Oral Squamous Cell Carcinoma Detection: A Comparison of CNNs in Histopathological Images

João O. B. Diniz, Breno A. Tamini,
André F. Alevino, Luiz O. O. Souza Jr.

Innovation Factory - Federal Institute of Maranhão (IFMA)
Grajaú — MA — Brasil
Email: joao.bandeira@ifma.edu.br

Luana Batista da Cruz

Federal University of Cariri (UFCA)
Juazeiro do Norte, CE - Brasil — MA — Brasil
Email: luana.batista@ufca.edu.br

***Abstract*—Oral Squamous Cell Carcinoma (OSCC) is the most prevalent form of oral cancer, accounting for approximately 90% of cases worldwide. Early and accurate diagnosis is critical for improving patient outcomes, yet traditional histopathological analysis remains subjective and time-consuming. This study presents a systematic benchmark comparing the performance of eleven convolutional neural network (CNN) architectures for the automated classification of OSCC in histopathological images. Using a publicly available dataset, all models were trained and evaluated under uniform conditions. Metrics such as accuracy, recall, precision, F1-score, and AUC-ROC were employed to assess performance. Results demonstrated that EfficientNetB3 achieved the highest accuracy (94.12%) and overall robustness, outperforming other architectures. This study provides a comprehensive comparative analysis of CNNs for OSCC detection, highlighting the potential of modern architectures like EfficientNet to enhance diagnostic precision. The standardized benchmark established here facilitates future research and clinical integration of deep learning tools for histopathological cancer diagnosis.**

I. Introduction

Oral cancer represents a major global public health challenge, with high morbidity and mortality rates due to late diagnosis [1], [2]. In Brazil, this malignancy ranks among the most prevalent cancers for the 2023-2025 triennium according to National Cancer Institute (INCA) estimates [3]. Among its variants, Oral Squamous Cell Carcinoma (OSCC) accounts for approximately 90% of cases, primarily affecting men over 60 with histories of tobacco use, alcohol consumption, and prolonged sun exposure [2], [4].

Despite its clinical significance, diagnosis still heavily relies on histopathological analysis, which is susceptible to inter- and intra-observer variability [5]. In this context, deep learning techniques applied to histopathological images have emerged as a promising solution to support pathologists, reducing diagnostic errors and enabling earlier detection [6]–[9].

Convolutional Neural Networks (CNNs), have demonstrated remarkable performance in classification of histopathological images within Digital Pathology, achieving results comparable to expert pathologists [10]–[12]. A critical factor in these models' success is their generalization capability, often enhanced through data augmentation strategies that simulate tissue variability while reducing overfitting risks [1], [13].

This work presents a systematic analysis of CNN architectures for automated OSCC detection in histopathological images. We compare eleven state-of-the-art architectures under standardized experimental conditions to identify the optimal accuracy-robustness trade-off. Our key contributions include:

- A evaluation of eleven CNN architectures for OSCC classification in histopathology, ensuring fair and standardized experimental comparisons across all models.
- A critical assessment of architectural suitability for digital pathology applications, including clinical implementation challenges and computational efficiency analysis.

II. Related Work

Recent research has significantly advanced the application of CNNs for automated detection of OSCC in histopathological images. The work by [14] demonstrated promising results by integrating multiple CNN architectures (VGG16, AlexNet, ResNet50, and InceptionV3) with Binary Particle Swarm Optimization (BPSO) for feature selection and XGBoost for classification, achieving an accuracy of 96.3%. In a different approach, [15] utilized transfer learning with AlexNet to distinguish between normal and OSCC images, attaining 90.06% test accuracy.

The field has progressively evolved to evaluate more sophisticated architectures. [10] conducted a comprehensive comparison between traditional CNNs, transformer-based models, and few-shot learning techniques, with DenseNet-121 emerging as the most effective model at 91.91% accuracy. Expanding on this, [11] developed an ensemble method combining multiple CNNs (AlexNet, ResNet, Inception, and Xception) that achieved 97.88% accuracy. Parallel research by [12] implemented an enhanced ResNet architecture with advanced data augmentation techniques, reaching 96.72% accuracy.

Recent advances in [1] systematically compared four modern architectures (VGG16, ResNet101, MobileNetV2, and InceptionV3), identifying ResNet101 as the top performer with 97.21% accuracy. Notably, studies by [16] and [4] explored the EfficientNet-B3 architecture, reporting unprecedented accuracy levels up to 98.00%, solidifying its position as a leading model for biomedical image analysis.

Despite these significant advancements, current literature reveals two critical limitations that our study addresses. First, most existing works concentrate on limited architecture selections or specific combinations while employing inconsistent experimental protocols, making direct comparisons unreliable. Second, no prior study has conducted a comprehensive evaluation of diverse, state-of-the-art CNNs under standardized training and validation conditions.

Our research directly addresses these gaps by implementing a systematic benchmark of eleven distinct CNN architectures. This unified evaluation identifies the optimal balance between performance and robustness and establishes the first standardized comparative analysis for OSCC classification in histopathology, providing valuable insights for both the computer vision and medical research communities.

III. Materials and Methods

Our method comprises four key steps, as illustrated in Fig. 1. The following subsections detail each stage.

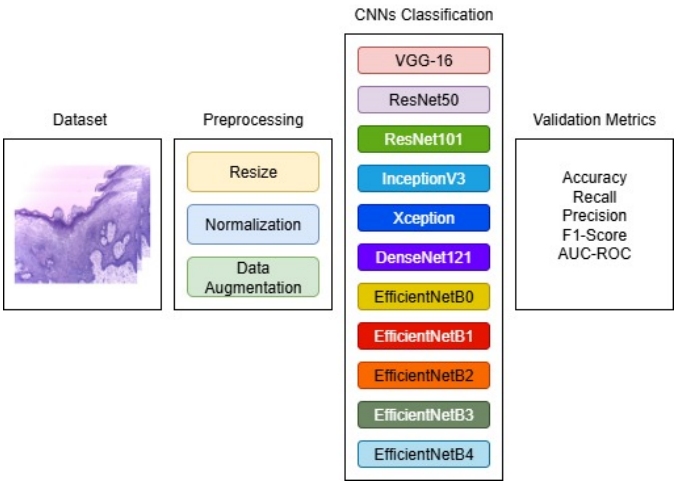

Fig. 1. Method workflow overview.

A. Dataset

We developed our method using a public histopathological image dataset of OSCC available on Kaggle [17], containing Hematoxylin and Eosin (H&E)-stained samples classified as Normal and OSCC. Certified pathologists collected the original high-resolution images (2040 $\times$ 1536 pixels), which we stratified into training (70%), validation (15%), and test (15%) sets. Figure 2 shows representative samples from each class.

The original dataset included pre-established data augmentation, yielding 2,435 Normal and 2,511 OSCC images. However, the unaugmented dataset contained only 231 Normal and 747 OSCC cases, revealing disproportionate augmentation - the Normal class was artificially increased 10$\times$ compared to just 3$\times$ for OSCC, potentially introducing deep learning model bias.

To ensure a replicable benchmark with rigorous experimentation, we implemented the following protocol:

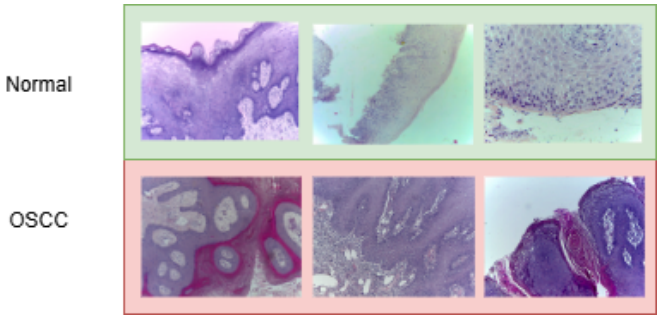

Fig. 2. Representative H&E-stained histopathology samples: (A) Normal oral mucosa showing intact epithelial layers, (B) OSCC demonstrating invasive tumor islands with keratin pearls. Scale bar: 500μm.

- Used only the original 978 samples (231 Normal + 747 OSCC).
- Applied data augmentation exclusively to the minority class during preprocessing.
- Eliminated artificial samples from validation and test sets.

The following section details our preprocessing and data augmentation pipeline, specifically designed to prevent model bias while maintaining histopathological validity.

B. Preprocessing

We resized images to 224 $\times$ 224 pixels, the standard input dimension for CNN's [18], [19], and normalized pixel values to the [0,1] to ensure compatibility with ImageNet pretrained weights. To address the 1:3 class imbalance, we applied data augmentation exclusively to the minority class (Normal) until matching the majority class (OSCC) sample count.

Our augmentation pipeline included: Random rotations; Horizontal and vertical flips; Brightness variations; and Zoom transformations.

Unlike the original dataset's disproportionate augmentation (10$\times$ oversampling of Normal class), our controlled approach:

- Balanced class distribution without bias introduction.
- Preserved histopathological validity through medically plausible transformations.
- Excluded augmented samples from validation/test sets.

This step generated sufficient training diversity while preventing model bias from artificial sample overrepresentation, ultimately improving generalization capability on real clinical samples.

C. CNNs Classification

We evaluated 11 representative CNN architectures spanning different generations of deep networks, all initialized with ImageNet pretrained weights [20] and adapted for binary classification. Training employed the Focal Loss [21], which emphasizes learning from more challenging instances, enhancing the model's robustness. The architectures comprise:

- **VGG16**: a classical architecture with 16 convolutional layers in sequential blocks, widely adopted in medical imaging tasks. Its uniform depth serves as an important baseline for comparison with newer models.

- **ResNet50/101**: Residual networks (50 and 101 layers) featuring skip connections to address vanishing gradients. The deeper ResNet101 variant tests whether additional depth improves the detection of subtle histopathological patterns.
- **InceptionV3**: utilizes parallel multi-scale convolutional filters (inception modules) to capture hierarchical features across different magnification levels.
- **DenseNet121**: implements dense connectivity where each layer receives feature maps from all preceding layers, promoting feature reuse and reducing gradient issues which is particularly advantageous for complex histopathology textures.
- **Xception**: an Inception evolution using depthwise separable convolutions for improved computational efficiency without accuracy loss.
- **EfficientNet Family**: a series of models (B0-B4) optimized through compound scaling:
  - **B0-B2**: progressively more complex variants (B0: baseline; B1-B2: intermediate).
  - **B3**: the optimal balance (48 layers) with squeeze-and-excitation attention for critical local feature detection.
  - **B4**: the largest tested variant evaluating marginal gains from added complexity.

This selection enables comparison of absolute model performance and highlights clinical trade-offs. These include model complexity, computational efficiency, and generalization capability in real-world diagnostic scenarios.

D. Validation Metrics

We evaluated model performance using well-established metrics in medical decision-support systems [21], including accuracy (overall proportion of correct classifications), recall or sensitivity (ability to correctly identify positive OSCC cases), precision (proportion of true positives among samples classified as positive), F1-score (harmonic mean of precision and recall), and the area under the ROC curve (AUC), which measures the separability between classes.

IV. Results and Discussion

This section presents the training environment, the results obtained for the 11 architectures, and a comparison with the literature.

A. Training Environment

We conducted all experiments on a machine equipped with an Intel® Core™ i7 processor running at 2.90 GHz, an NVIDIA RTX 4060 GPU with 8 GB of dedicated memory, 16 GB of RAM, and Windows 11 Pro as the operating system.

Before training, all images underwent the preprocessing and data augmentation steps described in Section III-B, ensuring experimental consistency across models. We trained all architectures under the same conditions: 20 epochs, Adam optimizer, Focal Loss function, and monitoring of the best F1-score on the validation set.

B. CNN Classification

Table I reports the results obtained on the test set for the 11 evaluated architectures.

TABLE I
Comparison of CNN Architectures.

Architectures	Accuracy	Recall	Precision	F1-score	AUC-ROC
VGG16	79.17%	98.91%	79.13%	87.92%	69.72%
ResNet50	86.67%	98.91%	85.85%	91.92%	80.94%
ResNet101	83.33%	97.82%	83.33%	90.00%	83.54%
InceptionV3	91.67%	93.47%	95.56%	94.51%	91.65%
Xception	92.50%	94.56%	95.60%	95.08%	91.73%
DenseNet121	90.83%	92.39%	95.51%	93.92%	93.28%
EfficientNetB0	83.33%	100.00%	82.14%	90.20%	82.14%
EfficientNetB1	84.17%	92.39%	87.63%	89.95%	83.35%
EfficientNetB2	92.33%	94.85%	93.57%	94.13%	93.71%
EfficientNetB3	94.12%	95.93%	95.23%	95.54%	95.06%
EfficientNetB4	93.41%	95.12%	94.26%	94.63%	94.49%

We observed that modern EfficientNets (B2–B4) achieved accuracy above 92%, with EfficientNetB3 standing out as the best overall performer. Classical models, such as VGG16, maintained high recall but compromised precision and AUC-ROC, indicating lower robustness for OSCC detection.

Beyond individual results, this study establishes a uniform benchmark for comparing different CNN architectures. All networks were evaluated under the same experimental conditions, using the same test set, preprocessing steps, and training protocols. This consistent framework provides a comprehensive view of performance evolution across generations of CNNs.

Therefore, the study not only identifies the most promising architectures but also delivers a robust benchmark for future comparisons, validations, and improvements, consistently contributing to advances in computer-assisted diagnosis of histopathological images.

C. Comparison with the Literature

Table II summarizes the main related works, along with the best result obtained in this study, enabling a direct comparison between methods and architectures.

TABLE II
Comparison with the literature.

Author(s)	Architecture(s)	Accuracy	Recall	Precision
[14]	InceptionV3	96.30%	—	—
[15]	AlexNet	90.06%	—	—
[10]	DenseNet-121	91.91%	91.93%	91.93%
[11]	ResNet + Xception	97.88%	—	—
[12]	ResNet	96.72%	—	—
[1]	ResNet101	97.21%	—	—
[16]	EfficientNetB3	98.00%	—	—
[4]	EfficientNetB3	98.00%	98.00%	98.00%
Best Architecture	EfficientNetB3	94.12%	95.93%	95.23%

Although none of the tested architectures reached the maximum values reported in the literature, EfficientNetB3 achieved the best performance among the networks evaluated in this study, confirming a trend similar to that observed in related works. The main contribution of this study lies in the evaluation of 11 different architectures under standardized conditions, creating a robust benchmark. This approach allows

not only the assessment of each network's performance but also an objective comparison across different generations of CNNs. By limiting augmentation to the minority class, our approach reduces the risk of data leakage and enables a more rigorous assessment of model generalization.

Furthermore, the main contribution of this study lies in the systematic evaluation of 11 CNN architectures under standardized conditions, establishing a robust benchmark. This approach enables not only the assessment of each network's performance but also a controlled comparison across different CNN generations, highlighting trade-offs between recall, precision, and AUC-ROC. By providing such a framework, the study offers a solid reference for future research on OSCC detection and medical computer vision applications.

D. Conclusion and Future Work

The study demonstrated that modern CNN architectures, particularly EfficientNetB3, outperform classical networks in OSCC detection, consistently balancing recall, precision, and AUC-ROC. The main value of this work lies in creating a robust benchmark by evaluating 11 architectures under standardized conditions. This benchmark provides a clear reference for selecting networks in clinical applications and contributes to the standardization of future comparisons in medical computer vision.

Future work may explore ensembles of architectures to further enhance accuracy and robustness, as well as the integration of attention mechanisms or Transformer-based approaches tailored for medical imaging. Expanding the benchmark to include different types of oral lesions and additional datasets will also enable the evaluation of model generalization in broader clinical scenarios.

ACKNOWLEDGMENT

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Fundação de Amparo à Pesquisa do Maranhão (FAPEMA), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We also acknowledge LLM use for spell-checking, grammar correction, and assistance in translating specific terms.

REFERENCES

[1] A. S. R. C. Murthy, G. Mercy, L. J. Prakash, and K. S. Bose, "Histopath-dl-oc: Deep learning for oral cancer prediction from histopathology data," in *2025 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2025, pp. 1190–1197.

[2] WHO, "World health organization: Oral cancer," https://www.who.int/news-room/fact-sheets/detail/oral-health, 2024, accessed on: June 26. 2025.

[3] INCA, "Instituto nacional de câncer, estimativa 2023: Incidência de câncer no brasil," https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil, 2023, accessed on: June 26. 2025.

[4] A. Kumar and L. Nelson, "Enhancing oral squamous cell carcinoma detection using efficientnetb3 from histopathologic images," in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*. IEEE, 2025, pp. 950–956.

[5] S. R. Bisht, P. Mishra, D. Yadav, R. Rawal, and K. P. Mercado-Shekhar, "Current and emerging techniques for oral cancer screening and diagnosis: a review," *Progress in Biomedical Engineering*, vol. 3, no. 4, p. 042003, 2021.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] E. D. Carvalho, O. Antonio Filho, R. R. Silva, F. H. Araujo, J. O. Diniz, A. C. Silva, A. C. Paiva, and M. Gattass, "Breast cancer diagnosis from histopathological images using textural features and cbir," *Artificial intelligence in medicine*, vol. 105, p. 101845, 2020.

[8] N. P. Ribeiro, F. R. Teles, J. O. B. Diniz, L. B. da Cruz, D. A. Dias Jr, G. Braz Junior, J. D. de Almeida, and A. C. de Paiva, "Improving colorectal cancer diagnosis using mirnet and inceptionv3 on histopathological images," in *Brazilian Conference on Intelligent Systems*. Springer, 2024, pp. 321–334.

[9] J. O. Diniz, N. P. Ribeiro, D. A. D. Junior, L. B. da Cruz, A. O. de Carvalho Filho, D. L. Gomes Jr, A. C. Silva, and A. C. de Paiva, "Efficientxyz-deepfeatures: seleção de esquema de cor e arquitetura deep features na classificação de câncer de cólon em imagens histopatológicas," in *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. SBC, 2024, pp. 82–93.

[10] B. M. S. Maia, M. C. F. R. de Assis, L. M. de Lima, M. B. Rocha, H. G. Calente, M. L. A. Correa, D. R. Camisasca, and R. A. Krohling, "Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer," *Expert Systems with Applications*, vol. 241, p. 122418, 2024.

[11] M. Das, R. Dash, S. K. Mishra, and A. K. Dalai, "An ensemble deep learning model for oral squamous cell carcinoma detection using histopathological image analysis," *IEEE Access*, 2024.

[12] D. Raval, A. Patel, J. N. Undavia, A. Shukla, and U. Patel, "Oral cancer detection with convolutional neural networks and transfer learning: A resnet-based approach," in *International Conference on Data Analytics & Management*. Springer, 2024, pp. 237–245.

[13] B. A. Tamanini, V. G. Sousa, L. P. Rodrigues, D. M. Oliveira, C. X. Dias, L. B. da Cruz, J. O. Diniz, and L. O. S. Júnior, "Classificaçao de carcinoma endometrioide de ovário por transformaçao de esquema de cor e radiomics em imagens histopatológicas," in *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. SBC, 2025, pp. 68–79.

[14] M. A. Deif, H. Attar, A. Amer, I. A. Elhaty, M. R. Khosravi, and A. A. Solyman, "Diagnosis of oral squamous cell carcinoma using deep neural networks and binary particle swarm optimization on histopathological images: an aiomt approach," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 6364102, 2022.

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

[16] R. L. d. Prado, J. A. Marsicano, A. K. Frois, and J. D. Brancher, "The use of machine learning to support the diagnosis of oral alterations," *Pesquisa Brasileira em Odontopediatria e Clínica Integrada*, vol. 25, p. e240048, 2025.

[17] A. F. Kebede, "Histopathologic oral cancer detection using cnns," https://www.kaggle.com/datasets/ashenafifasilkebede/dataset, 2022, accessed on: June 26. 2025.

[18] D. A. D. Júnior, L. B. da Cruz, J. O. B. Diniz, G. L. F. da Silva, G. B. Junior, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic method for classifying covid-19 patients based on chest x-ray images, using deep features and pso-optimized xgboost," *Expert Systems with Applications*, vol. 183, p. 115452, 2021.

[19] J. O. B. Diniz, N. P. Ribeiro, D. A. Dias Jr, L. B. da Cruz, G. L. da Silva, D. L. Gomes Jr, A. C. de Paiva, and A. C. Silva, "Anisotropicbreast-vit: Breast cancer classification in ultrasound images using anisotropic filtering and vision transformer," in *Brazilian Conference on Intelligent Systems*. Springer, 2024, pp. 95–109.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[21] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.