2026-08

Deep learning-based classification of leukoplakia and OSCC using histopathological images: a performance evaluation study

Chowdhury, Md Sazedul Islam

Independent University, Bangladesh (IUB)

Deep Learning-Based Classification of Leukoplakia and OSCC Using Histopathological Images: A Performance Evaluation Study

August 2026

Prepared by:

Md Sazedul Islam Chowdhury
ID: 2210888

Iffat Jahan Ishika
ID: 2222715

Maruf Al Hasan
ID: 2130208

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by:

Md. Rashedur Rahman, D. Eng.
Assistant Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

A Final Year Design Project (FYDP) report submitted for the Degree of Bachelor's of Computer Science & Engineering

Attestation

We are aware of the fact that plagiarism is strictly prohibited and that it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted materials and models developed by other researchers in our project, which have been properly cited in accordance with international standards and appropriate academic guidelines.

Author Name:

.....

Signature:

.....

Author Name:

.....

Signature:

.....

Author Name:

.....

Signature:

.....

Evaluation Committee

Supervisor

Name: _____ Signature _____

Internal Examiner 1

Name: _____ Signature _____

Internal Examiner 2

Name: _____ Signature _____

External Examiner

Name: _____ Signature _____

Acknowledgement

We would like to express our sincere gratitude to all those who have supported and contributed to the successful completion of this thesis.

First and foremost, we would like to express our deepest gratitude to our thesis supervisor, Md. Rashedur Rahman, D. Eng., Assistant Professor, Department of Computer Science and Engineering, Independent University, Bangladesh, for his invaluable guidance, continuous encouragement, constructive feedback and unwavering support throughout this research. His expertise, thoughtful suggestions and commitment to academic excellence have played a significant role in shaping this study and helping us overcome the challenges encountered throughout the research process.

We would also like to extend our sincere appreciation to Siam Tahsin Bhuiyan, Samiul Karim Mazumder and Fatin Fawjia Tanha, for their valuable guidance, technical support, constructive feedback and encouragement throughout the course of this work. Their insights and willingness to share their knowledge greatly contributed to the development and refinement of this research.

We also seek to thank Independent University, Bangladesh (IUB) and the Center for Computational & Data Sciences (CCDS) for providing a supportive academic and research environment, along with the necessary resources and opportunities that facilitated the successful completion of this thesis.

Also, we want to express our heartfelt gratitude to our family and friends for their unconditional support, patience, encouragement, and understanding throughout our academic journey. Their constant motivation, reassurance, and belief in us have been an invaluable source of strength, particularly during the challenging stages of this research.

Finally, we additionally aim to acknowledge the use of ChatGPT (OpenAI) and Claude (Anthropic**)** as supplementary writing tools during the preparation of this thesis. These tools were used solely to improve the clarity, grammar, organization and overall presentation of the written content. They were not used to generate, modify, interpret or alter the research results, experimental findings, data analysis or conclusions of this study. All research decisions, analyses, results and academic judgments presented in this thesis remain the responsibility of the authors.

List of Publications

M. S. I. Chowdhury, F. F. Tanha, I. J. Ishika, M. Al Hasan, S. K. Mazumder, R. Islam, R. Rahman, A. Islam, and S. B. Alam , “Deep Learning-Based Classification of Leukoplakia and OSCC Using Histopathological Images: A Performance Evaluation Study,” 2026 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), 04-05 September 2026, IUBAT-International University of Business Agriculture and Technology. (Presented)

ABSTRACT

Oral Squamous Cell Carcinoma (OSCC) is a malignant cancer that develops in the squamous cell lining in the oral cavity which is one of the most common and clinically significant malignancies of the oral cavity. In contrast, Oral Leukoplakia is a white oral mucosal lesion considered a potentially malignant disorder that may undergo malignant transformation and progress to OSCC. Early and accurate diagnosis is essential for improving patient outcomes; however, differentiating OSCC from Oral Leukoplakia remains challenging because the two lesions often exhibit similar histopathological characteristics. Since these conditions require different clinical management strategies, reliable discrimination between them is of considerable diagnostic importance.

This thesis investigates the application of deep learning for the histopathological classification of Oral Leukoplakia and OSCC using the publicly available NDB-UFES dataset. A comparative study was conducted using six Convolutional Neural Network (CNN) architectures: ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt-Base. To ensure a fair and consistent evaluation, all models were trained and tested under identical experimental conditions using the same preprocessing pipeline, transfer learning strategy and evaluation framework.

Model performance was assessed using Accuracy, class-wise Recall, Area Under the Receiver Operating Characteristic Curve (AUROC) and Matthews Correlation Coefficient (MCC). Among the evaluated architectures, ResNet-101 achieved the strongest overall performance, obtaining an Accuracy of 89.58%, an AUROC of 0.9481 and an MCC of 0.7810 while maintaining balanced Recall across both diagnostic classes.

To further examine model interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to visualize the image regions influencing the classification decisions. The generated heatmaps indicated that the models focused on diagnostically relevant tissue structures, providing additional confidence in the validity of the learned representations.

Overall, the findings suggest that the residual learning mechanism of ResNet-101 provides a robust and reliable foundation for the histopathological classification of Oral Leukoplakia and OSCC. This study contributes a comprehensive comparison of modern CNN architectures and establishes a strong baseline for future research in artificial intelligence-assisted oral cancer diagnosis.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1 Background

Oral cancer poses as a major concern for public health on a global scale and is a crucial contributor to cancer-related diseases and deaths worldwide. Each year, more than 300,000 new cases of oral cancer are reported worldwide, making it one of the most common cancers globally [1]. Late diagnosis of oral lesion results in advanced disease by decreasing the chance of survival by 30.00% [1][2]. Early-stage oral cancer carries a generally more favorable prognosis, underscoring the critical importance of timely and accurate detection in determining patient outcomes.

Figure 2: Leukoplakia of the floor of the mouth [36].

Figure 3: A vast leukoplakic lesion of the right tongue border (2.5 × 2.5 cm) with an OSCC of 1 cm in greatest dimension in its context [37].

Figure 4: An OSCC patient with mucosal changes of wide field of cancerization (cotton white discoloration) [38].

Figure 1: Clinical images Oral Leukoplakia and OSCC

Oral Leukoplakia (OL) occupies a central and clinically important position among oral potentially malignant disorders. It appears as thick, painless white or grey patches with the capacity to develop into cancer and is considered one of the most frequently observed premalignant disorders of the oral cavity [2]. The malignant transformation rate of Leukoplakia into OSCC ranges between 0.100% and 36.40% [2] which indicates that not all Leukoplakic lesions changes to carcinoma. It can be challenging to determine whether a lesion remains in the premalignant stage or has already developed into invasive carcinoma as both conditions may present as white or thickened patches in the oral mucosa and can share similar histopathological features such as epithelial alterations with increased cellular activity and abnormal tissue structures [2]. This similarity complicates treatment planning which makes precise histopathological diagnosis vital for early detection and prevention of malignant transformation.

In medical image analysis, Convolutional Neural Networks (CNNs) have proved outstanding performance for classification tasks as they can automatically extract discriminative features without complicated and costly traditional feature engineering, and have produced promising results across multiple medical imaging fields [3]. CNNs have notably shown strong performance in classifying histopathological images specifically, which indicates their potential for automated pathological analysis along with their capacity to learn complex

textures, structures and other structural features from high-resolution images makes them well suited to oral histopathological implementations [4]. Deep learning more broadly is an effective paradigm for tasks like image classification, segmentation and feature extraction motivating an extensive research scope in this area [5].

Despite these encouraging developments, significant challenges remain in the practical deployment of AI-based oral cancer diagnostic systems, especially the availability of well-annotated datasets, consistent comparison of candidate CNN architectures and the interpretability of model predictions for clinical use [6].

1.2 Problem Statement

The precise and early diagnosis of OL and OSCC presents a set of connected clinical and computational challenges that are not fully mentioned by current diagnostic approaches. Histopathological examination remains the gold standard for diagnosis, however distinguishing OL from OSCC can be difficult because these lesions often exhibit similar clinical presentations and overlapping microscopic characteristics that includes comparable epithelial changes and abnormal tissue architectures [2].

Recent advances in artificial intelligence (AI), particularly deep learning (DL), have shown considerable potential for assisting the analysis of oral histopathological images. However, existing studies have several limitations. Many approaches focus on OSCC versus normal or non-cancerous tissues, while relatively few specifically address the more challenging OL versus OSCC classification [10][11]. In addition, existing studies often evaluate different architectures, modified networks, or hybrid approaches under varying datasets and experimental settings [7][9]. Consequently, their reported performances cannot be directly compared to determine which established CNN architecture is most suitable for this classification task.

Along with methodological limitations, the traditional diagnosis process using multiple tests and tissue examination is usually time-consuming and stressful for clinicians [8]. However, delays in diagnosis may postpone treatment initiation and adversely affect patient outcomes. So, there is a growing need for reliable computer aided diagnostic systems that can help clinicians in the early classification of oral lesions by supporting diagnostic decision-making and reducing clinical workload and facilitate timely patient management without replacing expert pathological assessment [1][9]. Along with improved predictive performance, another significant challenge limiting the clinical adoption of AI-based diagnostic systems is the limited interpretability of deep learning models. Many high-performing CNN models operate as "black-box" systems providing accurate predictions without sufficiently explaining the underlying reasoning behind their decisions. This lack of transparency reduces clinician confidence and complicates model validation and also hinders integration into routine clinical practice. Consequently, improving model interpretability and ensuring that predictions are supported by diagnostically meaningful histopathological features have become important research priorities to develop trustworthy and clinically applicable AI systems for oral cancer diagnosis [6][12].

1.3 Research Motivation

Deep learning is particularly well suited to histopathological image analysis as CNNs can automatically learn complex textures, morphological patterns and discriminative tissue features directly from high-resolution images without manually engineered descriptors [3][4]. Owing to their ability to perform feature extraction and classification in an end-to-end manner, deep learning models have become an effective and versatile approach for medical image analysis such as image classification, segmentation and feature extraction tasks motivating their continued application to oral cancer diagnosis [5].

Since relatively few existing studies have investigated the direct discrimination of OL and OSCC as opposed to the more common task of distinguishing OSCC from normal tissue, a main motivation of this research is to identify the most effective CNN architectures specifically for this more clinically demanding classification problem [5]. Addressing this challenge has the potential to improve diagnostic decision-making by reducing human-related diagnostic variability and supporting early and more accurate identification of suspicious oral lesions. Because, improved diagnostic accuracy may help decrease unnecessary treatment costs related with incorrect clinical assessments and inappropriate treatment decisions [8].

A further motivation is methodological. Previous studies have generally evaluated CNN architectures individually or compared only a limited subset of models without employing a controlled and identical experimental setup [7][8][9]. So there is a clear need for a thorough, side-by-side comparison of multiple modern CNN architectures under identical experimental conditions. Such comparative evaluation is needed for drawing reliable conclusions regarding which architectural design is most suitable for distinguishing OL from OSCC using histopathological images.

Explainability is also an important motivation for this research. Clinicians must be able to check that model predictions are based on diagnostically meaningful and structurally accurate tissue areas rather than false correlations or imaging artifacts. Incorporating interpretability techniques can improve the transparency of deep learning models which will strengthen clinician confidence in AI-assisted diagnostic systems and also support their meaningful integration into routine oral medicine clinical workflows.

1.4 Research Objectives

This thesis pursues the following specific research objectives:

1. To systematically compare six CNN architectures for the binary classification of OL and OSCC using histopathological images from the NDB-UFES dataset.

2. To identify the best-performing state-of-the-art architecture using Accuracy, class-wise Recall, AUROC and Matthews Correlation Coefficient (MCC).

3. To analyze the trade-off between model complexity (parameter count) and classification performance across the six implemented architectures.

1.5 Research Questions

1. Which of the six evaluated CNN architectures, ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt-Base achieves the highest classification performance for binary discrimination between OL and OSCC on the NDB-UFES dataset?

2. How do the evaluated architectures compare in terms of class-wise Recall for leukoplakia and OSCC and what does this reveal about each model's reliability in a clinical screening context?

3. What is the relationship between model complexity as measured by parameter count and classification performance among the evaluated architectures?

1.6 Scope of Study

1. The study addresses a binary classification problem which discriminates between OL and OSCC.

2. The image data used consists of histopathological images drawn from the NDB-UFES dataset.

3. Six CNN architectures are evaluated as image classification backbones: ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt-Base and all are evaluated under identical experimental conditions, training settings, and hyperparameters.

4. Model evaluation employs Accuracy, class-wise Recall (for OL and for OSCC), AUROC and MCC. Precision, Specificity and F1-Score are not used since both classes represent pathological conditions and there is no distinct non-disease category.

5. Interpretability analysis is conducted using Grad-CAM visualization applied to all six evaluated CNN models.

6. The study does not contain multi-class oral lesion classification, lesion segmentation, object detection, survival prediction or the integration of structured clinical metadata into the classification model.

1.7 Research Contributions

1. **Systematic CNN Comparative Analysis:** A comprehensive comparison of six modern CNN architectures for binary OL versus OSCC classification conducted under identical experimental conditions.

2. **Identification of Optimal CNN Architecture:** identification of the best performing architecture through quantitative evaluation across Accuracy, class-wise Recall, AUROC and MCC.

3. **Complexity-Performance Analysis:** analysis of the relationship between parameter count and classification performance across the six architectures.

1.8 Thesis Organization

Chapter 1 – The introduction discusses oral squamous cell carcinoma (OSCC) and leukoplakia, their associated challenges and current diagnostic limitations, the research problem, objectives, research questions, and the motivation behind the study.

Chapter 2 - Literature Review covers the clinical characteristics and histopathology of OL and OSCC, the evolution of AI in oral cancer diagnosis, descriptions of the six evaluated CNN architectures, transfer learning in medical imaging, explainable AI methods, a review of previous studies, and the identified research gap.

Chapter 3 - Dataset describes the NDB-UFES dataset.

Chapter 4 - Methodology describes preprocessing pipeline, experimental design, training protocol, and evaluation metrics.

Chapter 5 - Results and Discussion presents quantitative results for all six CNN models, comparative analysis, the parameter counts versus performance trade-off, and Grad-CAM visualization results. Then the findings are interpreted in context of the research questions, discusses clinical implications, addresses limitations, and identifies future work.

Chapter 6 - Limitation and Future Work defines the constraints of the study beside suggesting the future work.

Chapter 7 - Conclusion summarizes key findings, restates contributions, and offers concluding remarks.

CHAPTER 2: LITERATURE REVIEW

This chapter presents a review of previous studies related to OL, OSCC and the use of artificial intelligence in their diagnosis. It discusses the clinical and histopathological characteristics of these conditions, along with the existing diagnostic methods. It also reviews the use of traditional machine learning, deep learning and transfer learning in oral cancer research. Finally, the limitations of existing studies are discussed to identify the research gap addressed in this thesis.

2.1 Oral Leukoplakia

OL is one of the most clinically important potentially malignant disorders because it is visible on examination but still its biological behavior is variable and not fully predictable from appearance alone. In the literature, OL is continuously treated as a lesion that needs careful clinical examination because it may remain stable, recur after treatment or undergo malignant transformation depending on its clinical type, anatomical site and histopathological grade and also patient-related risk factors. This uncertainty is precisely what makes OL a relevant and convincing target for computational analysis. When a lesion's risk cannot be derived reliably from clinical inspection alone, decision support systems based on imaging and metadata become especially valuable. Even with constantly improving histopathological analysis, common features between leukoplakia and OSCC may still cause major difficulties for oral healthcare and diagnosis which further reinforces the need for automated intelligent diagnostic support systems [8].

2.1.1 Clinical Characteristics

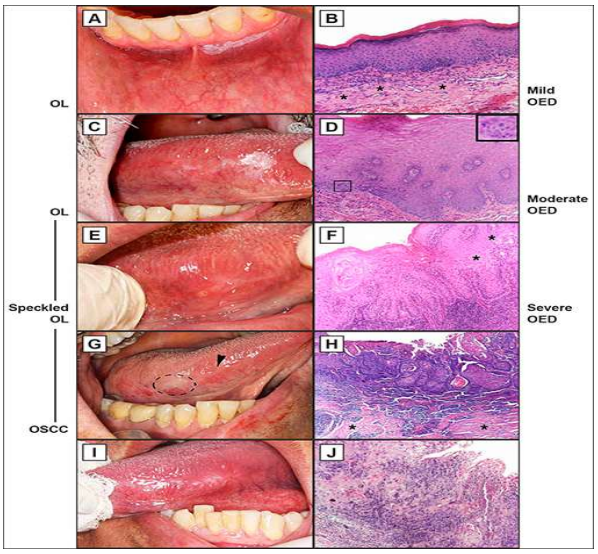

Figure 2: Corresponding clinical and microscopic aspects of oral leukoplakia and OSCC [39].

Clinically, OL is generally described as a white patch or plaque that cannot be characterized as any other definable disease which remains after other causes of oral white lesions have been excluded. This definition is important as the lesion is a diagnosis of exclusion rather than a single etiologic entity. In practice, leukoplakia may present as homogeneous or non-homogeneous, flat or verrucous, localized or diffuse and it may occur at several places inside the mouth. The tongue and floor of mouth are mainly identified as high-risk site in the clinical literature while lesions at other sites may show different patterns of relapse and advancement. The clinical challenge is that leukoplakia mostly overlaps visually with frictional keratosis, oral candidiasis, lichen planus, lichenoid reactions and early carcinoma. Due to this reason, only visual inspection is not enough for accurate risk stratification. A recent study on OL classification showed that CNN could differentiate leukoplakia from other white lesions using clinical photographs and thus supports the concept that image-based pattern recognition can assist at this early diagnostic stage [13]. Similarly, a photograph-based deep learning system successfully predicted epithelial dysplasia in OL and showed performance capable of supporting biopsy selection and patient self-monitoring [14]. These findings highlight a central clinical reality that is, the practical diagnostic problem is not only lesion recognition but also identifying which lesions require closer follow-up or immediate tissue diagnosis.

As medical images represent the largest part of clinical data, lack of resources and technologies make early detection of oral lesions particularly challenging in routine clinical environments [8]. The traditional diagnosis process involves multiple clinical tests and tissue examination is usually time-consuming and stressful for both clinicians and patients, this builds significant demands on specialist pathological services [1]. Therefore, a system capable of classifying oral cancer early without manual involvement is essential to reduce the serious clinical consequences of delayed diagnosis [1][9].

A further clinically important characteristic of OL is its tendency to recur after treatment. In a study focused specifically on tongue leukoplakia, recurrence was common after carbon dioxide laser surgery and malignant transformation still occurred during post-treatment follow-up [15]. This finding indicates that even when lesions are successfully treated, the underlying biological risk may stay. OL is hence best understood not as a static lesion responsive to single-intervention cure but as a condition needing longitudinal and sustained clinical management.

2.1.2 Histopathological Assessment of Oral Leukoplakia

Histopathologically, OL may range from hyperkeratosis without dysplasia to lesions showing mild, moderate or severe epithelial dysplasia. The presence and grade of dysplasia are primary clinical parameters because they are related with the probability of malignant progression. Still histopathologic interpretation is not always straightforward or consistent. Inter-observer variability is a well-documented problem in the grading of oral epithelial dysplasia and this limitation has directly motivated the development of automated techniques that attempt to standardize and objectify morphological assessment.

One recent study developed a fully automated algorithm which predicts malignant transformation in oral epithelial dysplasia using whole-slide histopathology images by implementing a segmentation-based computational pipeline to detect cell nuclei and define the epithelium before extracting clear morphological and spatial features [16]. This study is

especially notable because it addresses not only diagnostic classification but also prognostic risk prediction from histological structure which is a considerably more clinically meaningful output. A second study introduced an objective deep learning model to detect pathological features and grade oral epithelial dysplasia in leukoplakia which further illustrates the productive movement toward computational support in microscopic pathological interpretation [17].

At a biological level, the progression from leukoplakia to invasive carcinoma is not determined solely by the presence or grade of epithelial atypia. The tumor microenvironment, the character of the inflammatory cellular infiltrate and nuclear morphology and the degree of architectural disturbance all contribute to malignant risk. The histology-driven AI literature increasingly reflects this biological complexity by moving from simple binary lesion classification toward feature-based prognostic modeling that captures multiple tissue-level characteristics simultaneously.

2.2 Oral Squamous Cell Carcinoma (OSCC)

OSCC is the most common histologic type of oral cancer which continues to represent a major global health burden with continuous poor overall survival rates. The literature repeatedly focuses that OSCC outcomes are highly dependent on the stage at which diagnosis is established, making early and accurate lesion detection and classification clinically vital [1]. Because histopathology remains to be the definitive diagnostic standard, OSCC is well suited to image-based machine learning analysis. The principal challenge is that microscopic diagnosis needs specialized expertise and considerable time along with the ability to achieve consistent interpretation across highly heterogeneous tissue patterns.

2.2.1 Clinical Overview

OSCC typically develops against a background of chronic risk factor exposure, especially including tobacco use, alcohol consumption and in specific geographic and demographic settings, additional carcinogenic influences. Clinically, OSCC may happen from visibly suspicious mucosal lesions that attract clinical attention or from lesions that are initially subtle in appearance and easily overlooked during routine examination. The critical importance of early clinical recognition is repeatedly highlighted throughout the AI-based oral cancer literature because delayed diagnosis is consistently associated with worse patient survival and more extensive and morbid treatment interventions [1].

A comprehensive scoping review of AI applications in oral cancer resulted that AI portrays meaningful potential for improving early detection, risk modeling, image phenotyping and prognostic assessment [1]. This broad perspective is clinically important because OSCC is not a single imaging problem suitable for a single computational solution, it is a multistage disease requiring detection, histopathological confirmation, accurate staging and individualized treatment planning. The oral lesion detection and classification literature frequently frames OSCC classification together with oral potentially malignant disorders rather than as an isolated diagnostic category [19][13].

An additional clinically significant feature of OSCC is that diagnosis may be notably delayed due to limited patient access to specialist oral medicine services particularly in resource-constrained healthcare locations. This access gap makes automated image-based diagnostic evaluation an especially attractive research direction. Deep learning systems applied to oral photographic and histopathological images have showed the feasibility of classifying suspicious lesions and identifying malignant-appearing tissue which suggests a meaningful role for AI-assisted screening before formal biopsy or specialist referral. The clinical importance of this research direction lies in the recognition that AI is not intended to replace the oral pathologist but rather to reduce the frequency of missed early lesions and support a more efficient and fair diagnostic workflow [8][9].

2.2.2 Histopathological Assessment of OSCC

Histopathologically, OSCC is defined by malignant epithelial expansion with observable invasive growth into the underlying connective tissue stroma. Diagnostically important histological features include altered nuclear morphology with prominent nucleoli, atypical and abnormal mitotic figures, loss of normal epithelial maturation and stratification, architectural disorganization and variable degrees of keratinization including keratin pearls. Tumor heterogeneity is a major and constant obstacle to consistent histopathological diagnosis since different regions of the same tissue section may show varying degrees of differentiation and different patterns of stromal interaction. This spatial and morphological heterogeneity is one of the primary reasons deep learning has become so relevant to OSCC histopathology and convolutional architectures are well suited to learning complex spatial patterns across large and heterogeneous histological images [4][20].

Several important studies in the literature focus specifically on OSCC histopathology. A transfer learning-based investigation using pretrained CNN architectures showed that ResNet-50 performed strongly in OSCC histopathological classification which confirms the value of pretrained feature extractors for medical histological image analysis [20]. Another study implemented a hybrid learning strategy combining deep feature extraction with handcrafted classical descriptors reported high overall performance when deep features were combined with supplementary color, texture and shape representations [22]. Mandal evaluated several supervised machine learning methods for OSCC identification using ResNet-50-derived features, reporting a maximum accuracy of 89.00% for KNN and 81.00% for Linear SVM, however that study did not directly compare the six CNN architectures considered as baseline in the present thesis, and it focused on OL subtype and OSCC classification rather than broader OL-OSCC differentiation [7]. These results together indicate that OSCC histology contains both local microstructural features and broader textural patterns which can be effectively used through computational methods.

Explainable AI studies further demonstrate that histopathological deep learning predictions can be made more transparent and clinically meaningful. A study of head and neck cancer histopathology utilized Grad-CAM and related gradient-based visualization methods to confirm that the network's classification decisions were related with pathologist-relevant tissue regions [23]. This correspondence between computational attention and expert pathological

judgment is clinically important because OSCC diagnosis must be defensible, auditable and consistent with established diagnostic criteria.

2.2.3 Current Diagnostic Procedures

The current diagnostic pathway for OSCC starts with clinical examination of the oral mucosa followed by tissue biopsy and formal histopathological evaluation when clinically suspicious lesions are identified. Histological examination remains the definitive diagnostic gold standard but is time-consuming and dependent upon the availability and consistent judgment of experienced oral pathologists [20][22]. Complementary imaging modalities and computer-aided diagnostic systems are therefore being actively tested as ways of improving diagnostic efficiency and reducing the burden on specialist pathological services.

In routine clinical practice, histopathological examination is limited by institutional workload demands, variation among observers in borderline cases and the inherent subjectivity involved in morphological assessment of diagnostically challenging specimens. AI methods have been proposed to systematically address these limitations by providing fast, reproducible and quantitatively robust diagnostic support [8][9]. A multiclass oral lesion classification study using clinical photographic images portrayed that CNNs can distinguish OSCC from other diagnostic categories with strong AUC values which suggests that even before formal biopsy, image analysis can meaningfully support clinical screening and referral decisions [19].

More technically advanced approaches include fully automated and explainable computational systems for malignant transformation prediction in oral epithelial dysplasia [16] and attention-guided classification models for oral lesion analysis [24]. These methodological developments together indicate that the oral cancer AI diagnostic pipeline may originate from simple lesion recognition toward comprehensive decision support spanning initial assessment, dysplasia grading and long-term prognostic assessment.

2.3 Malignant Transformation Risk

The malignant transformation risk related with OL is one of the most clinically important reasons it attracts sustained research attention. Not every leukoplakic lesion transforms to carcinoma but a clinically meaningful minority does, and the ability to identify high-risk lesions prospectively has significant implications for patient management. In a tongue leukoplakia cohort, postoperative malignant transformation occurred in a measurable subset of patients and the annual transformation rate was reported as 2.28% [15]. This confirms that OL cannot responsibly be rejected as a benign surface abnormality, rather it should be considered a risk-bearing lesion with variable and partially unpredictable outcome progression.

Malignant transformation risk is caused by multiple interacting factors. The tongue leukoplakia study reported relations between recurrence or malignant change and variables including lesion size, clinical appearance, histopathological findings and history of head and neck cancer with lesion size and previous cancer history occuring as particularly important prognostic factors [15]. A deep learning system that predicted dysplasia probability from oral photographs further showed that image-based computational methods could estimate risk before biopsy and

potentially enable more targeted and clinically appropriate patient selection for histopathological investigation [14].

Overall, the clinical importance of malignant transformation risk is that it creates a diagnostically and prognostically meaningful computational target that is beyond simple binary lesion classification. Models capable of distinguishing leukoplakia from other white lesions are clinically helpful but models that estimate dysplasia grade or quantify transformation risk are more directly connected with the practical clinical decisions that oral medicine specialists must make.

2.4 Artificial Intelligence in Oral Cancer Diagnosis

AI in oral cancer diagnosis has evolved progressively from conventional machine learning methods built on manually engineered features toward deep learning architectures that learn representations directly from image data [1][22]. The literature shows a clear and coherent development across these methodological generations, each successive generation offering improvements in automation and representational power along with clinical relevance.

2.4.1 Traditional Machine Learning

Before the widespread adoption of deep learning in medical image analysis, oral cancer research commonly used traditional machine learning pipelines built upon manually extracted visual features which typically relies on color histograms, texture descriptors, local binary patterns, wavelet transform coefficients and shape-based measures. In one representative OSCC study, hybrid systems combined CNN-derived deep features with handcrafted descriptors that included fuzzy color histograms, discrete wavelet transform representations, local binary patterns and gray-level co-occurrence matrix features, then classified the resulting fused feature vectors using conventional machine learning classifiers [22]. Although the best results in that study were produced by hybrid deep learning pipelines, the work clearly illustrates the earlier machine learning paradigm that continues to inform modern feature engineering strategies.

Traditional machine learning approaches maintain practical utsefulness in contexts where datasets are small, where expert-defined visual descriptors are believed to capture clinically meaningful diagnostic structure or where model interpretability and computational efficiency are crucial concerns. However, they generally need careful manual feature design and domain expertise and may not generalize as effectively as deep architectures when the feature space is high-dimensional and complex. Mandal's finding in which KNN outperformed Linear SVM (89.00% versus 81.00% accuracy) for OSCC identification illustrates both the capabilities and the limitations of traditional classification approaches when applied to complex histopathological image data [7]. The progressive transition toward deep learning in oral cancer research hence reflects a fundamental shift from human-engineered feature representation to automatically learned hierarchical representation.

2.4.2 Deep Learning and Transfer Learning

Deep learning has become an important computational approach in oral cancer research because it can automatically learn hierarchical visual representations directly from image data without requiring extensive manual feature engineering. It has been widely applied to image classification, segmentation, feature extraction, and prognostic analysis in medical imaging [5]. A systematic review reported a substantial increase in deep learning studies in oral cancer and identified promising performance across classification, detection, segmentation, and prognostic prediction tasks [19]. In oral lesion analysis, deep learning has been applied to both clinical photographs and digitized histopathological images. For example, a study using oral photographic images employed DenseNet-169, ResNet-101, SqueezeNet, and Swin-S architectures to classify OSCC and oral potentially malignant disorders and reported strong AUC values for lesion recognition [19]. Other studies have explored architectural and feature-learning strategies to improve oral cancer classification. Metablock fusion with ResNetV2 achieved a balanced accuracy of 83.24%, This approach showed an improvement of 30.68% in performance compared to other implemented models [18], while a guided attention framework demonstrated the potential of architectural modifications to improve the interpretability and clinical relevance of deep learning models [6][24]. In histopathological analysis, deep CNNs have demonstrated the ability to distinguish OSCC from benign tissue and to support grading and progression prediction [16][20][23].

The effectiveness of deep learning for histopathological image analysis is largely attributed to its ability to learn hierarchical and multi-scale representations. CNN-based models can automatically learn local visual patterns, such as nuclear atypia, abnormal epithelial organization, keratinization, and tissue texture, while progressively capturing broader spatial relationships among tissue structures [3][4][5]. This ability reduces the dependence on manually designed feature descriptors and makes CNNs particularly suitable for histopathological image classification. CNN-derived features have also been combined with conventional machine learning and handcrafted descriptors. Ahmad et al. [10], for instance, proposed a hybrid approach that combined CNN-based deep features with handcrafted features, including Gray-Level Co-occurrence Matrix (GLCM), Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP), followed by Support Vector Machine (SVM) classification for early OSCC diagnosis. Their hybrid feature-fusion approach achieved 97.00% accuracy and an AUROC of 0.9680, outperforming the evaluated pretrained models, including DenseNet201, InceptionV3, InceptionResNetV2, NASNetLarge, and Xception [10]. Similarly, Yaduvanshi et al. [11] incorporated local binary pattern-based spatial information into a deep convolutional framework for OSCC classification and reported an accuracy of up to 94.44% on histopathological images [11]. Although these studies demonstrate the effectiveness of deep and handcrafted feature representations, they primarily focused on binary classification between normal and OSCC tissue using patch-based microscopic datasets. They therefore did not directly investigate the distinction between OSCC and leukoplakia, which represents the specific classification problem addressed in the present thesis.

Despite the strong performance of deep learning models, training deep neural networks from scratch generally requires a large number of labeled images. This can be challenging in medical imaging because the collection of histopathological images and expert annotation are costly and time-consuming. Transfer learning provides a practical solution to this limitation by adapting a model that has previously learned visual representations from a large dataset to a new, domain-specific classification task. In medical image analysis, pretrained CNNs,

commonly initialized using large datasets such as ImageNet, can provide useful feature representations that can subsequently be adapted to relatively small biomedical datasets [20][22].

Several studies have demonstrated the effectiveness of transfer learning in oral histopathological classification. A representative study compared pretrained VGG16, VGG19, ResNet-50, InceptionV3, and MobileNet architectures for OSCC histopathological classification and found that fine-tuned ResNet-50 achieved strong performance [20]. Another study developed a modified DenseNet201 architecture with additional custom layers and dropout regularization and reported an accuracy of 91.25%, outperforming NASNetLarge, InceptionNet, and Xception [9]. However, these pretrained models were evaluated using different architectural modifications and were not directly compared with their original, unmodified architectures, limiting their usefulness as a consistent architectural benchmark. Another investigation demonstrated the potential of pretrained architectures through ResNet-101-based feature fusion for OSCC classification [22]. These findings indicate that pretrained CNNs can provide effective representations for oral histopathological images, particularly when the available target dataset is limited.

However, existing studies have often evaluated individual CNN architectures, modified versions of pretrained models or hybrid feature-fusion approaches under different datasets, preprocessing procedures, training strategies, and evaluation settings [9][10][11][20][22]. Consequently, direct comparison of reported performance across architectures is difficult. Moreover, much of the existing oral histopathology literature focuses on distinguishing normal tissue from OSCC rather than distinguishing OSCC from an intermediate potentially malignant lesion such as leukoplakia. These limitations highlight the need for a systematic evaluation of multiple pretrained CNN architectures under a common experimental setting. Therefore, the present study evaluates multiple pretrained CNN architectures using the same dataset, preprocessing pipeline, training strategy, and evaluation framework for the classification of OL and OSCC.

2.5 Explainable AI and Grad-CAM

Explainability is a fundamental requirement for the responsible clinical deployment of deep learning diagnostic systems. In oral pathology and oral medicine, practitioners must be able to understand not only the final classification output of an AI model but also the specific reasoning process and image evidence that generated that output.

Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the widely used approaches for visualizing spatial regions that contribute to CNN classification decisions [25]. Grad-CAM computes gradients of the target class score with respect to feature maps from a convolutional layer and uses these gradients to determine the importance of different spatial regions. The resulting activation map provides a visual representation of areas that most strongly influence the model's prediction [25].

In oral cancer histopathology, such visualization methods can help determine whether a model is focusing on diagnostically meaningful regions, such as abnormal epithelial structures, tumor-associated tissue, dysplastic regions, or other morphological characteristics, rather than

irrelevant background or imaging artifacts [23]. A study of head and neck cancer histopathology demonstrated the usefulness of gradient-based visualization methods for examining whether deep learning predictions corresponded to pathologist-relevant tissue regions [23]. Similarly, attention-guided oral lesion classification approaches have emphasized the importance of directing model attention toward informative regions of the image [24].

2.6 Research Gaps

The reviewed literature provides strong evidence that deep learning has become an effective approach for oral lesion classification and OSCC detection from histopathological images along with the prediction of dysplasia or malignant transformation risk. But still, several important research gaps remain.

First, although numerous CNN-based models have been proposed for oral lesion and OSCC classification, most studies evaluate only one or a limited number of architectures. As a result, there is insufficient evidence to determine which CNN architecture provides the most reliable performance for distinguishing OL from OSCC under identical experimental conditions.

Second, previous studies often use different datasets, preprocessing techniques, augmentation strategies, training protocols and evaluation metrics, making direct comparison between published models difficult and limiting the objective assessment of CNN architectures for oral histopathological image classification.

Third, while modern architectures such as EfficientNet and ConvNeXt generated promising performance in various medical image analysis tasks, comprehensive comparisons involving both established architectures (ResNet-50, ResNet-101, DenseNet121) and more recent architectures (EfficientNet-B0, EfficientNet-B2, ConvNeXt-Base) on the same histopathological dataset remain limited. The studies discussed in Sections 2.4.2 and 2.5 report competitive individual performance but, as mentioned throughout this chapter, it does not provide a systematic comparison across these six widely used CNN architectures under standardized experimental settings [8][9].

Fourth, although explainable artificial intelligence techniques such as Grad-CAM have increasingly been incorporated into medical image analysis, they are not consistently applied alongside comparative evaluations of multiple CNN architectures, meaning limited attention has been given to assessing whether different architectures focus on clinically meaningful histopathological regions. These research gaps together motivate the present study. This thesis performs a systematic comparative evaluation of six state-of-the-art CNN architectures ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt-Base for the binary classification of OL and OSCC using the NDB-UFES histopathological dataset. All models are trained and evaluated under identical experimental conditions using the same preprocessing pipeline, training strategy and evaluation metrics. Grad-CAM is further implemented to provide visual explanations of model predictions enabling an assessment of whether the learned representations correspond to clinically meaningful histopathological regions.

CHAPTER 3: DATASET

3.1 Dataset Overview

The dataset used in this research is the NDB-UFES dataset which is a publicly available histopathological image dataset created by de Lima et al. through the Oral Diagnosis Project (NDB) at the Federal University of Espírito Santo (UFES), Brazil. The dataset was specifically developed to support research in artificial intelligence-assisted diagnosis of oral lesions and contains histopathological images obtained from patients diagnosed with Oral Leukoplakia (OL) and Oral Squamous Cell Carcinoma (OSCC). Due to its clinical relevance and standardized structure, the dataset has become an important resource for evaluating deep learning approaches in oral pathology [30].

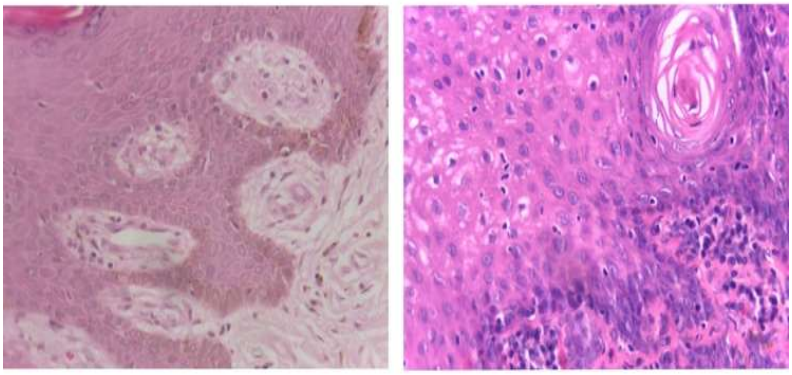

Figure 3: Sample Histopathological Images with Leukoplakia (Left), OSCC (Right).

A significant advantage of the dataset is that it comes from real clinical practice rather than laboratory-generated or synthetic samples. So, the images contain natural variations in staining intensity, tissue morphology, lesion presentation and image acquisition conditions. Such variability introduces realistic challenges that a practical diagnostic system must overcome and then contribute to the reliability of the evaluation process.

Overall, the NDB-UFES dataset serves as the primary experimental foundation of this research and provides the necessary histopathological image data to explore deep learning-based classification of oral lesions.

3.2 Histopathological Image Description

The histopathological images contained within the NDB-UFES dataset were obtained from tissue specimens which were collected during routine oral biopsy procedures. Following biopsy collection, the tissue samples underwent standard histopathological processing which included fixation, sectioning and hematoxylin-eosin (H&E) staining. Hematoxylin-eosin staining is one

of the most widely used techniques in pathology because it highlights cellular and tissue structures with sufficient contrast to help microscopic examination and disease diagnosis.

The images were acquired using a microscope-mounted digital imaging system under standardized laboratory conditions. Both low-magnification and high-magnification views were included in the dataset which allowed visualization of tissue architecture as well as detailed cellular morphology. This multi-scale representation is mainly important in oral pathology as diagnostic decisions often depend on both large-scale structural abnormalities and fine-grained cellular characteristics.

An important characteristic of the dataset is that all images come from real clinical cases diagnosed by experienced oral pathologists. So, the dataset reflects the diversity and complexity occurred in routine pathological practice. Variations in tissue preparation, staining intensity and lesion morphology along with image acquisition conditions contribute to the heterogeneity of the dataset and create a more realistic evaluation environment for deep learning models.

3.3 Dataset Distribution

The NDB-UFES dataset contains a total of 237 histopathological images obtained from 69 patients. Among these images, 146 belong to the OL class and 91 belong to the OSCC class. Table 1 summarizes the overall class distribution used throughout this study. It also includes a Comma Separated Values (CSV) file consisting of information shown in Table 2, along with the unique ID for each patient.

Table 1: Overall Dataset Distribution

Class	**Number of Images**
Oral Leukoplakia	146
Oral Squamous Cell Carcinoma	91
Total	237

Table 2: Dataset Metadata

Category	**Attributes**
Sociodemographic Data	Age, Biological Sex, Complexion
Clinical Data	Use of Tobacco, Consumption of Alcohol, Exposure to Sun, Fundamental Lesion, Biopsy Type, Color of Lesion, Type of Lesion Surface, Lesion Diagnosis

The class distribution shows a moderate degree of imbalance with Oral Leukoplakia accounting for approximately 61.6% of all images and OSCC accounting for approximately 38.4%.

The dataset also shows variability in lesion appearance, tissue morphology, staining characteristics and image acquisition conditions. Such diversity contributes positively to the

experimental evaluation because it challenges the models to learn generalized representations rather than memorizing dataset-specific patterns. So, strong performance on this dataset is more suggestive of a model's ability to generalize to previously unseen clinical cases.

The distribution of images across the two diagnostic categories provides a balanced foundation for evaluating deep learning models while also maintaining the realistic characteristics of clinical oral pathology data. As a result, the dataset offers a suitable benchmark for comparing the performance of different CNN architectures in the task of automated oral lesion classification.

3.4 Train-Validation-Test Split

A thorough dataset partitioning strategy is essential for obtaining reliable estimates of model performance and ensures that experimental results accurately reflect real-world generalization capability. To achieve this objective, the NDB-UFES dataset was divided into three non-overlapping subsets including a training set, a validation set and a test set.

The dataset contains a total of 237 histopathological images distributed across the OL and OSCC classes. Sixty percent of the available images were assigned to the training set, while the remaining forty percent were divided equally between the validation and test sets.

Table 3: Dataset Partitioning Across Training, Validation, and Test Sets

Dataset Partition	**Leukoplakia**	**OSCC**	**Total**
Training Set	86	56	142
Validation Set	30	17	47
Test Set	30	18	48
Total	**146**	**91**	**237**

The training set consisting of 142 images was used for model learning and parameter optimization. During training, network weights were iteratively adjusted through forward propagation, loss calculation, backpropagation and gradient-based optimization.

The validation set consisting of 47 images was used to monitor model performance during the training process. Validation results were used for hyperparameter tuning, learning rate scheduling, model selection and detection of potential overfitting. At the end of each epoch, validation loss and validation accuracy were computed to judge the model's ability to generalize beyond the training data.

The test set consisting of 48 images was kept exclusively for final model evaluation. No test images were used during training or hyperparameter optimization. Following completion of

training, the best-performing model checkpoint was loaded and evaluated on the test set to get the final reported performance metrics.

Maintaining a clear separation between training, validation and testing data is mainly important in medical image analysis because overly optimistic results can arise when information from the evaluation dataset unintentionally influences model development. By reserving the test set exclusively for final evaluation, the study ensures that reported results give an unbiased estimate of model performance on previously unseen data.

The use of independent validation and testing subsets also allows a more reliable comparison among the six evaluated CNN architectures. Since all models were trained and evaluated using identical data partitions, the observed performance differences can be primarily due to architectural characteristics rather than inconsistencies in dataset allocation.

Consequently, the adopted partitioning strategy offers a solid experimental foundation for assessing the effectiveness of deep learning models in oral lesion classification

CHAPTER 4: Methodology

4.1 Overall Research Workflow

This chapter presents the methodology used for the classification of OL and OSCC using histopathological images. The study focuses on comparing the performance of six pretrained CNN architectures under identical experimental conditions to find the most suitable model for distinguishing between these two clinically significant oral lesions.

The used methodology contains several sequential stages, beginning with dataset preparation and image preprocessing, followed by transfer learning using pretrained CNN models, model training, performance evaluation and interpretability analysis. Each CNN architecture was trained using the same dataset, training configuration, optimizer, learning rate and batch size, also evaluation protocol. Maintaining identical experimental settings across all models ensures that the observed performance differences are due to the network architectures rather than variations in the training procedure.

The motivation for this comparative framework comes from the diagnostic challenges related with OL and OSCC. Although histopathological examination continues to be the gold standard for diagnosis, differentiating these lesions can be difficult because they often show similar microscopic characteristics [2]. Furthermore, manual examination is time-consuming and may be affected by inter-observer variability among pathologists [3]. Recent advances in deep learning have showed that CNN can automatically learn discriminative image features directly from medical images which makes them highly suitable for medical image analysis [5]. As a result, evaluating different CNN architectures is important to identify a model that provides reliable classification performance, as well as maintain computational efficiency.

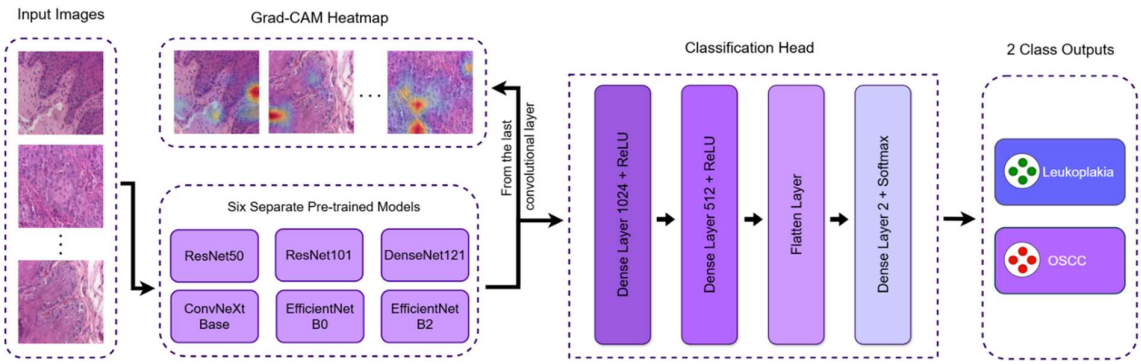

Figure 4: Overall methodology of the oral lesion classification

In Figure 4 method, the histopathological images are first resized to a fixed input dimension and converted into PyTorch tensors before being supplied to the CNN models. Each pretrained architecture acts as a feature extractor while its original ImageNet classification layer is replaced with a custom classifier designed specifically for the binary classification task. During training, only the newly added classifier is updated and the pretrained convolutional backbone

remains frozen cause training deep CNNs with millions of parameters on a limited number of medical images can increase the risk of overfitting and reduce generalization to unseen data. Therefore, pretrained models were used to leverage knowledge learned from a large source dataset and adapt it to the target classification task [31]. This transfer learning strategy allows the models to use generic visual features learned from the large-scale ImageNet dataset by decreasing the risk of overfitting caused by the relatively limited size of the oral histopathology dataset.

After training, the performance of each model is evaluated using multiple quantitative metrics, including accuracy, class-wise recall, Area Under the Receiver Operating Characteristic Curve (AUROC) and Matthews Correlation Coefficient (MCC). These complementary metrics provide a comprehensive assessment of the classification performance from different perspectives rather than relying solely on overall accuracy.

To improve the interpretability of the proposed framework, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to visualize the image regions that heavily contribute to the prediction of each model. These visual explanations help determine whether the models focus on diagnostically meaningful tissue structures instead of irrelevant background regions, hence increases confidence in the reliability of the classification results.

Overall, the methodology presented in this chapter provides a systematic and unbiased framework for comparing multiple state-of-the-art CNN architectures for oral histopathological image classification. The experimental design ensures fair comparison among the selected models and simultaneously evaluates both predictive performance and model interpretability.

4.2 Selection of CNN Architectures

To identify the most suitable CNN architecture for distinguishing OL from OSCC, six pretrained CNN models were selected for comparative evaluation: ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt Base. These architectures were chosen since they represent different generations and design philosophies of deep convolutional networks ranging from classical residual learning and dense connectivity to compound scaling and modernized convolutional designs.

Each architecture has unique structural characteristics that influence feature extraction capability, computational complexity and learning behavior. Evaluating multiple architectures under identical experimental conditions allows a fair comparison of their effectiveness for histopathological image classification and eliminates bias caused by differences in training strategy or hyper parameter settings.

Table 4: Parameter Comparison and Justification of the Selected CNN Architectures.

Model	An approximate of Parameters	Justification of Model Selection
ResNet-50	26 Million	Strong residual learning baseline capable of extracting discriminative histopathological features.
ResNet-101	45 Million	Deeper residual architecture selected to investigate whether increased network depth improves feature representation.
DenseNet121	8.5 Million	Dense feature reuse enables efficient learning of fine cellular and tissue structures with relatively few parameters.
EfficientNet-B0	5.3 Million	Lightweight architecture designed to achieve high accuracy with low computational complexity.
EfficientNet-B2	9.2 Million	Larger EfficientNet variant used to evaluate the effect of compound scaling on classification performance.
ConvNeXt-Base	89 Million	Modern CNN architecture incorporating recent architectural improvements for enhanced feature representation.

The selected architectures span a wide range of model sizes from approximately 5.3 million parameters in EfficientNet-B0 to nearly 89 million parameters in ConvNeXt Base. This diversity allows the investigation of how network depth, architectural design and model complexity influence classification performance on histopathological images. Such comparisons are mainly relevant in medical imaging because larger models do not always provide superior performance when training data are limited. Therefore, evaluating models with substantially different complexities helps to identify an architecture that achieves an appropriate balance between predictive accuracy, computational efficiency and generalization.

4.2.1 ResNet-50 and ResNet-101

ResNet-50 was proposed by He et al. [26], it is a 50-layer CNN that introduced the concept of residual learning to overcome the degradation and vanishing gradient problems faced when training very deep neural networks. Instead of learning a direct mapping between the input and output of each layer, ResNet learns a residual function that is added to the original input through an identity shortcut connection.

The shortcut connection allows gradients to flow directly through the network during backpropagation which makes optimization easier and allows substantially deeper networks to

be trained without suffering from degradation. ResNet-50 is constructed using bottleneck residual blocks consisting of successive 1×1, 3×3 and 1×1 convolutional layers with batch normalization and ReLU activation applied throughout the network. Global average pooling is performed before the final classification layer to produce a compact feature representation.

ResNet-50 was selected in this study as it provides a well-established baseline for medical image classification. Its residual connections allow effective learning of complex visual features, as well as maintain stable optimization during training. Many medical imaging studies have demonstrated that ResNet-50 performs reliably across a wide range of histopathological image classification tasks which makes it an appropriate reference architecture for comparative evaluation.

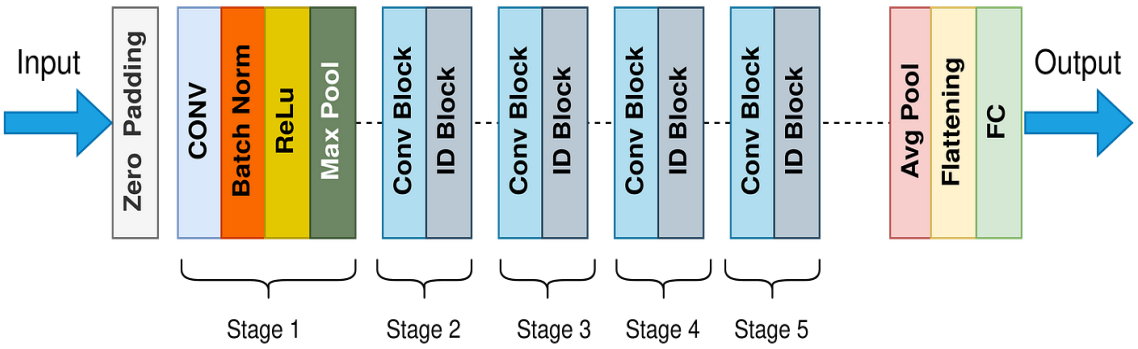

Figure 5: ResNet Model Architecture [32]

On the other hand, ResNet-101 is a deeper variant of the ResNet architecture that extends the residual learning framework to 101 convolutional layers [26]. It follows the same residual learning principle as ResNet-50 but contains considerably more bottleneck residual block which enables the network to learn more abstract and hierarchical feature representations.

The deeper architecture allows successive convolutional layers to capture information at multiple levels of abstraction. Early layers primarily detect simple visual features like edges and color transitions, whereas deeper layers progressively learn complex tissue morphology and cellular organization along with architectural patterns. This hierarchical representation is particularly useful for histopathological image analysis because distinguishing leukoplakia from OSCC often requires recognition of both fine cellular characteristics and larger tissue-level structures.

Like ResNet-50, ResNet-101 employs identity shortcut connections that facilitate gradient propagation throughout the network by reducing optimization difficulties connected with increasing network depth. Although the deeper architecture introduces additional computational cost, it also provides greater representational capacity for learning subtle pathological features.

ResNet-101 was included in this study to investigate whether increasing network depth improves classification performance for oral histopathological images compared with the shallower ResNet-50 architecture.

4.2.2 DenseNet121

DenseNet121 was introduced by Huang et al. [27], it employs a fundamentally different connectivity strategy from conventional CNNs. Instead of connecting each layer only to the next layer, DenseNet establishes direct connections between every layer within a dense block. So the output of each layer is concatenated with the outputs of all preceding layers before being forwarded to subsequent layers.

Dense connectivity offers several important advantages. First, it improves gradient propagation throughout the network which makes optimization more stable. Second, previously learned features are reused rather than repeatedly relearned which result in greater parameter efficiency. Third, the network preserves both low-level and high-level information simultaneously which is beneficial for histopathological image analysis where fine cellular details remain important throughout the classification process. DenseNet variants have demonstrated consistently strong performance in multiple oral lesion studies including both clinical photograph classification and OSCC histopathological detection tasks [19][21].

DenseNet121 contains approximately 8.5 million parameters which is considerably fewer than ResNet-50 and ResNet-101, but still it often achieves competitive performance because of its efficient feature reuse strategy. For this reason, DenseNet121 was selected to evaluate whether dense feature propagation can effectively capture subtle morphological differences between leukoplakia and OSCC while maintaining relatively low computational complexity.

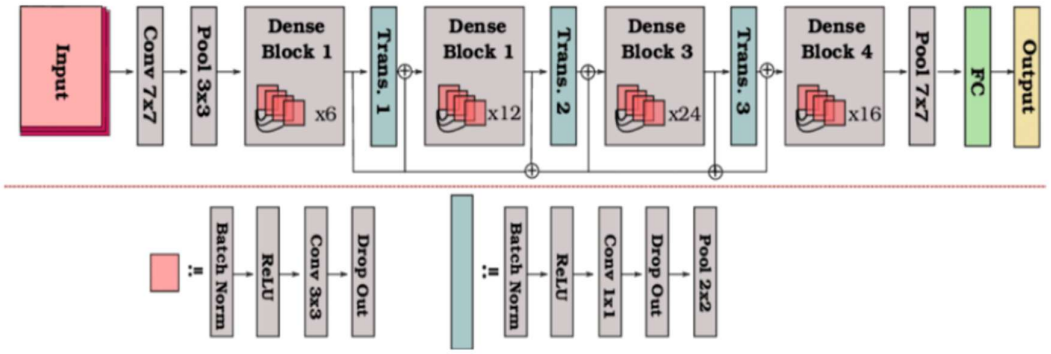

Figure 6: Densenet121 Model Architecture [33]

4.2.3 EfficientNet-B0 and EfficientNet-B2

EfficientNet-B0, proposed by Tan and Le [28], introduced the concept of compound model scaling which simultaneously scales network depth, width and input image resolution using a single scaling coefficient. Instead of increasing only one dimension of the network, EfficientNet balances all three dimensions to achieve improved accuracy with substantially fewer parameters.

EfficientNet-B0 uses MBConv (Mobile Inverted Bottleneck Convolution) blocks together with squeeze-and-excitation modules which enables efficient feature extraction while maintaining low computational cost. Despite containing only approximately 5.3 million parameters, EfficientNet-B0 has showed strong performance across numerous image classification tasks.

In this study, EfficientNet-B0 was selected as a lightweight architecture to evaluate whether efficient network design can achieve competitive performance for oral histopathological image classification with fewer parameters than deeper residual networks.

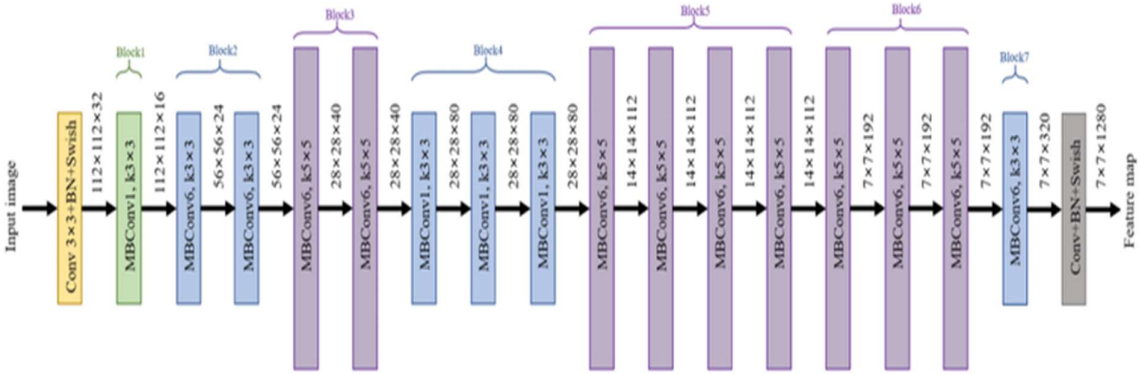

Figure 7: EfficientNet-B0 Model Architecture [34]

EfficientNet-B2 is a higher-capacity variant of the EfficientNet family proposed by Tan and Le [28]. Similar to EfficientNet-B0, it is based on the compound scaling principle where the network depth, width and input resolution are increased simultaneously in a balanced manner. In comparison to EfficientNet-B0, EfficientNet-B2 employs a deeper architecture, wider convolutional layers and a higher input resolution which enables the network to learn more detailed and discriminative feature representations, as well as maintain relatively high computational efficiency.

The architecture retains the Mobile Inverted Bottleneck Convolution (MBConv) blocks and squeeze-and-excitation (SE) modules used in EfficientNet-B0. MBConv blocks reduce computational complexity by replacing standard convolutions with depthwise separable convolutions while the SE modules improve feature representation by adaptively recalibrating channel-wise feature responses. Together, these components allow EfficientNet-B2 to capture both local texture information and higher-level semantic features efficiently.

EfficientNet-B2 contains approximately 9.2 million trainable parameters which makes it larger than EfficientNet-B0 but still considerably smaller than ResNet-101 and ConvNeXt-Base. The increased model capacity provides a greater ability to capture subtle histopathological characteristics like variations in epithelial organization, nuclear morphology, keratinization patterns and tissue architecture, that are important for differentiating OL from OSCC.

EfficientNet-B2 was included in this comparative study to investigate whether increasing the capacity of the EfficientNet architecture through compound scaling provides measurable improvements in classification performance while maintaining relatively low computational requirements. Comparing EfficientNet-B2 with EfficientNet-B0 also allows evaluation of how model scaling influences feature learning for histopathological image classification.

4.2.4 ConvNeXt-Base

ConvNeXt-Base was introduced by Liu et al. [29], this represents a modern CNN architecture that incorporates several design principles inspired by Vision Transformers (ViTs) while retaining the efficiency and simplicity of convolutional operations. Instead of replacing CNNs with transformer architectures, ConvNeXt demonstrates that conventional convolutional networks can achieve competitive performance through carefully designed architectural improvements.

Several modifications distinguish ConvNeXt from traditional CNN architectures. First, standard ReLU activation functions are replaced by the Gaussian Error Linear Unit (GELU) which provides smoother activation behavior and improves optimization. Second, batch normalization is replaced with layer normalization which improves feature normalization throughout the network. Third, the architecture uses large-kernel depthwise convolutions using 7×7 kernels which allows the network to capture a broader spatial context than conventional 3×3 convolutions. Finally, ConvNeXt adopts an inverted bottleneck structure that expands feature dimensions before projection improving feature representation while maintaining computational efficiency.

These architectural modifications significantly increase the receptive field of the network which allows ConvNeXt to capture long-range spatial dependencies within histopathological images. Such characteristics are mainly beneficial for oral tissue analysis since the diagnosis often depends not only on individual cellular morphology but also on the spatial relationships between epithelial layers, connective tissue, inflammatory infiltrates and invasive tumor regions.

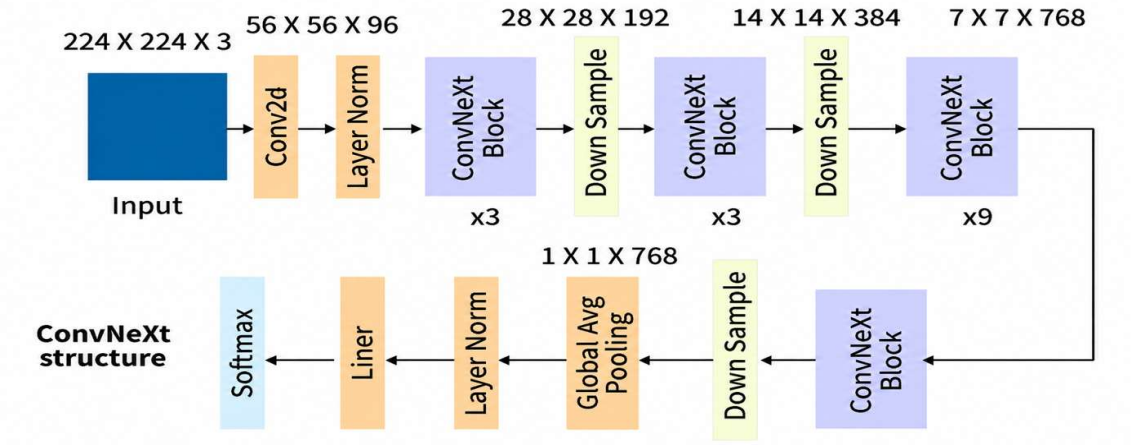

Figure 8: ConvNext Model Architecture [35]

ConvNeXt-Base contains approximately 89 million parameters which makes it the largest and most computationally demanding architecture evaluated in this study. Its relatively higher capacity enables the learning of highly complex visual representations, however larger models also carry an increased risk of overfitting when trained on relatively small medical datasets. So, ConvNeXt-Base was included to examine whether recent advances in CNN architecture

provide measurable advantages over more established models for histopathological classification of oral lesions.

4.3 Hyperparameter Configuration

To ensure consistency throughout the comparative evaluation, all CNN architectures were trained using the same hyperparameter configuration. Keeping the training settings consistent provides a fair basis for comparing the performance of the different architectures.

Table 5: Training Hyperparameters Used for CNN Model Development.

Hyperparameter	**Value**
Input image size	224 × 224
Number of classes	2
Transfer learning	ImageNet pretrained weights
Backbone	Frozen
Optimizer	Adam
Loss function	CrossEntropyLoss
Learning rate	0.0001
Batch size	32
Number of epochs	50
Best model selection	Lowest validation loss

Each hyperparameter was selected based on commonly adopted practices in transfer learning for medical image classification. The relatively small learning rate facilitates stable optimization, while the batch size of 32 provides efficient GPU utilization without exceeding memory limitations. Training for 50 epochs allows sufficient optimization and validation monitoring helps prevent overfitting.

4.4 Evaluation Metrics

The performance of each CNN architecture was evaluated using four quantitative metrics: Accuracy, Recall, Area Under the Receiver Operating Characteristic Curve (AUROC) and Matthews Correlation Coefficient (MCC). These metrics were selected because they provide complementary information about model performance and together offer a comprehensive assessment of binary classification models. As the dataset contains two disease classes, OL and OSCC rather than a healthy control group, the evaluation focuses the model's ability to correctly distinguish between the two pathological conditions.

As our dataset is relatively small and not perfectly balanced, so relying on Accuracy alone may not give a complete picture of model performance. Accuracy shows the overall percentage of correct predictions, but it can be affected by the class distribution and may not clearly show

how well the model performs for each disease class. As a result AUROC was also used to evaluate how well the models can distinguish between OL and OSCC across different classification thresholds. This gives a broader view of the model's discrimination ability rather than relying on only one decision threshold.

Class-wise Recall was used to determine how many actual OL and OSCC cases were correctly identified by each model. This is particularly important for OSCC because a false negative means that an actual cancer case is incorrectly classified as OL.

Finally, MCC was used to assess the overall classification performance because it considers TP, TN, FP, and FN together. This provides a more balanced evaluation of the model than Accuracy alone.

In the following equations, TP (True Positive) represents correctly classified OSCC images and TN (True Negative) represents correctly classified leukoplakia images. FP (False Positive) represents leukoplakia images incorrectly classified as OSCC and FN (False Negative) represents OSCC images incorrectly classified as leukoplakia.

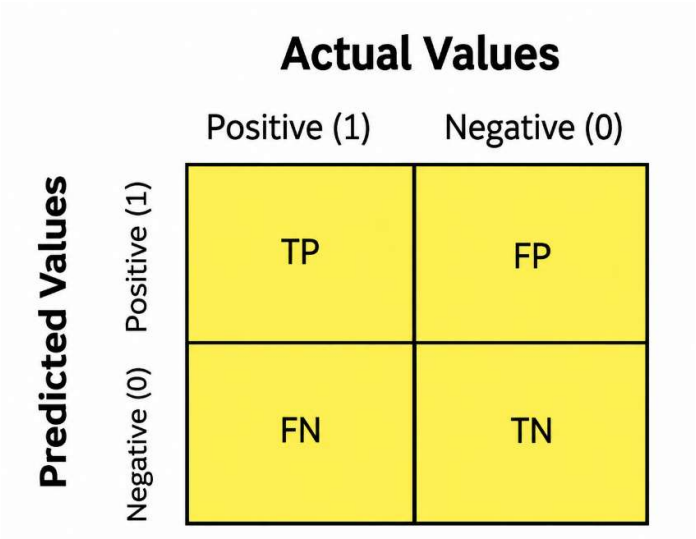

Figure 9: Confusion matrix showing the quadrants for true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

4.4.1 Accuracy

Accuracy represents the proportion of correctly classified samples among all evaluated samples. It provides an overall indication of the classification performance and is calculated as

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

Although accuracy is one of the most commonly reported performance measures, it should not be evaluated independently because it may produce misleading conclusions when class distributions are imbalanced. Therefore, additional evaluation metrics were also included to provide a more balanced assessment of model performance.

4.4.2 Recall

Recall measures the ability of the classifier to correctly identify samples belonging to each class. In this study, recall was calculated separately for leukoplakia and OSCC because the dataset contains only these two clinically relevant classes and does not include a normal or healthy class. Therefore, class-wise recall provides a clearer assessment of the model's ability to correctly identify each target condition, particularly its ability to detect OSCC and leukoplakia independently.

The recall for the OSCC class is defined as

$$\text{Recall}_{OSCC} = \frac{TP}{TP + FN}$$

this represents the proportion of malignant lesions that are correctly identified.

Similarly, the recall for the leukoplakia class is calculated as

$$\text{Recall}_{Leukoplakia} = \frac{TN}{TN + FP}$$

A high OSCC recall is particularly important in clinical practice because failing to detect malignant lesions may delay treatment and negatively affect patient outcomes. Meanwhile, high recall for leukoplakia shows that premalignant lesions are correctly identified without being unnecessarily classified as carcinoma.

4.4.3 Area Under the Receiver Operating Characteristic Curve (AUROC)

The Area Under the Receiver Operating Characteristic Curve (AUROC) evaluates the ability of a classifier to distinguish between the two classes across all possible decision thresholds. Unlike accuracy, which is calculated using a single classification threshold, AUROC measures the overall discriminative ability of the model independent of a specific threshold.

The Receiver Operating Characteristic (ROC) curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different decision thresholds. The area under this curve was computed using the trapezoidal rule,

$$AUROC = (\frac{TPR_{i+1} + TPR_i}{2}) \times (FPR_{i+1} - FPR_i)$$

where each point on the ROC curve corresponds to a different classification threshold.

AUROC values range between 0 and 1, where a value of 0.5 indicates random classification and a value of 1.0 represents perfect discrimination. Higher AUROC values indicate better separation between leukoplakia and OSCC across different operating thresholds.

4.4.4 Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) provides a balanced evaluation of binary classification performance by simultaneously considering all four components of the confusion matrix. Unlike accuracy, MCC remains informative even in unequal class distributions.

It is calculated as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

MCC considers true positives, true negatives, false positives and false negatives simultaneously. It provides a reliable overall assessment of classifier performance and is mainly suitable for medical image classification tasks.

4.5 Grad-CAM Visualization

Although quantitative evaluation metrics indicate how accurately a model performs, they do not explain why the model makes a particular prediction. For medical image analysis, interpretability is an important consideration because clinicians must be able to understand whether the model is focusing on diagnostically meaningful tissue regions.

To improve the interpretability of the CNN models, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed as a post hoc visualization technique. Grad-CAM generates a heatmap highlighting the image regions that contribute most strongly to the predicted class.

In this study, Grad-CAM [25] was applied after model training to visualize the regions that influenced the classification decisions of each CNN architecture. The generated heatmaps enable qualitative assessment of whether the models focus on histopathological structures that are considered diagnostically relevant by pathologists, such as epithelial abnormalities, cellular atypia, keratinization patterns, and invasive tissue regions, rather than irrelevant background areas.

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Results Evaluation Workflow

This chapter presents the results obtained from training and evaluating six CNN architectures: ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt Base for the binary classification of OL and OSCC using histopathological images from the NDB-UFES dataset. All models were trained and tested under the same experimental settings described in Chapter 4.

To provide a comprehensive evaluation, the results are examined from several perspectives. First, the overall classification performance of each model is compared using Accuracy, class-wise Recall for leukoplakia and OSCC, Area Under the Receiver Operating Characteristic Curve (AUROC) and Matthews Correlation Coefficient (MCC). These metrics give a balanced assessment of performance and identify strengths and weaknesses beyond simple accuracy values.

The analysis then moves beyond overall metrics by examining the confusion matrices of the two highest performing models. This allows a closer assessment of the types of errors made by each model which helps determine how effectively they distinguish between the two disease classes. In case of oral cancer diagnosis, this is mainly important because incorrectly classifying an OSCC case as leukoplakia may have more serious clinical consequences than the opposite type of error.

The Receiver Operating Characteristic (ROC) curves of all six models are also compared to assess their ability to separate the two classes across a range of classification thresholds. Unlike accuracy which depends on a single decision threshold, ROC analysis provides a threshold-independent view of model performance and offers additional insight into discriminative capability.

Along with predictive performance, model complexity is considered. The relationship between parameter count and classification accuracy is tested to see whether larger and more computationally demanding architectures actually provide a meaningful performance advantage. Understanding this trade-off is important because highly complex models may not always be the most practical choice, especially in medical applications where computational efficiency can also be a consideration.

Finally, Grad-CAM visualizations are analyzed to investigate the regions of the histopathological images that resulted the models' predictions. While high classification accuracy is important, interpretability is equally valuable in medical image analysis. Examining the visual attention patterns of the models helps to understand whether predictions are based on clinically meaningful tissue structures or on unrelated image characteristics.

The chapter ends by discussing the overall findings and relating them back to the research objectives established in Chapter 1. Through quantitative evaluation, error analysis, model

complexity assessment and visual interpretability analysis, this chapter provides a comprehensive understanding of the strengths and limitations of the evaluated CNN architectures for oral lesion classification.

5.2 Overall Classification Performance

The overall classification performance of the six evaluated CNN architectures is presented in Table 6. The results are generated using Accuracy, class-wise Recall for OL and OSCC, AUROC and MCC. Together, these metrics provide a more complete picture of model performance than accuracy alone, which allows both classification effectiveness and class-specific behavior to be examined.

Table 6: Overall Classification Performance of Baseline Models

Model	Accuracy (%)	Leukoplakia Recall (%)	OSCC Recall (%)	AUROC	MCC
ResNet-50	81.25	80.00	83.33	0.9037	0.6181
ResNet-101	**89.58**	90.00	**88.89**	**0.9481**	**0.7810**
DenseNet121	83.33	80.00	**88.89**	0.9407	0.6693
EfficientNet-B0	79.17	86.67	66.67	0.9407	0.5477
EfficientNet-B2	87.50	**93.33**	77.78	0.9185	0.7303
ConvNeXt Base	83.33	83.33	83.33	0.9241	0.6547

The results in Table 6 show noticeable differences in performance among the six evaluated architectures. Among the evaluated models, ResNet-101 achieved the strongest overall performance, with an Accuracy of 89.58%, a Leukoplakia Recall of 90.00%, an OSCC Recall of 88.89%, an AUROC of 0.9481, and an MCC of 0.7810. These results indicate that ResNet-101 provided a strong and relatively balanced performance across the evaluated metrics.

EfficientNet-B2 emerged as the second-best performer overall. It achieved an Accuracy of 87.50% and an MCC of 0.7303, both of which were second only to ResNet-101. Interestingly, EfficientNet-B2 produced the highest Leukoplakia Recall among all models at 93.33% which suggest that it was particularly effective at identifying leukoplakia cases. However, this strength was accompanied by a lower OSCC Recall of 77.78%, indicating that some carcinoma cases were missed during classification.

DenseNet121 and ConvNeXt Base both achieved an Accuracy of 83.33%, hence placed them in the middle of the performance ranking. Although their overall accuracies were identical, their class-wise behavior differed. DenseNet121 showed stronger performance in detecting OSCC, achieving an OSCC Recall of 88.89% and its Leukoplakia Recall remained at 80.00%.

On the other hand, ConvNeXt Base showed perfectly balanced recall values of 83.33% for both classes. This suggests that while the two models reached the same overall accuracy, they approached the classification task in different ways.

ResNet-50 achieved an Accuracy of 81.25% placing it below DenseNet121 and ConvNeXt Base but still above EfficientNet-B0. Its recall values were relatively balanced between the two classes with 80.00% for Leukoplakia and 83.33% for OSCC. Although the model performed reasonably well, it was consistently outperformed by the deeper ResNet-101 architecture across all reported metrics.

EfficientNet-B0 recorded the lowest overall Accuracy at 79.17% and the lowest MCC at 0.5477. While its Leukoplakia Recall of 86.67% was respectable, its OSCC Recall dropped to only 66.67%. This makes it the weakest OSCC detection performance among all six architectures. This imbalance had a significant impact on its overall effectiveness and highlights the challenge the model faced when distinguishing malignant cases from leukoplakia.

The superior performance of ResNet-101 can likely be due to its deeper residual architecture. The model uses skip connections that allow information to flow directly across layers which help to address the degradation problem that often affects very deep neural networks. As ResNet-101 contains substantially more layers than ResNet-50, it is able to learn richer hierarchical feature representations ranging from low-level texture information to more complex tissue-level patterns related with disease progression. Histopathological differences between leukoplakia and OSCC are often subtle and may occur at multiple spatial scales, making a deep architecture particularly advantageous for this task.

What makes ResNet-101 especially notable is that its performance advantage was not limited to a single metric. It achieved the highest Accuracy, AUROC and MCC along with maintaining strong Recall values for both classes. This consistency suggests that the model was not simply favoring one class over the other but was able to distinguish both leukoplakia and OSCC reliably. The high MCC value further supports this observation, as MCC is generally regarded as one of the most informative measures for binary classification because it considers all elements of the confusion matrix rather than focusing only on correctly classified samples.

EfficientNet-B2 showed a different but equally interesting performance profile. Although it did not surpass ResNet-101 overall, it achieved the highest Leukoplakia Recall among all models. One possible explanation can be due to EfficientNet's compound scaling strategy, which balances network depth, width and resolution simultaneously. This design allows the model to capture fine-grained visual details efficiently. Since leukoplakia often presents with subtle epithelial and surface-level changes rather than the more obvious structural disruption linked with OSCC, EfficientNet-B2 may have been particularly effective at learning these finer histopathological patterns.

However, this advantage came with a trade-off. While EfficientNet-B2 identified leukoplakia cases exceptionally well, its OSCC Recall was noticeably lower than that of ResNet-101. From a clinical perspective, this difference is important because failing to identify a carcinoma case may have more serious consequences than incorrectly classifying a benign lesion as malignant. Therefore, although EfficientNet-B2 performed strongly overall, its lower OSCC detection rate limits its practical usefulness compared to ResNet-101.

At the lower end of the performance spectrum, EfficientNet-B0's results suggest that model size and representational capacity also play a huge role in this classification task. As the smallest architecture evaluated in this study, EfficientNet-B0 contains fewer parameters and thus a more limited ability to learn complex histopathological features. This limitation appears most clearly in its OSCC Recall which was noticeably lower than that of the other models. While the architecture remains computationally efficient, the lower classification performance indicates that its capacity may not be enough for capturing the subtle morphological differences needed for reliable oral lesion classification.

A similar pattern can be observed when comparing ResNet-50 and ResNet-101. Although both models share the same residual learning framework, the deeper ResNet-101 consistently outperformed ResNet-50 across every evaluation metric. This comparison suggests that, within the context of the NDB-UFES dataset, additional network depth contributed meaningfully to improved feature extraction and classification performance among these two.

5.3 Confusion Matrix Analysis

While overall performance metrics provide a useful summary of model behavior, they do not reveal the specific types of classification errors made by a model. Due to this, the confusion matrices of the two best-performing architectures, ResNet-101 and EfficientNet-B2 were examined in greater detail. Confusion matrix analysis is mainly important in medical image classification because different types of errors can have very different clinical consequences.

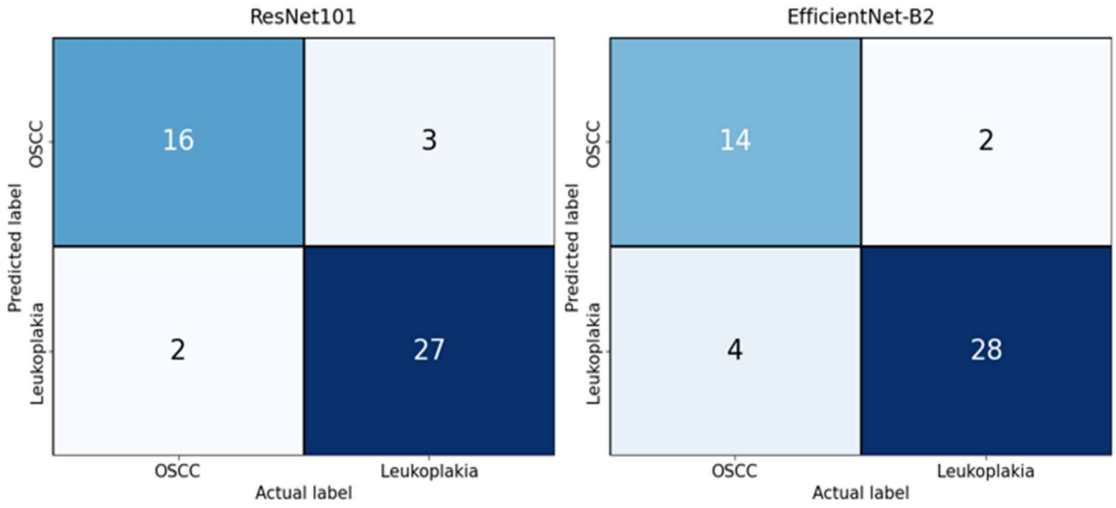

Figure 10: Confusion Matrices of ResNet-101 and EfficientNet-B2

From Figure 10 we can see that on the test set of 48 images, ResNet-101 produced 27 true negatives, 16 true positives, 3 false positives and 2 false negatives. These results correspond to the overall Accuracy of 89.58% reported in Table 6. Under the same testing conditions, EfficientNet-B2 produced 28 true negatives, 14 true positives, 2 false positives and 4 false negatives resulting in an overall Accuracy of 87.50%.

At first glance, the performance of the two models appears quite similar. Both models correctly classified the majority of test samples and generated only a small number of misclassifications. The difference in true negatives and false positives between the two architectures is relatively small which suggest that both models were effective at identifying leukoplakia cases. However, a closer examination reveals a more important distinction in their handling of OSCC cases.

The most clinically significant difference lies in the number of false negatives. ResNet-101 produced only 2 false negatives but EfficientNet-B2 produced 4. In practical terms, a false negative occurs when an image belonging to the OSCC class is incorrectly classified as leukoplakia. This type of error is concerning because it represents a missed cancer diagnosis.

From a clinical perspective, false negatives and false positives do not carry the same level of risk. A false positive occurs when a leukoplakia case is incorrectly classified as OSCC. Although such error may lead to additional investigations, unnecessary concern for the patient or increased healthcare costs, it is generally less harmful as further clinical examination and pathological review would likely identify the mistake.

However, a false negative can have much more serious consequences. If a carcinoma case is incorrectly identified as leukoplakia, there is a risk that treatment could be delayed or that the patient may not receive the level of clinical attention required for a malignant lesion. Since the prognosis of OSCC is strongly influenced by the stage at which the disease is detected and treated, reducing false negatives is an important objective of computer-aided diagnostic systems.

When viewed from this perspective, the advantage of ResNet-101 becomes even more clear. Although the difference in overall Accuracy between ResNet-101 and EfficientNet-B2 is only 2.08 percentage points, the difference in false negatives is much more clinically meaningful. ResNet-101 reduced the number of missed OSCC cases by half compared to EfficientNet-B2 representing a substantial improvement in diagnostic reliability.

Another way to interpret these results is through the class-wise Recall values reported earlier. ResNet-101 achieved an OSCC Recall of 88.89%, whereas EfficientNet-B2 achieved 77.78%. The confusion matrices provide a direct visual explanation for this difference. As ResNet-101 generated fewer false negatives, it was able to identify a greater proportion of carcinoma cases correctly. This reinforces the observation that ResNet-101 maintained a stronger balance between detecting leukoplakia and detecting OSCC.

The confusion matrix analysis also highlights an important lesson regarding model selection in medical image analysis. If Accuracy alone had been used to compare the two models, the performance gap might appear relatively small. However, investigating the confusion matrices reveals differences that are not immediately visible through Accuracy values alone. The decrease in false negatives achieved by ResNet-101 provides additional evidence that it is the more suitable architecture for this classification task.

Overall, the confusion matrix analysis supports the findings presented in the previous section. Both ResNet-101 and EfficientNet-B2 demonstrated strong classification performance but ResNet-101 showed a clearer advantage when clinically important error types were considered. Its ability to reduce false negatives and maintaining high overall accuracy strengthens its position as the most reliable model among the six architectures evaluated in this study.

5.4 ROC Curve Analysis

Although accuracy and recall provide valuable information about classification performance, they are calculated using a single decision threshold, therefore do not fully capture a model's ability to distinguish between classes across different operating conditions. To obtain a threshold-independent evaluation, Receiver Operating Characteristic (ROC) curve analysis was performed for all six CNN architectures. The ROC curves are presented in Figure 11 and the corresponding AUROC values were previously reported in Table 5.

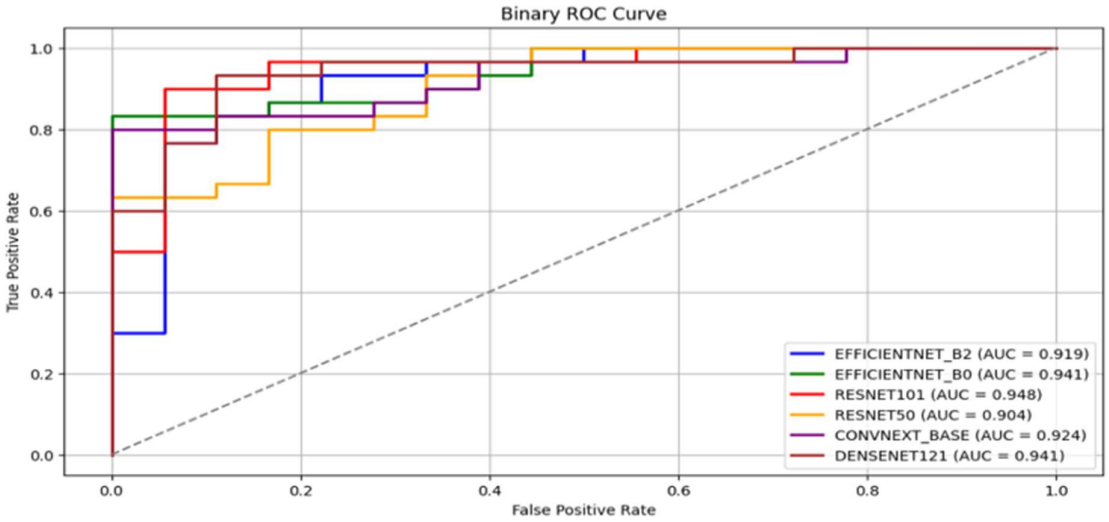

Figure 11: ROC Curves of the Six Evaluated CNN Model

The ROC curves provide a visual representation of the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) at different classification thresholds. In general, a curve that remains closer to the upper-left corner of the graph means stronger discriminative capability because it achieves high sensitivity while maintaining a low false positive rate. The Area Under the Receiver Operating Characteristic Curve (AUROC) proves this behavior into a single numerical value where values closer to 1 indicate better class separation.

Among the evaluated architectures, ResNet-101 demonstrated the strongest ROC performance, achieving the highest AUROC of 0.9481. Its ROC curve remained close to the upper-left region, indicating a strong ability to distinguish between OL and OSCC across different classification thresholds. This finding is consistent with the results presented in Table 6, where ResNet-101 also achieved the highest Accuracy (89.58%) and MCC (0.7810). The consistency across these evaluation metrics indicates that ResNet-101 provided the most reliable overall discriminative performance among the evaluated models.

The remaining architectures generally demonstrated strong class-separation capability, with AUROC values ranging from 0.9185 to 0.9407. DenseNet121 and EfficientNet-B0 achieved AUROC values of 0.9407, followed by ConvNeXt Base (0.9241) and EfficientNet-B2 (0.9185). Although these models showed differences in Accuracy and class-wise Recall, their

relatively high AUROC values indicate that they were generally capable of distinguishing leukoplakia from OSCC across different classification thresholds. In particular, the high AUROC of EfficientNet-B0 despite its lower Accuracy suggests that its overall class-separation capability was stronger than its single-threshold classification performance might indicate.

ResNet-50 exhibited the lowest ROC performance, with an AUROC of 0.9037. Although this still represents reasonable discriminative capability, it was lower than the AUROC values obtained by the other five architectures. This result, together with its comparatively lower Accuracy (81.25%) and MCC (0.6181), indicates that ResNet-50 provided the weakest overall performance among the evaluated models.

Overall, the ROC analysis further supports ResNet-101 as the best-performing architecture in this study. Its highest AUROC, combined with its superior Accuracy and MCC, demonstrates consistent classification and class-separation capability. The results also show that AUROC provides complementary information to threshold-dependent metrics such as Accuracy, allowing the discriminative behavior of the models to be evaluated across a range of classification thresholds.

5.5 Parameter Count versus Performance

While classification accuracy is often the primary metric used to evaluate deep learning models, it is not the only factor that should be taken into consideration during architecture selection for practical deployment. Computational complexity is another important factor, particularly in medical applications where available hardware resources may be limited. To investigate this matter, the relationship between model size and classification performance was examined by comparing the number of trainable parameters in each architecture with its corresponding test Accuracy in figure 12. In the figure, each model is represented as a bubble whose position reflects its parameter count and test Accuracy, while the bubble size is also scaled according to the total number of parameters.

ResNet-101 achieved the highest Accuracy among the evaluated models, reaching 89.58% with approximately 45.12 million trainable parameters. Although ResNet-101 has a relatively high parameter count, its higher classification performance indicates that the additional model capacity was useful for this classification task. This makes ResNet-101 a relatively balanced model in terms of classification performance and model complexity within the evaluated architectures.

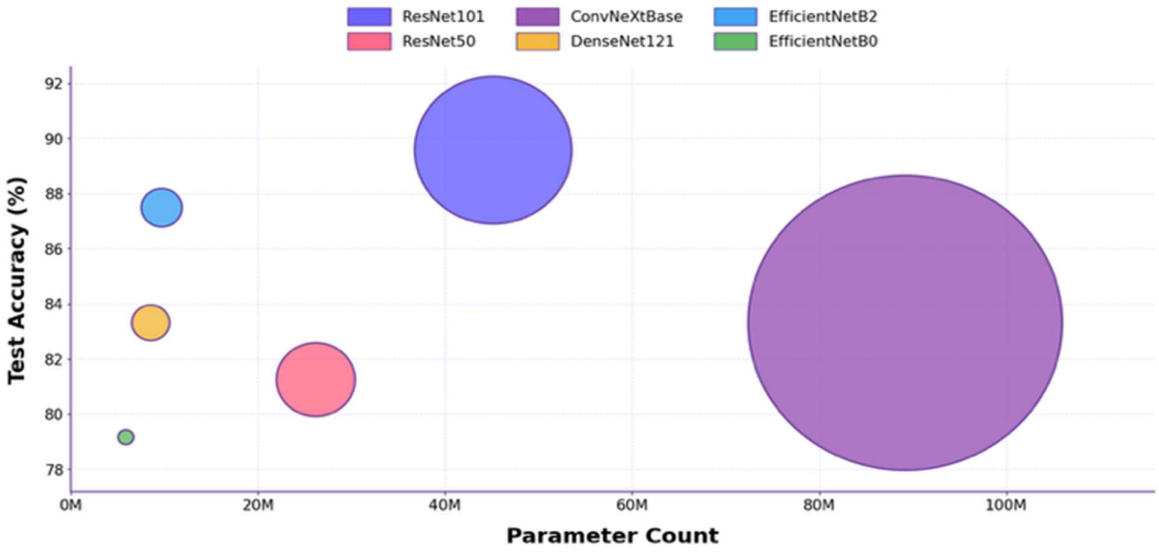

Figure 12: Parameter Count versus Classification Accuracy Bubble Chart

The remaining models show that parameter count does not consistently correspond to classification performance. In particular, some models with fewer parameters achieved competitive performance, while increasing model size did not always result in a corresponding improvement in Accuracy. These results suggest that the effectiveness of a model depends not only on the number of parameters but also on how its architecture represents and learns the relevant histopathological features.

ConvNeXt Base had the highest parameter count, with approximately 89.14 million trainable parameters, but achieved an Accuracy of 83.33%. Despite having substantially more parameters than the other evaluated models, its Accuracy was lower than that of ResNet-101. This indicates that a larger model does not necessarily provide better classification performance always, particularly when working with a relatively small dataset. Overall, the parameter-performance comparison suggests that model size alone is not a sufficient indicator of classification performance and an appropriate balance between model complexity and predictive performance should be considered.

5.6 Grad-CAM Interpretability Analysis

High classification accuracy alone does not necessarily mean that a model is making decisions for the right reasons. In medical image analysis, understanding why a model arrives at a particular prediction is almost as important as the prediction itself. A model may achieve strong quantitative performance while relying on irrelevant image regions, background artifacts or dataset-specific biases. Due to this, interpretability analysis was performed using Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the image regions that contributed most strongly to the classification decisions made by each CNN architecture.

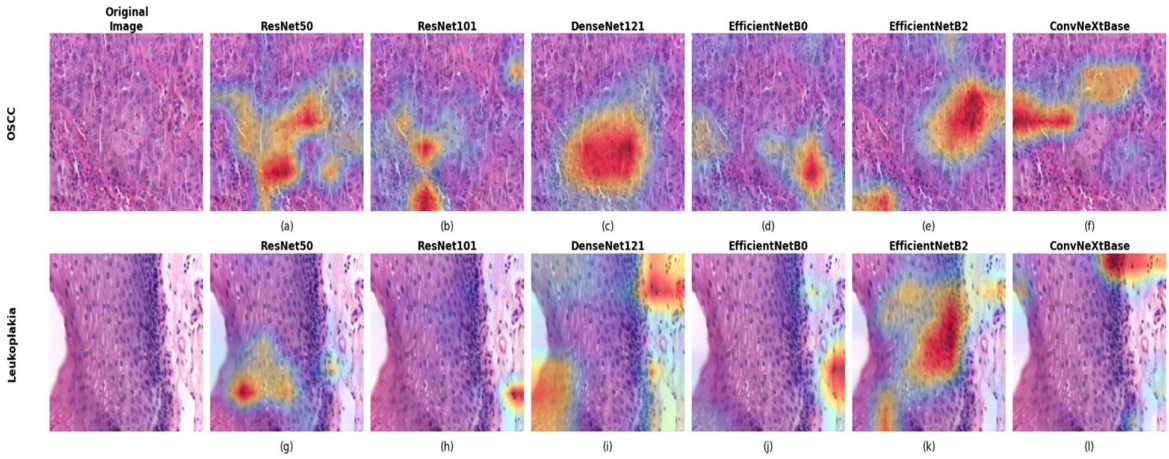

Figure 13: Grad-CAM Visualizations for OSCC and Leukoplakia Across All Six CNN Models. (a, b, c, d, e, f) OSCC and (g, h, i, j, k, l) Leukoplakia histopathological images with corresponding Grad-CAM heatmaps generated by ResNet-50, ResNet-101, DenseNet121, EfficienNet-B0, EfficientNet-B2 and ConvNext-Base, respectively

Grad-CAM is a widely used explainable artificial intelligence (XAI) technique that generates a heatmap which highlights the regions of an image that mostly influence a model's prediction. The method works by computing the gradients flowing into the final convolutional layer of a trained CNN and projecting this information back onto the input image. Regions that contribute strongly to the prediction appear as high-intensity areas in the heatmap and less important regions appear with lower intensity. By overlaying the heatmap on the original histopathological image, it becomes possible to visually inspect if the model is focusing on clinically meaningful tissue structures.

To evaluate the interpretability of the six CNN architectures, Grad-CAM visualizations were generated for representative samples from both diagnostic classes. Figure 13 presents the resulting heatmaps for OSCC and OL images across ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2 and ConvNeXt Base. In the OSCC samples, the activation maps generally show stronger responses around tissue regions containing noticeable cellular and structural variations. In comparison, the Leukoplakia samples generally show more distributed activation patterns across the tissue regions. Differences in the intensity and location of the highlighted regions can also be observed among the different CNN architectures.

Some variation between the architectures can also be observed. ResNet-50 and EfficientNet-B0 generally produced broader activation regions, while ResNet-101 and EfficientNet-B2 showed relatively more focused patterns. These differences suggest that the models may attend to different tissue regions when making their predictions. Across the representative samples, the heatmaps were mainly observed within the tissue regions, with relatively limited activation in the surrounding background.

Overall, the Grad-CAM visualizations provide an additional perspective on the model predictions by showing the regions that may have contributed to the classification. The patterns observed in the representative images also show that different architectures may focus on somewhat different regions while producing their predictions. Thus, Grad-CAM complements the evaluation by providing a visual view of the model's prediction behavior.

5.6 Summary

Overall, ResNet-101 achieved the best overall performance among the six evaluated CNN models for OL and OSCC classification. It achieved the highest Accuracy, AUROC, and MCC, while maintaining high and balanced Recall for both classes. EfficientNet-B2 achieved the highest Leukoplakia Recall, but its OSCC Recall was lower than ResNet-101, with four false-negative OSCC cases compared with two for ResNet-101.

The comparison also showed that a higher parameter count did not necessarily result in better performance. ConvNeXt Base had the largest number of parameters but achieved lower Accuracy than ResNet-101 and EfficientNet-B2. This suggests that architectural design and suitability for the dataset may be more important than model size alone.

The ROC and Grad-CAM analyses provided additional perspectives on model performance and prediction behavior. ResNet-101 achieved the highest AUROC, while Grad-CAM provided a visual representation of the regions associated with the model predictions. Overall, the findings indicate that ResNet-101 was the strongest-performing model within the experimental setting of this study, demonstrating the importance of comparing different architectures using multiple complementary evaluation metrics.

CHAPTER 6: LIMITATIONS AND FUTURE WORK

While the results presented in Chapter 5 demonstrate that CNN-based models, particularly ResNet-101 can achieve promising performance in classifying OL and OSCC with truly encouraging accuracy, there are several limitations that should be acknowledged. This chapter discusses the key limitations of the proposed approach and identifies potential directions for future research. These considerations provide a basis for further improving the robustness, generalizability and clinical applicability of CNN-based classification models for oral lesion diagnosis.

6.1 Limitations of the Study

Narrow class scope- This study focuses on binary classification of leukoplakia and OSCC. Real-world clinical practice involves a wider range of oral lesions and conditions including frictional keratosis, candidiasis, lichen planus and other dysplastic conditions that may have similar histopathological characteristics. Since the models developed in this study were solely trained on OL and OSCC samples, their ability to distinguish other classes remain a challenge. This limitation highlights the need for future studies to incorporate a more diverse set of oral lesion classes to improve the robustness and generalizability of the classification models.

Small dataset size- The dataset used in this study consists of only 237 images from 69 patients, which is relatively small by deep learning standards. Deeper architectures like ResNet-101 and ConvNeXt Base generally benefit from larger datasets by utilizing their representational capacity. The small dataset size also increases the risk of overfitting. The model may learn image-specific artifacts or dataset-specific characteristics instead of generalized patterns that are applicable across various samples.

Single-institution data- All images included in the study were obtained from a single institution, the Universidade Federal do Espírito Santo (UFES) in Brazil, which reflects a specific microscope setup, staining protocol and patient population. Although, the use of a single dataset is a reasonable and widely adopted approach for an initial study, it limits the ability to determine whether the developed models can maintain their performance across images acquired from different institutions, imaging systems, staining protocols, and patient populations. Hence, the generalizability of the proposed models beyond the current dataset remains uncertain and requires further investigation through external validation.

No clinical metadata used- The models in this study relied exclusively on histopathological image data and did not include additional clinical information. Variables such as lesion location, lesion size and patient risk factors, including tobacco and alcohol use, which have been shown to influence disease risk and clinical decision-making were not available to the models. In clinical practice, pathologists interpret histopathological findings along with such

contextual information. As a result, the absence of these additional information may have limited the models' ability to fully capture the complexity of real-world diagnostic assessment.

No cross-validation- All performance metrics reported in this study were derived from a single fixed training, validation and testing split rather than k-fold cross-validation. This approach was intentionally used to ensure a fair and consistent comparison across all six models. However, the reported results reflect performance on a single partition of the dataset. Considering the limited dataset size, alternative data splits could reasonably yield different performance outcomes. Using k-fold cross-validation would have provided a more robust and stable assessment by averaging model performance across multiple data partitions, thus offering a more reliable estimate of their generalization capability.

Class imbalance- The dataset shows a moderate class imbalance including 146 leukoplakia images and 91 OSCC images (approximately 62% and 38%, respectively). To eliminate this imbalance, evaluation metrics such as class-wise Recall and the Matthews Correlation Coefficient (MCC) were emphasized as they provide a more balanced assessment of model performance than overall accuracy alone. Nevertheless, class imbalance may still show a bias toward the majority class during model training. Accordingly, this limitation should be considered when interpreting the class-wise Recall differences discussed in Chapter 5.

No data augmentation and resolution reduction- The study did not apply data augmentation techniques, which may have limited the model's ability to learn variations in tissue appearance and reduced its robustness to unseen images. In addition, the original high-resolution histopathological images were resized to 224 × 224 pixels before being provided to the CNN models. This substantial reduction in spatial resolution may result in the loss of fine-grained cellular and tissue-level features that could be important for distinguishing leukoplakia from OSCC.

6.2 Future Work

The findings of this study suggest several future research directions. One of them is the exploration of multimodal learning approaches that combine histopathological image information with complementary clinical data available within the dataset. While the present work focused entirely on image-based classification, incorporating additional patient-related information may provide a more comprehensive representation of disease characteristics and potentially improve diagnostic performance. Future investigations may also examine advanced feature integration strategies, alternative deep learning architectures and more advanced learning frameworks designed to capture relationships between different data modalities. Such approaches could better reflect the multifaceted nature of clinical decision-making and contribute to the development of more robust computer-aided diagnostic systems for oral lesion assessment.

Additionally, further validation on larger and diverse datasets would be valuable for assessing the generalizability and clinical applicability of the proposed method. Continued research in these directions may help advance automated oral lesion classification and support more reliable diagnostic assistance in real-world clinical settings.

CHAPTER 7: CONCLUSION

7.1 Summary of the Research

This thesis investigated the application of deep learning for the classification of OL and OSCC using histopathological images. The primary objective was to evaluate and compare six modern CNN architectures to determine their suitability for distinguishing between these two clinically important oral lesions. In addition to identifying the best-performing architecture, the study evaluated model performance from multiple perspectives, including overall classification accuracy, class-wise recall, discrimination capability, model complexity, and interpretability.

The clinical motivation for this research was established by the diagnostic challenges associated with OSCC and OL. OSCC represents a major global health burden, and delayed diagnosis can lead to poorer clinical outcomes. OL, as a potentially malignant disorder, may exhibit histopathological characteristics that overlap with those observed in OSCC, making reliable differentiation challenging. These challenges highlight the potential value of computational approaches that can provide supportive and consistent analysis of histopathological images.

The literature review demonstrated that although deep learning and CNN-based approaches have been increasingly applied to oral cancer detection and histopathological image analysis, relatively few studies have systematically compared multiple CNN architectures under a common experimental framework for distinguishing OL from OSCC. Furthermore, many existing studies have focused on classification tasks such as normal versus cancerous tissue rather than the more challenging distinction between a potentially malignant lesion and established carcinoma. These observations provided the motivation for the present study.

The NDB-UFES dataset was used to conduct the experimental evaluation. Six CNN architectures—ResNet-50, ResNet-101, DenseNet121, EfficientNet-B0, EfficientNet-B2, and ConvNeXt-Base—were evaluated using a consistent preprocessing pipeline, transfer learning strategy, training protocol, and evaluation framework. Model performance was assessed using Accuracy, class-wise Recall, Area Under the Receiver Operating Characteristic Curve (AUROC), and Matthews Correlation Coefficient (MCC), together with confusion matrix and Grad-CAM analyses.

The experimental results demonstrated that ResNet-101 achieved the strongest overall performance among the six evaluated architectures. It attained the highest Accuracy (89.58%), AUROC (0.9481), and MCC (0.7810), while achieving an OSCC Recall of 88.89%, comparable to the highest value observed among the evaluated models. Furthermore, its confusion matrix indicated only two false-negative OSCC predictions, reflecting a high level of sensitivity toward the malignant class. Taken together, these findings indicate that ResNet-101 provided the most consistent and reliable overall performance for the binary classification of oral leukoplakia and OSCC within the investigated dataset.

The model-complexity analysis further demonstrated that a larger number of parameters does not necessarily result in superior classification performance. ConvNeXt-Base, which contained substantially more parameters than the other evaluated architectures, did not outperform ResNet-101. This finding suggests that architectural design and the suitability of learned representations may be more important than model size alone when working with relatively small histopathological datasets.

Grad-CAM analysis provided an additional perspective on model behavior by showing that the CNNs frequently concentrated their activation on tissue regions containing visible pathological and structural characteristics. OSCC samples generally showed stronger activation around abnormal tissue structures, while leukoplakia samples exhibited more distributed activation associated with epithelial and tissue-level changes. These visualizations provide supportive evidence that the models were responding to meaningful histopathological regions, although activation maps alone cannot establish clinical correctness.

Overall, the findings demonstrate the potential of transfer learning-based CNNs for distinguishing OL from OSCC using histopathological images. Among the evaluated architectures, ResNet-101 provided the most favorable balance between predictive performance and model complexity, making it the strongest-performing model within the experimental setting of this study.

7.2 Key Findings

Several important findings emerged from this study.

First, ResNet-101 demonstrated the strongest overall performance. It achieved the highest Accuracy (89.58%), AUROC (0.9481), and MCC (0.7810), together with the highest OSCC Recall (88.89%) among the six evaluated architectures. It also produced the fewest false-negative OSCC predictions, with only two OSCC cases incorrectly classified as leukoplakia. This is particularly important because missed OSCC cases may have serious clinical consequences.

Second, overall Accuracy alone was not sufficient to evaluate model performance. Although EfficientNet-B2 achieved the second-highest Accuracy (87.50%) and the highest Leukoplakia Recall (93.33%), it produced four false-negative OSCC predictions compared with two for ResNet-101. This demonstrates the importance of considering class-wise Recall and confusion matrices alongside overall Accuracy when evaluating models for clinically relevant classification tasks.

Third, greater model complexity did not necessarily lead to better classification performance. ConvNeXt-Base contained substantially more parameters than the other evaluated architectures but did not outperform ResNet-101 or EfficientNet-B2. This suggests that architectural suitability and the quality of learned feature representations may be more influential than parameter count alone, particularly when working with relatively small medical imaging datasets.

Fourth, Grad-CAM analysis provided evidence of meaningful model attention. The activation maps showed that the models generally concentrated on regions containing relevant histopathological structures. OSCC images tended to exhibit concentrated activation around abnormal tissue structures, including nuclear clustering, epithelial irregularity, and tissue disruption, whereas leukoplakia images generally showed more distributed activation associated with epithelial and tissue-level changes. This supports the use of explainability analysis as a complementary tool for examining model behavior.

Taken together, these findings indicate that ResNet-101 was the most suitable architecture among the evaluated models for the OL-versus-OSCC classification task under the experimental conditions of this study. The results also demonstrate that model selection should consider multiple dimensions of performance, including class-wise sensitivity, discrimination ability, correlation-based measures, complexity, and interpretability, rather than relying on Accuracy alone.

Bibliography

[1] V. Vinay et al., “Artificial Intelligence in Oral Cancer: A Comprehensive Scoping Review of Diagnostic and Prognostic Applications,” Diagnostics, vol. 15, no. 3, pp. 280–280, Jan. 2025, doi: https://doi.org/10.3390/diagnostics15030280.

[2] I. Evren *et al.*, ‘Associations Between Clinical and Histopathological Characteristics in Oral Leukoplakia’, *Oral Diseases*, vol. 29, no. 2, pp. 696–706, Oct. 2021, doi: 10.1111/odi.14038.

[3] Yadav, Samir S., and Shivajirao M. Jadhav. “Deep Convolutional Neural Network Based Medical Image Classification for Disease Diagnosis.” Journal of Big Data, vol. 6, no. 1, Dec. 2019, https://doi.org/10.1186/s40537-019-0276-2.

[4] Refika Sultan Doğan, and Bülent Yılmaz. “Histopathology Image Classification: Highlighting the Gap between Manual Analysis and AI Automation.” Frontiers in Oncology, vol. 13, 17 Jan. 2024, https://doi.org/10.3389/fonc.2023.1325271.

[5] H. Laçi, K. Sevrani, and S. Iqbal, “Deep learning approaches for classification tasks in medical X-ray, MRI, and ultrasound images: a scoping review,” BMC Medical Imaging, vol. 25, no. 1, May 2025, doi: https://doi.org/10.1186/s12880-025-01701-5.

[6] W. Huang, “Artificial Intelligence and Its Application in Early Oral Cancer Screening: A Systematic Review,” *Frontiers in Oncology*, vol. 16, Mar. 2026, doi: https://doi.org/10.3389/fonc.2026.1789708.

[7] R. Mandal, “Analysis Of Supervised Learning Approaches For Identification Of Oral Squamous Cell Carcinoma: A Multimodal Approach,” Cuestiones de Fisioterapia, vol. 54, no. 4, pp. 5299–5309, 2025, doi: https://doi.org/10.48047/h1j9b871.

[8] X. Wang et al., “A novel deep learning model for automated diagnosis of oral squamous cell carcinoma and related leukoplakia in pathological images,” Journal of Stomatology Oral and Maxillofacial Surgery, vol. 127, no. 4, p. 102743, Feb. 2026, doi: https://doi.org/10.1016/j.jormas.2026.102743.

[9] S. H. Begum and P. Vidyullatha, “Deep Learning Model for Automatic Detection of Oral squamous cell carcinoma (OSCC) using Histopathological Images,” International Journal of Computing and Digital Systems, vol. 13, no. 1, pp. 889–899, Apr. 2023, doi: https://doi.org/10.12785/ijcds/130170.

[10] M. Ahmad et al., “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, pp. 5247–5247, Oct. 2023, doi: https://doi.org/10.3390/cancers15215247.

[11] V. Yaduvanshi, R. Murugan, and T. Goel, “Automatic oral cancer detection and classification using modified local texture descriptor and machine learning algorithms,” Multimedia Tools and Applications, vol. 84, no. 2, pp. 1031–1055, Apr. 2024, doi: https://doi.org/10.1007/s11042-024-19040-y.

[12] Payam Mirfendereski, G. Y. Li, A. T. Pearson, and A. R. Kerr, “Artificial Intelligence and the Diagnosis of Oral Cavity Cancer and Oral Potentially Malignant Disorders From

Clinical Photographs: A Narrative Review," *Frontiers in Oral Health*, vol. 6, Mar. 2025, doi: https://doi.org/10.3389/froh.2025.1569567.

[13] E. Ramesh, A. Ganesan, Krithika Chandrasekar Lakshmi, and P. M. Natarajan, "Artificial Intelligence—Based Diagnosis of Oral Leukoplakia Using Deep Convolutional Neural Networks Xception and MobileNet-v2," *Frontiers in Oral Health*, vol. 6, Mar. 2025, doi: https://doi.org/10.3389/froh.2025.1414524.

[14] J. Adeoye *et al.*, "A Deep Learning System to Predict Epithelial Dysplasia in Oral Leukoplakia," *Journal of Dental Research*, vol. 103, no. 12, pp. 1218–1226, Oct. 2024, doi: https://doi.org/10.1177/00220345241272048.

[15] S.-W. Yang, Y.-S. Lee, L.-C. Chang, C.-H. Yang, C.-M. Luo, and P.-W. Wu, "Oral Tongue Leukoplakia: Analysis of Clinicopathological Characteristics, Treatment Outcomes, and Factors Related to Recurrence and Malignant Transformation," Clinical Oral Investigations, vol. 25, no. 6, pp. 4045–4058, Jan. 2021, doi: https://doi.org/10.1007/s00784-020-03735-1.

[16] A. J. Shephard *et al.*, "A Fully Automated and Explainable Algorithm for Predicting Malignant Transformation in Oral Epithelial Dysplasia," *npj Precision Oncology*, vol. 8, no. 1, Jun. 2024, doi: https://doi.org/10.1038/s41698-024-00624-8.

[17] J. Peng *et al.*, "Oral Epithelial Dysplasia Detection and Grading in Oral Leukoplakia Using Deep Learning," *BMC Oral Health*, vol. 24, no. 1, Apr. 2024, doi: https://doi.org/10.1186/s12903-024-04191-z.

[18] Muniz, Leandro, et al. "Importance of Complementary Data to Histopathological Image Analysis of Oral Leukoplakia and Carcinoma Using Deep Neural Networks." Intelligent Medicine, vol. 3, no. 4, 6 Feb. 2023, pp. 258–266, https://doi.org/10.1016/j.imed.2023.01.004.

[19] K. Warin and Siriwan Suebnukarn, "Deep Learning in Oral Cancer- a Systematic Review," *BMC Oral Health*, vol. 24, no. 1, Feb. 2024, doi: https://doi.org/10.1186/s12903-024-03993-5.

[20] 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. 3, p. e13444, Mar. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e13444.

[21] D. Li, X. Wang, J. Liu, and J. Sun, "Deep Learning-Based Automated Detection of Oral Leukoplakia in Clinical Imaging," *Cureus*, May 2025, doi: https://doi.org/10.7759/cureus.83368.

[22] S. M. Fati, E. M. Senan, and Y. Javed, "Early Diagnosis of Oral Squamous Cell Carcinoma Based on Histopathological Images Using Deep and Hybrid Learning Approaches," *Diagnostics*, vol. 12, no. 8, p. 1899, Aug. 2022, doi: https://doi.org/10.3390/diagnostics12081899.

[23] M. Dörrich *et al.*, "Explainable Convolutional Neural Networks for Assessing Head and Neck Cancer Histopathology," *Diagnostic Pathology*, vol. 18, no. 1, Nov. 2023, doi: https://doi.org/10.1186/s13000-023-01407-8.

[24] K. Figueroa *et al.*, "Interpretable Deep Learning Approach for Oral Cancer Classification Using Guided Attention Inference Network," *Journal of Biomedical Optics*, vol. 27, no. 1, p. 015001, Jan. 2022, doi: https://doi.org/10.1117/1.JBO.27.1.015001.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 618–626, https://doi.org/10.1109/iccv.2017.74.

[26] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778, https://doi.org/10.1109/cvpr.2016.90.

[27] Huang, Gao, et al. "Densely Connected Convolutional Networks." Openaccess.thecvf.com, 2017, doi.org/10.48550/arXiv.1608.06993.

[28] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," PMLR, pp. 6105–6114, May 2019, Available: https://proceedings.mlr.press/v97/tan19a.html?ref=ji.

[29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," arXiv:2201.03545 [cs], Jan. 2022, doi: https://doi.org/10.48550/arXiv.2201.03545.

[30] M. C. F. Ribeiro-de-Assis et al., "NDB-UFES: An oral cancer and leukoplakia dataset composed of histopathological images and patient data," Data in Brief, vol. 48, p. 109128, Apr. 2023, doi: https://doi.org/10.1016/j.dib.2023.109128.

[31] S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, "Transfer Learning for Cancer Diagnosis in Histopathological Images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, p. 129, Mar. 2022, doi: https://doi.org/10.11591/ijai.v11.i1.pp129-136.

[32] S. Mukherjee, "The Annotated ResNet-50 - TDS Archive - Medium," Medium. Accessed: Aug. 30, 2026. [Online]. Available: https://medium.com/data-science/the-annotated-resnet-50-a6c536034758.

[33] R Chetan, D. V. Ashoka, and A. Prakash, "HybridTransferNet: Advancing Soil Image Classification Through Comprehensive Evaluation of Hybrid Transfer Learning," Research Square (Research Square), Jun. 2023, doi: 10.21203/rs.3.rs-3032907/v1.

[34] A. Zhou et al., "Multi-Head Attention-Based Two-Stream EfficientNet for Action Recognition," Multimedia Systems, Jun. 2022, doi: 10.1007/s00530-022-00961-3.

[35] GeeksforGeeks, "ConvNeXt," GeeksforGeeks. Accessed: Aug. 31, 2026. [Online]. Available: https://www.geeksforgeeks.org/computer-vision/convnext/

[36] I. van der Waal, "Oral Leukoplakia: Present Views on Diagnosis, Management, Communication With Patients, and Research," Current Oral Health Reports, vol. 6, no. 1, pp. 9–13, Jan. 2019, doi: 10.1007/s40496-019-0204-8.

[37] C. Bramati et al., "Early Diagnosis of Oral Squamous Cell Carcinoma May Ensure Better Prognosis: A Case Series," Clinical Case Reports, vol. 9, no. 10, Oct. 2021, doi: 10.1002/ccr3.5004.

[38] Y. Eltohami and A. Suleiman, “Survival Analysis of Sudanese Oral Squamous Cell Carcinoma Patients With Field of Cancerization,” BMC cancer, vol. 24, no. 1, p. 473, Apr. 2024, doi: 10.1186/s12885-024-12197-7.

[39] R. E. Fanchiotti, S. E. M. Reid, B. L. R. de Azevedo, W. G. Bautz, L. A. P. de Barros, and da G. de S. L. Nogueira, “Fibrosis in Oral Carcinoma and Leukoplakia: An Immunohistochemical Study,” Journal of Oral & Maxillofacial Research, vol. 15, no. 3, p. e1, Sep. 2024, doi: 10.5037/jomr.2024.15303.