PAS-Net: Rapid Prediction of Antibiotic Susceptibility from Fluorescence Images of Bacterial Cells Using Parallel Dual-Branch Network

Wei Xiong[1], Kaiwei Yu[2], Liang Yang[2], and Baiying Lei[1](✉)

[1] National-Regional Key Technology Engineering Laboratory for MedicalUltrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China
leiby@szu.edu.cn

[2] School of Medicine, Southern University of Science and Technology, Shenzhen 518055, China

Abstract. In recent years, the emergence and rapid spread of multi-drug resistant bacteria has become a serious threat to global public health. Antibiotic susceptibility testing (AST) is used clinically to determine the susceptibility of bacteria to antibiotics, thereby guiding physicians in the rational use of drugs as well as slowing down the process of bacterial resistance. However, traditional phenotypic AST methods based on bacterial culture are time-consuming and laborious (usually 24–72 h). Because delayed identification of drug-resistant bacteria increases patient morbidity and mortality, there is an urgent clinical need for a rapid AST method that allows physicians to prescribe appropriate antibiotics promptly. In this paper, we present a parallel dual-branch network (i.e., PAS-Net) to predict bacterial antibiotic susceptibility from fluorescent images. Specifically, we use the feature interaction unit (FIU) as a connecting bridge to align and fuse the local features from the convolutional neural network (CNN) branch (C-branch) and the global representations from the Transformer branch (T-branch) interactively and effectively. Moreover, we propose a new hierarchical multi-head self-attention (HMSA) module that reduces the computational overhead while maintaining the global relationship modeling capability of the T-branch. PAS-Net is experimented on a fluorescent image dataset of clinically isolated Pseudomonas aeruginosa (PA) with promising prediction performance. Also, we verify the generalization performance of our algorithm in fluorescence image classification on two HEp-2 cell public datasets.

Keywords: Parallel dual-branch network · Feature interaction unit · Hierarchical multi-head self-attention · Antibiotic susceptibility prediction

1 Introduction

In recent years, the overuse and misuse of antibiotics have led to an increase in the rate of bacterial antibiotic resistance worldwide [1, 2]. The increasing number of multi-drug resistant strains not only poses a serious threat to human health, but also poses great

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 580–591, 2023.
https://doi.org/10.1007/978-3-031-43993-3_56

difficulties in clinical anti-infection treatment [3]. To address this issue, clinicians rely on antibiotic susceptibility testing (AST) to determine bacterial susceptibility to antibiotics, thus guiding rational drug use. However, the traditional AST method requires overnight culture of the bacteria in the presence of antibiotics, which is time-consuming and laborious (usually 24–72 h). Such delays prevent physicians from determining effective antibiotic treatments promptly. Therefore, there is an urgent clinical need for a rapid AST method that allows physicians to prescribe appropriate antibiotics in an informed manner, which is essential to improve patient outcomes, shorten the treatment duration, and slow down the progression of bacterial resistance.

In this paper, we take *Pseudomonas aeruginosa* (PA) as the research object and observe the difference in shape and distribution of bacterial aggregates formed by sensitive and multi-drug resistant bacteria through fluorescent images, so we want to use image recognition technology to distinguish these two types of bacteria for the purpose of rapid prediction of antibiotic susceptibility. However, we recognize that this classification task presents several challenges (as shown in Fig. 1). Firstly, in images of sensitive PA and multi-drug resistant Pseudomonas aeruginosa (MDRPA), inter-class variation is low, but intra-class variation is high. Secondly, some images have exposure problems due to the high intensity of bacterial aggregation. Lastly, there are low signal-to-noise ratio of the images, coupled with possible image artifacts resulting from inhomogeneous staining or inappropriate manipulation.

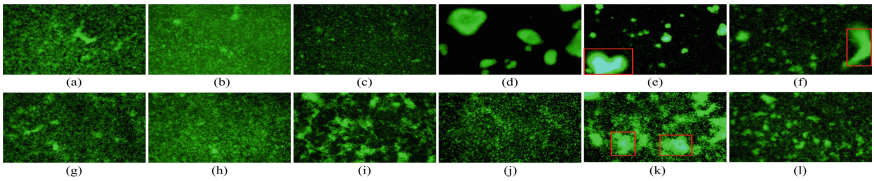

Fig. 1. Examples of fluorescent images of sensitive PA (first row) and MDRPA (second row).

In recent years, deep learning techniques have made a splash in the field of image recognition with their impressive performance and have provided powerful support for a wide range of applications in biomedical research and clinical practice. Notably, deep learning methods based on convolutional neural network (CNN, e.g., ResNet [4], ResNeXt [5], ResNeSt [6]) are widely used in microscopic image classification tasks. For instance, Waisman *et al.* [7] utilized transmission light microscopy images to train a CNN to distinguish pluripotent stem cells from early differentiated cells. Riasatian *et al.* [8] proposed a novel network based on DenseNet [9] and fine-tuned and trained it with various configurations of histopathology images. Recently, due to the successful application of ViT [10] to image classification tasks, many research efforts (e.g., DeiT [11], PVT [12], Swin Transformer [13]) have attempted to introduce the power of self-attention mechanism [14] into computer vision. For example, He *et al.* [15] applied a spatial pyramidal Transformer network to learn long-range contextual information for skin lesion analysis for skin disease classification.

The above studies show that two deep learning frameworks, CNN and Transformer, are effective in microscopy image classification tasks. CNN is good at extracting local

features, but its receptive field is limited by the size of the convolution kernel and cannot effectively capture the global information in the image. Meanwhile, in visual Transformer, its self-attention module is good at capturing feature dependencies over long distances, but ignores local feature information. However, these two kinds of feature information are very important for the classification of microscope images with complex features. To tackle this issue, this paper builds a hybrid model that maximizes the advantages of CNN and Transformer, thus enhancing the feature representation of the network. To achieve the complementary advantages of these two techniques, we propose a parallel dual-branch network named PAS-Net, specifically designed to enable rapid prediction of bacterial antibiotic susceptibility. The main contributions of this study are as follows:

1) We develop a parallel dual-branch classification network to realize the interactive learning of features throughout the whole process through feature interaction unit (FIU), which can better integrate local features of CNN branch (C-branch) and global representations of Transformer branch (T-branch).
2) We propose a more efficient hierarchical multi-head self-attention (HMSA) module, which utilizes a local-to-global attention mechanism to simulate the global information of an image, while effectively reducing the computational costs and memory consumption.

To the best of our knowledge, this study represents the first attempt to use deep learning techniques to realize rapid AST based on PA fluorescence images, which provides a new perspective for predicting bacterial antibiotic susceptibility.

1.1 Method

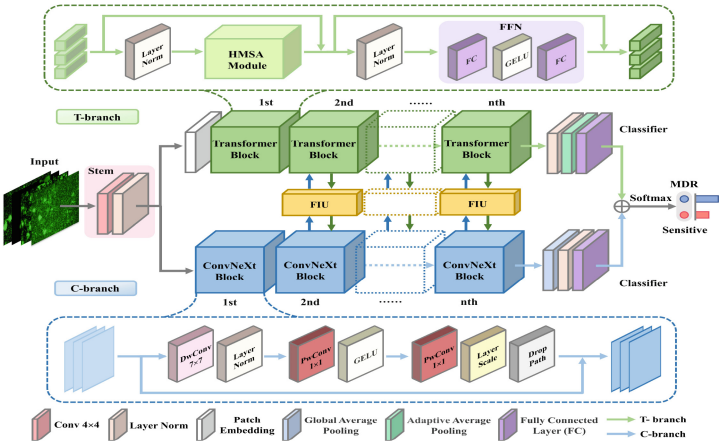

Fig. 2. Overall architecture of the proposed PAS-Net. DwConv: Depthwise convolution. PwConv: Pointwise convolution. FFN: Feed forward network.

Figure 2 shows the overview of our proposed PAS-Net. The model consists of four parts: the Stem module, the parallel C-branch and T-branch, and the FIU connecting the

dual branches. The Stem module, which is a 4×4 convolution with stride 4 followed by a layer normalization for quadruple downsampling of the input image. The C-branch and T-branch are stacked with 12 ConvNeXt [16] blocks and Transformer blocks respectively. FIU is applied from the second feature extraction layer, because the initial features of the first feature extraction layer are the same and all come from Stem module, so there is no need for interaction.

1.2 Feature Interaction Unit

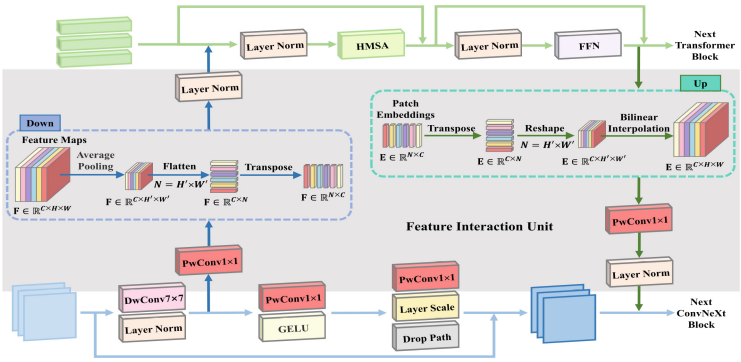

Fig. 3. FIU implementation details. The data stream into FIU is divided into two directions, denoted as CNN → Transformer and Transformer → CNN.

Feature dimension mismatch exists between feature map from C-branch and vector sequence from T-branch. Therefore, our network use FIU as a bridge to effectively combine the local features and the global representation in an interactive manner to eliminate the misalignment between the two features, as shown in Fig. 3.

CNN→Transformer: The feature map is first aligned with the dimensions of the patch embedding by 1×1 convolution. Then, the feature resolution is adjusted using the down-sampling module to complete the alignment of the spatial dimensions. Finally, the feature maps are summed with the patch embedding of the T-branch.

Transformer→CNN: After going through the HMSA module and FFN, the patch embedding is fed back from the T-branch to the C-branch. An up-sampling module needs to be used first for the patch embedding to align the spatial scales. The patch embedding is then aligned to the number of channels of the feature map by 1×1 convolution, and added to the feature map of the C-branch.

1.3 Hierarchical Multi-head Self-attention

Figure 4 shows the detailed structure of HMSA module. To be able to compute attention in a hierarchical manner, we reshape the input patch embedding $\boldsymbol{E}$ back to the patch map $\boldsymbol{E}_p$. Firstly, the patch map $\boldsymbol{E}_p$ is divided into small grids of size $G \times G$, *i.e.*, each

grid contains $G \times G$ (set $G = 4$ in this paper) pixel points. Then, a 1×1 pointwise convolution is performed on $\boldsymbol{E}_{p'}$ to obtain three matrices $\boldsymbol{Q} = \boldsymbol{E}_{p'}\boldsymbol{W}^Q$、$\boldsymbol{K} = \boldsymbol{E}_{p'}\boldsymbol{W}^K$ and $\boldsymbol{V} = \boldsymbol{E}_{p'}\boldsymbol{W}^V$, respectively, where $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$ and $\boldsymbol{W}^V$ are three learnable weight matrices with shared parameters that are updated together with the model parameters during training. After that, we compute local attention $\boldsymbol{A}_0$ within each small grid using the self-attention mechanism, which can be defined as:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}. \tag{1}$$

Then Eq. (1) is applied once more on the basis of $\boldsymbol{A}_0$ to obtain global attention $\boldsymbol{A}_1$. We reshape them back to the shape of the input $\boldsymbol{E}_{p'}$. The final output of HMSA is

$$HMSA(E) = \text{Transpose}(\text{Flatten}(\text{A}_1 + \text{A}_0 + \text{E}'_\text{p})). \tag{2}$$

The original MSA module computes attention map over the entire input feature, and its computational complexity scale quadratically with spatial dimension N, which can be calculated as:

$$\Omega(MSA) = 4ND^2 + 2N^2D. \tag{3}$$

In contrast, our HMSA module computes attention map in a hierarchical manner so that $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$ are computed within small $G \times G$ grids. The computational complexity of HMSA is

$$\Omega(HMSA) = 3ND^2 + 2NG^2D. \tag{4}$$

With this approach, only a limited number of image blocks need to be processed in each step, thus significantly reducing the computational effort of the module from $O(N^2)$ to $O(NG^2)$, where G^2 is much smaller than N. For example, the size of the input image is 224×224, if the patch is divided according to the size of 4×4, the division will get $(224 / 4)^2 = 3136$ patches, i.e., $N = 3136$. However, we set G to 4, so the computational complexity of the HMSA module is greatly reduced.

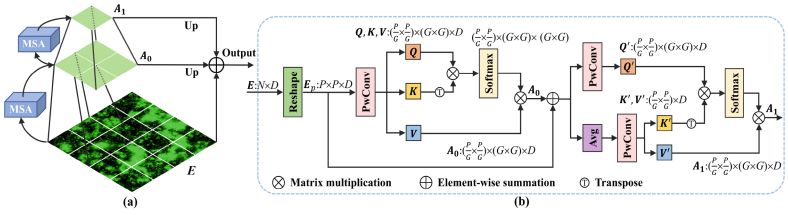

Fig. 4. Illustration of proposed HMSA. (a) Hierarchical structure of the HMSA module. (b) Implementation details of the HMSA module.

2 Experiments and Results

2.1 Experimental Setup

The fluorescent images of PA come from a local medical school. We screen out 12 multi-drug resistant strains and 11 sensitive strains. Our dataset has 2625 fluorescent images of PA, 1233 images of sensitive PA and 1392 images of MDRPA. We randomly divide the data into a training set and a test set in a ratio of 9:1. To better train the network model and prevent overfitting, we perform five data enhancement operations on each image, including horizontal flip, vertical flip and rotation at different angles (90°, 180°, 270°). Finally, our data volume is expanded to 15,750 images, including 14,178 training images and 1,572 test images.

To achieve comprehensive and objective assessment of the classification performance of the proposed method, we select eight classification evaluation metrics, including accuracy (Acc), precision (Pre), recall, specificity (Spec), F1-score (F1), Kappa, area under the receiver operating characteristic (ROC) curve (AUC). All experiments are implemented by configuring the PyTorch framework on NVIDIA GTX 2080Ti GPU with 11 GB of memory.

2.2 Results

In this paper, we adopt Conformer [17] as the baseline of our network, and then make adjustments and improvements to optimize its performance on the PA fluorescence image dataset. Table 1 shows the results of the ablation experiments for different modules in the network. Among them, "Baseline + CB" indicates that the ResNet block in the original C-branch is replaced by the ConvNeXt block, reflecting the impact of the performance-enhanced C-branch on the classification performance. "Baseline + CB + Stem" replaces the convolutional module of the standard ResNet network in baseline with the Stem module on top of the modified C-branch. "Baseline + CB + Stem + HMSA" represents the replacement of the traditional MSA module in Baseline with the efficient HMSA module proposed in this paper on the basis of "Baseline + CB + Stem". The proposed HMSA module replaces the traditional MSA module in baseline, which achieves the improvement of network efficiency and classification performance.

In order to evaluate the classification performance of the proposed method, we choose ten state-of-the-art image classification methods for comparison, including 5 CNN networks: ResNet50 [4], ResNeXt50 [5], ResNeSt50 [6], ConvNeXt-T [16] and DenseNet121 [9], and 5 Transformer-related networks: ViT-B/16 [10], DeiT-S [11], PVT-M [12], Swin-T [13] and CeiT-S [18]. The results of the comparative experiment are illustrated in Table 2. We can observe that our dual-branch network achieves the best performance on our dataset, and outperforms the CeiT-S by 7.08%, 6.7%, and 6.66% in accuracy, recall and F1-score, respectively.

To further analyze and compare the computational complexity of different methods, we compare the number of model parameters (#Param) and the number of floating-point operations per second (FLOPs). In general, the higher the number of parameters and operations, the higher the performance of the model, but at the same time, the greater the computational and storage overhead. It can be seen that the accuracy of ViT is 5% lower

Table 1. Ablation experiments of different modules in PAS-Net (%).

Model	Acc	Pre	Recall	Spec	F1	Kappa	Youden	AUC
Baseline	91.72 ± 2.71	92.59 ± 2.98	91.83 ± 4.38	91.59 ± 3.75	92.14 ± 2.64	83.38 ± 5.41	83.42 ± 5.34	97.46 ± 1.30
Baseline + CB	94.08 ± 2.39	93.46 ± 3.23	95.63 ± 2.74	92.32 ± 4.14	94.49 ± 2.19	88.09 ± 4.82	87.95 ± 4.90	98.48 ± 1.04
Baseline + CB + Stem	94.19 ± 1.64	93.50 ± 1.84	95.73 ± 2.19	92.45 ± 2.25	94.59 ± 1.55	88.32 ± 3.30	88.18 ± 3.28	98.11 ± 0.45
Baseline + CB + Stem + HMSA	**96.04 ± 1.35**	**95.81 ± 1.59**	**96.80 ± 2.36**	**95.18 ± 1.91**	**96.28 ± 1.30**	**92.04 ± 2.70**	**91.97 ± 2.64**	**99.42 ± 0.37**

Table 2. Classification performance camparison of state-of-the-art methods on the test set (%).

Model	Acc	Pre	Recall	Spec	F1	Kappa	AUC	#Param	FLOPs
ResNet50	87.97 ± 1.74	88.85 ± 2.89	88.66 ± 5.15	87.19 ± 4.17	88.62 ± 1.94	75.86 ± 3.43	95.58 ± 1.23	23.5M	8.45G
ResNeXt50	89.62 ± 1.50	89.40 ± 1.73	91.31 ± 3.03	87.71 ± 2.41	90.31 ± 1.51	79.14 ± 2.98	95.50 ± 0.71	23.0M	8.78G
ResNeSt50	90.67 ± 1.90	89.88 ± 2.31	92.96 ± 3.76	88.07 ± 3.10	91.34 ± 1.90	81.23 ± 3.79	96.38 ± 0.75	25.4M	11.02G
ConvNeXt-T	87.91 ± 0.93	84.63 ± 1.43	94.41 ± 1.91	80.57 ± 2.38	89.23 ± 0.83	75.55 ± 1.88	95.85 ± 0.44	27.8M	8.90G
DenseNet121	89.92 ± 1.89	89.77 ± 2.53	91.65 ± 6.09	87.95 ± 3.90	90.53 ± 2.18	79.74 ± 3.68	96.31 ± 0.99	7.0M	5.94G
ViT-B/16	82.63 ± 2.47	83.20 ± 4.03	84.75 ± 4.72	80.25 ± 6.72	83.81 ± 2.18	65.08 ± 5.07	90.84 ± 1.23	86.0M	37.30G
DeiT-S	85.67 ± 1.57	88.44 ± 4.81	84.52 ± 4.93	86.98 ± 6.54	86.21 ± 1.48	71.31 ± 3.21	94.14 ± 1.12	21.7M	10.13G
PVT-M	86.72 ± 0.58	84.16 ± 1.68	92.46 ± 2.37	80.23 ± 2.94	88.07 ± 0.54	73.16 ± 1.19	93.52 ± 0.54	43.7M	14.09G
Swin-T	88.57 ± 1.68	90.07 ± 3.67	88.58 ± 5.92	88.58 ± 4.64	89.09 ± 2.11	77.09 ± 3.26	95.52 ± 0.49	29.8M	4.45G
CeiT-S	88.96 ± 1.29	89.24 ± 1.43	90.10 ± 3.21	87.67 ± 2.08	89.62 ± 1.40	77.82 ± 2.55	95.41 ± 0.55	23.9M	10.51G
Ours	**96.04 ± 1.35**	**95.81 ± 1.59**	**96.80 ± 2.36**	**95.18 ± 1.91**	**96.28 ± 1.30**	**92.04 ± 2.70**	**99.42 ± 0.37**	43.4M	23.37G

than that of ResNet50, but its model complexity is about three times higher. The number of model parameters of PVT-M is similar to that of our PAS-Net, but the accuracy is much worse. The number of parameters of our proposed PAS-Net is 43.4M and FLOPs is 23.37G, indicating that the network achieves a good balance between the number of parameters, FLOPs, accuracy and classification consistency.

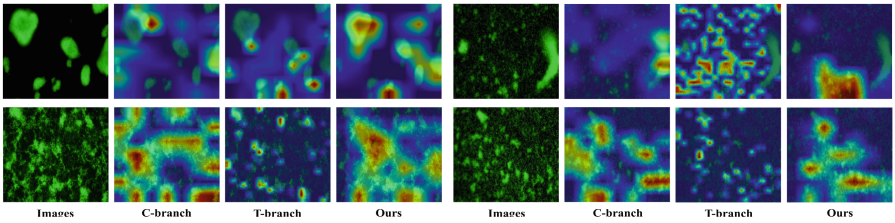

Fig. 5. Grad-cam is used to highlight the discriminant regions of interest for predicting sensitive and multi-drug resistant bacteria. The first column shows four images from the test set, the first two for sensitive PA and the last two for MDRPA.

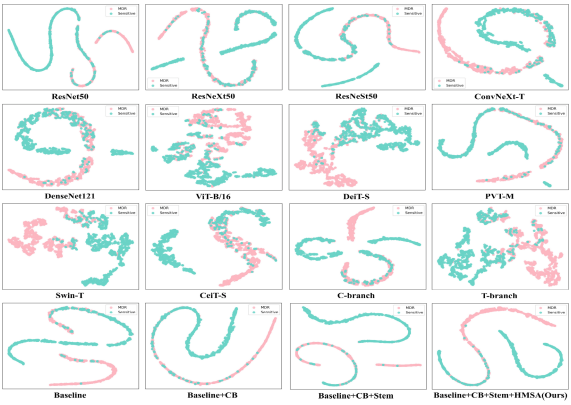

Fig. 6. Two-dimensional t-SNE maps of different models on our test set. Pink points represent MDRPA, blue points represent sensitive PA.

To verify the interpretability of the proposed PAS-Net and understand its classification effect more intuitively and effectively, we visualize the results using Grad-CAM, as shown in Fig. 5. From the second column vertically, we can see that the C-branch only focuses on local edge parts or incorrectly highlights some regions that are not relevant to the discrimination, as shown in the heat map in the first and second rows. From the third column, we can see that the T-branch can obtain the global attention map, but at the same time it produces some worthless and redundant features. A side-by-side comparison shows that the heat map in the fourth column can focus well on some discriminative regions with distinct features and reflect the correlation between local regions. For example, our network can effectively capture the bacterial aggregates with clear edges and largest area in the first image, and also establish the long-range

feature dependencies among the three small bacterial aggregates near the lower right corner; for the second image, our dual-branch network corrects the error of focusing the C-branch to the exposure position because of the Transformer's ability to learn the global feature representation For the third and fourth images, the network nicely combines the discriminative regions focused on by the C-branch and T-branch, capturing both the local features of larger bacterial aggregates and learning the distributional dependencies among bacterial aggregates. This shows to some extent that our proposed model effectively exploits the advantages of CNN and Transformer and maximizes the retention of local features and global representation.

We also use the t-SNE dimensionality reduction algorithm to map the feature vectors learned from the last feature extraction layer of different networks onto a two-dimensional plane, as shown in Fig. 6. The visualization allows us to observe the clustering of the image features extracted by these networks. Compared with other models, the features extracted by our proposed PAS-Net can better distinguish the sensitive bacteria (blue) from the multi-drug resistant bacteria (pink).

2.3 Robustness to HEp-2 Dataset

To further verify the effectiveness of PAS-Net in fluorescent image classification tasks, we also apply our method to two HEp-2 cell public datasets, ICPR 2012 and I3A Task1. ICPR 2012 dataset uses average class accuracy (ACA) as the evaluation metric, which is the same concept as the accuracy mentioned above, while I3A Task1 uses mean class accuracy (MCA). We select four deep learning techniques for classification of HEp-2 cells for comparison, respectively, and the results are shown in Table 3. Without using pre-trained weights for migration learning and data augmentation, our network achieves 81.61% and 98.71% accuracy on ICPR 2012 dataset and I3A Task1 dataset, respectively, and the experimental results demonstrate the generalizability of the proposed PAS-Net for fluorescent image classification tasks.

Table 3: Algorithm comparison on ICPR2012 dataset and I3A Task1 dataset (%).

ICPR 2012	Method	ACA	I3A Task1	Method	MCA
Gao et al. [19]	Seven layers CNN	74.8	Gao et al. [19]	Seven layers CNN	96.76
Phan et al. [20]	VGG-16 + SVM	77.1	Jia et al. [21]	VGG-like network	98.26
Jia et al. [21]	VGG-like network	79.29	Li et al. [22]	Deep residual inception model	98.37
Liu et al. [23]	DACN	81.2	Lei et al. [24]	Cross-modal transfer learning	98.42
Ours	PAS-Net	81.61	Ours	PAS-Net	98.71

3 Conclusion

In this paper, we develop a PAS-Net framework for rapid prediction of antibiotic susceptibility from bacterial fluorescence images only. PAS-Net is a parallel dual-branch feature interaction network. FIU is a connecting bridge to align and fuse the local features from the C-branch and the global representation from the T-branch, which enhances the feature representation ability of the network. We design a HMSA module with less computational overhead to improve the computational efficiency of the model. The experimental results demonstrate that our method is feasible and effective in PA fluorescence image classification task, and can assist clinicians in determining bacterial antibiotic susceptibility.

Acknowledgement. This work was supported National Natural Science Foundation of China (Nos. 62101338, 61871274, 32270196 and U1902209), National Natural Science Foundation of Guangdong Province (2019A1515111205), Shenzhen Key Basic Research Project (KCXFZ20201221173213036, JCYJ20220818095809021, SGDX202011030958020–07, JCYJ201908081556188–06, and JCYJ20190808145011 -259), Shenzhen Peacock Plan Team Project (grants number KQTD20200909113758–004).

References

1. Holmes, A.H., et al.: Understanding the mechanisms and drivers of antimicrobial resistance. Lancet **387**, 176–187 (2016)
2. Dadgostar, P.: Antimicrobial resistance: implications and costs. Infect. Drug Resist. **12**, 3903–3910 (2019)
3. Ferri, M., Ranucci, E., Romagnoli, P., Giaccone, V.: Antimicrobial resistance: a global emerging threat to public health systems. Crit. Rev. Food Sci. Nutr. **57**, 2857–2876 (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
6. Zhang, H., et al.: Resnest: split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2736–2746 (2022)
7. Waisman, A., et al.: Deep learning neural networks highly predict very early onset of pluripotent stem cell differentiation. Stem Cell Rep. **12**, 845–859 (2019)
8. Riasatian, A., et al.: Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. Med. Image Anal. **70**, 102032 (2021)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
10. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:.11929 (2020)
11. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)

12. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
14. Vaswani, A., et al.: Attention is all you need. Adv. neural inf. Process. Syst. **30** (2017)
15. He, X., Tan, E.-L., Bi, H., Zhang, X., Zhao, S., Lei, B.: Fully transformer network for skin lesion analysis. Med. Image Anal. **77**, 102357 (2022)
16. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
17. Peng, Z., et al.: Conformer: local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 367–376 (2021)
18. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 579–588 (2021)
19. Gao, Z., Wang, L., Zhou, L., Zhang, J.: HEp-2 cell image classification with deep convolutional neural networks. IEEE j. Biomed. Health Inform. **21**, 416–428 (2016)
20. Phan, H.T.H., Kumar, A., Kim, J., Feng, D.: Transfer learning of a convolutional neural network for HEp-2 cell image classification. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1208–1211. IEEE (2016)
21. Jia, X., Shen, L., Zhou, X., Yu, S.: Deep convolutional neural network based HEp-2 cell classification. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 77–80. IEEE (2016)
22. Li, Y., Shen, L.: A deep residual inception network for HEp-2 cell classification. In: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R.S., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 12–20. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_2
23. Liu, J., Xu, B., Shen, L., Garibaldi, J., Qiu, G.: HEp-2 cell classification based on a deep autoencoding-classification convolutional neural network. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 1019–1023. IEEE (2017)
24. Lei, H., et al.: A deeply supervised residual network for HEp-2 cell classification via cross-modal transfer learning. Pattern Recogn. **79**, 290–302 (2018)