JEI Journal of Electronic Imaging

RESEARCH PAPER

Oral cancer detection using transfer learning-based framework from histopathology images

Dawit Kiros Redie,* Saurabh Bilgaiyan, and Santwana Sagnika
KIIT University, School of Computer Engineering, Bhubaneswar, Odisha, India

ABSTRACT. Oral cancer is a serious worldwide health problem. It might be seen in the face, oral glands, neck, or mouth of the patient. Cancer detection utilizing histopathology images aids in easing and forecasting abnormality. Furthermore, it delivers better outcomes if biological procedures are implemented appropriately, however, there are large opportunities for human errors and blunders during physical examinations. The advancement of deep learning techniques may open the door to a better detection of oral cancer from histopathology images, which will be advantageous to lab staff and medical professionals. Our study presents an in-depth analysis of 10 pre-trained deep convolutional neural network models using transfer learning approach for the detection of oral cancer. The experiment was performed for two classes, i.e., normal and oral squamous cell carcinoma. The results show that the VGG19 model with data augmentation was able to attain the highest classification accuracy of 96.26% using the transfer learning technique. We have also introduced an approach by integrating the VGG19 pre-trained model with a custom naïve inception block. This fusion of two well-known models harnesses their complementary strengths and results in a more robust architecture. The incorporation of the inception block addresses limitations observed in the VGG19 framework, such as vanishing gradient issues and excessive computational requirements. The proposed model for detection of oral cancer has undergone a thorough block-wise fine tuning. Our results demonstrate the superior performance of our deep learning architecture compared to existing literature, highlighting its potential to enhance the detection and diagnosis of oral cancer.

© 2023 SPIE and IS&T [DOI: 10.1117/1.JEI.32.5.053004]

Keywords: transfer learning; deep learning; oral squamous cell carcinoma; medical imaging; data augmentation; convolutional neural network

Paper 230004G received Jan. 6, 2023; revised Jul. 29, 2023; accepted Aug. 28, 2023; published Sep. 7, 2023.

1 Introduction

Oral cancer is one of the most frequent cancers in the world, with a high mortality and late detection rate. It is a subcategory of neck and head cancers, with around 475,000 new cases identified annually across the globe.[1] The premature disease has an ~80% survival rate, but the late-stage disease has a survival rate of fewer than 20%.[1] Squamous cell carcinoma is the most prevalent kind of head and neck cancer. Ninety percent of head and neck cancers begin in squamous cells, which line the oral, nasal, and pharyngeal cavities.[2] Multiple risk factors are associated with oral cancer, and the survival rate after treatment is also unexpectedly low.[3] Potentially malignant lesions, such as oral submucosal fibrosis, erythroplasia, and leukoplakia, are frequent

*Address all correspondence to Dawit Kiros Redie, 2163004@kiit.ac.in

1017-9909/2023/$28.00 © 2023 SPIE and IS&T

types of precancer disease that may leads to oral cancer. Additionally, it is essential to differentiate between benign and malignant tumours. The prognosis of oral cancer may be influenced by age, gender, and smoking history.[4] An open sore is the most prevalent symptom of oral cancer when it does not heal and can be painful. In addition, red or white lesions around the mouth, swollen jaw, and loss of tooth are symptoms of oral cancer.[2] There are numerous procedures for diagnosing this serious malignancy, including a biopsy in which microscopic tissue samples are removed from the mouth and inspected under a sterile microscope. These images that are obtained from the microscope are called histopathological images. Histopathological images are often used to distinguish between lesions that are benign, precancerous, and cancerous in the mouth. Using histopathological images for cancer screening assists in alleviating and predicting abnormalities. Moreover, it yields superior results if biological methods are appropriately conducted. However, this method is constrained by arbitrary interpretations and insufficient diagnosis. There are also obstacles associated with manually detecting cancer using histopathological biomedical images owing to the high possibility of incorrect identification due to human error. Thus understanding technological breakthroughs, such as artificial intelligence, may aid in resolving healthcare issues.[5] A computer-aided cancer diagnostic system that automates the procedure effectively is necessary to eliminate the laborious, time-consuming effort of diagnosing cancer. Deep learning using convolutional neural networks (CNNs) is a cutting-edge technology for processing and analyzing many medical images.[6] CNNs are currently the most effective form of a deep learning model for image analysis. CNNs comprise several layers that alter their input using small-scale convolution filters and analyze multiple arrays containing feature information extracted from given input. Convolutional neurons analyze the input once the images are formulated into a feature map during development. CNN networks can be broadly divided into two training approaches. The first approach involves training networks from scratch, starting with randomly initialized weights. In contrast, the second approach utilizes transfer learning, where pre-trained networks,[7] trained on large datasets, are employed to transfer knowledge from one domain to another. Transfer learning enables the leveraging of information learned from previous tasks, improving performance and efficiency in new domains. The transfer learning strategy, which utilizes pre-trained models, offers the advantage of transferring trained weights from a model learned in one discipline to another medical area. This approach aids in reducing the training time and computational expenses required to train the network from scratch. We have performed our experiment in two categories. In the first category, to save the training time and computing power, we have done our experiment based on the transfer learning strategy. We have taken the following well-known pre-trained networks, i.e., DenseNet121, DenseNet169, DenseNet201, MobileNetv2, ResNet50, ResNet101, Xception, AlexNet, VGG16, and VGG19. We have frozen most of the layers of each model to prevent it from training. In addition, we have used fine-tuning to adjust the weights and parameters of each layer.

Further in the second category, we introduce an approach by integrating the VGG19 pre-trained model with a custom naïve inception block. This fusion of two well-known models allows us to harness their complementary strengths and overcome limitations observed in the VGG19 framework. The inclusion of the inception block addresses challenges, such as vanishing gradient issues and excessive computation, resulting in improved learning capacity and performance of our proposed model. By combining these contributions, we present a robust deep learning-based model for oral cancer detection. Our approach enhances the accuracy, efficiency, and generalization capabilities of the detection model, facilitating early diagnosis, personalized treatment planning, and improved patient care. We believe that the integration of tailored data preprocessing techniques, comprehensive model evaluation, and the fusion of pre-trained models with a custom inception block sets our research apart and contributes to the advancement of oral cancer detection methodologies.

The remainder of this research is structured as follows. Section 2 outlines the current research works of deep learning techniques for the detection of oral cancer. In Sec. 3, the datasets used in the experiment are presented and the learning mechanism employed in the proposed methodology are explained in depth. In addition, Sec. 4 presents the theoretical background of our proposed method. Furthermore, experimental setup and analysis are given in Sec. 5 and Sec. 6 explicates the result and discussion of the proposed model. Finally, the conclusion and future work are presented in Sec. 7.

2 Related Works

A vast number of research studies based on deep learning models were carried out regarding the analysis and diagnosis of oral cancer utilizing histopathology images over the past few years.

Alhazmi et al.[8] used an ANN-based prediction model to detect oral cancer. They studied a total of 29 variables associated with oral cancer patients. They achieved an accuracy of 78.95% for oral cancer prediction. However, their study had limitations, such as the absence of data pre-processing and the lack of a robust predictive model to improve oral cancer screening and diagnosis.

Furthermore, Bansal et al.[9] proposed deep hybrid transfer learning techniques for classifying and predicting oral cancer. They employed five pre-trained models and applied them to real-time and histopathological datasets. They implemented Gaussian blur for image pre-processing and tested different optimizers such as, SGD, ADAM, and RMSprop to maximize model performance. They achieved an accuracy of 92.41% for oral cancer detection from histopathological datasets.

Additionally, Welikala et al.[10] proposed two deep learning-based methods for the automated diagnosis and classification of oral cancer. They used Faster R-CNN for object identification and ResNet-101 for image classification. The $F1$ score for image classification was 87.07% for identifying images with lesions and 78.30% for identifying images that required referral. They achieved an $F1$ score of 41.18% for detecting lesions that needed referral. Their study involved data pre-processing and decision criteria for learning.

Moreover, Shavlokhova et al.[11] utilized a CNN-based model called MobileNet for the classification of oral cancer. They achieved a specificity and sensitivity of 0.96 and 0.47, respectively. Aubreville et al.[12] employed a deep learning-based model for oral cancer detection, achieving an accuracy of 88.3%.

Furthermore, Tanriver et al.[13] proposed a two-stage deep learning framework for detecting and classifying oral potentially malignant disorders. They used YOLO5I for lesion detection and EfficientNet-b4 for classifying the detected lesions. They reported classification results for different models used in their experiments and achieved an $F1$-score of 85.8% using the EfficientNet-b4 model.

The biggest challenge of employing deep learning models is gathering enough accurately recorded items for effective training and avoiding unbalanced data at the time of training. Most of the above studies require data pre-processing, which is the main step while implementing a deep learning model. In addition to this, most of the above models are not robust, so we cannot reuse them on any other dataset or any medical research area.[14] Hence, we propose a transfer learning-based model that overcomes the above limitations of existing literatures. The main contributions of our proposed study are the followings.

1. Data pre-processing is done at the initial stage, and we have introduced a deep learning model, which is based on a transfer learning approach that can correctly classify oral cancer. We have combined the VGG19 pre-trained model with naïve inception block to take the advantage of the goodness of both models.
2. A detailed comparative study of 10 pre-trained models that are trained on ImageNet has been conducted through applying them on our histopathological images. The results obtained from the above experiment were compared with the proposed model. We were able to achieve better results.
3. The proposed model has been trained and tested using histopathological images. The classification result shows that there is a significant enhancement in the performance of the model in comparison to other well-known methods.
4. In order to minimize the computational complexity and increase the resilience of the proposed model, the proposed network has been constructed by combining the appropriate mix of layers in a well-organized method.

3 Materials and Methods

3.1 Dataset Description

Every area of medical image diagnosis is now extremely simple to analyze and anticipate with the use of medical radiography thanks to the advancement of machine learning and deep learning

Table 1 Overall summary of histopathological images used without data augmentation.

Image category	Train	Test
OSCC	747	187
Normal	58	232

Table 2 Overall summary of histopathological images used with data augmentation.

Image category	Train	Test
OSCC	2158	540
Normal	1995	499

techniques. While there are many machine-learning-based approaches available to identify medical conditions, deep learning produces good findings with a very efficient methodology. Deep learning plays a significant role in every part of medical image analysis with excellent validation accuracy and real outcomes. In this research, the diagnosis of oral cancer was accomplished by the use of histopathological images obtained from various sources. The first source is a histopathological image repository of oral squamous cell carcinoma.[15] There are 1224 images available in this dataset. The images were obtained using a Leica ICC50 HD microscope using tissue slides that had been gathered, processed, and classified by medical professionals from 230 individuals. Additional images were collected from Kaggle repository.[16] The dataset in Ref. 15 was highly unbalanced in which the number of images in OSCC class were 934, whereas the number of images in normal class were just 290. To avoid this highly unbalanced problem and in order to substantially increase the number of input images, we have applied a data augmentation approach on the training dataset that will also enhance the model accuracy through minimizing the overfitting problem. We have also resized image input to match the size of an image input layer that can enhance desired features and reduce artifacts that can bias the model. In addition, Gaussian blur and other noise-reduction preprocessing methods were applied to the images. The overall summary of our dataset is presented in Tables 1 and 2. The sample for the histopathological images of each case is shown in Fig. 1.

3.2 Transfer Learning

In recent years, deep learning algorithms have achieved extremely good performance in almost all areas of challenges, including recommendation systems,[17] emotion identification,[18] natural language processing,[19] audio recognition,[20] image recognition,[6] and image segmentation.[21] Deep learning is a learning approach that eliminates complicated hand-crafted extraction of features. As the layer of abstraction grows, it discovers features at various levels of abstraction. A CNN is a deep learning architecture designed for processing and analyzing structured grid-like data, such as images. It consists of multiple layers of interconnected neurons organized in a hierarchical manner. The key component of a CNN is the convolutional layer, which applies filters to input data, extracting relevant features through convolutional operations. These features are then combined and passed through subsequent layers, such as pooling layers for downsampling and nonlinear activation functions. CNN is the most well-known deep learning method and often used neural network for image classification. Convolutional, pooling, and fully connected layers constitute most of its layers. Nevertheless, in order for CNN to train well, a large, labelled dataset like ImageNet[22] is required, which is a difficult challenge in the area of oral pathology. Aside from that, the effectiveness of CNN on small datasets is not promising because of overfitting. Transfer learning is an approach that includes using previously trained networks to initialize CNN weights.[23–26] It has been discovered that the effectiveness of transfer learning

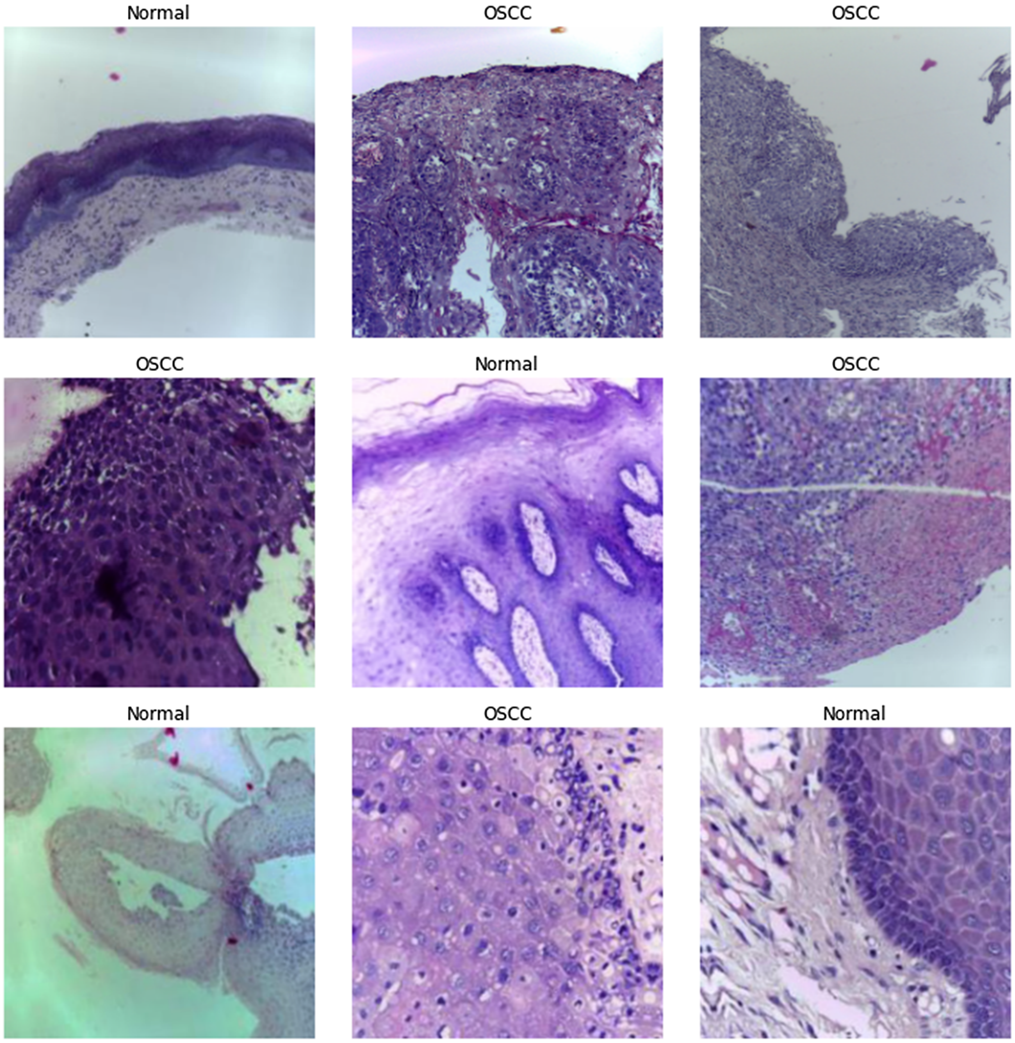

Fig. 1 Sample histopathological images from our database.

outperforms training the model from scratch on a small dataset. This is because transfer learning makes use of previously learned information. We employed 10 deep CNN models, each of which was pre-trained on a large ImageNet dataset. These are DenseNet121,[27] DenseNet169,[27] DenseNet201,[27] MobileNetv2,[28] ResNet50,[29] ResNet101,[29] Xception,[30] AlexNet,[31] VGG16,[32] and VGG19.[32] DenseNet[27] is one of the network architectures that combines thick layers. The basic diagram of DenseNet is presented in Fig. 2. In each dense layer, convolutional and pooling operations are present. DenseNet is capable of reusing the feature, resulting in the model having fewer parameters than other models. Across the thick layers, a certain interconnection arrangement is used to effectively handle the flow of data. This is accomplished by building a meaningful connection between the connecting layers and the feature map. Each convolutional block's output will be utilized as the input for the succeeding layer, whereas MobileNets are a tiny network that suits the constraints of having minimal resources while optimizing the latency. MobileNets have been developed specifically for use in mobile and visual applications. MobileNetv2 is an extraordinarily efficient feature extractor for segmentation of images, object recognition, and image classification. ResNet[29] was developed to perform image identification problems more reliably owing to its high numbers of layers. ResNet offers many architectures, including ResNet101, ResNet50, and ResNet50V2. As shown in Fig. 3, the residual block is made up of 1×1 convolutional layers.

The ResNet50 and ResNet101 models were used in this research and the ReLU activation function and batchnorm layers are added after each 3×3 convolutional layer. Xception[30] is based on InceptionV3, which employs linear layer of depth wise convolutional layers along with

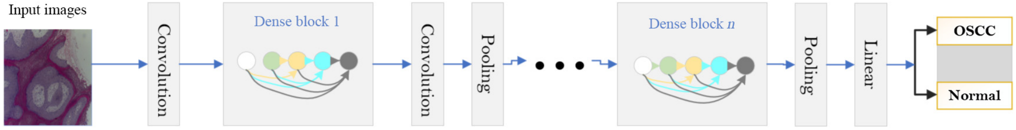

Fig. 2 Basic diagram of DenseNet with n dense blocks.

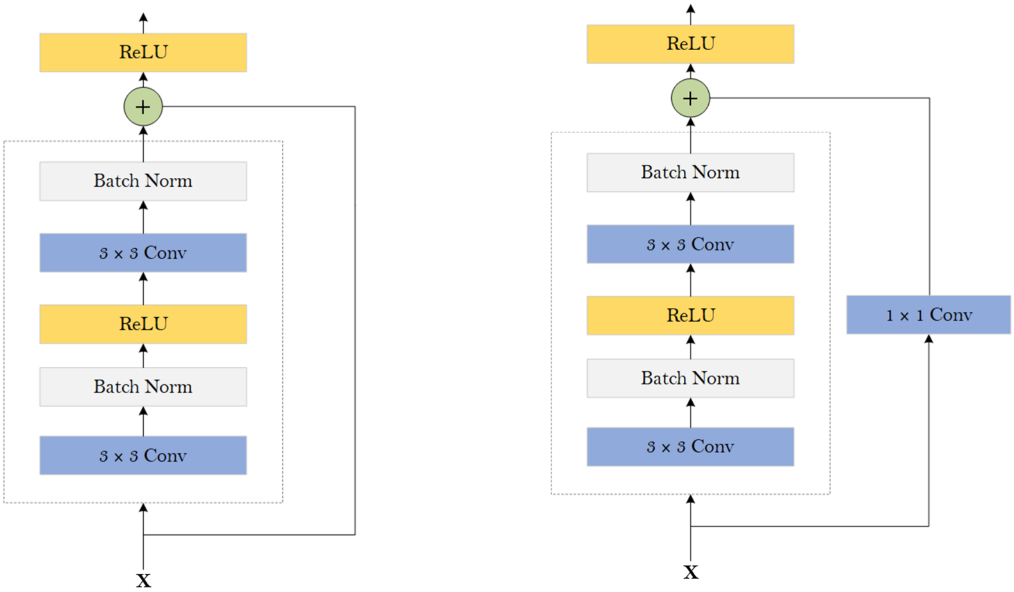

Fig. 3 Basic diagram of residual networks (a) without 1 × 1 convolution and (b) with 1 × 1 convolution.

residual connections to decrease computational complexity. It can be employed to detect and classify different types of disease in a medical sector. The CNN structure known as VGG19[32] is a deep learning model with 19 layers and a very conventional architecture. The model makes predictions using a set of weights that have been previously trained using ImageNet. The default size for the input of an RGB image is 224 × 224 pixels, and it has three channels. VGG network training utilizes a model with 3 × 3 convolutional layers with stride 1 and one max pooling layer with two 2 × 2 stride 2 filters. The depth of these layers increases as they are stacked. The basic block diagram of VGG networks is given in Fig. 4. Generally, the transfer learning approach was used to address the issues of insufficient data and processing time through utilizing ImageNet database. For each model, the weights from the ImageNet training were obtained. Input images are abstracted into feature maps, which are then used as input shapes for different layers during training stage. Table 3 displays the pre-trained model with their output shape and number of parameters for transfer learning. The input shape represents the dimensions or structure of the data that the pre-trained model expects as input. It specifies the size, channels, and other relevant properties of the input data that should be provided to the model for accurate and consistent processing. Following extensive investigation, the fine-tuning criteria were devised. Certain layers of the model, which are untrainable layers beginning at the bottom of the CNN, were frozen. This is advantageous since the weights of these layers are not

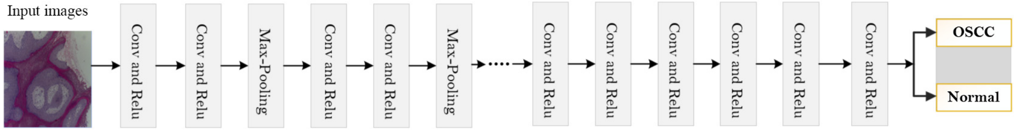

Fig. 4 Block diagram representation of VGG networks.

Table 3 Pre-trained model with their output shape and number of parameters for transfer learning.

Pre-trained model	Input shape	Trainable parameters	Untrainable parameters
DenseNet121	[112, 112]	1,138,882	6,870,208
DenseNet169	[112, 112]	1,871,554	12,326,080
DenseNet201	[112, 112]	2,205,378	17,863,872
MobileNetv2	[112, 112]	1,352,514	2,189,760
ResNet50	[112, 112]	2,161,026	23,454,912
ResNet101	[112, 112]	2,213,250	42,394,816
Xception	[299, 299]	724,384	10,024,384
AlexNet	[55, 55]	265,730	2,469,696
VGG16	[224, 224]	569,906	32,024,384
VGG19	[224, 224]	539,906	20,024,384

anticipated to change during the training phase of the model. In addition, in the pretraining step, the final feature map is flattened and sent to a fully connected layer.

4 Proposed Work

A unique deep learning-based model is proposed by combining the VGG19[32] network with naïve inception module.[33] We have selected VGG19 from the rest pre-trained networks because it achieved the highest accuracy.

The proposed architecture is built as shown in Fig. 5 by first stacking the VGG19 layers, beginning with the VGG19 pre-process layer and continuing through the fourth block of pool layers. The VGG19 pre-trained network receives the $224 \times 224 \times 3$ image as input. We have considered up to the fourth block of VGG19 because it is useful to identify the most important initial features of the input data. Additionally, the inclusion of blocks after the fourth block of VGG19 would simply make the computation more complex and would not contribute to the performance of the model as per our experimental findings. Furthermore, the layers of the naïve inception block are concatenated with the essential bottleneck features that were collected up to the fourth block of pool layer of the VGG19 model. The convolutional layers of the naïve inception block include filter sizes of 1×1, 5×5, and 3×3, with each layer having a stride size of 3×3 and a leaky rectified linear unit activation function. Additionally, the naïve inception block includes a maxpool 2D layer. LeakyReLU is a modification of ReLU activation function, which is used to avoid dying neurons using a small epsilon value in the negative half of its derivative.

In this research, we have created a more robust architecture that successfully handles the class imbalance issue by taking into account the merits of both of the VGG19 and inception. We used VGG19 pre-trained layers in the first stage of our network because the VGG19 model works well for image recognition task and is good at handling various kinds of images. The main

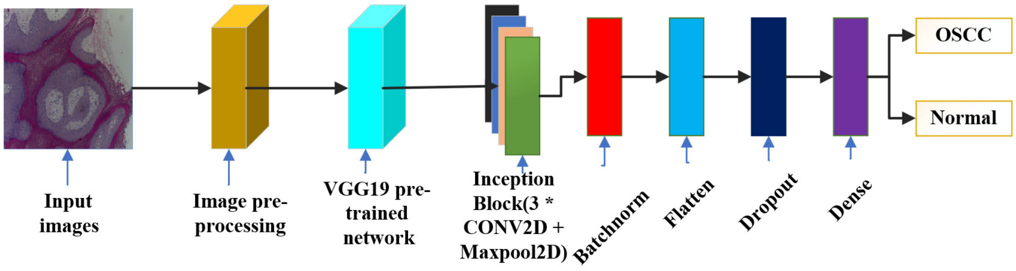

Fig. 5 Detailed architecture of proposed model.

strategy of the proposed framework is to use the most relevant part of pre-trained models by combining two well-known models and use different techniques to extract better features so that the model could perform better with histopathological datasets.

Considering the benefit of both models and to circumvent the difficulties associated with the deployment of VGG networks, we have altered the VGG19 framework. These processing issues are difficult to avoid even on powerful processors because of its enormous memory usage. Therefore, we have retrieved the key features using the VGG19 model until the fourth block and eliminated layers that come after the fourth block that caused excessive computation and complexity issues. Furthermore, we have included the naïve inception block in the higher level of proposed network in order to compensate for the limitations that were observed in VGG19 framework. The addition of inception block can be used to solve the vanishing gradient issue. In general, the addition of the naïve module has been done with the intention of improving the CNN's capacity for learning as well as dealing with larger number of filters, both of which were discovered to be deficiencies in VGG networks. Moreover, other benefit of adding the inception block is that it can still achieve good performance even with just one fully connected layer. In addition, a number of higher layers with random initialization are added, including the dense layer as well as the batch normalisation, flatten, and dropout layers to enhance the performance of the model.

5 Experimental Setup and Implementation

The experiments in this study were implemented using the Python programming language. All tests were done on Google Collaboratory using a Tesla K80 GPU graphics card, an Intel i5-core at 3.8 GHz CPU, and 8 GB RAM running on 64-bit Windows 11.

Beforehand, we conducted many experiments on a histopathology dataset of oral cancer samples fed into 10 well-known pre-trained CNN models, namely, DenseNet121, DenseNet169, DenseNet201, MobileNetv2, ResNet50, ResNet101, Xception, AlexNet, VGG16, and VGG19, each with a varying number of trainable convolution blocks and selection of frozen layers. All models underwent fivefold cross-validation training. Additionally, we selected each model's greatest performance on its own, and then we assessed each model on the test set. From our histopathological dataset, 80% of images were used for training our model, whereas the rest 20% were used for testing purpose. The training period was excessively long due to the more than 5000 histopathological images utilized for binary classification. Therefore, for binary classification, 50 epochs were employed. We also employed regularization and early stopping approaches to avoid overfitting.

The various hyperparameters were chosen to get the best possible performance out of the model. The adam optimizer was used. We have used the learning rate finder from fastai library to get the optimal learning rate. It is used to reduce the amount of guessing involved in selecting an appropriate beginning learning rate. Flattened loss was used as loss function and the batch size of 64 was chosen for training. Several assessment criteria, including accuracy, precision, $F1$ score, and kappa score, were used to determine the model's effectiveness. These metrics depend on the true positives, true negatives, false positives, and false negatives parameters. They are explained as follows. The number of inputs classified as true, which are true predictions, are called true positives; whereas true negatives are the number of inputs classified as false, which are correct predictions. False positives are the number of inputs predicted as true, which are wrong predictions. False negatives are the number of inputs predicted as false, which are wrong predictions.

Precision, as shown in Eq. (2), is defined as the ratio of correctly categorized positive inputs to the whole dataset's positive inputs. The ratio of accurate predictions to all the available data items in the dataset is known as accuracy. The mathematical equation of accuracy is presented in Eq. (1). It indicates the overall effectiveness of the model:

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total samples}}, \tag{1}$$

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}, \tag{2}$$

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}, \tag{3}$$

$$F1\text{-score} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}, \tag{4}$$

Furthermore, sensitivity, also known as recall or true positive rate, measures the proportion of actual positive cases correctly identified by a model. $F1$-score is a metric that combines precision and recall to provide a balanced measure of a model's performance. The mathematical equation for sensitivity is presented in Eq. (3), and the equation for $F1$-score is presented in Eq. (4) in this paper.

The kappa score contrasts the model's accuracy with the efficiency of a random system. It controls objects that may be accurately categorized by accident by measuring the correspondence between the classified items and the ones designated as ground truth. A kappa number of 0 indicates there is no match, whereas a kappa value of 1 indicates a perfect match. The formula for kappa score is given by the following equation:

$$\text{kappa score} = \frac{T(\text{TP} + \text{TN}) - Z}{T^2 - Z}, \tag{5}$$

where $Z = (\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) + (\text{TN} + \text{FN}) * (\text{TN} + \text{FP})$. TP denotes true positives, TN denotes true negatives, FP denotes false positives, FN denotes false negatives, and T denotes the total number of input images in the dataset.

In addition to the above performance parameters, we have incorporated the confusion matrix (CM) to represent the experimental result in a better way. The CM provides a more comprehensive view of the outcomes of a predictive model, and it also shows the classes that are being classified correctly or wrongly. The accuracy of a classification model may be measured by comparing the number of test records that were classified correctly versus the number that were guessed wrong.

6 Results and Discussion

We have modified VGG19 to further increase the performance of the model. We have kept the initial layers of VGG19 up to fourth block and removed the upper layers. We have added an inception block at the upper layer. Our proposed model is designed in such way that tackles class imbalance problem and decrease the complexity of VGG19.

The experimental results of various evaluation metrices, such as $F1$-score, precision, sensitivity, and accuracy of the proposed model for oral cancer classification, are given in Table 4. It can be seen from Table 4 that the proposed model achieved an average accuracy of 98.64% and the obtained average $F1$-score, precision, and sensitivity values of 98.7%, 98.59%, and 98.81%, respectively. The results of confusion matrices are also shown in Fig. 6.

In addition, Fig. 7 shows how the validation accuracy increased with respect to the increasing of the number of batches processed for the proposed model. The train and validation loss with

Table 4 Result of various evaluation metrics of the proposed model for the classification of OSCC (%).

Folds	Accuracy	$F1$-score	Precision	Sensitivity	Kappa score
Firstfold	98.74	98.79	98.88	98.70	0.975
Secondfold	98.43	98.52	98.18	98.88	0.969
Thirdfold	98.55	98.61	98.70	98.52	0.971
Fourthfold	98.55	98.60	98.33	98.88	0.971
Fifthfold	98.94	98.97	98.88	99.07	0.980
Average	**98.64**	**98.70**	**98.59**	**98.81**	**0.973**

Note: Bold values indicate the average values.

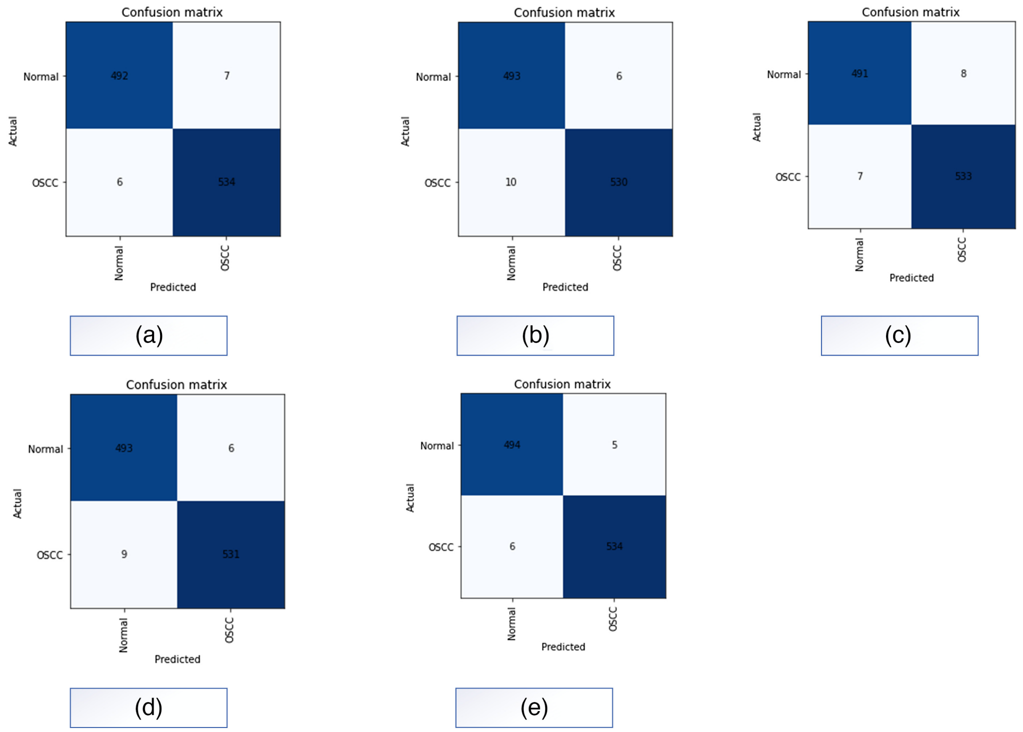

Fig. 6 Confusion matrix for classification of OSCC using the proposed model: (a) firstfold CM, (b) secondfold CM, (c) thirdfold CM, (d) fourthfold CM, and (e) fifthfold CM.

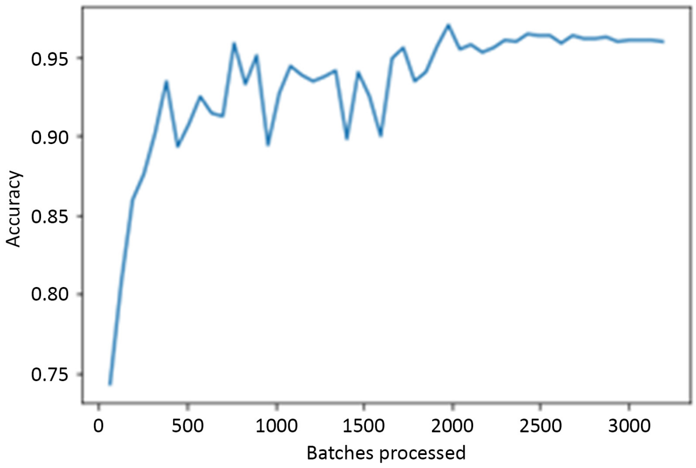

Fig. 7 Validation accuracy with respect to the number of batches processed.

respect to the number of batches processed is given in Fig. 8. It can be noted from the graph that both the validation and training losses were decreasing with the number of batches increasing. Both the losses were higher at the start but decreased exponentially. This shows how the proposed model was learning very effectively with higher number of training batches.

We have also compared the experimental results of the proposed model with the results of pre-trained models. The overall comparative experiment results of pre-trained models and our proposed work are presented in Table 5. The results given in this table are the average of five-folds. It can be seen from the experimental result that VGG19 has achieved the highest accuracy of 96.26% from the rest nine pre-trained models for detection of oral cancer using histopathology images. This is the main reason why we have selected VGG19 and modified it to get better results. It can be seen from the results that the proposed model achieved a better performance

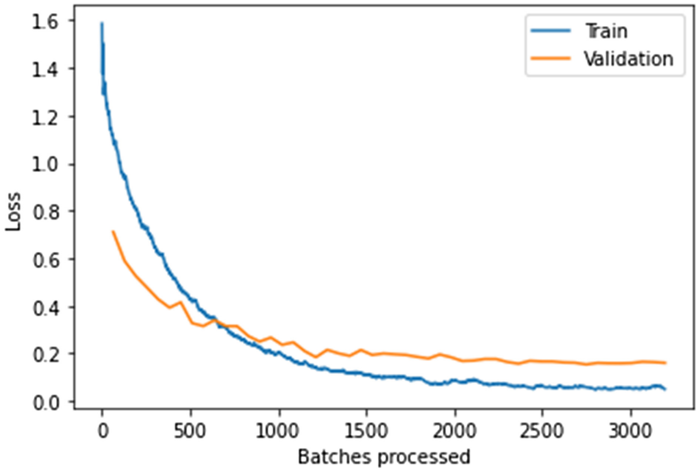

Fig. 8 Validation and training losses with respect to the number of batches processed.

Table 5 Experimental results of various metrics for different pre-trained models and proposed model with the best value in bold (%).

Pre-trained network	Accuracy	$F1$-score	Precision	Recall	Kappa score
DenseNet121	95.41	95.5	95.4	95.5	0.917
DenseNet169	96.01	96.2	96.1	96.04	0.936
DenseNet201	96.17	96.3	96.2	96.19	0.935
MobileNetv2	95.47	95.6	95.4	95.5	0.922
ResNet50	95.894	95.9	95.84	96	0.928
ResNet101	95.86	95.9	95.9	95.7	0.927
Xception	95.65	95.8	95.7	95.8	0.925
AlexNet	95.72	95.9	95.9	95.7	0.926
VGG16	96.12	96.2	96.2	96.19	0.934
VGG19	96.26	96.3	96.3	96.3	0.939
Proposed method	**98.64**	**98.70**	**98.59**	**98.81**	**0.973**

Note: Bold values indicate the values obtained from the proposed model outperform the others.

from the rest pre-trained models. We have proposed a model that is less complex from other models and yields superior performance.

In addition, the performance parameter results of different pre-trained models and proposed work are presented graphically in Fig. 9. It can be seen from the chart that almost all the pre-trained models achieved an accuracy of minimum 95.4% and maximum of 96.4%. This consistence result shows that with minor modification of the deep learning models, we can even achieve better results to detect oral cancer.

Figure 10 demonstrates the confusion matrices of VGG19. It displays consistency between expected and actual values, signifying superior performance from the rest pre-trained models.

Figure 11 depicts the box plots of the kappa scores of the 10 pre-trained networks and the proposed work. Considering the frequency of each class, the kappa score of the model reveals how much better it performs than a classifier that makes predictions at random. A higher kappa score reflects greater confidence in the classification model. A side-by-side comparison reveals that the proposed model provides a larger range of kappa scores for the supplied dataset. The box plot shows the kappa score sliding toward 1 as the training set size rises, which indicates greater confidence in the model. The proposed model has a median kappa score value of 0.973. This

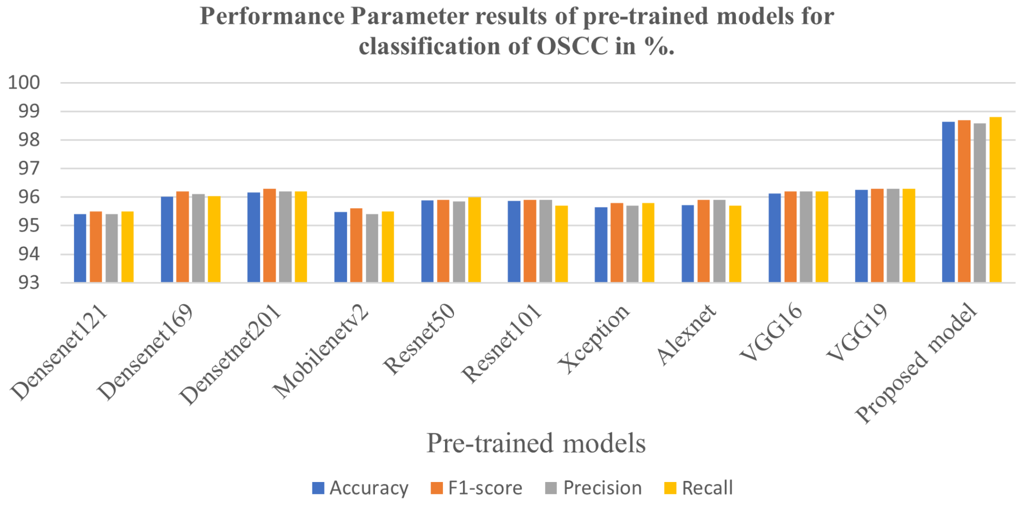

Fig. 9 Result of performance parameters using chart.

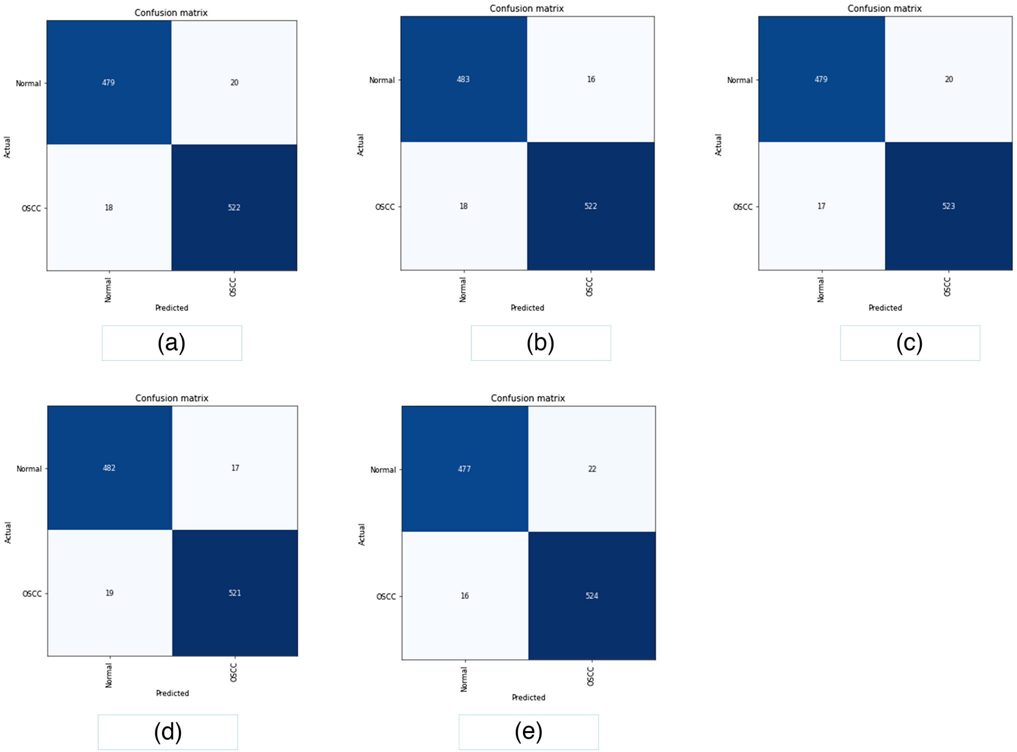

Fig. 10 Confusion matrix for classification of OSCC using VGG19 model: (a) firstfold CM, (b) secondfold CM, (c) thirdfold CM, (d) fourthfold CM, and (e) fifthfold CM.

number emphasizes the validity of the proposed model and demonstrates how the model's performance improves with increasing training set size.

Comparative analysis of our proposed model with other existing related works is shown in Table 6. There were very limited number of research works available in the area of employing deep learning techniques for the detection of oral cancer. Most of these works have used histopathology images but they lack data pre-processing stage, and their model is not robust. In this table, we have presented the comparison of our proposed model with other existed models considering the accuracy performance parameter. It is clearly evident that our proposed model has achieved a better accuracy for the detection of oral cancer from histopathological images.

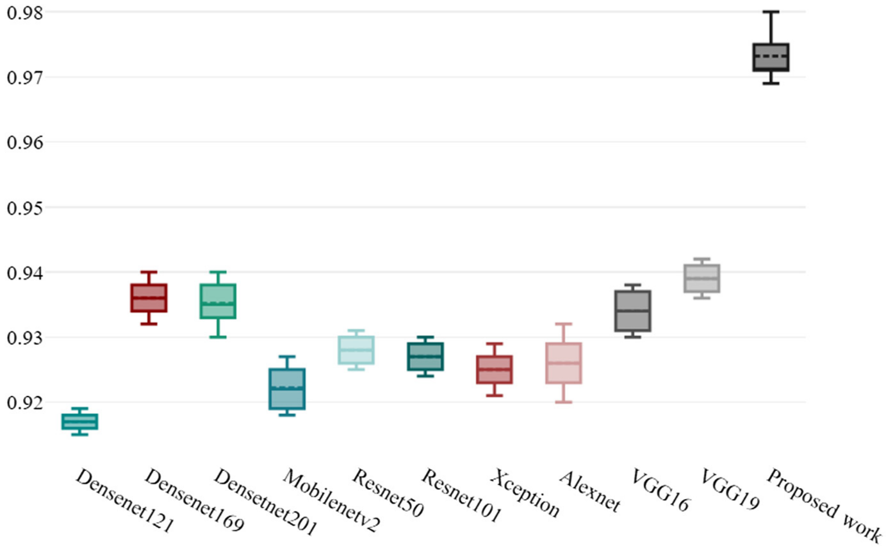

Fig. 11 Box plots of kappa score for the pre-trained networks and proposed work.

Table 6 Comparative analysis of the proposed model with previous research works.

Publication	Method	Dataset	Accuracy (%)
Aberville[12]	InceptionV3	Confocal laser endomicroscopy images	80.01
Alhazmi[8]	ANN	Pathologic reports, taken 29 variables	78.95
Welikala[10]	ResNet101	Real-world oral cancer images	78.30
Shavlokhova[11]	MobileNet	*Ex vivo* fluorescence confocal microscopy images	77.89
Rahman et al.[14]	AlexNet	Histopathology images	90.06
Palaskar et al.[23]	MobileNet and ResNet	Histopathology images	83.66
Bansal et al.[9]	Transfer learning	Histopathology images	92.41
Panigrahi et al.[34]	Transfer learning	Histopathology images	96.6
Proposed model	Transfer learning (VGG19)	Histopathology images	96.264
	Modified VGG19 model		**98.64**

A more comprehensive study requires a greater quantity of patient dataset. In order to be effective, deep-learning models need to be trained on more than a million images, which is difficult to do in the medical sector. In addition, training deep neural networks on a restricted dataset may result in overfitting problem and inhibit its generalization. Therefore, in the future work, we are planning to incorporate real-world images of oral cancer to enhance the robustness of the model.

The proposed model in this study has significant potential for real-life applications across various domains. By harnessing deep-learning techniques and transfer learning, the model demonstrates its efficacy in detecting oral cancer from histopathology images. Accurately identifying cancerous tissues holds immense implications for clinical practice and patient care. The followings are some potential real-life use cases and advantages of the proposed model.

- *Early detection and diagnosis*. Timely identification of oral cancer at an early stage is critical for prompt intervention and improved patient outcomes. The proposed model can aid medical professionals in identifying suspicious tissue samples, facilitating early diagnosis, and timely treatment.

- *Pathology assistance*. Histopathology analysis plays a pivotal role in cancer diagnosis. The proposed model can serve as a valuable tool for pathologists, offering an automated and reliable method for analyzing histopathology images. This can help alleviate the workload of pathologists and potentially enhance the efficiency and accuracy of cancer detection.
- *Treatment planning and monitoring*. Accurate diagnosis of oral cancer is essential for devising appropriate treatment plans. The proposed model can assist clinicians in assessing the extent of cancerous tissues, guiding them in developing personalized treatment strategies. Additionally, the model can be employed for monitoring treatment response over time, enabling necessary adjustments.

It is important to emphasize that while the proposed model demonstrates promise in real-life applications, further validation and integration into clinical practice are crucial.

7 Conclusion and Future Work

A transfer learning-based framework is proposed by employing pre-trained networks to automatically classify oral cancer from histopathological images. We conducted numerous experiments on a histopathology dataset of oral cancer samples. These input images were fed into 10 well-known pre-trained CNN models, namely, DenseNet121, DenseNet169, DenseNet201, MobileNetv2, ResNet50, ResNet101, Xception, AlexNet, VGG16, and VGG19. Data augmentation was also utilized to enhance the accuracy, and the performance of the above networks was compared. Evaluation metrics, such as the accuracy, $F1$-score, precision, sensitivity, and kappa score were used. The experimental results show that the VGG19 model with data augmentation was able to attain the highest classification accuracy of 96.26%. We have also modified the VGG19 model to improve its performance by adding naïve inception block module at the higher level. The proposed model has achieved an average accuracy of 98.64%. Our model is not only simple but also robust, which solved data imbalance problem and achieved better performance than other similar models. In the future, we are planning to add real-world photographic images of oral cancer to further improve the robustness of our model and show that our model can also handle transfer learning technique. Additionally, several transfer learning techniques could be employed with ensemble classifiers to further enhance the efficiency of the proposed model.

Research Data Policy and Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

No funding was received for conducting this study. There are no conflicts of interest declared by any of the authors.

References

1. N. Sinevici and J. O'Sullivan, "Oral cancer: deregulated molecular events and their use as biomarkers," *Oral Oncol.* **61**, 12–18 (2016).
2. E. E. Vokes et al., "Head and neck cancer," *N. Engl. J. Med.* **328**(3), 184–194 (1993).
3. K. Dhanuthai et al., "Oral cancer: a multicenter study," *Med. Oral, Patol. Oral Cir. Bucal* **23**(1), e23 (2018).
4. L. Lavanya and J. Chandra, "Oral cancer analysis using machine learning techniques," *Int. J. Eng. Res. Technol* **12**, 596–601 (2019).
5. V. Kearney et al., "The application of artificial intelligence in the IMRT planning process for head and neck cancer," *Oral Oncol.* **87**, 111–116 (2018).
6. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).

7. K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW)*, IEEE, pp. 1–6 (2015).
8. A. Alhazmi et al., "Application of artificial intelligence and machine learning for prediction of oral cancer risk," *J. Oral Pathol. Med.* **50**(5), 444–450 (2021).
9. K. Bansal, R. Bathla, and Y. Kumar, "Deep transfer learning techniques with hybrid optimization in early prediction and diagnosis of different types of oral cancer," *Soft Comput.* **26**(21), 11153–11184 (2022).
10. R. A. Welikala et al., "Automated detection and classification of oral lesions using deep learning for early detection of oral cancer," *IEEE Access* **8**, 132677–132693 (2020).
11. V. Shavlokhova et al., "Deep learning on oral squamous cell carcinoma ex vivo fluorescent confocal microscopy data: a feasibility study," *J. Clin. Med.* **10**(22), 5326 (2021).
12. M. Aubreville et al., "Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning," *Sci. Rep.* **7**(1), 11979 (2017).
13. G. Tanriver, M. Soluk Tekkesin, and O. Ergen, "Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders," *Cancers* **13**(11), 2766 (2021).
14. A.-U. Rahman et al., "Histopathologic oral cancer prediction using oral squamous cell carcinoma biopsy empowered with transfer learning," *Sensors* **22**(10), 3833 (2022).
15. T. Rahman, "A histopathological image repository of normal epithelium of oral cavity and oral squamous cell carcinoma," *Mendeley Data* **1** (2019).
16. A. F. Kebede, "Histopathologic oral cancer detection using CNNs," 2022, https://www.kaggle.com/datasets/ashenafifasilkebede/dataset (accessed September 30, 2010).
17. H. Zhang et al., "DBNCF: personalized courses recommendation system based on DBN in MOOC environment," in *Int. Symp. Educ. Technol. (ISET)*, 2017-January, IEEE, pp. 106–108 (2017).
18. M. Mohammadpour et al., "Facial emotion recognition using deep convolutional networks," in *IEEE 4th Int. Conf. Knowl.-Based Eng. and Innov. (KBEI)*, 2017-January, IEEE, pp. 17–21 (2017).
19. T. Young et al., "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.* **13**(3), 55–75 (2018).
20. O. Abdel-Hamid et al., "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(10), 1533–1545 (2014).
21. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
22. J. Deng et al., "ImageNet: a large-scale hierarchical image database," *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 248–255 (2009).
23. R. Palaskar et al., "Transfer learning for oral cancer detection using microscopic images," in *Comput. Vision Pattern Recognit.*, pp. 1–8 2020).
24. H. Xie et al., "A lightweight 2D CNN model with dual attention mechanism for heartbeat classification," *Appl. Intell.* **46**(1), 1–16 (2022).
25. K. M. Knausgård et al., "Temperate fish detection and classification: a deep learning based approach," *Appl. Intell.* **52**(6), 6988–7001 (2022).
26. F. Younas, M. Usman, and W. Q. Yan, "A deep ensemble learning method for colorectal polyp classification with optimized network parameters," *Appl. Intell.* **53**, 2410 –2433 (2023).
27. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4700–4708 (2017).
28. A. G. Howard et al., "MobileNets: efficient convolutional neural networks for mobile vision applications," https://arxiv.org/abs/1704.04861 (2017).
29. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
30. F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1251–1258 (2017).
31. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM* **60**(6), 84–90 (2017).
32. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," https://arxiv.org/abs/1409.1556 (2014).
33. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2818–2826 (2016).
34. S. Panigrahi et al., "Classifying histopathological images of oral squamous cell carcinoma using deep transfer learning," *Heliyon* **9**(3), e13444 (2023).

Dawit Kiros Redie received his master's degree (MTech) in software engineering from KIIT University, Bhubaneswar, India. He has completed his bachelor's degree (BTech) in electronics and computer science engineering. He has received several internships including a software engineering internship from Johnson Controls. He is an enthusiastic learner and programmer. His area of interests includes big data, Apache Spark, deep learning, software engineering, and machine learning.

Saurabh Bilgaiyan is currently working as an assistant professor at KIIT University, Bhubaneswar, India, since 2016. He completed his master's degree and PhD in computer science engineering from KIIT University in 2014 and 2018, respectively, and his BE degree in information technology from Bansal Institute of Research and Technology, Bhopal, India, in 2012. He has published more than 44 research papers in various reputed international journals, conferences, and edited books. His areas of interest include soft computing, software engineering, cloud computing, image processing, and machine learning.

Santwana Sagnika is an assistant professor at the School of Computer Engineering of KIIT University. She completed her PhD in computer science and engineering in the field of natural language processing. Her areas of interest include natural language processing, deep learning, machine learning, image processing, and optimization.