CPNet: A Hybrid Neural Network for Identification of Carcinoma Pathological Slices

Runwei Guan*†, Yanhua Fei‡, Xiaohui Zhu†, Shanliang Yao*†, Yong Yue†, Jieming Ma†
*Faculty of Science and Engineering, University of Liverpool, Liverpool, United Kingdom
Email: {Runwei.Guan, S.Yao17}@liverpool.ac.uk
†School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China
Email: {xiaohui.zhu, yong.yue, jieming.ma}@xjtlu.edu.cn
‡Department of Oncology, The Affiliated Jiangyin Hospital of Southeast University Medical College, Wuxi, China
Email: gyqjy68@sohu.com

***Abstract*—In the medical field, pathological carcinoma images look much more complicated than other medical images. Identifying carcinoma pathology images is a time-consuming and error-prone task for regular doctors and even for some specialists. Nowadays, deep learning has been widely applied in medicine, which could significantly reduce the time cost and improve accuracy. To save time and improve the accuracy of identifying pathological carcinoma slices, we propose a novel ViT-CNN hybrid neural network called CPNet, specially for the classification of different categories of carcinoma pathological slices. CPNet achieves the state-of-the-art performance in *PatchCamelyon* and our own dataset. We adopt a transfer learning method to identify degrees of malignancy using a few samples. Furthermore, we design and develop a fast medical decision system, where we deploy the CPNet in it. The system could effectively assist doctors in identifying the cancer pathology images with high accuracy and speed. The code of CPNet is in https://github.com/GuanRunwei/CPNet.**

***Index Terms*—intelligent medicine, pathological image identification, deep learning, CNN-ViT hybrid NN**

I. Introduction

In recent years, with the gradual maturity of image processing technology and the rapid development of machine learning, machine learning and deep learning technologies have begun to enter the medical field to replace medical staff to complete many difficult and time-consuming tasks [1].

As one of the diseases with the largest number of patients in the world, cancer has naturally attracted the attention of deep learning technology, and many phased results have been achieved so far [2].

Emine CENGIL et al. [3] proposed a method based on 3D convolutional neural network to identify lung CT images to diagnose lung cancer. Siddharth Bhatia et al. [4] proposed a method of using Resnet to extract image features, ensembling two models which are XGBoost and random forest for lung CT image diagnosis classification. A. Asuntha et al. [5] proposed a method that uses HoG, LBP, and SIFT for feature extraction, then uses FPSO for feature selection, and finally uses a novel FPSOCNN to classify lung CT images. Lakshmanaprabu S.K et al. [6] proposed a method for lung cancer image classification using ODNN for feature extraction and then using LDA model to classify. Zhihua Zhou et al. [7] proposed a method for identifying lung cancer cells based on integrated artificial neural networks.

Pathology denotes the process and principle of the occurrence and development of diseases [8]. That is, the causes of the disease and the changes in the structure, function and metabolism of cells, tissues, and organs during the disease process and their laws. Compared to other medical images such as CTs and MRNs, pathology images are more complicated and confusing in their structures and forms [9].

Coundray et al. [10] adopted Inception v3 to classify lung cancer pathology images. Wang et al. [11] used tranfer learning and CNN-based models to analyze the lung cancer pathology. AlZubaidi [12] used a CNN to extract features and used a series of machine learning models to classify different lung cancer pathology images. Luo et al. [13] proposes a statistics machine learning framework to analyze the lung cancer pathological images.

It could be seen that the methods of machine learning and deep learning are gradually mature. However, machine learning models need high-quality features extracted by pretrained neural networks, which have high cost and low generalization. For deep learning, CNNs are widely used, but CNNs could not model the global contextual feature due to the locality in its inductive bias. It means CNNs could not learn and model the potential connection and correlation of features. Moreover, no matter ML models or CNNs, they both have weak robustness, which means they are sensitive to noises [14].

Vision transformer(ViT) [15] is robust to noises and has high generalization. However, ViTs discard the inductive bias in CNNs and have large number of parameters, which are highly time-consuming for training and inference.

In this paper, we focus on the identification of 3 most common cancers and their pathology images, including lung cancer, liver cancer and colon cancer. Furthermore, we research combining CNN with ViT, which could accelerate training and inference compared with the pure ViT. Our contributions are summarized as

- We propose a ViT-CNN hybrid neural network called

This research was supported by the Suzhou Science and Technology Project (SYG202122), the Key Program Special Fund of Xi'an Jiaotong-Liverpool University (XJTLU) (KSF-A-19, KSF-E-65, KSF-P-02, KSF-E-54) and the Research Development Fund of XJTLU (RDF-19-02-23).

978-1-6654-6733-9/22/$31.00 ©2022 IEEE

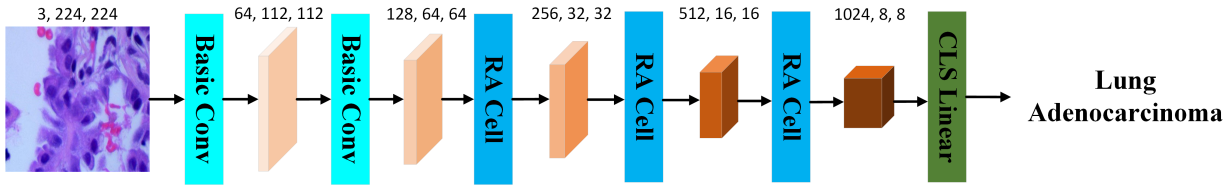

Fig. 1. The structure of CPNet. A CPNet includes 2 Basic Conv blocks, 3 RA Cells and 1 CLS Linear block.

CPNet, which is specifically for classification of cancer pathology images. CPNet achieves 99.8% accuracy in *PatchCamelyon* and 99.2% accuracy in our base dataset.
- We propose a transfer learning method to transfer the trained CPNet to train in a few labeled samples for identification of malignancy degrees. We get 92.8% accuracy with this transfer learning method.
- We design and develop a fast diagnosis system, assisting physicians in decision.

II. Related Work

A. ViT-CNN Hybrid Neural Network

There are many NNs combining CNN with ViT, which could introduce the inductive bias to ViT. It could speed up the training and inference. ViTs can calculate the feature similarity and model the global context, which is the weakness of CNN due to its locality. Moreover, CNNs are high-pass filters while ViTs are low-pass filters [16]. It means these two filters are complementary. ViTAE [17] concatenates CNN in the residual side of the vision transformer block. BoTNet [18] replaces one basic convolution module with a multi-head self-attention(MHSA) module in a residual block. Visformer [19] combines MHSAs and CNNs in each stage with much less FLOPs than pure ViTs.

B. Transfer Learning

Deep learning needs a large amount of data. In some fields, the labelled data is few for the high labelling cost. Transfer learning could let the model converge earlier in the training process, with the help of pretraining the model in similar data fields. The common transfer learning fields including network fine-tuning, few-shot learning, weak-shot learning, etc [20].

III. CPNet

For images of pathological carcinoma, different pathological carcinomas may have the similar feature in some local areas. Under these circumstances, we need to find an approach to model the global feature, which could dramatically reduce the error rate. Therefore, we propose CPNet as the backbone for the recognition of pathological carcinoma images. The structure of CPNet is shown in Fig.1. CPNet consists of two basic convolution blocks (BC), three residual attention cells (RA) and one linear block for classification.

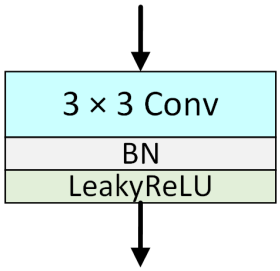

Fig. 2. Basic Convolution Block

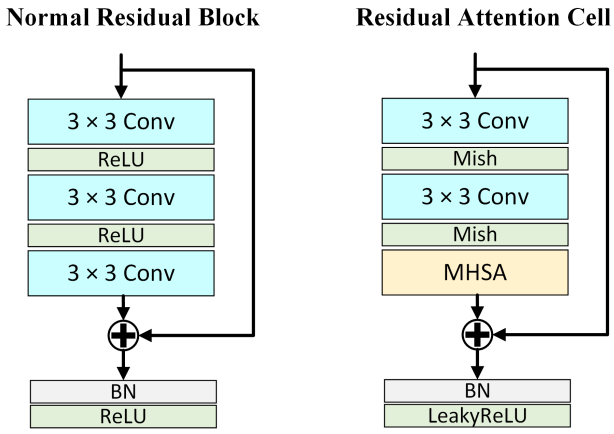

Fig. 3. Residual Attention Cell. The left one is a normal residual block in ResNet. The right one is a residual attention cell in CPNet proposed in this paper.

A. Basic Convolution

As Fig.2 shows, basic convolution (BC) is to extract the feature of image in the early stage and deepen the network. It consists of a convolution with 3 × 3 kernels, a batch-norm (BN) and a LeakyReLU for activation. Assuming the input map is $x_i \in R^{C\times W\times H}$, the operation of BC is show in Eq.1.

$$x_{i+1} = LeakyReLU(BN(Conv(x_i))), x_{i+1} \in R^{C\times \frac{W}{2}\times \frac{H}{2}} \tag{1}$$

B. Residual Attention Cell

Residual Attention Cell (RA Cell) is located after two BCs instead of directly used at the beginning. The reason can be concluded as:

- For the receptive field of shallow layers is small, the features that shallow layers extract almost have no semantic

information. It is meaningless to use MHSA at the early stage, which would cause over-fitting.

- Stacking BCs could deepen the network, making the feature from low-level to high-level, so MHSA in late stages could aggregate the high-level semantic feature.

As Fig.3 shows, compared with the normal residual block in ResNet [21], firstly, we replace ReLU with Mish [22] for activation. We empirically argue that ReLU would cause the death of partial neurons for its characteristics, which is adverse for training and inference some time, because we do not know whether the partial killed neurons are useful or not. We empirically argue that Mish has better generalization.

Secondly, Park et al. [16] argued that inserting the attention mechanism at the end of each stage can improve the predictive performance of NN, so we add MHSA after 2 convolution layers. Conventional vision transformer flatten the map as a sequence, so vision transformer needs to learn the position information by itself, which would prolong the training time. In RA Cell, we directly use the feature map extracted by convolution. Assuming there is a feature map $x \in R^{C\times W\times H}$, we make three copies of it as $x_q \in R^{C\times W\times H}, x_k \in R^{C\times W\times H}$ and $x_v \in R^{C\times W\times H}$. The generation process of self-attention map is shown in Fig.4 and Eq.2.

$$
\begin{aligned}
M_{Att} &= W^q x_q \otimes W^k x_k, M_{Att} \in R^{C\times W\times H} \\
M^{'}_{Att} &= M_{Att} \oplus RPE, M^{'}_{Att} \in R^{C\times W\times H} \\
M^{''}_{Att} &= softmax(\frac{M^{'}_{Att}}{\sqrt{d_k}}), M^{''}_{Att} \in R^{C\times W\times H} \\
x_{self-att} &= M^{''}_{Att} \otimes W^v x_v, x_{self-att} \in R^{C\times W\times H}
\end{aligned} \tag{2}
$$

where W^q, W^k and W^v are three learnable weights. $\otimes$ means element-wise multiply. $\oplus$ means element-wise addition. M_{Att} denotes the attention matrix. RPE denotes the relative position encoding, the position encoding will be changed according to the map's shape in different bottlenecks, which is more flexible and general than the absolute position encoding. $\sqrt{d_k}$ indicates the dimension of $M^{'}_{Att}$. $M^{''}_{Att}$ is the matrix of attention, which indicates the position needs to be payed attention to. MHSA stacks multiple $x_{self-att}$. Each $x_{self-att}$ pays attention to n channels. MHSA is shown in Eq.3.

$$
mhsa(x) = x^1_{self-att} \copyright x^2_{self-att} \cdots \copyright x^i_{self-att} \tag{3}
$$

where © denotes the stack operation.

Like the normal residual block, a RA Cell also consists of a residual side, which could alleviate the gradient explosion or vanishing. Assuming the input map is $x \in R^{C\times W\times H}$, the output of MHSA is $x_3 \in R^{C\times W\times H}$. The whole process is in Eq.4.

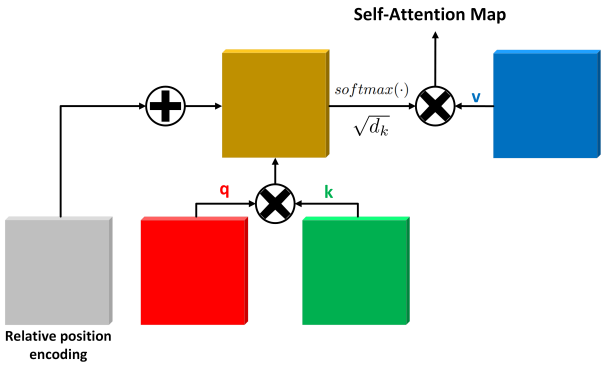

Fig. 4. Self Attention in Residual Attention Cell

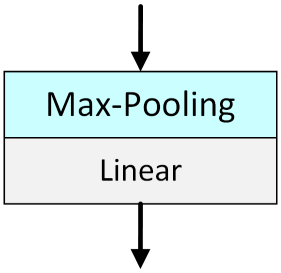

Fig. 5. CLS Linear Block

$$
\begin{aligned}
x_1 &= Mish(Conv(x)), x_1 \in R^{C\times W\times H} \\
x_2 &= Mish(Conv(x_1)), x_2 \in R^{C\times W\times H} \\
x_3 &= MHSA(x_2), x_3 \in R^{C\times W\times H} \\
x^{res}_3 &= x \oplus x_3, x^{res}_3 \in R^{C\times W\times H}
\end{aligned} \tag{4}
$$

Like the normal residual block in ResNet, there is a batch-norm and an activation function at the last stage. However, we replace ReLU with LeakyReLU to avoid the death of partial neurons. Given an input map $x \in R^{C\times W\times H}$, the process is shown in Eq.5.

$$
\begin{aligned}
x_1 &= BN(x), x_1 \in R^{C\times W\times H} \\
x_2 &= LeakyReLU(x_1), x_2 \in R^{C\times W\times H}
\end{aligned} \tag{5}
$$

C. CLS Linear

As Fig.5 shows, a CLS Linear block consists of a max-pooling operation for down-sampling and a linear layer for classification. Given an input map $x \in R^{C\times W\times H}$, the process is shown in Eq.6.

$$
\begin{aligned}
x_1 &= maxpool(x), x_1 \in R^{C\times \frac{W}{2}\times \frac{H}{2}} \\
y &= linear(x_1), y \in R^{cls\times 1\times 1}
\end{aligned} \tag{6}
$$

where cls denotes the number of categories. y is a one-dimension array for classification.

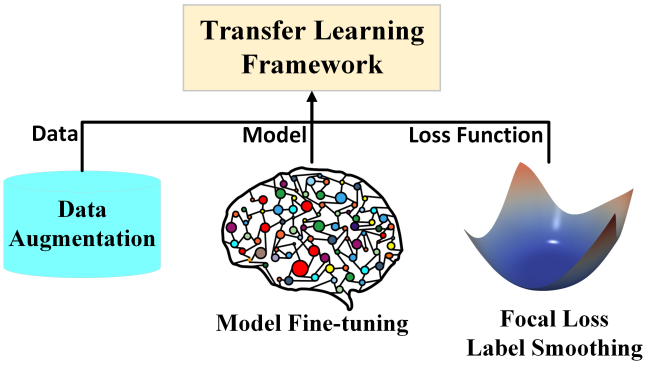

Fig. 6. Transfer Learning Framework for Malignancy Degree Identification

TABLE I
DATA AUGMENTATION METHODS

Method	IsUsed
Mixup [23]	✓
CenterCrop	✓
RandomHorizontalFlip	✓
RandomVerticalFlip	✓
RandomRotation	✓
GaussianBlur	✓
CenterCrop	✓
RandomPerspective	✓
ColorJitter	✓

To sum up, CPNet models features of locality by CNNs and global context by MHSAs, which enhances the feature representation of the image.

IV. TRANSFER LEARNING FOR MALIGNANCY DEGREE IDENTIFICATION

Labelling malignancy degrees of cancer pathology images is a time-consuming and tough job for doctors, but deep learning is a data-hungry game. Therefore, it is challenging and significant to use few training samples to train a qualified model. We propose a transfer learning framework to transfer the cancer-category-prediction model to malignancy-degree-prediction model. As Fig.6, we aggregate different approaches in our framework from perspectives of data, model and loss function.

A. Data

Since the number of training samples is limited, we need to use some augmentation methods to extend the dataset indirectly. Moreover, the augmentation could enhance the generalization and alleviate over-fitting.

The augmentation methods in the framework are shown in Table I.

Fig.7 visualizes the result of mixup.

B. Model

For the same training samples, we empirically argue that features learned in shallow layers could be shared in other downstream tasks. It means we could just train the parameters

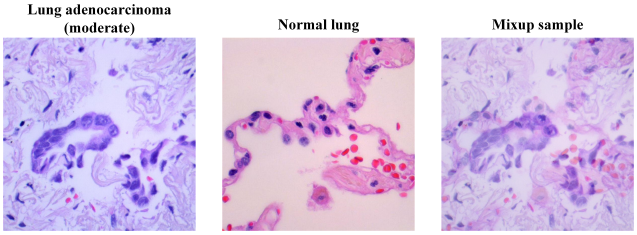

Fig. 7. Mixup

in last RA Cell and CLS Linear layer, for the parameters of other layers, we just freeze them. Thus, the training data we need would be much less than training a model from scratch. Fig.8 shows the fine-tuned model trained for malignancy degree identification.

C. Loss Function

To let the model converge as fast as possible, we combine focal loss [24] with label smoothing [25]. Label smoothing could alleviate the over-confidence due to few samples. Focal loss could enlarge the loss value of hard samples and minish the loss value of easy samples, which could make the model pay attention to hard samples. Eq.7 shows the label smoothing operation.

$$p_i = \frac{e^{\frac{x_i}{T}}}{\sum_{j=1}^{K} e^{\frac{x_j}{T}}}, \forall 1, 2, \ldots, K \tag{7}$$

where p_i is the normalized predicted value. x_i indicates the predicted value. x_j indicates all the values in the softmax matrix. T is a constant above 1.

Eq.8 shows the focal loss.

$$Focal(p_i) = -(1 - p_i)^{\gamma} log(p_i) \tag{8}$$

where we set γ to 2 empirically.

V. EXPERIMENTS

A. Classification of Pathological Carcinoma Images

1) Dataset: The dataset consists of two parts, the first one is *lung-and-colon-cancer-histopathological-images* [26] in Kaggle while the second is collected from clinical patients, totally 30 thousand images, 70% for training and 30% for test. We call it base dataset. The dataset has 8 categories, including lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon benign tissue, colon adenocarcinoma, liver benign tissue, liver adenocarcinoma and liver squamous cell carcinoma. Some samples are shown in Fig.9.

2) Implementation Details: Implementation details are shown in TABLE II. We train the model for 15 epochs with the batch size of 16 on one RTX 3060 GPU. We set the initial learning rate to 0.001 and use Adam as the optimizer, which is under the cosine scheduler. We set the weight decay to 5e-4.

We train CPNet as well as other state-of-the-art models in our base dataset. We select models whose parameter numbers are close to our CPNet. As TABLE III shows, CPNet performs

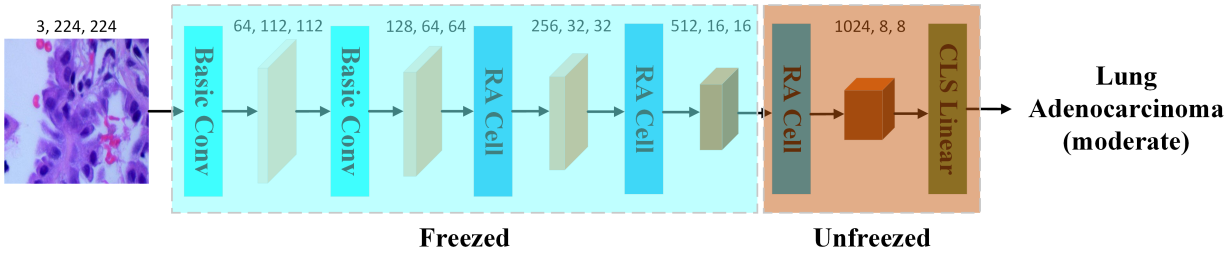

Fig. 8. Fine-tuned Model for Malignancy Degree Identification

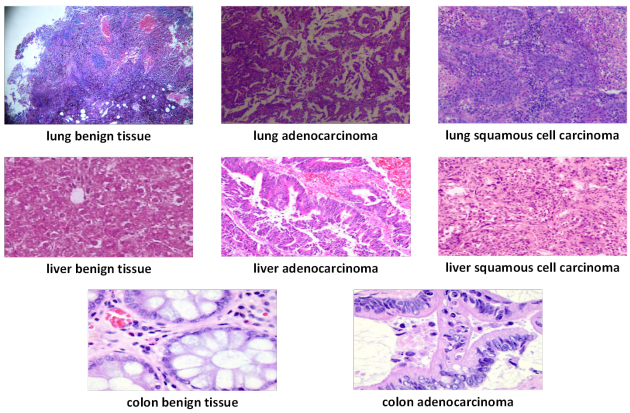

Fig. 9. Some samples in base dataset. It has eight classes totally, including lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon benign tissue, colon adenocarcinoma, liver benign tissue, liver adenocarcinoma and liver squamous cell carcinoma.

TABLE II
IMPLEMENTATION DETAILS OF CLASSIFICATION OF CANCER PATHOLOGY IMAGES

Epochs	Batch Size	Initial LR	Optimizer	Weight Decay
15	16	0.001	Adam	5e-4

better than other 4 models, which are 2 ViT-based and 2 CNN-based models.

Moreover, we also train and test CPNet in *PatchCamelyon* benchmark [30]. As TABLE IV shows, CPNet achieves the state-of-the-art results compared with other models. CPNet gets the accuracy of 99.8%, which is 0.7% higher than TNT-S and 1.1% higher than EfficientNet-b4.

TABLE III
COMPARISON OF DIFFERENT MODELS ON OUR BASE DATASET

Model	Params(M)	Accuracy(%)
CPNet (ours)	**24**	**99.2**
DeiT-S [27]	22	98.4
TNT-S [28]	24	98.1
EfficientNet-b4 [29]	19	97.8
ResNet-50 [21]	25	97.2

TABLE IV
COMPARISON OF DIFFERENT MODELS IN *PatchCamelyon* [30]

Model	Params(M)	Accuracy(%)
CPNet (ours)	**24**	**99.8**
TNT-S [28]	24	99.1
EfficientNet-b4 [29]	19	98.7
DeiT-S [27]	22	98.4
ResNet-50 [21]	25	96.1

TABLE V
IMPLEMENTATION DETAILS OF CLASSIFICATION OF MALIGNANCY DEGREE

Epochs	Batch Size	Initial LR	Optimizer	Weight Decay
5	8	0.005	Adam	5e-4

B. Transfer Learning for Malignancy Degree Identification

1) Dataset: We select 1200 images from the base dataset and label them into 3 categories: mild, moderate and severe. Among them, 1000 images for training and 200 images for test.

2) Implementation Details: Implementation details are shown in TABLE V. We use the model pretrained in the base dataset. We train the model for 5 epochs with the batch size of 8 on one RTX 3060 GPU. We set the initial learning rate to 0.005 and use Adam as the optimizer, which is under the cosine scheduler. We set the weight decay to 5e-4.

As TABLE VI shows, the model trained under the transfer learning framework has the accuracy of 92.8%, which is 11.2% higher than the model out of the transfer learning framework.

C. Diagnosis System

We design and develop a diagnosis system based on .Net Core and Django, where .NET Core is for the client and Django is for the back-end service. Fig.10 shows the window of diagnosis result in the client. It takes only 15ms for the

TABLE VI
USING AND NOT USING THE TRANSFER LEARNING FRAMEWORK

IsTransfered	Accuracy(%)
✓	**92.8**
✗	81.6

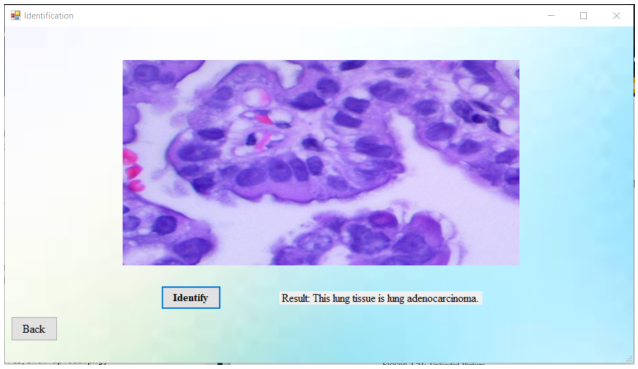

Fig. 10. Identification Result

system to identify one image. Moreover, if the value of the class with the highest probability predicted by the model and the class with the second highest probability are too close, the system will alert the doctor.

VI. Conclusion

In this paper, we propose a ViT-CNN hybrid neural network called CPNet to identify carcinoma pathological slices. CPNet achieves 99.8% accuracy in *PatchCamelyon* and 99.2% accuracy in our base dataset, outperforming other models. We also propose a practical transfer learning framework to train the model in a few samples to identify malignancy degrees, which achieves 92.8%, 11.2% higher than directly training. Last but not least, we design and develop a fast diagnosis system, which could effectively help doctors improve the diagnosis accuracy and speed.

Declarations

Data availability The datasets used during and/or analyzed during the current study are available from their official websites and the first author on a reasonable request.

Contribution Runwei Guan and Yanhua Fei contributes to this paper equally.

References

[1] Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual review of biomedical engineering, 19, 221-248.
[2] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical image analysis, 2017, 42: 60-88.
[3] Cengil, E., & Cinar, A. (2018, September). A deep learning based approach to lung cancer identification. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP) (pp. 1-5). IEEE.
[4] Bhatia, S., Sinha, Y., & Goel, L. (2019). Lung cancer detection: a deep learning approach. In Soft Computing for Problem Solving (pp. 699-705). Springer, Singapore.
[5] Asuntha, A., & Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications, 79(11), 7731-7762.
[6] Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, 374-382.
[7] Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. Artificial intelligence in medicine, 24(1), 25-36.
[8] Travis, W. D. (2011). Pathology of lung cancer. Clinics in chest medicine, 32(4), 669-692.
[9] Cengil, E., & Cinar, A. (2018, September). A deep learning based approach to lung cancer identification. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP) (pp. 1-5). IEEE.
[10] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., ... & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine, 24(10), 1559-1567.
[11] J. Wang, S., Yang, D. M., Rong, R., Zhan, X., Fujimoto, J., Liu, H., ... & Xiao, G. (2019). Artificial intelligence in lung cancer pathology image analysis. Cancers, 11(11), 1673.
[12] AlZubaidi, A. K., Sideseq, F. B., Faeq, A., & Basil, M. (2017, March). Computer aided diagnosis in digital pathology application: Review and perspective approach in lung cancer classification. In 2017 annual conference on new trends in information & Communications technology applications (NTICT) (pp. 219-224). IEEE.
[13] Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J., ... & Xiao, G. (2017). Comprehensive computational pathological image analysis predicts lung cancer prognosis. Journal of Thoracic Oncology, 12(3), 501-509.
[14] Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., & Yang, M. H. (2021). Intriguing properties of vision transformers. Advances in Neural Information Processing Systems, 34.
[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[16] Park, N., & Kim, S. (2022). How Do Vision Transformers Work?. arXiv preprint arXiv:2202.06709.
[17] Xu, Y., Zhang, Q., Zhang, J., & Tao, D. (2021). Vitae: Vision transformer advanced by exploring intrinsic inductive bias. Advances in Neural Information Processing Systems, 34.
[18] Srinivas, A., Lin, T. Y., Parmar, N., Shlens, J., Abbeel, P., & Vaswani, A. (2021). Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16519-16529).
[19] Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., & Tian, Q. (2021). Visformer: The vision-friendly transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 589-598).
[20] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.
[21] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[22] Misra, D. (2019). Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681.
[23] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.
[24] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).
[25] Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. Advances in neural information processing systems, 32.
[26] https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images
[27] Jia, D., Han, K., Wang, Y., Tang, Y., Guo, J., Zhang, C., & Tao, D. (2021). Efficient vision transformers via fine-grained manifold distillation. arXiv preprint arXiv:2107.01378.
[28] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. Advances in Neural Information Processing Systems, 34.
[29] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.
[30] Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018, September). Rotation equivariant CNNs for digital pathology. In International Conference on Medical image computing and computer-assisted intervention (pp. 210-218). Springer, Cham.