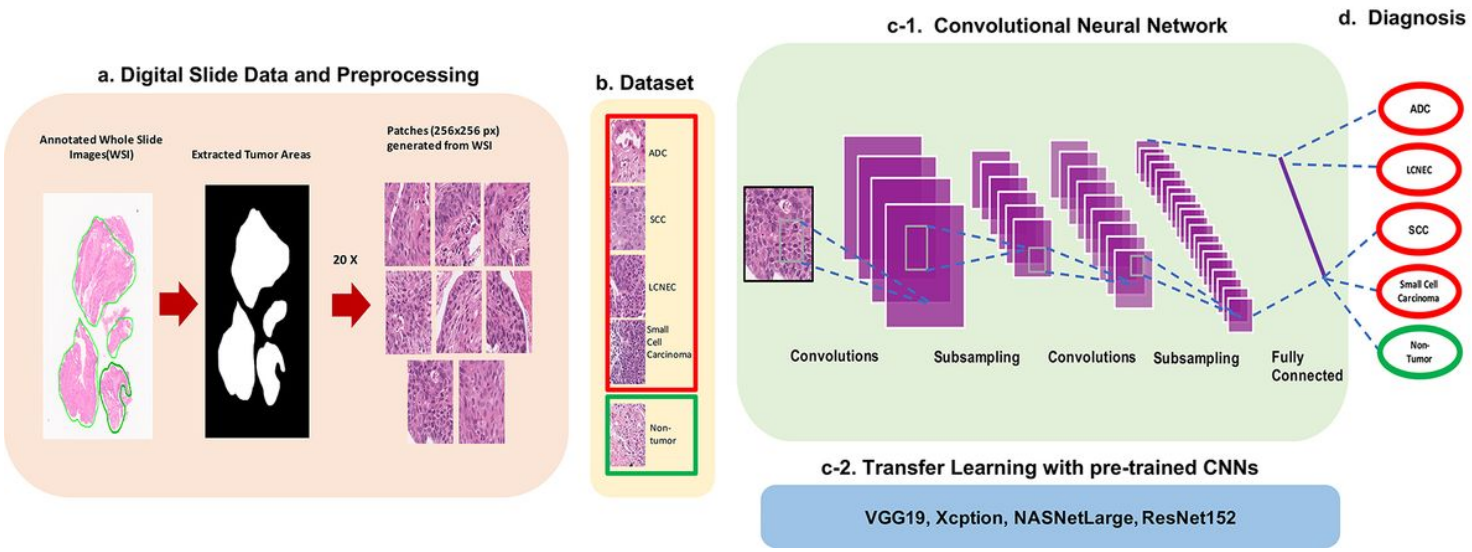

Figure 1

Workflow of AI model for multinomial pattern classification in lung cancer biopsies. a. Digital slide data and preprocessing. Cancer regions on the whole slide images (WSIs) were annotated by pathologists, then the tumor areas were extracted and used to generate nonoverlapping patches 256×256 pixels in size at a magnification of 20×. b. Dataset. The dataset contains a total of 5 classes with 18 ADC, 12 LCNEC, 19 SCC, and 18 small cell carcinoma WSIs, as well as 123 non-tumor WSIs. 10235 patches were generated from the original 190 WSIs containing the four lung cancer subtypes and negative cases. Out of 10235 patches, 7366 patches were used to construct the training sets (72%), 816 patches were in the validation sets (8%) and 2053patches were in the test sets (20%). c-1. Convolutional Neural Network. Our model takes a tensor with dimensions of (244, 244, 3) as input, the CNN consists of four convolution blocks with a max pool layer in each. The 1st and 2nd hidden layer have 16 and 32 filters, respectively, with a kernel size of (2, 2) and use a rectified linear unit (ReLU) as their activation functions. The 3rd and 4th hidden layer have 64 filters with a kernel size of (2, 2) and also use rectified linear units (ReLU). The CNN also contains a fully connected dense layer with 5 units that uses softmax as its activation function. When compiling the model, Nadam and categorial cross entropy were chosen as the optimizer and loss function, respectively. c-2. Transfer Learning with pre-trained CNNs. Four pre-trained convolution neural networks based on Xception[Chollet 2016], VGG19[Simonyan 2015], and NasNetLarge[Zoph 2018] were evaluated in this study. After the pre-trained models were chosen, we repurposed their already learned knowledge and carried out fine-tuning for our task. d. Diagnosis. The AI models diagnosed the input images from the test sets as either one of four types of lung cancer or as a negative case.

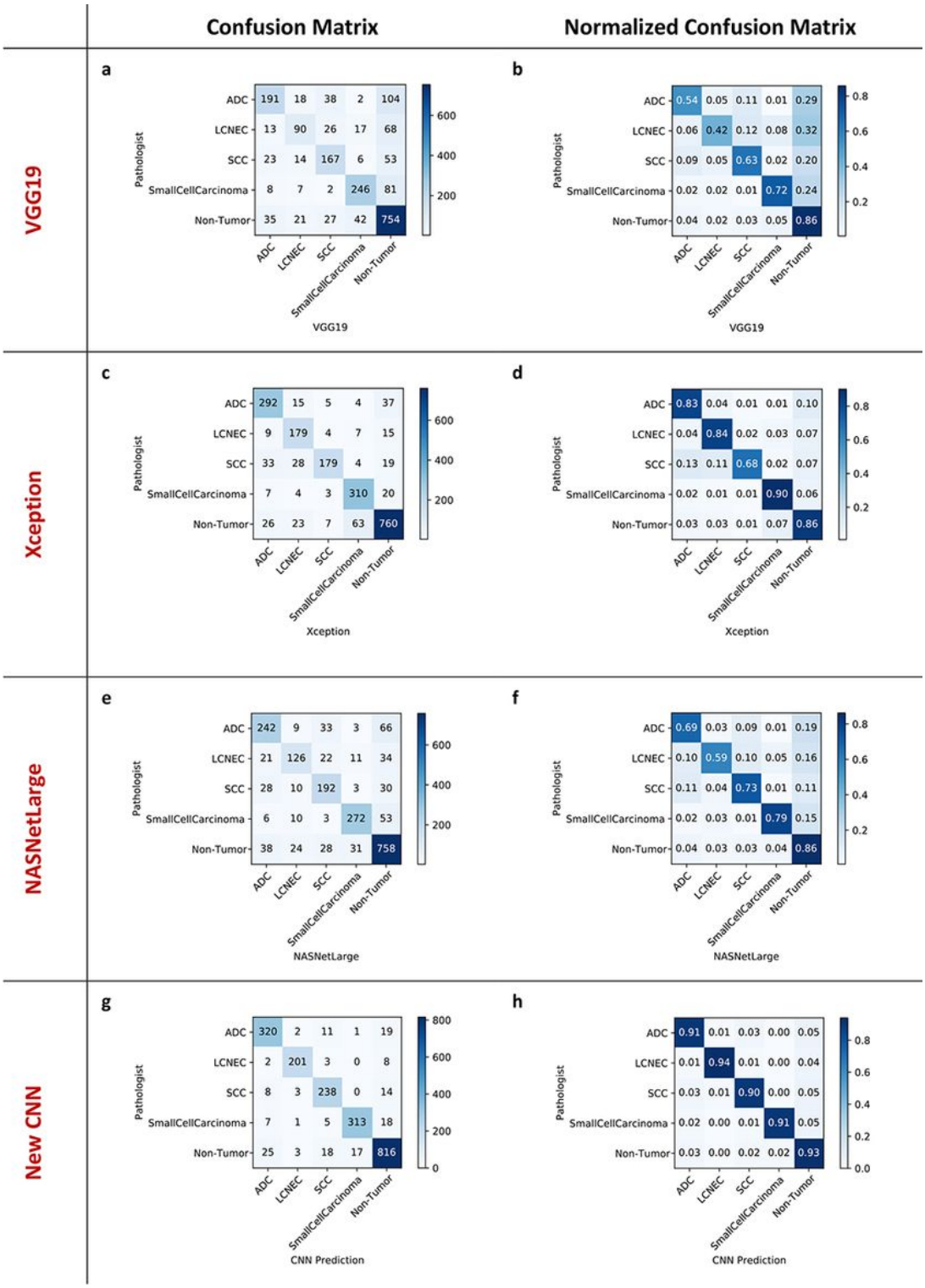

Figure 2

Confusion matrices of the multi-classification outputs from the AI models To evaluate the performance of each multi-classification model, a confusion matrix and a normalized confusion matrix were generated for each model [VGG19 (a, b), Xception (c, d), NASNetLarge (e, f), new CNN (g, h)]. Each row of each matrix represents the number of patches in a predicted class by the corresponding AI model, while each column represents the actual instances in each class according to pathologists. The test sets consisted

of 2053 patches; 353 of adenocarcinoma (ADC), 214 of large cell neuroendocrine carcinoma (LCNEC), 263 of squamous cell carcinoma (SCC), 344 of small cell carcinoma, and 344 from non-tumor cases.

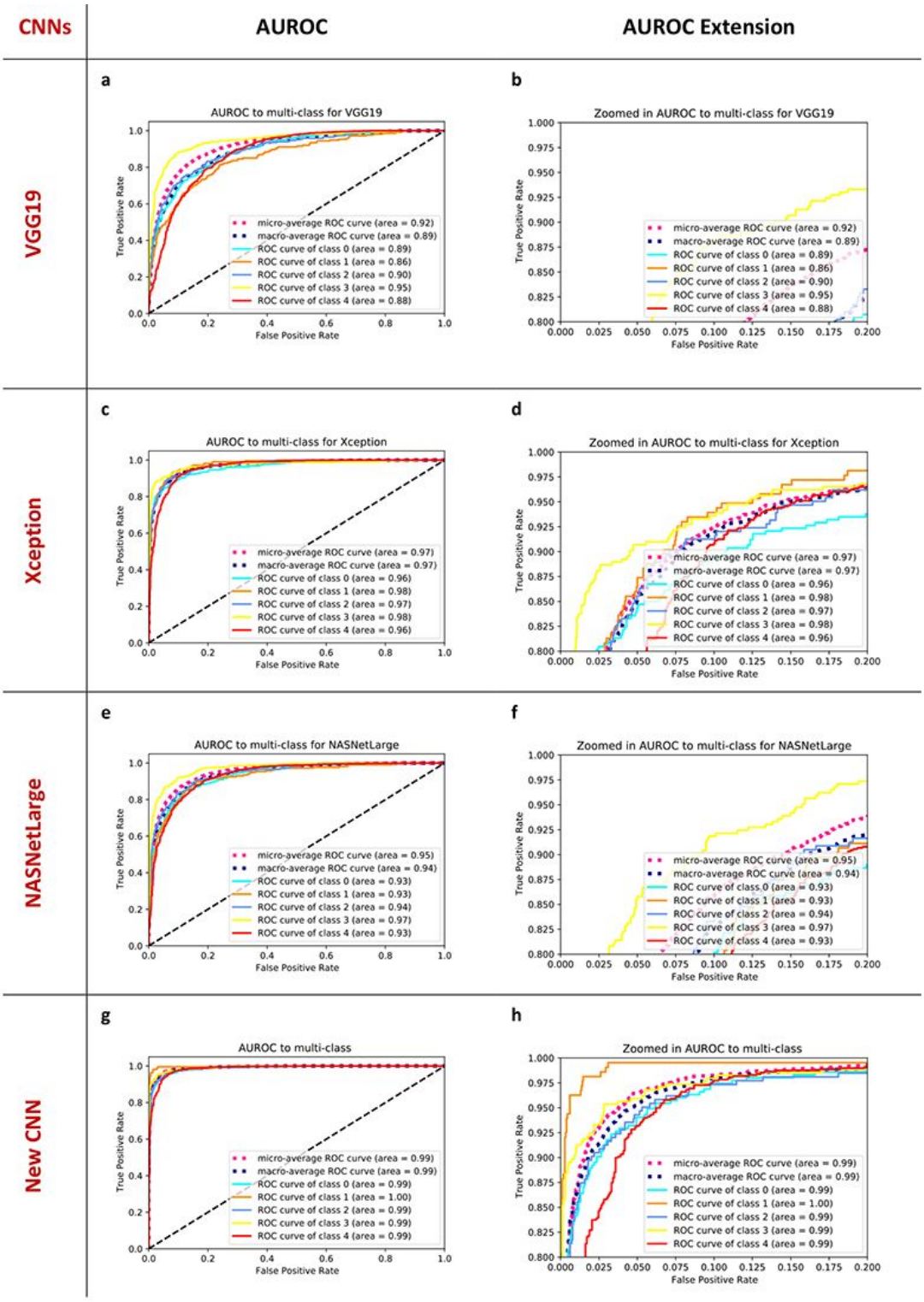

Figure 3

Areas under the curve (AUC) for the receiver operation characteristic curve (ROC) achieved by the AI models. The diagnostic performance of each AI model in multinomial classification of lung cancer subtypes was evaluated using three test sets. The area under the curve (AUC) of receiver operation

characteristic curve (ROC) and extended AUROCs results achieved by each model are shown here [VGG19 (a, b), Xception (c, d), NASNetLarge (e, f), new CNN (g, h)]. Class 0: ADC, adenocarcinoma; class 1: LCNEC, large cell neuroendocrine carcinoma; class 2: SCC, squamous cell carcinoma; class 3: small cell carcinoma; class 4: non-tumor.