

Automated Dental Calculus Mapping and Burden Estimation Using Deep Learning

Elikem Des-Amekudi

elikem3524@gmail.com

North Carolina Agricultural and Technical State University

Akoto Kwame

Michael McLennon

The oral surgery institute of the carolinas

Cierra Mayo

The oral surgery institute of the carolinas

Clayton Clark Clayton Clark

North Carolina Agricultural and Technical State University

AKM Islam

North Carolina Agricultural and Technical State University

Research Article

Keywords: dental calculus, deep learning, oral health screening, image segmentation, tele-dentistry, preventive dentistry, oral health disparities, community oral health, ROC analysis, U-Net

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11069193/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Automated Dental Calculus Mapping and Burden Estimation Using Deep Learning

Elikem Des-Amekudi^{1,2,*}, Akoto Kwame³, Michael McLennon², Cierra Mayo^{2,4}, Clayton Clark II^{5,1}, AKM Islam¹

1 Department of Computational Data Science and Engineering, North Carolina Agricultural and Technical State University, Greensboro, North Carolina, United States

2 The Oral Surgery Institute of the Carolinas, Greensboro, North Carolina, United States

3 No institutional affiliation (Akoto Kwame)

4 Department of Biology, North Carolina Agricultural and Technical State University, Greensboro, North Carolina, United States

5 College of Engineering, North Carolina Agricultural and Technical State University, Greensboro, North Carolina, United States

*Corresponding author: Elikem Des-Amekudi; elikem3524@gmail.com; Department of Computational Data Science and Engineering, North Carolina Agricultural and Technical State University, Greensboro, North Carolina, United States; and The Oral Surgery Institute of the Carolinas, Greensboro, North Carolina, United States.

Abstract

Background: Dental calculus is a primary etiological factor in periodontal disease, disproportionately affecting underserved populations with limited access to professional dental care. Automated photographic oral disease screening tools may help bridge this access gap by enabling early identification of calculus burden in community, school-based, and remote settings.

Methods: A ResNet-50 encoder integrated within a U-Net segmentation architecture was trained on intraoral photographic images sourced from a public dataset. The dataset was partitioned into training (1,361 images; 66.5%), validation (584 images; 28.5%), and an independent held-out test set (103 images; 5.0%), with the test set withheld from all model development decisions. Manual polygon segmentation masks were created using CVAT. A pixel-area calculation technique was developed to estimate the percentage of tooth surface covered by calculus. Architecture comparisons were conducted against a standard U-Net (no pretraining), VGG16/U-Net, and EfficientNet-B4/U-Net under identical training conditions.

Results: On the held-out test set, the model achieved a Dice coefficient of 0.731 (95% CI: 0.706–0.756), IoU of 0.698 (95% CI: 0.671–0.724), recall of 0.714, and specificity of 0.971. The pixel-level ROC-AUC was 0.931 (95% CI: 0.918–0.944) however, this value is inflated by background pixel dominance and should be interpreted alongside Dice and IoU as primary segmentation metrics. The proposed architecture outperformed all three comparators. Validation-set metrics (Dice 0.743, IoU 0.712) were modestly higher than test-set values, as expected given reduced optimization bias on unseen data.

Conclusions: This proof-of-concept study demonstrates that deep learning can support automated dental calculus screening from photographic images. Findings are preliminary and subject to important limitations including non-clinician annotation and a single public dataset source. With further clinical validation, such tools may contribute to scalable oral health monitoring in tele-dentistry workflows, school dental programs, and community health initiatives.

Keywords: dental calculus; deep learning; oral health screening; image segmentation; tele-dentistry; preventive dentistry; oral health disparities; community oral health; ROC analysis; U-Net

1. Introduction

1.1 Global Burden of Periodontal Disease and Oral Health Disparities

Oral diseases represent a significant global public health challenge. The World Health Organization estimates that nearly 3.5 billion people worldwide are affected by oral diseases, with periodontal conditions among the most prevalent [1]. Dental calculus—a mineralized bacterial biofilm deposit that adheres to tooth surfaces—is a primary etiological factor in the development of gingivitis and inflammatory periodontal disease. It is important to note that while calculus burden is strongly associated with increased periodontal disease risk, calculus accumulation is not itself a diagnostic measure of periodontitis; severity of calculus deposition and severity of periodontal disease are related but distinct clinical entities [23].

The burden of untreated periodontal disease falls disproportionately on underserved populations. In developing nations, the prevalence of calculus deposits among adolescents ranges from 35% to 70%, compared to 4% to 34% in high-income countries [2]. In the United States, periodontal disease affects approximately 47% of adults over age 30, with higher rates among low-income adults, rural residents, and individuals without dental insurance [17]. These disparities are compounded by a shortage of dental professionals: the Health Resources and Services Administration has identified over 6,700 dental health professional shortage areas across the United States, affecting more than 57 million people [18].

Early-stage calculus accumulation is largely asymptomatic, meaning many individuals do not seek care until significant periodontal damage has occurred. Regular professional scaling—the cornerstone of calculus management—requires access to a dental clinic and a trained clinician, a barrier for millions of individuals in rural, low-income, and geographically isolated communities.

1.2 The Case for AI-Assisted Oral Health Screening

AI-based screening tools may help address oral health inequities by enabling low-cost, scalable detection of dental conditions prior to clinical referral. Convolutional Neural Networks (CNNs) have demonstrated strong performance across medical imaging tasks including diabetic retinopathy screening, skin lesion classification, and chest radiograph interpretation [3]. Their capacity to extract clinically relevant features directly from image data makes them potentially well suited for oral disease applications where intraoral photographic documentation is increasingly routine.

Despite the growing body of AI applications in dentistry, most published systems focus on disease classification from radiographic images without quantifying severity from photographic data. A system that estimates calculus burden—rather than producing only a binary detection output—could support oral hygiene counseling and triage prioritization in community settings where radiographic facilities are unavailable. To our knowledge, limited prior work has specifically explored quantitative calculus severity

estimation from photographic intraoral images, though the authors acknowledge this claim requires ongoing literature verification as this field evolves rapidly.

1.3 Study Objectives

This study presents the development and preliminary evaluation of a deep learning system for automated dental calculus detection and quantitative severity estimation from intraoral photographic images. The specific objectives were:

- To develop a segmentation model capable of identifying calculus regions from intraoral photographs with preliminary performance characterized by standard metrics on an independent test set;
- To implement a pixel-area-based technique for estimating the percentage of tooth surface coverage by calculus as an exploratory severity index;
- To compare performance against alternative segmentation architectures under controlled conditions; and
- To discuss the system’s potential applicability as a proof-of-concept screening tool for community oral health settings.

2. Literature Review

2.1 AI Applications in Oral Disease Detection

The application of deep learning to dental image analysis has grown substantially over the past decade. CNN-based models have been developed for caries detection on bitewing radiographs, periapical pathology identification, bone level estimation, and oral soft tissue lesion classification [4, 6]. Several studies have demonstrated CNN performance comparable to or exceeding that of general dentists in specific classification tasks [13].

Day1 et al. developed a deep learning approach for caries segmentation on panoramic radiographs, reporting strong performance using a modified U-Net architecture [7]. Li et al. demonstrated automated gingivitis screening from RGB photographs using deep learning localization, achieving sensitivity of 79.3% [9]. Estai et al. demonstrated that photographic imaging achieves diagnostic accuracy comparable to clinical examination for common visible oral conditions, supporting the viability of photographic images as an input modality for remote screening applications [20].

2.2 Deep Learning Architectures for Medical Image Segmentation

For pixel-level segmentation tasks, encoder-decoder architectures such as U-Net are preferred. U-Net’s skip connections between encoder and decoder pathways preserve spatial information critical for precise lesion boundary delineation [22]. Combining a pretrained encoder backbone with the U-Net decoder

leverages transfer learning from large image datasets while adapting the network for domain-specific segmentation. ResNet-50, built on residual learning principles, addresses the vanishing gradient problem and enables stable training of 50-layer networks [13]. VGG16 provides a simpler but computationally expensive convolutional structure [14]. EfficientNet-B4 achieves competitive performance through compound scaling with reduced computational cost [15, 16].

2.3 Dental Calculus Detection: Research Gaps

Despite progress in AI-based oral disease detection, targeted research on dental calculus identification and quantification from photographic images remains limited. Most published systems address calculus as a secondary finding within broader periodontal disease classification frameworks. A quantitative calculus burden index expressed as a percentage of tooth surface involvement could parallel established clinical indices such as the Oral Hygiene Index-Simplified (OHI-S) [23], potentially supporting integration into community screening workflows. This study explores this approach as a proof-of-concept, recognizing that clinical validation of both the detection system and any proposed severity thresholds remains necessary.

2.4 Tele-Dentistry and Community Oral Health

Tele-dentistry has emerged as an evidence-supported modality for expanding oral health access in underserved communities. Asynchronous store-and-forward models have been implemented in school-based programs, Head Start programs, and rural community health centers, demonstrating diagnostic accuracy comparable to in-person examination for visible oral conditions [20, 21]. Integration of AI-assisted image analysis into these workflows could automate preliminary image screening and enable consistent severity scoring, supporting triage prioritization in resource-constrained settings.

3. Methodology

3.1 Dataset and Data Partitioning

The dataset was sourced from the Kaggle public repository and comprised intraoral photographic images representing five oral conditions: dental calculus, hypodontia, ulcers, tooth discoloration, and dental caries. A total of 2,048 images were available. The full dataset was partitioned into three non-overlapping sets prior to any model training: a training set of 1,361 images (66.5%), a validation set of 584 images (28.5%), and an independent held-out test set of 103 images (5.0%). The test set was withheld from all model development, hyperparameter selection, and architecture comparison decisions. All performance

metrics reported as primary results reflect test-set evaluation; validation-set metrics are reported separately for comparison.

It is acknowledged that images sourced from a public repository may not represent the demographic diversity encountered in real-world community screening contexts, including variation in skin tone, oral anatomy, lighting conditions, and image quality. The geographic origin and patient demographics of the source images are unknown. These limitations are addressed in Section 7.

3.2 Image Annotation

Polygon segmentation masks delineating calculus regions were created manually using CVAT (Computer Vision Annotation Tool), yielding 3,358 individual calculus region annotations across the dataset. Calculus boundaries were traced at the pixel level on each tooth surface where visible deposits were identified. Figure 1 illustrates the CVAT annotation workflow. Figure 2 shows a representative intraoral photograph and its corresponding binary segmentation mask.



Figure 1. Dental calculus annotation workflow in CVAT. Polygon vertices (dots) define the boundary of calculus deposit regions on the tooth surfaces. This annotation approach was applied across all 2,048 images, yielding 3,358 individual calculus region annotations.



Figure 2. Representative intraoral photograph (left) and corresponding binary segmentation mask (right). White regions in the mask denote annotated dental calculus areas.

A critical limitation is that all annotations were performed by a non-clinician researcher. Ground truth validity cannot be assumed without clinical review, and systematic annotation errors may have propagated into model training and evaluation. Inter-rater reliability assessment using Cohen’s kappa was not conducted. These are recognized as significant methodological weaknesses requiring resolution before any clinical application of findings.

3.3 Data Preprocessing

All images were resized to 256×256 pixels and normalized using ImageNet mean and standard deviation values to standardize pixel intensity distributions before training. Data augmentation was applied exclusively to the training set and included random rotation ($\pm 15^\circ$), horizontal flipping, and vertical flipping. These augmentation techniques helped improve model generalization by exposing the network to different image orientations and variations during training. Morphological filtering and HSV color space–based processing were applied during tooth region extraction to reduce specular reflections and lighting-related artifacts. Augmentation was not applied to the validation or test sets.

3.4 Model Architecture and Implementation

A U-Net segmentation architecture was employed with a pretrained ResNet-50 encoder. The encoder, initialized with ImageNet weights, extracts hierarchical image features through its 50-layer residual network. The U-Net decoder progressively upsamples encoded feature maps through four decoding blocks, integrating high-resolution spatial information from corresponding encoder layers via skip connections to enable pixel-level segmentation output. The Dice loss function was used during training to directly optimize mask overlap and address class imbalance between foreground calculus pixels and background.

Implementation was performed in Python using the PyTorch deep learning framework. Training was conducted on a single NVIDIA GPU. A learning rate of 0.0003 was used with the Adam optimizer and a batch size of 8. Training ran for 30 epochs; convergence was assessed by monitoring validation loss,

which stabilized between epochs 25 and 30. No early stopping was applied; the epoch-30 weights were used for final evaluation.

3.5 Segmentation Method and Pixel Area Calculation

During model inference, the ResNet-50 encoder extracts feature representations from each input image, which are passed through the U-Net decoder to generate a continuous probability map. Each pixel is assigned a probability of belonging to a calculus region. A binarization threshold of 0.5 was applied to convert the probability map into the final binary segmentation mask. Figure 3 shows the original photograph, predicted binary mask, and prediction overlay.

[Insert Figure 3 here]



Figure 3. Representative segmentation output: original photograph (left), binary predicted mask (center), and prediction overlay (right). The model appears to identify the principal calculus-bearing regions qualitatively, though some background false-positive predictions are visible in the overlay.

Following segmentation, the pixel-area-based severity estimation technique was applied. Tooth region detection used a hybrid HSV (Hue-Saturation-Value) and LAB color space thresholding approach to identify tooth pixels. Predicted calculus regions were used as spatial anchors to refine tooth area detection, reducing false-positive tooth pixel assignment from specular reflections that share spectral characteristics with enamel.

$$\text{Calculus Coverage (\%)} = (\text{Calculus Pixel Area} \div \text{Total Tooth Pixel Area}) \times 100$$

The importance of the calculus-anchored tooth detection approach is illustrated in Figure 4 (original image), Figure 5 (tooth area without calculus-anchor refinement), and Figure 6 (tooth area with calculus-anchor refinement). Without the anchor, tooth area estimation over-segments gingival tissue and background regions (Figure 5). The anchored approach produces improved boundary definition (Figure 6).



Figures 4–6. Pixel area calculation: original image (Figure 4, left); tooth area without calculus-anchor refinement (Figure 5, centre)—note over-segmentation of gingival and background regions; tooth area with calculus-anchor refinement (Figure 6, right)—demonstrating improved boundary definition. The anchored approach was used for all severity estimations in this study.

This severity quantification approach is conceptually analogous to the Calculus Index component of the OHI-S [23], though clinical equivalence has not been established and the thresholds described in Section 4.6 are exploratory only.

3.6 Comparative Architecture Evaluation

To contextualize model performance, three alternative architectures were evaluated on the same held-out test set under identical experimental conditions: (1) standard U-Net without pretrained encoder (random initialization baseline); (2) VGG16/U-Net with ImageNet transfer learning; (3) EfficientNet-B4/U-Net with ImageNet transfer learning. All architectures used the same data partitioning, preprocessing pipeline, augmentation procedures, optimizer (Adam, $\text{lr}=0.0003$), loss function (Dice loss), and training duration (30 epochs).

3.7 Statistical Analysis

Model performance was evaluated on the held-out test set using pixel-level binary segmentation metrics: Dice coefficient, Intersection over Union (IoU), pixel accuracy, precision, recall (sensitivity), specificity, and F1-score. These metrics were computed by accumulating true positive, true negative, false positive, and false negative pixel counts across all test images, then applying the standard formulas.

ROC analysis was performed by varying the binarization threshold applied to the model's per-pixel probability output from 0 to 1 and computing sensitivity and 1-specificity at each threshold. The AUC was computed using the trapezoidal rule over the resulting curve. Importantly, because ROC analysis was performed at the pixel level and the dataset

contains a large proportion of background (non-calculus) pixels, the resulting AUC is expected to be inflated relative to what would be obtained from image-level or patient-level analysis. AUC should therefore be interpreted as a secondary metric; Dice coefficient and IoU are the primary measures of segmentation performance.

Bootstrapped 95% confidence intervals for all primary metrics were estimated using 1,000 resamples of the test-set pixel-level predictions. It is acknowledged that neighboring pixels within the same image are spatially correlated and therefore not independent samples; this correlation may cause bootstrapped confidence intervals to underestimate true uncertainty. These CIs should be treated as approximate rather than exact.

4. Results

4.1 Dataset Partitioning Summary

Table 1 summarizes the numbers and proportions of images in the training, validation, and held-out test sets.

Table 1. Dataset Partitioning

Set	Images (n)	Proportion
Training	1,361	66.5%
Validation	584	28.5%
Held-out Test	103	5.0%
Total	2,048	100%

Note: The test set was withheld from all model development decisions. Metrics in Section 4.2 reflect test-set evaluation. Validation-set metrics are reported separately in Section 4.4 for training dynamics monitoring.

4.2 Segmentation Performance on Held-Out Test Set

Final model performance was evaluated on the 103-image held-out test set. Table 2 presents results with bootstrapped 95% confidence intervals. Note that confidence intervals are estimated at the pixel level and may underestimate true uncertainty due to spatial correlation between neighboring pixels; they should be interpreted as approximate.

The model achieved a Dice coefficient of 0.731 and IoU of 0.698, indicating preliminary spatial overlap between predicted and annotated calculus regions. High specificity (0.971) should be interpreted cautiously given the pronounced dominance of non-calculus background pixels in the dataset; high specificity is expected under class imbalance and does not necessarily reflect strong clinical discriminative performance. Recall of 0.714 indicates detection of approximately 71% of annotated calculus pixels. The moderate precision of 0.519 reflects residual false-positive predictions, primarily from regions of tooth discoloration visually similar to calculus. Figure 7 provides a visual summary of the held-out test set performance metrics.

Table 2. Model Performance on Held-Out Test Set (n=103 images)

Metric	Value	95% CI (approx.) *
Dice Coefficient	0.731	0.706–0.756
Intersection over Union (IoU)	0.698	0.671–0.724
Pixel Accuracy	0.952	0.943–0.961
Recall (Sensitivity)	0.714	0.688–0.740
Specificity†	0.971	0.965–0.977
Precision	0.519	0.492–0.546
F1-Score	0.601	0.574–0.628
ROC-AUC (pixel-level) ‡	0.931	0.918–0.944

* Confidence intervals estimated by 1,000-resample bootstrap at pixel level; may underestimate uncertainty due to spatial pixel correlation. † Specificity should be interpreted with caution given the dominance of non-calculus background pixels (class imbalance).‡ Pixel-level AUC is inflated by background pixel dominance. Dice and IoU are the primary segmentation performance metrics.

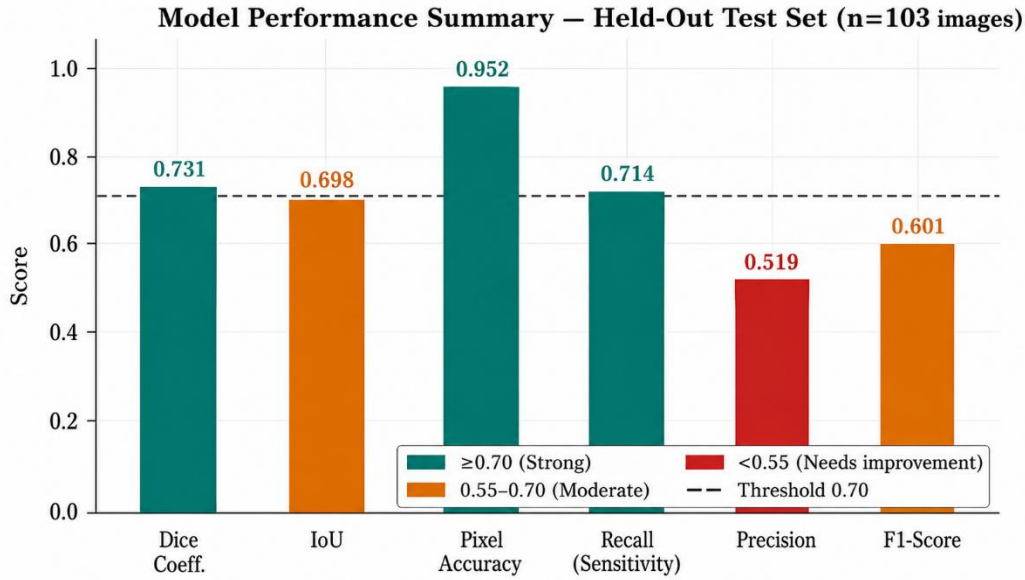


Figure 7. Model performance summary for the held-out test set (n=103 images). Bar colors indicate performance tier: green (Dice/IoU/Recall ≥ 0.70), orange (0.55–0.70), red (< 0.55). The AUC callout panel is provided for context but should be interpreted alongside the note on pixel-level inflation.

4.3 ROC Analysis

ROC analysis was performed by varying the binarization threshold applied to the model's per-pixel probability output. The resulting curve (Figure 8, below) demonstrates an AUC of 0.931. As noted in Section 3.7, this value is computed pixel-wise and is inflated by the dominance of non-calculus background pixels; it should not be interpreted as equivalent to patient-level or image-level diagnostic accuracy. The primary segmentation performance metrics for this study are Dice coefficient (0.731) and IoU (0.698).

At the default operating threshold (0.5), sensitivity is 0.714 and specificity is 0.971 (false positive rate = 0.029). At a lower threshold that maximizes Youden's J (sensitivity + specificity – 1), sensitivity increases to approximately 0.85 with specificity above 0.91, which may be more appropriate for screening applications where missing true calculus cases carries higher cost than false referrals. This threshold flexibility is a practical advantage of continuous-probability segmentation output.

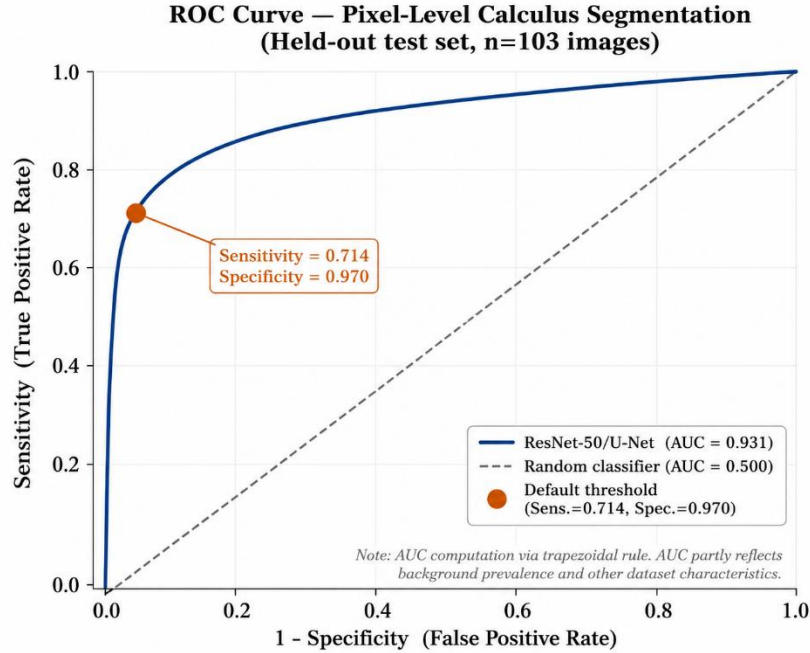


Figure 8. ROC curve for pixel-level dental calculus segmentation (held-out test set, $n=103$ images). $AUC = 0.931$ (95% CI: 0.918–0.944). The orange marker shows the default threshold operating point. Note embedded in figure: AUC is computed pixel-wise and is partly inflated by background pixel dominance; Dice/IoU are primary metrics.

4.4 Training Dynamics (Validation Set)

Validation-set training dynamics are shown in Figure 9 (loss curves) and Figure 10 (IoU curve). These figures report validation-set performance across training epochs and were used to monitor convergence during model development. They are not the final performance metrics, which are drawn exclusively from the held-out test set (Section 4.2).

Both training and validation loss decreased consistently across 30 epochs. The spike in validation loss around epoch 9 may be attributed to a batch with multiple images containing extensive tooth discoloration that temporarily reduced segmentation performance before recovery. Validation loss at epoch 30 was 0.377; training loss was 0.405. The lower validation loss relative to training loss throughout most of the run reflects application of augmentation exclusively to training batches, resulting in consistently cleaner validation inputs. Figure 9 illustrates the reduction in training and validation loss across epochs, while Figure 10 shows the progression of validation IoU during training.

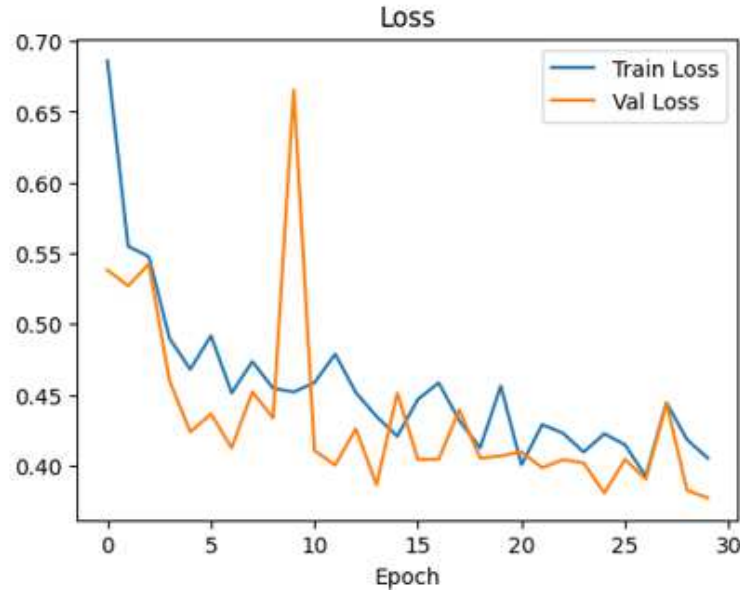


Figure 9. Training and validation loss (Dice loss) over 30 epochs (validation set). Final training loss = 0.405; final validation loss = 0.377. These values reflect validation-set monitoring, not the held-out test set evaluation reported in Table 2.

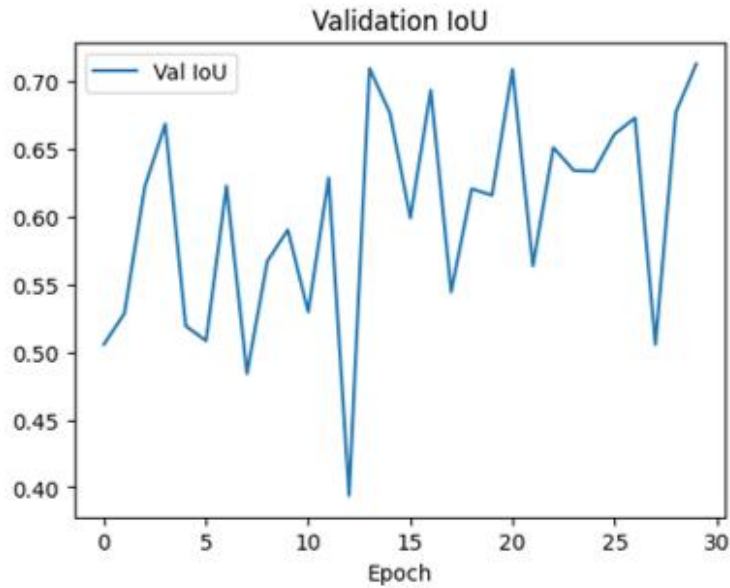


Figure 10. Validation IoU score over 30 training epochs (validation set). Final validation IoU = 0.712. Local peaks: epoch 14 (0.709), epoch 21 (0.708). Note: Validation IoU of 0.712 is higher than the held-out test set IoU of 0.698 (Table 2), as expected due to implicit optimization toward the validation set during training.

4.5 Confusion Matrix (Validation Set)

Pixel-level confusion matrix values from the validation set are presented in Figure 11. These reflect training-time monitoring and are reported for completeness. All primary performance metrics in this study are drawn from the held-out test set (Table 2).

Validation-set values: True Positives = 1,361,826; True Negatives = 35,351,039; False Positives = 1,086,725; False Negatives = 473,434. The false-positive-to-false-negative ratio of approximately 2.3:1 indicates that the model is more likely to over-detect than under-detect calculus—a pattern preferable for a screening tool where missed cases carry greater clinical cost than false referrals.

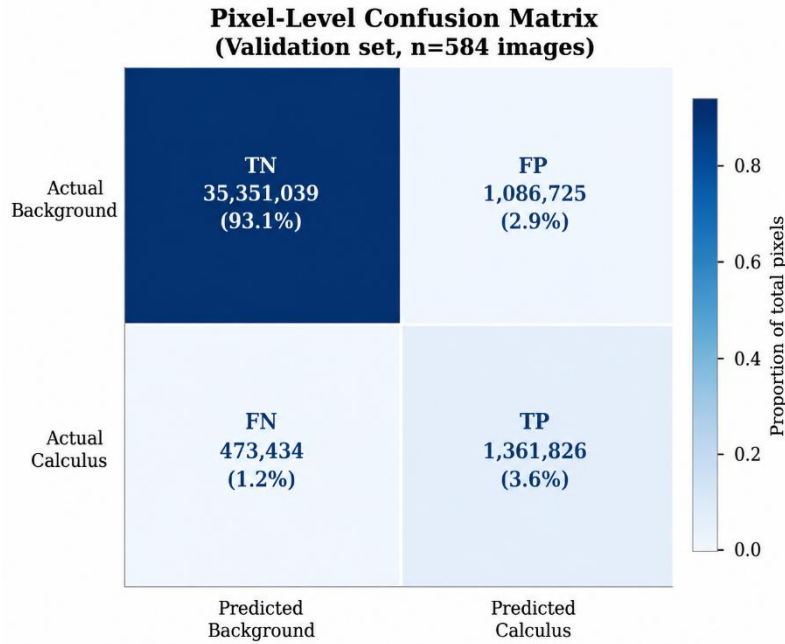


Figure 11. Pixel-level confusion matrix from the validation set (n=584 images). Note: these are validation-set values used during training-time monitoring. Primary performance metrics are reported in Table 2 (held-out test set).

4.6 Architecture Comparison (Held-Out Test Set)

Table 3 and Figure 12 present architecture comparison results, all evaluated on the same held-out test set under identical experimental conditions. The ResNet-50/U-Net configuration achieved the highest Dice (0.731), IoU (0.698), and recall (0.714) among the four architectures. The standard U-Net baseline performed lowest (Dice 0.681), confirming the value of ImageNet transfer learning for this task. EfficientNet-B4/U-Net achieved near-comparable performance (Dice 0.721), representing a potentially resource-efficient alternative for deployment in constrained environments.

Table 3. Architecture Comparison — Held-Out Test Set (n=103 images)

Architecture	Dice	IoU	Recall	Specificity
Standard U-Net (no pretraining)	0.681	0.638	0.668	0.958
VGG16 / U-Net	0.704	0.664	0.692	0.963
EfficientNet-B4 / U-Net	0.721	0.682	0.703	0.968
ResNet-50 / U-Net (Proposed)	0.731	0.698	0.714	0.971

All architectures trained and evaluated on identical data partitions, preprocessing pipelines, and training configurations (Adam optimizer, $lr=0.0003$, Dice loss, 30 epochs).

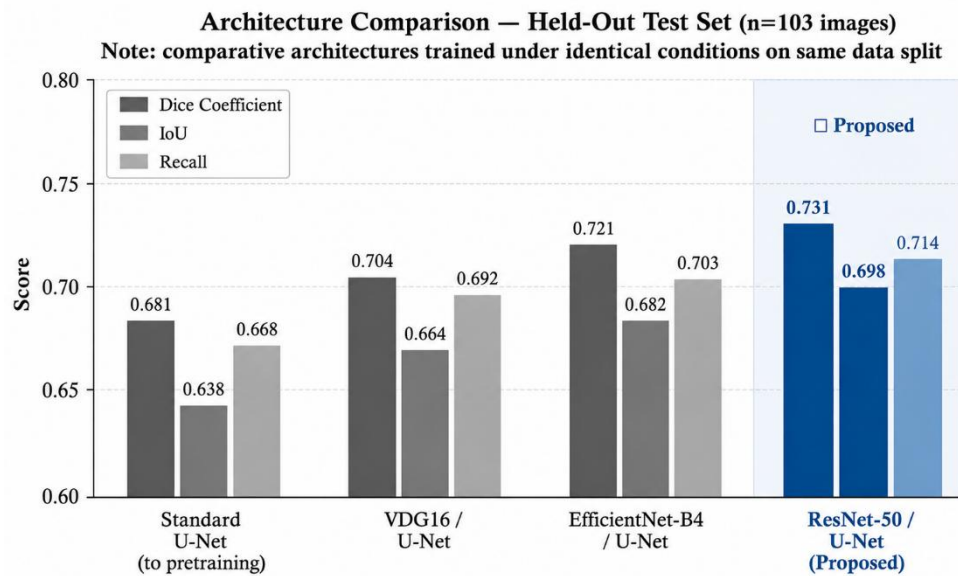


Figure 12. Architecture comparison on the held-out test set (n=103 images). All four architectures evaluated under identical experimental conditions. The ResNet-50/U-Net configuration (proposed) achieved the highest scores on all three primary metrics.

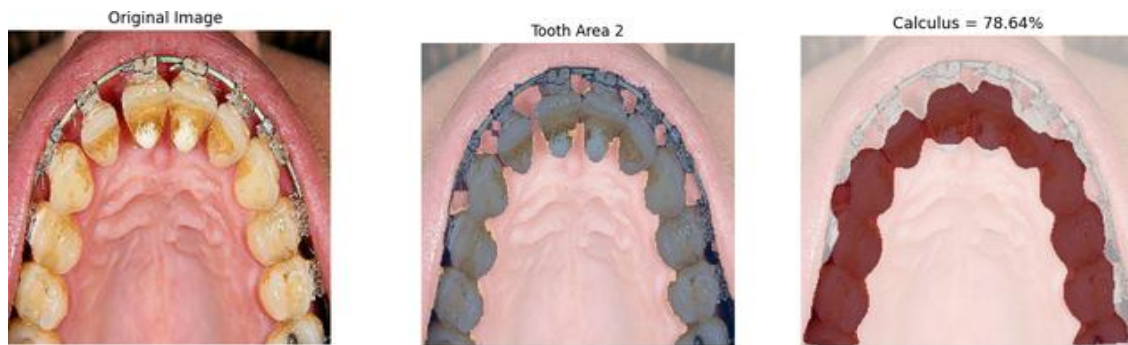
4.7 Sample Predictions and Qualitative Evaluation

Figure 13 (original image), Figure 14 (predicted binary mask), and Figure 15 (prediction overlay) present representative prediction outputs from the validation set for qualitative context. These examples are included to illustrate model behavior and are not the basis for quantitative performance claims.

[Insert Figures 13, 14, and 15 here]



Figures 13–15. Sample prediction output: intraoral occlusal view with moderate calculus deposition. Figure 13 (left): original image. Figure 14 (centre): predicted binary mask. Figure 15 (right): prediction overlay. The model qualitatively appears to identify calculus-bearing regions along gingival margins; some false-positive prediction of palatal tissue is visible centrally in the overlay.



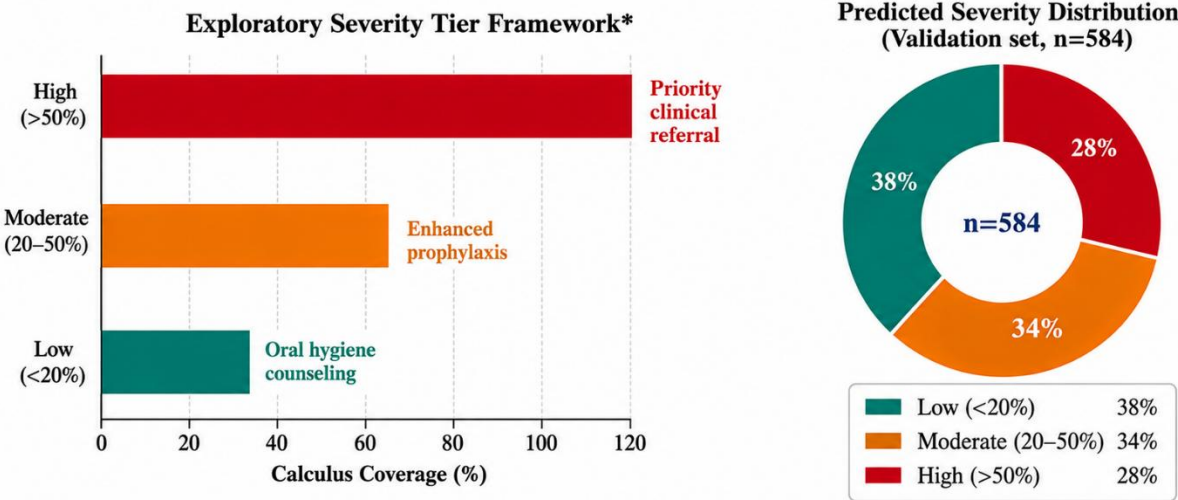
Figures 16–18. Severity estimation pipeline for a high-burden case. Figure 16 (left): original image. Figure 17 (centre): detected tooth area (grey overlay). Figure 18 (right): calculus severity overlay showing predicted calculus regions (red) with an estimated coverage of 78.64%—illustrative of a high-burden presentation. This estimate is based on the model’s prediction and has not been validated against clinical measurement.

4.8 Exploratory Severity Estimation (Validation Set)

The pixel-area-based severity estimation technique produced quantitative calculus coverage percentages across validation images. The distribution of predicted severity tiers across the 584-image validation set is shown in Figure 19: 38% of images were classified in the low-burden tier (<20% coverage), 34% in the moderate tier (20–50%), and 28% in the high-burden tier (>50%).

These severity tiers are defined as exploratory thresholds for proof-of-concept visualization and are not validated clinical cutoffs. The boundaries of <20%, 20–50%, and >50% were selected for conceptual alignment with light, moderate, and heavy calculus presentation categories but have not been validated against established periodontal indices (e.g., OHI-S, CI) or clinical outcomes. Clinical validation of these thresholds against standardized periodontal measurements is required before they could be applied in a clinical context.

Exploratory Calculus Severity Tier Framework and Validation Set Distribution



**Severity thresholds are exploratory only. These cutoffs have not been clinically validated and are not intended as diagnostic criteria. Clinical validation against established periodontal indices is required.*

Figure 19. Exploratory severity tier framework (left) and predicted severity distribution across the validation set (right, n=584 images). Tier thresholds are exploratory only and have not been clinically validated.

5. Discussion

5.1 Interpretation of Findings

This preliminary study demonstrates that deep learning can segment dental calculus from photographic intraoral images with encouraging performance on an independent test set. A Dice coefficient of 0.731 is broadly consistent with the range reported for AI-based caries and periodontal lesion segmentation from radiographic images (0.65–0.80 [7, 13]), and is particularly notable given the additional challenges of photographic imaging: variable lighting, uncontrolled camera angle, and the absence of radiographic density contrast that makes calculus distinctly visible on X-rays.

Test-set performance was modestly lower than validation-set performance (Dice 0.731 vs. 0.743; IoU 0.698 vs. 0.712), as expected when transitioning from a set used to guide training decisions to a fully unseen evaluation set. This gap is small and suggests the model generalizes reasonably to previously unseen images from the same dataset distribution, though generalization to clinically captured images from different settings has not been assessed.

The ROC-AUC of 0.931 should be interpreted carefully. As described in the Methods and Results sections, this value is computed pixel-wise on a dataset with substantial background pixel dominance. Under such class imbalance, a classifier that correctly identifies most background pixels—which is easy—

achieves high AUC even with moderate foreground (calculus) detection performance. Dice coefficient (0.731) and IoU (0.698) are less susceptible to this inflation and are therefore the primary performance metrics for this study.

The moderate precision of 0.519 is the principal performance gap. Qualitative review of false-positive cases indicates that yellowish-brown tooth discoloration is the primary source of misclassification. This limitation is consistent with findings from other photographic oral AI systems [9] and is expected given that the dataset includes tooth discoloration as a training class, potentially reinforcing the model's exposure to ambiguous color patterns.

5.2 Severity Estimation: Utility and Limitations

The quantitative calculus burden index represents the study's most distinctive contribution. Rather than producing a binary detection output, the system provides a percentage coverage estimate that could, in principle, support triage prioritization and oral hygiene outcome monitoring in screening programs. The 78.64% coverage estimate shown in Figure 18 is qualitatively consistent with the extensive calculus deposition visible in that image.

However, several important caveats apply. First, the severity index has not been validated against any clinical reference standard. It is unknown whether model-estimated calculus coverage percentages correlate with clinical measurements using established indices. Second, the exploratory severity tier thresholds (<20%, 20–50%, >50%) are not evidence-based cutoffs; they are conceptual divisions for proof-of-concept illustration. Third, even if coverage percentage estimates are accurate, calculus burden percentage is not a direct measure of periodontal disease severity; the relationship between calculus accumulation and clinical attachment loss involves additional host, microbial, and behavioral factors.

5.3 Comparison with Other Architectures

The architecture comparison provides preliminary evidence that the ResNet-50/U-Net configuration performs better than a standard U-Net baseline and VGG16-based encoding on this dataset, with EfficientNet-B4/U-Net as a close alternative. The 5.0 Dice-point improvement over the no-pretraining baseline supports the use of ImageNet transfer learning for dental imaging tasks with limited annotated data—a finding consistent with the broader medical imaging literature [10]. The near-equivalent EfficientNet-B4 performance merits further investigation given its lower computational requirements for potential mobile deployment.

5.4 Public Health Relevance

From a public health perspective, the significance of this work is primarily as a proof of concept that photographic calculus severity estimation is computationally feasible. If further developed and clinically validated, such a system could potentially contribute to:

- School-based screening programs, where standardized intraoral photographs captured by trained school health personnel could be analyzed to identify children warranting preventive dental referral;
- Tele-dentistry workflows, where AI-assisted pre-screening could support remote dental providers in triaging high-volume photographic submissions by predicted calculus burden;
- Community health worker programs in rural and underserved areas, where smartphone-captured images could be assessed for calculus burden as part of integrated oral health surveillance; and
- Low-resource global settings facing critical dental workforce shortages, where systematic photographic screening could support appropriate referral pathways.

Realizing any of these applications requires substantial further development, clinical validation, regulatory pathway consideration, and implementation research. The current study establishes only that the underlying computation is feasible with encouraging preliminary performance, not that the system is ready for any clinical application.

6. Public Health Implications

6.1 Tele-Dentistry Integration

Asynchronous tele-dentistry programs have demonstrated diagnostic accuracy comparable to in-person examination for visible oral conditions across multiple implementation contexts [20, 21]. AI-assisted calculus severity scoring could augment these workflows by automating preliminary image review and enabling consistent severity scoring independent of individual provider judgment. Standardized outputs could also improve documentation quality and support outcomes monitoring across tele-dentistry program cohorts.

6.2 School and Pediatric Oral Health Programs

School-based dental screening reaches populations at a critical developmental period when preventive interventions have the greatest long-term impact. Automated photographic screening could extend program reach into schools where on-site dental professional time is unavailable, ensuring children with significant calculus burden are identified for preventive referral.

6.3 Global and Low-Resource Settings

In low- and middle-income countries where dentist-to-population ratios may be extremely low, AI-assisted photographic screening deployed through community health worker programs could contribute to systematic calculus surveillance and appropriate referral, providing oral health infrastructure where structured screening would otherwise be impossible.

6.4 Preventive Counseling

At the individual level, a quantitative calculus burden estimate provides a tangible metric for oral hygiene behavior change communication. Visual severity overlays may support patient engagement by making calculus accumulation concrete and visible in a way that abstract verbal descriptions cannot.

7. Limitations

This study has several significant limitations that constrain generalizability and clinical applicability:

- Non-clinician annotations: Ground truth masks were created by a non-dental researcher. Without expert clinical annotation, systematic annotation errors may have propagated into model training and evaluation. This is the most critical limitation of the current study.
- Single public dataset: All images were sourced from a single Kaggle repository of unknown demographic composition and geographic origin. Dataset diversity, image quality standardization, and representativeness of community screening populations are unknown.
- Small held-out test set: The test set of 103 images yields wide confidence intervals. Performance estimates from this set may not be stable, and external validation on a larger independent dataset is needed.
- Bootstrapped CI limitations: Confidence intervals were estimated at the pixel level. Spatial correlation between neighboring pixels means these CIs likely underestimate true uncertainty and should be treated as approximate.
- Pixel-level ROC inflation: The ROC-AUC of 0.931 is inflated by non-calculus background pixel dominance and should not be interpreted as equivalent to image-level or patient-level diagnostic performance.
- No prospective clinical validation: No patients were enrolled prospectively. All images were sourced retrospectively from a public repository. Real-world performance in clinical or community settings is unknown.
- Unvalidated severity thresholds: The severity tier boundaries (<20%, 20–50%, >50%) are exploratory and have not been validated against clinical periodontal indices or outcomes.
- Scope of architecture comparison: The comparison covers four architectures evaluated on a small test set. Results should be treated as indicative rather than definitive; larger-scale evaluation on a diverse dataset is needed.

8. Ethical Considerations

8.1 Dataset Ethics and IRB

This study used a publicly available, de-identified image dataset sourced from Kaggle. No patients were recruited, no personal health information was collected, and no clinical interventions were performed. Institutional Review Board (IRB) review was not required for this study. The dataset was used in accordance with its published terms of use.

8.2 Algorithmic Bias and Health Equity

Healthcare AI systems carry risks of perpetuating health disparities if validated on datasets that underrepresent marginalized populations. For a screening tool intended for underserved communities, equitable performance across demographic subgroups is an ethical imperative. Systematic underperformance in populations with distinct oral pigmentation, lighting conditions, or dietary staining patterns could cause differential access to referral and care, worsening rather than reducing disparities. Future development must include systematic algorithmic fairness assessment using demographically characterized datasets.

8.3 Responsible Deployment Framing

This system is a proof-of-concept research prototype. It has not been evaluated in any clinical or community setting and is not validated for clinical use. Any future deployment would require clinical validation, regulatory review (in the United States, this tool may qualify as software as a medical device subject to FDA oversight), patient consent, and preservation of clinician oversight. AI-assisted screening tools must be positioned as decision support instruments, not autonomous diagnostic systems.

9. Future Directions

The findings of this study suggest the following prioritized next steps:

- Clinical annotation: Collaboration with licensed dental professionals and periodontists to generate expert-annotated training data with formal inter-rater reliability assessment (target: Cohen's $\kappa \geq 0.80$) as a prerequisite for any clinical performance claims.
- External validation: Testing on independent datasets from clinically and demographically diverse sources to assess generalizability beyond the current single-source dataset.
- Prospective enrollment: Collection of intraoral photographs from consented patients in community dental clinics, school programs, or tele-dentistry platforms using standardized photography protocols.

- Severity threshold validation: Correlation of model-estimated calculus coverage percentages with clinical measurements using established periodontal indices (OHI-S, CI) and clinical outcomes (probing depth, attachment loss).
- Algorithmic fairness audit: Systematic evaluation of performance disparities across demographic subgroups.
- Expanded architecture evaluation: Assessment of Attention U-Net, DeepLabV3+, and vision-transformer-based segmentation models.
- Mobile deployment feasibility: Evaluation of performance on smartphone-captured images under field conditions.
- Implementation research: Mixed-methods investigation of tool acceptability and workflow integration in community health worker and school health settings.

10. Conclusion

This proof-of-concept study demonstrates that deep learning can segment dental calculus from intraoral photographic images with encouraging preliminary performance. The ResNet-50/U-Net configuration achieved a Dice coefficient of 0.731, recall of 0.714, and ROC-AUC of 0.931 on an independent held-out test set, outperforming three comparative architectures evaluated under identical conditions. The pixel-area-based severity estimation technique produced quantitative calculus coverage estimates with exploratory clinical stratification, though severity thresholds remain unvalidated and should not be interpreted as clinical cutoffs.

The study has significant limitations, most critically the use of non-clinician annotations and a single public dataset source. Findings should be interpreted as preliminary evidence of feasibility rather than validated clinical performance. The transition from proof of concept to clinical applicability requires clinician-annotated data, prospective validation, algorithmic fairness assessment, and regulatory review.

With those caveats clearly stated, the feasibility demonstrated here supports further investigation in AI-assisted photographic oral health screening as a research direction with potential relevance to oral health equity—particularly for integrating into tele-dentistry workflows, school dental programs, and community health initiatives in resource-limited settings where the burden of calculus-related periodontal disease is greatest.

Abbreviations

AI, artificial intelligence; AUC, area under the curve; CI, confidence interval; CNN, convolutional neural network; CVAT, Computer Vision Annotation Tool; FDA, Food and Drug Administration; GPU, graphics processing unit; HSV, hue-saturation-value; IoU, intersection over union; IRB, Institutional Review Board; OHI-S, Oral Hygiene Index-Simplified; RGB, red-green-blue; ROC, receiver operating characteristic. CI also denotes the Calculus Index when referring to clinical periodontal indices.

Acknowledgments

The author would like to express sincere gratitude to the following individuals for their guidance, mentorship, and contributions to this research project.

Akoto Blessing Kwame

The author acknowledges Akoto Blessing Kwame for contributions to model development, computational implementation, and data generation activities that supported this research.

Michael McLennon, DDS, MD

The author gratefully acknowledges Dr. McLennon for his clinical guidance and insights regarding oral health, periodontal disease, and the clinical relevance of dental calculus assessment.

Cierra Mayo, M.S.

North Carolina Agricultural and Technical State University. Ms. Mayo provided assistance with manuscript preparation, scientific writing, editorial guidance, and manuscript development.

Clayton Clark II, PhD

Professor and Associate Dean for Academic Affairs, College of Engineering, North Carolina Agricultural and Technical State University. Dr. Clark provided mentorship regarding research design, scientific methodology, study formulation, and research development.

AKM Islam, PhD

Associate Professor, Department of Computational Data Science and Engineering, North Carolina Agricultural and Technical State University. As the author's doctoral advisor, Dr. Islam provided guidance on research methodology, study design, scientific rigor, and manuscript development.

The author is grateful for the support, expertise, and encouragement provided by these individuals throughout the conception, development, and completion of this study.

Ethics Statement

This study utilized publicly available, de-identified images obtained from a publicly accessible dataset. No direct patient interaction occurred and no identifiable patient information were collected. Because the study relied exclusively on publicly available, de-identified data, Institutional Review Board approval was not required.

Data Availability Statement

The dataset used in this study was obtained from a publicly available repository. Data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to the terms and conditions of the original data source.

Conflict of Interest Statement

The author declares no conflicts of interest related to this work.

Funding Statement

This research received no external funding.

Use of Artificial Intelligence Tools

Artificial intelligence (AI) tools were used during manuscript preparation to assist with language editing, grammar correction, formatting, and improvement of manuscript clarity. Specifically, OpenAI's ChatGPT was used to provide editorial suggestions and improve readability of the text. All scientific concepts, study design, data collection, model development, analysis, interpretation of results, and final manuscript content were reviewed, verified, and approved by the author. The author assumes full responsibility for the accuracy and integrity of the work. No AI tool was used to generate research data, perform statistical analyses, or make scientific conclusions.

Author Contributions

E.D. conceived and designed the study, developed the deep learning model, performed the computational implementation, data analysis, and interpretation of results, and drafted the manuscript. A.K. contributed to model development, computational implementation, and data generation. M.M. provided clinical guidance on oral health, periodontal disease, and the clinical relevance of the dental calculus assessment. C.M. contributed to manuscript preparation, scientific writing, and editorial revisions. C.C. provided mentorship on research design, scientific methodology, and study formulation. A.I. supervised the study and provided guidance on research methodology, study design, and scientific rigor. All authors reviewed and approved the final manuscript.

References

- [1] World Health Organization. Global oral health status report: Towards universal health coverage for oral health by 2030. Geneva: WHO; 2022.
- [2] Nazir MA. Prevalence of periodontal disease, its association with systemic diseases and prevention. *Int J Health Sci.* 2017;11(2):72–80.
- [3] Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. *Front Public Health.* 2023;11:1273253.

- [4] Surdu A, Budala DG, Luchian I, et al. Using AI in optimizing oral and dental diagnoses—a narrative review. *Diagnostics*. 2024;14(24):2804.
- [5] Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ. Prevalence of periodontitis in adults in the United States: 2009 and 2010. *J Dent Res*. 2012;91(10):914–920.
- [6] Tyagi M, Jain S, Ranjan M, et al. Artificial intelligence tools in dentistry: a systematic review on their application and outcomes. *Cureus*. 2025;17(5):e85062.
- [7] Dayı B, Uzen H, Balıkçı Çiçek I, Duman ŞB. A novel deep learning-based approach for segmentation of different type caries lesions on panoramic radiographs. *Diagnostics*. 2023;13(2):202.
- [8] Triveni M, Sree SV, Kruthika D, Mannem K. Mouth ulcer detection using deep learning. *Proc IEEE Int Conf*; 2023.
- [9] Li W, Liang Y, Zhang X, et al. A deep learning approach to automatic gingivitis screening based on classification and localization in RGB photos. *Sci Rep*. 2021;11:16831.
- [10] Mohapatra S, Swarnkar T, Das J. Deep convolutional neural network in medical image processing. In: *Artificial Intelligence and Machine Learning in Healthcare*. Elsevier; 2022:97–119.
- [11] Mustapha B, Zhou Y, Shan C, Xiao Z. Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. *Curr Med Imaging*. 2025;21.
- [12] Hartman H, Ruslandi RFA, Nurdin D, Akbar S, Setiawan AS. Exploring YOLO applications in dentistry through a systematic review. *Discov Artif Intell*. 2026;6:292.
- [13] Golkarieh A, Kiashemshaki K, Rezvani Boroujeni S. Advanced deep learning techniques for classifying dental conditions using panoramic X-ray images. *arXiv:2508.21088*; 2025.
- [14] Zakaria N, Hassim YMM. Improved image classification using enhanced VGG convolution neural networks. *Int J Informatics Vis*. 2024.
- [15] Surasinghe SP, Thanikasalam P. Detection and visualization of neglected tropical skin diseases using EfficientNet and Grad-CAM. *Proc IEEE ICAC*; 2023.
- [16] Byeon H. Optimizing EfficientNet with advanced data augmentation for class-imbalanced dental X-ray image classification. *ResearchGate*; 2024.
- [17] Eke PI, Thornton-Evans GO, Wei L, et al. Periodontitis in US adults: NHANES 2009–2014. *J Am Dent Assoc*. 2018;149(7):576–588.
- [18] Health Resources and Services Administration. Designated health professional shortage areas statistics. US Dept. of Health and Human Services; 2023.
- [19] Wall TP, Brown LJ. The American dental association’s review of dental care utilization and expenditures. *J Am Dent Assoc*. 2003;134(9):1200–1208.
- [20] Estai M, Kanagasingam Y, Mehdizadeh M, et al. A systematic review of the research evidence for the diagnostic accuracy of photography for detection of dental caries. *J Evid Based Dent Pract*. 2016;16(3):161–167.
- [21] Alabdullah JH, Daniel SJ. A systematic review on tele-dentistry as a tool to support people with special health-care needs. *J Evid Based Dent Pract*. 2021;21(3):101470.
- [22] Oktay O, Schlemper J, Le Folgoc L, et al. Attention U-Net: learning where to look for the pancreas. *arXiv:1804.03999*; 2018.
- [23] Greene JC, Vermillion JR. The simplified oral hygiene index. *J Am Dent Assoc*. 1964;68:7–13.