Technology and Health Care 32 (2024) S465–S475
DOI 10.3233/THC-248041
IOS Press

Intelligent deep learning supports biomedical image detection and classification of oral cancer

Rongcan Chen[a,1], Qinglian Wang[b,1] and Xiaoyuan Huang[c,*]
[a]*School of Informatics, Xiamen University, Xiamen, Fujian, China*
[b]*Department of Stomatology, Zhongshan Hospital (Xiamen), Fudan University, Xiamen, Fujian, China*
[c]*Department of Stomatology, Zhongshan Hospital, Xiamen University, Xiamen, Fujian, China*

Abstract.
BACKGROUND: Oral cancer is a malignant tumor that usually occurs within the tissues of the mouth. This type of cancer mainly includes tumors in the lining of the mouth, tongue, lips, buccal mucosa and gums. Oral cancer is on the rise globally, especially in some specific risk groups. The early stage of oral cancer is usually asymptomatic, while the late stage may present with ulcers, lumps, bleeding, etc.
OBJECTIVE: The objective of this paper is to propose an effective and accurate method for the identification and classification of oral cancer.
METHODS: We applied two deep learning methods, CNN and Transformers. First, we propose a new CANet classification model for oral cancer, which uses attention mechanisms combined with neglected location information to explore the complex combination of attention mechanisms and deep networks, and fully tap the potential of attention mechanisms. Secondly, we design a classification model based on Swim transform. The image is segmented into a series of two-dimensional image blocks, which are then processed by multiple layers of conversion blocks.
RESULTS: The proposed classification model was trained and predicted on Kaggle Oral Cancer Images Dataset, and satisfactory results were obtained. The average accuracy, sensitivity, specificity and F1-Socre of Swin transformer architecture are 94.95%, 95.37%, 95.52% and 94.66%, respectively. The average accuracy, sensitivity, specificity and F1-Score of CANet model were 97.00%, 97.82%, 97.82% and 96.61%, respectively.
CONCLUSIONS: We studied different deep learning algorithms for oral cancer classification, including convolutional neural networks, converters, etc. Our Attention module in CANet leverages the benefits of channel attention to model the relationships between channels while encoding precise location information that captures the long-term dependencies of the network. The model achieves a high classification effect with an accuracy of 97.00%, which can be used in the automatic recognition and classification of oral cancer.

Keywords: Oral cancer classification, attention mechanism, transformer

1. Introduction

Oral cancer is a cancer that occurs in the tissues of the mouth, usually originating in the mucosal cells lining the mouth. This cancer may involve different parts of the mouth, including the lips, floor of the

[1]Contributed equally to this work.
*Corresponding author: Xiaoyuan Huang, Department of Stomatology, Zhongshan Hospital, Xiamen University, Xiamen, China. E-mail: huangxiaoyuan6531@163.com. ORCID: 0009-0006-2161-664X.

mouth, tongue, buccal mucosa, gums, etc. Early and timely identification of OSCC is crucial to improve diagnosis, treatment, and survival. Commonly used detection methods are tissue biopsy and imaging examination. Diagnosis usually requires a tissue biopsy to check for abnormal cells in the affected area. X-rays, CT scans, and MRIs can help understand the size and spread of tumors. Despite our remarkable progress in understanding the molecular mechanisms of cancer, late-stage diagnosis remains a major constraint to the development of precision medicine. The application of deep learning technology in the field of image has achieved great success, greatly promoting the development of image processing and computer vision, including image classification, segmentation, object detection, image generation and so on. CNN is one of the most popular models in deep learning and is particularly suitable for image processing tasks. It captures the local features in the image effectively through the convolution layer, and reduces the data dimension through the pooling layer, so as to achieve efficient learning and classification of the image.

Therefore, deep learning techniques are widely used to improve recognition accuracy and reduce cancer-specific mortality and prevalence [1]. Automated image examination is clearly of great significance in helping pathologists and clinicians detect OSCC early and make treatment decisions. However, detection by healthcare professionals is extremely complicated due to the significant heterogeneity of cancers. In addition, early oral precancerous lesions are often asymptomatic. The lesions often appear to be small and seemingly harmless, leading to late onset of symptoms in patients, ultimately leading to delayed diagnosis [2,3]. The continued development of deep learning provides an effective way for assistive technologies to automate the screening of oral cancer and provide timely feedback to patients and professionals as early as possible when a patient is examined. Oral cancer is detected. This has important implications for improving diagnosis, treatment and improving survival rates.

Computer-aided diagnosis has the ability to analyze large amounts of medical data for tumor detection and classification. It can usually provide accurate classification results in a shorter time and is more accurate than manual diagnosis. Due to the limited availability of public datasets on oral cancer images, there is a lack of extensive research on various deep learning (DL) methods in oral cancer. The aim of this paper is to study the application of several DL algorithms in the classification of oral cancer images. In particular, we apply two deep learning methods, CNN and Transformers. We propose a new CANet classification model for oral cancer, which uses attention mechanisms to encode neglected location information, explores the complex combination of attention mechanisms and deep networks, and fully taps the potential of attention mechanisms. Experimental results show that our CANet has achieved leading performance in oral cancer classification.

2. Literature review

In the early stages of OSCC, computer-automated image analysis can effectively help clinicians and provide informed decisions for cancer management [4]. Bhandari et al. [11] focused on improving the classification and detection performance of oral tumors in the shortest processing time. Their proposed techniques include convolutional neural networks with adaptive loss functions to support multi-class classification and minimize data overfitting, thereby reducing errors in oral tumor classification and prediction. Lu et al. [12] proposed an automatic diagnosis method for oral tumors based on cytopathological images. The method covers focus selection of each cell, classification based on CNNs, and identification of cell nuclei based on fully convolutional regression. The method enables faster determination of the focus of each cell with human-level accuracy. Song et al. [14] introduced a deep learning (DL) image classification method. Through the fusion, extraction and calculation of the data, it provides input to the

deep learning neural network. Next, this study compared the effects of regularization, transfer learning techniques and CNNs on the classification efficiency of oral tumors.

Figueroa et al. [5] designed an interpretable deep learning (DL) model and directed the network to focus on the tumor region and accurately depict the location in the image. Lim et al. [6] developed the D 'OraCa architecture specifically for the classification of oral lesions. They innovatively proposed a method for identifying oral lesions, which improved the classification accuracy. Shamim et al. [7] used transfer learning to identify precancerous lesions in a relatively small dataset and evaluated the performance of six deep convolutional neural network (DCNN) models for oral cancer classification. These DCNN models were able to effectively differentiate between the five types of tongue cancers and performed well in identifying benign and precancerous lesions.

Chan et al. [8] proposed an innovative DCNN that uses texture mapping to identify cancer areas and uses a method to automatically mark regions of interest (ROI). The method consists of two collaborative branches: the upper branch detects oral cancer, and the lower branch marks ROI and segments the lesions. Through this network method, the tumor area can be accurately extracted, and the synergy of the upper and lower branches makes the identification of the tumor area more accurate. This method can also calculate the standard deviation value of the texture image through a sliding window. The considerable heterogeneity of oral lesions makes the detection process difficult and is thought to be a major reason for delayed referrals to oral cancer specialists. In addition, early OSCC lesions are often asymptomatic and may appear as small, seemingly harmless lesions that cause patients to develop symptoms late, ultimately leading to a delayed diagnosis. Therefore, it is particularly important to design an effective classification model for oral cancer. However, some limitations of CNNs in image classification have been identified, especially in the medical field. Examples include the lack of data sets, the uneven distribution of the number of disease images, and the lack of ROI description.

Due to the great success of transformer-based architectures in natural language processing (NLP) tasks, many scholars have recently studied the application of transformers in the field of computer vision [9,10]. The emergence of Visual Transformer (ViT) [23] shows that purely transformer-based architecture is also an effective solution to solve vision problems. The design of ViT is heavily based on the original Transformer model [14] and proposes a novel view of treating image patches as visual words. The transformer's self-attention mechanism helps to increase the weight of important features while suppressing noise-causing features. The converter has also achieved notable success in medical image classification. Behnaz Gheflati et al. [15] designed a VIT-based breast ultrasound image classification model, whose performance can be compared with convolutional neural networks (CNNS), or even better than convolutional neural networks.

3. Materials and methods

3.1. Oral cancer dataset

The data set selected for this article is the Oral Cancer Images classification data set provided on Kaggle, which is a widely trusted and widely used data set in the field of oral cancer classification. This dataset covers images of lips and tongues, divided into two categories: cancerous group and non-cancerous group. The images were obtained from various ENT hospitals in Ahmedabad and the images were classified with the help of doctors. Table 1 showcases the details of the dataset, there are 44 non-cancer images and 87 cancer images, for a total of 131 images. The selection of this data set provides strong support for studying oral cancer classification problems and has wide applications in the field of oral medicine. Figure 1 shows some examples.

Table 1
Dataset details

Class	No. of samples
Cancer	87
Non-Cancer	44
Total Number of Samples	131

Fig. 1. Sample images: The first and second columns are pictures of cancer, and the third and fourth columns are pictures without cancer.

3.2. Image augmentations

Deep learning models rely heavily on sufficient training data to achieve high accuracy. Insufficient data can lead to overfitting problems, which affects model performance. However, in many computer vision tasks, training data is quite limited. In this case, data augmentation techniques become very helpful. Image enhancement generates new samples from existing training samples by introducing a series of random changes, such as rotation, translation, flipping, etc., effectively reducing the model's dependence on certain specific attributes.

In this study, we adopted the Albumentation library for image enhancement. Albumentations [16] is a powerful data enhancement tool that contains rich data enhancement methods, such as image resizing, scaling, rotation, flipping, etc. We randomly extracted 30 rectangular image patches from the official training images at different ratios (3/5 and 4/5 of the original image size), and then resized them to 224 $\times$ 224 pixels using bilinear interpolation. Next, we use Albumentations to augment the data, including random rotation ($[-10°, 10°]$), scaling (90%–110% of width and height), and horizontal and vertical flipping to expand the training data set.

3.3. CANet

Attention mechanisms have been shown to be of great help in image classification and image segmentation [21]. The attention mechanism allows the neural network to selectively focus on a subset of input data when processing it, rather than treating all information equally. This allows the network to concentrate on processing information relevant to the task, improving computational efficiency and model performance. The attention mechanism allows the neural network to dynamically adjust its attention to different regions to better capture important features [14].

We introduced the attention module based on ResNet to form our CANet, as shown in Fig. 2. CANet consists of a classification layer, a global averaging pooling layer and multiple CA (Coordinate Attention)

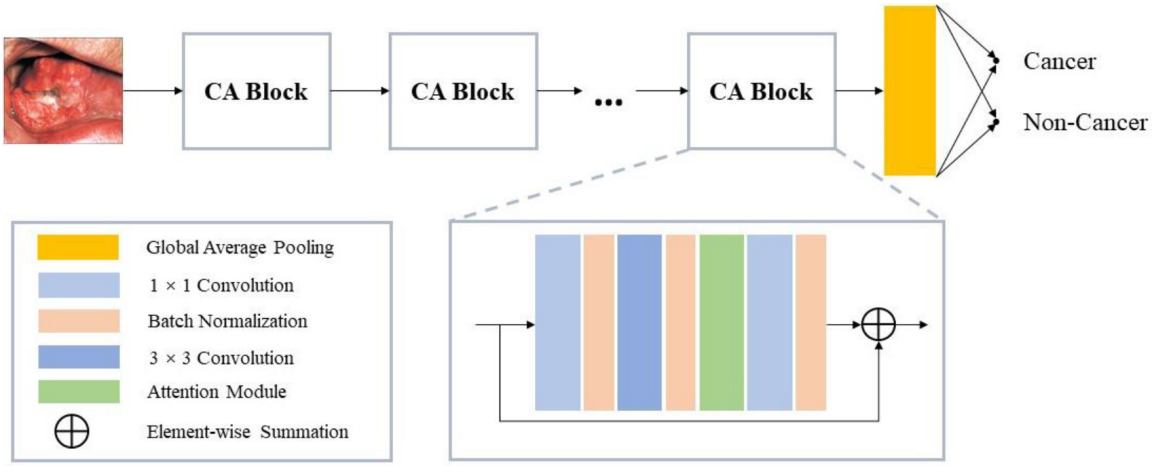

Fig. 2. Architecture of the proposed CANet model.

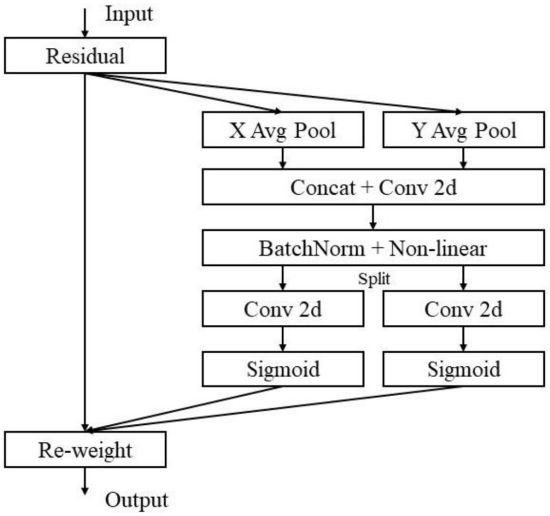

Fig. 3. Schematic of the proposed attention block.

blocks. The CA block is based on the Residual Block and incorporates the attention module. Our attention module captures the long-range relationships of the network by embedding precise location information into channel relationships. A schematic diagram of the attention module can be found in Fig. 3. In the following description, we will explain it in detail.

In order to promote the attention block to better utilize precise location information to capture long-distance interactions in space, we split the global pooling into two independent 1D feature encoding operations, and expression Eq. (1) is transformed into this Right operation. Specifically, for a given input X, we encode each channel separately along the horizontal with one averaging pooling operation and vertically with another averaging pooling operation. Therefore, the output of the c-th channel at height h can be represented by the following formula:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leqslant i \leqslant W} x_c(h, i) \tag{1}$$

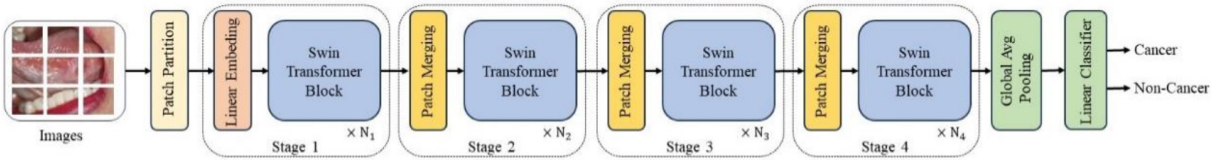

Fig. 4. The architecture of the Swin transformer model for oral cancer classification.

Similarly, the output of the c-th channel at width w can be expressed as:

$$z_c^w(w) = \frac{1}{H}\sum\nolimits_{0 \leqslant j \leqslant H} x_c(j, w) \tag{2}$$

Our method aggregates feature along different spatial directions to generate a pair of direction-aware feature maps. Our attention blocks retain precise location information while capturing remote dependencies. This helps the network locate objects of interest more accurately, which improves the model's understanding and perception of spatial structure.

Next, we concatenate the feature maps produced by Eqs (1) and (2) and send them into a shared 1×1 convolution transformation function F_1, resulting in:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{3}$$

where $[\cdot, \cdot]$ represents the cascade operation along the spatial dimension, δ is the nonlinear activation function, and $f \in \mathrm{R}^{C/r \times (H+W)}$ is the intermediate feature map in the horizontal and vertical directions, where r is the reduction ratio used to control the block size. We then split f along the spatial dimension into two separate tensors f^h and f^w. Two other 1×1 convolution transformations F_h and F_w are used to transform f^h and f^w respectively into tensors with the same number of channels as the input X, resulting in:

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{4}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{5}$$

where σ is the sigmoid function. The output g^h and g^w are then respectively expanded and used as attention weights. Finally, the output of our attention block F_{ca} can be written as:

$$F_{ca}(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{6}$$

3.4. Transformer

The Vision Transformer (ViT) [23] stands out as the pioneering transformer model that employs exclusive self-attention for image recognition. This model divides the image into a sequence of 2D patches, incorporates positional markers, and individually processes each patch through multiple transformer layers, multi-head self-attention modules (MSA), and multi-layer perceptrons (MLP). While ViT necessitates pre-training on extensive datasets, it excels in various computer vision tasks. In addressing the challenges of ViT training, [17] introduced the Data Efficient Image Transformer (DeiT), merging the transformer model with a distillation approach. Notably, the DeiT model showcases the transformer's effectiveness in training on moderately sized datasets. Recent advancements in research, such as the ViT-based token-to-token ViT (T2T-ViT) [18], Pyramid Vision Transformer (PVT) [19], and Swin Transformer [20], have further elevated performance standards in computer vision.

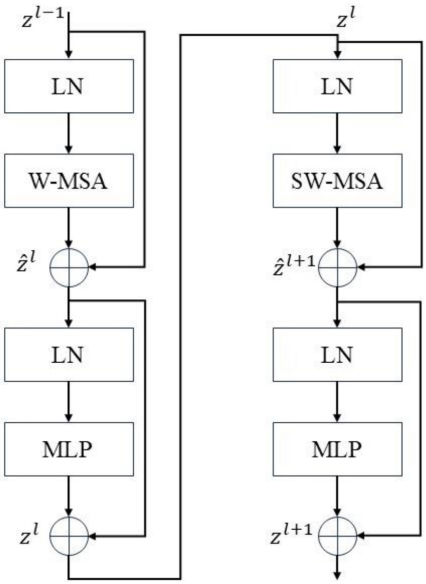

Fig. 5. Two consecutive Swin transformer blocks.

Figure 4 shows our design of Transformer-based oral cancer classification architecture. The Swin Transformer model consists of four stages, each stage consists of multiple Swin-Transformer blocks, which are connected through patch merging layers. The input size is $H \times W \times 3$ non-overlapping patches of size 4×4 are extracted through the patch segmentation module to obtain patch tokens of $\frac{H}{4} \times \frac{W}{4}$.

Next, these patch tokens pass through multiple Swin Transformer Blocks (STBs), as shown in Fig. 5. This Swin Transformer block together with the linear embedding layer is called "Stage 1". In Phase 1, the number of tokens remains $\frac{H}{4} \times \frac{W}{4}$. The feature map after generating this layer. At the beginning of the stage, the patch merging layer is responsible for reducing the number of tokens by concatenating the features of 2 patches and changing the dimensionality of the output to 2C. Next, apply the following STB for feature transformation. Successive patch merging and STB are denoted as "Phase 2". At this stage, the resolution of the output feature map is maintained at $\frac{H}{8} \times \frac{W}{8}$. "Phase 2" is repeated twice to form "Phase 3" and "Phase 4", in which the constructed feature maps are $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$ respectively.

The computational unit in the Swin Transformer comprises two alternating window-based multi-head self-attention (W-MSA) blocks and a shift window-based multi-head self-attention (SW-MSA) block. Alongside MSA, it incorporates two MLP layers, with layer normalization (LN) applied before each layer in both MSA and MLP. A residual connection is established between these two modules. The configuration of the STB is visually represented in Fig. 5.

4. Results and discussion

In our experiments, we divide the data set in an 8:1:1 ratio for training, validation, and testing. We choose the SGD algorithm with batch size 32 to optimize the classification network. The initial learning rate is set to 0.0001 and is reduced by half every 20 epochs. The maximum epoch is set to 100. In addition, the CANet model is initialized using ImageNet pre-trained weights, which provides reasonable initial weights for the encoder and avoids over-fitting problems.

Table 2
Result analysis of CANet approach with distinct class labels

Label	Accuracy	Sensitivity	Specificity	F1-score
Cancer	97.10	95.64	100.0	98.10
Non-Cancer	96.90	100.0	95.64	95.12
Average	97.00	97.82	97.82	96.61

Table 3
Result analysis of Swin transformer approach with distinct class labels

Label	Accuracy	Sensitivity	Specificity	F1-score
Cancer	95.10	93.64	96.90	94.10
Non-Cancer	94.80	97.10	94.14	95.22
Average	94.95	95.37	95.52	94.66

Quantitative evaluation: We use four performance indicators to evaluate the obtained classification results, including F1-Score, accuracy (Acc), sensitivity (SE), and specificity (SP), the indicators are defined as follows:

$$F1 = 2 \times \frac{TP}{2 \times TP + FP + FN} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

Among them, *TP*, *FN*, *TN* and *FP* represent the number of true positives, false negatives, true negatives and false positives respectively, and *X0* and *X1* are the confidence scores of negatives and positive examples.

Qualitative evaluation: We adopt class activation mapping (CAM) [22] to visualize the attention regions in the obtained feature maps. Since the GAP feature vector comes directly from the feature map, the weight can be regarded as the contribution of the feature map to the target category score, and the CAM can be obtained by weighting the sum. We define the output of the GAP layer as $M_c\,(x, y)$, where c represents the number of categories, x and y represent the position of the feature map, and represents the CAM of class c as M_c, then each spatial element is given by:

$$M_c\,(x, y) = \sum\nolimits_k w_k^c f_k\,(x, y) \tag{11}$$

Table 2 provides the classification results of oral cancer by CANet model. The CANet model identified cancer cases with 97.10% accuracy, 95.64% sensitivity, 100.00% specificity and 98.10% F1-Socre, respectively. In addition, the CANet model was used to classify non-cancer cases, and the accuracy, sensitivity, specificity and F1-Socre were 96.90%, 100.00%, 95.64% and 95.12, respectively. In addition, the average accuracy, sensitivity, specificity and F1-Score of CANet model were 97.00%, 97.82%, 97.82% and 96.61, respectively.

Table 3 shows the classification results of Swin transformer for oral cancer. The average accuracy, sensitivity, specificity and F1-Socre were 94.95%, 95.37%, 95.52% and 94.66%, respectively. The experimental results show that Swin transformer pre-trained on ImageNet achieves good results in the classification of oral cancer, but the performance is slightly inferior to our CANet. While the Swin

Table 4
Result on Oral Cancer Dataset with proposed model compared to the existing models. Optimal performance is indicated in bold

Model	Accuracy	Sensitivity	Specificity	F1-score
Swin Transformer	94.95	95.37	95.52	94.66
OIDCNN-OPMDD	96.30	97.37	97.37	95.71
IDL-OSCDC	92.86	90.00	95.00	91.81
CANet	**97.00**	**97.82**	**97.82**	**96.61**

Table 5
Ablation experiment

Description	Accuracy	Sensitivity	Specificity	F1-score
Baseline	93.95	92.37	93.52	92.66
+ X Attention	95.10	93.48	94.12	93.68
+ Y Attention	95.10	93.37	93.98	93.32
+ Coor Attention	97.00	97.82	97.82	96.61

Transformer model combines the strengths of both convolutional neural networks and transformers to enhance the sensitivity in classifying oral cancer lesions, the nature of transformers mandates training with a substantial amount of data. However, our original oral cancer data set contains only 131 images, which is a small data set. Not taking full advantage of transformer.

In addition, to verify the effectiveness of our network, we compared the performance of the CANet model with that of OIDCNN-OPMDD [24] and IDL-OSCDC [25]. In [24], the author used IDCNN model to carry out feature extraction and classification process, and OIDCNN-OPMDD model combined Inception module with DCNN. At the same time, in order to improve the classification performance of IDCNN model, moth flame optimization (MFO) technology is used. IDL-OSCDC uses Gabor filtering as preprocessing to eliminate the noise content in the image. Detailed comparison results are shown in Table 4. It can be clearly seen that the CANet model using attention mechanism proposed in this paper achieves satisfactory performance and is superior to other models in classification accuracy.

To demonstrate the performance of the proposed attention, an ablation experiment was conducted, and the corresponding results are shown in Table 5. We remove horizontal attention or vertical attention from attention to understand the importance of coordinate coded information. As shown in Table 5, the performance of the model with vertical or horizontal attention is similar However, the best results are achieved when both are considered. Experimental results show that coordinate information embedding is more helpful to image classification.

On the basis of these findings, it would be worthwhile to explore broader applications of the CANet model in various medical fields. For example, its application in breast cancer, lung cancer and other high-risk cancers. The attention module enables the network to focus on processing task-relevant information. Our neural network is resistant to noise and variation in input data. By focusing on parts that are more important to the task, the network can more easily ignore noise or irrelevant information, improving robustness. Although our model is robust, its predictions may be affected by non-medical factors. Image quality may be affected by device performance, environmental factors, etc. Low-quality images can make it difficult for the model to classify correctly. Medical images may come from different medical devices, institutions or doctors. These different sources may lead to changes in data distribution, thereby affecting the model's generalization performance. Inconsistent labels result from manual annotation errors, subjective judgments, and other reasons. Inconsistent annotations may have an impact on model learning and performance evaluation.

Although our model exhibits superior predictive performance, challenges remain in interpreting the model's decision-making process. In future research, we can further explore how to enhance the interpretability of the model so that doctors and patients can better understand the prediction results. Overall, this study provides new insights into the application of deep learning in oral cancer classification, but there is still room for further improvement.

5. Conclusion

In this study, we explored in depth different deep learning architectures and methods in the field of oral cancer classification, including convolutional neural networks and converters. Experimental results show that the average accuracy, sensitivity, specificity and F1-Score of Swin transformer architecture reach 94.95%, 95.37%, 95.52% and 94.66%, respectively. In contrast, the average accuracy, sensitivity, specificity and F1-Score of our proposed CANet model were significantly improved to 97.00%, 97.82%, 97.82% and 96.61%. The attention module in the CANet model gives full play to the advantage of channel attention, and models the relationship between channels, capturing the long-term dependency with accurate location information. Therefore, the CANet model performs well and can be widely used in the automatic identification and classification of oral cancers. In the future, we will further investigate deep instance segmentation of oral cancers and combine it with the CANet model to improve the accuracy of the overall classification.

Looking forward, it is necessary to discuss possible future research directions to build on our findings and further advance the field of medical prediction. Future research efforts can explore improvements to the CANet model to obtain higher classification accuracy. It would be valuable to investigate ways to enhance the interpretability of deep learning models in medical settings, allowing physicians and patients to better understand the decision-making processes behind predictions. Furthermore, class imbalance in the dataset may cause the model to perform poorly on a few classes. Reasonable handling of sample imbalance is to ensure that the model has good performance for all categories and is also the focus of future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 62002304) and the China Fundamental Research Funds for the Central Universities (Grant no. 20720210053).

Conflict of interest

None to report.

References

[1] Singh A, Sahu A, Verma S. Computer intelligence in detection of malignant or premalignant oral lesions: the story so far. Computational Intelligence in Oncology: Applications in Diagnosis. Prognosis and Therapeutics of Cancers. 2022: 187-200.

[2] Siddiqui SY, Haider A, Ghazal TM, et al. IoMT cloud-based intelligent prediction of breast cancer stages empowered with deep learning. IEEE Access. 2021; 9: 146478-146491.

[3] Mansour RF, Alfar NM, Abdel-Khalek S, et al. Optimal deep learning based fusion model for biomedical image classification. Expert Systems. 2022; 39(3): e12764.
[4] Azimi S, Ghorbani Z, Tennant M, et al. Population survey of knowledge about oral cancer and related factors in the capital of Iran. Journal of Cancer Education. 2019; 34: 116-123.
[5] Figueroa KC, Song B, Sunny S, et al. Interpretable deep learning approach for oral cancer classification using guided attention inference network. Journal of Biomedical Optics. 2022; 27(1): 015001-015001.
[6] Lim JH, Tan CS, Chan CS, et al. D'OraCa: deep learning-based classification of oral lesions with mouth landmark guidance for early detection of oral cancer//Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25. Springer International Publishing. 2021: 408-422.
[7] Shamim MZM, Syed S, Shiblee M, et al. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. The Computer Journal. 2022; 65(1): 91-104.
[8] Chan CH, Huang TT, Chen CY, et al. Texture-map-based branch-collaborative network for oral cancer detection. IEEE Transactions on Biomedical Circuits and Systems. 2019; 13(4): 766-780.
[9] Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022; 45(1): 87-110.
[10] Xu Y, Wei H, Lin M, et al. Transformers in computational visual media: A survey. Computational Visual Media. 2022; 8: 33-62.
[11] Bhandari B, Alsadoon A, Prasad PWC, et al. Deep learning neural network for texture feature extraction in oral cancer: Enhanced loss function. Multimedia Tools and Applications. 2020; 79: 27867-27890.
[12] Lu J, Sladoje N, Runow Stark C, et al. A deep learning based pipeline for efficient oral cancer screening on whole slide images//International Conference on Image Analysis and Recognition. Cham: Springer International Publishing. 2020: 249-261.
[13] Song B, Sunny S, Uthoff RD, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. Biomedical Optics Express. 2018; 9(11): 5318-5329.
[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017: 30.
[15] Gheflati B, Rivaz H. Vision transformers for classification of breast ultrasound images//2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE. 2022: 480-483.
[16] Buslaev A, Iglovikov VI, Khvedchenya E, et al. Albumentations: fast and flexible image augmentations. Information. 2020; 11(2): 125.
[17] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention// International conference on machine learning. PMLR. 2021: 10347-10357.
[18] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 558-567.
[19] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568-578.
[20] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
[21] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.
[22] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.
[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. arXiv preprint arXiv2010.11929, 2020.
[24] Alabdan R, Alruban A, Hilal AM, et al. Artificial-Intelligence-Based Decision Making for Oral Potentially Malignant Disorder Diagnosis in Internet of Medical Things Environment. Healthcare. MDPI. 2022; 11(1): 113.
[25] Alanazi AA, Khayyat MM, Khayyat MM, et al. Intelligent deep learning enabled oral squamous cell carcinoma detection and classification using biomedical images. Computational Intelligence and Neuroscience. 2022; 2022.