Journal of Dentistry Indonesia

Volume 30 | Number 1 Article 10

4-28-2023

Head and Neck Tumor Histopathological Image Representation with Pre- Trained Convolutional Neural Network and Vision Transformer

Ranny Rahaningrum Herdiantoputri
Department of Oral Pathology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), Tokyo, Japan, rannyrh.mpa@tmd.ac.jp

Daisuke Komura
Department of Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, kdais-prm@m.u-tokyo.ac.jp

Tohru Ikeda
Department of Oral Pathology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), Tokyo, Japan, tohrupth.mpa@tmd.ac.jp

Shumpei Ishikawa
Department of Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, shum-phm@md.u-tokyo.ac.jp

Recommended Citation

Herdiantoputri, R. R., Komura, D., Ikeda, T., & Ishikawa, S. Head and Neck Tumor Histopathological Image Representation with Pre- Trained Convolutional Neural Network and Vision Transformer. J Dent Indones. 2023;30(1): 41-47

Journal of Dentistry Indonesia 2023, Vol. 30, No. 1, 41-47
doi: 10.14693/jdi.v30i1.1501

ORIGINAL ARTICLE

Head and Neck Tumor Histopathological Image Representation with Pre-Trained Convolutional Neural Network and Vision Transformer

Ranny Rahaningrum Herdiantoputri[1], Daisuke Komura[2*], Tohru Ikeda[1], Shumpei Ishikawa[2*]

[1]Department of Oral Pathology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), Tokyo, Japan
[2]Department of Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
**Correspondence e-mail to: ishum-prm@m.u-tokyo.ac.jp and kdais-prm@m.u-tokyo.ac.jp*

ABSTRACT

Image representation via machine learning is an approach to quantitatively represent histopathological images of head and neck tumors for future applications of artificial intelligence-assisted pathological diagnosis systems. **Objective:** This study compares image representations produced by a pre-trained convolutional neural network (VGG16) to those produced by a vision transformer (ViT-L/14) in terms of the classification performance of head and neck tumors. **Methods:** Whole-slide images of five oral tumor categories (n = 319 cases) were analyzed. Image patches were created from manually annotated regions at 4096, 2048, and 1024 pixels and rescaled to 256 pixels. Image representations were classified by logistic regression or multiclass Support Vector Machines for binary or multiclass classifications, respectively. **Results:** VGG16 with 1024 pixels performed best for benign and malignant salivary gland tumors (BSGT and MSGT) (F1 = 0.703 and 0.803). VGG16 outperformed ViT for BSGT and MSGT with all magnification levels. However, ViT outperformed VGG16 for maxillofacial bone tumors (MBTs), odontogenic cysts (OCs), and odontogenic tumors (OTs) with all magnification levels (F1 = 0.780; 0.874; 0.751). **Conclusion:** Being more texture-biased, VGG16 performs better in representing BSGT and MSGT in high magnification while the more shape-biased ViT-L/14 performs better in representing MBT, OC, and OT.

Key words: artificial intelligence, digital pathology, head and neck tumors

How to cite this article: Herdiantoputri RR, Komura D, Ikeda T, Ishikawa S. Head and neck tumor histopathological image representation with pre-trained convolutional neural network and vision transformer. J Dent Indones. 2023;30(1): 41-47

INTRODUCTION

A histopathological analysis is crucial for tumor diagnosis and studying oral disease pathophysiology. This process involves staining tissue specimens with Hematoxylin & Eosin and carefully examining them under a microscope. However, this observation tends to be time-consuming and subjective except for highly experienced pathologists. Therefore, developing a diagnostic method to help accelerate diagnosis and objectively classify tumor cases is necessary, especially in the presence of a limited number of pathologists. To develop such a method, it is important to make histopathological images comparable to one another by quantifying the features within the image. Producing image representations using machine learning helps to quantitatively represent histopathological images of head and neck tumors, which can be valuable for future applications of artificial intelligence(AI)-assisted histopathological diagnosis systems.[1,2]

Technological advancements in digital images have facilitated the objective quantification of histopathological images beyond image recognition.[3–6] Encoding the features of histopathological images with image representation produced by a bilinear convolutional neural network (B-CNN) has successfully captured the histological characteristics of many types of cancer.[2] The bilinear operation takes the spatial invariance of histopathological images into account. It means that the B-CNN considers the relative locations of cellular components in accordance with other components instead of their relative positions within the images.[7,8] However, CNN

exhibits a strong texture bias, which means the network extracts information about an image from its local texture more than it does from its global shape.[9] CNN may perform well in representing histopathological images that comprise multiple object repetitions or contain spatially stationary statistics, which can be described as texture. This may not be the case when histopathological images exhibit elements of contours that describe the form of an object or elements that have non-stationary statistics, which can be described as shapes. In such instances, a vision transformer (ViT) may perform better. ViT has been demonstrated to be able to capture long-distance information and is less sensitive to local texture information than CNN. In other words, it exhibits a stronger shape bias than CNN.[10] Considering the variety of tumor types and the surrounding tissue of the tumor in the head and neck area, ViT provides a promising possibility to create a better image representation for the histopathological image of the oral region.

Therefore, this study aims to compare image representation produced by pre-trained B-CNN to that produced by pre-trained ViT with different magnification levels. We chose VGG16 to represent B-CNN due to its previous success as a histopathological image classifier in several studies[11–13] and ViT-L/14 to represent ViT due to its recent success as a large-scale image encoder.[14] The results presented in this study provide the insight required to move forward with the development of AI-assisted histopathological diagnostic systems.

METHODS

This study was approved by the Research Ethical Review Committee of Tokyo Medical and Dental University Hospital (Reference Number: D2019-087). We collected histopathology slides of the following five oral tumor categories: benign salivary gland tumor (BSGT), malignant salivary gland tumor (MSGT), maxillofacial bone tumor (MBT), odontogenic cyst (OC), and odontogenic tumor (OT) from the Pathology Division of TMDU Hospital (n = 319 cases). The cases included in these categories were the most common cases that we encountered from 2001 to 2021 to ensure the quality of the tissue staining and the final image (Table 1).

Table 1. Cases included in the analysis throughout this study.

Category	Diagnosis	Number of cases
Benign_ salivary_gland_ tumor	Pleomorphic_adenoma	16
	Myoepithelioma	12
	Basal_cell_adenoma	10
	Warthin_tumour	11
Malignant_ salivary_gland_ tumor	Mucoepidermoid_ carcinoma	10
	Adenoid_cystic_ carcinoma	10
	Basal_cell_ adenocarcinoma	10
	Carcinoma_ex_ pleomorphic_adenoma	10
	Salivary_duct_ carcinoma	15
Maxillofacial_ bone_tumor	Osteosarcoma	12
	Ossifying_fibroma	15
	Fibrous_dysplasia	13
	Cemento_osseous_ dysplasia	12
Odontogenic_ cyst	Radicular_cyst	19
	Inflammatory_ collateral_cyst	11
	Dentigerous_cyst	21
	Odontogenic_keratocyst	14
	Calcifying_ odontogenic_cyst	11
	Orthokeratinized_ odontogenic_cyst	11
Odontogenic_ tumor	Ameloblastoma	25
	Adenomatoid_ odontogenic_tumor	10
	Ameloblastic_fibroma	10
	Odontogenic_fibroma	11
	Odontogenic_myxoma_ myxofibroma	10
	Cemento_ossifying_ fibroma	10

Image preprocessing

Histopathology slides were scanned at 40× magnification using a NanoZoomer S210 Digital slide scanner, (C13239-01, Hamamatsu Photonics, Hamamatsu, Japan) to create whole-slide images (WSIs). We then chose and annotated the regions of interest (ROIs) that represent the area of the tumor, excluding the normal tissue and areas containing artifacts (e.g., torn or folded tissue). Strictly from the ROIs, square image patches of 4096, 2048, and 1024 pixels (905, 453, and 226 μm) were extracted in a random orientation to ensure rotational invariance. Image patch size represents the magnification level of the WSI: a smaller image means a higher magnification level. All image patches were resized into 256 pixels.

Supervised learning

The image representation was calculated using VGG16 block4_conv3 layer and ViT-L/14 (Figure 1). These networks were trained to recognize and predict the classification of images before their usage in our study: VGG16 was pre-trained on 1.2 million images from ImageNet in Tensorflow[15] and ViT-L/14 was pre-trained on 400 million image-text pairs in PyTorch.[14] For calculations with ViT-L/14, the image patches were further resized to 336 pixels. To calculate the image representation, image patches were automatically

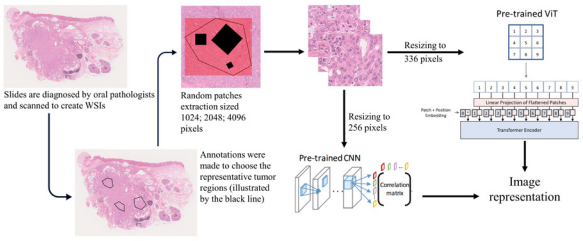

Figure 1. Production of a histopathology image representation.

broken down into pixels. Each pixel was passed through the layers of the network, which contain pre-defined weights and act as filters to extract the features from the simple ones (such as dots and lines) to the more complex ones (such as squares and circles). By going through these layers, the original image patches were transformed into sets of feature values that we refer to as image representation.[16] From the image representation, models for multiclass classification were created using multiclass Support Vector Machines (SVM). All image data were divided into training and test sets in a 3:1 ratio. The performance of each category was assessed using precision, recall, and the F1 score at the case level.

Precision = True Positive/(True Positive + False Positive)
Recall = True Positive/(True Positive + False Negative)
F1 score = 2x (Precision x Recall)/(Precision + Recall)
The F1 score represents the balance between precision and recall. Therefore, the higher the F1 score, the more distinguishable one category is from the others.[17] By comparing the F1 scores of images represented by VGG16 and ViT-L/14, we were able to conclude, which network performs better for the representation of each tumor category.

RESULTS

Our supervised learning models divide the image representations into the original five categories (Figure 2). MSGT has been shown to be the most distinguishable by VGG16, especially for 1024-pixel and 2048-pixel images (Figures 2A and 2B), while BSGT is hardly distinguishable from MSGT, especially for 2048-pixel images (167 images are correctly classified into BSGT and 115 images are classified into MSGT) (Figure 2B). Generally, OT is hardly distinguishable from MBT when using VGG16 (Figures 2A–2C). On the other hand, BSGT is shown to be indistinguishable from MSGT by ViT-L/14 (Figures 2D–2F). From the calculated F1 score, the highest performance of VGG16 in representing BSGT and MSGT is achieved with 1024-pixel images (F1 = 0.703 and 0.803, respectively). The lowest F1 scores of VGG16 for BSGT and MSGT (F1 = 0.565 and 0.696)

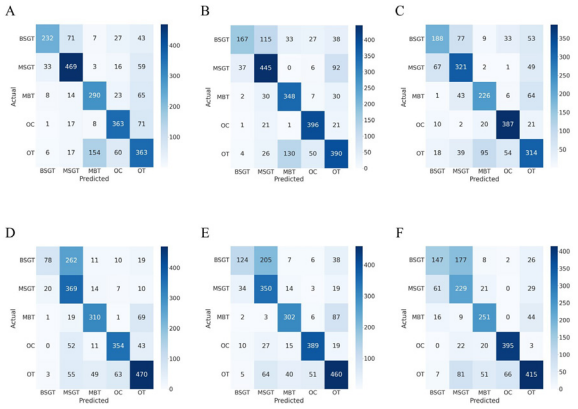

Figure 2. Confusion matrices of supervised learning models. The numbers in the matrix represent the number of images classified in each category. (A) Image representation from 1024-pixel image patches generated by VGG16. (B) Image representation from 2048-pixel image patches generated by VGG16. (C) Image representation from 4096-pixel image patches generated by VGG16. (D) Image representation from 1024-pixel image patches generated by ViT-L/14. (E) Image representation from 2048-pixel image patches generated by ViT-L/14. (F) Image representation from 4096-pixel image patches generated by ViT-L/14.

are higher than those of ViT-L/14 for both categories with all magnification levels. The highest F1 scores of VGG16 for MBTs, OCs, and OTs are achieved with 2048 pixels (F1 = 0.749; 0.855, and 0.666). However, these scores are lower than those of ViT-L/14 for all three categories with all magnification levels. The highest F1 score of ViT-L/14 for MBT is 0.780; that for OC is 0.87, and that for OT is 0.751. We found that the performances of VGG16 and ViT-L/14 may differ according to the characteristics of each tumor category (Table 2).

The scores in bold font represent the best F1 scores for each magnification level in the category. The scores in bold font with asterisks (*) represent the best F1 score across all magnification levels in the category (BSGT: benign salivary gland tumor; MSGT: malignant salivary gland tumor; MBT: maxillofacial bone tumor; OC: odontogenic cyst; OT: odontogenic tumor).

DISCUSSION

Our study reports the performance of machine learning models in representing the most diverse category of head and neck tumors to date, including salivary gland tumors, OTs, and OCs, at the same time. However, in computational pathology, most supervised learning models have been developed using conventional methods, including feature extraction, data training, testing, and the prediction of image classification. Sakamoto et al. reported that fine-tuning VGG16 with image patches of odontogenic keratocysts, radicular cysts, and stroma successfully produced a model to predict odontogenic keratocysts or other OCs with

Table 2. The performances of VGG16 and ViT-L/14 for five tumor categories at different magnification levels.

VGG16	1024 px			2048 px			4096 px		
	Precision	**Recall**	**F1_score**	**Precision**	**Recall**	**F1_score**	**Precision**	**Recall**	**F1_score**
BSGT	0.829	0.610	**0.703***	0.791	0.439	**0.565**	0.662	0.522	**0.584**
MSGT	0.798	0.809	**0.803***	0.699	0.767	**0.731**	0.666	0.729	**0.696**
MBT	0.628	0.725	0.673	0.680	0.834	0.749	0.642	0.665	0.653
OC	0.742	0.789	0.765	0.815	0.900	**0.855**	0.805	0.879	0.840
OT	0.604	0.605	0.604	0.683	0.650	0.666	0.627	0.604	0.615
ViT-L/14	**1024 px**			**2048 px**			**4096 px**		
	Precision	**Recall**	**F1_score**	**Precision**	**Recall**	**F1_score**	**Precision**	**Recall**	**F1_score**
BSGT	0.765	0.205	0.324	0.709	0.326	0.447	0.636	0.408	0.497
MSGT	0.487	0.879	0.627	0.539	0.833	0.655	0.442	0.673	0.534
MBT	0.785	0.775	**0.780***	0.799	0.755	**0.776**	0.715	0.784	**0.748**
OC	0.814	0.770	**0.791**	0.855	0.846	0.850	0.853	0.898	**0.875***
OT	0.769	0.734	**0.751***	0.738	0.742	**0.740**	0.803	0.669	**0.730**

an AUC of 0.997.[11] A study by Liu et al. employed a similar method by fine-tuning VGG16 with image data augmentation, image denoising, and sharpening of breast cancer histopathology images to predict benign or malignant breast cancer cases, successfully achieving 98.89% accuracy.[12] Although this method guarantees high-accuracy predictions between limited classes, much data is required to train and test the network.

Head and neck tumors, especially those in the oral region, are not always common. With a small amount of data, generating image representations with a pre-trained network as we did in our study reduces the possibility of overfittingg.[2,16] This approach is termed "few-shot learning" or, more specifically, "few-shot image classification." We used pre-trained networks to automatize feature extraction, then incorporated a machine learning algorithm (SVM) to learn the association between the pre-defined image features and the tumor category. Once the image representations are generated, they can be correlated to any biological and clinical parameters.[6,16] The image representation has been used to objectively analyze histopathological images for purposes other than diagnosis. Xu et al. reported how image representation generated by CNN helps pathologists discover patterns in colon cancer and brain tumors with biological insights.[18] Another study by Komura et al. demonstrated that image representation is useful for histological profiling and subtyping and correlates morphological features with genomic mutations with examples of more than 30 types of cancer.[2]

Our results demonstrated that pre-trained networks can create objective and comparable oral histopathological image representation; however, the ability to do so depends on the properties of the network itself. There are some differences that are worth noticing between the networks that we used in our study. The input size for VGG16 is 256 pixels, while that for ViT-L/14 is 336 pixels. Moreover, VGG16 was pre-trained with approximately 1.2 million images, whereas ViT-L/14 was pre-trained with 400 million image-text pairs. These differences may account for the difference in performance that was presented in the previous section.

Our dataset includes histopathological images with a wide range of visual characteristics. Solid tumors in BSGT and MSGT, such as myoepithelioma, warthin tumors, and basal cell adenocarcinoma, exhibit texture-like features of tumor cells that are arranged in a repetitive manner (Figure 3). MBT and OT include tumors in the jawbone, such as fibrous dysplasia, cemento-osseous dysplasia, and cemento-ossifying fibroma. These categories may include mineralized tissues that exhibit defined contours. OCs, while rarely including mineralized tissue, are differentiated from other types of head and neck tumors and from one another mostly by the lining epithelium, which forms a contour-like feature along the image, separating the lumen area and the connective tissue (Figure 4). The higher F1 scores showed by VGG16 in representing BSGT and MSGT may be associated with its high texture bias. This hypothesis is supported by the fact that the image representation produced by VGG16 of these categories that are correctly classified exhibit texture-like features (Figure 5). In their study, Geirhos et al. reported that after testing CNN's recognition accuracy for different image styles in comparison to humans, both can recognize images with their original texture and shape equally well. However, when the

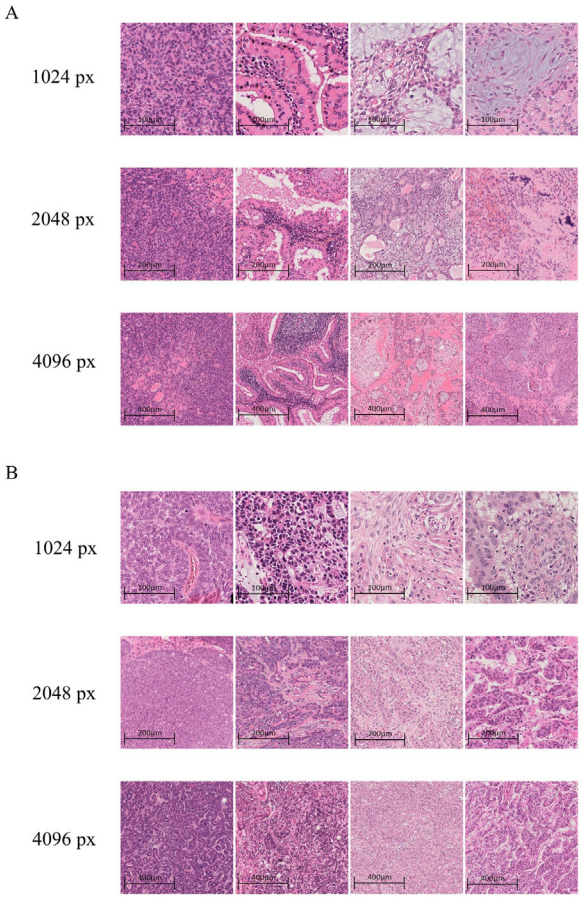

Figure 3. Images correctly classified by VGG16. Solid tumors generally show texture-like features from two categories: (A) BSGT and (B) MSGT.

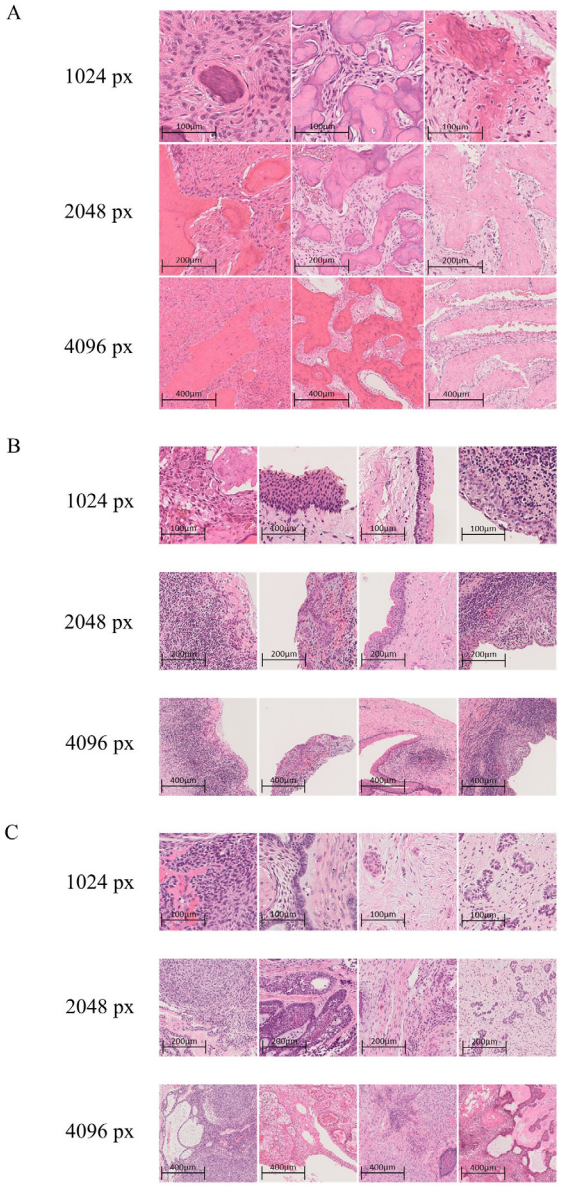

Figure 4. Images correctly classified by ViT-L/14. Images exhibiting definitive contour-like features separating different types of tissue. (A) MBT, (B) OC, and (C) OT. Scale bars: 100 µm for 1024-pixel images, 200 µm for 2048-pixel images, 400 µm for 4096-pixel image.

object was outlined to create a more definitive shape, CNN's recognition accuracy decreased (9). With higher magnification, the texture-like features of BSGT and MSGT become more homogenous, which may result in a better representation with VGG16. On the other hand, ViT-L/14 showed higher performance for MBT, OT, and OC, which may be associated with ViT's larger receptive field, resulting in its stronger shape bias, which is more similar to the human visual ability than that of CNN. This hypothesis is supported by the fact that the image representation produced by ViT-L/14 of these categories that are correctly classified do not exhibit texture-like features but rather exhibit more definitive contours (Figure 5). Our report shows that VGG16 generally classifies images based on the texture-like features seen throughout the image, while ViT-L/14 gives importance to definitive contours, which is in line with the results of the study by Naseer et al., with a similar analysis done by Geirhos et al., providing evidence that ViT gives higher importance to global shape than CNN, which exploits local textures to recognize an image.[10] More studies are needed to find the middle ground between these two types of bias for further use in computational pathology.

Despite the novelty of our findings, our study focuses on the machine-learning representation of an image patch, which makes its findings highly dependent on the manually selected ROIs. Furthermore, our dataset only included tumor cases that are commonly found in Japan. A multicenter cohort study including rare cases and cases with rare subtypes in the dataset will better ensure the usefulness of the model in the clinical setting.

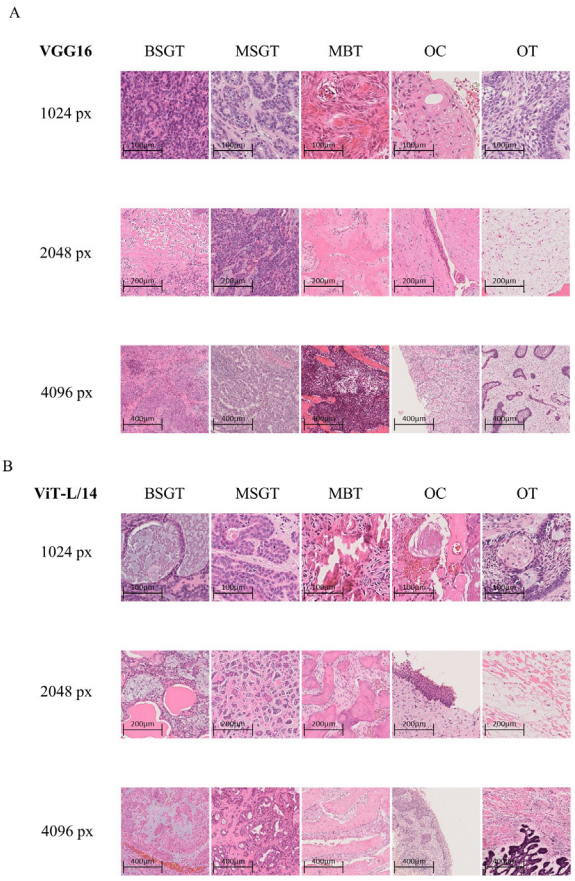

Figure 5. Comparisons of correctly classified images in both networks for each category. (A) Images correctly classified by VGG16 for each category are compared directly with (B) images correctly classified by ViT-L/14 of the same category and diagnosis. Scale bars: 100 µm for 1024-pixel images, 200 µm for 2048-pixel images, 400 µm for 4096-pixel image.

CONCLUSION

VGG16, a CNN known to have a high texture bias, performs better in representing BSGT and MSGT in high magnification, while ViT-L/14, which has the ability to encode contextual cues and has stronger shape bias than VGG16, outperforms it in representing MBT, OC, and OT.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. Elazab N, Soliman H, El-Sappagh S, Islam SMR, Elmogy M. Objective diagnosis for histopathological images based on machine learning techniques: Classical approaches and new trends. Mathematics. 2020; 8(11):1863.
2. Komura D, Kawabe A, Fukuta K, Sano K, Umezaki T, Koda H, Suzuki R, Tominaga K, Ochi M, Konishi H, Masakado F, Saito N, Sato Y, Onoyama T, Nishida S, Furuya G, Katoh H, Yamashita H, Kakimi K, Seto Y, Ushiku T, Fukayama M, Ishikawa S. Universal encoding of pan-cancer histology by deep texture representations. Cell Rep. 2022; 38(9):110424.
3. Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, Fei B. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Sci Rep. 2019; 9(1):14043.
4. Lou P, Wang C, Guo R, Yao L, Zhang G, Yang J, Yuan Y, Dong Y, Gao Z, Gong T, Li C. HistoML, a markup language for representation and exchange of histopathological features in pathology images. Sci Data. 2022; 9(1):387.
5. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F, Olson N, Smith JL, Mohtashamian A, Wren JH, Corrado GS, MacDonald R, Peng LH, Amin MB, Evans AJ, Sangoi AR, Mermel CH, Hipp JD, Stumpe MC. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med. 2019; 2:48.
6. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology. Nat Cancer. 2022; 3(9):1026-38.
7. Lin TY, Maji S. Visualizing and Understanding Deep Texture Representations [Internet]. arXiv; 2016 [cited 2022 Oct 12]. Available from: http://arxiv.org/abs/1511.05197.
8. Xu Y, Li F, Chen Z, Liang J, Quan Y. Encoding Spatial Distribution of Convolutional Features for Texture Representation. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2021. p. 22732-44.
9. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness [Internet]. arXiv; 2019 [cited 2022 Oct 12]. Available from: http://arxiv.org/abs/1811.12231.
10. Naseer M, Ranasinghe K, Khan S, Hayat M, Khan FS, Yang MH. Intriguing Properties of Vision Transformers [Internet]. arXiv; 2021 [cited 2022 Oct 12]. Available from: http://arxiv.org/abs/2105.10497.
11. akamoto K. Morita K ichi, Ikeda T, Kayamori K. Deep-learning-based identification of odontogenic keratocysts in hematoxylin- and eosin-stained jaw cyst specimens [Internet]. arXiv; 2019 [cited

2022 Oct 12]. Available from: http://arxiv.org/abs/1901.03857.
12. Liu M, Yi M, Wu M, Wang J, He Y. Breast pathological image classification based on VGG16 feature concatenation. J Shanghai Jiaotong Univ (Sci). 2022; 27(4):473-84.
13. Zhou P, Cao Y, Li M, Ma Y, Chen C, Gan X, Wu J, Lv X, Chen C. HCCANet: Histopathological image grading of colorectal cancer using CNN based on multichannel fusion attention mechanism. Sci Rep. 2022; 12(1):15103.
14. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S. Learning Transferable Visual Models From Natural Language Supervision [Internet]. arXiv; 2021 [cited 2022 Oct 12]. Available from: http://arxiv.org/abs/2103.00020.
15. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014 [cited 2022 Oct 17]; Available from: https://arxiv.org/abs/1409.1556.
16. Chowdhury A, Jiang M, Chaudhuri S, Jermaine C. Few-shot Image Classification: Just Use a Library of Pre-trained Feature Extractors and a Simple Classifier. 2021 [cited 2022 Oct 17]; Available from: https://arxiv.org/abs/2101.00562.
17. Seo S, Kim Y, Han HJ, Son WC, Hong ZY, Sohn I, Shim J, Hwang C. Predicting successes and failures of clinical trials with outer product-based convolutional neural network. Front Pharmacol. 2021; 12:670670.
18. Xu Y, Jia Z, Wang LB, Ai Y, Zhang F, Lai M, Chang EI. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics. 2017; 18(1):281.

(Received January 22, 2023; Accepted March 28, 2023)