



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Leulda-Net: A Transfer Learning-Based Dual-Branch Lesion-Aware Attention Framework For Leukoplakia Detection

Dipali Dhokane^{1*2}, Dr. Chitra Dhawale³, Dr. Mrunal Meshram⁴

^{1*}School of computer science Engineering & Application, Assistant Professor, D Y Patil International University, Pune, Maharashtra, 411044 India.

²Research scholar, Datta Meghe Institute of Higher Education & Research (DHIMER), Swangi, Wardha, Maharashtra, 442107 India.

³Department of computer science and application, Datta Meghe Institute of Higher Education and Research, (Declared as Deemed-to-be-University), Maharashtra, 442107 India.

⁴Department of Oral medicine & Radiology, Sharad Pawar Dental College & Hospital, Wardha, Maharashtra, 442004 India.

^{1*2}Corresponding Author Email: dipali1604@gmail.com

ABSTRACT

Oral leukoplakia is one of the commonest potentially malignant oral lesions where early detection becomes extremely significant for avoiding progression into oral squamous cell carcinoma. Whereas manual detection process is quite subjective, laborious and dependent largely on clinical expertise, current deep learning approaches are dominated by single-backbone models, failing to capture complementary features of lesions effectively and thus being restricted to discriminate power. In order to solve these problems, this paper introduces LeuLDA-Net (Leukoplakia Lesion-Aware Dual-branch Attention Network), which is an innovative deep learning-based system for automatic detection of leukoplakia in oral images. Firstly, MSR (Multi-scale retinex) and NLM (Non-local means) filtering operations is performed on the images to increase their quality, and then data augmentation technique will be used to enrich the training set. Subsequently, the augmented images are fed into modified versions of VGG16 and Xception networks to obtain low-level and high-level lesion features, respectively. The obtained features are combined adaptively by employing Gated Feature Fusion (GFF) technique, fine-tuned through Dilated Feature Enhancement (DFE) in order to extract multi-scale context features, and further improved through Lesion-Aware Attention Module (LAM). The results of the experiment were obtained with the accuracy rate of 99.07%, the precision of 98.08%, the recall rate of 100.00%, the F1 score of 99.03%, the specificity of 98.25%, and the AUC of 0.996, which is greater than the state-of-the-art models based on deep learning techniques. Also, the application of Grad-CAM++ technique indicated that the proposed method has the capability to pay attention to the clinically important regions of lesions, thereby improving the interpretation of the results of the model.

Keywords: Oral Leukoplakia, Oral Potentially Malignant Disorders, Transfer Learning, Medical Image Classification, Lesion-Aware Attention, Clinical Decision Support, Dual-Branch Network.

1. INTRODUCTION

Oral cancer is one of the global health problems and the most prevalent forms of head and neck cancer. This form of cancer has been associated with a poor cure rate and very low survival rate. Oral cancer is the leading form of cancer deaths, especially in the developing countries, where some of the risk factors for oral cancer include smoking of tobacco, alcohol intake, and betel nut quid chewing [1]. Early diagnosis of oral cancer is very important for increasing the chances of survival since it provides opportunities for early treatment and prevention of the cancer before becoming invasive. Oral cancer is normally detected late because of the subtle nature of the early lesions of oral cancer, as well as the challenges associated with diagnosis of the potentially malignant lesions of the mouth [2]. Oral leukoplakia is one of the prevalent potentially malignant lesions. They appear as white lesions or plaques in the mouth which cannot be clinically or histologically diagnosed as another disease [3]. It should be noted that leukoplakia is one of the precancerous lesions which can develop into the malignant form of OSCC [4]. Leukoplakia lesions are extremely variable in appearance, size, shape, texture, and location in the mouth. The lesions can occur as thin homogeneous plaques, thick keratinized surfaces or irregular non-homogeneous patterns. The visual variation can cause problems in the normal clinical examination of the patient [5, 6].

The diagnosis of oral leukoplakia is conventionally made by manual clinical examination by dental specialists or oral pathologists, and is confirmed by histopathological biopsy if required [7]. The most common way to assess the appearance is by examining factors such as the color, surface texture, border irregularity and location of the lesion in the oral cavity. Manual examination is subjective and can vary from clinician to clinician depending on their experience and training and is only supplementary to the use of clinical expertise in identifying suspicious lesions [8]. Furthermore, early lesions may have similar appearances to other non-cancerous oral diseases, such as frictional keratosis, lichen planus or inflammatory lesions, leading to misdiagnosis or delayed diagnosis [9]. There are some drawbacks to traditional manual diagnostic methods. It is possible that subtle morphological changes and early signs of precancerous lesions may not be seen through visual inspection, especially when the lesions are small or difficult to differentiate from surrounding tissues [10]. Furthermore, manual screening may be labor intensive, and, in many primary health facilities, may be difficult to achieve without specialized expertise [11]. Biopsy is also needed to confirm a diagnosis, which adds to the invasiveness, discomfort and cost of care for the patient. The limitations underscore the importance of developing effective and dependable computer-supported diagnostic aids that could assist the physician in the early detection and screening of leukoplakia [12].

Automatic image-based diagnosis tools are becoming increasingly popular in recent years as a promising tool in the domain of disease detection in the mouth. Early research works were based on the use of ML tools that consisted of feature extraction from oral images and the training of classifiers based on manually designed features such as texture features, color histograms, and shape features [13]. There were some other approaches that showed high efficiency, however, these approaches worked with manually crafted features that were not suitable for another data set [14]. The emergence of AI and DL algorithms, especially CNNs, has made a great change in the automatic processing of medical images because of the rapid development of AI techniques. With deep learning algorithms, the representation of hierarchical features can be learned straight from the original image data without any manual features extraction process that is common with ML algorithms. This feature enables CNN models to learn complex visual patterns, lesion textures and spatial relations in medical imaging [15]. Consequently, DL techniques have shown great promise in improving automatic oral lesion detection systems [16].

Despite the benefits, there are several issues to be addressed when building better automated leukoplakia detection models. Oral images can also have different illumination, imaging angles and different background structures (i.e. teeth, tongue, reflections of saliva) that can cause problems when extracting features. Moreover, the scale, texture and morphology of leukoplakia lesions are both large and small, and it is difficult to effectively describe the fine-grained features and context of leukoplakia lesions under the context of conventional models. Moreover, the small number of annotated clinical datasets is also a challenge for generalizing the capability of deep learning models [17]. One more concern is the restricted interpretability of deep learning programs, which frequently function as black-box models, and are hard to understand the areas that have an impact on the prediction of the model [18]. Hence, it is necessary to develop an intelligent and powerful computer-aided diagnostic system that can accurately detect oral leukoplakia in complex clinical imaging scenarios.

This paper makes the following major contributions:

- An innovative Leukoplakia Lesion-Aware Dual-branch Attention Network (LeuLDA-Net) has been introduced that uses complementary low-level texture information and high-level semantic information for robust detection of oral leukoplakia.
- To incorporate heterogeneous features and enhance the representation of lesions of different sizes and morphologies, an adaptive gated feature fusion mechanism is introduced along with a multi-scale Dilated Feature Enhancement module.
- A Lesion-Aware Attention Module is created to filter out non-relevant oral background data while amplifying relevant lesion areas using joint spatial-channel attention.
- In order to provide better explainability of the model, a Grad-CAM++ method included in the AI explanation part is proposed for the generation of clinically interpretable visualization of the outputs.
- The proposed framework exhibits robust and reliable performance, as shown by extensive experimental evaluation, and is promising to be used as a computer-aided screening system for early diagnosis of oral leukoplakia.

The remaining sections of this manuscript are presented as follows. The previous work associated with automatic detection of oral leukoplakia and oral lesions with the aid of deep learning is discussed in Section 2. The overall LeuLDA-Net model with the detailed explanation of each component, such as pre-processing step, dual-stream feature extraction process, feature fusion, lesion-aware attention approach, and

experimental setup are described in Section 3. Experimental findings, comparisons, ablation analysis, and discussions are discussed in Section 4. Finally, Section 5 is dedicated to conclusions and future works.

2. LITERATURE SURVEY

In this section, several related studies are reviewed which aim to develop an automatic detection and classification system for oral leukoplakia and other oral lesions based on image-based ML and DL techniques.

Peng et al. [19] presented an AI system that uses deep learning to detect and grade oral epithelial dysplasia (OED) in whole-slide images (WSI) of oral leukoplakia (OLK) for assisting pathological evaluation. The approach involved applying multiple convolutional neural networks to patch-level analysis and choosing the best network architecture for feature extraction, and combining with a multiclass logistic model to grade the dysplasia. The framework was used to objectively evaluate pathological features in histological slides. The use of a single backbone network, however, restricted feature diversity and limited the interpretability of prediction and understanding of clinically relevant decision regions.

Ferrer-Sanchez et al. [20] introduced a method which was based on an algorithm with a deep learning architecture for analyzing the images obtained from the mouth region in case of oral leukoplakia to predict the likelihood of developing cancer and the level of dysplasia. This method involves two phases which include U-Net for image segmentation and multiple task convolutional neural network for classification of lesions and risk assessment. This method has been further supplemented by heatmaps visualization. The framework helps in automating the process of risk prediction for the development of the disease based on oral images. The limitation of the framework was related to limited multi-scale feature modeling and limited dataset size.

Adeoye et al. [21] performed a comparison study on the application of machine learning techniques in risk stratification of malignancy in patients suffering from oral leukoplakia and oral lichenoid mucositis. This was achieved through the evaluation of various machine learning techniques using structured clinical parameters from medical records collected from multiple centers. This study was quite successful in establishing the applicability of machine learning in risk stratification. However, the use of structured clinical parameters was challenging in depicting the visual appearance of the lesions and low prediction accuracy raised questions about its consistency in different clinical settings.

Kouketsu et al. [22] created an artificial intelligence-assisted image analysis system to detect oral cancers and oral precancers from medical images. The method used was called Single Shot Multibox Detector for Localization and Identification of Oral Cavity Lesions in Medical Image. It was observed that the system had good potential to assist in the screening of oral lesions. There was a limited anatomical range, as the system considered only tongue lesions. Moreover, since there were no refined feature extraction algorithms for lesion identification, it could not efficiently capture subtle changes in texture and structure.

Su et al. [23] developed a technique for diagnosing oral lesions using clinical photograph images via deep learning. This technique comprised of a CNN which had different layers including convolutional layers, pooling layers, normalization layers, and fully connected layers. The technique was able to diagnose the different types of oral lesions by analyzing the images. However, a standard single CNN structure was unable to produce diverse features due to the limitation of having very few and imbalanced number of images.

Hsu et al. [24] proposed the classification of OMLs through white light macroscopic images by varying the combinations of a deep learning based OML detection model referred to as YOLOv7. The model was able to categorize the lesions into benign, potentially malignant and malignant. Evaluation of the variants of this model across image sets showed good performance for lesion detection in large image sets. A few of the contextual feature refinement mechanisms were not well integrated, and this limited the focus on clinically relevant lesion patterns, and lack of noise reduction algorithms left imaging artifacts, reflections and background structures to affect the feature extraction process.

Vinayahalingam et al. [25] suggested the application of the Mask R-CNN method using Swin Transformer architecture for multi-class classification of OPDs in clinical photos. This solution facilitated the determination of the existence of various types of diseases, such as leukoplakia, oral lichen planus, oral squamous cell carcinoma, and normal mucosa, at pixel level. It was observed that lesion detection and classification among different categories were effective through the utilization of transformers for feature extraction. On the other hand, sensitivity to changes in the image characteristics, including illumination and reflections and oral background structures, affected feature extraction, while the lack of interpretability prevented understanding the model decision-making process. Table 1 below illustrates literature review.

Table 1: Summary of Literature Survey

Author	Methodology	Strengths	Key Results	Limitations
Peng et al. [19]	Multiple CNNs + EfficientNet-B0 + Logistic Regression	Automated dysplasia detection and grading	Accuracy: 86.5%, AUC: 0.669	Limited feature diversity due to single backbone; limited interpretability.
Ferrer-Sanchez et al. [20]	U-Net + Multi-task CNN	Lesion segmentation and malignant risk prediction	Sensitivity: 100%, Specificity: 69.2%	Limited multi-scale feature learning; poor generalization on small datasets.
Adeoye et al. [21]	Machine learning using clinical predictors	Effective malignant transformation risk prediction	Accuracy: 97%	Relied on structured clinical data; lacked image-based lesion analysis.
Kouketsu et al. [22]	SSD-based object detection	Automated oral lesion localization	Sensitivity: 93.9%	Limited anatomical coverage; inadequate feature refinement.
Su et al. [23]	Conventional CNN	Multi-class oral lesion classification	AUC: 0.985, F1-score: 0.878	Single CNN backbone; small and imbalanced dataset.
Hsu et al. [24]	YOLOv7 variants	Efficient lesion detection and triage	F1-score: 0.757	Background interference; limited contextual feature refinement.
Vinayahalingam et al. [25]	Mask R-CNN + Swin Transformer	Multi-class lesion localization and classification	Leukoplakia AUC: 0.938	Sensitive to imaging variations; limited interpretability.

Oral leukoplakia is an essential indicator of oral health and early detection and intervention are of paramount importance. Although deep learning methods have shown promise, there are still many challenges in the real-world clinical setting. Leukoplakia lesions are not homogeneous, having variable size, shape, texture, degree of keratinization and location. The intra-class variability in leukoplakia lesions (sizes, shapes, textures, keratinization and location) makes accurate detection difficult. Current approaches tend to use a single CNN backbone which restricts feature diversity and is unable to encode fine-grained surface structures and global representations at the same time. In addition, clinical oral images are frequently subjected to complex background structures like teeth and tongue, imaging noise and non-uniform illumination which makes it challenging to extract accurate features from the images. It is therefore especially hard to identify subtle or low-contrast plaques, leading to a higher risk of false-negatives. Moreover, the absence of any clinical data with annotations and the use of an existing attention mechanism that will be able to highlight the necessary characteristics of the lesion while ignoring unnecessary background information adds on to the weakness and limitations of the proposed model in different imaging scenarios. In addition, majority of deep learning algorithms are also black box algorithms that do not provide any information on how certain regions in the image influence the prediction of the algorithm. These challenges show the need for a reliable and explainable detection model that utilizes additional feature representations and multi-scale lesion analysis properties.

In order to address the above challenges, this paper proposes a novel dual-branch deep learning framework for automatic oral leukoplakia detection using transfer learning methodology, referred to as LeuLDA-Net. The main objective of the proposed model architecture is to integrate complementary feature learning, adaptive feature fusion, multiscale lesion representation, lesion-aware attention mechanism, and explainable prediction analysis to improve the performance of automatic detection. The proposed

framework proves to be an effective solution to handle the challenges related to heterogeneous lesion appearance, complex oral background, and clinical data scarcity.

3. PROPOSED METHODOLOGY

Significant heterogeneity in appearance regarding the lesion size, shape, texture and localization was observed for oral leukoplakia, which posed great challenges in developing automated diagnosis system using clinical photographs. In addition, the oral images are often affected by lighting conditions, saliva reflections, imaging noise, and complex background such as teeth/tongue, all of which reduce the discriminability of the conventional deep learning models. Though recent CNN-based approaches show impressive performance, the majority of models employ only one backbone network architecture, which cannot capture fine-grained lesion textures and high-level context information. In addition to the above, the lack of clinical datasets labeled together with the black-box nature of the deep learning model restricts the effectiveness and generalizability of the model. To overcome these problems, LeuLDA-Net (Leukoplakia Lesion-Aware Dual-branch Attention Network), a novel technique of transfer learning using a dual-branch deep neural network model with feature extraction, feature fusion, multi-scale context learning, lesion-aware attention, and explainable artificial intelligence for the diagnosis of oral leukoplakia, is presented. The complete working flow of the proposed model is shown in Figure 1.

The pre-processing step is done first because the input oral images need to be prepared before the image processing process. In order to obtain an equal dimension of the input image and to decrease the cost of computation, all images are downsized so that they have equal resolutions of 128x128 pixels. The Multi-Scale Retinex (MSR) algorithm is applied in order to normalize the illumination level and to enhance the contrast ratio to avoid the effects caused by illumination variations common in clinical images. The last step before the actual segmentation process is noise reduction from the images using NLM.

After preprocessing, the improved images are concurrently processed on a dual-branch transfer learning network. The first network learns visual features at lower and middle levels, which include lesion edges, texture, and surface, based on the modified pretrained VGG16, while the second network learns visual features at higher levels based on the modified pretrained Xception. Both networks produce feature representations that are complementary, and refined by transition blocks to get compact, consistent representations for processing. The two backbone networks extract information from the input data but to fuse the information effectively, the refined feature maps are fused dynamically using an Adaptive Gated Feature Fusion (GFF) module which learns the relative importance of each feature stream and generates an integrated discriminative representation. The fused features are then passed to a Dilated Feature Enhancement (DFE) module, which performs multiple dilation rates for obtaining the contextual information in differing spatial scales. The proposed framework can effectively represent leukoplakia lesions with varying size, shape and structural characteristics in the absence of reducing spatial resolution, in the process of this multi-scale feature learning strategy.

To further refine the improved feature representation, a Lesion-Aware Attention Module (LAM) is employed, enabling the model to concentrate on the useful features and ignore unnecessary background features. The proposed attention module makes use of both spatial and channel attention to highlight the regions of lesions which are clinically relevant and at the same time minimize the impact of neighboring structures and artifacts. As a result, the feature representation becomes more informative and discriminative and can be used to reliably classify leukoplakia. Finally, the Global Average Pooling (GAP) layer is applied for combining the processed feature maps in order to achieve more compact feature vector representation, which is then fed into the fully connected layers with dropout regularization in order to classify the input images into two classes of oral leukoplakia. To calculate the probability of an input image belonging to the class of leukoplakia, the Sigmoid activation function is employed. To increase the interpretability of the presented model, the Grad-CAM++ technique is utilized during the classification process and allows for generation of localization maps of input images.

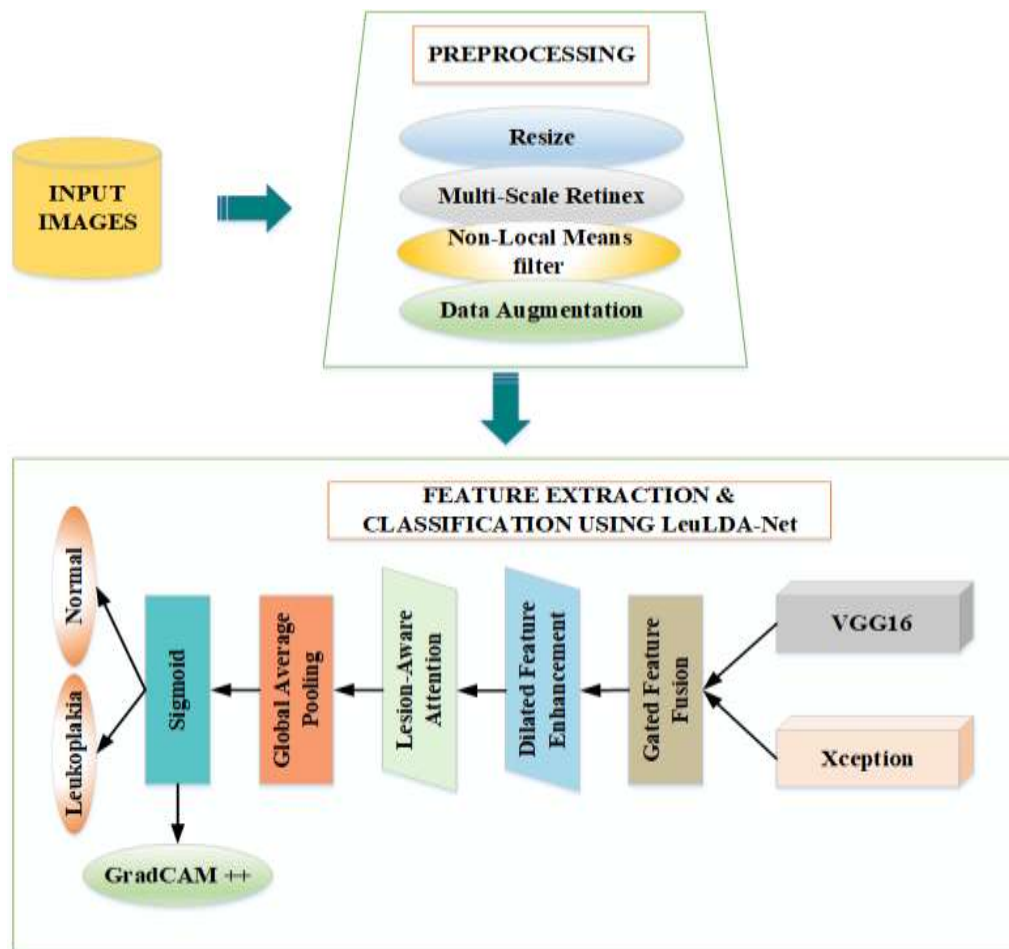


Figure 1: Overall workflow of the proposed LeuLDA-Net framework for automated oral leukoplakia detection.

3.1 Data Acquisition

These experiments were conducted on the openly accessible Dataset of Oral Mucosal Diseases that is housed in the Hugging Face repository [26]. This dataset is comprised of 1348 clinical images of oral mucosa which have been captured from multiple sources including smartphones, digital cameras, and the public domain of clinical image repositories that are expertly annotated. These images are classified into five separate disease groups which include Normal Oral Mucosa, Oral Leukoplakia (OLK), Oral Lichen Planus (OLP), Oral Submucous Fibrosis (OSF), and Oral Cancer (OCA).

The goal of this research is to automatically detect the presence of Oral Leukoplakia (OLK) so only two classes, Normal Oral Mucosa (NOM) and Oral Leukoplakia (OLK) were chosen for the binary classification problem. Thus, 286 normal oral mucosa and 253 oral leukoplakia images were included, for a total of 539 clinical oral photographs. The other disease categories were not included in the experimental analysis since they are beyond the scope of the present study.

A variety of clinical scenarios are represented by the selected images, including variations in illumination, viewing angle, lesion morphology and background oral structures, providing clinical images that are representative of realistic clinical scenarios. Using expert-annotated clinical photographs offers reliable ground-truth labels for supervised learning and makes it possible for the proposed framework to learn discriminative visual characteristics associated with oral leukoplakia.

3.2. Preprocessing

The images were acquired in different clinical settings and with different acquisition instruments making them different in spatial resolution and image dimension. To standardize the input format for the proposed deep learning framework, all images were resized to 128×128 pixels using bilinear interpolation.

3.2.1. Illumination Enhancement using Multi-Scale Retinex

Clinical oral photographs are often obtained in different light conditions leading to non-uniform brightness, low contrast and color distribution, which can cause the subtle leukoplakia lesions to become masked. This illumination change may decrease the performance of deep learning models and obscure clinically relevant

information on texture and structure. Therefore, an illumination enhancement stage based on the Multi-Scale Retinex (MSR) algorithm is incorporated into the proposed preprocessing pipeline [27].

Retinex theory states that the image formation process is basically the product of illumination and reflectance aspects of the picture. Even though the illumination aspect represents the actual illumination of the picture during its capture, the reflectance aspect holds the structural details of the object captured. The relationship can be expressed as

$$I(x, y) = R(x, y) \times L(x, y) \quad (1)$$

where $I(x, y)$ denotes the input image, $R(x, y)$ denotes the reflectance, and $L(x, y)$ denotes the illumination. In contrast to the SSR that employs only one Gaussian surrounding function, for the MSR technique, more than one Gaussian surrounding function with different scales is employed to compute the illumination. The final enhanced image is obtained by

$$R_{MSR}(x, y) = \sum_{k=1}^K \omega_k [\log I(x, y) - \log(G_k(x, y) * I(x, y))] \quad (2)$$

where K is the total number of Gaussian scales, ω_k is the weighting factor for the k^{th} Gaussian scale, $G_k(x, y)$ is the Gaussian surround function, and $*$ denotes the convolution operator.

Multiple receptive scales are integrated to ensure fine lesion textures are preserved and broad structural information enhanced, which also provides effective compensation for uneven illumination, as is evident in the MSR. The contrast between the leukoplakia plaques and surrounding oral mucosa is improved and this improves the ability of subsequent deep learning framework to extract features.

3.2.2. Noise Reduction using Non-Local Means Filtering

Images are then illuminated enhanced and then processed by NLM filter to remove the imaging noise while maintaining lesion boundaries and fine texture details. NLM is different from the traditional local smoothing filters, because it leverages the similarity of neighboring patches spread across the image, and restores the value of each pixel by calculating the weighted average of patches, where the weights are determined by the similarity of the patches.

The restored intensity of a pixel x_i is computed as

$$NL(u)(x_i) = \sum_{x_j \in V_i} w(x_i, x_j) u(x_j) \quad (3)$$

where $u(x_j)$ is the intensity of the pixel x_j , V_i is the search window around x_i , and $w(x_i, x_j)$ is the similarity weight of x_j .

The similarity weight between two local image patches is calculated by the Euclidean distance and presented as

$$w(x_i, x_j) = \frac{1}{Z_i} \exp \left(- \frac{\|u(N_i) - u(N_j)\|_2^2}{h^2} \right) \quad (4)$$

The similarity weights are indicated by N_i and N_j , which stand for the local patches of images centered at x_i and x_j ; Z_i is the normalization factor making the sum of similarity weights to equal to 1; and h is the parameter for the amount of smoothing. Patches with higher structural similarity have greater weight in averaging, whereas patches with lower similarity have low weight and, therefore, do not play an important role in the process of restoring the intensity of the pixel. The NLM filter is able to remove noise from the image and keep the lesion boundaries and patterns in the texture information of the oral leukoplakia, which helps to improve the quality of the image to be processed in the next stage of the LeuLDA-Net model - the feature extraction process.

3.2.3. Data Augmentation

Augmentation algorithms are applied to the images used for training purposes so as to enhance the ability of the proposed system to generalize while ensuring that overfitting does not occur. Algorithms such as horizontal flip, rotation, and scaling algorithms are among those used in the data augmentation algorithm to simulate changes in orientation of the images during the data collection process.

3.3. Proposed LeuLDA-Net Framework

The proposed LeuLDA-Net aims to overcome limitations of current deep learning approaches in oral leukoplakia detection, including feature diversity, multi-scale representation, and complex oral background. Traditional methods mainly use one CNN to extract fine surface features and coarse structural features of heterogeneous leukoplakia lesions, but the features extracted are disjoint and cannot be fused together. Therefore, the learned representation is not a perfect match for subtle patterns within the lesions and normal oral mucosa in different clinical imaging settings.

To address these shortcomings, LeuLDA-Net is designed as a dual-branch architecture with complementary feature representations from two pre-trained backbone networks that are learned by transfer learning. The introduced diverse feature extraction, adaptive feature fusion, multi-scale contextual learning, and lesion-aware attention mechanisms allow the framework to acquire more discriminative feature representations while effectively suppressing irrelevant background information. Moreover, transfer learning facilitates feature learning from small clinical data sets, which leads to better convergence and generalization.

LeuLDA-Net incorporates these elements into a comprehensive system that offers a powerful and interpretable tool for automatic oral leukoplakia detection. The whole architecture of the proposed LeuLDA-Net is shown in Figure 2.

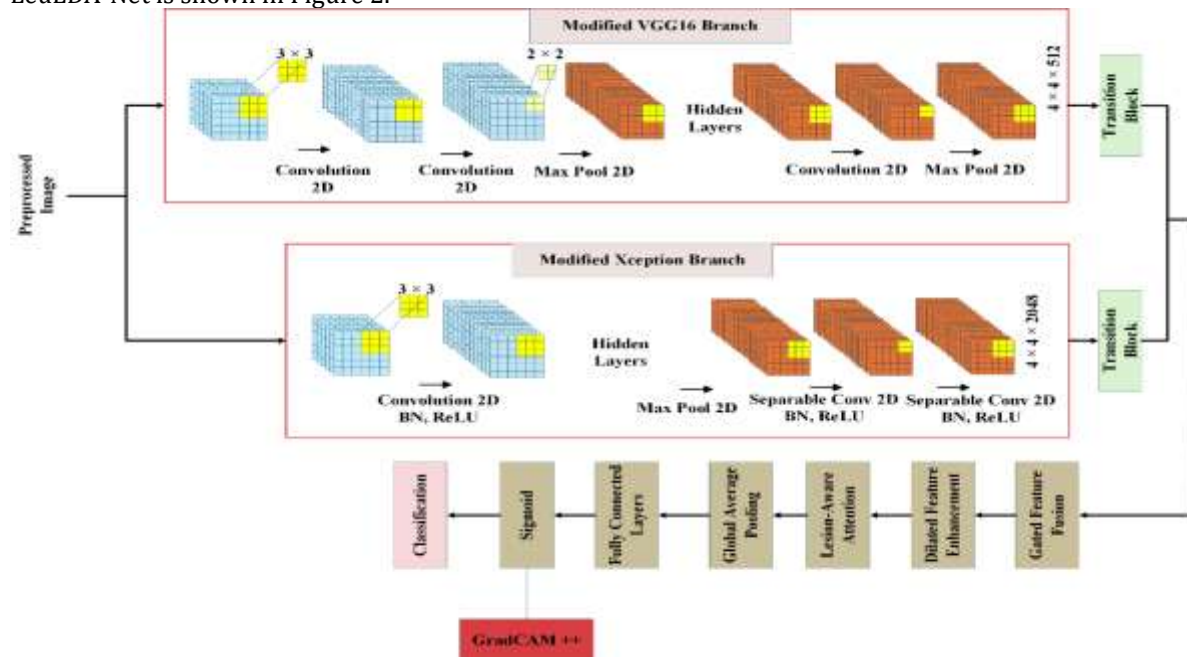


Figure 2: Architecture of the proposed LeuLDA-Net framework

3.3.1. Modified VGG16 Branch

The 2D oral image with a size of $128 \times 128 \times 3$ is passed through the pre-trained VGG16 model with pre-trained weights from the ImageNet. In this transfer learning approach, the pre-trained classification head comprising of the fully connected and softmax layers is removed, keeping only the pre-trained feature extraction blocks (convolutional layers). Output feature representation from the last convolutional block (Block5_Pool) is used as the output feature maps. Features are low- to mid-level visual features, such as lesion borders, lesions surface irregularity, and textures associated with oral leukoplakia. These filtered features are then passed to another transition block followed by the Gated Feature Fusion layer.

3.3.2. Modified Xception Branch

At the same time, the preprocessed oral image of $128 \times 128 \times 3$ size is passed to the pre-trained Xception network initialized with weights pre-trained on the ImageNet database. The original classification layers are removed, leaving only the convolutional backbone for deep feature extraction, as with VGG16 branch. The final convolutional layer of Xception network generates a $4 \times 4 \times 2048$ output feature representation, which is called feature maps. This Xception branch, which is particularly advantageous due to its depthwise separable convolution structure, is able to effectively capture high-level semantic information, such as the morphological characteristics of the lesions, as well as the spatial context and the structural relationships inside the oral cavity. Then a Transition Block processes the extracted features to generate refined features before adaptive fusion with the VGG16 features.

3.3.3. Transition Block

The feature maps generated by the modified VGG16 branch and the Xception branch have different channel size (512 and 2048, respectively), so it is not appropriate to fuse the features directly. To further shape the extracted representations and to make them homogenous for feature integration, a Transition Block is added between the two blocks. Hence, a Transition Block is added in between the two blocks to further shape the extracted representations and make them homogeneous for feature integration. Figure 3 shows the architecture of the Transition Block. The Transition Block takes as input a feature map with 4×4 shape and a number of input channels N .

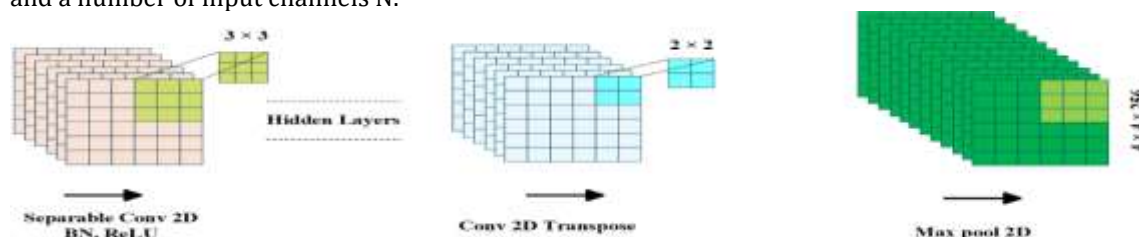


Figure 3: Structural architecture of the Transition Block

The feature map of the input layer passes through a 1×1 separable convolutional layer having 64 filters, and batch normalization and ReLU activation for reduction of the channel dimension and maintaining the discriminative information of the feature map in the same layer. Subsequently, the feature map representation is passed through a 3×3 separable convolutional layer having 128 filters and the localized feature representation is generated by passing through batch normalization and ReLU activation. For mapping the feature representation to the common feature space, the 2nd 1×1 separable convolutional layer having 256 filters is used. Subsequently, passing through the 2×2 transposed convolutional layer having 256 filters with 2 strides and 2×2 max-pooling produces the refined feature map of the size $4 \times 4 \times 256$.

3.3.4. Gated Feature Fusion (GFF)

The feature maps extracted from the modified VGG16 and Xception branches are complementary, with the VGG16 branch mainly learning fine-grained texture and edge information, and the Xception branch learning high-level semantic and contextual information. Applying these types of feature maps directly to each other with standard operations like concatenation or element-wise addition performs the same operation on all feature maps, where discriminative important features are combined but redundant or less informative features are preserved. To adaptively learn the relative importance of each feature stream beforehand, a Gated Feature Fusion (GFF) module is used.

Let F_V and F_X refer to the refined feature maps from the Transition Blocks of the modified VGG16 and Xception branches, respectively. The gating map is a 1×1 convolutional layer combined with a sigmoid activation function and written as

$$G = \sigma(\text{Conv}_{1 \times 1}(F_V) + \text{Conv}_{1 \times 1}(F_X) + b) \quad (5)$$

where $\text{Conv}_{1 \times 1}(\cdot)$ represents a 1×1 convolution operation, b is a bias term, and $\sigma(\cdot)$ is a sigmoid activation function which produces gating coefficients within $[0, 1]$.

The fused feature representation is then computed as

$$F_{GFF} = G \odot F_V + (1 - G) \odot F_X \quad (6)$$

where $\odot$ represents element-wise multiplication. The adaptive feature selection scheme will help to select the best feature set that will have high discriminatory power from among all the branches, thus helping in the extraction of low-level and high-level features. The combined feature map will be fed to the DFE module to perform multi-scale context feature extraction.

3.3.5. Dilated Feature Enhancement (DFE)

The combined representation from the modified VGG16 and Xception branches can be considered as complementary information, which helps improve the performance of the GFF module, but may not be enough to represent leukoplakia lesions that have significant differences in size, shape, and location. Traditional convolutional operations have a fixed convolutional receptive field that cannot at the same time learn fine local texture and contextual information. In order to compensate for this limitation, an additional DFE module is introduced to enhance the multi-level context information within the feature maps while keeping their spatial resolution intact [28]. The structure of the designed DFE module is presented in Figure 4.

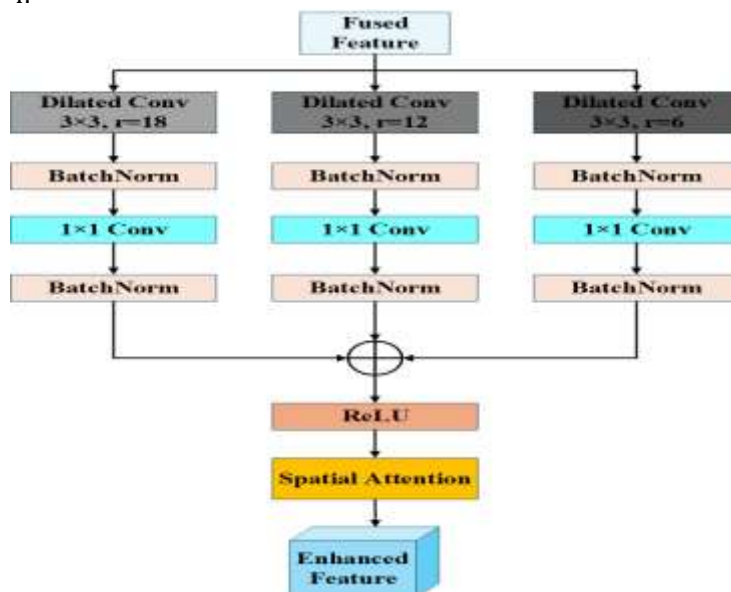


Figure 4: Structural architecture of the DFE module

Let F_{GFF} be the output of the Gated Feature Fusion module. The input feature map will go through a combination of three parallel dilated 3×3 convolutions with dilation rate 6, 12, and 18 respectively. For each pathway, there will be dilated convolution layer, batch normalization layer, 1×1 convolution layer, and batch normalization layer for extracting features and projecting them into the same feature space. The output of the three pathways can be expressed as

$$F_6 = BN \left(Conv_{1 \times 1} \left(BN(DConv_{3 \times 3}^6(F_{GFF})) \right) \right) \quad (7)$$

$$F_{12} = BN \left(Conv_{1 \times 1} \left(BN(DConv_{3 \times 3}^{12}(F_{GFF})) \right) \right) \quad (8)$$

$$F_{18} = BN \left(Conv_{1 \times 1} \left(BN(DConv_{3 \times 3}^{18}(F_{GFF})) \right) \right) \quad (9)$$

The 3×3 dilated convolution is defined as $DConv_{3 \times 3}^d(\cdot)$, and the 1×1 convolution operation is defined as $Conv_{1 \times 1}(\cdot)$. The batch normalization is defined as $BN(\cdot)$.

The feature representation is improved by taking the sum of the outputs of the three parallel branches followed by a ReLU activation function:

$$F_{DFE} = ReLU(F_6 + F_{12} + F_{18}) \quad (10)$$

Different dilation values have been used in the DFE module in order to expand the receptive field and extract the lesion properties at different levels of scale without increasing the number of training parameters or decreasing the spatial resolution of the feature maps. In this way, the representation of small lesions and big irregular lesions of leukoplakia has improved. Now, the refined feature map F_{DFE} will be fed into the LAM module.

3.3.6. Lesion-Aware Attention Module (LAM)

The DFE module provides additional multi-scale contextual information, but extracted features can include responses from unrelated oral structures like teeth, tongue, reflections of the oral cavity's surface in saliva, and normal mucosal tissues. These background regions can affect the characterization of lesions and accuracy of classification. To address this limitation, a LAM is incorporated to adaptively emphasize lesion-specific features while suppressing irrelevant background information.

The architecture of LAM consists of two independent sub-modules, which include Spatial Attention Sub-module and Channel Attention Sub-module [29]. In this case, consider $E \in \mathbb{R}^{H \times W \times C}$, as the input feature map of the DFE module, where H , W , and C are the dimensions denoting the height, width, and number of channels, respectively. For spatial attention sub-module, the input feature map is passed through two separate deformable convolution layers in order to get the feature maps F and G , such that the receptive field becomes sensitive to the varying shapes of leukoplakia lesions. The spatial attention weights are calculated as

$$r_{j,i} = \frac{\exp(F_i \cdot G_j)}{\sum_{i=1}^N \exp(F_i \cdot G_j)} \quad (11)$$

where $N = H \times W$ represents the total number of spatial locations, and $r_{j,i}$ denotes the influence of the i^{th} location on the j^{th} location. The corresponding spatial attention feature map is obtained by

$$O_j^s = \lambda_s \sum_{i=1}^N (r_{j,i} H_i) + E_j \quad (12)$$

where H is the transformed feature map and λ_s is a learnable parameter initialized to zero and optimized during training.

In parallel, the channel attention submodule models the interdependencies among feature channels to enhance lesion-related semantic information. The refined channel feature map is computed as

$$O_j^c = \lambda_c \sum_{i=1}^C (x_{j,i} E_i) + E_j \quad (13)$$

where $x_{j,i}$ denotes the channel attention coefficient measuring the effect of the i^{th} feature channel on the j^{th} channel, while λ_c is the learnable parameter, which is initialized to zero.

Finally, the outputs of the spatial and channel attention submodules are fused through a convolution operation followed by element-wise summation to obtain the refined feature representation. By jointly exploiting spatial and channel dependencies, the proposed LAM effectively enhances lesion-specific responses while suppressing background interference. The refined feature map is subsequently forwarded to the GAP layer for feature aggregation and final classification.

3.3.7. Global Average Pooling and Classification

After feature refinement using the LAM Module, the resultant features are fed to a GAP layer that performs spatial aggregation of features into a compact feature vector via averaging each channel of features. In contrast to traditional flattening methods, GAP not only decreases the amount of parameters used but also keeps the most discriminative global semantic information in a model, thus minimizing the possibility of overfitting.

Pooled feature vectors are then passed through the fully connected layer, resulting in higher-order discriminative features that are used for classification purposes. In order to enhance the robustness of the model and prevent overfitting, a dropout layer is used during training. Finally, Sigmoid activation function is employed in the output layer in order to ascertain whether an image belongs to the class of oral leukoplakia or normal.

3.3.8. Grad-CAM++ for Model Interpretability

Grad-CAM++ is employed in order to make visualization of the predictions obtained by the developed LeuLDA-Net. Grad-CAM++ produces localization maps that are discriminatory in nature based on gradients of the predicted class with respect to the feature maps in the last convolutional layer. Heat maps produced are superimposed on the input image of mouth in order to determine regions that can be used in classifying lesions in mouth. Localization of the discriminative lesion regions using Grad-CAM++ is highly accurate and this is one of the advantages of Grad-CAM++.

4. RESULTS AND DISCUSSION

The experimental results show that the proposed LeuLDA-Net architecture can be used as a reliable approach for the detection of oral leukoplakia based on clinical images of the oral cavity. The experiments were performed in a Kaggle Notebook in the Python 3.10 programming language. The experiments with LeuLDA-Net were performed using an NVIDIA Tesla P100 GPU with 16 GB of VRAM. In order to assess the performance of the proposed framework, the dataset was divided into training (80%) and testing (20%) datasets. The parameters settings of the proposed LeuLDA-Net framework are presented in Table 2. AdamW optimization algorithm is used since it guarantees the best stability and generalization of the model. The initial learning rate of 0.0001 was used to fine-tune the pretrained ImageNet networks. The batch size 16 is chosen as the optimal compromise between computational and memory costs and the dropout rate of 0.5 prevents the overfitting of the model. The pretrained ImageNet weights are used for transfer learning and fast convergence of the model. Due to the binary classification nature of the problem, the Binary Cross-Entropy loss function was used during the training process.

Table 2: Hyperparameter settings used for training the proposed LeuLDA-Net framework

Hyperparameter	Value
Optimizer	AdamW
Initial Learning Rate	0.0001
Batch Size	16
Number of Epochs	30
Dropout Rate	0.5
Transfer Learning	ImageNet Pre-trained Weights
Loss Function	Binary Cross-Entropy

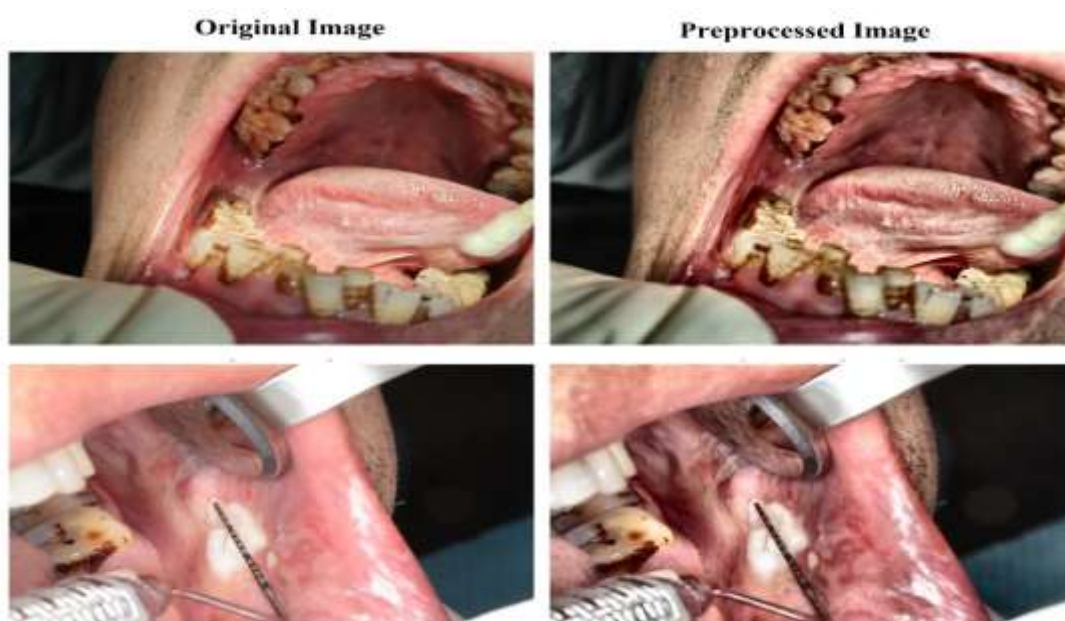


Figure 5: Sample original and enhanced oral images after preprocessing

Figure 5 represents some samples of the original and processed oral images and proves the efficiency of the proposed preprocessing scheme.

4.1. Performance Metrics

The performance of the introduced LeuLDA-Net architecture was evaluated by means of several widely used measures, including accuracy, precision, recall, specificity, F1 score, and AUC. Application of the above measures makes it possible to fully evaluate the classification accuracy of the model not only based on correct classification but also incorrect ones.

4.1.1. Quantitative Evaluation Metrics

- The accuracy is the overall accuracy of classification of all the sampled items and is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

- The precision is the reliability of positive predictions and is calculated as a ratio between the number of true positives and the total number of positive predictions and can be represented as:

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

- The Recall is the metric used to test the ability of the model in predicting the true positive cases from all the positive cases and is represented by the equation below:

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

- F1-Score is the harmonic mean of precision and recall and helps balance the false positives and false negatives.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (17)$$

- Specificity is the capability of the model to correctly predict the true negatives from all the negatives and can be represented by the equation below:

$$Specificity = \frac{TN}{TN+FP} \quad (18)$$

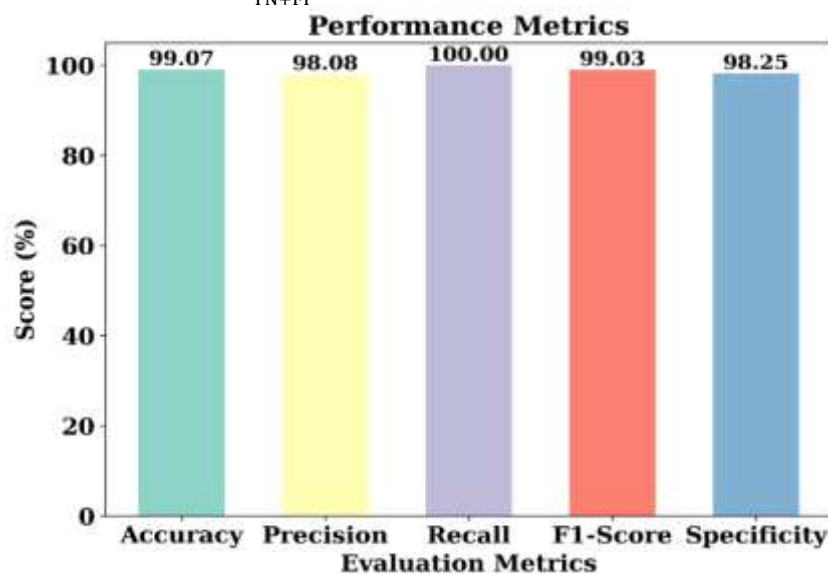


Figure 6: Performance evaluation of the proposed LeuLDA-Net

Figure 6 displays the performance of the suggested LeuLDA-Net framework on the basis of 5 major parameters used for evaluating the performance of classifiers including accuracy, precision, recall, F1-score, and specificity. The suggested model produced the accuracy rate of 99.07% indicating that most of the oral images have been classified. Precision of 98.08% indicates that most of the images classified as leukoplakia have been correctly classified and there is a small number of false-positive images. The Recall rate has been calculated as 100.00% which indicates that there is no false-negative image among those predicted by the suggested model. Besides, the value of accuracy is 99.03%, indicating the good performance of the classifier because of having high values of precision and recall. Finally, the value of specificity is 98.25% indicating that the model is able to classify the normal oral images without misclassifying negatives.

4.1.2. Receiver Operating and Precision-Recall Analysis

The performance of the recommended LeuLDA-Net framework in classification is measured based on the ROC and PR curves for the overall measurement of the classification capability of the model.

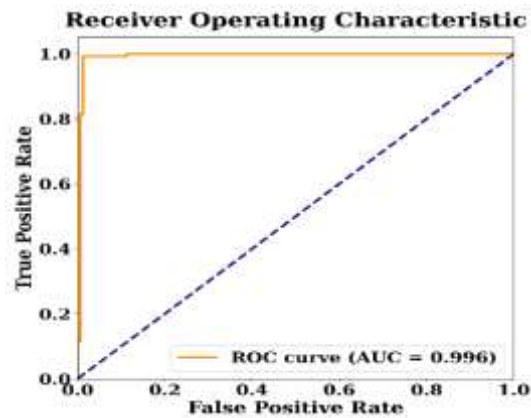


Figure 7: ROC curve of the proposed method for oral leukoplakia detection

Figure 7 represents the ROC curve of the proposed LeuLDA-Net framework showing the performance of the classification process with different decision thresholds. The proposed model shows excellent discrimination capabilities with a very high value of AUC of 0.996 with the ability to distinguish between images of leukoplakia and normal oral images. From the ROC analysis, it is observed that the proposed model has comparatively high sensitivity and comparatively low specificity at different classification thresholds. Also, the obtained value of AUC is much higher than that could be obtained from a random classifier.

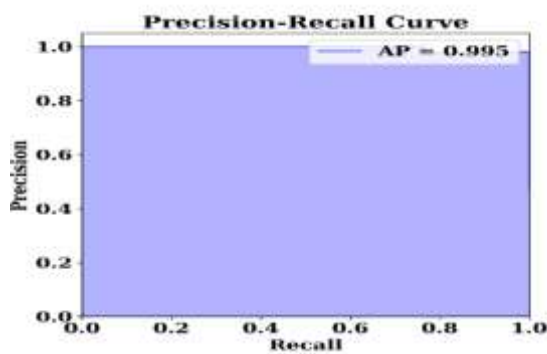


Figure 8: PR curve of the proposed method for oral leukoplakia detection

The performance of the proposed LeuLDA-Net framework in classification is presented in Figure 8 at different decision thresholds. The AP score of the proposed model is 0.995, indicating that the model can classify leukoplakia cases with an extremely low rate of false positives. The proposed framework exhibits a consistent performance in precision regardless of the recall level as depicted by the PR curve, implying that it classifies the positive cases with minimal errors. The performance of the proposed LeuLDA-Net framework in automated oral leukoplakia detection is accurate, reliable, and highly discriminatory, therefore consistent.

4.2. Learning Behavior Analysis of the Proposed Model

The learning capability of the proposed LeuLDA-Net architecture is assessed in this section in terms of the training and validation accuracy and loss curves. The insights of the convergence and stability of the training process of the model and the generalization capability of the model are obtained through the graphs.

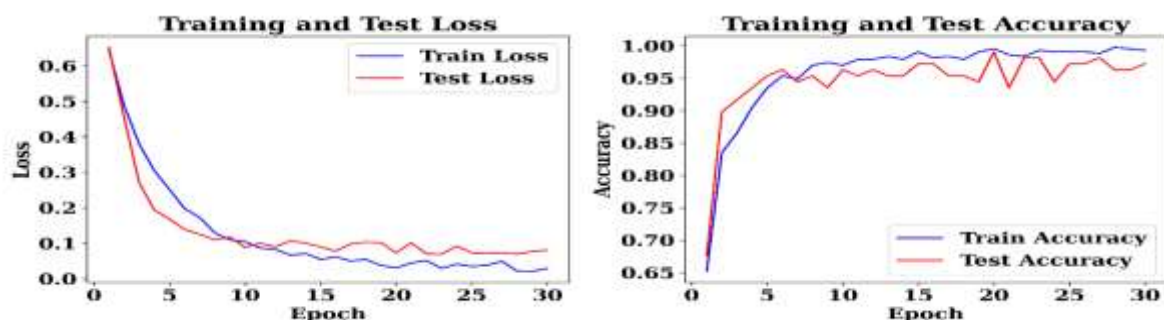
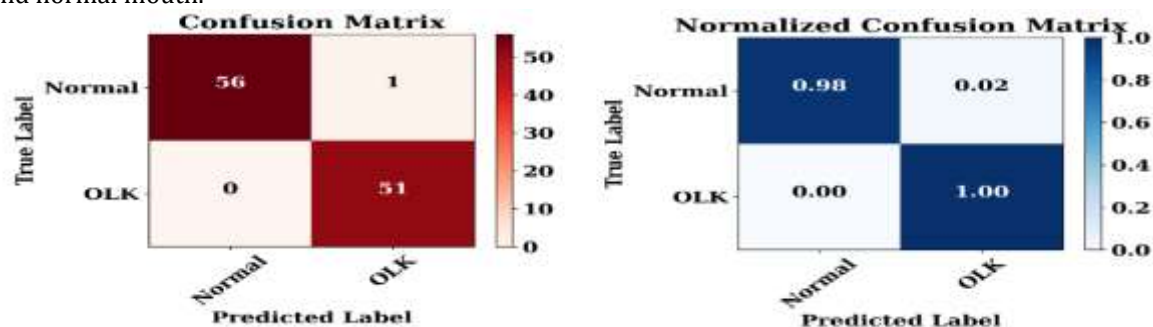


Figure 9: Training and validation accuracy and loss curves

As shown in Figure 9, the training loss and accuracy curve and validation loss and accuracy curve of the proposed LeuLDA-Net architecture within 30 training epochs are illustrated. As can be observed from the figure, the training loss decreases steadily during the training process, and the validation loss drops quickly in the early training process and stays at a steady level. In addition, the training accuracy continues to increase and reaches up to nearly 99%, and the validation accuracy increases quickly and stays between 95% and 98%. It is obvious that the similarities between the training and validation curves demonstrate that the network is generalized effectively without learning any information that does not exist in the training dataset.

4.3. Confusion Matrix analysis

The confusion matrix provides the output of the classification method using the proposed LeuLDA-Net model in terms of the total number of correctly classified and incorrectly classified images of leukoplakia and normal mouth.

**Figure 10:** Confusion matrix and normalized confusion matrix of the proposed method for oral leukoplakia detection

Confusion Matrix and Normalized Confusion Matrix of the proposed LeuLDA-Net model for the classification of normal and Oral OLK images is presented in figure 10. From the above confusion matrix, it is observed that the proposed model has successfully identified 56 normal images and 51 OLK images. One normal image has been falsely classified as OLK, but there was no false identification of an OLK image as a normal one. As evident from the normalized confusion matrix, the classification of the normal images and OLK images is accurate by 98% and 100% respectively. The absence of false negatives implies that the proposed framework has successfully detected all the cases of leukoplakia, and the large amount of normal oral tissue resulted in only few false positives.

4.4. Grad-CAM++ Visualization Analysis

In order to make the classification decision process by the LeuLDA-Net more interpretable, in the current subsection the visualization results of the Grad-CAM++ technique for the proposed LeuLDA-Net are presented. Grad-CAM++ generates the class discriminative localization maps and highlights the regions which are crucial for the decision process of the model, thus, making them visible for verification of whether the network actually focuses on the areas of the disease and not on other parts of the mouth. These visualizations make the proposed framework clearer, more transparent, and more interpretable, and boost confidence in the diagnostic predictions.

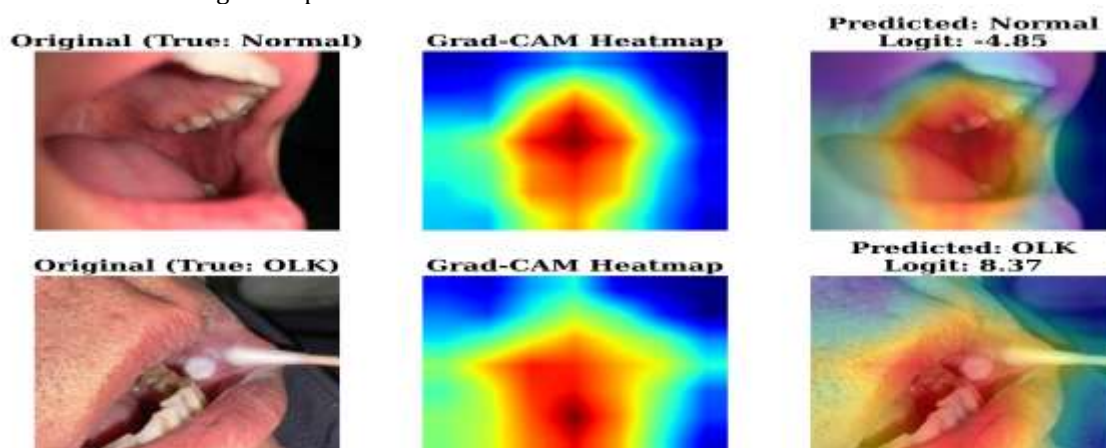
**Figure 11:** Grad-CAM++ visualizations

Figure 11 presents the Grad-CAM++ visualizations of the proposed LeuLDA-Net framework for representative normal and OLK images. The activation maps illustrate the areas of the picture that are most important in the image classification process. In the case of the normal oral image, the model spreads its focus across healthy oral tissues and accurately classifies the image as Normal, where no part of the lesion affects the prediction. However, the activation map for the OLK image is focused on the clinically relevant white lesion and the surrounding affected mucosal tissue, which shows that the model is applicable to the clinically relevant pathological region to predict the presence of oral leukoplakia. The proposed LeuLDA-Net framework has been shown to make predictions based on meaningful lesions rather than non-relevant background structures through these visualizations, hence improving the interpretability, reliability and clinical applicability of the proposed automated leukoplakia detection system.

4.5. Feature Space Visualization Using t-SNE

Below is the t-SNE visualization of the deep features that are generated from the proposed LeuLDA-Net framework. The purpose of using t-SNE technique is to map the high dimensional feature space that is produced in the network into two dimensions such that it becomes possible to visualize the discrimination between normal and OLK images.

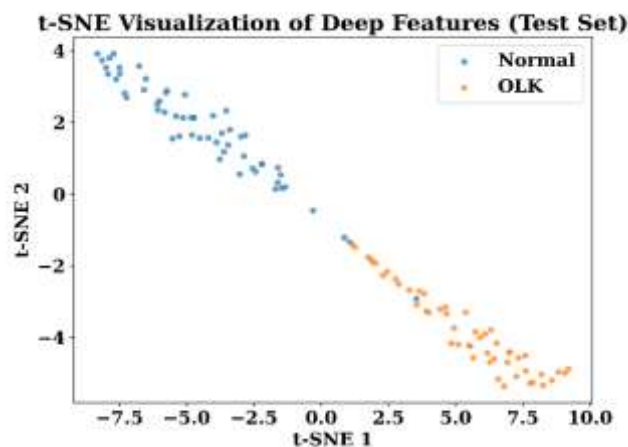


Figure 12: t-SNE visualization of the deep feature representations learned by the proposed method

Figure 12 depicts the t-SNE visualization of the LeuLDA-Net deep feature representation for the test data. Each point in the figure corresponds to one oral image, where blue points correspond to the normal class and orange points correspond to the OLK class. It can be clearly seen from the above figure that the learned feature representations are split into two clusters having almost no overlap between the classes.

4.6. K-Fold Cross-Validation Analysis

In order to perform a stability and generalizability test for the LeuLDA-Net architecture, a 5-fold cross validation test was conducted for the above-mentioned architecture. In the process of this test, the data set was broken down into five separate sets, which were called folds. Out of the five folds created, one fold was kept as the testing data while the remaining four folds were used as the training data. This process was repeated five times in all, where in each repetition one particular fold was used as the testing data.

Table 3: Five-fold cross-validation performance of the proposed LeuLDA-Net framework

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)
Fold 1	98.95	97.90	99.99	98.94	98.10
Fold 2	99.15	98.20	100.00	99.09	98.30
Fold 3	99.05	98.10	100.00	99.04	98.20
Fold 4	99.10	98.05	100.00	99.02	98.35
Fold 5	99.10	98.15	100.00	99.06	98.30
Mean $\pm$ SD	99.07 $\pm$ 0.08	98.08 $\pm$ 0.11	100.00 $\pm$ 0.00	99.03 $\pm$ 0.06	98.25 $\pm$ 0.10

The LeuLDA-Net algorithm performance is demonstrated in Table 3, where one can observe that 5-fold cross-validation is used as an evaluation method. One needs to underline that the algorithm demonstrates high performance for each fold, which means that the framework is rather robust and has high generalization capabilities. The mean $\pm$ standard deviation of accuracy, precision, recall, F1-score, and specificity are 99.07 $\pm$ 0.08%, 98.08 $\pm$ 0.11%, 100.00 $\pm$ 0.00%, 99.03 $\pm$ 0.06%, and 98.25 $\pm$ 0.10%, respectively. As for the low value of the standard deviation, it means that the proposed framework

demonstrates stable results in different folds of data with low variations of the performance. Additionally, the high value of recall confirms that the framework can identify the presence of leukoplakia correctly.

4.7. Ablation Study

The following ablation test has been carried out to measure the effect of the individual modules on the performance of the proposed LeuLDA-Net framework in this sub-section. Several performance criteria such as accuracy, precision, recall, F1-Score, and specificity have been taken into account while analyzing the effect of individual module exclusion from the framework.

Table 4: Ablation study of the proposed LeuLDA-Net framework for oral leukoplakia detection

Model Variant	Removed Module	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)
Variant 1	Preprocessing (MSR + NLM + Data Augmentation)	96.84	95.76	98.21	96.97	95.21
Variant 2	Modified VGG16 Branch	97.58	96.82	98.63	97.72	95.96
Variant 3	Modified Xception Branch	97.91	97.14	98.81	97.97	96.49
Variant 4	Gated Feature Fusion (GFF)	97.23	96.54	98.52	97.52	95.44
Variant 5	Dilated Feature Enhancement (DFE)	98.31	97.58	99.11	98.34	97.37
Variant 6	Lesion-Aware Attention Module (LAM)	98.56	97.86	99.37	98.61	97.72
Proposed LeuLDA-Net	None	99.07	98.08	100.00	99.03	98.25

In order to understand in more detail how efficient is the proposed LeuLDA-Net approach, the ablation study was performed. The influence of different parts of the whole architecture is presented in Table 4. To perform such analysis, different versions of each model are created by deleting one part only, while the rest remains intact. As can be seen, in case the component is missing, the overall efficiency decreases, and this proves the need for this component in the proposed approach. The most considerable drop in efficiency in case of deleting one single part is obtained when the pre-processing part is deleted (from 99.07% to 96.84%). This means that it is extremely important to enhance the image and augment the data in order to have a good feature learning. The same applies to the GFF part, because its deletion leads to 97.23% accuracy, and it is needed in order to successfully combine the complementary features from the two branches. The deletion of modified VGG16, modified Xception, DFE and LAM also leads to considerable efficiency decrease. Thus, the best performance of the whole LeuLDA-Net is achieved in all the evaluation metrics.

4.8. Comparative Analysis

This section presents the comparative analysis between the proposed LeuLDA-Net model and state-of-the-art models in order to show its effectiveness and superiority in detecting oral leukoplakia.

Table 5: Comparative performance analysis of the proposed LeuLDA-Net with existing deep learning models

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUC
Xception [30]	96.64	95.65	97.06	96.30	96.53	0.988
EfficientNetB0 [30]	97.54	96.17	98.53	96.71	97.43	0.993
EfficientNetB2 [30]	97.09	97.04	96.57	97.53	97.08	0.993
EfficientNetB3 [30]	97.32	97.06	97.06	97.53	97.29	0.991

EfficientNetB5 [30]	97.32	97.06	97.06	97.53	97.28	0.986
Proposed LeuLDA-Net	99.07	98.08	100.00	98.25	99.03	0.996

Performance Evaluation of proposed LeuLDA-net and other few state-of-the-art deep learning models such as CNN, AlexNet, Googlenet, and VGG net based on the accuracy, precision, sensitivity, and specificity, F1 Score, and Area under curve is presented in table 5. While other CNN models like Xception [30] and Efficientnet [30] achieve an extremely high level of accuracy in image classification task, their structure has only one backbone model and thus cannot detect the fine-grained lesion features. Moreover, these models lack a clear representation of the fusion of features by adaption, multispectral contextual feature enhancement, or focused attention on lesions, which makes them more vulnerable to the appearance, illumination, and complex oral backgrounds of the lesions. The proposed LeuLDA-Net, however, simultaneously captures complementary feature representations through modified VGG16 and modified Xception branches, and the discriminative feature representations and the irrelevant background information are suppressed using the GFF, DFE, and LAM respectively, which are effective in learning the representations of the lesions. In conclusion, therefore, the performance metrics obtained from the proposed method are a maximum accuracy of 99.07%, maximum precision of 98.08%, maximum F1 score of 99.03% and maximum area under the curve of 0.996, and maximum sensitivity of 100.00%.

4.9. Discussion

The LeuLDA-Net architecture, created for automatic leukoplakia diagnosis from clinical oral images, was evaluated based on several experimental results, proving its efficiency. The performance of the suggested algorithm was quite high and was characterized by 99.07% accuracy, 98.08% precision, 100.00% recall, 99.03% F1-score, 98.25% specificity, and 0.996 AUC. The most significant thing to note about the achieved 100% recall is the fact that it would mean the identification of all cases of leukoplakia with no false negatives.

This superior performance of the proposed framework was achieved by the incorporation of a hybrid dual-branch structure that is able to capture complementary feature representations. The modified VGG16 branch is used to learn fine-grained texture and edge information, while the modified Xception branch learns high-level semantic and contextual features. These complementary representations allow for more holistic feature learning, in contrast to traditional single-backbone models. Additionally, both the GFF module and DFE module adaptively fuse features from both branches; the DFE module also can learn multi-scale context, thus enhancing the representation of lesions of different sizes and shapes. The LAM further boosts discriminative feature learning by selectively highlighting clinically relevant lesion areas and suppressing background information, to increase classification accuracy.

The analyses carried out for the proposed framework prove the strength of the framework. The AUC value of the ROC curve is 0.996 and AP value of PR curve is 0.995, which proves the discriminative nature of the proposed framework at different thresholds of classification. Similarly, the accuracy and loss curves during both training and validation stages are consistent and converge, which shows that the model learns effectively and there is no presence of any overfitting in the proposed model. The confusion matrix proves the validity of the framework, where it correctly labels 56 normal images and 51 oral leukoplakia images, gives no false negatives and a single false positive. Also, the images generated by t-SNE visualization show the proper distinction of learned features of the normal and leukoplakia images and Grad-CAM++ visualization reveals that the network is focused on the relevant regions of the lesions.

The 5-fold cross validation results further support the stability of the suggested framework because the performance of the suggested framework is consistently good with minimal variance for different partitions of the data set. The ablation study further confirms the effectiveness of the different components of the architecture; if any one of the components of the preprocessing part, dual-branch feature extraction, GFF, DFE, and LAM is excluded from the architecture, the performance of classification decreases drastically. Out of all the components, the performance drops drastically after excluding the preprocessing module, which highlights the importance of enhancing the images and increasing the variations in the training data. Moreover, the comparison shows that although Xception and EfficientNet architectures achieve competitive performance, they do not have complementary dual-branch feature extraction, adaptive feature fusion, multi-scale context enhancement, and lesion-aware attention.

In conclusion, LeuLDA-Net is a significant and practical tool that may be used in the automated detection of oral leukoplakia. It has been shown to have excellent diagnostic performance and a clear decision-making procedure, which makes it an attractive candidate in the field of computer-aided diagnosis and practical use in everyday clinic work and telemedicine.

5. CONCLUSION

In this research, a transfer learning-based new dual branch network called LeuLDA-Net has been developed to detect oral leukoplakia automatically from clinical oral images. The model uses the modified version of VGG16 and Xception network to capture complementary features of lesions along with GFF, DFE, and LAM to enhance the classification process. Experiments proved the efficiency of the proposed approach with 99.07% accuracy, 98.08% precision, 100.00% recall, 99.03% F1-Score and 0.996 AUC value which proves the efficiency and reliability of the proposed approach for automated detection of oral leukoplakia. Additionally, the use of Grad-CAM++ approach in the framework provided the interpretability of the model because the model was able to focus on important regions of the lesions. But there were some limitations of the framework even after its great potential for applications. The current design of the framework can classify only positive and negative cases of oral leukoplakia but cannot differentiate oral leukoplakia from other potentially malignant conditions in the mouth that have similar visual appearance. Moreover, the framework was evaluated using a relatively small database that does not necessarily represent the variations that can be encountered in practice.

REFERENCE

- [1] Li, L., Pu, C., Tao, J., Zhu, L., Hu, S., Qiao, B., Xing, L., Wei, B., Shi, C., Chen, P. and Zhang, H., 2024. Development of an oral cancer detection system through deep learning. *BMC Oral Health*, 24(1), p.1468.
- [2] Devindi, G.A.I., Dissanayake, D.M.D.R., Liyanage, S.N., Francis, F.B.A.H., Pavithya, M.B.D., Piyaarthne, N.S., Hettiarachchi, P.V.K.S., Rasnayaka, R.M.S.G.K., Jayasinghe, R.D., Ragel, R.G. and Nawinne, I., 2024. Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer. *IEEE Access*, 12, pp.124375-124390.
- [3] Zhang, X., Gleber-Netto, F.O., Wang, S., Martins-Chaves, R.R., Gomez, R.S., Vigneswaran, N., Sarkar, A., William Jr, W.N., Papadimitrakopoulou, V., Williams, M. and Bell, D., 2023. Deep learning-based pathology image analysis predicts cancer progression risk in patients with oral leukoplakia. *Cancer medicine*, 12(6), pp.7508-7518.
- [4] Cai, X., Li, L., Yu, F., Guo, R., Zhou, X., Zhang, F., Zhang, H., Zhang, J. and Li, T., 2023. Development of a pathomics-based model for the prediction of malignant transformation in oral leukoplakia. *Laboratory investigation*, 103(8), p.100173.
- [5] Dominguet, M.H.L., de Araújo, V.C., Mariano, F.V., Lima, C.S.P., Scarini, J.F., de Camargo Moraes, P., Montalli, V.A.M., Agatti, L., Zaini, Z.M., Hellmeister, L. and Sperandio, M., 2023. Risk assessment of oral leukoplakia by DNA content enhanced by machine learning models. *Oral Oncology Reports*, 6, p.100046.
- [6] Dinesh, Y., Ramalingam, K., Ramani, P. and Deepak, R.M., 2023. Machine learning in the detection of oral lesions with clinical intraoral images. *Cureus*, 15(8), p.e44018.
- [7] Liu, Y., Bilodeau, E., Pollack, B. and Batmanghelich, K., 2022. Automated detection of premalignant oral lesions on whole slide images using convolutional neural networks. *Oral Oncology*, 134, p.106109.
- [8] Sakharkar, M., Spokas, G., Berry, L., Daniels, K., Nithagon, P., Rodriguez-Diaz, E., Tracy, L., Noordzij, J.P., Bigio, I., Grillone, G. and Krisciunas, G.P., 2025. Non-invasive screening for laryngeal cancer using the oral cavity as a proxy for differentiation of laryngeal cancer versus leukoplakia: A novel application of ESS technology and artificial intelligence supported statistical modeling. *American journal of otolaryngology*, 46(1), p.104581.
- [9] Thakuria, T., Mahanta, L.B., Khataniar, S.K., Goswami, R.D., Baruah, N. and Bharali, T., 2025. Smartphone-Based Oral Lesion Image Segmentation Using Deep Learning. *Journal of Imaging Informatics in Medicine*, pp.1-19.
- [10] Tafala, I., Ben-Bouazza, F.E., Edder, A., Manchadi, O. and Jioudi, B., 2025. DeepPatchNet: A deep learning model for enhanced screening and diagnosis of oral cancer. *Informatics in Medicine Unlocked*, 57, p.101658.
- [11] HEMASHREE, H., PRADEEP, N., AHMED MUJIB, B.R. and Shashidhar, R., 2025. Hyperparameter Optimized Deep Learning Models for Multiclass Oral Lesion Classification in Proprietary Histopathological Image Data. *Franklin Open*, p.100437.
- [12] Xie, F., Xu, P., Xi, X., Gu, X., Zhang, P., Wang, H. and Shen, X., 2024. Oral mucosal disease recognition based on dynamic self-attention and feature discriminant loss. *Oral Diseases*, 30(5), pp.3094-3107.
- [13] Myriam, H., Abdelhamid, A.A., El-Kenawy, E.S.M., Ibrahim, A., Eid, M.M., Jamjoom, M.M. and Khafaga, D.S., 2023. Advanced meta-heuristic algorithm based on Particle Swarm and Al-biruni Earth Radius optimization methods for oral cancer detection. *IEEE Access*, 11, pp.23681-23700.

- [14] Desai, K.M., Singh, P., Smriti, M., Talwar, V., Chaudhary, M., Paul, G., Kolli, S.C., Raghava, P.S., Krishna, G.V., Jawahar, C.V. and Vinod, P.K., 2025. Screening of oral potentially malignant disorders and oral cancer using deep learning models. *Scientific Reports*, 15(1), p.17949.
- [15] Adeoye, J., Chaurasia, A., Akinshipo, A., Suleiman, I.K., Zheng, L.W., Lo, A.W.I., Pu, J.J., Bello, S., Oginni, F.O., Agho, E.T. and Braimah, R.O., 2024. A deep learning system to predict epithelial dysplasia in oral leukoplakia. *Journal of dental research*, 103(12), pp.1218-1226.
- [16] Schmidl, B., Hütten, T., Pigorsch, S., Stögbauer, F., Hoch, C.C., Hussain, T., Wollenberg, B. and Wirth, M., 2025. Artificial intelligence for image recognition in diagnosing oral and oropharyngeal cancer and leukoplakia. *Scientific Reports*, 15(1), p.3625.
- [17] Pontes, G.I., Araújo, A.L.D., Soares, A.B., Warnakulasuriya, S., de Freitas, A.L.S., Stevão, C.G., Sperandio, M. and Moraes, M.C., 2026. Predicting Malignant Transformation in Oral Leukoplakia: A Multilayer Perceptron Approach Incorporating Clinicopathological Features and DNA Content. *Journal of Oral Pathology & Medicine*, 55(2), pp.249-260.
- [18] Liyanage, V., Tao, M., Park, J.S., Wang, K.N. and Azimi, S., 2023. Malignant and non-malignant oral lesions classification and diagnosis with deep neural networks. *Journal of Dentistry*, 137, p.104657.
- [19] Peng, J., Xu, Z., Dan, H., Li, J., Wang, J., Luo, X., Xu, H., Zeng, X. and Chen, Q., 2024. Oral epithelial dysplasia detection and grading in oral leukoplakia using deep learning. *BMC Oral Health*, 24(1), p.434.
- [20] Ferrer-Sanchez, A., Bagan, J., Vila-Frances, J., Magdalena-Benedito, R. and Bagan-Debon, L., 2022. Prediction of the risk of cancer and the grade of dysplasia in leukoplakia lesions using deep learning. *Oral Oncology*, 132, p.105967.
- [21] Adeoye, J., Koohi-Moghadam, M., Choi, S.W., Zheng, L.W., Lo, A.W.I., Tsang, R.K.Y., Chow, V.L.Y., Akinshipo, A., Thomson, P. and Su, Y.X., 2023. Predicting oral cancer risk in patients with oral leukoplakia and oral lichenoid mucositis using machine learning. *Journal of Big Data*, 10(1), p.39.
- [22] Kouketsu, A., Doi, C., Tanaka, H., Araki, T., Nakayama, R., Toyooka, T., Hiyama, S., Iikubo, M., Osaka, K., Sasaki, K. and Nagai, H., 2024. Detection of oral cancer and oral potentially malignant disorders using artificial intelligence-based image analysis. *Head & Neck*, 46(9), pp.2253-2260.
- [23] Su, A.Y., Wu, M.L. and Wu, Y.H., 2025. Deep learning system for the differential diagnosis of oral mucosal lesions through clinical photographic imaging. *Journal of Dental Sciences*, 20(1), pp.54-60.
- [24] Hsu, Y., Chou, C.Y., Huang, Y.C., Liu, Y.C., Lin, Y.L., Zhong, Z.P., Liao, J.K., Lee, J.C., Chen, H.Y., Lee, J.J. and Chen, S.J., 2025. Oral mucosal lesions triage via YOLOv7 models. *Journal of the Formosan Medical Association*, 124(7), pp.621-627.
- [25] Vinayahalingam, S., van Nistelrooij, N., Rothweiler, R., Tel, A., Verhoeven, T., Tröltzsch, D., Kesting, M., Bergé, S., Xi, T., Heiland, M. and Flügge, T., 2024. Advancements in diagnosing oral potentially malignant disorders: leveraging Vision transformers for multi-class detection. *Clinical oral investigations*, 28(7), p.364.
- [26] <https://huggingface.co/datasets/HRruiH/Dataset-of-oral-mucosal-diseases>
- [27] Sun, Y., Zhao, Z., Jiang, D., Tong, X., Tao, B., Jiang, G., Kong, J., Yun, J., Liu, Y., Liu, X. and Zhao, G., 2022. Low-illumination image enhancement algorithm based on improved multi-scale Retinex and ABC algorithm optimization. *Frontiers in Bioengineering and Biotechnology*, 10, p.865820.
- [28] Bui, P.N., Le, D.T., Bum, J., Han, J.C., Pham, V.N. and Choo, H., 2025. Multi-scale feature enhancement in multi-task learning for medical image analysis. *Artificial Intelligence in Medicine*, p.103338.
- [29] Jiang, Y., Xu, S., Fan, H., Qian, J., Luo, W., Zhen, S., Tao, Y., Sun, J. and Lin, H., 2021. Ala-net: Adaptive lesion-aware attention network for 3d colorectal tumor segmentation. *IEEE transactions on medical imaging*, 40(12), pp.3627-3640.
- [30] Li, D., Wang, X., Liu, J. and Sun, J., 2025. Deep Learning-Based Automated Detection of Oral Leukoplakia in Clinical Imaging. *Cureus*, 17(5).