Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 26 March 2026
Revised 27 April 2026
Accepted 20 May 2026

Natural Resources for Human Health 2026, Vol. 6 (2s): 759–776

KEYWORDS:
Oral Squamous Cell Carcinoma, OSCC, Deep Learning, Mixup Augmentation, Focal Loss, EfficientNet, EfficientNetV2, MobileNetV3, ResNet-50, Medical Image Analysis, Cancer Detection, Histopathology

Detection of Oral Squamous Cell Carcinoma Using Mixup Augmentation and Deep Learning

Kuruvikulam Chandrasekaran Arun[1], Balakrishnan Koustubha Madhavi[2], Pravin Madhukar Dhanrao[3], Debabrata Swain[4]*, Viraj Bhagiya[5], Om Deshmukh[6], Manish Kumar[7], Davinder Paul Singh[8]

[1, 2]School of Technology, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia. Email: 1Kchandran.arun@gmail.com, 2kousmadhu717@gmail.com
[3]Department of Computer Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India. Email: dhanraopravincomp@sanjivani.org.in
[4*,5,6,8]Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, India. Email: 4debabrata.swain7@yahoo.com, 5virajbhagiya@gmail.com, 6omdeshmukh170905@gmail.com, 8devsingh0071@gmail.com
[7]Department of Electronics and Communication Engineering, Pandit Deendayal Energy University, Gandhinagar, India, Email: manish.kumar@sot.pdpu.ac.in
*Corresponding Author: Debabrata Swain, debabrata.swain7@yahoo.com

ABSTRACT: Oral Squamous Cell Carcinoma (OSCC) is a significant global health challenge where early and accurate detection is critical for improving patient outcomes. Histopathological examination is a method for diagnosing OSCC, but the issue is the labour, time spent and the observer variability. In this research, deep learning based OSCC identification from histopathological images was explored. Mixup data augmentation and Focal Loss to improve model outcomes. The study compared CNN architectures, EfficientNet, EfficientNetV2, MobileNetV3 and ResNet-50. Evaluation process included 5-fold stratified cross validation to get unbiased results. EfficientNetV2-Medium achieved the best overall performance, with an Accuracy of 0.9914, Precision of 0.9988, Recall of 0.9840, and an F1-score of 0.9913. These findings show that combining advanced data augmentation with modern CNNs can produce a robust and highly accurate tool to support pathologists in OSCC diagnosis.

1. INTRODUCTION

According to 2022 statistics around 389,000 fresh cases were detected and 188,000 mortalities were reported[1]. Over 90% of oral cancers are of the forms known as oral Squamous Cell Carcinoma (OSCC) [2], [3]. OSCC originates in the squamous epithelial lining of the oral cavity and oropharynx. Some of the factors that cause this problem are excessive tobacco and alcohol consumption and betel quid chewing [4], [5]. Infections with HPV strains are also related to OSCC [4]. From the global occurrences it is observed that geography along with the stated habits cause this issue [3].

The diagnosis of a patient also plays an important role in prognosis. Patients detected of OSCC In early stages (Stage 1/ Stag2) are observed to have 5-year survival rates exceeding 80% [6]. Diagnosis at later stages (Stage III/IV), have larger tumours or metastasis, resulting in a much poorer prognosis, with survival rates often below 50% [2], [7]. Clinically, OSCC can appear as non-healing ulcers, growths, or discoloured patches (erythroplakia or leukoplakia), but early lesions may be subtle or asymptomatic [8]. These factors being very common and unknown to patients delay the diagnosis process [6].

The definitive diagnosis of OSCC depends on the histopathological examination of a tissue biopsy from a suspicious lesion

[9]. Haematoxylin and Eosin (H&E) stained tissue are analysed by pathologists to identify problematic features like cellular atypia, abnormal stratification, dyskeratosis, increased mitotic activity and stromal invasion [9].

Histopathological assessment faces several challenges. The human based review of slides is labour-intensive and time consuming. Biopsy-based diagnosis is prone to sampling errors, and the assessment is largely qualitative. Imaging techniques like CT, MRI and PET play an important role for staging but lack the cellular level detail for primary diagnosis which depends on histopathology [10].

The use of Whole-Slide Imaging (WSI) in digitization of entire glass biopsy slides has made the basis for computational pathology [11]. Such large datasets are ideal for deep learning algorithms, particularly Convolutional Neural Networks (CNNs). CNNs could learn difficult features and patterns in images directly from the pixels like humans [12].

Considering the effectiveness of CNNs in tasks like diabetic retinopathy and skin cancer classification. DL can help overcome limitations of conventional methodology for OSCC histopathology.

The challenges encountered during large digital OSCC image datasets are patient privacy, cost and time-consuming annotation. Data augmentation helps overcome data limitations. Mixup augmentation [17] creates new training samples by combining pairs of images and their labels, increasing data diversity to help the model learn better. The study focuses in the effectiveness of combining Mixup augmentation with modern CNN architectures for automated OSCC detection. The models chosen for the research are CNN models, EfficientNet, EfficientNetV2, MobileNetV3, and ResNet-50. A 5-fold cross-validation is also used to ensure reliable evaluation.

The contribution of this research is in the methodological pipeline that was followed. The points below illustrate them:

1. The pipeline integrated image standardization, transfer learning, two-phase fine tuning, model targeted threshold selection into one framework.

2. Dual stage data augmentation combined offline augmentation to have a balanced dataset with online augmentation to improve generalization.

3. Mixup augmentation was used to generate synthetic training samples that increased the ability of the models to fetch greater results.

4. A unified framework was developed by integrating Focal Loss, AdamW optimization, Cosine Annealing learning rate scheduling, early stopping and gradient clipping.

5. The study employed Youden's J statistic to get an optimal threshold than the fixed 0.5 classification barrier.

6. A common custom classification head was designed and integrated with all pre-trained CNN backbones for consistent transfer learning. The classifier combines Layer Normalization, Dropout, SILU activation and fully connected layers and finally a Sigmoid output.

The rest of the paper follows this section wise roadmap. Section 2 summarises the reviewed research and the gaps observed in them. Section 3 explains the approach of the research. The following section aims to discuss about results those were achieved during the study. Section 5 compares the employed mechanism with the referred literatures. Finally, the conclusion derived are stated by the last section.

2. RESEARCH REVIEW

The motivation for this study is to investigate computationally efficient models and apply pre-processing steps that can achieve enhanced results. The use of advanced augmentation, optimization and transfer learning techniques to provide a fair evaluation was the main objective.

A.-u. Rahman et. Al [18] depicted a transfer learning based OSCC detection system using customized AlexNet architecture. It used an OSCC dataset from Kaggle with around 2511 affected and 2435 normal images. All the images were resized to 227x227 pixels as required by the AlexNet's input. It modified the final layer of the pre-trained AlexNet. ReLU was used as an activation function. Softmax output layer for binary classification was used. The model was trained for 70 epochs using learning rate equal to 0.001 and a 70:30 train-test split was used. The approach achieved the highest accuracy of 97.66%.

S. Panigrahi et. al [19] presented a DL based Computer Aided Diagnosis (CAD) system for classifying OSCC images. The

study aimed to bifurcate between benign and malignant lesions for early oral cancer diagnosis. Addreses the challenge of limited availability of large annotated medical datasets by using transfer learning. The research proposed 2 approaches with pre-trained DCNN architectures: VGG16, VGG19, ResNet-50, InceptionV3, MobileNet. And the second prepared a baseline CNN trained from scratch consisting of 10 convolutional layers. ResNet-50 Emerged as the best performing model with highest accuracy of 96.6%. transfer learning outperformed the CNN architecture.

A. Yadav et al. [20] evaluated 4 CNN architectures ResNet-18, AlexNet, DenseNet-169, DenseNet-201. It incorporated Cyclic Learning Rate (CLR) strategy to adjust dynamically the LR during training. DenseNet-169 and DenseNet-201 used dense feature reuse (dense connectivity). The study did an overall comparative study to identify the most suitable CNN architectures for OSCC classification. DenseNet-169 achieved the highest classification accuracy 95% outperforming the rest.

K. Dhanya et al. [21] used pre-trained CNNs for feature extraction, and a tuned CNN performed the final classification. The dataset was divided in training with 4141 images and 1014 for validation and 126 for testing. Image processing and data augmentation including resizing, random zoom, random rotation and horizontal flipping to reduce overfitting. The research used 3 architectures ResNet-50, VGG16 and InceptionV3. Randomized Hyperparameter tuning was used for network optimization and searching for the best learning rate. The model performances were compared with the help of standard metrics and also sensitivity. ResNet-50+ tuned CNN had the best overall performance with an accuracy of 90% and sensitivity of 97%.

A. Soni et al. [22] evaluated 17-pretrained architectures and performed two stage statistical analysis for the same problem statement. EfficientNetB0 was chosen as the baseline architecture among the evaluated models. The research enhanced the EfficientNetB0 by incorporating Dual Attention Network (DAN) to focus on spatial and channel wise features in the images. Here metrics such as Mathews correlation, Cohen's Kappa score and computational Time were used apart from the normally used statistical scores. EfficientNet and DAN model achieved CK of 88.8%, Accuracy of 91.1% and computational Time of 66.41%. ReLu gets stuck at local minima

E. Albalawi et al. [23] proposed EfficientNetB3 based DL system for automatic classification of OSCC images. A public dataset of 1224 H&E-stained biopsy pictures collected from 230 patients. Image preprocessing including resizing and ImageDataGenerator augmentation was used to improve the model output. Exploratory data analyses was done with the help of visualisation, null value analysis and inspection of the data prior to the training. For reliable evaluation training, testing and validation was done. The system employed ImageNet pretrained weights and added Normalization, Dense and Dropout layers for binary classification. To prevent overfitting and to dynamically adjust the LR during training Early Stopping and ReduceLROnPlateau callbacks was applied. Achieved a test accuracy of 98.73% and precision of 99%.

X. Wei et al. [24] considered 4 cancer types including OSCC that comprised of 75% of the data. Four tumour grades were considered. It addressed the gaps in traditional ML methods, also emphasized on the problem of gradient descent getting stuck in local maxima. The aim of the research was using an Improved Tunicate Swarm Algorithm (ITSA) to improve OSCC detection from tongue and lip images. The problem was the small dataset size. SdSMOTE was used for data augmentation to deal with class imbalances and to generate synthetic minority class samples. A fitness function evaluation to calculate the computational complexity. CNN-ITSA achieved the highest accuracy of 98.70%.

M. Ahmad et al. [25] compared DL, CNN+SVM and CNN+ handcrafted features+ SVM to identify the best model. The preprocessing included RGB colour normalization for consistency. Artifact removal, contrast enhancement and ROI enhancement were also included. Gaussian Filter and Laplacian Filter to remove noise and to sharpen the lesion edges were recruited. 5 transfer learning models were used Xception, InceptionV3, InceptionResNetV2, InceptionResNetV2, NASNetLarge, DenseNet201. NASNetLarge achieved the accuracy of 94.44%. (DenseNet201 + GLCM + HOG + LBP + SVM) achieved the highest overall performance with 97.00% accuracy.

N. W. Asnake et al. [26] pre-trained CNN models combined with XGBoost. To determine the best model for accurate OSCC classification. Preprocessing steps included color normalization to remove staining variations and Bilateral Filtering to remove noise. To improve the image contrast and to prevent over-amplification of noise CLAHE (Contrast Limited Adaptive Histogram Equalization) was also included. InceptionV3, MobileNetV3, AlexNet were used as feature extraction. 5 experiments were done combining XGBoost with each of the three and using softmax with InceptionV3 before and after preprocessing. AlexNet+XGBoost achieved the highest accuracy score of 99%.

N. Das et al. [27] aimed to classify the OSCC images into 4 classes according to Broder's grading system (Normal, well differentiated, moderately differentiated and poorly differentiated). The paper showed 2 approaches one was using transfer learning and the other was making their own CNN from scratch. Data augmentation was done to increase the training patches from 8321 to 37816. It experimented with optimizers: SGD, RMSProp and Adam and 4 learning rates. CNN performed the best with 97.5% and from the Transfer Learning models VGG19 with 92%.

T. Majeed et al. [28] addressed the challenge of class imbalance in histopathology images. It also stated the time consuming analysis of H&E whole slides. Different imbalance ratios were used ranging from 0.10 to 0.30. The study combines transfer learning with sampling techniques. Three sampling methods were sued: SMOTE, Deep SMOTE and ADASYN.

S. K. D. Sharma et al. [29] Dataset contained 5192 images with almost equal distribution of both normal and OSCC classes. Image preprocessing were done before training including manual cropping, resize all images to 224x224x3. Data augmentation was applied to only training data to increase data diversity and better generalization. Five pre-trained models were employed out which EfficientNetB3 got the highest accuracy 98.33%. According to the research EfficientNetB3 performed the best due to strong gradient propagation and compound scaling strategy.

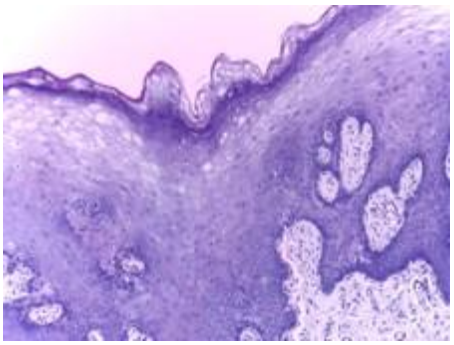
3. RESEARCH METHODOLOGY

3.1. Data Collection

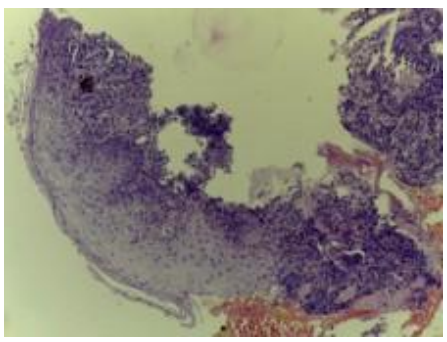
The strategy used a dataset published by Rahman (2019) [31] which is available for public use. The dataset had 89 unaffected tissue images and 439 affected images. The images were captured at 100 times magnification with the help of a microscope called Leica ICC50 HD after H&E straining. The samples were collected from 230 different patients, medical professionals organized the slides before making the dataset.

H&E is a technique used to colour the nucleus of the cell with blue or purple with the Haematoxylin. Eosin is used to stain the cytoplasm and surrounding tissue pink. Due to the high image resolution and clear microscopic details, the dataset is appropriate for training and evaluating classification models.

Figure 1 Images in the dataset



(b) Normal Image



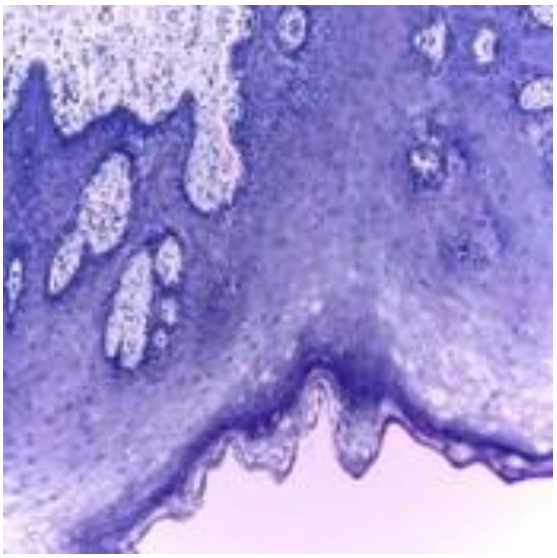
(b) OSCC Image

3.2. Preprocessing of data

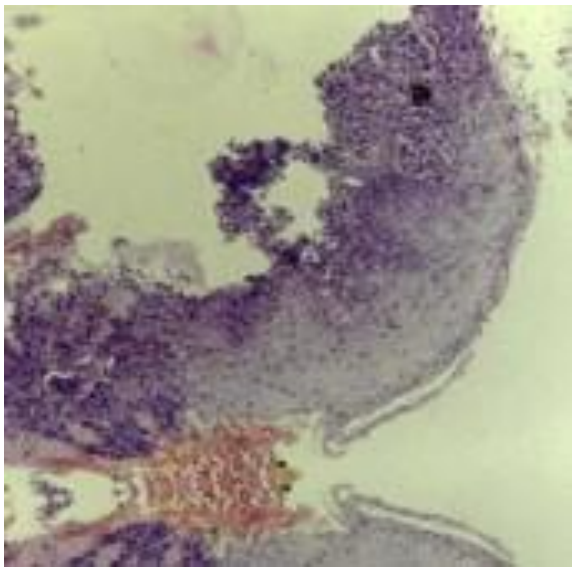
3.2.1. Handling dataset imbalance

The procured dataset was imbalanced as there were more samples of OSCC images than normal images. To overcome data imbalance offline augmentation was done before model training and the augmented images were saved. The goal was to create a balance between dataset with normal 2500 samples of each class. All the original patches were retained in the final dataset. To create varied training samples geometric augmentations like rotation, shifting, zooming and flipping were applied. A custom colour augmentation was applied in which natural H&E staining features like hue, saturation and contrast were changed slightly and some images were kept as it is. Reflection padding helped in dealing empty region to avoid introduction of artificial padding.

Figure 2 Augmented pictures



(a) Augmented Normal Image



(b) Augmented OSCC Image

3.2.2. Image Standardization

The image standardization process for this investigation was designed to ensure compatibility with pre-trained models, involving two critical steps. The image patches were changed to a size common to a resolution that matched the input size

required by the architecture. ImageNet mean and standard deviation was used to normalize pixel values [32]. This normalization was used to make the distribution of the histopathological pictures like the data prior to training.

3.3.2. Mixup Augmentation

The study applied online data augmentation during training withing each fold of the cross-validation process to decrease overfitting. These changes were applied only to training data generating new image variations dynamically in each training iteration. Standard geometric augmentations were applied with probability of p=0.5 for each transformation. The transformations included random horizontal and vertical scaling, translation and rotation to simulate natural variations of histopathological images.

The study included Mixup augmentation as a data regularization technique to improve robustness and generalization of the model [17]. Mixup generated synthetic images by combining two randomly selected images linearly with corresponding labels. Given two image-label pairs (xi, yi) and (xj , yj).

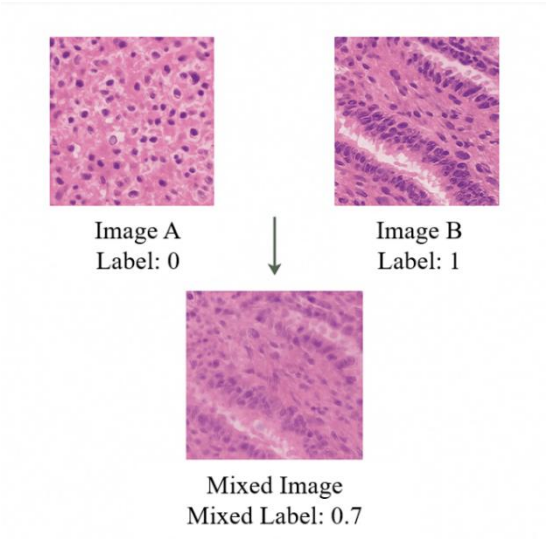
{(x',y'): new training sample , refer (3) and (4) }

$$x' = \mathrm{m}.x_i + (1-\mathrm{m})x_j \quad (3)$$

$$y' = \mathrm{m}.y_i + (1-\mathrm{m})y_j \quad (4)$$

m is the mixing coefficient sampled from symmetric Beta distribution with $\alpha = 0.4$ based on the references. 'm' was restricted to the range [0.1, 0.9] to avoid unrealistic distortion of images and both the type of images contribute meaningfully. This augmentation applied for the training enhance the learning process. For the validation images were only resized and normalized.

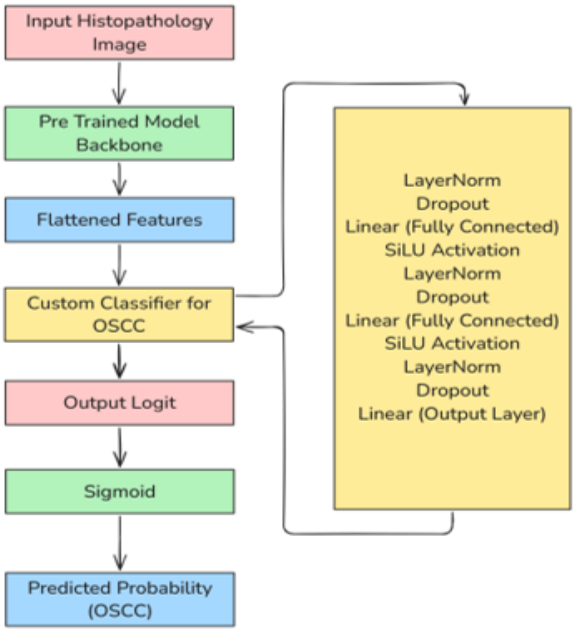
Figure 3



3.3. Model Training

The system is built on transfer learning, it utilizes CNN architecture that have been trained on general visual features from large scale datasets. These CNN models are pre-trained in the Imagenet dataset containing images across 1000 categories making them ideal for learning features as textures, shapes and patterns. The task of binary OSCC detection the pre-trained models are changes by replacing the original classification layer with custom classification head for binary classification. The diverse set of models enables a comparison to determine which CNN architecture shows best performance to given scenarios.

Figure 4 Custom DL model



3.3.1. ResNet-50

ResNet-50 was used as a baseline model due to it performance across varied applications to compare new CNN models [30]. It is a deep neural network consisting of 50 trainable layers. ReLu function is applied to each convolutional layer to introduce nonlinearity. Resnet uses skip connections (residual blocks) to overcome a problem called Vanishing Gradient generally observed in deep neural networks. These skip connections provide a direct information path and gradients to flow through the network to improve learning stability during training.

The residual block operation is defined by equation (1):

$$b = f(a, W_i) + a \quad (1)$$

In the residual block, the input is added to the output of the block's transformations allowing the network to learn only residual mapping instead of the complete transformation to simplify optimization. It was chosen as baseline model for reference to compare with other model performances.

3.3.2. EfficientNet

Beyond the baseline model the research included the EfficientNet family to determine whether it could provide greater performance as well as computational efficiency. EfficientNet uses a compound method of scaling to overcome problems of increasing network depth or width to have effective model scaling [33]. Unlike CNNs that independently increase network depth or width, EfficientNet uniformly scales the network dimensions and image resolution using compound scaling method.

To achieve balanced model growth uniform scaling of depth, width and input image resolution. The balanced scaling strategy follows fewer FLOPs (Floating Point Operations) compared to CNN. The aims to evaluate whether the MBConv blocks could provide some computational efficiency with good performance.

3.3.3. EfficientNetV2

EfficientNetV2 was employed to check whether it could give better results than EfficientNet. It uses Fused-MBConv layers in the early stages of the network it reduces memory access overhead and speed up computation, however the later layers continue to use MBConv. The Fused-MBConv block replaces the depth wise convolutional and expansion layer with a single regular convolutional layer. EfficientNetV2 uses a progressive learning strategy in which the input resolution is gradually increased during training to stabilize the learning procedure. EfficientNet V2 was chosen because it provided an

balance between classification accuracy, efficiency and training speed.[34]

3.3.4. MobileNetV3

MobileNetV3 was included in the research to check whether it can effectively be deployed for less resource-based environment [35]. It was developed using Neural Architecture Search (NAS) to achieve an optimal balance low latency and high classification accuracy mobile processors. Its architecture incorporates SE (squeeze & excitation) module. The mathematical formulation of the h-swish activation function:

$$h - swish(s) = s \cdot \frac{ReLU6(s+3)}{6} \qquad (2)$$

h-swish function is designed to be compatible with mobile hardware and maintains the advantages of standard nonlinear functions. The h-swish activation function provides a computationally reliable approximation of the Swish function. The architecture was optimized using Neural Architecture Search (NAS) together with the NetAdapt algorithm to identify an optimal trade-off between latency, model size, and classification accuracy. The inclusion of this model for comparison with others is to explore low-cost diagnostic.

3.4. Training and Optimization Components

The study carefully selected the loss function, optimizer, and learning rate scheduler due to their important role. These training components were specifically selected to address problems associated with histopathological image classification.

3.4.1. Cross-Validation Setup

The study used 5-fold stratified cross validation for enhanced performance. The 5000-image dataset was divided into 5 folds using stratified-k -Fold ensuring each fold having equal class distribution. During each iteration 4 folds were used for training and 1 was used for testing.

3.4.2. Activation Function (SiLU)

The study employed sigmoid weighted Linear Unit (SiLU) also known as Swish. Equation 5 defines the SiLU function. Here σ(x) represents the sigmoid function. Unlike ReLU which has non differentiable point at zero SiLU is smooth and continuously differentiable activation function. The non-monotonic behavior of SiLU provides greater representational capability than ReLU.

$$\text{SiLU}(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}} \qquad (5)$$

3.4.3. Focal Loss function

The loss function calculates the difference between the predicted model and the actual labels and guiding it reduce errors during the training period. In OSCC classification some samples are more difficult to classify like early-stage lesions than obvious cancer cases, so using loss functions like Binary Cross Entropy (BCE). Focal Loss was employed to conquer this problem [36] which emphasises on the hard to classify samples. Focal loss introduces a factor called modulating factor to reduce the influence of easy samples and making the model focused on difficult cases during training. The Focal Loss is defined as:

$$L(q) = -\beta(1-q)r\log(q) \qquad (6)$$

{ L(q): Focal loss value, q: estimated probability of the true class, β: balancing coefficient, r: focusing parameter}

Here, pt is the model's predicted probability for the correct class, and γ is the focusing parameter that controls the easy samples contribution.

3.4.4. Optimizer (AdamW)

The optimizer is responsible for updating the model parameters during training. It combines the strength of AdaGrad and RMSProp by using 1st and 2nd moments of the gradients for adaptive parameter updates. The normal Adam optimizer uses weight decay (L2 regularization) by adding it directly into gradient update, which reduces the effectiveness especially for large gradients. Therefore, to avoid such an issue the study chose AdamW [37] which applies weight decay independently from the gradient based parameter update. Equation 7 presents the normal Adam update rule in which the weight decay term is coupled with the gradient.

$$\theta_{t+1} = \theta_t - \eta\left(\frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon} + \lambda\theta_t\right) \quad (7)$$

{θ_t: model parameters, λ: weight decay coefficient, η: learning rate, $\hat{m}_t$ and $\hat{v}_t$: the bias-corrected first and second moment estimates}

Equation 8 illustrates the AdamW update rule, where weight decay is applied separately from the gradient update. By decoupling weigh decay from the optimization step it provides more consistent regularization and stable training making it suitable for DL models.

$$\theta_{t+1} = \theta_t - \eta\left(\frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon}\right) - \eta\lambda\theta_t \quad (8)$$

3.4.5. Cosine Annealing

The learning rate (LR) is an important hyper-parameter because it determines the size of the parameter updates by the optimizer during training period. The study employed a Cosine Annealing LR Scheduler. Cosine Annealing gradually decreases the learning from maximum in the beginning to minimum value during the training phase by following cosine curve for smooth transition. Equation 9 shows the learning rate calculation at each training epoch. e_"max" and e_"min" represent the maximum and minimum learning rates, Tcur and Tmax present the current epoch and the total number of training epochs.

$$e_t = e_{min} + \frac{1}{2}(e_{max} - e_{min})\left(1 + \cos\left(\frac{T_{cur}}{T_{max}}\pi\right)\right) \quad (9)$$

This smooth, gradual reduction helps the model converge more robustly and settle into wider, more stable minima in the loss landscape, which is strongly correlated with better generalization to unseen data.

3.5 Evaluation Protocol

This part of the research focuses on the evaluation of the methods and architectures and the strategy to evaluate the models chosen.

3.5.1. Two-Phase Training Strategy

To effectively adapt already trained CNNs to the task of OSCC identification, a 2-phase training method was employed. In the initial phase, referred to as feature extraction head training, the primary body of the pre-trained model commonly known as the backbone was frozen. This means that the weights obtained from pre-training on the ImageNet dataset were kept unchanged. Only the parameters of the newly added custom classification head were updated during this stage. By restricting training to the head, the model was able to interpret the general-purpose features extracted by the backbone in the context of OSCC histopathology images. A relatively high initial LR of $1 \times 10{-}3$ was used in this phase to promote rapid adaptation of the classification head to the new domain.

Following this, the second phase involved full network fine-tuning, during which the entire model including the previously frozen backbone was unfrozen, allowing all parameters to be trainable. A low initial LR of $1{\times}10{-}5$ was chosen. The small LR helps the model to make gradual adjustments to pre-trained weights. This is done to avoid overwriting useful features learned from ImageNet. A batch size of 32 was chosen to balance training stability and memory usage. Early stopping with a patience of 5 epochs was recruited to reduce the overfitting chances.

The validation area under ROC was used as a criterion to monitor the training process to terminate automatically once the model's performance on unseen data plateau. Gradient clipping was used with maximum norm of 1.0 to maintain stability during backpropagation.

3.5.2. Performance assessment

The model performance assessment was done using the best performing model from each cross-validation fold. A set of binary classification metrics were employed to test the diagnostic accuracy. These metrics use four outcomes of a binary classifier. Four outcomes used to evaluate are: True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN). TP refers to a case in which OSCC image is correctly identified and on the other hand TN indicates that a normal sample is correctly classified as normal. False Positives and False Negatives are the cases where the predicted outcome is contradictory to the true output. These are used to calculate further metrics that are used in this study.

{True Positives: Tp, True Negatives: Tn, False Negative: Fn, False Positive: Fp}

Accuracy: The ratio of all the true predictions out of a all the predictions.

$$\text{Acc} = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (10)$$

Precision: It is the fraction of True Positives predicted out of all the positives predicted. It is used to measure the reliability of a positive OSCC prediction.

$$\text{Prec} = \frac{d}{d + c} \quad (11)$$

{d: true positives, c: false positives}

Recall (Sensitivity): Actual positives detected divided by sum of true positives and false negatives. This metric is important in medical screening as it represents low false negative rate.

$$\text{Rec} = \frac{y}{y + z} \quad (12)$$

{y: true positives, z: False negatives}

F1-Score: The harmonic mean of Precision and Recall, providing a single metric that balances the two.

$$\text{F1-Score} = 2 \times \frac{\text{p} \times \text{q}}{p + q} \quad (13)$$

{Precision: q, Recall: p}

Area Under the ROC Curve (AUC): This curve is a unit graph between False Positive Rates and True Positive Rates.

The area under this curve is a numeric measure of the ROC.

To convert the model's raw probability outputs (from 0 to 1) into a definitive binary classification (Normal or OSCC), a classification threshold must be selected. To do this optimally for each model, determined a model-specific threshold from the validation set predictions by maximizing Youden's J statistic. This statistic balances sensitivity and specificity to find the threshold that achieves the best trade-off between the two. It is defined as:

$$J = a + b - 1 \quad (14)$$

{Sensetivity: a; Specificity: b}

4. OUTCOMES & DISCUSSION

This part of the study details the results of the evaluated architectures for OSCC detection, based on the 5-fold cross-validation protocol.

Table I summarizes the key performance metrics, averaged across the 5 test folds, for each of the nine CNN architectures. A visual comparison of key metrics across all models is presented in Figure 5.

The results in Table I and visualized in Figure 5 depict the highest performer model EfficientNetV2-Medium. It demonstrated highest mean Accuracy (0.9914 ± 0.0019), Precision (0.9988 ± 0.0010), Recall (0.9840 ± 0.0034), and F1-Score (0.9913 ± 0.0019). Furthermore, it tied with EfficientNetV2-Small for the highest mean AUC (0.9995 ± 0.0002), indicating outstanding discriminative capability. The reported standard deviations reported across the folds are notably small for all key metrics, indicating very stable and reliable performance of the models.

The competitive results were achieved because of the training methodology. The training method uses two techniques together: Mixup Augmentation and focal Loss. The research used two methods because they help solve lack of diverse data and unequal class distribution. Mixup creates new training samples, making the dataset more diverse and helping the model to learn patterns that work well on new images. After the Mixup improves the data, Focal Loss helps the model spend more time learning from difficult and confusing image patches. Using Mixup and Focal Loss together was the main reason for achieving higher results and this approach can be used to other image classification problems.

The research uses ROC along with numerical evaluation metrics for visual representation to assess best performing model. ROC illustrates relationship between True Positive Rate and False Positive Rate. According to the curve which lies very close to the top left corner indicates that the model is effective for the classification process. High AUC score was obtained for the best fold and with overall average AUC confirming the model's performance. The model maintained low false positive rate. Figure7 depicts the confusion matrix of the best performing model. OSCC and Normal samples had equal share in the 1000 test set images. From this we get to know that 499 Normal images and 494 OSCC images with 0 false positives and 6 false negatives was achieved depicting great performance.

Table 1: Results of 5-Fold Cross-Validation On the model

Model	Acc	Prec	Rec	F1
Efficient Net B0	0.9868 ± 0.0018	0.9927 ± 0.0016	0.9808 ± 0.0057	0.9867 ± 0.0019
Efficient Net B4	0.9852 ± 0.0033	0.9935 ± 0.0035	0.9768 ± 0.0055	0.9851 ± 0.0034
Efficient Net B7	0.9800 ± 0.0055	0.9875 ± 0.0093	0.9724 ± 0.0062	0.9788 ± 0.0055
Efficient NetV2 Small	0.9908 ± 0.0026	0.9980 ± 0.0026	0.9836 ± 0.0039	0.9907 ± 0.0027
Efficient NetV2 Medium	0.9914 ± 0.0019	0.9988 ± 0.0010	0.9840 ±0.003 4	0.9913 ±0.001 9
Efficient NetV2 Large	0.9889 ± 0.0020	0.9968 ± 0.0030	0.9828 ± 0.0070	0.9874 ± 0.0021
MobileNetV3 Small	0.9870 ± 0.0026	0.9935 ± 0.0050	0.9804 ± 0.0020	0.9869 ± 0.0026

MobileNetV3 Large	0.9838 ± 0.0030	0.9939 ± 0.0029	0.9736 ± 0.0053	0.9836 ± 0.0031
ResNet-50	0.9800 ± 0.0024	0.9926 ± 0.0037	0.9672 ± 0.0058	0.9797 ± 0.0025

Figure 5: comparison of results

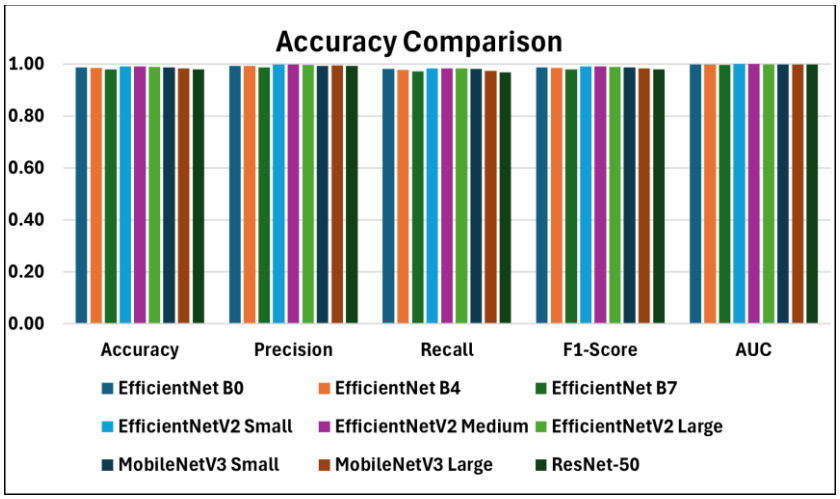


Figure 8 shows the images that EfficicentNetV2-Medium identified correctly for each of both classes. The correctly classified OSCC patches exhibited characteristics pathological features like nuclear atypia, pleomorphism and disorganized epithelial architecture, on the other hand Normal patches showed regular epithelial stratification and uniform nuclei. The comparison of 9 CNN architectures showed high performance, EfficientNetV2-Medium outperforming the rest. It achieved accuracy of 0.9914, precision of 0.9988, Recall (0.9840), and F1-score (0.9913). It had an AUC of 0.9995 and EfficientNetV2-small had the same result. These outcomes show the effectiveness of EfficientNetV2 model family with architectural such as Fused-MBConv blocks and compound scaling contributing to feature extraction from OSCC histopathological images. MobileNetV3-Small achieved AUC of 0.9986 and F1 score 0.9869 indicating the computational efficiency architectures can also achieve great performance. The model could provide a faster second opinion. The study and the developments only developed a help for medical professionals and not to replace them.

Figure 6: ROC of EfficientNetV2-Medium (best performer)

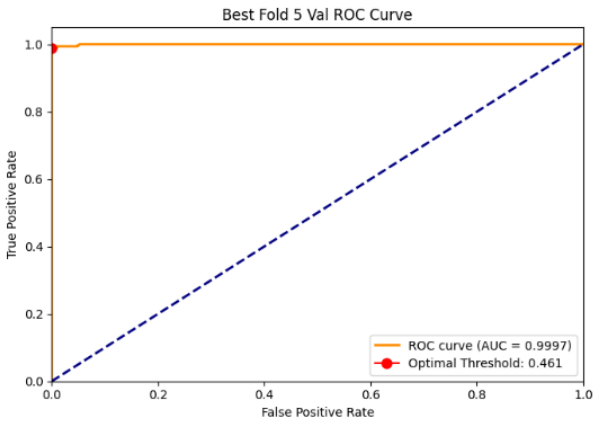


Figure 7: Confusion matrix

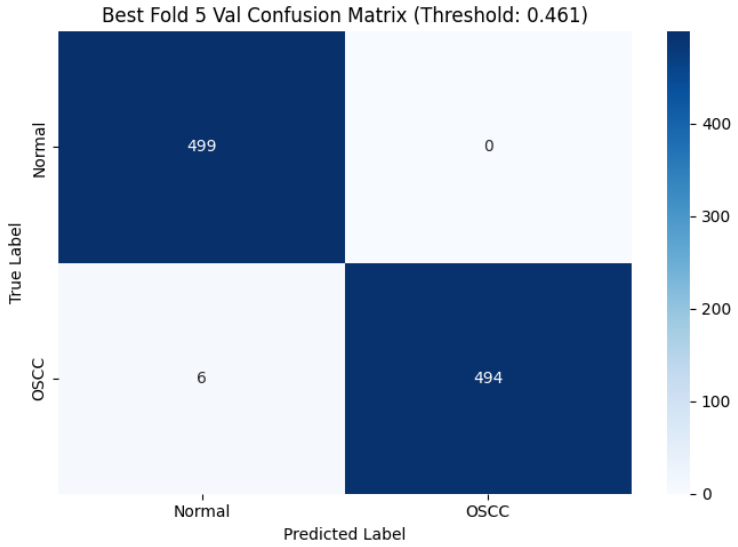
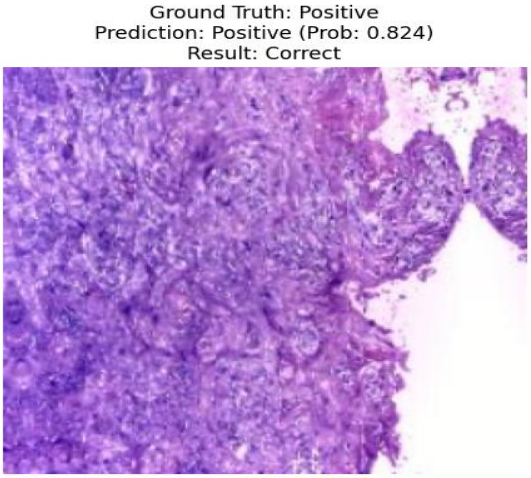
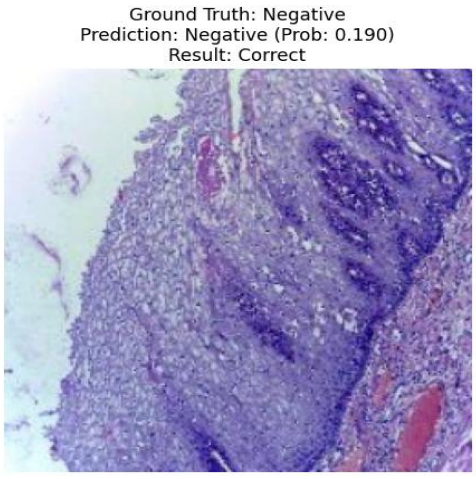


Figure 8: Correctly classified images by EfficientNetV2-Medium: (a) True Positive (OSCC), (b) True Negative (Normal)



(a) True Positive (OSCC)



(b) True Negative (Normal)

5. COMPARATIVE ANALYSIS

The results are placed in the context of the current study by comparing the framework against listed recent studies on OSCC detection. Selected works that employ a variety of methods from custom CNNs optimized with metaheuristic algorithms to transfer learning with both modern and legacy systems which allows to compare the performance against diverse approaches.

[18] There was no use of focal loss. Also, the dataset has almost balanced dataset therefore no class imbalance handling and advanced augmentation technique. [19] there was no custom attention mechanisms and no modern augmentation techniques. [20] did not depict transfer learning improvements beyond the pretrained. [20] compared 4 CNN architectures with Cyclic Learning Rate yet the best performing model but had no unique augmentation method and no ensemble-based learning rate. [21] employed CNNs for feature extraction but the custom CNN classifier had limited feature learning. The dataset used was small achieved the accuracy of 90%. [22] compared 17 pretrained models combined with Dual Attention network. The use of MCC and CK score added novelty to the research. However, use ReLu activation function and no latest optimization technique.

[23] had no comparison with modern CNN architecture of EfficientNet. There was no method equivalent to Mixup augmentation was used to deal with requirement of diversified data. There was no 5-fold cross validation done and achieved 98.73%. [24] study emphasized more on optimization rather than architecture and had a very high computation cost. It focused on multi cancer classification and handled data imbalance with the help of SdMOTE. Moreover, had a small dataset yet achieved a competitive accuracy score of 98.70%. [25] No unique activation function was used such as presented by the study. Several conventional ML models with DL were combined with RGB colour normalization and Region Of Interest. [26] There was no augmentation done on the dataset as it had perfectly balanced case scenarios. For pre-processing CLAHE method was used to improve image contrast but no Mixup augmentation was used, and no novel optimizer was used like AdamW. The achieved accuracy was 99% which may be caused due to overfitting. [27] the research undertook 4 classes rather than binary classification with normal image augmentation and experimented with Hyper parameter tuning with 3 optimizers and the learning rates and presented that transfer learning showed underfitting. [28] focused more on the concept of over sampling and under sampling and oversampling techniques and their accuracy. [29] there was no attention mechanism and class imbalance and feature engineering was done and achieved a competitive score.

Table II provides a summary of the comparison between the referred works and the proposed methodology. The evaluation of the references included the models used, the core methodology and the final accuracy.

TABLE II: Comparative Analysis of Proposed referenced works

Author(s) & Reference	Model Used	Methodology	Accuracy
[21]	ResNet-50, VGG16, InceptionV3 + Tuned CNN.	Used pre-trained models as feature extractors and fed the features into a separate, tuned CNN for classification.	90.00%
[22]	EfficientNet + DAN	17-pretreined architectures +2 stage analysis	91.1%
[20]	DenseNet-169	4 CNN architectures+ Cyclic Learning rate	95.0%
[19]	ResNet-50	Pretrained + limited dataset	96.6%
[25]	DenseNet201	Laplacian and	97.0%

	+ GLCM + HOG + LBP + SVM	Gaussian filter+5 transfer learning model+ RGB color normalization+ ROI enhancement	
[27]	Custom CNN architecture	Batch Normalization, + ReLU + Max Pooling + Dropout	97.5%
[29]	EfficientNetB3.	Applied a single, modern pre-trained deep learning model (EfficientNetB3) to classify histopathological images.	98.33%
[24]	Custom CNN optimized with ITSA.	Developed a custom CNN and optimized its performance using an Improved Tunicate Swarm Algorithm (ITSA), a metaheuristic approach.	98.70%
[23]	EfficientNetB3	Image preprocessing+ ImageDataGenerator + ReduceLROnPlateau	98.73%
[26]	AlexNet + XGBoost	bilateral filtering+ color normalization,	99.0%
Proposed Study	EfficientNet, EfficientNetV2, MobileNetV3, ResNet-50.	Systematic Comparison of multiple modern CNNs within a 5-fold cross-validation framework, enhanced with Mixup augmentation and a Focal Loss function.	99.14%

The framework achieved the highest accuracy 99.14% higher than the other frameworks listed in table 2. The outcomes suggest that the entire pipeline and not just the model architecture, is important to get higher accuracy in OSCC identification.

6. CONCLUSION

The work provides a methodology for developing such diagnostic tools, but clinical translation requires further steps. The further studies should explore model interpretability methods to give pathologists with visual explanations for diagnostic predictions to develop clinical trust. The promising performance of models like MobileNetV3 proves that such models can be deployed for low cost, point of care . The above research could contribute in rapid, accurate OSCC screening more accessible in limited resources and helping to improve outcomes through earlier detection.

7. DECLARATIONS

All authors in the research declare that they do not have any conflict of interest

REFERENCES

[1] F. Bray et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834

[2] C. Rivera, "Essentials of oral cancer," International Journal of Clinical and Experimental Pathology, vol. 8, no. 9, pp. 11884–11894, Sep. 2015. [Online]. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4637760

[3] S. Warnakulasuriya, "Global epidemiology of oral and oropharyngeal cancer," Oral Oncol., vol. 45, no. 4, pp. 309–316, 2009, doi: 10.1016/j.oraloncology.2008.06.002. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S1368837508001838

[4] S. Marur and A. A. Forastiere, "Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment," Mayo Clin. Proc., vol. 91, no. 3, pp. 386–396, Mar. 2016, doi: 10.1016/j.mayocp.2015.12.017. [Online]. Available: https://doi.org/10.1016/j.mayocp.2015.12.017

[5] R. Mehrotra, M. Singh, R. K. Gupta, M. Singh, and A. K. Kapoor, "Trends of prevalence and pathological spectrum of head and neck cancers in North India," Indian J. Cancer, vol. 42, no. 2, pp. 89–93, 2005.

[6] D. Swaminathan, N. A. George, S. Thomas, and E. M. Iype, "Factors associated with delay in diagnosis of oral cancers," Cancer Treat. Res. Commun., vol. 40, p. 100831, 2024, doi: 10.1016/j.ctarc.2024.100831. [Online].Available:https://www.sciencedirect.com/science/article/pii/S2468294224000431

[7] J. Massano, F. S. Regateiro, G. Janu´ario, and A. Ferreira, "Oral squamous cell carcinoma: Review of prognostic and predictive factors,"Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontics, vol. 102, no. 1, pp. 67–76, Jul. 2006. [Online].Available:https://doi.org/10.1016/j.tripleo.2005.07.038

[8] Alsarraf and R. Alrumaih, "Oral cancer clinical presentations – An illustrative review," International Journal of Dental Sciences and Research, vol. 12, no. 2, pp. 19–24, 2024. [Online]. Available: https://pubs.sciepub.com/ijdsr/12/2/1

[9] J. J. Pindborg et al., Histological Typing of Cancer and Precancer of the Oral Mucosa, 2nd ed. Berlin, Germany: Springer, 1997, International Histological Classification of Tumours.

[10] Z. Rumboldt, L. Gordon, R. Bonsall, and S. Ackermann, "Imaging in head and neck cancer," Current Treatment Options in Oncology, vol. 7, no. 1, pp.23–34, Feb. 2006. [Online]. Available: https://doi.org/10.1007/s11864-006-0029-2

[11] V. Pereira-Prado, F. Martins-Silveira, E. Sicco, J. Hochmann, M. A. Isiordia-Espinoza, R. G. Gonz´alez, D. Pandiar, and R. Bologna-Molina, "Artificial intelligence for image analysis in oral squamous cell carcinoma: A review," Diagnostics, vol. 13, no. 14, Art. no. 2416, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/14/2416

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015. [Online].Available: https://doi.org/10.1038/nature14539

[13] Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb.2017.[Online].Available: https://doi.org/10.1038/nature21056

[14] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. S´anchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[15] Y. Zhao, J. Zhang, D. Hu, H. Qu, Y. Tian, and X. Cui, "Application of Deep Learning in Histopathology Images of Breast Cancer: A Review,"Micromachines, vol. 13, no. 12, Art. no. 2197, 2022. [Online]. Available:https://www.mdpi.com/2072-666X/13/12/2197

[16] D. Komura, M. Ochi, and S. Ishikawa, "Machine learning methods for histopathological image analysis: Updates in 2024," Computational and Structural Biotechnology Journal, vol. 27, pp. 383–400, Jan. 2025. doi: 10.1016/j.csbj.2024.12.033.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," 2018, arXiv:1710.09412. [Online]. Available: https://arxiv.org/abs/1710.09412

[18] A.-u. Rahman, A. Alqahtani, N. Aldhafferi, M. U. Nasir, M. F. Khan, M. A. Khan, and A. Mosavi, "Histopathologic oral cancer prediction using oral squamous cell carcinoma biopsy empowered with transfer learning," Sensors, vol. 22, no. 10, p. 3833, 2022. doi: 10.3390/s22103833.

[19] S. Panigrahi, B. S. Nanda, R. Bhuyan, K. Kumar, S. Ghosh, and T. Swarnkar, "Classifying histopathological images of oral squamous cell carcinoma using deep transfer learning," Heliyon, vol. 9, no. 2, p. e13444, 2023.

[20] Yadav and S. Yadav, "Enhancing Oral Squamous Cell Carcinoma Detection: A Transfer Learning Perspective on Histopathological Analysis Using ResNet-18, AlexNet, DenseNet-169, and DenseNet-201 with Cyclic Learning Rate," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 17s, pp. 689–699, 2024.

[21] K. Dhanya, D. Venkata Vara Prasad, and L. Y. Venkataramana, "Detection of oral squamous cell carcinoma using pre-trained deep learning models," Experimental Oncology, vol. 46, no. 2, pp. 119–128, 2024.

[22] Soni, P. K. Sethy, A. K. Dewangan, A. Nanthaamornphong, S. K. Behera, and B. Devi, "Enhancing oral squamous cell carcinoma detection: a novel approach using improved EfficientNet architecture," BMC Oral Health, vol. 24, no. 1, p. 601, 2024.

[23] E. Albalawi, A. Thakur, M. T. Ramakrishna, S. B. Khan, S. SankaraNarayanan, B. Almarri, and T. H. Hadi, "Oral squamous cell carcinoma detection using EfficientNet on histopathological images," Frontiers in Medicine, vol. 10, p. 1349336, 2024.

[24] X. Wei, L. Chanjuan, J. Ke, Y. Linyun, G. Jinxing, and W. Quanbing, "Convolutional neural network for oral cancer detection combined with improved tunicate swarm algorithm to detect oral cancer," Scientific Reports, vol. 14, no. 1, p. 28675, 2024.

[25] M. Ahmad, M. A. Irfan, U. Sadique, I. ul Haq, A. Jan, M. I. Khattak, Y. Y. Ghadi, and H. Aljuaid, "Multi-Method Analysis of Histopathological Image for Early Diagnosis of Oral Squamous Cell Carcinoma Using Deep Learning and Hybrid Techniques," Cancers, vol. 15, no. 21, p. 5247, 2023.

[26] N. W. Asnake, A. M. Ayalew, and A. A. Engda, "Detection of oral squamous cell carcinoma cancer using AlexNet on histopathological images," Discover Applied Sciences, vol. 7, no. 1, p. 155, 2025.

[27] N. Das, E. Hussain, and L. B. Mahanta, "Automated classification of cells into multiple classes in epithelial tissue of oral squamous cell carcinoma using transfer learning and convolutional neural network," Neural Networks, vol. 128, pp. 47–60, 2020.

[28] T. Majeed, T. A. Masoodi, M. A. Macha, M. R. Bhat, K. Muzaffar, and A. Assad, "Addressing data imbalance challenges in oral cavity histopathological whole slide images with advanced deep learning techniques," International Journal of System Assurance Engineering and Management, Jul. 2024. [Online]. Available: https://doi.org/10.1007/s13198-024-02440-6

[29] S. K. D. Sharma, "Oral squamous cell detection using deep learning," arXiv preprint arXiv:2408.08939, 2024. [Online]. Available: https://arxiv.org/abs/2408.08939

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[31] Rahman, Tabassum Yesmin (2019), "A histopathological image repository of normal epithelium of Oral Cavity and Oral Squamous Cell Carcinoma", Mendeley Data, V1, doi: 10.17632/ftmp4cvtmb.

[32] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[33] M. Tan and Q. Lse, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, Jun. 2019, pp. 6105–6114.

[34] M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," in Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, Jul. 2021, vol. 139, pp. 10096–10106.

[35] Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Korea, Oct. 2019, pp. 1314–1324.

[36] Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," 2019, arXiv:1711.05101. [Online]. Available: https://arxiv.org/abs/1711.05101

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 2980–2988.