International Journal on Computational Modelling Applications
Vol. 01, Iss. 01, S. No. 003, pp. 32-44, July 2024
ISSN (E): 3048-8516

TransEns-Network: An Optimized Light-weight Transformer and Feature Fusion Based Approach of Deep Learning Models for the Classification of Oral Cancer

Karnika Dwivedi[1], Bharti Chugh[2], Anugrah Srivastava[3], Jai Prakash Pandey[4]

[1,2]Department of Computer Science & Engineering, KIET Group of Institutions, Ghaziabad, Delhi-NCR, India
[3]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India
[4]Dr A P J Abdul Kalam Technical, University, Lucknow, Uttar Pradesh, India
dwivedikarnika1995@gmail.com*

How to cite this paper: Karnika Dwivedi, Bharti Chugh, Anugrah Srivastava, Jai Prakash Pandey, "TransEns-Network: An Optimized Light-weight Transformer and Feature Fusion Based Approach of Deep Learning Models for the Classification of Oral Cancer International Journal on Computational Modelling Applications , Vol. no. 01, Iss. No. 01, S No. 003, pp. 32–44, July 2024.

Received: 07/06/2024
Revised: 30/06/2024
Accepted: 10/07/2024
Published: 31/07/2024

Abstract

Oral cancer is one of the most dangerous types of cancer that can threaten human health. The diagnosis of cancer and its possible disorder at an early stage is required to increase the survival rate. The transformer models are so popular in computer vision application because of their ability to extract long range relationships among datapoints. The combination of CNN and transformer has attracted extensive study in classifying medical images on the limited dataset. In this work, a lightweight, fast and robust automatic transformer-based network has been designed. The proposed network utilizes the capabilities of CNN models for feature extraction with transformer model to create a fusion of transformer and convolution model for the classification of oral cancer images. The transformer network can capture both the local and global dependence of the image features. The joint efforts of the transformer and CNN in extracting the features from images reduce the computation cost and complexity as well as increase the performance of the presented model. The performance of the model is tested on an unseen test set of a publicly available dataset of oral cancer. The result findings are proving that the combined structure of CNN and transformer model can extract more discriminatory features which helps in improving the classification performance of the model. The comparative analysis with other state-of-the-art models determines that the proposed model achieved competitive performance among these considered models. In addition, the presented approach can be beneficial in the diagnosis system for the detection of oral cancer at an early stage as it is effective and can give more accurate predictions.

Keywords

Deep learning; classification; feature fusion; vision transformer, oral cancer

1. Introduction

Oral cancer is a worldwide health concern. Approximately 3% of cancer cases diagnosed globally are oral cancer [1]. According to the report of the World Health Organization 2020, more than 370000 cases were affected by oral cancer [2]. Oral cancer is a general term that includes cancer of the tongue, gingiva, palate, buccal mucosa, floor of the mouth and lips. Many studies have proven that tongue cancer is one of the most common types of oral cancer [3]. Oral cancer is one of the most common cancers in Asia and primarily affects people from Asia approximately (65.8% of cases) [4] because of their choice of lifestyle which includes strong risk factors for oral cancer such as chain-smoking, alcohol consumption, and betel quid chewing. According to most of the studies, oral cancer is a type of cancer that spreads quickly and affects teeth, two-thirds of the tongue, lip liner, the roof of the mouth, sections of the oral cavity and other parts of the face [5]. Initially, oral cancer appears in red and white patches in the mouth. Tooth displacement and unusually heavy bleeding from the mouth are two main indicators of oral cancer [6]. Early detection of oral cancer presents its growth in early diagnosis and quick action in treatment selection which makes it easy to cure it and rescue a person from cancer and lowers mortality.

Traditionally, the primary factor influencing cancer treatment is the tumour's grade. However, the inconsistent grading makes the prognosis for oral cancer patients even more uncertain [7]. Improved prognostic and diagnostic accuracy helps the physician in making decisions about the best treatment for survival. The use of machine learning techniques in the analysis of oral cancer reported better prognostication when compared to the traditional methods [8]. Recently there has been a significant increase in research on AI-based medical imaging and diagnostic technologies. AI is popular in oncology because it has the potential to increase accuracy in cancer screening [9]. Machine learning-based systems are good at identifying oral, lung and breast cancer. The integration of these methods into the diagnostic system specifically for disease screening in resource-constrained conditions may give better results. Analyzing massive datasets to find cancerous tumors can save time and effort with the help of artificial intelligence [10]. Further, more research is required on the application of artificial intelligence in the early diagnosis of oral cancer along with its effectiveness in recognizing and detecting oral cancer at an early stage.

Song et al. [11] presented a Bayesian deep network to evaluate the uncertainty for assessing oral cancer and utilised a huge intraoral cheek mucosa image for demonstration. A research has been proposed by Tanriver et al.[12] which explains about the possibility of automated system for the diagnosis of oral with the help of computer vision and deep learning methods to identify the type of oral cancer using oral cancer images A deep learning-based method was proposed by Camalan et al. [13] for classifying images as "suspicious" and "normal" by implementing transfer learning on Inception-ResNet-V2. An automatic heat map has been made to indicate the area of the image that most likely contains the decision-making. A novel deep-learning architecture D'OraCa was developed by Lim et al. [14] to categorize oral lesions from the images. This method evaluated the effectiveness of five different deep neural networks and MobileNetV2 was chosen as the feature extractor from the mouth images. The authors Lin et al. [15] proposed a successful smartphone-based image-processing method that employs deep-learning techniques to address the problem of automatically identifying oral disease with a simple and efficient rule-based image-capture method.

There are a number of alternative approaches have been discussed which can monitor the oral cavity automatically and can provide feedback to the medical professionals and as well as to the patient during the identification process of oral cancer. This has become possible with the help of deep learning models and computer vision as it has ability to capture and understand the pattern of diseases from the images and model can learn those patterns which gives the accurate predictions. A number of research has been done for the classification of oral cancer using computer vision task and also including the specialized technology for capturing those images like autofluorescence image, hyperspectral images, and optical coherence tomography image which have been addressed as presented in the study [17-19].In contrast, a number of studies conducted using light weight photography which concentrated on the detection of specific types of oral lesions. Research showed that the deep learning algorithms can outperform the human experts in several aspects related to identification of illnesses. Deep learning techniques have additionally presented the promising results for automatic diagnosis and detection of oral cancer using fluorescence images and confocal laser endomicroscopy (CLE) images [20]. Author presented in the research [21] where two stage method has been proposed for the identification of oral cancer at different stages. In this work segmentation and classification using random forest classifier have been performed to achieve the best performance of 93.24%. A deep learning-based method presented in [22] where a mobile connected to the device for acquiring the florescence oral cancer images and classify them in different categories with the highest accuracy of 86.9%. Although a lot of research have been done in the field of deep learning and computer vision for the diagnosis of oral can-

cer still there is need to design an automated, light weighted model which can provide accurate prediction with the effective computational cost and robustness. As all the existing research require expensive devices for capturing the oral cancer images which is not easily accessible for all and also, they require to modernize for processing so that patient can get the diagnosis report at an early stage without vising to the hospital using automatic system.

With the rapid development of both imaging and intelligent technologies it has become possible for identifying chronic disease with the help of intelligent systems at an early stage with the support of expert. The deep learning architectures specifically CNNs are very popular for providing accurate prediction on image datasets by performing convolution operation and give an intelligent decision. Recently deep learning algorithms have shown promising results to feature-based approaches in medical imaging analysis. A light weight and robust transformer-based network has been designed [23] for the multiclass classification using microscopic images. The presented network is very light, fast, efficient and automatic to perform microscopic image classification with optimized performance. A deep learning based sequential network has been presented [24] for the blood cell classification which cab help in the automatic diagnosis of blood cancer and various haematological malignancies. Therefore, it has been analysed that deep learning-based architectures are very effective in classification and designing ensemble of these architecture may have significant impact for improving the classification performance to identify oral cancer disease.

As mentioned in this section, a lot of efforts have been made to address the challenges of identifying oral malignancy with the help of traditional AI-based methods. The traditional machine learning-based methods are time-consuming and require a lot of pre-processing steps to identify oral cancer.

The main key contribution of the presented study regarding oral cancer classification has considered number of contributions which are discussed as follows:

1. The combination of transformer model with deep learning models VGG16 and ResNet50, presents a new feature fusion technique network which can extract the features from the oral cancer images by fusing the local and global features from different branches.
2. The serial concatenation technique applied at fusion layer enhances the classification accuracy of the proposed model.
3. Different augmentation technique have been employed to increase the sample size of the dataset and to solve the data unbalancing problem.
4. The hybrid approach of deep learning classification, combining various techniques with transformer and other CNNs, is more accurate and efficient than traditional approaches as it can accelerate the diagnosis of oral cancer at early stage.
5. The obtained results are compared with existing state-of the-art deep learning models to prove the effectiveness of the model.
6. The evaluated results of the proposed TransEns network is outperforming the other models which is proving the effectiveness of the model and also significance of the model for the diagnosis of oral cancer at early stage.

This study introduces an automatic deep-learning-based feature fusion technique to classify oral cancer at an early stage using lip and mouth images. The proposed method performs a fusion-based feature extraction process using VGG16, DenseNet121 and ResNet50 models. The features extracted from these deep learning models are concatenated to make the final classification prediction. Moreover, this effective parameter selection has been done to perform hyperparameter tuning well by implementing different optimization algorithms. To examine the effectiveness of the proposed fusion-based method, its performance compared with different pre-trained deep-learning models which proves that the proposed fusion technique outperforms all other models.

The structure organization of remaining work is defines as follows: section 2 illustrates about the materials and methods used for implementing the proposed study with the details of the datasets and defined network. The experimental analysis id presented in section 3 which defines the experiments performed to calculate the performance of the presented network along with their comparisons with other deep learning models. Section 4 conclude the overall about the presented study for oral cancer diagnosis.

2. Materials and Methods

2.1 Dataset Description: The dataset used in this study to implement the presented model is defined in this section. It is a publicly available dataset of oral cancer [16] which consists of images of mouth lip and tongue for the diagnosis of oral cancer. The considered dataset consists of a total of 131 images into two categories cancer and non-cancer. The sample images of the dataset are given in Fig. 1 where (a) represents the cancer class and (b) represents images of the non-cancer class. The images in the considered dataset are in jpg format and have different challenging conditions such as images being inconsistent and having different variations.

Figure 1. Sample images of OCI dataset (a) Cancer (b)Non-cancer

The category-wise distribution of the images among the two classes is defined in Table 1 which have been used for training and testing of the proposed model. The training set of the considered dataset is augmented to increase the size of the number of sample images while training the model and reduce the chance of overfitting.

Table 1. The distribution of oral cancer images among two classes

Category	Number of images
Cancer	87
Non-Cancer	44

2.2 Methodology: The proposed automatic method for oral cancer classification using mouth lip and tongue images is comprehensively defined in this section. The block diagram of the proposed method is presented in Fig.1. which first collected the dataset of the considered problem and performed some pre-pre-processing steps to prepare the dataset for training the proposed model. Data augmentation has been performed to increase the number of samples for better training of the model and to avoid the problem of data unbalancing. After augmentation, the augmented images were fed as input while training the model.

The training images are split into patches as the proposed model is a combination of the transformer network and CNN network. The transformer-based network divides the input images into patches to extract the relevant feature maps whereas the CNN models take the input images directly in the form of pixels without splitting them into patches and can extract the meaningful information. In this study, a combined approach has been used which employs the fusion of a vision transformer network with VGG16 and ResNet50 model for extracting the more discriminating information and then based on their predictions a serial fusion technique has been applied for concatenation and making the final classification.

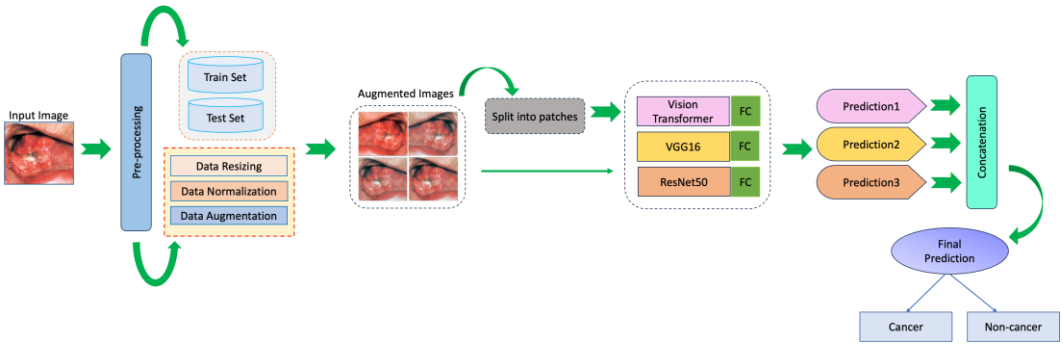

Figure 2. Schematic diagram of the proposed automatic oral cancer classification system

2.3 Proposed Architecture: The overall structure of the proposed model given in this section is presented in Fig. 2 which is an end-to-end network. The presented network is mainly a combination of a vision transformer network and two CNN networks VGG16 and ResNet50 which have been used as feature extractor for extracting the more important features from the images. The features extracted from each model are combined by performing the feature concatenation. The concatenated feature map is used to make the final prediction by employing the SoftMax activation function in the final layer of the network. The SoftMax function is very useful because of its compatible with cross-entropy loss while training the model and it encourages the model to produce probability distributions that are close to the true distribution.

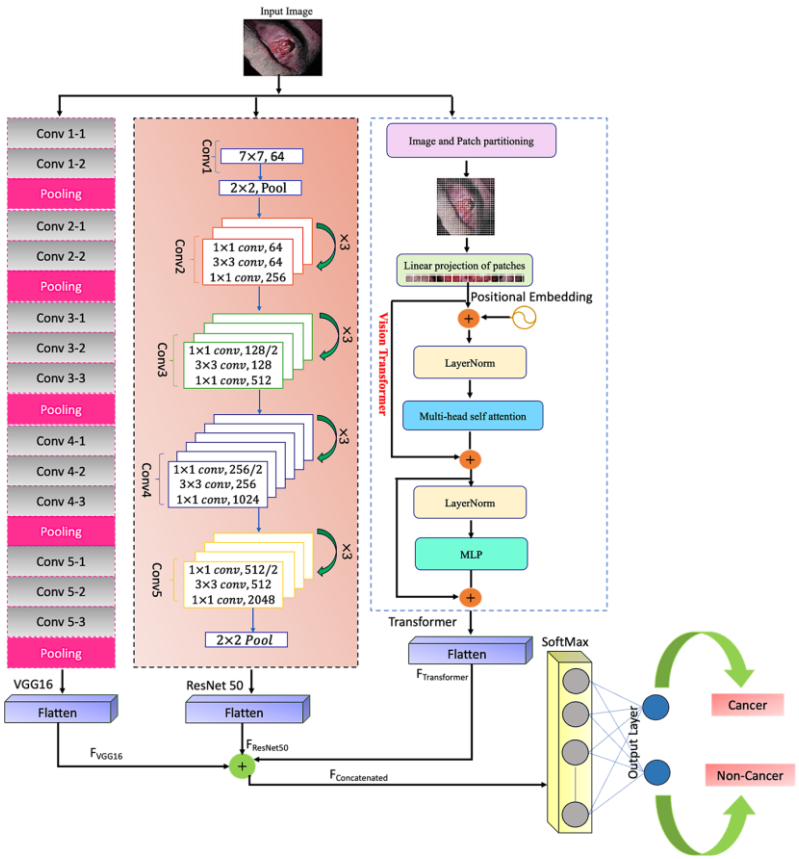

Figure 3. TransEns: Architecture of the proposed transformer based fusion nwteork

The proposed model utilized feature fusion technique to get more relevant information from the images by combining different features extracted from different deep learning models to provide a more comprehensive feature vector, which can perform binary classification effectively for identifying the different classes of oral cancer. Serial and parallel feature fusion are the two most frequently used feature fusion techniques. The serial feature fusion technique has been employed in this model for concatenating the features maps. The feature vectors acquired by the various deep learning models VGG16, ResNet50 and transformer networks are $F_{VGG16:512}$, $F_{ResNet50:2048}$, and $F_{Transformer}$ respectively which are mainly concatenated for the final prediction. The applied concatenation operation is defined in equation no. 1 where $fn(i)$ defines the nth feature vector for the i^{th} samples of $f(i)$ and $\cup$ defines the concatenation operation.

$$f(i) = \cup_{i=0}^{2} fn(i) \quad (1)$$

The proposed methodology is combination of two modules feature extraction module and feature classification module which is focusing on extracting relevant features and the make classification to identify the oral cancer class.

2.4 Feature Extraction Module: The presented network utilizes the kernel size of $\mathbf{3 \times 3}$ and stride of size 2 to perform the convolution operation with input image dimension $\mathbf{224 \times 224 \times 3}$. The features extracted from the transformer and CNN networks are then utilized for feature fusion. Since the dimension of the feature extracted from the transformer network and CNN networks are different so feature concatenation has been applied for fusing the extracted features. The CNN model extracted the features of dimension $\boldsymbol{H \times W \times C}$ where H, W and C stands for height, weight and channel. The dimensions of the feature shape extracted from transformer module is $\boldsymbol{(N+1) \times D}$ where N represents the number of patches and D is for output dimensions. After this the extracted feature map are customized from convolution layer to transformer model and transformed into $\mathbf{7 \times 7}$ patches then these patches down sample through the linear layer to perform the position embedding for feature fusion. The extracted global features from the multi-head self-attention are then projected with convolution layer which helps in extracting the relevant local and global features.

2.5 Feature Classification Module: In this structure the features extracted from different networks are fused to obtain local and global feature and finally the fused feature vector is generated by average pool and the SoftMax layer is applied for the final prediction to generate the classification results as cancer or non-cancer. The proposed fusion model was trained for 100 number of iterations with considering different optimizers SGD, RMSProp and Adam. Out of these models trained with Adam optimizer giving the best classification results with batch size 32 and learning rate 0.0001. The batch normalization is layer is employed to achieve the faster convergence.

3. Experiments and Analysis

This section defines the implementation of proposed transformer based fusion network. To start the experiments first the dataset is collected and prepared as per training requirement and then parameter setting are specified for the classification. Three deep learning models have been considered for designing the fusion of the model. The model having the best performance are then combined with transformer network to perform the feature fusion task. The ensemble of these model is giving best performance as it is giving more accurate prediction with higher classification accuracy and lower computation cost. The system configuration used to perform all the experiments is defined in the Table. 2 which is used for training the proposed network.

Table 2. System configuration used to perform the experiments of the proposed model

Name	Parameter
Computer Operating System	MAC OS
Processor	Apple M1- Chip
RAM	8 GB
Language version	Python 3.6.5
Development environment	Jupyter Notebook, Python 3.7.9, Tensorflow, Keras, Pandas, OpenCV, Numpy, and Matplotlib.

Input dataset	Oral cancer dataset
Input Image Dimension	224×224
Batch size	32
Loss function	Categorical Cross Entropy
Activation function	SoftMax
Step size	0.0001
Number of epochs	100
Optimizer	Adam

3.1 Dataset preparation:

The detailed description of dataset considered for training and designing of the proposed model is defined here in this section. A publicly available dataset OCI of oral cancer having cancer and non-cancer is carried out for the classification task. The dataset consists of 131 images of mouth and lips in which some of them are cancerous and rest are non-cancerous. To prevent the issues that could impact negatively the performance of the proposed model and to maintain the generalizability of the fusion model some pre-processing steps have been performed. Since the images of the OCI dataset have different dimensions, so image resizing has been performed to convert all the images into dimensions of $\mathbf{224 \times 224}$. Data normalization has been performed to standardize the input features which helps certain features from dominating the learning process. The dataset is split in different training and testing ratios for analyzing best training set. Dataset distribution is given in Table.3 where it has been defined that different ratio proportions have been taken into consideration for analyzing the performance of the proposed model on the best sample. On evaluating the performance, it has been examined that the split ratio 7:3 giving the best classification in comparison to other sample distribution. Thus, it has been taken for the final implementation of the proposed model.

Table 3. The sample distribution in train test folder according to considered split ratio

Ratio	**60:40**	**70:30**	**80:20**	**90:10**
Train set	79	92	105	118
Test set	52	39	26	13

Since the dataset is very small in size so different data augmentation techniques have been employed to increase the size of training sample 6 times of the original training set. The data augmentation also helps to handle the problem of data unbalancing which improves the generalizability of the model. The training is done on an augmented train set and the parameter tuning and model evaluation is done with the validation set. Finally, the performance of the proposed model is evaluated on a test set. The sample of augmented images of the dataset for cancer class is represented in Fig.4.

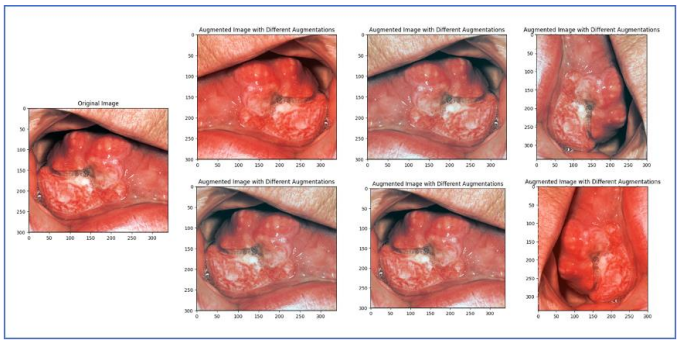

Figure 4. Sample images of oral cancer after augmentation for cancer class

3.2 Evaluation metrics: The performance of the proposed model is calculated on the basis of the considered metrics precision, recall, f1-score and accuracy which are very popular for evaluating performance of a deep learning classifier. Precision defines the ratio of true positive samples to total number of positive samples presents. Whereas the sensitivity defines the number of samples predicted positive correctly out of total positive samples. The f1-score defines the harmonic mean of precision and recall. The accuracy of a network defines the relationship between true positive and true negative to the overall samples. True positive (TP) defines the number of samples predicted positive correctly which are actually positive. True negative (TN) is known for the samples which actually negative and predicted as negative. False positive (FP) is the number of samples which are actually negative and predicted as positive and false negative (FN) defines the number of objects which are actually positive and predicted as negative**.**

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

3.3 Evaluated results:

The calculated results of the proposed fusion network are discussed in this section. The confusion matrix derived from the proposed model for the different training and test sets of the oral dataset is represented in Fig.5.which shows that the best classification results obtained for the with the test set of 30% of total sample. Confusion matrix is a performance evaluation matrix which represents the accuracy of a classification model by calculating the true positive, true negative, false positive and false negative of the data sample. The true positive (TP) is the number of samples classified as positive out of the total positive sample and the true negative (TN) is the number of samples that are not classified correctly. False positive (FP) is the number of samples other than the target class classified correctly. False negative (FN) is the number of samples that are other than the target class and classified incorrectly.

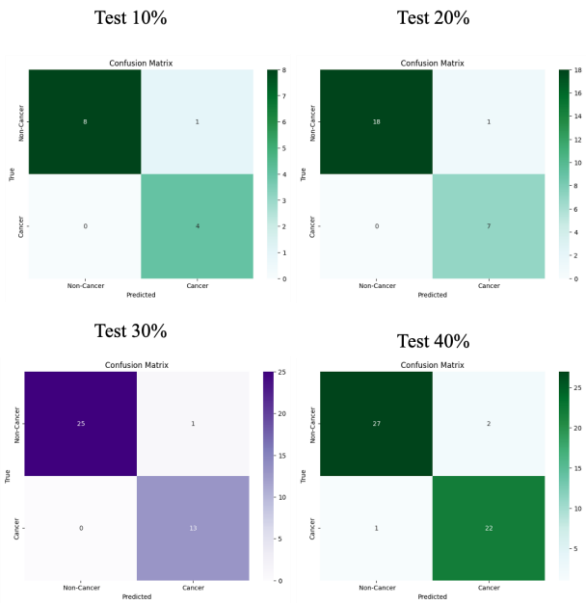

Figure 5. Confusion matrix of the proposed model for different training and test sets

The performance of the proposed fusion network is compared with different deep learning models VGG16, DenseNet121, ResNet50, InceptionV3 and ViT as shown in Table 4 with different evaluation matrices considered for performance evaluation. The calculated results show that the out of all these considered deep learning models ResNet50 and VGG16 are giving best classification accuracy.

Table 4. Comparative analysis of the results obtained through pre-trained deep learning models

Model	Precision	Recall	F1-score	Test Accuracy
ResNet50	81%	100%	90%	87.18%
VGG16	76%	96%	85%	76.92%
DenseNet 121	78%	81%	79%	71.79%
InceptionV3	73%	83%	78%	69.23%
ViT	70%	81%	75%	66.67%
Proposed Fusion Model: TransEns (VGG16+ResNet50+ViT)	96%	98%	97%	**97.43%**

After analyzing the performance of each individual model, a fusion of best performing models is designed to get more discriminatory features of local and global lesions and then realizing the classification performance by aggregating these features. Since ResNet 50 and VGG16 models are on top in terms of performance, hence these models are integrated with transformer model to create transformer based fusion network which help to get more accurate prediction.

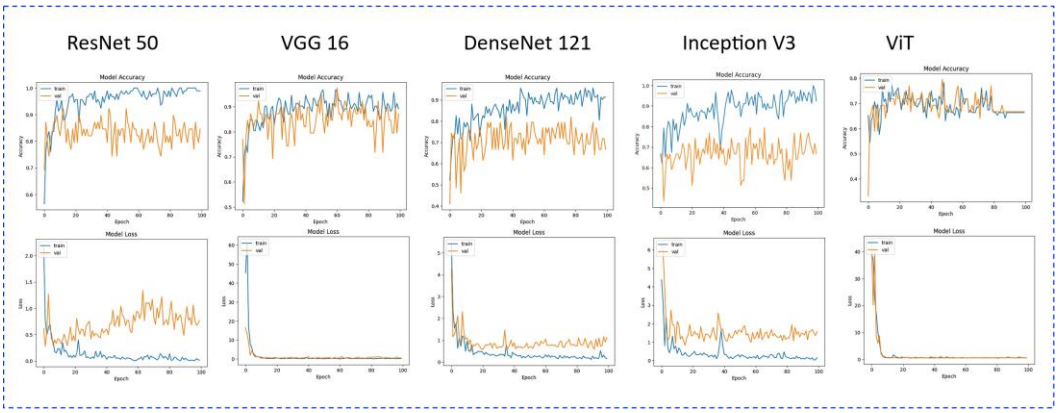

Figure 6. Loss and accuracy curves for different deep learning models

The classification results obtained through these models are presented in Fig.6 which shows the loss and accuracy curve of these considered pre-trained models for individual on which it has been determined the best performing model. Overall, the performance of the proposed transformer based fusion network and comparison with other state of the art deep neural networks have been presented in Table 5 which explains that the performance obtained by the fusion network is best and more accurate in comparison to other models.

Table 5. Performance comparison of the proposed fusion network with other pre-trained models in terms of accuracy

Model	Validation Accuracy	Test Accuracy
ResNet50	92.31%	87.18%
VGG16	92.31%	76.92%

DenseNet 121	79.42%	71.79%
Inception V3	82.05%	69.23%
ViT	79.49%	66.67%
Proposed Fusion Model: TransEns (VGG16+ ResNet50+ViT)	**98.91%**	**97.43%**

3.4 Discussion: Overall discussion about the effectiveness of the proposed fusion network and the impact of each module with their combinations have been discussed in this section. In this study a dataset consisting of oral cancer images of mouth and lip has been considered for the implementation of designed fusion network. The dataset consists of two different classes having names of classes as cancer and non-cancer. A fusion of transformer-based model and deep neural network has been proposed in the work to perform binary classification for early diagnosis of oral cancer. For experimental analysis dataset has been prepared by performing some pre-processing to maintain the consistency of the dataset. To select the best sample size of the dataset, it has been divided in multiple ratios, out of which best combination of training and testing folder (7:3) has been chosen for further experiments and comparison. To solve the problem of data unbalancing different augmentation techniques have been employed.

The rationale behind choosing the VGG16 and ResNet50 for designing the fusion network is that instead of taking large number of parameters these models are taking small size kernel and stride 1 with same padding and specifically implemented for RGB images of size 224×224. So, these models were tuned up for the considered dataset of oral cancer where it has been seen that both the models are giving best classification results. The transformer networks excel in handling the long range dependencies between input sequences which enable the parallel processing. This making them highly effective in the classification task. Designing a fusion network from multiple modalities gives the advantage of the complementarity of data in order to provide a better performance by improving imaging quality while preserving the specific features for accurate diagnosis. The hybrid approach combines the feature extraction capabilities of CNNs with the long-range relationship-capturing ability of transformers. The proposed model achieves high classification accuracy with minimal computational cost, demonstrating its potential for early diagnosis of oral cancer.

For evaluating the performance of the presented model different performance measures have been considered and the compared with other existing methods. The performance of the proposed fusion network is compared with different deep learning models VGG16 [25], DenseNet121 [26], DenseNet169 [26], ResNet50 [27] and InceptionV3 [28] to analyze the performance of each individual model for classification of oral cancer images so that an efficient fusion network can be prepared. The results obtained through the proposed network is presented in Fig.7 with confusion matrix and loss accuracy curve for the classification of oral cancer. It has been determined that the performance of the proposed feature fusion network have achieved the best classification results with validation accuracy of 98.91% and test accuracy of 97.43% by outperforming the existing method. The outcomes of the proposed model with its confusion matrix defines that the proposed fusion model is providing best results by predicting 38 samples positive out of total 39 test sample as presented in confusion matrix.

To get generalized and optimized performance of the model different parameters have been considered randomly on which performance of the model is analysed. For the experiment purpose, different kernel sizes having different learning rates of 0.001, 0.0001 with different batch sizes 16, 32 and different optimizers have been considered. The hyperparameter tunning has been done for evaluating the best performance of the model. The proposed fusion model is comprised of many layers in different configurations of filters to capture high-level and low-level features which makes the model more capable to perform binary classification for categorizing oral cancer images. The proposed fusion model performs well on unseen test sets of the dataset which makes the model more generalized and effective for making accurate predictions. The results validated the performance of the presented model for accurate automatic classification of oral cancer using mouth and tongue images.

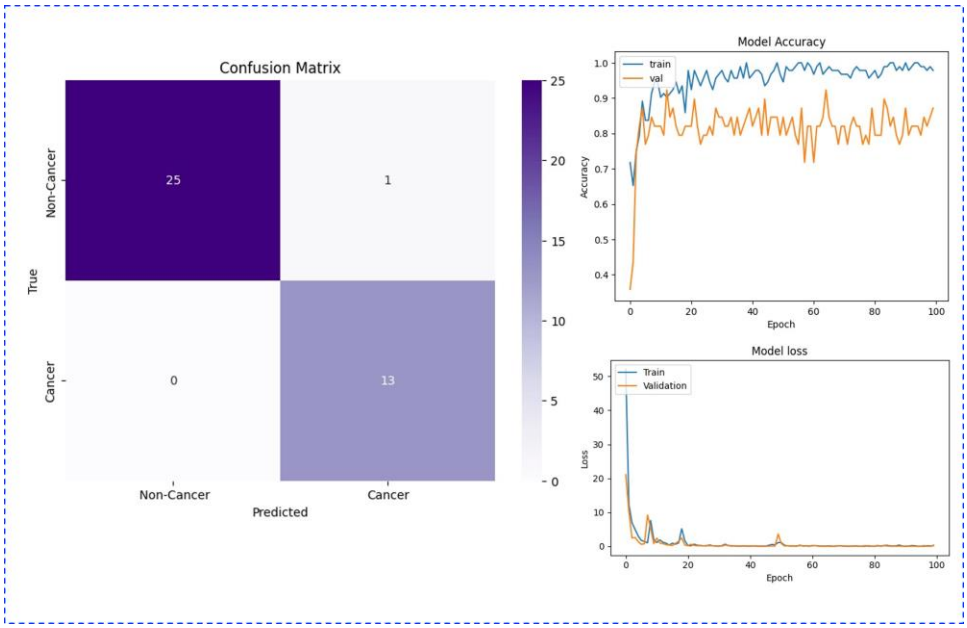

Figure 7. Results of the proposed fusion model with loss accuracy curve and confusion matrix for automatic oral cancer classification

4. Conclusion

A transformer based deep fusion network TransEns has been proposed in this work for the binary classification of oral cancer using mouth lip and tongue images. The presented model is performing well on a public dataset of OCI having mouth lip and tongue images due to the distinct feature extraction capabilities of CNN and transformer. CNN models are effective at extracting the local features while the integration of transformer network increases the model's ability to extract more relevant information by detecting the global features. Different augmentation techniques have been employed to solve the problem of data unbalancing. To design the fusion network mainly three deep learning-based model have been considered are VGG16, ResNet50 and vision transformer model because of their feature extraction capability and classification performance. Before designing ensemble of these models, features have been extracted for each individual model for getting more prominent information. Finally, the extracted features are concatenated using serial concatenation operation for creating the fusion of the model and giving the more accurate prediction. To make the model more reliable and robust hyperparameter optimization has been done which is optimizing the number of trainable parameters of the fusion model for the binary classification. The batch normalization layer is applied to reduce the overfitting of the network and to achieve the faster convergence. Overall, the results obtained through the proposed network has been compared with state-of-the-art works to analyze the significant impact of the proposed transformer based fusion network over the other existing method for the classification of oral cancer using mouth lip and tongue images. The results finding reflects that the proposed model has achieved the best classification results with an accuracy of 98.91% which is outperforming the existing methods. The implementation of this proposed network for the real-time operations may help in medical diagnosis for identification of oral cancer at early stage.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394-424. https://doi.org/10.3322/caac.21492 (2018).

[2] Ren, Z. H., Hu, C. Y., He, H. R., Li, Y. J., & Lyu, J. (2020). Global and regional burdens of oral cancer from 1990 to 2017: Results from the global burden of disease study. *Cancer Communications*, *40*(2-3), 81-92 https://doi.org/10.1002/cac2.12009 (2020).

[3] Ibayashi, H., Pham, T. M., Fujino, Y., Kubo, T., Ozasa, K., Matsuda, S., & Yoshimura, T. (2011). Estimation of premature mortality from oral cancer in Japan, 1995 and 2005. *Cancer Epidemiology*, *35*(4), 342-344.https://doi.org/10.1016/j.canep.2011.01.010 (2011).

[4] Rao, S. V. K., Mejia, G., Roberts-Thomson, K., & Logan, R. (2013). Epidemiology of oral cancer in Asia in the past decade-an update (2000-2012). *Asian Pacific journal of cancer prevention*, *14*(10), 5567-5577.https://doi.org/10.7314/apjcp.2013.14.10.5567 (2013).

[5] Razmjooy, N., Sheykhahmad, F. R., & Ghadimi, N. (2018). A hybrid neural network–world cup optimization algorithm for melanoma detection. *Open Medicine*, *13*(1), 9-16.

[6] Huang, Q., Ding, H., & Razmjooy, N. (2023). Optimal deep learning neural network using ISSA for diagnosing the oral cancer. *Biomedical Signal Processing and Control*, *84*, 104749. https://doi.org/10.1016/j.bspc.2023.104749

[7] Kirubabai, M. P., & Arumugam, G. (2021). Deep learning classification method to detect and diagnose the cancer regions in oral MRI images. *Med. Leg. Update*, *21*, 462-468.

[8] Gupta, R. K., & Manhas, J. (2021). Improved classification of cancerous histopathology images using color channel separation and deep learning. *Journal of Multimedia Information System*, *8*(3), 175-182.

[9] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ... & Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, *24*(9), 1342-1350.

[10] Ilhan, B., Lin, K., Guneri, P., & Wilder-Smith, P. (2020). Improving oral cancer outcomes with imaging and artificial intelligence. *Journal of dental research*, *99*(3), 241-248.

[11] Song, B., Sunny, S., Li, S., Gurushanth, K., Mendonca, P., Mukhia, N., ... & Liang, R. (2021). Bayesian deep learning for reliable oral cancer image classification. *Biomedical Optics Express*, *12*(10), 6422-6430.

[12] Tanriver, G., Soluk Tekkesin, M., & Ergen, O. (2021). Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers*, *13*(11), 2766.

[13] Camalan, S., Mahmood, H., Binol, H., Araujo, A. L. D., Santos-Silva, A. R., Vargas, P. A., ... & Gurcan, M. N. (2021). Convolutional neural network-based clinical predictors of oral dysplasia: Class activation map analysis of deep learning results. *Cancers*, *13*(6), 1291.

[14] Lim, J. H., Tan, C. S., Chan, C. S., Welikala, R. A., Remagnino, P., Rajendran, S., ... & Barman, S. A. (2021). D'OraCa: deep learning-based classification of oral lesions with mouth landmark guidance for early detection of oral cancer. In *Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25* (pp. 408-422). Springer International Publishing.

[15] Lin, H., Chen, H., Weng, L., Shao, J., & Lin, J. (2021). Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. *Journal of Biomedical Optics*, *26*(8), 086007-086007.

[16] https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images

[17] Wilder-Smith, Petra, et al. "In vivo diagnosis of oral dysplasia and malignancy using optical coherence tomography: preliminary studies in 50 patients." *Lasers in Surgery and Medicine: The Official Journal of the American Society for Laser Medicine and Surgery* 41.5 (2009): 353-357.

[18] Heidari, Andrew Emon, et al. "Optical coherence tomography as an oral cancer screening adjunct in a low resource settings." *IEEE Journal of Selected Topics in Quantum Electronics* 25.1 (2018): 1-8.

[19] Jeyaraj, Pandia Rajan, and Edward Rajan Samuel Nadar. "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm." *Journal of cancer research and clinical oncology* 145 (2019): 829-837.

[20] Tschandl, Philipp, et al. "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study." *The lancet oncology* 20.7 (2019): 938-947.

[21] Das, Dev Kumar, et al. "Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis." *Tissue and Cell* 53 (2018): 111-119.

[22] Song, Bofan, et al. "Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning." *Biomedical optics express* 9.11 (2018): 5318-5329.

[23] Dwivedi, Karnika, Malay Kishore Dutta, and Jay Prakash Pandey. "EMViT-Net: A novel transformer-based network utilizing CNN and multilayer perceptron for the classification of environmental microorganisms using microscopic images." *Ecological Informatics* 79 (2024): 102451.
[24] Dwivedi, Karnika, and Malay Kishore Dutta. "Microcell-Net: A deep neural network for multi-class classification of microscopic blood cell images." *Expert Systems* 40.7 (2023): e13295.
[25] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
[26] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[27] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.