MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

Computer-Aided Pathologic Diagnosis of Nasopharyngeal Carcinoma Based on Deep Learning

Songhui Diao,^{*†} Jiaxin Hou,^{*‡} Hong Yu,[§] Xia Zhao,[§] Yikang Sun,[§] Ricardo Lewis Lambo,^{*} Yaoqin Xie,^{*} Lei Liu,[§] Wenjian Qin,^{*} and Weiren Luo[§]

From the Shenzhen Institutes of Advanced Technology,^{*} Chinese Academy of Science, Shenzhen; the Department of Pathology,[§] Cancer Research Institute, The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen Third People's Hospital, National Clinical Research Center for Infectious Diseases, Shenzhen; the Shenzhen College of Advanced Technology,[†] University of Chinese Academy of Sciences, Shenzhen; and the College of Medicine and Biological Information Engineering,[‡] Northeastern University, Shenyang, China

Accepted for publication
April 9, 2020.

Address correspondence to Wenjian Qin, Ph.D., Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, 1068 Xueyuan Ave., Shenzhen, 518055, P.R. China; or Weiren Luo, M.D., Ph.D., Department of Pathology, Cancer Research Institute, The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen Third People's Hospital, National Clinical Research Center for Infectious Diseases, 29 Bulan Rd., Shenzhen, 518112, P.R. China. E-mail: wj.qin@siat.ac.cn or binglike-01@szsy.sustech.edu.cn.

The pathologic diagnosis of nasopharyngeal carcinoma (NPC) by different pathologists is often inefficient and inconsistent. We have therefore introduced a deep learning algorithm into this process and compared the performance of the model with that of three pathologists with different levels of experience to demonstrate its clinical value. In this retrospective study, a total of 1970 whole slide images of 731 cases were collected and divided into training, validation, and testing sets. Inception-v3, which is a state-of-the-art convolutional neural network, was trained to classify images into three categories: chronic nasopharyngeal inflammation, lymphoid hyperplasia, and NPC. The mean area under the curve (AUC) of the deep learning model is 0.936 based on the testing set, and its AUCs for the three image categories are 0.905, 0.972, and 0.930, respectively. In the comparison with the three pathologists, the model outperforms the junior and intermediate pathologists, and has only a slightly lower performance than the senior pathologist when considered in terms of accuracy, specificity, sensitivity, AUC, and consistency. To our knowledge, this is the first study about the application of deep learning to NPC pathologic diagnosis. In clinical practice, the deep learning model can potentially assist pathologists by providing a second opinion on their NPC diagnoses. (*Am J Pathol* 2020, 190: 1691–1700; <https://doi.org/10.1016/j.ajpath.2020.04.008>)

Nasopharyngeal carcinoma (NPC) is an uncommon malignant tumor with a special geographic distribution.¹ According to statistics, in 2018, there were 129,079 new cases and 72,987 deaths worldwide.² The incidence of NPC is especially high in the south of China. Because of the nonspecific symptoms of NPC, most patients are only diagnosed when it has reached its intermediate or terminal stage.³ In the diagnosis process followed by pathologists, inflammation and lymphoid hyperplasia are easily confused with malignant tumors.^{4,5} Some examples of these three categories are shown in Figure 1. Diagnosis of NPC will have an important influence on the selection of the therapy regimen.

Histopathology in which a positive biopsy is taken from the tumor is the gold standard for diagnosis. The accuracy of histopathology diagnosis largely depends on the pathologist's experience. It usually takes over 10 years to cultivate a senior pathologist, and their small numbers lead to a high

Supported by the National Natural Science Foundation of China grants 61901463 and 81872202; Shenzhen Science and Technology Program of China grant JCYJ20170818160306270; Natural Science Foundation of Guangdong Province grants 2015A030313263 and 2018A030313778; and Guangdong Provincial Key Laboratory of Cell Microenvironment and Disease Research grant 2017B030301018.

S.D., J.H., and H.Y. contributed equally to this work.

Disclosures: None declared.

Figure 1 Examples of three categories of whole slide images in the diagnosis of nasopharyngeal lesions: chronic nasopharyngeal inflammation (A–D), lymphoid hyperplasia (E–H), and nasopharyngeal carcinoma (I–L) (from top to bottom). Scale bars = 5 mm (A–L).

workload for those available. Meanwhile, discordant diagnostic results may arise among different pathologists, especially in complex cases, because diagnosis based on morphology is subjective. In some previous studies, researchers have given some mechanisms to predict NPC and proposed various methods to assist in NPC diagnosis.^{6–8} However, in these methods, either the process of analysis is very complex, or they lack clear quantitative indicators.

The growing availability of whole slide images (WSIs) for histopathology diagnosis has led to the trend of digital analysis.^{9,10} Nonetheless, it is still a challenge for computers to handle gigapixel WSIs that contain large amounts of complex information.¹¹ Deep learning is a branch of machine learning that combines low-level features to form more abstract high-level representations of the attributed categories to discover distributed representations of data.¹² With the recent development of convolutional neural networks, deep learning continues to mature and find uses in medical image processing. Through adequate training using images of pathology and their labels, convolutional neural networks can diagnose new images automatically to improve the accuracy

of diagnoses and to significantly reduce the workload of pathologists.¹³ Moreover, deep learning has already been investigated and applied to diagnosis, and has shown its superiority to previous approaches.¹⁴

In the pathologic diagnosis of NPC, some complex cases are easily misdiagnosed by inexperienced pathologists. For example, reactive lymphoid hyperplasia is often found in cases of chronic nasopharyngeal inflammation (hereinafter referred to as inflammation) and NPC, making it sometimes difficult to distinguish between the two. At the same time, the existence of interobserver variability also reduces the reliability of the diagnosis. We have thus applied deep learning to solve these problems. A state-of-the-art convolutional neural network architecture was chosen to accomplish classification of WSIs, and its results were compared with those of pathologists. The following are the main contributions of this paper: i) introduction of deep learning into the pathologic diagnosis of NPC based on WSIs, ii) reduction of the dependence of personal experience in the diagnostic process, which will assist junior and intermediate pathologists, and iii) achievement of

Table 1 Case Characteristics in Training, Validation, and Testing Sets

Training examinations					Validation examinations				Testing examinations			
		Lymphoid					Lymphoid				Lymphoid	
Data set	Inflammation	hyperplasia	NPC		Data set	Inflammation	hyperplasia	NPC	Data set	Inflammation	hyperplasia	NPC
Cases	481 (100)	219 (46)	74 (15)	188 (39)	136 (100)	58 (43)	30 (22)	48 (35)	114 (100)	39 (34)	34 (30)	41 (36)
WSIs	1244 (100)	644 (52)	130 (10)	470 (38)	612 (100)	315 (52)	63 (10)	234 (38)	114 (100)	39 (34)	34 (30)	41 (36)

Data are expressed as *n* (%).
Inflammation refers to chronic nasopharyngeal inflammation.
NPC, nasopharyngeal carcinoma; WSI, whole slide image.

Figure 2 Flowchart of the experimental process. The data were divided into training set, validation set, and testing set. The model was then trained as illustrated in Figure 3. Finally, the performance of the model was compared with that of three pathologists (Y.S., X.Z., H.Y.) with different levels of experience who made the diagnoses of the testing set.

unambiguous diagnosis of NPCs. Our code is available online (GitHub, <https://github.com/SH-Diao123/NPC-diagnosis-based-on-deep-learning>, last accessed March 28, 2020).

Materials and Methods

Image Data Acquisition

Between April 2004 and September 2018, 1970 WSIs of 731 biopsied cases were collected in order to make a definitive diagnosis at the Department of Pathology, the People's Hospital of Gaozhou and Shenzhen Third People's Hospital. The age range of the patients who provided these

samples varied from 18 to 71 years, with an average age of 43 years. All cancer samples were classified as non-keratinizing carcinoma according to the World Health Organization histologic classification. The collected WSIs consist of three categories: 316 cases of inflammation, 138 cases of lymphoid hyperplasia, and 277 cases of NPC. Informed consent was obtained from the institutional research ethics committee. These WSIs were acquired by scanning formalin-fixed, paraffin-embedded hematoxylin and eosin–stained tissues using a Motic VM1000 scanner (Motic, Xiamen, China). They were annotated by two board-certified pathologists with at least 15 years of clinical experience (H.Y. and W.L.) to generate the ground truth in a single-blinded fashion according to the guidelines of the

Figure 3 Overview of the deep learning method used in this study for diagnosis of nasopharyngeal carcinoma. First, whole slide images (WSIs) consisting of chronic nasopharyngeal inflammation, lymphoid hyperplasia, and nasopharyngeal carcinoma were annotated by pathologists (Y.S., X.Z., H.Y.) as 0, 1, or 2. These WSIs were separated into three datasets in appropriate proportions: training set, validation set, and testing set. Secondly, patches were extracted from annotated WSIs, each with a size of 600 pixels $\times$ 600 pixels $\times$ 3 channels (height $\times$ width $\times$ channel) and given the same label as the parent WSI. The pre-trained deep learning model (Inception-v3) was trained with patches from the training set multiple times to obtain a finely tuned model. The full connection layer of Inception-v3 outputs the final classification. The parameters were constantly fine-tuned by feedback from the loss function to improve network performance. The patches are finally classified into the three categories mentioned above.

Figure 4 Example of patches from different categories of whole slide images (WSIs), each with a size of 600 × 600 pixels. The patches in the first row were cut from chronic nasopharyngeal inflammation WSIs; the patches in the second row were cut from lymphoid hyperplasia WSIs; the patches in the third row were cut from nasopharyngeal carcinoma WSIs. For complex cases, it is difficult to determine which category of WSI they belong to based solely on morphology. These patches belong to the training set that was used to train the model. Scale bars = 125 μ m.

*WHO Classification of Head and Neck Tumors, 4th Edition*¹⁵ that is, a pathologist would read the scan once, and no consensus or review between the pathologists was performed. Moreover, an immunohistochemical (PanCK, AE1/AE3, CD3, CD20; Zymed Laboratories, San Diego, CA) test was performed for final confirmation in the cases in which it was difficult to differentiate inflammation and lymphoid hyperplasia from NPC. WSIs are generally stored in a pyramidal structure with a magnification varying from $\times 4$ to $\times 40$. On the basis of previous experience,¹⁶ images were chosen with a magnification of $\times 20$ to execute these experiments. The WSIs were divided into three datasets: a training set (1244 WSIs from 481 cases), a validation set (612 WSIs from 136 cases), and a testing set (114 WSIs from 114 cases), with each set in an appropriate ratio that is discussed in the *Patients Characteristics* section of *Results*. The details of the datasets are shown in *Table 1*.

Deep Learning Algorithm

Inception-v3 is one of the state-of-the-art deep learning models for classification that decomposes convolution and uses

regularization to get superior results with fewer parameters.¹⁷ Its performance was verified by the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 (ImageNet, <http://image-net.org/challenges/LSVRC/2012/index>, last accessed May 4, 2019), which is a commonly used dataset containing 1000 categories of daily objects used to prove the visual object recognition ability of the model.¹⁸ Inception-v3 was therefore chosen for pathology image classification, which was implemented with PyTorch version 1.0 in the Python language on an Intel Xeon processor E5-2620 v4 (512GB, 2.10 GHz; Intel, Santa Clara, CA) and Nvidia GPU Tesla V100 (32GB; Nvidia, Santa Clara, CA).

It is well known that deep learning model training needs a large amount of data. Compared with natural images, medical images are relatively few, so the transfer learning technique was used in this study to obtain an effectively trained deep learning model based on images of pathology. The model was thus first trained on ILSVRC2012 to obtain initial parameters. Keeping the initial parameters, the model was then trained using the pathologic image dataset to fine-tune some of these parameters. Multiple trainings were performed to get the best performance model.

Table 2 Result of Training and Validation Sets of the Model

Evaluation indicator	Accuracy	Category	AUC	Sensitivity	Specificity
Training set	0.935	Inflammation	0.920	0.988	0.867
		Lymphoid hyperplasia	0.980	0.999	0.892
		NPC	0.950	0.977	0.896
Validation set	0.905	Inflammation	0.912	0.929	0.869
		Lymphoid hyperplasia	0.946	0.931	0.898
		NPC	0.869	0.929	0.801

Inflammation refers to chronic nasopharyngeal inflammation.
AUC, area under the curve; NPC, nasopharyngeal carcinoma.

Figure 5 Receiver-operating characteristic curves (ROCs) for the prediction of nasopharyngeal carcinoma. The area under the curve (AUC) reflects performance: the larger the area, the better the performance. Green curves indicate chronic nasopharyngeal inflammation; red curves, lymphoid hyperplasia; and blue curves, nasopharyngeal carcinoma. **A:** AUCs of the model for predicting the three categories based on the training set (482 patients, 1244 images). **Dashed line** indicates the reference line. **B:** AUCs of the model for predicting three categories based on the validation set (136 patients, 612 images). **Dashed line** indicates the reference line. **C:** AUCs of the model for predicting the three categories based on the testing set (114 patients, 114 images). **D:** AUCs of the junior pathologist's (Y.S.) diagnoses based on the testing set. **E:** AUCs of the intermediate pathologist's (X.Z.) diagnoses based on the testing set. **F:** AUCs of the senior pathologist's (H.Y.) diagnoses based on the testing set. mAUC, mean of the area under the curve.

The task of distinguishing three categories of WSIs is achieved by the network model to predict the probability of each category and make a diagnosis. Among them, the tumor region is the most difficult to determine. The weight of the tumor region parameters during training was increased so that the model could learn more from tumor regions and thus improve the accuracy of the diagnosis.

Experiments

The experiments that were performed include training and testing the deep learning model and comparing the manual diagnoses with computer-aided diagnoses. An illustration of these processes is shown in Figure 2.

Consistency evaluation of the experiment

Figure 6 Comparison of the evaluation consistency of the model and the pathologists (Y.S., X.Z., H.Y.) using the Jaccard index, Euclidean distance, and kappa factor. The closer the values of Jaccard index and kappa factor are to 1, the more the diagnoses agree; whereas the closer the value of the Euclidean distance is to 0, the more the diagnoses agree. For the three indices, the senior pathologist (H.Y.) has the highest evaluation consistency, whereas the model takes second place. AUC, area under the curve.

Figure 7 Visual evaluation of consistency of the model and pathologists (Y.S., X.Z., H.Y.). The lighter the color is, the higher the consistency. **A:** Yellow indicates the 1 reference line, and purple indicates the 0 reference line. Most evaluations of the model and pathologists are green to yellow. Although some differences are not obviously significant, they are important in clinical practice. **B:** A closer look at part of the axis (from 0.80 to 1.00) to make the differences more obvious. It is easy to see that the consistency of the model is better than that of the junior (Y.S.) and intermediate pathologists (X.Z.).

The model-training part of the experiment is shown in Figure 3. The three categories of the collected WSIs were annotated by pathologists with a 0, 1, or 2. Because the size of the WSIs was too large for a computer to deal with

directly, they were cut into smaller patches, each with a size of 600 pixels $\times$ 600 pixels $\times$ 3 channels (height $\times$ width $\times$ channel) and a label that was the same as that of the parent WSI. Examples of patches from the three categories are shown in Figure 4. Training the pre-trained Inception-v3 model multiple times with patches from the training set produced a finely tuned architecture. The full connection layer of Inception-v3 outputs the final classification. The parameters are constantly updated by feedback from the cross-entropy loss function to improve network performance.

The contrast experiments are executed on the testing set. The pathologists involved in the experiment have different levels of experience and are classified as junior, intermediate, and senior. A junior pathologist is defined as a recent hire who has diagnosed only a few pathology sections. An intermediate pathologist has several years of experience in diagnosis. A senior pathologist has more than 10 years of experience in diagnosis. For the model, the junior pathologist, the intermediate pathologist, and the senior pathologist each made an independent diagnosis and then evaluated the quality of his/her diagnosis using unified indicators for comparative analysis.

Evaluation and Analysis

The performance of the model was evaluated based on accuracy, sensitivity, specificity, the receiver-operating characteristic curve, and the area under the curve (AUC). The AUC varies from 0 to 1, with 1 indicating perfect classifier performance and 0.5 indicating that classifier performance is no better than random. The receiver-operating characteristic curves and AUCs were used as metrics to compare the diagnostic performances of the pathologists and the model in classifying the WSIs.

The consistency of the model was also analyzed using the Jaccard index, Euclidean distance, and kappa factor, and compared with those of the pathologists. The Jaccard index ranges from 0 to 1, with high values indicating a high similarity. The Euclidean distance also ranges from 0 to 1, with low values indicating high similarity. In the case of the kappa value, a negative value indicates poor agreement, 0 to 0.20 indicates slight agreement, 0.21 to 0.40 indicates fair agreement, 0.41 to 0.60 indicates moderate agreement, 0.61 to 0.80 indicates substantial agreement, and 0.81 to 1.00 indicates perfect agreement.¹⁹

Results

Patient Characteristics

There were 316 inflammation cases (43%), 138 lymphoid hyperplasia cases (19%), and 277 NPC cases (38%) in total. In the training and validation sets, the ratio of the three categories was about 5:1:4, respectively. In the testing set,

Figure 8 A comparison of the accuracy of the model and the pathologists (Y.S., X.Z., H.Y.). The areas indicate the quality of their diagnoses. In order of size, their respective areas are as follows: senior pathologist (H.Y.) > the model > intermediate pathologist (X.Z.) > junior pathologist (Y.S.).

the ratio was close to 1:1:1 in order to assess the classification performance of each category more fairly.

Performance Evaluation

The performances of the trained model based on the training and validation sets are listed in Table 2. Based on the training dataset, for the diagnosis of inflammation, lymphoid hyperplasia, and NPC, the respective AUCs are 0.920, 0.980, and 0.950, the respective sensitivity values are 0.988, 0.999, and 0.977, and the respective specificity values are 0.867, 0.892, and 0.896. The mean accuracy of the diagnosis is 0.935 using the training dataset. Based on the validation dataset, for the diagnosis of inflammation, lymphoid hyperplasia, and NPC, the respective AUCs are 0.912, 0.946, and 0.869, the respective sensitivity values are 0.929, 0.931, and 0.929, and the respective specificity values are 0.869, 0.898, and 0.801. The mean accuracy of the diagnosis is 0.905 using the validation dataset. It can be seen that the model performs well with both the training set and the validation set.

The results of the model based on the testing set are shown in Figure 5C. The AUCs for the diagnosis of inflammation, lymphoid hyperplasia, and NPC are 0.905, 0.972, and 0.930, respectively, and the mean AUC is 0.936. Using the AUC as an evaluation standard, the model can be said to perform well. The pathologists made their diagnoses on the same testing dataset, and the doctors' diagnoses were evaluated in the same way. As shown in Figure 5, D–F, for the junior pathologist, the AUCs for the diagnoses of inflammation, lymphoid hyperplasia, and NPC are 0.815,

0.961, and 0.933, respectively, and the mean AUC is 0.903. For the intermediate pathologist, the AUCs for the diagnoses of inflammation, lymphoid hyperplasia, and NPC are 0.851, 0.981, and 0.900, respectively, and the mean AUC is 0.909. For the senior pathologist, the AUCs for the diagnoses of inflammation, lymphoid hyperplasia, and NPC are 0.910, 0.982, and 0.975, respectively, and the mean AUC is 0.956. As expected, the overall quality of the diagnoses of the senior pathologist is higher than that of the other pathologists. In summary, the model performs better than the junior and the intermediate pathologists, and the value of the evaluation index is only slightly lower than that of the senior pathologist.

Consistency Evaluation

The consistency of the model and the pathologists were also analyzed using three indices. The results are shown in Figure 6. For the model, the values of the Jaccard index, Euclidean distance, and kappa factor are 0.879, 0.242, and 0.815, respectively. For the junior pathologist, the values of the Jaccard index, Euclidean distance, and kappa factor are 0.825, 0.296, and 0.735, respectively. For the intermediate pathologist, the values of the Jaccard index, Euclidean distance, and kappa factor are 0.860, 0.265, and 0.842, respectively. For the senior pathologist, the values of the Jaccard index, Euclidean distance, and Kappa factor are 0.895, 0.230, and 0.842, respectively. Meanwhile, the kappa factor for the combination of all pathologists is 0.806. The visual evaluation of consistency is shown in Figure 7. Although the difference in color intensity is not significant

Figure 9 Comparison of the model's and pathologists' (Y.S., X.Z., H.Y.) percentages of correct and incorrect diagnoses. **A:** The comparison between the model and the junior pathologist (Y.S.). **B:** The comparison between the model and the intermediate pathologist (X.Z.). **C:** The comparison between the model and the senior pathologist (H.Y.). According to the results, the percentage of incorrect diagnoses falls significantly when the pathologists are paired with the model.

in Figure 7A, a small difference in diagnostic quality can have a significant impact on clinical diagnosis. By taking a closer look at part of the axis, one can see that the difference becomes more apparent as shown in Figure 7B in which a light color indicates high consistency. It is obvious that all of the evaluation indices of the senior doctor ranked first and that the model takes second place.

Discussion

In this paper, an automated pathologic diagnostic model for NPC was developed that shows a performance comparable

to that of pathologists. In earlier studies of NPC diagnosis, it was a challenging problem that had no perfect solution. For medical imaging diagnosis, most studies are based on computed tomography, magnetic resonance imaging, positron emission tomography, ultrasound, and endoscopy images.^{20–22} Some researchers have also included biomarkers for diagnosis.^{23,24} However, the definitive diagnosis must be made according to biopsy. As far as we know, deep learning has shown great potential in the application of medical diagnosis. In a study on the pathologic diagnosis of melanoma,²⁵ the model based on deep learning outperformed 11 board-certified pathologists, demonstrating the great potential of the model in assisting pathologists to perform diagnoses. But before the present study, deep learning had not yet been applied to the pathologic diagnosis of NPC. These results suggest that the current study model can act as a reliable reference for a pathologist's diagnosis.

The greatest contribution of our model is to reduce dependency on the pathologist's experience. According to the performance evaluations, the model is superior to the junior and intermediate pathologists and is approximately equal to the senior pathologist in diagnosing inflammation. In the diagnosis of lymphoid hyperplasia, the model's performance is also better than that of the junior and the intermediate pathologists. The model can therefore play an excellent auxiliary role in the classification of inflammation and lymphoid hyperplasia. Figure 8 is a radar plot in which the accuracy of each diagnostician is represented by surface area. It can be seen intuitively that the area corresponding to the model is larger than that corresponding to the junior and intermediate pathologists, and is similar to that corresponding to the senior pathologist. This indicates that the overall performance of the model is only slightly below that of a senior pathologist, and its diagnostic results should have considerable reference value for junior and intermediate pathologists.

By using the deep learning model, we have also achieved a fairly objective and quantitative pathologic diagnosis. In the evaluation of consistency, the model outperformed junior and intermediate pathologists, and the results were better than that of the combined diagnoses of pathologists with different levels of experience. Many studies have proposed that computer analysis can significantly reduce the inconsistencies in the diagnoses of several pathologists,^{26,27} which this study indeed confirms for the case of computer-aided diagnosis of NPC.

Moreover, the work efficiency and accuracy of pathologists can be significantly increased with the aid of the model. For some easily identifiable cases, the model can achieve high diagnostic accuracy thus freeing pathologists to focus on complex areas. In the process of manual diagnosis, detecting the region of interest and making a detailed diagnosis usually costs 2 to 3 days, whereas the model requires only a few seconds to give a definite result.

In order to demonstrate the advantages mentioned above, a more detailed statistical analysis was performed based on

the testing dataset. Figure 9 compares the results of the model with those of the pathologists. As shown in Figure 9A, in the comparison between the junior pathologist and the model, 82.4% of cases were diagnosed correctly by the pathologist; in the remaining 17.6% of cases that the junior pathologist diagnosed incorrectly, 82.4% were diagnosed correctly by the model. In the end, the junior pathologist and the model together diagnosed incorrectly only 1.80% of cases. For the case of the intermediate pathologist paired with the model, the percentage of cases that were diagnosed incorrectly was 1.80%. Meanwhile, for the senior pathologist paired with the model, the percentage of cases that were diagnosed incorrectly was 0.90%. These percentages are illustrated in Figure 9, B and C. It can therefore be concluded that the incidence of misdiagnosis dropped dramatically for all three pathologists when they were paired with the model.

This study also has some limitations, and the model presented in this paper still has room for improvement. First, the network structure, weighting strategy, and training methods can be improved to deal specifically with the characteristics of NPC WSIs. At present, the accuracy of digital pathologic image analysis of some other cancers has reached a high level. For example, in the study of Liu et al,²⁸ the AUC of breast cancer nodal metastasis diagnosis was 0.99. Work should therefore be done in automated algorithmic analysis of NPC to bring its AUCs to the same level. Secondly, the classification used in this study determines whether the tumor is cancerous or not but provides no further classification of the cancer subtype. In clinical practice, subtype diagnosis of NPC plays a decisive role in the selection of treatment methods and prognosis. In the diagnosis of non-small cell lung cancer, Coudray et al²⁹ have distinguished adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) by using deep learning with an average AUC of 0.97. Yan et al³⁰ accomplished a four-class classification task of breast cancer and obtained an accuracy of 91.3%, which outperformed the state-of-the-art method. Therefore, as a next step, we should try to achieve fine-grained classification to improve the clinical value of the model. Third, the sample selected in this study was very limited and cannot represent the whole population of NPC patients, so multicenter studies with more samples are needed to validate this model. Campanella et al³¹ evaluated their framework based on a dataset that contains 44,732 WSIs from 15,187 patients, and their results were even more convincing. Experiments based on a large amount of data can provide stronger confirmation of the robustness and generalizability of the model.

In conclusion, this paper has proved that a deep learning model has a diagnostic ability comparable to that of pathologists according to various performance metrics. Our model analyzes WSIs quantitatively and objectively to reduce the dependence on the pathologists' experience. The model's consistency is higher than that of relatively inexperienced pathologists. In clinical diagnosis, a deep learning

model can be used as an auxiliary tool to provide pathologists with a diagnostic reference, thus reducing the workload of pathologists and improving the efficiency and quality of clinical diagnoses.

References

1. Chua M, Wee J, Hui E, Chan A: Nasopharyngeal carcinoma. *Lancet* 2016, 387:1012–1024
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018, 68:394–424
3. Wei WI, Sham JST: Nasopharyngeal carcinoma. *Lancet* 2005, 365: 2041–2054
4. Li XH, Chang H, Xu BQ, Tao YL, Gao J, Chen C, Qu C, Zhou S, Liu SR, Wang XH, Zhang WW, Yang X, Zhou SL, Xia YF: An inflammatory biomarker-based nomogram to predict prognosis of patients with nasopharyngeal carcinoma: an analysis of a prospective study. *Cancer Med* 2017, 6:310–319
5. Ai Q, King A, Chan J, Chen W, Chan K, Woo J, Zee B, Chan A, Poon D, Ma B, Hui E, Ahuja A, Vlantis A, Yuan J: Distinguishing early-stage nasopharyngeal carcinoma from benign hyperplasia using intravoxel incoherent motion diffusion-weighted MRI. *Eur Radiol* 2019, 29:5627–5634
6. Luo WR, Fang W, Li S, Yao K: Aberrant expression of nuclear vimentin and related epithelial–mesenchymal transition markers in nasopharyngeal carcinoma. *Int J Cancer* 2012, 131:1863–1873
7. Luo W, Chen XY, Li SY, Wu AB, Yao KT: Neoplastic spindle cells in nasopharyngeal carcinoma show features of epithelial–mesenchymal transition. *Histopathology* 2012, 61:113–122
8. Luo W, Yao K: Molecular characterization and clinical implications of spindle cells in nasopharyngeal carcinoma: a novel molecule-morphology model of tumor progression proposed. *PLoS One* 2013, 8:e83135
9. Snead DRJ, Tsang Y-W, Meskiri A, Kimani PK, Crossman R, Rajpoot NM, Blessing E, Chen K, Gopalakrishnan K, Matthews P, Momtahn N, Read-Jones S, Sah S, Simmons E, Sinha B, Suortamo S, Yeo Y, El Daly H, Cree IA: Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016, 68: 1063–1072
10. Laurinavicius A, Laurinaviciene A, Dasevicius D, Elie N, Plancoulaine B, Bor C, Herlin P: Digital image analysis in pathology: benefits and obligation. *Anal Cell Pathol (Amst)* 2012, 35:75–78
11. Komura D, Ishikawa S: Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J* 2018, 16:34–42
12. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew M: Deep learning for visual understanding: a review. *Neurocomputing* 2016, 187:27–48
13. Tizhoosh HR, Pantanowitz L: Artificial intelligence and digital pathology: challenges and opportunities. *J Pathol Inform* 2018, 9:38
14. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI: A survey on deep learning in medical image analysis. *Med Image Anal* 2017, 42:60–88
15. El-Naggar AK, Chan JKC, Grandis JR, Takata T, Slootweg PJ (Eds): WHO Classification of Head and Neck Tumors. 4th ed. Geneva, Switzerland: World Health Organization, 2017
16. Diao S, Luo W, Hou J, Yu H, Chen Y, Xiong J, Xie Y, Qin W: Computer aided cancer regions detection of hepatocellular carcinoma in whole-slide pathological images based on deep learning. 2019 International Conference on Medical Imaging Physics and Engineering (ICMIPE) 2019:1–6
17. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: Rethinking the inception architecture for computer vision. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016:2818–2826

18. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Li F-F: ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015, 115: 211–252
19. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977, 33:159
20. Mohammed MA, Ghani MKA, Hamed RI, Ibrahim DA: Analysis of an electronic methods for nasopharyngeal carcinoma: prevalence, diagnosis, challenges and technologies. *J Comput Sci* 2017, 21:241–254
21. King AD, Vlantis AC, Yuen TWC, Law BKH, Bhatia KS, Zee BCY, Woo JSK, Chan ATC, Chan KCA, Ahuja AT: Detection of nasopharyngeal carcinoma by MR imaging: diagnostic accuracy of MRI compared with endoscopy and endoscopic biopsy based on long-term follow-up. *AJNR Am J Neuroradiol* 2015, 36:2380–2385
22. Wei J, Pei S, Zhu X: Comparison of (18)F-FDG PET/CT, MRI and SPECT in the diagnosis of local residual/recurrent nasopharyngeal carcinoma: a meta-analysis. *Oral Oncol* 2016, 52:11–17
23. Jiang W, Cai R, Chen Q-Q: DNA methylation biomarkers for nasopharyngeal carcinoma: diagnostic and prognostic tools. *Asian Pac J Cancer Prev* 2015, 16:8059–8065
24. Jia W, Ren C, Wang L, Zhu B, Jia W, Gao M, Zeng F, Zeng L, Xia X, Zhang X, Fu T, Li S, Du C, Jiang X, Chen Y, Tan W, Zhao Z, Liu W: CD109 is identified as a potential nasopharyngeal carcinoma biomarker using aptamer selected by cell-SELEX. *Oncotarget* 2016, 7: 55328–55342
25. Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, Schadendorf D, Sondermann W, Franklin C, Bestvater F, Flaig MJ, Krah D, von Kalle C, Fröhling S, Brinker TJ: Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur J Cancer* 2019, 118:91–96
26. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A: Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019, 16:703–715
27. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, van der Laak J, Hulsbergen-van de Kaa C, Litjens G: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020, 21: 233–241
28. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, Olson N, Peng LH, Hipp JD, Stumpe MC: Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Arch Pathol Lab Med* 2019, 143: 859–868
29. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018, 24:1559–1567
30. Yan R, Ren F, Wang Z, Wang L, Zhang T, Liu Y, Rao X, Zheng C, Zhang F: Breast cancer histopathological image classification using a hybrid deep neural network. *Methods* 2020, 173:52–60
31. Campanella G, Hanna MG, Geneslaw L, Mirafior A, Silva VWK, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019, 25:1301–1309