The Prediction of Oral Cancer through Deep Learning-based Classification Using Mouth Landmarks

Under the Guidance of: Prof. Sagari S M
Adarsh B, Vinayak R Nayak, Tarun Patil, Tejaswi Rajendra Hegde

Department of Computer Science Engineering DSCE

Abstract

In this research, we propose a novel solution for early detection of oral cancer in low- and middle-income nations. Early detection of oral cancer can save lives, and automated techniques can help. On the basis of photographic images, we aim to develop a deep learning classification system for oral lesions. To achieve this, we propose a model for identifying landmarks in oral images, which is then used to guide the categorization of oral lesions. The effectiveness of this approach is evaluated using a variety of deep convolutional neural networks. The proposed system has the potential to improve the accuracy of oral lesion classification and aid in the early detection of oral cancer.

Introduction

In low- and middle-income countries, oral cancer is a prevalent and dangerous form of cancer. In 2018, there were an estimated 354,864 new cases of oral cancer and 177,384 deaths. There are three major risk factors for oral cancer: smoking, drinking alcohol, and chewing betel quid. Many people are not aware of the symptoms of oral cancer, making it essential to detect it early for better survival. Oral cancer can often be preceded by oral potentially malignant disorders (OPMDs) which are visible and can be detected without special instruments.

AI has been implemented in various fields to increase efficiency and cut costs. Recent advancements in deep learning have led to improved performance of AI models in various areas, and they even surpass human performance in some cognitive tasks. The use of deep learning in the medical field has grown in popularity and has made significant progress in tasks such as diagnosing skin lesions, detecting pneumonia from chest X-rays, and improving visualization of pathologies. While the results from deep learning in the medical field have been impressive, it is not intended to replace human professionals, but rather to assist them and improve efficiency.

Researchers have recently utilized deep learning techniques to detect oral cancer in its early stages. These techniques involve training deep learning algorithms to identify intricate details of oral lesions and recognize distinctive visual patterns of oral cancer. Previous studies in this area have mainly focused on utilizing various types of images, such as multidimensional hyperspectral images, computed tomography (CT) images, microscopic images, autofluorescence images, and photographic images. In our research, we propose a new deep learning framework for classifying oral lesions from photographic images into four different referral decision classes. Our framework consists of two modules: a mouth landmark detection module and an oral lesion classification module. We have also developed a new mouth landmark model that identifies the location of the mouth, which is utilized as a feature to guide the classification model.

Deep Learning

Deep learning is a branch of machine learning which uses deep neural networks, which are composed of many layers of artificial neurons, to learn from large amounts of data. These networks are designed to learn and improve their performance with more data, unlike traditional M.L algorithms which have a finite capacity to learn. This makes deep learning particularly useful for tasks where the data is complex and varied such as image and speech recognition, natural language processing, and autonomous systems.

Deep learning networks are trained by a process called back-propagation, this involves adjusting weights of the neurons in the network, based on the errors of the previous training iteration. The network is repeatedly exposed to vast amounts of data until it can proficiently classify or predict the intended output.

Deep learning has been used in many different applications such as image and speech recognition, natural language processing, self-driving cars, and medical diagnosis. The ability of deep learning networks to learn and improve with more data makes them suitable for tasks where traditional machine learning algorithms struggle, such as in cases where the data is unstructured and complex.

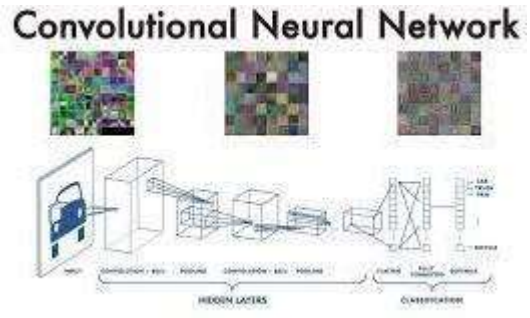
In summary, deep learning is a subfield of machine learning that uses deep neural networks to learn from large amounts of data, making them suitable for complex tasks like image, speech and natural language processing, autonomous systems, and medical diagnosis. The networks are designed to improve their performance with more data, unlike traditional machine learning algorithms.

CNN (Convolutional Neural Networks)

Convolutional Neural Networks (CNN) are a specific type of deep learning architecture that is designed to process and recognize visual patterns in images. They use a mathematical operation called convolution, which is a linear operation that multiplies two functions to create a third function that expresses how the shape of one function can be changed by the other. In the case of CNNs, this operation is applied to two images represented as matrices, to extract meaningful information from the image.

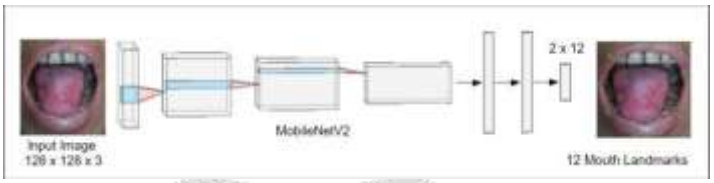
CNNs are similar to other types of neural networks, but they have an added layer of complexity due to the use of convolutional layers, which are essential for the network's function. These layers help the network identify and extract important features from the image, such as edges, textures, and shapes, which are then passed on to the next layer for further processing. This hierarchical process allows CNNs to learn and recognize complex visual patterns in images with high accuracy.

Additionally, CNNs are trained with huge amounts of labeled data and can be used for various image-related tasks such as object detection, 1image classification, semantic segmentation and more. Also, CNNs are widely used in computer vision, image and video analysis, and natural language processing tasks.



Mouth Landmark

The detection of oral cancer can be made more efficient with the assistance of mouth landmark detection. As of now there is not much research on mouth landmarks, but facial landmark detection method is used to detect the region of mouth. According to this research the facial landmarks detect overall parts on the face i,e eyes, nose, eyebrow, mouth. In this research the detection of the mouth uses 12 dotted pattern to identify the region of mouth.



1. Detection of Oral Cancer using Deep Learning Techniques.

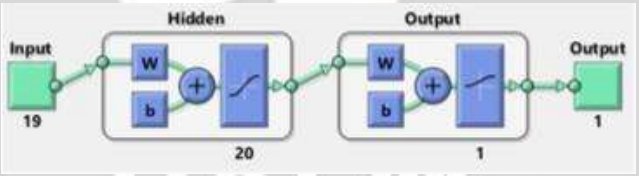
An automated system has been created to diagnose oral malignancies using machine learning and deep neural network models. This system includes various classification models such as Naive Bayes, KNN, SVM, ANN, and CNN. A new CNN network with 43 deep layers has been designed for this system, with its structure inspired by the VGG16 Network.

For the purpose of classification, a dataset of 630 oral images was used to train the system, which included images obtained from the internet as well as those collected from hospitals with the aid of oral specialists. Of the 1200 lesion images used, 600 were normal and 600 were malignant.

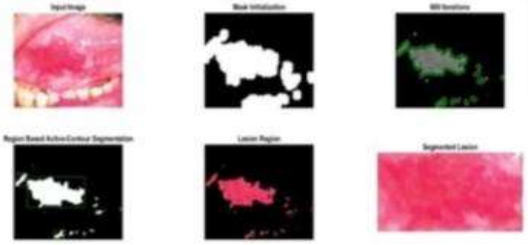
Methods

The input is an RGB oral image to the disease diagnosis system.

Segmentation:



We first convert the RGB image into the YCbCr color space
and then use the blue difference Cb and Cr intensity values to create a mask for the oral lesion. Lesions in cancerous oral images are typically characterized by either red or white patches. By calculating the mean Cr value of the input images, we can determine whether the image is more likely to have white or red patches. If the mean Cr value falls below a predetermined threshold, the image is classified as having white patches, whereas if the mean Cr value is above this threshold, the image is classified as having red patches.

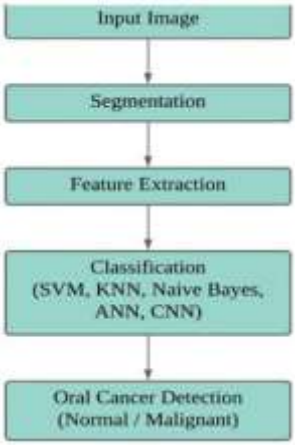


Feature Extraction

44 features feature vector is created by extracting

Feature Selection

Relevant features are selected by using the statistical feature selection methods. Based on the ranks by feature selection methods the top 19 features are selected.



Classification

To train the model a dataset of nineteen selected features are used.

Classification Techniques employed are:

Support Vector Machine (SVM)

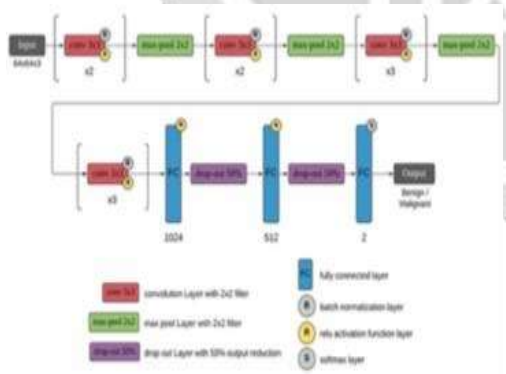
The Support Vector Machine model is used for classification and employs a method ,which is the kernel trick to alter data and identify the optimal boundary among the possible outputs.

Artificial Neural Network (ANN*)*

Ninteen features from dataset are extracted from lesion images used to train network. The network has twenty hidden layers a, single output layer

Convolutional Neural Network (CNN)

A novel convolutional neural network with 43 layers has been developed for detecting oral diseases. This network comprises ten convolutional layers, with batch normalization layers used to normalize the output of each convolutional layer, and ReLU layers used as the activation function. Max-pool layers are utilized for pooling. The network also includes three fully connected layers, and the softmax layer is used to generate output predictions. The network takes 64x64 RGB images as input and classifies them as either benign or malignant.



K-Nearest Neighbor (KNN)

K-nearest neighbor is used for classification of image based on selected features on lesion region.
For the current work, a value of ten is chosen for K.

Naïve Bayes

A NB classifier is a probabilistic classifier that uses Bayes theorem
For the present classification, a Bayesian classifier with different kernel densities is used.

Table 4: Training accuracies of ML models

Models	Accuracy
Naive Bayes	[illegible]
KNN	[illegible]
SVM	[illegible]
ANN	[illegible]
CNN	[illegible]

Conclusion:

Through experimentation, it was found that various classifiers can effectively identify lesions of oral cancer. In particular, the deep learning Convolution neural network model demonstrated high accuracy for distinguishing between normal, cancerous images.

Table 5: Testing performance measures of ML models

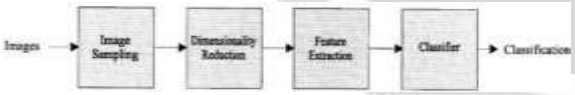
Evaluation measure	Naive Bayes	KNN	SVM	ANN	CNN
Precision	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Recall	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
F1-score	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Specificity	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Accuracy	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

2.Convolutional Neural Network Conceptualization: A Deep Learning Approach

Convolutional Neural Networks (CNN) are a popular deep learning technique that can solve complex problems by extracting important features automatically from raw input data. The model is composed of multiple processing layers that learn different features at varying levels of abstraction, enabling the network to recognize features at different levels. Unlike traditional machine learning models, deep learning models eliminate the need for pre-selected features. CNN has various applications, including image and speech recognition, natural language understanding, signal processing, face recognition, and small molecule bioactivity prediction. Other types of deep learning architectures include Recurrent Neural Networks and Deep Belief Networks. CNN is commonly used for tasks like image classification, face detection, speech recognition, and facial expression recognition. One of the main benefits of CNN is weight sharing, which reduces the number of trainable parameters and leads to smoother training.

General Model of Convolution Neural Network

The neural network model in question comprises a sole input layer, numerous hidden layers, and a single output layer. Each neuron in the network receives an input vector X and generates an output vector Y by subjecting it to a function F. The function F is defined as F(X, W)=Y, where W represents the weight vector that determines the inter-neuron connections between adjacent layers.

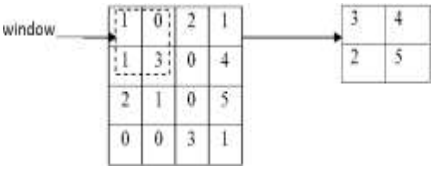


Convolution Layer

When an image is fed into a neural network to be classified, the input layer receives it and the predicted class label is generated based on the features extracted from the image. Each neuron in the following layer has a local connection with a limited number of neurons in the preceding layer, called the receptive field. The receptive field enables the neuron to extract local features from the input image. To create a feature map, a weight vector is formed by the receptive field of a neuron, which slides over the input vector to create a feature map by using a filter or kernel. By sliding the filter horizontally and vertically, the convolution operation is performed, which extracts N features from the input image and creates N filters and N feature maps, representing them as distinct features in a single layer. By using local receptive fields, the number of trainable parameters is reduced in the network. The output in the subsequent layer for the location (i,j) is calculated by the formula: output[i,j] = $\sigma(\sum(X[k,l] * W[k,l]) + b)$, where X is the input, W is the filter or kernel that slides over the input, b is the bias, * represents the convolution operation, and σ is the nonlinearity introduced in the network.

Pooling Layer

After the convolution layer in a CNN, a pooling operation is applied to further reduce the number of trainable parameters. A window is chosen to select input elements, which are then passed through a pooling function, producing an output vector. There are various types of pooling techniques, including average pooling and max-pooling. Max-pooling is the most commonly used technique, as it significantly reduces the size of the feature maps.



Fully Connected Layer

In the final layers of the network, fully connected layers are used. The output from the previous layer, typically a pooling layer, is flattened and passed through a fully connected layer. Each neuron in the fully connected layer is connected to every neuron in the previous layer. The dot product of the weight vector and input vector is calculated to produce the final output. Activation functions such as ReLU or softmax are applied to the output of the fully connected layer to determine the probability of the input belonging to a specific class.

Activation Function

The activation function of a neural network is responsible for determining whether a neuron should be activated or not, based on the weighted sum of its inputs and an added bias. By introducing nonlinearity to the output of neurons, activation functions allow neural networks to model complex relationships between input and output. One commonly used activation function is the rectified linear unit (ReLU), which is defined as f(x) = max(0,x), where x is the input to the neuron. The ReLU activation function has been shown to be effective in reducing the likelihood of vanishing gradients during backpropagation, leading to faster training and improved performance of neural networks.

.

3. Detection of oral cancer using smartphone-based images using deep learning techniques for early diagnosis

Oral cancer is a prevalent health problem worldwide. Early detection of cancerous tumors and potentially malignant conditions in the mouth can significantly increase survival rates. Smartphone-based image detection methods aim to improve the effectiveness of their methodologies. To achieve this, researchers have implemented a rule-based image capturing method for collecting oral cavity images and created a medium-sized dataset with five categories using this method.

They have also introduced a deep learning network to evaluate the accuracy of detection. The dataset was collected from hospital outpatients and inpatients using four different smartphones for training and testing. The outcome of this method shows an accuracy of 83%, specificity of 96%, precision of 84%, and F1 of 83% on images of over 400. The architectural tools which are used in this research are HRNet, DenseNet169, ResNet50, and VGG16.

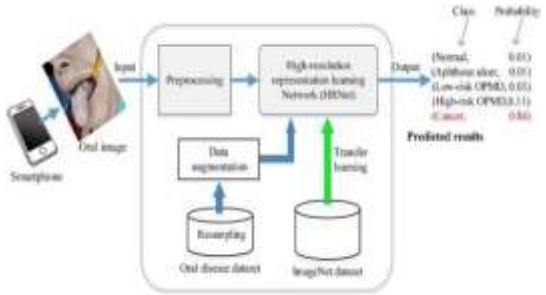


Table 4 Comparison of classification performance with different methods on the oral disease dataset.

Method	Params	SE_{macro}	SP_{macro}	Pr_{macro}	*F*1
VGG16	568M	0.728	0.950	0.767	0.745
ResNet50	24M	0.770	0.954	0.788	0.771
DenseNet169	**12M**	0.810	0.965	0.835	0.817
HRNet-W18	17M	**0.830**	**0.966**	**0.843**	**0.836**

Note: The values shown in bold represent the highest performance.

Fig. 9 Examples of oral images with visualization. The heatmaps highlight the class-discriminative regions used for oral classification.

4. A Detailed Look At CNN-based Approaches In Facial Landmark Detection

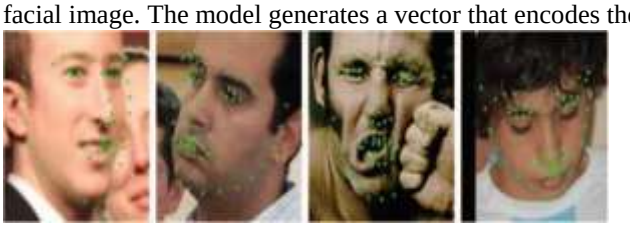
In this study, the authors uses convolutional neural networks (CNNs) for the purpose of identifying landmarks in images. The CNNs were classified into two categories: regression and heatmap techniques. The authors also used a pixel-wise classification (PWC) model. The research focuses on the use of facial landmarks for various applications, including face recognition, head pose estimation, facial expression recognition, facial component manipulation, and facial expression estimation. To train and test their model, the authors used six datasets containing images of facial landmarks, including the AFW dataset, Helen dataset, LFPW dataset, three hundred Faces in 300-W dataset, a subset of hundred and thirty five images with hard poses and expressions from the 300-W dataset , and COFW dataset

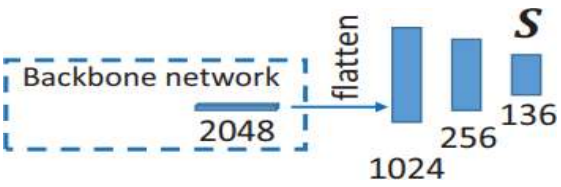
Approaches in Regression

Facial landmark detection involves using regression models, which can be broadly classified into two types: direct regression models and cascaded regression models. A direct regression model predicts the coordinates of landmarks directly from a facial image, using a vector to represent the landmarks. The size of the vector is twice the number of landmarks being detected. To illustrate, consider the example of a direct regression model for identifying sixty-eight facial landmarks, depicted in the provided image. The model takes a facial image as input and generates the coordinates of the landmarks as output.

Direct regression model: -

Facial landmark detection is performed using a model that predicts the coordinates of landmarks, represented as a vector, from a facial image. The vector size is twice the number of landmarks. As an illustration, the image showcases an instance of a direct regression model that identifies sixty-eight facial landmarks from the input facial image. The model generates a vector that encodes the landmark coordinates for the given facial image.





Cascaded regression models.

This model differs from direct regression model in that they continuously update a pre-detected landmark to detect the final landmarks. This is done by updating the landmark location in each stage using a vector generated by a sub-network called Si. The model produces the final landmark coordinates after n updates. The cascaded regression model can be represented by the formula Si = Si1 + Si1, where I = 1, 2, n. The model is assessed by using the L2 distance, which measures the point-by-point difference between the detected landmarks and the ground-truth landmarks. This difference is formulated as a regression loss: lossr

$$\text{ед} = 1\ |L| \sum l \in L\ kSl - \hat{}Sl\ k2$$

Heatmap Approaches

Heatmap approach identifies landmarks by generating a two-dimensional heatmap that indicates the position of the landmarks. Facial landmark detection models draw inspiration from fully convolutional networks (FCNs) which include a convolutional component that extracts semantic features from facial images. These semantic features are decoded into a series of heatmaps. Heatmap approaches for facial landmark detection can be classified into distribution models, heatmap regression models, and pixel-wise classification models based on the characteristics of the heatmap and the loss function employed. The choice of the heatmap approach is dependent on the application requirements and the specific characteristics of the facial images being processed.

5. A New Approach to Mouth Mouth Detection Using Landmark

In this study, a new method for detecting mouths in color images is proposed. The method involves segmenting images based on skin color, rather than lip color, and using specialized techniques to quickly identify mouth candidates and classify them using a neural network. This approach resulted to a correctness of 85.5%. Mouth detection has various applications, such as videoconferencing, language recognition in conversations, identifying specific emotions, and automatic speech recognition. There are two main approaches to mouth detection: direct and indirect. In the direct method, the mouth is detected by applying color-based segmentation in a particular color space (e.g. RGB, YCbCr, HSV) and assuming that the face and lip colors are known. In the indirect method, the face is first detected using a geometrical model, and then the mouth is located by projecting onto specific coordinates and extracting mouth features.

General Algorithm

There are 3 steps in this algorithm:

Segment image based on skin color (build binary image).
Find the biggest object and local complement image.
Validation by neural network.

Skin Segmentation

This method involves identifying the mouth by segmenting the image based on skin color in the YCbCr color space. This color space is useful for handling a range of lighting conditions and is commonly used in video compression standards. The method does not rely on lip color for detection**.**

Determining Mouth Candidates

After we segmented the image next is to detect the biggest component of the image that is face.

This process contains two stages:

1. Find all connected components in the segmented image.
1. Find the object that has maximum area in these connected components.

Locating the Mouth

By using features, we can distinguish between the mouth and other objects.

Features of mouth: -

Elongation feature: -Since different objects have different geometrical features, proportion between length and width between each object is different.

elongation= width/length;

Location feature: -We can determine the location of the connected component in the face based on this feature. With the fact that mouth is always located below one-half of the face.

Location= a/b;

Validation by Neural Network

A neural network is used because it is not possible to accurately classify the mouth based solely on mouth features. The input to the neural network is a three-dimensional feature vector, which includes the rate, location, and length of the mouth. The output is a binary value (either 0 or 1). Based on these features our neural network is trained to classify the input.

[6]Squeeze-Net

There has been a lot of focus on creating more complex models using large computer clusters to improve accuracy on the ImageNet challenge, but less attention has been paid to the potential of mobile phones and edge devices. The goal of this research is to make sophisticated high-end approaches more feasible for use on real-world devices, even if it means giving up on some of the calibre of the network's output. One example of this is Squeeze-Net, a compact neural network developed by Cornell University in 2016 that can run on device with low computational power, such as phones, while still achieving Alex-Net-level accuracy with fifty times fewer parameters , model size of 0.5MB.

Reducing the size of the network was the primary goal of this paper.

Fire modules

In each fire module of Squeeze-Net, the input is first compressed and then expanded twice using different techniques. The final output is created by coupling the output of the 2 expansion layers. This process reduces the amount of data that is passed through the network, which in-turn decreases number of parameters and the overall model size. However, it also means that some information is lost in the compression step. Squeeze-Net also uses fewer maxpool operations, which can further minimise the amount of data flowing through network. The authors of Squeeze-Net also applied techniques from another paper to further reduces the size of this model.

Model Pruning

Network pruning is a technique that can make model execution faster. In most of the cases, a minute portion of network's activations (e.g., 20%) contribute the majority of the results (e.g., 80%), which follows a pattern similar to Zipf's law. This implies that it is possible to create a smaller network with fewer activations by sacrificing some accuracy, a process known as sparsification or pruning.

Difference between Squeeze-Net 1.0 and 1.1

Squeeze-Net is a computer vision network from 2016 that is known for being able to achieve good results with few cycles and parameters than other networks. There are two versions of Squeeze-Net: v1.0 and v1.1. The first layer is the major difference between the 2 layers with v1.0 using a 7x7 stride and 96 filters and v1.1 using 3x3 strides and 64 filters. Mobile-Net is another family of networks that was specifically designed to be run on mobile devices and is more accurate than Squeeze-Net. It uses depth-wise separable convolutions to achieve this performance. These architectures are important to understand because they are optimized for use on mobile devices.

[7] MobileNet v1

Model is specially designed to run on mobile hardware, much better use of parameters and data space. Mobile-Net uses a combination of depth-wise convolutions and pointwise convolutions to process an input image and produce the desired output shape. This approach, called a depth-wise separable convolution, is less disruptive than Squeeze-Net and allows for most of the benefits of Squeeze-Net's compression method to be retained. Additionally, because depth-wise separable convolutions can be efficiently run-on mobile hardware, this approach is more cost-effective.

Output

By the author’s implementation the results are on par with Res-net 50 network, but this network is small in general and can be calculated faster at runtime and so is a good improvement for mobile devices.

[8] Mobile-Net v2

Mobile-Net v2 is an improved version of Mobile-Net v1 that was released in 2018. It is slightly more accurate and more efficient computationally. Mobile-Net v2 uses inverted residual blocks and linear bottleneck layers, which are modifications of the architecture used in Mobile-Net v1. In a residual block of inverted, the depth convolution is applied first, followed by a 1x1 convolution to compress the network back down at the end.
In our implementation of inverted skip connections, we made a small modification by adding a ReLU activation function to the merged output of the bottleneck layer and the input in the original ResNet network. When using the same training process and basic setup, this network was observed to perform better than the MobileNet v1 architecture.

[9] Detection of Oral Cancer using Deep Learning

Oral cancer is a widespread and complex disease that often has a poor prognosis. In India, it has been linked to around 130,000 deaths. While there are various diagnostic methods for oral cancer, they often struggle to accurately identify and classify cancer cells. Using advanced technologies and deep learning algorithms, we can improve early detection and classification. This work explores the use of wavelet features , the Zernike Moment as techniques for extracting three characteristics that may be useful in diagnosis.

Deep Learning

Deep-learning algorithms are designed to analyze data and draw conclusions in a way that is similar to how humans think. Neural networks, a multi-layered structure of algorithms that mimics the human brain's arrangement, is used by them to accomplish this. Similar to how our brains use pattern recognition and classification to differentiate between various types of information, neural networks can be trained to perform similar tasks on data. Each layer of a neural network can be viewed as a filter that starts with a broad perspective and progressively narrows down to specific features, increasing the probability of detecting and producing an accurate output. In this way, neural networks are capable of performing complex tasks such as image and speech recognition, natural language processing, and predictive analytics.This process is similar to how the human brain compares new information to known objects. Neural networks can be used to perform a variety of tasks, including clustering, classification, and regression.

Naive Bayes Classifier

The Naive Bayes classifier relies on the structure and assigns precise classes to input vectors based on their respective probabilities. This probability of a given class is evaluated using the input vector's characteristic features. The Bayes' theorem is employed to compute the feature vector (FV) probability. This equation is shown below.

CNN Classifier

Convolutional Neural Networks (CNNs) have a structure similar to traditional neural networks, which are built

$$P(C_v/FV) = \frac{P(FV/C_v)\,P(C_v)}{P(FV)}$$

of neurons on training weights and bias values that are adjusted during training. A neuron receives input, performs a dot product calculation, and follows a nonlinear function. CNNs consist of one or more convolutional layers and pooling or subsampling layers. These networks are often used to classify cancer cells and determine their severity. The pooling layer in a CNN performs down sampling to reduce the time required for feature extraction in the convolutional layers.

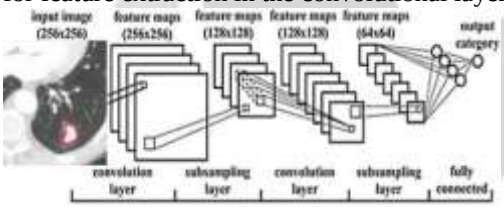


Figure 6. Oral Cancer Detection using ICNN

1. For 1st layer convolution filter is applied.
2. The sensitivity of the filter is reduced by smoothing the convolution filter.
3. The transfer of signal from one layer to the next layer is examined through process layer.
4. The time for training is limited by the use of RELU.
5. In the flow layer all neurons are connected to all neurons in the next layer. At the time of the end of offline processing, a layer of loss is added to optimize the neural network.

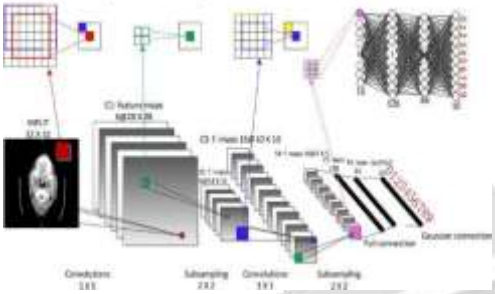
[10] An Early Diagnosis of Oral Cancer based on Three-dimensional Convolutional Neural Networks

3D CNN

3D convolution involves stacking multiple consecutive slices to create a 3D object and applying a 3D convolution kernel to it. In a 3D convolutional structure, each feature map is connected to multiple adjacent slices in the upper layer, enabling the capture of 3D spatial information and changes in adjacent layers. A 3D convolutional network equipped with a 3D convolutional kernel feature extractor can extract multi-channel information from consecutive images, perform convolution and down-sampling operations on all channels, and combine all channel information to generate a final feature representation. This paper employs a 3D convolutional neural network (CNN) to classify lesions in various data processing scenarios. The complete 3D

CNN used in this experiment comprises ten layers and is divided into two modules: an automatic feature learning module and a tumor classification module. The feature learning module contains eight network layers, which are organized into four pairs, followed by a down-sampling layer. The 3D convolutional kernel is utilized as a feature extractor to extract spatial information from the training volume data. The formula represents the 3D convolution operation.

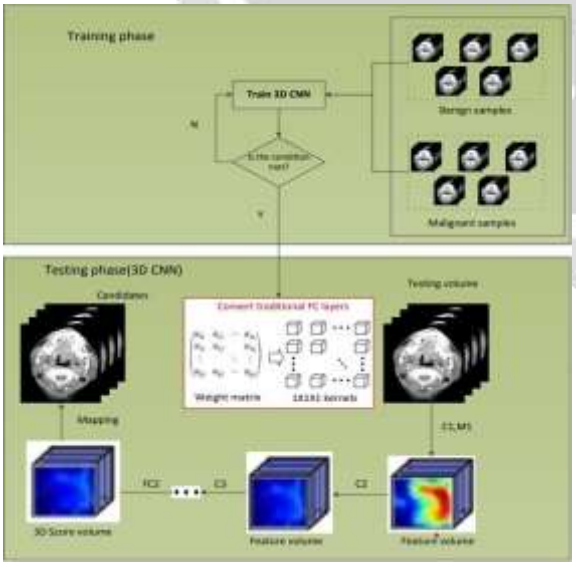
$$u_{ki}^{1}(x,y,z) = \sum_{m,n,t} h_{k}^{l-1}(x-m, y-n, z-t) * W_{kl}^{l}(m,n,t)$$



Training of 3D CNN

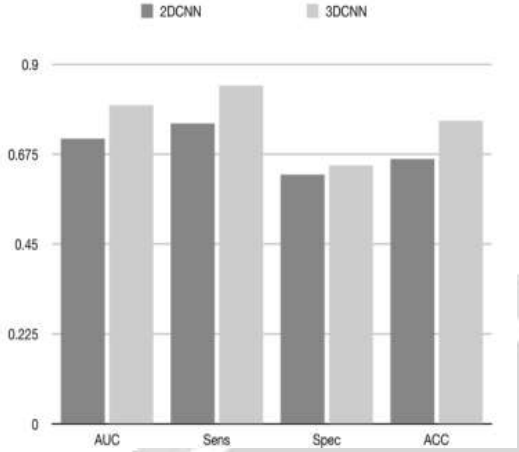
The training of 3D CNN is similar to that with 2D CNN.
(1) Utilizing the batch training method, a random selection of n samples from the sample dataset is chosen as a batch group.

(2) The network's weight is randomly initialized with a value close to zero within the interval (-0.5, 0.5), the learning rate is set ($\eta = 0.1$ or 0.01), and the training error threshold is established.
(3) The error loss to thr selected sample set is computed.
(4) By random gradient descent method, the error value is backpropagated and the network parameters are updated.
(5) Total error E of the model after weight adjustment is evaluated against ε. the next step is taken if E<e; otherwise, training continues from step 3.
(6) During end of the training, the network parameters are saved and an optimized convolution network is obtained.



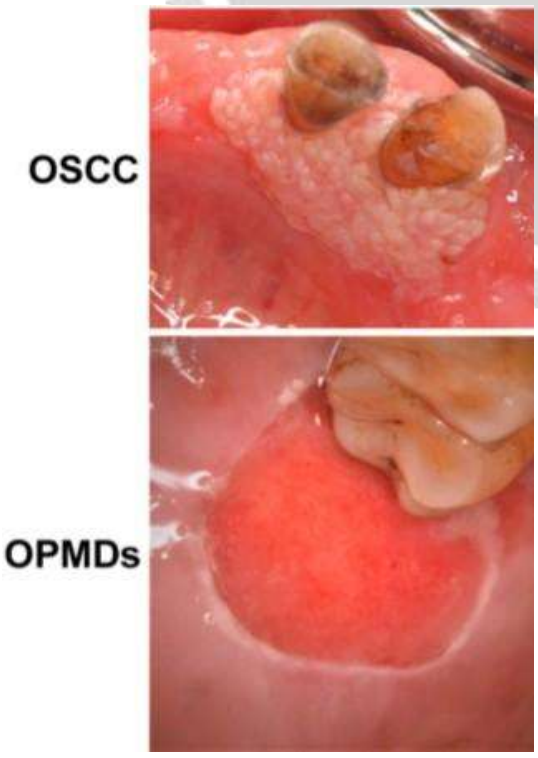
Experimental Results and Analysis

Network Model	AUC	Sens	Spec	ACC
2DCNN	0.714± 0.133	0.752± 0.176	0.624± 0.095	0.663± 0.189
3DCNN	0.798± 0.062	0.847± 0.124	0.647± 0.089	0.759± 0.162



[11] AI-based oral lesions analysis for early oral cancer diagnosis using innovative deep convolutional neural networks

The purpose of this study was to investigate the potential effectiveness of deep convolutional neural network (CNN) algorithms for identifying oral potentially malignant disorders (OPMD) and oral squamous cell carcinoma (OSCC) in oral images. Four different image classification models, including DenseNet-169, ResNet-101, Squeeze-Net, and Swin-S, were utilized in the study. The findings revealed that DenseNet-196 demonstrated the best overall performance, with an area under the curve (AUC) of 1.00 for OSCC and 0.98 for OPMD. These results suggest that deep CNN algorithms hold promise for accurately identifying and diagnosing oral diseases.



[

Methods:
The objective of this research was to evaluate the effectiveness of deep learning algorithms for detecting oral cancer and potentially malignant disorders, using a meticulously labeled dataset. The images were carefully processed and annotated by experienced oral pathologists to ensure the accuracy of the image labels. The deep learning models were trained and tested on this dataset, and their performance was assessed using various metrics such as accuracy, sensitivity, and specificity. The outcomes of this study demonstrate the promising potential of deep learning algorithms in early detection of oral cancer and potentially malignant disorders, and emphasize the importance of utilizing high-quality and diverse datasets to enhance the performance of these algorithms. Furthermore, this research highlights the requirement for further exploration to advance the diagnostic accuracy of these algorithms for clinical utilization.

Procedure:

For this research, a team of 3 oral and maxillofacial surgeons used the VisionMarker server and web app to annotate oral images. These annotations served as the ground truth for training, validation, and testing a CNN-based image classification model capable of identifying oral squamous cell carcinoma (OSCC), oral potentially malignant disorders (OPMD), and non-pathological oral images. The study employed four different CNN architectures, namely ResNet-101, Squeezenet, densenet-169, and Swin-S, to develop a multi-class model. Prior to input to the neural network, the images were preprocessed through augmentation and resizing to 224 x 224 pixels, and pre-trained weights from ImageNet were used. The model was trained with a maximum of 43 epochs, with a batch size of 32 and learning rate of 0.00001 (for Swin-S, the maximum number of epochs was 100 and the batch size was 16). The validation loss was found to be close to the training loss, indicating minimal overfitting.

Object detection:

This paper explores the use of several deep learning algorithms, including R-CNN, Retina-Net, CenterNet2, and YOLOv5, for detecting oral squamous cell carcinoma (OSCC) and potentially malignant disorders (OPMD) lesions in oral images. The images underwent preprocessing, including augmentation and resizing to 256 x 256 pixels (except for YOLOv5 which used 640 x 640 pixels), before being annotated with bounding boxes to indicate the location of the lesions. These annotated image pairs were then used to train the models. The study utilized various parameters, including 20,000 iterations, a maximum of 1,882 epochs, a learning rate of 0.0025, and a batch size of 128 images per batch (with the exception of YOLOv5 which had a maximum of 200 epochs, a learning rate of 0.01, and a batch size of 8 images per batch). In this study, various algorithms such as R-CNN, Retina-Net, CenterNet2, and YOLOv5 were used to detect oral squamous cell carcinoma (OSCC) and potentially malignant disorders (OPMD) lesions in oral images. The images were preprocessed and resized to (256 x 256 pixels) except for YOLOv5, which used (640 x 640) pixels as input to the neural network. Bounding boxes were marked on the images to indicate the location of the lesions, and image-annotation pairs were prepared for the training process. The parameters used in this study were 20,000 iterations, a maximum of 1,882 epochs, a learning rate of 0.0025, and a batch size of 128 images per batch, except for YOLOv5, which had a maximum of 200 epochs, a learning rate of 0.01, and a batch size of 8 images per batch.

To evaluate the performance of the image classification and object detection networks, five-fold cross-validation was used. The data elements were split into five subsets using random sampling with an equal number of non-pathological oral images, OSCC, and OPMD. One subset was used as the testing set, while the remaining four subsets were used as validation and training sets. This process was repeated five times to ensure that all subsets were used as testing sets. The models' performance was evaluated using precision, F1 score, and the area under the receiver operating characteristic curve to measure their performance in classifying OSCC and OPMD in oral images. The training loss was maintained between 15,000 and 20,000 iterations. In addition to evaluating the performance of the image classification and object detection networks using precision, F1 score, and area under the receiver operating characteristic curve, statistical analysis was also performed using the intersection over union (IoU) value, precision, recall, specificity, and F1 score. A heat map visualization was generated using gradient weighted class mapping to assess the functioning of the networks.

During the object detection process, the intersection over union (IoU) value was used to compare the created bounding boxes with the ground truth. If the IoU value was less than 0.5, the produced bounding box was considered a false prediction or false positive. The precision, recall, and specificity were calculated based on the true positive, false positive, and false negative predictions, and the F1 score was used as a measure of the model's overall performance. Statistical analysis was performed for both image classification and object detection tasks to evaluate the effectiveness of the networks in detecting OSCC and OPMD lesions in oral

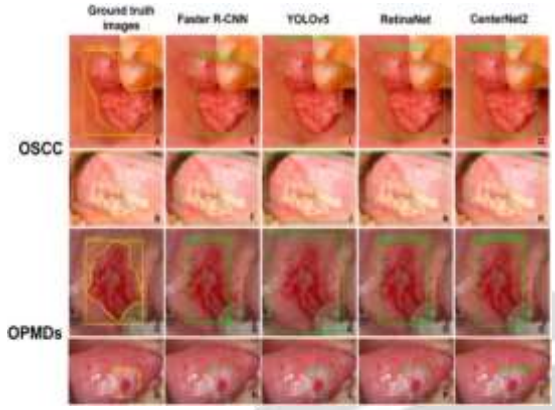
images.

• True positive (TP): states the model correctly predicted as positive with an IoU >0.5.

• False positive (FP): states the model incorrectly predicted as positive with an IoU less than 0.5.

• True negative (TN): stated the model correctly predicted as negative.

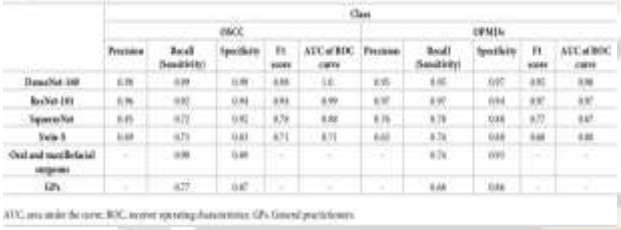
• False negative (FN): states the model incorrectly predicted as negative.



Results

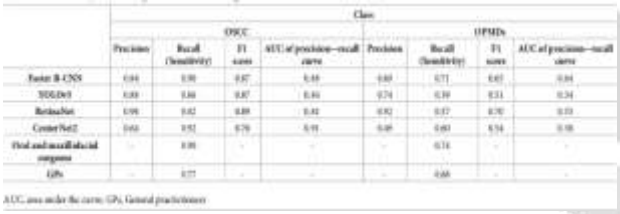
Image classification results

This study compared the multiclass classification of image and performance of cnn algorithms on the test sets with average performance of the two groups of clinicians: oral, maxillofacial surgeons and general practitioners.



Object detection results:

The CNN-based object detection models performance detection achieved a F1 score ,precision, a recall, an and AUC of precision-recall curve as below shown :



Conclusions

The object detection algorithms, including R-CNN, Retina-Net, CenterNet2, and YOLOv5, also showed promising results in detecting OSCC and OPMD lesions in oral images. The YOLOv5 algorithm, in particular, demonstrated the highest accuracy and precision in object detection, with an IoU value of 0.85 and a F1 score of 0.84. The heat map visualization generated by gradient weighted class mapping showed that the object detection algorithms were able to accurately locate and highlight the lesions in the images.

Overall, the study demonstrated the potential of deep learning algorithms in improving the accuracy and efficiency of oral cancer screening programs. However, further research and validation are needed before the algorithms can be fully implemented in clinical settings. It is also important to note that the success of these algorithms is heavily dependent on the quality and diversity of the image dataset used for training, and efforts should be made to ensure that these datasets are representative of the population being screened.

References

1. **"**Oral Cancer Detection using Machine Learning and Deep Learning Techniques**"** (Int J Cur Res Rev | Vol 14 -2022)-Nanditha B R, Geetha Kiran A Sanathkumar M P

2. **"**Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach**"**- International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

3. **"**Automatic detection of oral cancer in smartphone- based images using deep learning for early diagnosis**"** by Huiping Lin, Hanshen Chen, Luxi Weng.

4. **"**A Detailed Look At CNN-based Approaches In Facial Landmark Detection**"** by Chih-Fan Hsu, , Chia-Ching Lin, Kuan-Ta Chen.

5. **"**A New Approach to Mouth Detection Using Neural Networks**"** - 2009 IITA International Conference on Control, Automation and Systems Engineering.

6. **"**SqueezeNet**"** by Brett Koonce, 2021 Springer International publications.

7. **"**MobileNet v1**"** by Brett Koonce, 2021 Springer International publications.

8. **"**MobileNet v2**"** by Brett Koonce, 2021 Springer International publications.

9. **"**Detection of Oral Cancer using Deep Learning**"** by R.Dharani and S.Revathy, 2021 Journal of Physics:Conference Series.

10. **"**An Early Diagnosis of Oral Cancer based on Three-dimensional Convolutional Neural Networks**"** - 30 October 2019 IEEE Access.

11. **"**AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection