Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

AI-Assisted Multimodal Deep Learning Framework for Early Detection of Oral Cancer

Amrutasagar Kavarthapu
Department of CSE (AI & ML)
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, India

Vankayalapati Nikhita
Department of CSE (AI & ML)
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, India

Naragani Nivas
Department of CSE (AI & ML)
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, India

Puritipati Manasa
Department of CSE (AI & ML) Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, India

Purilla Mahidhar
Department of CSE (AI & ML) Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, India

Abstract: - Oral Potentially Malignant Disorders (OPMDs) and Oral Squamous Cell Carcinoma (OSCC) pose a serious global health challenge, where delayed diagnosis significantly reduces treatment effectiveness and patient survival rates. Early-stage oral lesions often exhibit subtle visual characteristics, making accurate screening difficult, particularly in primary care and resource-limited settings. Although recent advances in deep learning have demonstrated promising results for oral lesion classification using intraoral images, most existing approaches rely solely on image data and overlook the clinical risk factors routinely considered by clinicians. To address this limitation, this study proposes an AI-assisted multimodal oral cancer detection framework that integrates intraoral images with patient metadata to enhance diagnostic accuracy and clinical relevance. The proposed pipeline utilizes state-of-the-art convolutional neural networks for image feature extraction combined with a machine learning–based metadata classifier to model patient risk profiles. Multiple pre-trained image encoders, including MobileNetV3-Large, EfficientNet-B0, ResNet-50, DenseNet-121, and InceptionV3, are evaluated and compared under a unified experimental setup. An ensemble fusion strategy with adaptive threshold tuning is introduced to balance sensitivity and specificity, closely mimicking real-world clinical decision-making. Experimental results demonstrate that the proposed multimodal ensemble approach outperforms image-only models, achieving an overall accuracy of 83.3%, an F1-score of 88.9%, and a Matthews Correlation Coefficient (MCC) of 0.56 on the validation dataset. These findings confirm that combining visual cues with patient-specific clinical information significantly improves early detection of OPMDs and potential malignancies, supporting more reliable clinical screening and decision-making.

Keywords - Oral Cancer Detection, Oral Potentially Malignant Disorders (OPMD), Multimodal Deep Learning, Convolutional Neural Networks, Clinical Metadata Fusion, EfficientNet, Ensemble Learning, Medical Image Analysis, Artificial Intelligence in Healthcare, Early Cancer Screening.

I. INTRODUCTION

Oral cancer, predominantly Oral Squamous Cell Carcinoma (OSCC), is one of the most common malignancies affecting the head and neck region and Disorders (OPMDs), such as leukoplakia and erythroplakia, which present visible changes in the oral mucosa. Early identification of these conditions significantly improves treatment outcomes and survival rates; however, in many cases, diagnosis is delayed due to limited access to specialists, lack of awareness, and reliance on subjective visual examination.

Conventional oral cancer screening methods depend heavily on clinical expertise, histopathological analysis, and invaise biopsy procedures. While these approaches are effective, they are time-consuming, resource intensive, and not always feasible for large-scale or community-level screening. In recent years, advances in artificial intelligence (AI) and deep learning have enabled automated analysis of medical images, offering a promising alternative for early and non-invasive detection of oral lesions using standard photographic images.

Most existing deep-learning-based oral cancer detection systems primarily focus on image-only classification using convolutional neural networks (CNNs). Although such methods demonstrate encouraging performance, they often overlook clinically relevant patient information such as age, gender, and lifestyle habits (e.g., smoking, alcohol consumption, and betel quid chewing), which play a crucial role in oral cancer risk assessment. In real clinical practice, clinicians rely on a combination of visual examination and patient history rather than images alone. Ignoring this complementary information can limit the diagnostic reliability of image based models.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

To address this limitation, this study proposes a multimodal deep-learning framework that integrates oral lesion images with patient metadata to improve the accuracy and robustness of early oral cancer detection. The proposed approach employs multiple state-of-the-art pre-trained CNN architectures for feature extraction from Region of Interest (ROI) oral images, combined with a machine-learning-based metadata classifier. An ensemble fusion strategy is used to merge image-basedpredictions with clinical risk factors, closely emulating real-world diagnostic reasoning.

The primary contributions of this work include: (i) a comprehensive comparison of multiple deep-learning image encoders for oral lesion classification, (ii) incorporation of patient metadata to enhance predictive performance, (iii) an optimized ensemble fusion and thresholding strategy to improve clinical sensitivity, and (iv) deployment of the trained model as an interactive web-based application for real-time decision support. The experimental results demonstrate that the proposed multimodal ensemble framework outperforms single-modality approaches, highlighting its potential for assisting clinicians in early screening and improving oral cancer prognosis.

1.2 Motivation:

Despite significant advancements in medical imaging and artificial intelligence, early detection of oral cancer remains a challenging task, particularly in low-resource and rural settings. A major motivation for this research arises from the high mortality rate associated with Oral Squamous Cell Carcinoma (OSCC), which is largely attributed to late-stage diagnosis. Many patients initially present with subtle oral lesions that are often misinterpreted or ignored, leading to delayed clinical intervention and poorer outcomes.

Recent studies have demonstrated the effectiveness of deep learning models in classifying oral lesions from photographic images. However, most existing approaches rely solely on visual data and fail to incorporate patient-specific clinical information, such as age, gender, and lifestyle habits, which are well established risk factors for oral cancer. This image only strategy does not fully reflect real-world diagnostic practices, where clinicians combine visual inspection with patient history to arrive at informed decisions. Consequently, models that ignore metadata may produce unreliable predictions, particularly in cases where visual signs are ambiguous.

Another motivating factor is the growing availability of high-quality oral images captured using common devices such as smartphones and intraoral cameras. While these images provide valuable diagnostic cues, variations in lighting conditions, image quality, and lesion appearance can affect model performance. Furthermore, there is a strong need for explainable and clinically meaningful decision-support systems rather than black-box predictions. By incorporating metadata and applying an ensemble-based fusion strategy, the proposed framework produces more balanced and clinically interpretable outputs. This approach not only enhances predictive accuracy but also aligns with clinicians' reasoning processes, thereby increasing trust and adoption in real healthcare environments. Finally, the motivation of this work extends beyond model performance to practical applicability. The deployment of the proposed multimodal framework as a web-based application enables real-time screening support, making it suitable for preliminary assessment in community health centers and telemedicine platforms. This research is motivated by the goal of developing an accessible, reliable, and clinically relevant AI-assisted tool that supports early oral cancer detection and ultimately contributes to improved patient outcomes.

1.3 Objectives:

- To develop a multimodal AI-based system that integrates oral lesion images and patient metadata for early detection of oral potentially malignant disorders (OPMD).
- To evaluate and compare multiple pre-trained deep learning models for oral lesion image classification and identify the most effective architecture.
- To improve diagnostic accuracy and robustness by applying ensemble learning and optimized decision thresholding techniques.
- To design an image-priority fusion strategy that reflects clinical decision-making by emphasizing visual lesion evidence while incorporating patient risk factors.
- To deploy a practical, user-friendly web application that enables real-time oral cancer risk assessment using uploaded images and clinical inputs.

1.4 Related Work:

Early research on oral cancer detection primarily focused on machine learning techniques. Handcrafted features such as texture, color, and shape descriptors were extracted from oral cavity images and classified using algorithms like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) [1], [2]. Although these methods provided initial insights, their performance was highly dependent on features.

With the emergence of deep learning, convolutional neural networks (CNNs) became the dominant approach for oral lesion classification. Several studies employed pre-trained architectures such as ResNet, DenseNet, Inception, and MobileNet to distinguish between benign and malignant oral lesions using photographic images [3]–[6]. These models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

significantly improved classification accuracy by learning hierarchical image features; however, most approaches relied exclusively on image data without considering patient-specific clinical information

In summary, existing literature highlights the effectiveness of deep learning for oral cancer detection but reveals gaps in comprehensive multimodal integration, systematic model comparison, and sensitivity-oriented decision strategies. These limitations motivate the proposed multimodal ensemble framework designed to improve early detection accuracy and support reliable clinical screening.

II. PROPOSED SOLUTION

The proposed system presents an AI-assisted multimodal framework for the early detection of Oral Potentially Malignant Disorders (OPMD) by integrating visual information from oral cavity images with clinically relevant patient metadata. The objective of this approach is to replicate the diagnostic reasoning of medical professionals, who evaluate both observable lesion characteristics and patient risk factors before reaching a clinical decision.

In this framework, oral images are processed using multiple pre-trained convolutional neural network architectures through transfer learning. Models such as MobileNetV3-Large, EfficientNet-B0, ResNet-50, DenseNet-121, and InceptionV3 are fine-tuned to extract high-level discriminative features from region-of interest oral images. These networks are trained to distinguish between healthy tissue and potentially malignant lesions based on visual patterns such as color irregularities, texture variations, and lesion morphology. Parallelly, structured patient metadata including age, gender, smoking habit, betel quid chewing, and alcohol consumption—is utilized to capture non-visual risk factors strongly associated with oral cancer development. This metadata is modeled using a machine learning classifier to estimate the probability of malignancy based on clinical history.

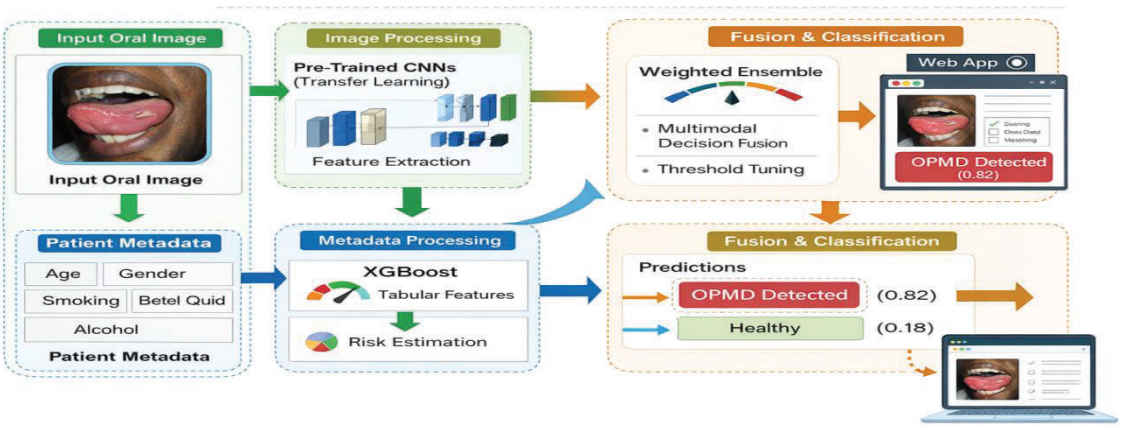

Fig 1: Flow Chart of Project

The outputs from the image-based deep learning model and the metadata-based classifier are combined using a decision-level ensemble fusion strategy. An image priority fusion mechanism is incorporated to ensure that strong visual evidence of potential malignancy is not suppressed by the absence of high-risk habits, reflecting real-world clinical practices. Additionally, threshold tuning is applied to the final prediction scores to optimize classification performance and improve sensitivity toward early-stage lesions. The proposed system is deployed as a web-based application using the Flask framework, allowing users to upload oral images and input patient details to obtain real-time diagnostic results along with confidence scores. This integrated solution demonstrates improved diagnostic accuracy and clinical relevance, making it suitable for early screening and decision support in oral cancer detection.

2.1 Dataset Collection:
The Dataset of Annotated Oral Cavity Images for Oral Cancer Detection, obtained from Zenodo. The dataset consists of 3,000 high-quality images of oral cavities taken with mobile phone cameras from the Sri Lankan population. The images are categorized into healthy, benign, oral potentially malignant disorders (OPMD), and oral cancer (OCA) by domain experts. Each image contains annotations for oral cavity and lesion boundaries in COCO format. Additionally, patient metadata, such as age, sex, diagnosis, and risk factors like smoking, alcohol consumption, and betel quid chewing, is included in the meta-data files. The dataset used in this study is a multimodal oral cancer dataset

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

consisting of oral cavity images and corresponding.

Alongside the image data, structured patient metadata was utilized to incorporate known clinical risk factors associated with oral cancer. The metadata includes five attributes: age, gender, smoking habit, alcohol consumption, and betel quid The dataset was curated to remove incomplete records, duplicate entries, and low-quality images to ensure data reliability. Following preprocessing, the dataset was divided into training and validation subsets while 2.2 Data Preprocessing and Preparation Effective data preprocessing is critical to ensure robustness and generalization of deep learning models, particularly in medical image analysis. In this study, both image data and patient metadata were preprocessed independently and then prepared for multimodal fusion to enable reliable learning. All oral cavity images were first standardized to ensure uniformity across the dataset. Each image was resized to a fixed spatial resolution of 224 × 224 pixels, matching the input requirements of the employed pre trained convolutional neural networks. The images were converted to RGB format to maintain consistency in color channels. Pixel intensity values were normalized to the range using min–max normalization, defined as:

The clinical metadata associated with each image was preprocessed to ensure compatibility with machine learning models. Categorical variables such as gender, smoking habit, alcohol consumption, and betel quid chewing were transformed into representations using binary encoding. numerical Continuous variables, such as age, were normalized using standard scaling to reduce bias caused by differing numerical ranges. This normalization is expressed as:

Clinical metadata consisting of patient age, gender, smoking status, alcohol consumption, and betel quid chewing habits were preprocessed prior to model maintaining class balance between healthy and OPMD After preprocessing, the dataset was divided into training and validation sets using a stratified split to preserve the class distribution between healthy and OPMD samples. Binary labels were encoded as 0 for Healthy and 1 for OPMD. This preparation ensured consistent supervision across both unimodal and multimodal learning stages. By applying structured preprocessing to both image and metadata modalities, the dataset was transformed into a form suitable for efficient training of deep learning and ensemble models, enabling the proposed system to learn complementary visual and clinical patterns effectively.

III. IMPLEMENTATION

The proposed AI-assisted oral cancer detection system employs a multimodal framework that integrates deep visual features from oral cavity images with structured clinical chewing habit. These factors are widely recognized in clinical literature as contributors to oral cancer risk and are routinely considered during medical diagnosis. Each metadata record is accurately aligned with its corresponding oral image, enabling effective multimodal learning.

metadata to enhance the early identification of Oral Potentially Malignant Disorders (OPMD). This approach mimics clinical diagnostic reasoning by prioritizing visual lesion characteristics while incorporating patient risk factors. The system architecture consists of image preprocessing, metadata preprocessing, independent model inference, probabilistic fusion, and threshold-based decision making to generate clinically interpretable outcomes.

All oral images were resized to a fixed spatial resolution of 224×224 pixels to ensure compatibility with pre-trained convolutional neural networks. Pixel intensity normalization was performed to scale values within the range [0, 1] is expressed as

$$I_{norm} = \frac{I}{255}$$

where I represents the original pixel intensity and I_{norm} denotes the normalized pixel value. This normalization improves numerical stability and accelerates model convergence. Feature extraction was performed using the EfficientNet-B0 architecture, which balances model complexity and performance through compound scaling. The final classification layer applies a sigmoid activation function to compute the probability of OPMD from image features, given by

$$P_{img} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

where z represents the output of the final fully connected layer.

Clinical metadata consisting of patient age, gender, smoking status, alcohol consumption, and betel quid chewing habits were preprocessed prior to model inference. Categorical variables were encoded into binary numerical representations, while continuous features were retained in their original scale. An Extreme Gradient Boosting (XGBoost) classifier was trained to model the non-linear relationship between clinical features and OPMD risk. The metadata-based probability prediction is expressed as

$$p_{meta} = f_{xgb}(X)$$

where X denotes the vector of patient metadata features and f_{xgb} represents the trained XGBoost classifier.

To obtain a unified prediction, a weighted multimodal fusion strategy was employed, assigning higher importance to image-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

derived probabilities to reflect the clinical relevance of visible oral lesions. The final probability of OPMD was computed as:

$$P_{final} = w_{img}.P_{img} + w_{meta}.P_{meta}$$

where w_{img} and w_{meta} are empirically determined weights for image and metadata contributions, respectively. Threshold tuning was performed to optimize classification performance, and the final decision rule was defined as:

$$Prediction = \begin{cases} OPMD\ Detected, & P_{final \geq \tau} \\ Healthy, & P_{fianl \leq \tau} \end{cases}$$

where τ denotes the optimized decision threshold.

The proposed implementation was evaluated using multiple performance metrics to ensure balanced assessment under class imbalance conditions. Accuracy was calculated as:

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

while the Matthews Correlation Coefficient (MCC), a robust metric for binary classification, was computed as:

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

IV. RESULTS ANDDISCUSSION

The validation accuracy comparison of multiple state of-the-art deep learning architectures and the proposed multimodal ensemble framework for oral lesion classification. Among the individual image-based models, EfficientNet-B0 achieved the highest validation accuracy of 79.37%, outperforming MobileNetV3-Large (77.51%), ResNet-50 (78.57%), and DenseNet-121 (75.66%), which can be attributed to Efficient Net's compound scaling strategy that balances depth, width, and resolution for efficient feature extraction. Although the image-only models demonstrated competitive performance, their predictions were limited in cases where visual cues alone were insufficient to distinguish potentially malignant disorders from benign conditions.

The proposed ensemble model, which integrates EfficientNet-B0 image predictions with patient metadata using a clinically guided fusion mechanism, achieved the highest validation accuracy of 83.33%, representing a significant improvement over all standalone image models. This performance gain highlights the effectiveness of incorporating clinical risk factors such as age, gender, and habit history alongside visual features, thereby mimicking real-world clinical decision-making and improving robustness in early oral cancer detection.

	Model	Validation Accuracy
0	MobileNetV3-Large	0.7751
1	ResNet50	0.7857
2	DenseNet121	0.7566
3	EfficientNetB0	0.7937
4	Proposed Ensemble (Image + Metadata)	0.8333

Fig 2: Models Trained

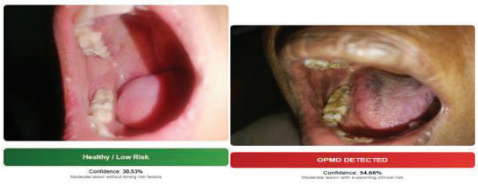

Fig 3: results

qualitative prediction examples generated by the proposed multimodal framework for oral cancer screening. The left image corresponds to a healthy oral cavity, where the model predicts a *Healthy / Low Risk* outcome with a low confidence score, indicating the absence of prominent visual abnormalities. In contrast, the right image illustrates an oral cavity exhibiting visible lesion characteristics, such as irregular tissue texture and discoloration, which are commonly associated with oral potentially malignant disorders (OPMD). The system correctly identifies this case as *OPMD Detected* with high confidence. These qualitative results complement the quantitative evaluation metrics by demonstrating the model's ability to differentiate benign and OPMD.

V. CONCLUSION

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

This work presented a multimodal AI-assisted framework for early detection of oral cancer and oral potentially malignant disorders (OPMD) by integrating oral cavity images with patient clinical metadata. Unlike traditional image-only models, the proposed approach combines visual lesion characteristics with key risk factors such as age, gender, smoking, betel quid chewing, and alcohol consumption, thereby reflecting real clinical diagnostic practices. Several deep learning architectures were evaluated for image-based classification, among which EfficientNetB0 achieved the highest standalone performance. Further improvement was obtained through a weighted ensemble strategy that fused image predictions with metadata-based risk estimation. The proposed ensemble model achieved a validation accuracy of 83.33%, outperforming individual CNN models, and demonstrating improved classification reliability, as supported by F1-score and MCC values. These results confirm that multimodal learning significantly enhances early oral cancer detection compared to unimodal approaches.

VI. FUTURE WORK

Future work will focus on expanding the dataset with a larger and more diverse population to improve model generalization and robustness. Incorporating explainable AI techniques, such as attention maps or gradient-based visualization methods, can enhance interpretability and clinical trust in the model's predictions. Additionally, extending the framework to support multi-class classification of different oral lesion types and integrating longitudinal patient data could improve disease progression analysis. Finally, deploying the system as a web-based or mobile screening tool integrated with telemedicine platforms would enable practical, real-time oral cancer screening, particularly in resource-constrained healthcare settings.

VII.REFERENCES:

[1] P. Warnakulasuriya, "Global epidemiology of oral and oropharyngeal cancer," *Oral Oncology*, vol. 45, no. 4–5, pp. 309–316, 2009.

[2] S. Uthoff, A. Song, and E. K. Lacy, "Deep learning for automated detection of oral potentially malignant disorders using clinical photographs," *IEEE Access*, vol. 8, pp. 155981–155990, 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017, pp. 4700–4708.

[5] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, 2016, pp. 2818–2826.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[7] J. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[8] T. Rajpurkar et al., "Deep learning for medical image analysis: Challenges and applications," *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[9] M. Aubreville et al., "Automatic classification of cancerous tissue in the oral cavity using convolutional neural networks," *IEEE Access*, vol. 8, pp. 102–113, 2020.

[10] Q. Fu et al., "Deep learning-based approach for oral cancer detection using photographic images," *Cancers*, vol. 12, no. 6, pp. 1–14, 2020.

[11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[12] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[13] S.-C. Huang et al., "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutionalnetworks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[15] C. Szegedy et al., "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[17] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1314–1324.