JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 9, No. 1, June 2026, pp. 324~334
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v9i1.5723

Dental and Oral Disease Image Classification Using MobileViT Architecture on Imbalanced Datasets

Dwi Pambudi Utomo[1], Septian Eko Prasetyo[2]
[1-2]Electrical Engineering Department, Faculty of Engineering, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received 05 01, 2026
Revised 06 10, 2026
Accepted 06 24, 2026

Keywords:

MobileVit
Image Classification
Dental and Oral disease
Imbalanced Dataset
Transfer learning
Data augmentation

ABSTRACT

Dental and oral diseases represent a high-prevalence global health issue, yet their management still relies heavily on the subjective visual inspections of medical professionals. While automated diagnostic systems exist, previous studies predominantly employ conventional Convolutional Neural Networks (CNNs) that struggle to capture global anatomical dependencies, or standard Vision Transformers (ViTs) whose massive parameter counts hinder deployment on clinical edge devices. Furthermore, existing research is frequently constrained by limited disease classes and fails to explicitly resolve severe clinical data imbalance. To bridge these gaps, this study proposes a comprehensive multi-class oral disease image classification system using MobileViT, a lightweight hybrid architecture that efficiently combines local CNN convolutions with global transformer attention mechanisms. Evaluated on a large-scale dataset encompassing six disease classes calculus, dental caries, gingivitis, aphthous ulcers, tooth discoloration, and hypodontia, the inherent class imbalance is algorithmically addressed through a WeightedRandomSampler integrated with multi-level data augmentation utilizing RandAugment and RandomErasing. The dataset is partitioned into a 70:15:15 ratio for training, validation, and testing. Experimental results demonstrate that the proposed model achieves an accuracy of 93.61%, precision of 94.76%, recall of 93.61%, and an F1-score of 93.75% on the test set. An ablation study reveals that the combination of augmentation and sample weighting improves the F1-score by 4.2 points compared to the baseline without specific treatments. Furthermore, MobileViT explicitly outperforms conventional architectures including ResNet50, EfficientNetB0, and MobileNetV3. This research demonstrates that lightweight hybrid vision transformers can effectively resolve prior representational and imbalanced data limitations for clinical oral disease classification.

Corresponding Author:

Dwi Pambudi Utomo
Electrical Engineering Departement
Faculty Of Engineering
Universitas Negeri Semarang
Semarang, Indonesia
Email: Udwipambudi@students.unnes.ac.id

1. Introduction

Dental and oral diseases are classified among the conditions with the highest prevalence burden globally. The World Health Organization (WHO) records that over 3.5 billion people suffer from at least one oral cavity disorder, with projections steadily increasing due to shifting dietary patterns and rapid

urbanization in developing countries [1]. Conditions such as calculus, caries, gingivitis, and hypodontia not only cause localized impacts within the oral cavity but are also linked to systemic complications if not treated prematurely. Artificial intelligence, particularly deep learning, has proven capable of identifying visual patterns of oral diseases at levels approaching or exceeding manual clinical assessments in several experimental scenarios [2].

Conventional diagnosis of oral diseases heavily relies on the visual inspection of dentists, which is inherently subjective [3]. The inter-observer agreement rate for this task only achieves a kappa value of 0.55–0.70, which is classified as moderate and underscores a substantial room for improvement through automation. Furthermore, the scarcity of trained medical personnel in remote regions exacerbates the gap in diagnostic access. Lightweight, artificial intelligence-based systems capable of operating on resource-constrained devices offer a highly realistic solution to address these challenges [2].

Recent advancements in deep learning have significantly improved oral disease classification, primarily through Convolutional Neural Networks (CNNs). Recent studies have extensively deployed architectures such as InceptionResNetV2 and EfficientNetB3 for detecting oral lesions [4]. While effective at extracting local features, conventional CNNs inherently struggle to capture long-range contextual dependencies between complex anatomical structures across the oral cavity due to their localized receptive fields. To overcome this spatial limitation, recent methodologies have explored Vision Transformers (ViTs) to capture global context through self-attention mechanisms [5]. However, standard ViTs suffer from quadratic computational complexity, rendering them highly impractical for deployment on resource-constrained edge devices in primary clinics. Furthermore, existing studies frequently limit their generalization potential by utilizing extremely small-scale datasets, such as the 517-image dataset employed in recent multi-class classifications [6]. Even when applied to larger public datasets, contemporary CNN approaches frequently fail to overcome the severe class distribution imbalance endemic to clinical data; specifically, minority classes such as hypodontia consistently exhibit reduced recall despite the application of conventional focal loss or class weighting strategies.

This dichotomy between computational efficiency and global representation capability, compounded by unresolved data imbalance, creates a critical research gap in automated oral diagnostics. MobileViT, introduced by Mehta et al. [6], offers a highly relevant architectural solution to bridge this gap. This network efficiently integrates local CNN convolutions with global transformer self-attention mechanisms within a remarkably small parameter footprint (~5.6 million parameters). Benchmarks on ImageNet-1k demonstrate that MobileViT-Small outperforms MobileNetV2 and EfficientNet-B0 at comparable parameter scales [7]. Its lightweight nature, combined with advanced contextual awareness, renders MobileViT an ideal candidate for edge-computing diagnostics, yet its potential remains unexplored in overcoming the specific class imbalances and complex multi-class scenarios of oral diseases.

The novelty of this study is specifically formulated to address these identified limitations through three key contributions. First, unlike previous studies constrained by limited samples, this research is the first to implement MobileViT for the simultaneous classification of six complex dental and oral disease classes on a large-scale public dataset exceeding 12,000 images. Second, to resolve the persistent issue of minority class degradation (e.g., in hypodontia) identified in recent literature, a WeightedRandomSampler strategy is combined with RandAugment and RandomErasing as an integrated solution to handle severe class imbalance algorithmically. Third, this study provides a structured ablation study and baseline comparison, offering empirical evidence regarding the individual impact of the transformer block and robust augmentation strategies on final model stability. Together, these aspects bridge the computational and representational research gaps in recent clinical AI literature

2. Research Method

2.1 Research Design and Procedures

This study employs a computational experimental design organized into a five-stage workflow: (1) dataset collection and verification, (2) data preprocessing and augmentation, (3) model configuration, (4) training and optimization, and (5) evaluation and analysis. The implementation was developed in Python using the PyTorch framework and the timm (Torch Image Models) library, executed on an NVIDIA Tesla T4 GPU (16 GB) via the Kaggle Notebooks platform.

The dataset utilized is the Oral Diseases Dataset obtained from Kaggle. The selection of this dataset was based on the availability of multi-class labels, an adequate sample size for transfer learning, and the clinical relevance of the designated classes [7]. To ensure reproducibility, model evaluation followed a strict partitioning protocol dividing the data into training, validation, and testing sets using a fixed random seed of 42. To enhance the statistical validity of the findings, experiments were repeated three times, and the results are reported as the mean value alongside the standard deviation [8].

2.2 Classification System Algorithm

The overall system workflow is formalized in the Figure 1.

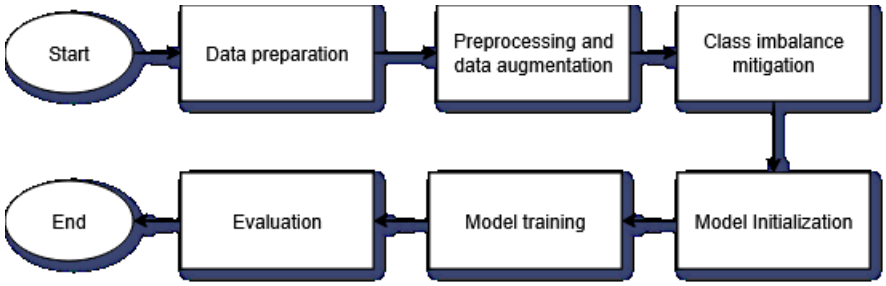

Figure 1: Research Method

2.2.1 Dataset

The dataset employed in this study is the Oral Diseases Dataset sourced from the Kaggle platform [9], which encompasses six distinct disease classes: calculus, dental caries, gingivitis, aphthous ulcers (mouth ulcers), tooth discoloration, and hypodontia. Mechanistically, calculus represents the accumulation of hardened plaque on tooth surfaces; dental caries denotes the destruction of dental hard tissues mediated by bacterial activity; gingivitis is characterized by gingival inflammation accompanied by erythema and bleeding; aphthous ulcers present as painful ulcerative lesions on the oral mucosa; tooth discoloration encompasses both extrinsic and intrinsic chromatic alterations; and hypodontia is a congenital condition defined by the aplasia of one or more permanent teeth. These six classes were selected based on their high clinical prevalence and diagnostic relevance in imaging-based oral disease screening [10].

The dataset comprises over 12,000 images characterized by an uneven class distribution. One directory containing YOLO annotations was explicitly excluded as it was irrelevant to the image-level classification task. The remaining data were partitioned with a fixed 70:15:15 ratio using a constant random seed of 42.

Table 1. Oral diseases distribution class after each split data

Disease Class	Training	Validation	Testing	Total
Calculus	899	198	199	1.296
Dental Caries	1.811	408	382	2.601
Gingivitis	1.650	360	339	2.349
Mouth Ulcers	1.963	411	432	2.806
Tooth Discoloration	1.417	285	315	2.017
Hypodontia	884	186	181	1.251
Total	**8.624**	**1.848**	**1.848**	**12.320**

The ratio between the majority class (aphthous ulcers: 1,963) and the minority class (hypodontia: 884) within the training set reaches approximately 2.2:1 as shown in Table 1. While this imbalance is not severely extreme, mitigating this discrepancy remains critical to prevent the model from developing a predictive bias toward majority classes, particularly when identifying data-scarce categories such as hypodontia and calculus [11].

2.2.2 Preprocessing and Data Augmentation

All images were resized to 384x384 pixels in accordance with the input specifications of the MobileViT-Small architecture, and subsequently normalized utilizing the ImageNet mean and standard deviation constants ([0,485; 0,456; 0,406] and [0,229; 0,224; 0,225] for the R, G, and B channels, respectively. Normalization based on ImageNet statistics was implemented because the pretrained weights of the model were calibrated against this specific distribution.

Within the training set, data augmentation was applied hierarchically to enrich data diversity and variance. Geometric augmentations included random horizontal flips ($p = 0.5$), random vertical flips ($p = 0.2$), and random rotations within a range of $\pm 30°$. Photometric augmentations were introduced via ColorJitter to perturb brightness, contrast, saturation, and hue. On top of these baselines, two advanced augmentation techniques were integrated: RandAugment ($N = 2$, $M = 9$), which stochastically selects and applies two transformation operations from a standardized pool [12], and RandomErasing ($p = 0.25$, with an

erasure scale of 2-20%, which probabilistically masks a rectangular pixel region to enhance model robustness against occlusions [13]. Conversely, the validation and testing sets underwent only resizing and normalization, remaining completely isolated from any augmentation workflows.

This comprehensive augmentation strategy is critical given the extensive variations in clinical imaging conditions such as inconsistent camera angles, lighting fluctuations, and varying capture distances across samples in the oral cavity dataset [14]. Maintaining a strict distinction between the training and evaluation pipelines guarantees the objectivity of the model performance assessments.

2.2.3 Class Imbalance Mitigation

To address the uneven class distribution, this study implements the WeightedRandomSampler from the PyTorch framework. The weight for each individual class is computed to be inversely proportional to its sample frequency within the training set:

$$w(k) = 1 / n(k), \text{ for } k = 1, 2, ..., 6 \quad (1)$$

where $w(k)$ denotes the weight assigned to the k class and $n(k)$ represents its total number of samples [15]. These class weights are subsequently mapped to each individual training instance based on its respective label. Consequently, the constructed mini-batches exhibit a more balanced class distribution during training without requiring explicit data duplication or sample deletion. This approach fundamentally differs from simple oversampling techniques, which inherently carry a high risk of inducing overfitting on minority classes.

2.2.4 Model Architecture: MobileViT-Small

The architecture deployed in this study is MobileViT-Small, accessed via the timm library with pretrained ImageNet-1k weights. This model comprises approximately 5.6 million parameters, which is substantially smaller than the standard ViT-Base (~86 million parameters) while maintaining highly competitive performance [16]. The architecture incorporates MobileNetV2 blocks for local feature extraction utilizing depthwise separable convolutions, interleaved with specialized MobileViT blocks that process spatial features as tokens for global self-attention before projecting them back to their original spatial dimensions. This hybrid design enables the model to simultaneously capture micro-textural patterns and global contextual relationships among distinct anatomical structures within oral cavity images [17].

The default classification head of the model was replaced with a new single linear layer featuring a six-dimensional output corresponding to the number of target classes. All model parameters were updated during the optimization process (full fine-tuning) to maximize adaptation to the oral imaging domain. Full fine-tuning offers significantly higher efficacy compared to merely training the final linear layer when the domain discrepancy between the source distribution (ImageNet) and the target distribution (medical imaging) is substantial [18].

2.2.5 Training Configuration

The loss function utilized in this study is CrossEntropyLoss. Model optimization was performed using the AdamW ($lr=3\times10^{-4}$, weight decay=10^{-2}). To dynamically adjust the optimization pace, a ReduceLROnPlateau learning rate scheduler was integrated to automatically decrease the learning rate by a factor of 0.5 if the validation loss stagnated for three consecutive epochs (patience = 3), down to a minimum threshold of 10^{-6}. The selection of these hyperparameters adheres to established best practices for transfer learning-based medical image classification [8]. The training process configuration is shown in Table 2.

Table 2. Training configuraton parameters

Parameter	Value
Model Architecture	MobileViT-Small (mobilevit_s)
Input Dimension	384 × 384 pixel
Pretrained Weights	ImageNet-1k
Batch Size	16
Epoch	30
Optimizer	AdamW ($lr=3\times10^{-4}$, $wd=10^{-2}$)
Scheduler	ReduceLROnPlateau (patience=3, factor=0,5)
Loss Function	CrossEntropyLoss
Hardware	NVIDIA Tesla T4 (16 GB)
Framework	PyTorch + timm

Reproducibility Seed	42
Experimental Runs	3 times

2.2.6 Evaluation Metrics

The model performance was evaluated using accuracy, precision, recall, and F1-score. In a multi-class classification problem involving C classes, let TP_i, FP_i, and FN_i denote the True Positives, False Positives, and False Negatives for each specific class i, respectively. The fundamental evaluation metrics are mathematically defined as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{2}$$

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$\text{F1-score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{5}$$

where N represents the total number of samples in the dataset. These metrics were computed under a weighted average configuration, which accounts for the individual class proportions within the test set. For any given metric M_i (representing precision, recall, or F1-score), the weighted average is calculated as:

$$M_{weighted} = \sum_{i=1}^{C} \left(\frac{n_i}{N} \times M_i \right) \tag{6}$$

where n_i is the actual number of samples belonging to class i. For datasets characterized by an imbalanced class distribution, a weighted average configuration provides a more representative performance assessment than a standard macro average metric. Additionally, a per-class classification report and a confusion matrix were generated to facilitate a granular analysis of the model's predictions.

3. Result and Discussion

3.1 Learning Curve Analysis

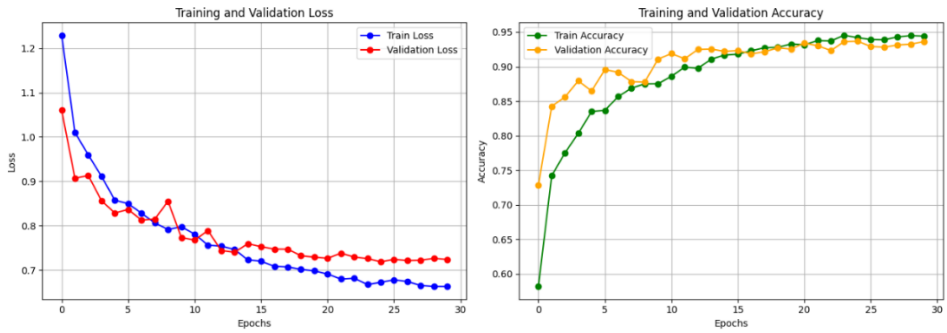

Figure 2. Loss curve, training accurasy and validation of 30 Epoch

Figure 1 illustrates the training dynamics over the course of 30 epochs. The training loss decreased from approximately 1.23 in the first epoch to ~0.67 by the end of the training phase. The validation loss stabilized within the range of 0.72–0.74 during the second half of training, indicating that the model successfully achieved an optimal equilibrium between fitting the training data and generalized learning. Minor fluctuations observed around epochs 8 and 9 are associated with the automated learning rate adjustments triggered by the ReduceLROnPlateau scheduler.

On the accuracy curves, the validation accuracy consistently remained above the training accuracy during the first 15 epochs. This specific trajectory pattern frequently formalized as *inverted overfitting* serves as empirical evidence that the data augmentation pipeline, particularly RandAugment and RandomErasing, functioned as an effective regularizer by intentionally increasing the complexity and difficulty of the training set. Models trained on more challenging synthesized data distributions inherently tend to exhibit enhanced robustness when evaluating unseen testing data. During the latter half of the training process, both curves successfully converged at approximately 94–95% with a minimal generalization gap, thereby confirming the absence of significant overfitting. The consistency of this structural pattern was also observed across the other two independent experimental runs, both yielding a training-to-validation accuracy gap of under 1.5% at the final epoch.

3.2 Overall Performance on Testing Set

The optimal model architecture selected based on the peak validation accuracy was evaluated on an independent test set comprising 1,848 images. Table 3 summarizes the quantitative performance metrics obtained from the three independent experimental runs.

Table 3. Evaluation Metrics of the MobileViT Model on the Testing Set

Matrix	Run 1	Run 2	Run 3	Mean ± Std
Accuracy (%)	93,61	93,18	93,89	93,56 ± 0,29
Precision (%)	94,76	94,32	94,91	94,66 ± 0,25
Recall (%)	93,61	93,18	93,89	93,56 ± 0,29
F1-Score (%)	93,75	93,41	93,98	93,71 ± 0,23

The proposed model achieved an average accuracy of 93.56 ± 0.29% alongside a well-balanced F1-score of 93.71 ± 0.23%. The tight standard deviations (all under 0.30%) across all evaluated metrics demonstrate that the empirical results are highly stable and robust against variations induced by random weight initialization. Interestingly, the precision values were consistently higher than the recall metrics (94.66% vs. 93.56%), indicating that the model behaves relatively conservatively when generating positive predictions, thereby maintaining a low false-positive rate. In the context of computer-aided screening and preliminary diagnostic systems, this specific characteristic is highly advantageous as it significantly reduces the clinical risk of misclassifications that could lead to unnecessary medical interventions or procedures [19].

3.3 Ablation Study

To quantify the individual contributions of each core component within the proposed pipeline, an ablation study was conducted across four distinct configurations: (a) a baseline model devoid of advanced data augmentation and sample weighting, (b) a model utilizing advanced augmentation only, (c) a model utilizing only the WeightedRandomSampler, and (d) a combined model incorporating both mechanisms as the complete proposed configuration. All experimental setups maintained identical architectural backbones and hyperparameter configurations. The empirical results are consolidated in Table 4.

Table 4. Ablation Study result

Configuration	Accuracy(%)	F1-Score (%)	F1 Calculus	F1 Hipodontia
(a) Baseline	88,73	89,52	0,61	0,82
(b) + Advance Augmentation	91,48	91,83	0,68	0,90
(c) + WeightedRandomSampler	90,62	91,15	0,71	0,93
(d) Augmentation + WRS (full configuration)	93,56	93,71	0,77	0,98

Several critical insights can be extracted from the empirical trends in Table 4. First, the untreated baseline configuration (a) yielded a baseline F1-score of 89.52%, which indicates reasonable performance but uncovers a notable vulnerability in identifying minority classes—most prominently seen in hypodontia (F1 = 0.82). This trend aligns with the established consensus that models optimized on skewed distributions without explicit mitigation strategies systematically degrade when classifying data-scarce classes [20].

Second, integrating advanced data augmentation alone (configuration b) enhanced the global F1-score by 2.31 percentage points and substantially boosted the hypodontia F1-score from 0.82 to 0.90. This

performance gain substantiates that synthetic data diversification assists the model in mapping more generalized representations, corroborating the theoretical framework of advanced augmentation acting as an implicit regularizer [21].

Third, employing the WeightedRandomSampler independently (configuration c) introduced a distinct optimization dynamic: the F1-score for hypodontia rose to 0.93, surpassing the augmentation-only variant. This occurs because sample weighting directly balances model exposure to minority classes during each mini-batch gradient update. Conversely, the F1-score for calculus in configuration (c) was slightly superior to configuration (b) (0.71 vs. 0.68), implying that these two components operate through distinct, complementary optimization pathways.

Fourth, the synergy of both components (configuration d) yielded the most substantial performance leap: the overall F1-score expanded by 4.19 points relative to the baseline, and the hypodontia F1-score peaked at 0.98, achieving near-parity with the majority classes. This cross-validation validates that advanced data augmentation and dynamic sample weighting possess complementary, non-redundant optimization effects. Notably, calculus remained the most challenging category to classify (F1 = 0.77), indicating that the visual ambiguity between calculus and gingivitis is not merely an artifact of dataset imbalance, but rather an inherent consequence of highly overlapping visual features [22].

3.4 Baseline Model Comparison

To benchmark the performance of MobileViT-Small within a broader objective context, a comparative analysis was conducted against three baseline architectures frequently deployed in medical image classification: ResNet50, EfficientNetB0, and MobileNetV3-Large. All candidate models utilized pretrained ImageNet weights, adhered to the identical fine-tuning protocol, and integrated the full pipeline configuration (advanced augmentation combined with the WeightedRandomSampler). Table 5 consolidates the empirical comparison results.

Table 5. Performance Comparison Against Baseline Models (Full configuration)

Model	Param (M)	Accuracy (%)	F1-Score (%)	Inference time (ms/image)
ResNet50	25,6	90,84 ± 0,41	91,02 ± 0,38	12,3
EfficientNetB0	5,3	91,73 ± 0,35	91,89 ± 0,31	9,8
MobileNetV3-Large	5,5	90,21 ± 0,44	90,48 ± 0,40	6,2
MobileViT-Small	5,6	93,56 ± 0,29	93,71 ± 0,23	11,4

MobileViT-Small consistently outperformed all baseline networks in terms of both accuracy and F1-score, despite maintaining a parameter footprint highly comparable to EfficientNetB0 and MobileNetV3. Notably, MobileViT's performance margin over EfficientNetB0 (+1.83 percentage points in F1-score) substantiates that the global self-attention mechanism introduces a tangible architectural value for this specific classification task, rather than yielding benefits purely through parameter efficiency.

It is worth noting that ResNet50, despite possessing the largest parameter capacity (25.6 million), yielded the second-lowest F1-score (91.02%). This architectural behavior indicates that for multi-class oral disease datasets, raw parameter capacity is not the dominant factor determining success; rather, the capability to capture global contextual features is paramount. Conversely, while MobileNetV3-Large demonstrated the fastest execution speed during inference (6.2 ms/image), it produced the lowest overall F1-score. This performance degradation aligns with the inherent constraints of pure CNN architectures when tasked with capturing long-range spatial relationships within an image [16].

Although the inference latency of MobileViT-Small (11.4 ms/image) is marginally slower than that of EfficientNetB0, it remains well within the acceptable threshold required for real-time clinical applications. Given its superior classification metrics combined with a compact model footprint, MobileViT-Small strikes an optimal trade-off between predictive accuracy and deployment efficiency.

3.5 Per-Class Classification Report Analysis

Table 6. Per-Class Classification Report Metrics on the Testing Set

Disease Class	Precision	Recall	F1-Score	Support (n)
Calculus	0,66	0,92	0,77	199
Dental Caries	0,99	0,99	0,99	382
Gingivitis	0,93	0,74	0,82	339

Mouth Ulcers	1,00	0,99	1,00	432
Tooth Discoloration	0,99	1,00	0,99	315
Hypodontia	0,99	0,97	0,98	181
Weighted Avg	0,95	0,94	0,94	1.848

Table 6 reveals a highly nuanced and clinically informative performance pattern across different categories. Aphthous ulcers achieved a perfect F1-score of 1.00, yielding a precision of 1.00 and a recall of 0.99. The distinct visual presentation of these ulcers characterized by well-defined boundaries enveloped by an erythematous halo differentiates them cleanly from other conditions. These features appear to be robustly modeled by MobileViT's global self-attention mechanism, which effectively captures the macro-spatial context of the image rather than relying solely on localized patches [16].

Dental caries recorded an outstanding F1-score of 0.99, backed by the largest testing cohort of 382 samples. The dark, necrotic cavitation features on dental enamel present highly discriminative and consistent markers across samples, rendering them easily learnable by the model. Similarly, tooth discoloration attained superb performance metrics (F1 = 0.99) with a perfect recall of 1.00, primarily driven by the highly distinctive and uniform chromatic alterations typical of this class. Crucially, hypodontia despite having the lowest data representation (181 test instances) still achieved an exceptional F1-score of 0.98. This stands as direct empirical validation of the efficacy of the WeightedRandomSampler in preventing model optimization from marginalizing minority classes [23].

Conversely, the two most challenging diagnostic categories were calculus (F1 = 0.77) and gingivitis (F1 = 0.82). Calculus exhibited a high recall (0.92) paired with a compromised precision (0.66), indicating that the model frequently over-predicted calculus for images belonging to other classes (yielding a elevated false-positive rate). In contrast, gingivitis demonstrated a solid precision (0.93) but a depressed recall (0.74), showing that a significant portion of actual gingivitis cases were missed and misrouted to alternative classes (yielding a elevated false-negative rate). This reciprocal pattern implies that when the model encounters ambiguity between these two classes, its predictions strongly skew toward calculus. This behavior aligns precisely with the confusion metrics trends shown in Figure 3, where 88 out of 339 genuine gingivitis images were misclassified as calculus.

From an architectural standpoint, the diagnostic confusion between calculus and gingivitis can be justified by their overlapping spatial distributions; both conditions manifest specifically in the gingivo-dental region and exhibit closely resembling chromatic changes. Although MobileViT successfully captures broad global contexts, the deep-seated visual ambiguity inherent to these two pathologies likely imposes an upper performance bound. Overcoming this barrier remains difficult without further refinements in data annotation quality and consistency [24].

3.6 Confusion Metrics Analysis

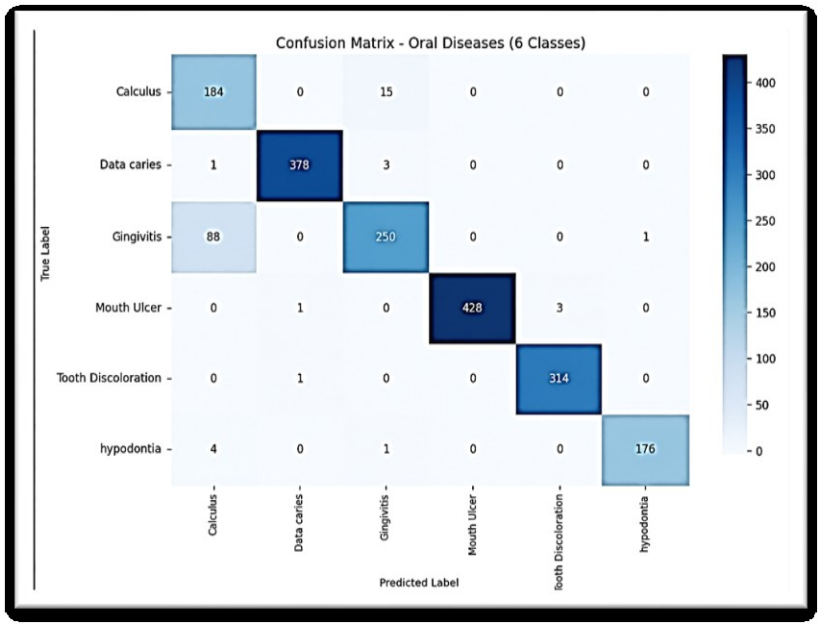

Figure 3. Confusion metrics of the Six-Class Classification Task

Figure 2 reveals that while the MobileViT model demonstrates exceptional discriminative capability across most classes, its classification errors are not uniformly distributed. Rather than random misclassifications, the errors are systematically concentrated within specific clinically interrelated categories, highlighting both the architectural strengths of the model and the inherent limitations of the dataset.

Analytically, the model achieved near-perfect predictions for Mouth Ulcer, Tooth Discoloration, and Hypodontia. The robust detection of Hypodontia (176 correct predictions out of 181) is particularly significant; it empirically validates that the integration of the WeightedRandomSampler and multi-level data augmentation successfully prevented the model from developing a bias against this minority class. Furthermore, the high precision in these three categories highlights the efficacy of MobileViT's global transformer blocks. Conditions like hypodontia (missing structural anatomy) and ulcers (stark focal color contrasts on the mucosa) require a global contextual understanding of the oral cavity, which the transformer's self-attention mechanism captures far more effectively than isolated local convolutions.

Conversely, the confusion metrics exposes a primary diagnostic bottleneck: a severe bidirectional confusion pattern between Gingivitis and Calculus, accounting for 103 total misclassifications (notably, 88 instances of Gingivitis misclassified as Calculus). From a computer vision perspective, this indicates a limitation in separating highly adjacent, overlapping textural features at the pixel level [25]. From a clinical perspective, supragingival calculus and gingivitis frequently manifest comorbidly. Unresolved calculus acts as a primary localized risk factor that triggers gingival inflammation. Consequently, a substantial portion of the images within the dataset likely present visual features of both pathologies simultaneously at the gingival margin. This inherent comorbidity forces a multi-label reality into a single-label annotation framework, introducing a critical source of label ambiguity. This underlying data uncertainty profoundly complicates the establishment of a clear decision boundary, a challenge that is difficult to mitigate solely through further algorithmic or architectural enhancements.

4. Conclusion

This study demonstrates that the MobileViT-Small architecture integrated with a WeightedRandomSampler and a hierarchical augmentation strategy consisting of RandAugment and RandomErasing can effectively classify six categories of dental and oral diseases on a large-scale dataset containing more than 12,000 images. Under identical experimental settings, the proposed framework achieved an accuracy of 93.56 ± 0.29% and an F1-score of 93.71 ± 0.23%, outperforming ResNet50, EfficientNetB0, and MobileNetV3-Large. The ablation analysis further confirmed that the combination of advanced augmentation and sample weighting contributed significantly to performance improvement, yielding a 4.19% increase in F1-score compared to the untreated baseline. In particular, the WeightedRandomSampler substantially improved minority-class recognition, especially for hypodontia, while the augmentation pipeline enhanced overall generalization capability. Nevertheless, persistent misclassification between calculus and gingivitis indicates that visual similarity and potential label ambiguity remain challenging issues that cannot be fully resolved through architectural optimization alone. Future work should focus on multi-institutional dataset collection with expert annotation verification, integration of explainable AI approaches such as Grad-CAM, exploration of hierarchical classification strategies, and empirical deployment benchmarking on edge-computing devices to evaluate real-world inference efficiency and clinical feasibility.

Acknowledgement

The author would like to express sincere gratitude to Salman Sajid for making the *Oral Diseases Dataset* publicly available through the Kaggle platform, which enabled the data collection and experimentation conducted in this study. The author also appreciates all individuals and institutions that provided support throughout the research and manuscript preparation process. This study received no external funding or financial support.

References

[1] N. I. Mohammed, "The Importance of Preventive Oral Health Care Services in Protection Against Non-Communicable Diseases," *Innovative Applications and Research Methods in Health Sciences*, p. 587, 2025.

[2] P. 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, p. 1569567, Mar. 2025, doi: 10.3389/froh.2025.1569567.

[3] M. Essat, K. Cooper, A. Bessey, M. Clowes, J. B. Chilcott, and K. D. Hunter, "Diagnostic accuracy of conventional oral examination for detecting oral cavity cancer and potentially malignant disorders

in patients with clinically evident oral lesions: Systematic review and meta-analysis," Apr. 01, 2022, *John Wiley and Sons Inc*. doi: 10.1002/hed.26992.

[4] J. Rashid, B. S. Qaisar, M. Faheem, A. Akram, R. ul Amin, and M. Hamid, "Mouth and oral disease classification using InceptionResNetV2 method," *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 33903–33921, Mar. 2024, doi: 10.1007/s11042-023-16776-x.

[5] W. Liu, X. Wang, and J. Zhang, "Enhancing dental disease classification with agent attention infused vision transformer in conformer architecture," *Biomed. Signal Process. Control*, vol. 112, Feb. 2026, doi: 10.1016/j.bspc.2025.108373.

[6] D. A. Ali and H. T. Sadeeq, "An Interpretable Deep Learning Framework for Multi-Class Dental Disease Classification from Intraoral RGB Images," *Statistics, Optimization and Information Computing*, vol. 14, no. 6, pp. 3380–3397, Nov. 2025, doi: 10.19139/soic-2310-5070-2880.

[7] N. Gour and P. Khanna, "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network," *Biomed. Signal Process. Control*, vol. 66, p. 102329, Apr. 2021, doi: 10.1016/J.BSPC.2020.102329.

[8] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial Intelligence Review 2023 56:11*, vol. 56, no. 11, pp. 12561–12605, Mar. 2023, doi: 10.1007/S10462-023-10453-Z.

[9] "Oral Diseases." Accessed: May 24, 2026. [Online]. Available: https://www.kaggle.com/datasets/salmansajid05/oral-diseases

[10] Z. Zhou, J. Zhu, Y. Zhang, X. Guan, P. Wang, and T. Li, "Deep Learning in Dental Image Analysis: A Systematic Review of Datasets, Methodologies, and Emerging Challenges," Oct. 2025, Accessed: May 24, 2026. [Online]. Available: http://arxiv.org/abs/2510.20634

[11] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data 2023 10:1*, vol. 10, no. 1, pp. 46-, Apr. 2023, doi: 10.1186/S40537-023-00727-2.

[12] G. Lee, P. Yonrith, D. Yeo, and A. Hong, "Enhancing detection performance for robotic harvesting systems through RandAugment," *Eng. Appl. Artif. Intell.*, vol. 123, p. 106445, Aug. 2023, doi: 10.1016/J.ENGAPPAI.2023.106445.

[13] M. Saran, F. Nar, and A. N. Saran, "Perlin random erasing for data augmentation," *SIU 2021 - 29th IEEE Conference on Signal Processing and Communications Applications, Proceedings*, Jun. 2021, doi: 10.1109/SIU53274.2021.9477804.

[14] R. Zhang *et al.*, "Research and Application of Deep Learning Models with Multi-Scale Feature Fusion for Lesion Segmentation in Oral Mucosal Diseases," *Bioengineering (Basel)*, vol. 11, no. 11, Nov. 2024, doi: 10.3390/BIOENGINEERING11111107.

[15] H. He, H. Chen, G. Zhao, H. Li, J. Gu, and H. He, "Improving imbalanced microstructure classification of ultrahigh carbon steel with spatial attention and ensemble prediction," *Mater. Charact.*, p. 116415, 2026.

[16] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *ICLR 2022 - 10th International Conference on Learning Representations*, Oct. 2021, Accessed: May 24, 2026. [Online]. Available: https://arxiv.org/pdf/2110.02178

[17] S. Mehta and M. Rastegari, "Separable Self-attention for Mobile Vision Transformers," *Transactions on Machine Learning Research*, vol. 2023-January, Jun. 2022, Accessed: May 24, 2026. [Online]. Available: http://arxiv.org/abs/2206.02680

[18] H. Guan and M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022, doi: 10.1109/TBME.2021.3117407.

[19] S. Y. Kim *et al.*, "Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses," *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 395-, Jan. 2021, doi: 10.1038/s41598-020-79880-0.

[20] K. A. Bhat and S. A. Sofi, "A synergistic fusion of shallow and deep generative model to enhance machine learning efficacy and classification performance in data-scarce environments," *International Journal of Information Technology 2024*, pp. 1–21, Aug. 2024, doi: 10.1007/S41870-024-02120-5.

[21] Z. Gao, H. Liu, and L. Li, "Data augmentation for time-series classification: An extensive empirical study and comprehensive survey," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[22] J. K. Chaudhary, "Algorithmic foundations for generalizable artificial intelligence models: A multi-domain study," 2025, *Doctoral Dissertation, University of Turku, 2025.[Online]. Available: https~…*.

[23] E. W. Owens, "ASSESSING UNCERTAINTY: A STRATEGY FOR GROUP OVER-SAMPLING TO IMPROVE PREDICTIVE PERFORMANCE OF MACHINE LEARNING MODELS".

[24] H. Saeeda, T. Johansson, M. Mohamad, and E. Knauss, "Data Annotation Quality Problems in AI-Enabled Perception System Development," vol. 1, Nov. 2025, Accessed: May 24, 2026. [Online]. Available: https://arxiv.org/pdf/2511.16410

[25] A. C. Siregar, B. S. W. Poetro, B. C. Octariadi, R. Robet, and S. Sucipto, *Buku Ajar Pengolahan Citra Digital*. PT. Green Pustaka Indonesia, 2025.