applied sciences

MDPI

Article

QEEF: A Quantitative Explainability Evaluation Framework for CNN and Vision Transformer-Based Segmentation Models in Dental Images

Vincent Majanga †, Ernest Mnkandla †, Yifan Luo † and Daniel Oladele *,†

Department of Computer Science, University of South Africa, Preller Street, Muckleneuk Ridge, Pretoria 1709, South Africa; majanvi@unisa.ac.za (V.M.); mnkane@unisa.ac.za (E.M.); luoy@unisa.ac.za (Y.L.)

* Correspondence: oladeda@unisa.ac.za

† These authors contributed equally to this work.

Abstract

Deep learning methods have exemplified performance in various dental image segmentation tasks. However, their interpretability inadequacy remains a hindrance to clinical adoption. This study introduces a quantitative explainability-driven evaluation framework to systematically assess and compare convolutional and transformer-based segmentation models in dental imaging. Specifically, we evaluate a convolutional Watershed Encoder–Decoder Neural Network (WEDN) and the hybrid Vision Transformer U-Net (ViTUNet) model using both qualitative and quantitative explainability metrics. The qualitative analysis employs Grad-CAM, occlusion sensitivity, and attribution-based visualization techniques to assess model decision patterns. Quantitatively, we define and apply explainability metrics, including Dice coefficient, boundary Dice, sparsity, and faithfulness metric scores, to independently evaluate segmentation performance and explanation quality. The experimental results show that ViTUNet achieves superior spatial localization, with higher Dice (0.765) and boundary Dice (0.421), compared to the WEDN model (0.664 and 0.221), indicating improved lesion delineation in dental images. Moreover, the ViTUNet model exhibits higher sparsity scores in attribution maps, indicating more concentrated and structured interpretability regions, while the WEDN model demonstrates higher faithfulness scores under perturbation-based evaluations, thus reflecting stronger dependence on localized feature representations. These findings highlight a key distinction between convolutional and transformer-based architectures with predictive behavior and explanation characteristics. Notably, the results further indicate that gradient-based explanation methods are more stable for transformer-based models, while perturbation-based methods better capture the decision logic of convolutional networks. In general, this study establishes an integrated framework for collectively evaluating segmentation performance and explainability, allowing more transparent and clinically reliable deployment of deep learning models in dental imaging.

Keywords: dental imaging; Vision Transformer; explainable artificial intelligence; model interpretability; explainability evaluation

Academic Editor: Thomas Lindner

Received: 17 June 2026
Revised: 9 July 2026
Accepted: 14 July 2026
Published: 16 July 2026

1. Introduction

Dental image segmentation plays a crucial role in computer-aided diagnosis by enabling the accurate identification and analysis of dental structures and oral diseases such as

dental caries. Precise segmentation facilitates the identification of affected regions, which then aid in clinical decision-making, early treatment planning, and disease monitoring.

Conventionally, deep learning techniques, especially convolutional neural networks (CNNs), have previously achieved significant success in medical image segmentation due to their automatic learning of hierarchical feature representations on medical images [1].

In contrast to traditional approaches that focus on domain-specific features, deep learning has emerged as the state-of-the-art (SOTA) framework in medical imaging, demonstrating strong performance across diverse clinical tasks [2].

Architectures such as U-Net [3] have become essential due to their encoder–decoder structure and their ability to capture multi-scale contextual information. However, CNN-based models are innately limited by localized receptive fields, constraining their ability to model global dependencies and contextual relationships [4].

Even though deeper frameworks partially resolve this limitation, they often remain inadequate for capturing complex spatial relationships in medical images. Therefore, CNN-based dental imaging methods have been widely used for segmentation, detection, and classification tasks, achieving strong performance in highlighting dental caries and other intra-oral conditions on medical images [5]. Nevertheless, their dependence on localized feature extraction constrains their ability to represent global anatomical contexts.

Recently, Vision Transformer (ViT) architectures [6] have emerged as a successful alternative to convolution-based methods. Unlike CNNs, ViTs use self-attention mechanisms to represent global contextual relationships across entire images. This enables transformers to capture complex spatial interactions more effectively than traditional convolutional neural networks.

Numerous studies have demonstrated that transformer-based models achieve significant performance in medical image segmentation by learning global representations that are complex for CNNs to represent [7]. However, transformers alone may struggle to preserve fine-grained local features, which are essential for medical imaging analysis. To resolve this limitation, hybrid CNN–transformer architectures have been proposed, combining CNNs' local feature extraction with transformers' global representation capabilities [8].

Representative hybrid approaches include the Fully Convolutional Transformers (FCTs), which combine convolution encoders with transformer-based modules to capture both local- and global-range dependencies in medical images [9].

The MISSFormer introduces a hierarchical U-Net architecture that improves both texture representation and global context learning via the encoder and bottleneck attention blocks. Similarly, the nnFormer [8] improves volumetric segmentation using local–global self attention and enhanced skip connections that replace conventional concatenation between the encoder and the decoder path.

Other prominent architectures include [10], which employs deformable transformer encoders (DeTrans-encoder) with multi-scale attention to upgrade long-range feature representation. The ScribFormer [11] also fuses CNN, transformer, and attention-focused channels for weak oversight. The hierarchical hybrid Vision Transformer (H2Former) [12] method integrates multi-scale channel attention with hierarchical transformers for improved combined feature extraction.

Additionally, models such as [13–15] and TransAttUNet [16] present a transformer-based attention neural network—namely, TransAttUNet—which displays the effectiveness of hybrid models by combining multi-scale features, attention-based enhancement, and cross-resolution representation.

Jointly, these architectures underscore that hybrid CNN–transformer architectures consistently outperform pure convolutional or transformer-based models in medical im-

age segmentation. This is achieved by stabilizing local feature information with global contextual reasoning.

Despite these advancements, deep learning models in clinical applications suffer from constrained interpretability. Both CNNs and transformer-based models are usually regarded as "black-box" models, raising concerns of transparency, reliability, and clinical trustworthiness [17]. Interpretability in medical image segmentation is vital because inter-pixel errors can directly affect early diagnosis and treatment planning [18].

To resolve these issues, explainable artificial intelligence (XAI) has become a vital component in the analysis of medical imaging systems. XAI methods aid in enhancing transparency by identifying regions that influence model predictions. Explanations in segmentation tasks must be fine-grained and aligned with clinically meaningful anatomical micro-structures [19].

Typical XAI techniques include gradient-based methods such as Grad-CAM [20], which produce explainable maps that identify significant regions. Perturbation-based methods, such as occlusion sensitivity analysis [21], quantify changes in model results when inputs are adjusted. Furthermore, transformer-based attention methods provide built-in interpretability by representing global dependencies across image regions [22].

Diverse studies have explored XAI techniques in medical imaging beyond just classification tasks. For instance, CNN-based XAI frameworks were used in [23] to improve transparency and evaluate model reliability in diagnostic applications. Other research works [24] categorize XAI methods into case-based, contextual, and complementary explanations, underscoring the diversity of interpretability strategies.

Despite these advances, most dental imaging works rely on segmentation accuracy metrics such as Dice coefficient and IoU, with little focus on explainability evaluation [25].

A key constraint in existing research is the lack of standardized quantitative evaluation of explanation quality. Whilst qualitative illustrations of saliency or attention maps are widely used, they do not guarantee that explanations will faithfully represent decision-making. As a result, visually acceptable explanations may not necessarily reflect exact model reasoning, thus impeding reliability in clinical setups.

To address this gap, this study focuses on two important quantitative characteristics of explainability: **faithfulness,** which measures how accurately explanations reflect the true influence of input features on model predictions; and **sparsity**, which evaluates how focused and clinically relevant the explanations are. These metrics are particularly vital in dental imaging, where anatomical structures are small, densely stacked, and require precise localization for meaningful interpretation.

In this study, we present an explainability-driven comparative analysis of two dental image segmentation models:a CNN-based convolutional Watershed Encoder–Decoder Network (WEDN) [26] and a hybrid Vision Transformer U-Net (ViTUNet) [27]. Both models were trained under similar conditions and evaluated using both qualitative and quantitative XAI methods, including Grad-CAM, occlusion sensitivity, and ensemble attribution techniques. Additionally, explainability-specific measures such as faithfulness and sparsity scores were incorporated to provide a thorough assessment of interpretability.

The main contributions of this proposed work are as follows:

- A methodical explainability-driven comparison between CNN-based WEDN and transformer-based ViTUNet model architectures for dental image segmentation.
- A comprehensive analysis of local versus global feature learning behavior in CNN and transformer architectures, along with its impact on interpretability.
- A unified evaluation framework combining segmentation metrics (Dice, boundary Dice) with quantitative explainability metrics (faithfulness and sparsity).

- An evaluation of gradient-based and perturbation-based XAI methods for both CNN and transformer segmentation models.

Ultimately, this work moves beyond conventional accuracy-centric evaluation by integrating explainability as a core dimension of model assessment. The proposed analysis provides a meticulous comprehension of model decision-making behavior and contributes towards more transparent, reliable, and clinically trustworthy dental imaging systems.

2. Proposed Explainability-Driven Framework

2.1. Overview

This proposed study consolidates an explainability-driven evaluation framework for the segmentation of carious lesions in bitewing dental images. The framework integrates four interconnected components:

(i) Data pre-processing and augmentation;
(ii) Watershed-based image enhancement and mask generation;
(iii) Deep learning segmentation (WEDN and ViTUNet);
(iv) Explainability evaluation using QEEF.

Two deep learning model architectures are investigated under similar experimental conditions:

- Convolution-based Watershed Encoder–Decoder Neural Network (WEDN) [26];
- Vision Transformer-based U-Net (ViTUNet) [27].

This framework is designed to jointly aid segmentation performance analysis and explainability assessment in a structured and replicable pipeline. The first component standardizes dental images via augmentation to enhance data diversity and model generalization. The second component uses watershed-based image processing techniques to generate edge-improved masks that emphasize lesion structures and anatomical boundaries. The third component handles carious lesion segmentation via the WEDN and ViTUNet architectures under similar experimental conditions. Lastly, the fourth component quantitatively evaluates model interpretability via the proposed QEEF metrics, namely, faithfulness, sparsity, and boundary Dice. Figure 1 shows the proposed Quantitative Explainability Evaluation Framework (QEEF) diagram overview.

2.2. Dataset and Materials

Dataset

This study uses a 100-image dataset of bitewing dental images, publicly sourced from the Mendeley dataset repository [28]. These images vary in resolution but are standardized to (128 × 128), and they are all of PNG format for computational consistency. Figure 2 shows a sample of the images in the dataset.

Even though data augmentation increases the dataset size to 3000 images, the original dataset consists of 100 unique bitewing radiographs. Consequently, the dataset may not sufficiently capture the variability found in real-world clinical environments, such as data acquisition protocols, image quality, differences in imaging devices, and patient population analysis.

2.3. Data Augmentation

To resolve the limited dataset size and improve model generalization, the dataset was expanded using Albumentations [29]. The augmentation pipeline was applied to both images and corresponding masks to preserve spatial alignment and model training only. Image transformations included the following:

- Affine transformations;
- Gaussian noise;
- Random brightness and contrast adjustment;
- Horizontal flipping;
- Hue, saturation, value compositions, and resizing operations.

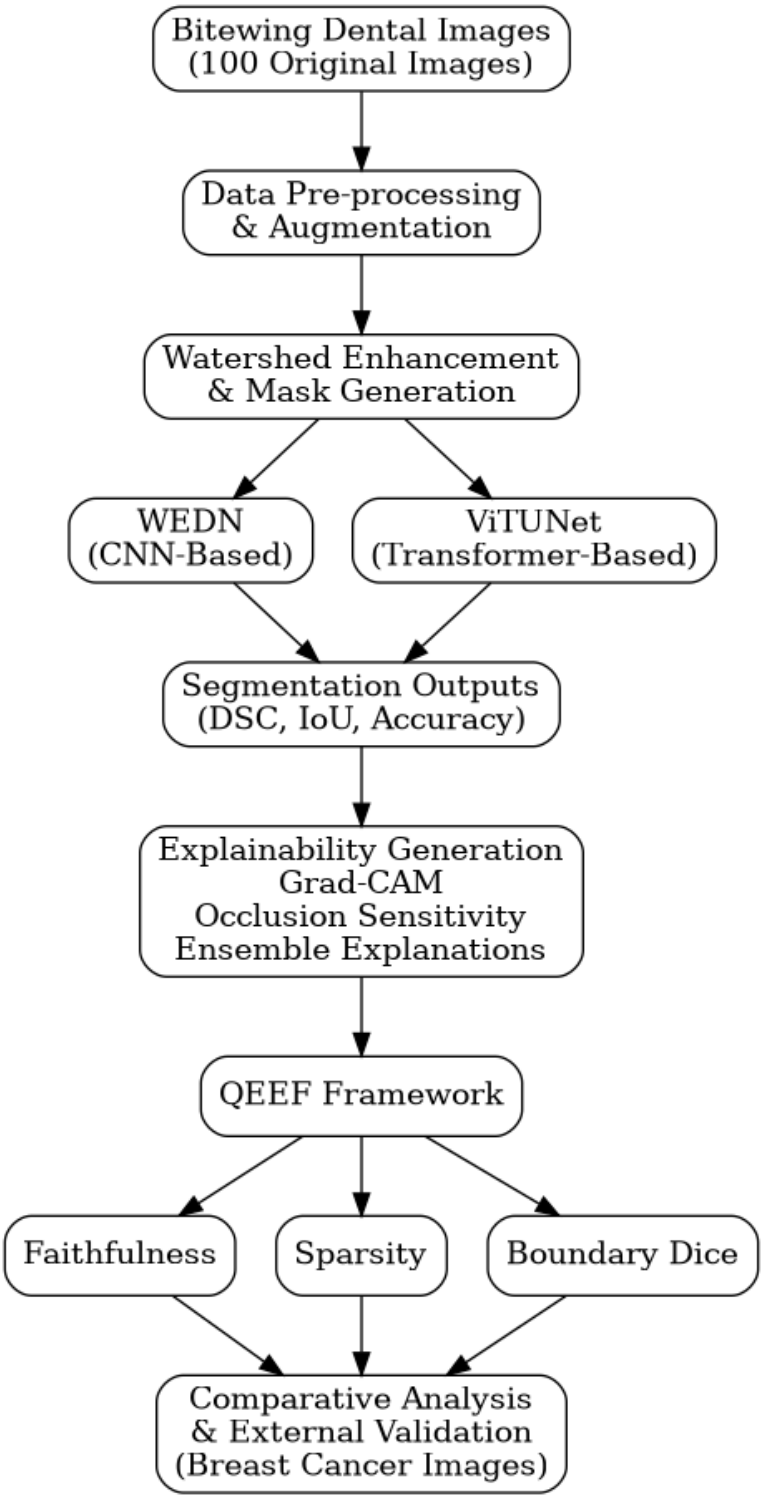

Figure 1. Quantitative Explainability Evaluation Framework (QEEF) framework diagram.

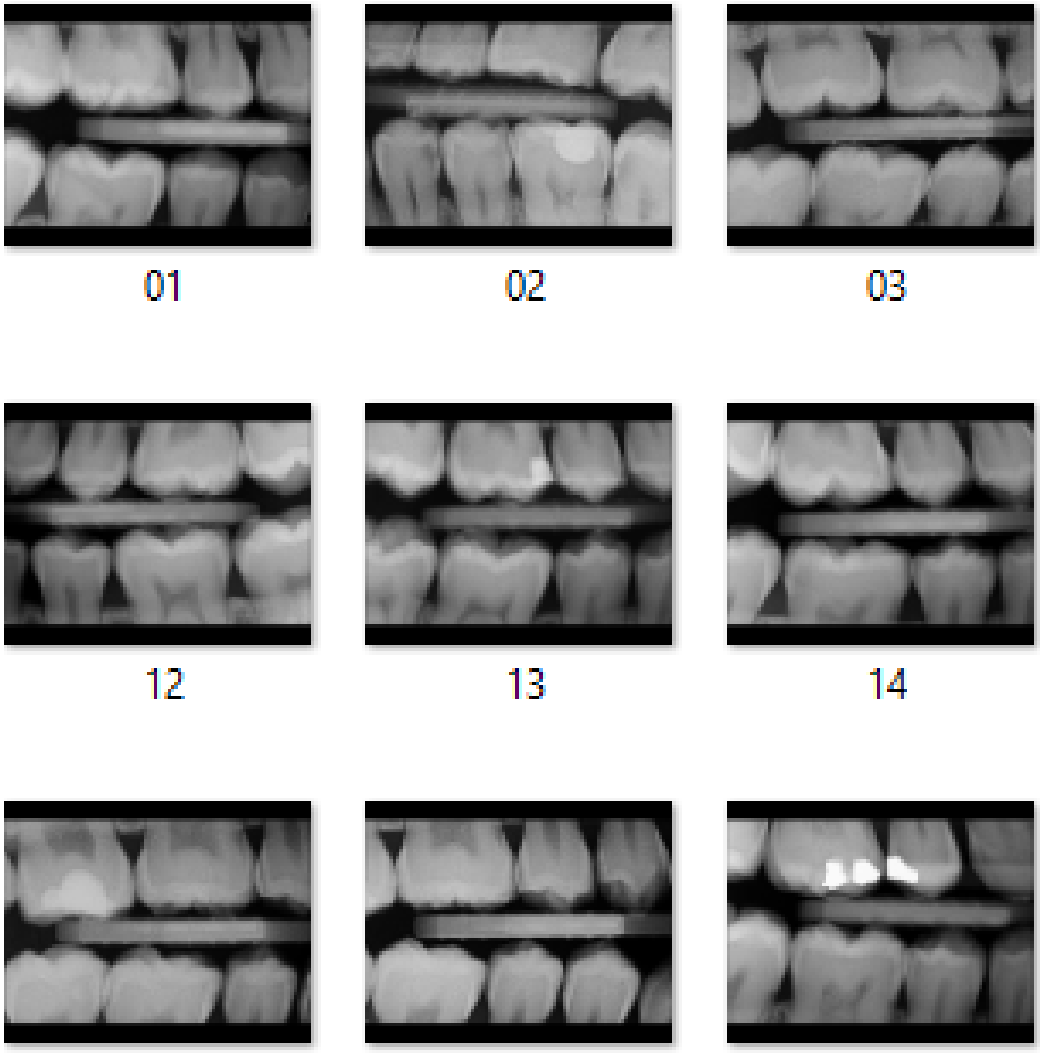

Figure 2. Some original bitewing dental images from the Mendeley dataset repository [28].

The original dataset was expanded from 100 to approximately 3000 paired image samples, ensuring robustness in training deep learning models under small data conditions. Figure 3 shows a sample of the augmented images in the dataset.

2.4. Image Enhancement and Watershed Mask Generation

A pre-processing pipeline was applied to produce refined watershed masks used to enhance boundary effective learning.

2.4.1. Thresholding

Otsu and binary thresholding were applied to segment foreground regions corresponding to potential carious lesions.

2.4.2. Morphology Operations

Operations such as erosion, dilation, opening, and closing were used to remove noise, improve object boundaries, and separate foreground and background structures.

2.4.3. Distance Transform and Connected Components

Distance transformation identifies sure foreground regions, whereas connected component analysis labels image regions into foreground, background, and unknown classes.

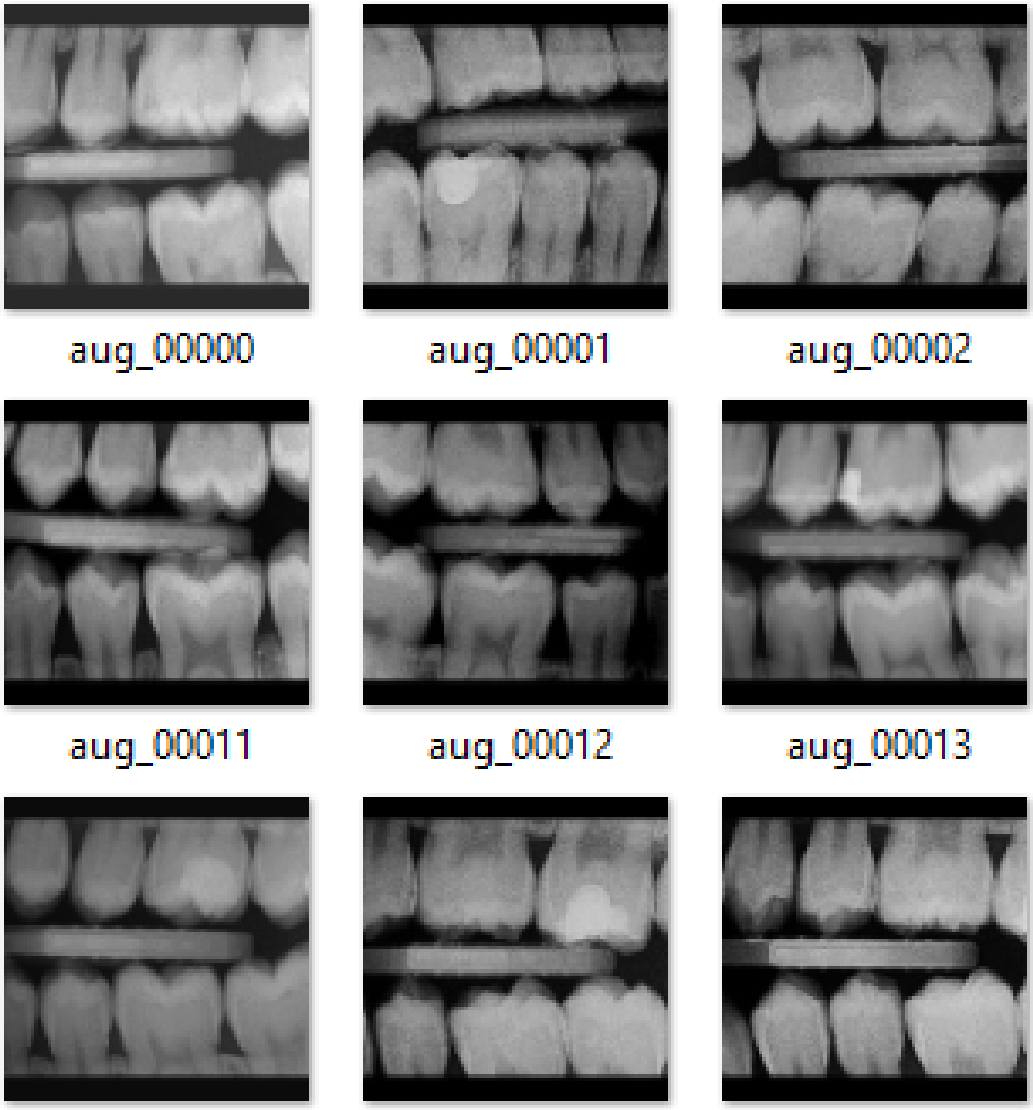

Figure 3. Original bitewing dental images after Albumentations.

2.4.4. Watershed Segmentation

The watershed transform was applied to the generated boundary masks that enhance lesion regions. The resulting dataset formed a watershed dataset used jointly with augmented images for model training.

Figure 4 visualizes the final mask results after the application of the watershed segmentation method.

2.5. Segmentation Models

2.5.1. Convolutional Watershed Encoder–Decoder Neural Network (WEDN)

This model, illustrated by Figure 5, is a convolutional encoder–decoder architecture enhanced with watershed-based boundary refinement to improve segmentation of low-contrast and intersecting dental structures.

Let $x \in \mathbb{R}^{H \times W}$ be the input image.

$F_e = \phi enc(x)$ for the features extracted at the encoder.

$\hat{y} = \phi dec(F_e)$ for the reconstruction at the decoder.

The WEDN model utilizes the 128 × 128-sized images from both the image and watershed-generated mask datasets as input.

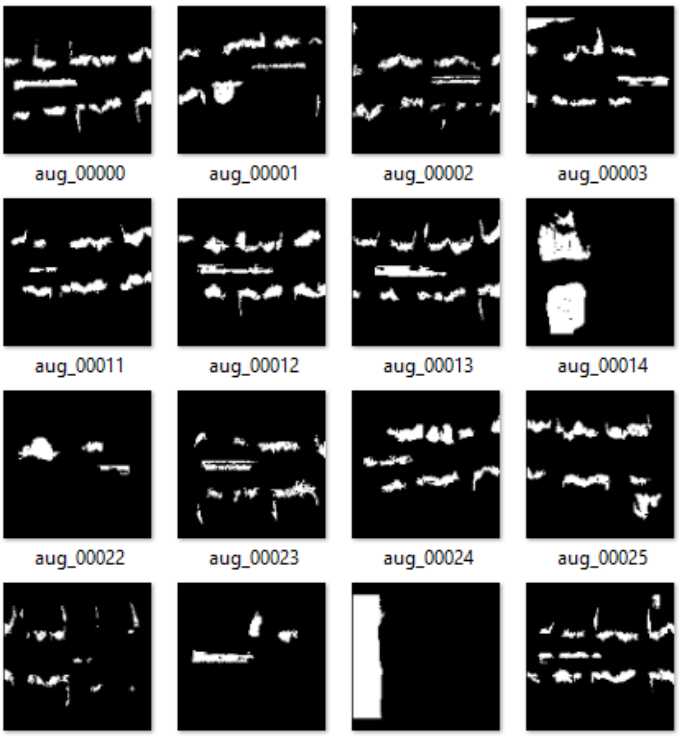

Figure 4. Image masks in the watershed mask dataset after application of the watershed segmentation method.

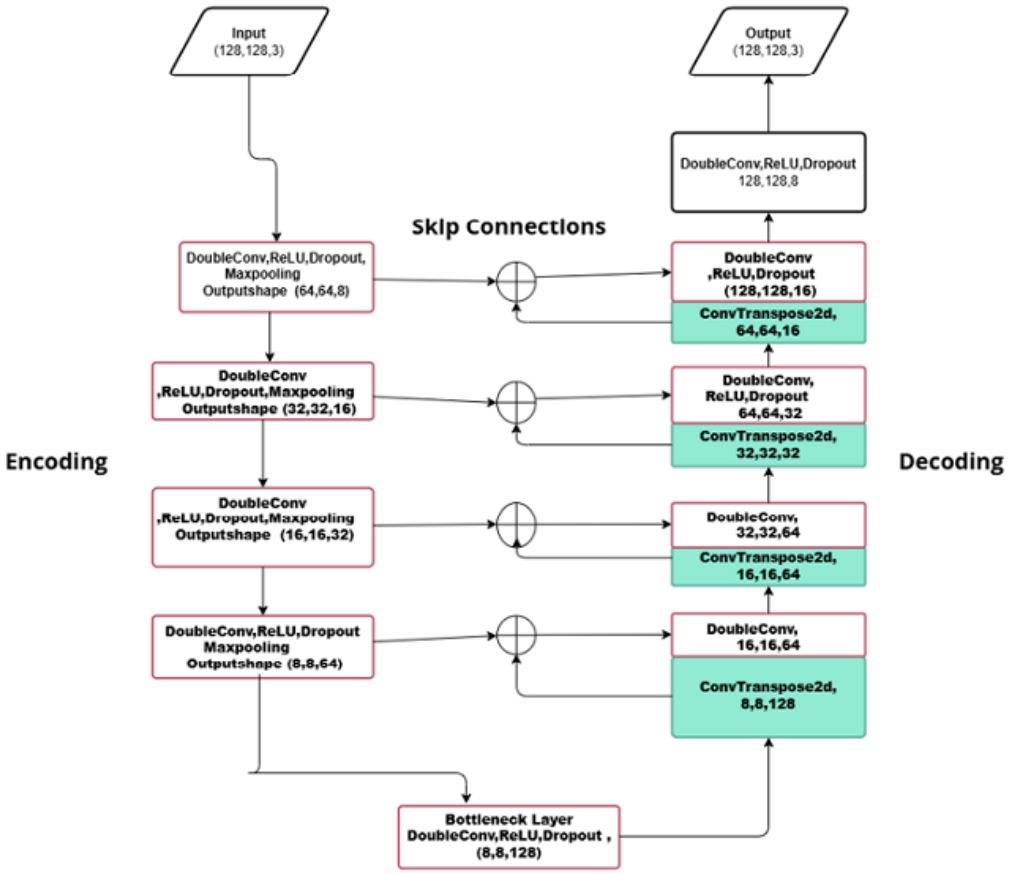

Figure 5. WEDN model architecture diagram.

2.5.2. Vision Transformer U-Net (ViTUNet)

This model, illustrated by Figure 6, integrates a convolutional encoder with a Vision Transformer encoder to capture both spatial features and global dependencies.

Patch Embedding:

Let $Z_0 = Patchembedding(x)$.

Transformer Encoding:

$Z_l = MSA(Z_{l-1}) + Z_{l-1}$.

Decoder Reconstruction:

$\hat{y} = \phi dec(Z_L)$.

Hence, this hybrid model architecture enables simultaneous learning of both global dependencies and local spatial features.

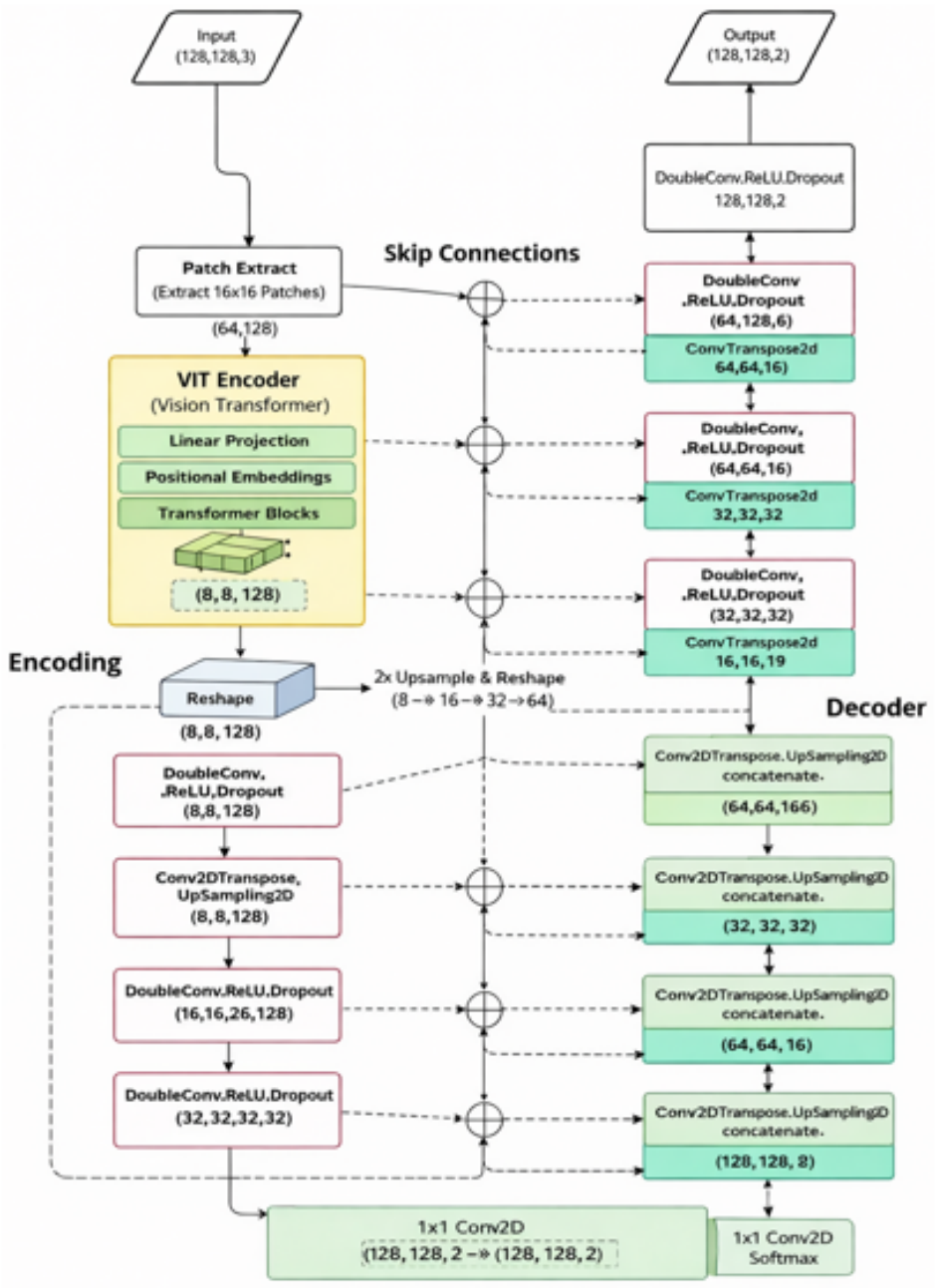

Figure 6. ViTUNet model architecture diagram.

Notably, from a computational standpoint, WEDN fundamentally depends on convolutions and localized operations, resulting in comparably lower memory requirements and

computational complexity. Contrarily, ViTUNet relies on self-attention to model long-range dependencies that generally acquire higher computational and memory requirements due to global contextual processing. Hence, the primary objective of this study is the quantitative evaluation of explainability rather than computational benchmarking; a comprehensive analysis of training time and computational efficiency is not in the scope of the present work.

2.6. Explainability Generation Module

For further interpretability analysis, subsequent explainability methods were applied to the trained models.

Let $f(x)$ signify a trained segmentation model and $E(x)$ the corresponding explanation map.

The following explainability methods were used:

- Gradient-based methods for convolutional layers (WEDN);
- Attention-based attribution methods for transformer layers (ViTUNet);
- Occlusion sensitivity analysis for robustness evaluation.

The explanation maps $E(x)$ were produced after model inference and stored for subsequent quantitative analysis using the QEEF metrics.

2.7. Quantitative Explainability Evaluation Framework (QEEF)

To evaluate interpretability in a systematic manner, this study describes the Quantitative Explainability Evaluation Framework (QEEF), which consists of three complementary metrics: faithfulness, sparsity, and boundary Dice. These metrics were selected because they measure explanation reliability, localization concentration, and anatomical boundary consistency, respectively. Rather than aggregating them into a single score, each metric is reported independently to preserve interpretability and avoid weighting bias.

Faithfulness:

$$\text{Faithfulness score} = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - f(x_i \setminus R_i)) \tag{1}$$

where N is the total number of test samples, $f(x_i)$ is the original prediction confidence for image x_i, and R_i is the important explanation region identified by the XAI method; $x_i \setminus R_i$ denotes the image after occluding the important region.

This was used to evaluate the extent to which accentuated highlighted regions truly influenced model predictions by measuring prediction degradation after perturbation.

Sparsity:

$$\text{Sparsity score} = 1 - \frac{\|M\|_0}{\|M\|_1} \tag{2}$$

where M represents the explanation/attention map, $\|M\|_0$ is the number of non-zero activated pixels, and $\|M\|_1$ is the sum of activation intensities. This measures the concentration of explanation activation regions, where higher values indicate more focused and localized explanations.

Boundary Dice Explanations:

$$\text{Boundary Dice} = \frac{2 \times |B_p \cap B_g|}{|B_p| + |B_g|} \tag{3}$$

where B_p is the predicted segmentation boundary, B_g is the ground-truth segmentation boundary, and $|B_p \cap B_g|$ represents the intersecting boundary pixels.

This evaluates the degree to which explanation-guided segmented boundaries align with the ground-truth anatomical boundaries. Boundaries were extracted using morphological contour operations following thresholding of the explanation maps. Therefore, boundary Dice values may vary due to threshold selection and boundary extraction method. Higher boundary Dice scores indicate improved anatomical edge alignment and boundary homogeneity.

2.8. Summary of the Proposed Framework

The proposed framework follows an aligned pipeline for explainability-driven segmentation of carious lesions in bitewing dental images. The process begins with data processing and augmentation to improve data diversity and model generalization. Subsequently, the watershed-based image enhancement and mask generation are used to improve the boundary delineation of lesion structures.

The resultant data are then used to train and evaluate two segmentation models, WEDN and ViTUNet, under similar experimental conditions to ensure fair comparative analysis. Both models produce pixel-wise segmentation outputs corresponding to carious lesion regions.

Lastly, subsequent model explanations are produced using gradient-based and perturbation-based methods, generating explanation maps that are quantitatively evaluated using the proposed Quantitative Explanation Evaluation Framework (QEEF), which assesses explanation quality via three complementary metrics: faithfulness, sparsity, and boundary Dice. This unified workflow enables a comprehensive evaluation of both segmentation performance and interpretability, forming the foundation for comparative analysis presented in the Results and Discussion section.

3. Experimental Results and Discussion

The experiments were conducted on 3000 image–watershed mask pairs, split into 2400 for training and 600 for testing. To obtain optimal model configurations, iterative experiments were carried out, and extensive model hyperparameter tuning was performed by varying the dropout rate, learning rate, batch size, and the number of training epochs. Through iterative experiments, both the WEDN and ViTUNet models achieved optimal performance using the Adam optimizer with a learning rate of 0.0004, dropout rates ranging from 0.1 to 0.4 across convolutional blocks, the Dice + categorical cross-entropy loss, and batch normalization. Weight regularization was also used via early stopping, reducing the learning rate when the training plateaus, and increasing the number of epochs from 25 to 50 and then 75. Additionally, the hardware and software configurations included random seed 42, CPU RAM = 16 GB, and Python version 3.6.0.

To comprehensively evaluate model performance and interpretability, segmentation metrics including Dice coefficient, intersection over union (IoU), and accuracy were utilized alongside the proposed Quantitative Explainability Evaluation Framework (QEEF), which evaluates explanation quality via faithfulness, sparsity, and boundary Dice scores.

During training, both models exhibited performance plateaus after several epochs, prompting the incorporation of early stopping and adaptive learning rate reduction strategies to mitigate overfitting and improve generalization.

3.1. Segmentation Performance Evaluation

To quantitatively evaluate segmentation performance, the Dice coefficient, IoU, and accuracy metrics were employed. Equations (4)–(6) define the IoU, Dice coefficient, and accuracy evaluation metrics, respectively.

$$\mathrm{IoU}_k = \frac{\sum_i \mathbf{1}\left[\hat{Y}_i = k \ \wedge \ Y_i = k\right]}{\sum_i \mathbf{1}\left[\hat{Y}_i = k \ \vee \ Y_i = k\right]}, \quad k = 1, 2, \ldots, K \tag{4}$$

Given that IoU_k is for class k, $\hat{Y}_i$ is the predicted class label at pixel i, Y_i the ground-truth class label at pixel i, and $1[\cdot]$ the indicator function (equals 1 if the condition is true), while $\wedge$ is the logical AND (intersection), $\vee$ the logical OR (union), and K the total number of classes.

$$\mathrm{Dice} = \frac{2 \cdot |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} = \frac{2 \sum_i \hat{Y}_i \cdot Y_i}{\sum_i \hat{Y}_i + \sum_i Y_i} \tag{5}$$

where $\hat{Y}$ is the predicted segmentation mask, Y the ground-truth mask, $\hat{Y}_i, Y_i \in 0, 1, 2$; Numerator: computes twice the intersection; Denominator: Sum of sizes (predicted + true).

$$\text{dental dataset accuracy} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{1}\left(y_{ij}^{n} = \hat{y}_{ij}^{n}\right)}{N \times H \times W} \tag{6}$$

Given the above, N is the total number of images in the dataset, $H \times W$ represents the image dimensions per pixel, y_{ij}^{n} and $\hat{y}_{ij}^{n}$ are the ground-truth and predicted labels of pixel i, j in image n, respectively, and $\mathbf{1}(\cdot)$ is the indicator function.

Table 1 summarizes the quantitative segmentation performance of the WEDN and ViTUNet on the testing dataset.

Table 1. Segmentation performance comparison between WEDN and ViTUNet models on bitewing dental images.

Authors	Segmentation Method	Image Types Used	DSC	IoU	Accuracy
majanga et al. [26]	WEDN	Bitewing dental radiographs	94.70%	89.99%	96.14%
majanga et al. [27]	Vision Transformer + UNet model	Bitewing dental radiographs	97.88%	95.87%	98.45%

Both models achieved strong segmentation performance on the testing dataset. Nonetheless, ViTUNet consistently outperformed WEDN across all evaluation metrics, achieving high Dice, IoU, and validation accuracy scores. These results indicate that both models successfully learned discriminative segmentation representations. At the same time, the transformer-based architecture demonstrated superior capability to capture the complex contextual relationships necessary for accurate anatomical delineation. However, segmentation performance alone does not explain how these predictions are produced, justifying the need for explainability evaluation via the proposed QEEF.

3.2. Explainability Analysis

Although the WEDN and ViTUNet models achieved high segmentation performance, traditional segmentation metrics do not provide insight into the reasoning process underlying model predictions. Hence, explainability analysis was conducted to investigate model decision-making behavior (black-box testing), feature localization characteristics, contextual logic, and the reliability of learned representations.

To accomplish this, qualitative explanation methods, such as Grad-CAM visualizations, occlusion sensitivity, and ensemble explanations, were fused with the proposed QEEF, which evaluates explanations using faithfulness, sparsity, and boundary Dice metrics. WEDN and ViTUNet were selected as representative CNN-based and transformer-based segmentation architectures to comparatively assess architectural interpretability behavior.

3.2.1. Qualitative Explainability

Grad-CAM Visual Analysis: Gradient-Weighted Class Activation Mapping (Grad-CAM) was utilized to display discriminative image regions contributing to segmentation predictions. The resulting heatmaps revealed considerable architectural differences between the WEDN and ViTUNet models.

WEDN produced highly localized activation maps concentrated around edge transitions and high-contrast anatomical structures. These activations were restricted to local pixel neighborhoods surrounding the segmented tooth regions, suggesting strong reliance on texture information and localized intensity variations.

Conversely, ViTUNet generated broader and spatially distributed attention regions extending across multiple anatomical structures. The transformer-based architecture captured long-range dependencies between neighboring teeth and surrounding anatomical regions, allowing the inclusion of global contextual information beyond isolated edge features. Hence, the ViTUNet explanations appeared smoother and more contextually understandable than the sharply localized activations observed in WEDN.

This characteristic was specifically evident in low-contrast regions and areas containing intersecting anatomical structures, where ViTUNet maintained stable contextual attention while WEDN remained heavily dependent on localized texture features.

Occlusion Sensitivity Analysis: These experiments were performed by methodically masking image regions and measuring the resulting changes in segmentation confidence. The results revealed that WEDN exhibited significant prediction degradation when small regions near tooth edges were occluded. This behavior confirms CNNs' strong dependence on localized edge information and their innate local receptive-field bias.

In contrast, ViTUNet exemplified greater robustness to localized occlusions. Prediction degradation occurred more progressively as larger image regions were masked, suggesting that the transformer architecture distributed feature importance across wider contextual regions. Moreover, ViTUNet maintained relatively stable predictions in anatomically vague regions, where WEDN exhibited increased sensitivity to missing local structural information.

Attention Map Evaluation: Preliminary experiments incorporated transformer attention maps to visualize internal self-attention behavior. The application of the proposed explainability framework on ViT transformer blocks produced attention rollout maps that were inconsistent, with blurry localization behavior and weak correlation to clinically meaningful segmentation regions. Figure 7 demonstrates how attention maps accentuate broad image regions unrelated to the final segmentation outputs, hindering interpretation for pixel-level medical segmentation tasks.

Therefore, Grad-CAM and occlusion sensitivity were retained as the key qualitative explainability methods because they provided more stable, spatially meaningful, and clinically interpretable visual explanations.

Ensemble Explainability Analysis: To improve explanation robustness, ensemble explanations combining Grad-CAM and occlusion sensitivity maps were explored. Ensemble explanations reduced isolated, noisy activations and improved consistency between highlighted regions and anatomical structures. For WEDN, ensemble explanations preserved edge-focused activations while suppressing irrelevant background regions. For ViTUNet, ensemble explanations improved contextual continuity across neighboring dental structures and produced smoother activation region distributions.

These findings indicate that combining complementary explanation techniques provides a more reliable interpretability than relying on a single explanation method. Figure 8 presents qualitative explainability visualizations for the WEDN model. Figures 8 and 9 present qualitative explainability results for WEDN and ViTUNet, respectively.

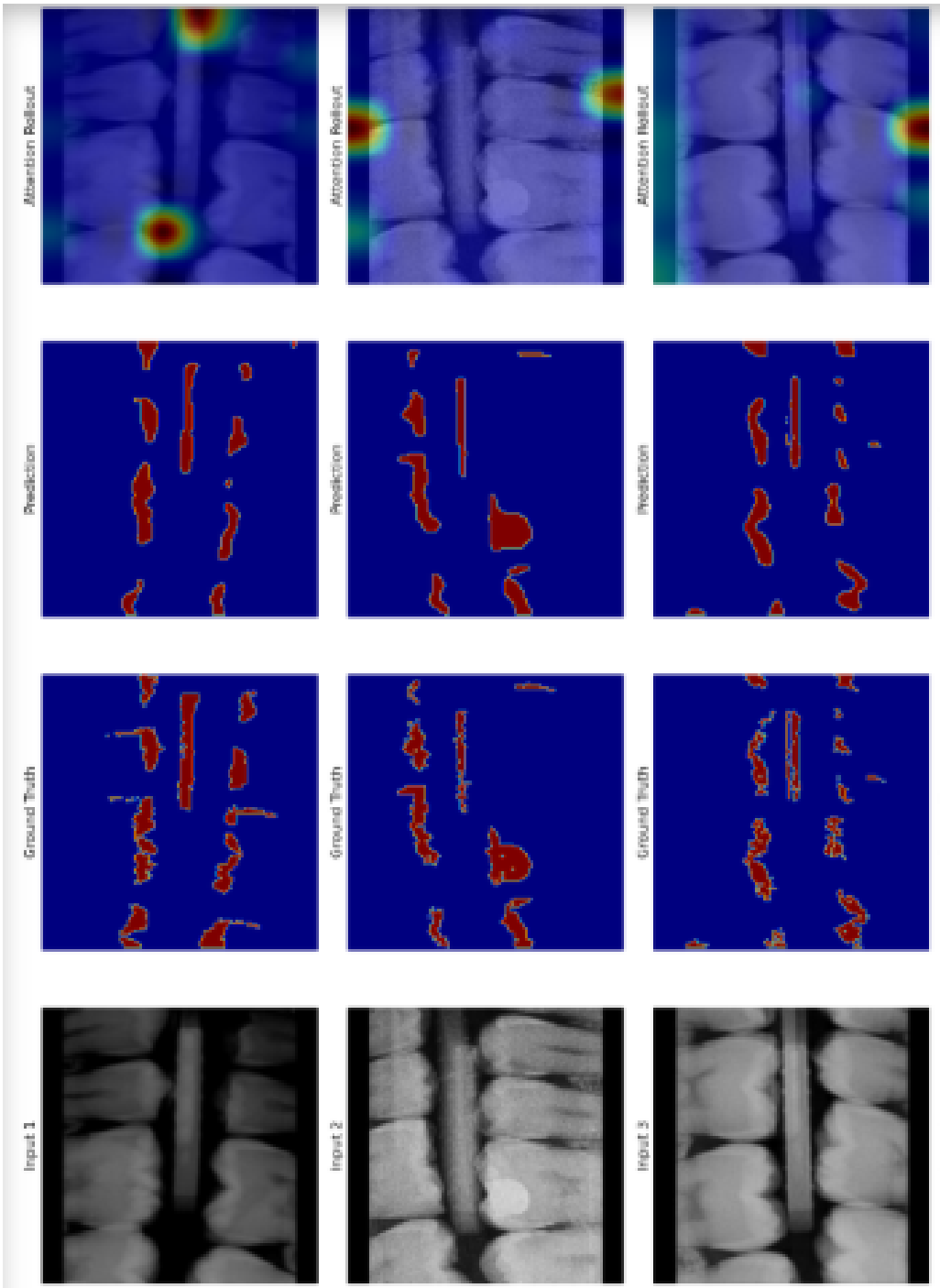

Figure 7. Attention rollout maps results for the ViTUNet model on dental images.

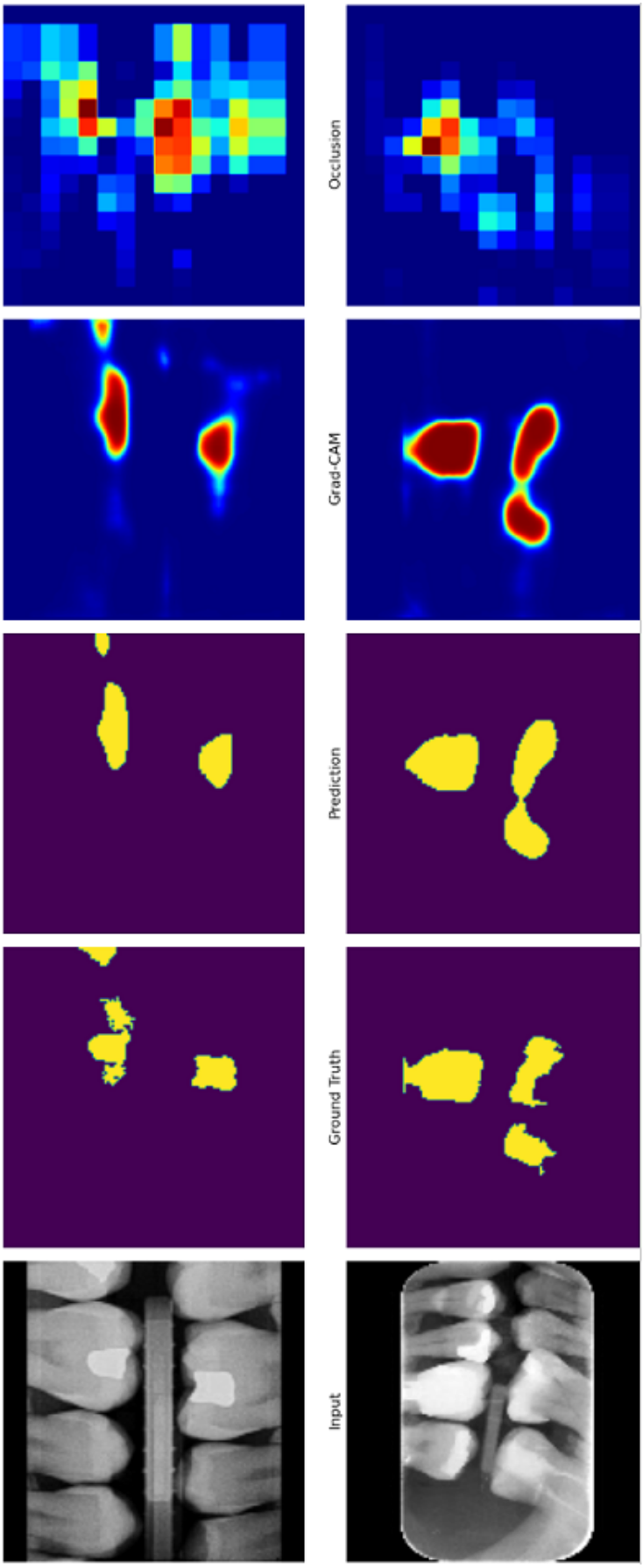

Figure 8. Qualitative explainability results for the WEDN model on dental images.

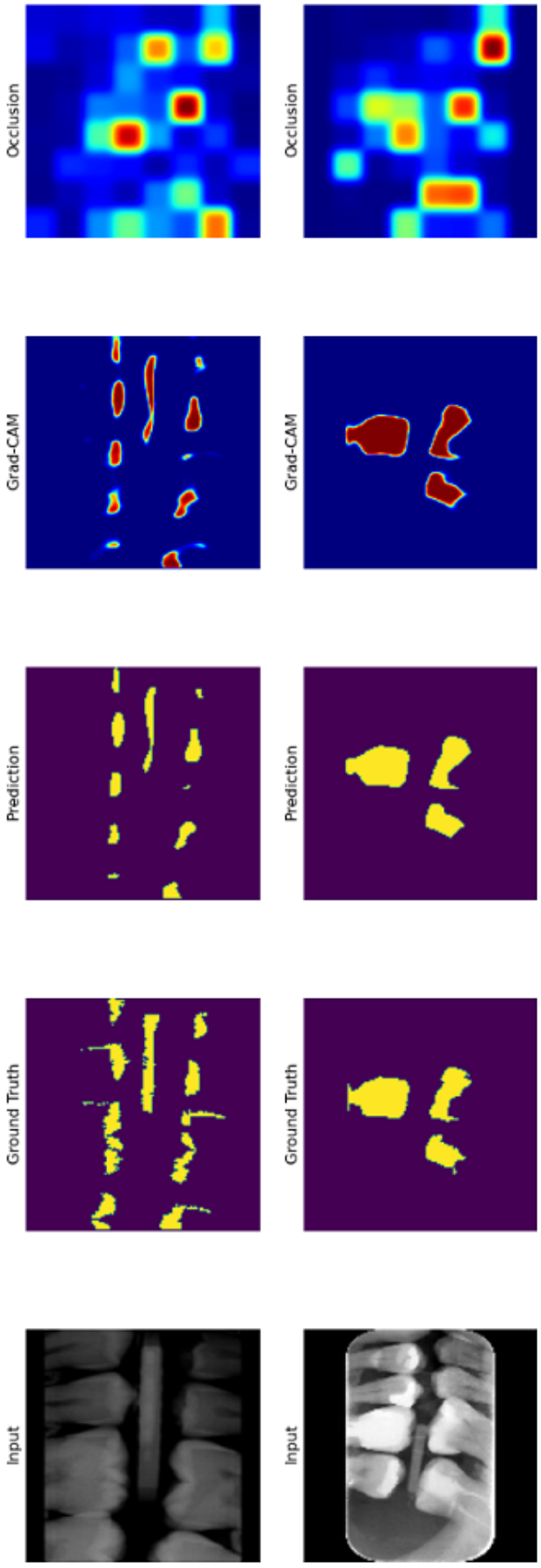

Figure 9. Qualitative explainability results for the ViTUNet model on dental images.

3.2.2. External Explainability Validation on Breast Cancer Images

To investigate whether the explainability behavior observed in dental images generalizes across imaging types, an external breast cancer dataset was used for qualitative validation. This evaluation was not intended to assess segmentation performance but, rather, to check the robustness and transferability of explanation behaviors under different domains.

Figure 10 presents Grad-CAM and occlusion sensitivity maps generated using the ViTUNet model on representative breast cancer images.

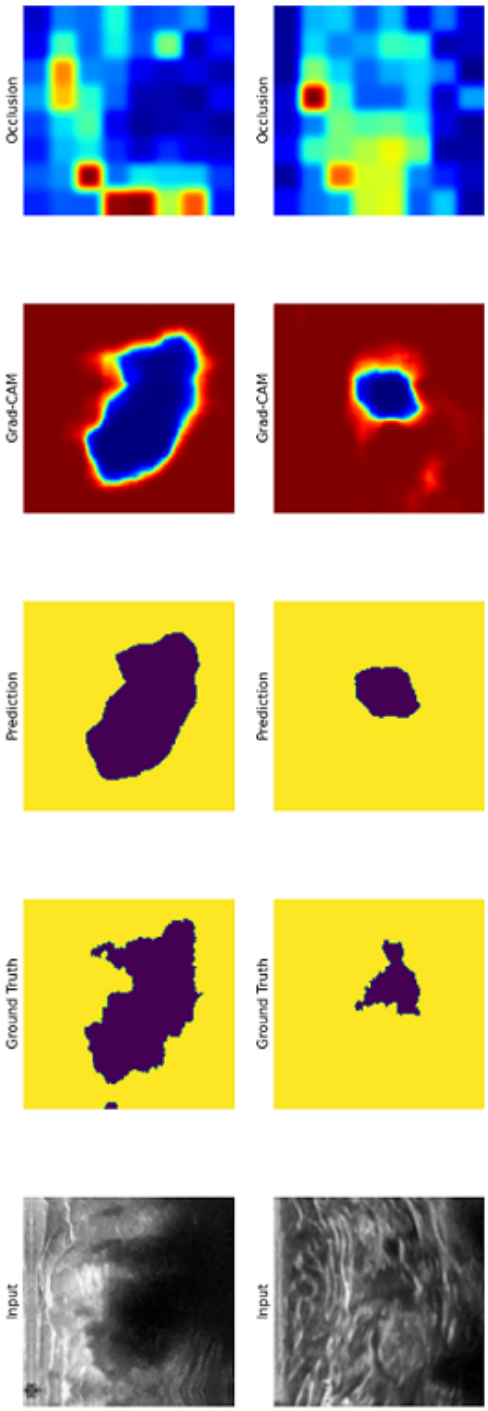

Figure 10. Qualitative explainability results from the ViTUNet model on breast cancer images.

The obtained explanation results via Grad-CAM generated localized activated regions concentrated around tumor structures, while occlusion sensitivity produced broader distributed regions reflecting perturbation-based feature importance.

Significantly, ViTUNet maintained contextual understanding across diverse tissue structures, highlighting both tumor regions and surrounding structure patterns. This indicates that the transformer-based architecture learns transferable contextual representations that remain interpretable across different medical imaging types. Moreover, the observed diffused localization behavior may partially be a result of limited training data, patch-based tokenization, and the specific ViTUNet implementation, rather than the inherent limitation of transformer model architectures.

3.2.3. Quantitative Explainability Evaluation Using the Proposed QEEF

Quantitative explainability evaluation was conducted using the proposed Quantitative Explainability Evaluation Framework (QEEF). As defined in the previous section, the QEEF evaluates explanation quality using three complementary metrics: faithfulness (Equation (1)), sparsity (Equation (2)), and boundary Dice (Equation (3)). Within the proposed QEEF, these metrics are treated as independent evaluations. Each metric captures a distinct characteristic of explanation quality, enabling a more comprehensive assessment of interpretability without relying on a single aggregated measure.

Table 2 summarizes the QEEF explainability results for WEDN and ViTUNet across Grad-CAM, occlusion sensitivity, and ensemble explanations. Negative faithfulness metric score values suggest weak relationships between highlighted regions and model prediction behavior. Under these circumstances, perturbation of highly activated regions does not regularly reduce prediction confidence and may sometimes increase it, indicating that the explanation method does not entirely capture the model's decision-making process.

Table 2. Quantitative explainability evaluation comparison between WEDN and ViTUNet models.

Model	XAI Method	Dice Score	Boundary Dice	Sparsity	Faithfulness
WEDN	Grad-CAM	0.6639	0.2210	0.9288	0.8072
	Occlusion Sensitivity	0.2197	0.0488	0.9383	1.0406
	Ensemble Explanation	0.5425	0.1644	0.9425	0.7409
ViTUNet	Grad-CAM	0.7648	0.4209	0.9413	0.1744
	Occlusion Sensitivity	0.2169	0.0517	0.8444	−0.3922
	Ensemble Explanation	0.7258	0.3777	0.9385	−0.2439

The QEEF results reveal clear interpretability behavioral differences between CNN-based and transformer-based segmentation architectures. WEDN achieved the highest faithfulness scores across most explanation methods, indicating strong dependence on localized discriminative regions and closer correspondence between highlighted regions and prediction behavior. Moreover, WEDN generated highly sparse explanations, confirming its reliance on short-range spatial dependencies.

In contrast, ViTUNet achieved significantly higher boundary Dice scores, particularly for Grad-CAM and ensemble explanations, indicating stronger alignment between explanation maps and anatomical structures. These findings suggest that transformer-based architectures capture more long-range coherent dependencies and structurally meaningful representations. Hence, it may provide more contextually coherent characteristics in complex medical imaging situations where contextual reasoning/understanding is key to robust segmentation performance.

Additionally, the results also show that Grad-CAM and ensemble explanations generally provide stronger edge alignment, whereas occlusion sensitivity regularly produces higher faithfulness due to its perturbation-based nature. These observations strengthen

the importance of evaluating multiple explainability aspects simultaneously rather than depending on a single metric.

To further evaluate the generalizability of the proposed QEEF, quantitative explainability experiments were carried out on the external breast cancer image dataset using the ViTUNet model.

The external validation results are illustrated in the Table 3 demonstrates that the explanation methods provided varying faithfulness, sparsity, and Dice scores. Grad-CAM and ensemble explanations were relatively structurally consistent, while occlusion sensitivity continued to present perturbation-based explanation characteristics.

Table 3. QEEF explainable results obtained from the ViTUNet model on the external breast cancer image dataset.

Model	XAI Method	Dice Score	Boundary Dice	Sparsity	Faithfulness
ViTUNet	Grad-CAM	0.9411	0.0408	0.0640	−0.0204
	Occlusion Sensitivity	0.3169	0.0254	0.2660	0.1003
	Ensemble Explanation	0.9218	0.0300	0.0557	0.0159

Even though similar qualitative patterns were observed across imaging domains, the absolute quantitative metric values differed substantially due to variations in lesion morphology, image texture, and explanation map distributions.

Figure 11 presents a visual comparison of the quantitative explainability results obtained from the WEDN and ViTUNet models on dental images.

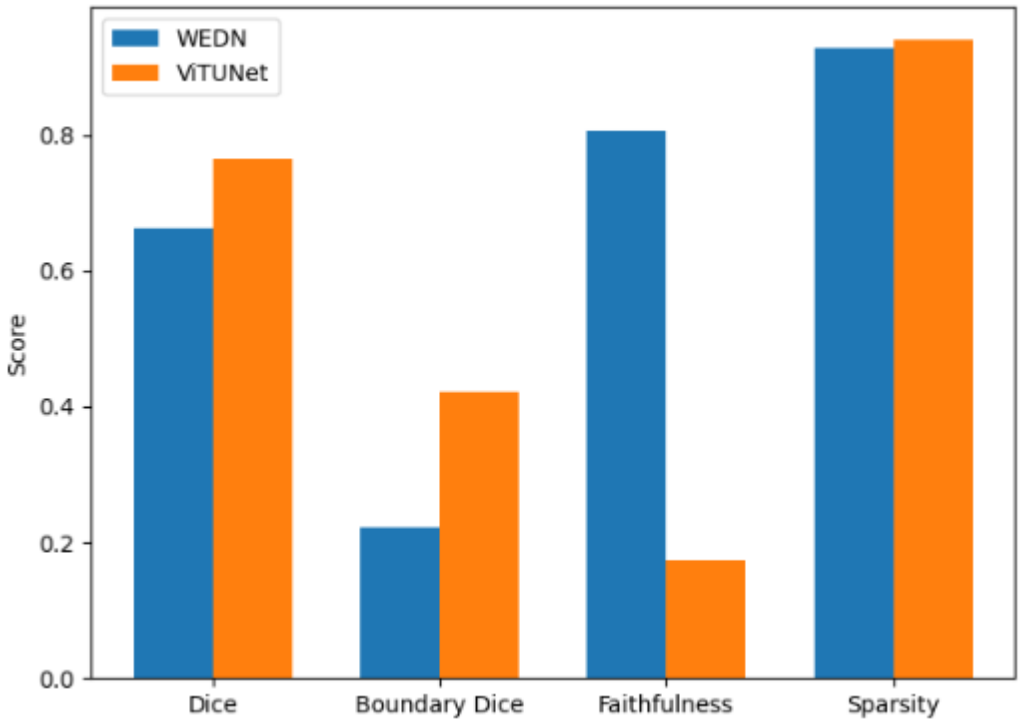

Figure 11. Visual QEEF explainability results for WEDN and ViTUNet models.

3.3. Discussion

The QEEF comparative analysis shows distinct architectural interpretability behaviors between CNN-based and transformer-based segmentation models. WEDN essentially depends on localized edge gradients, textural information, and short-range spatial rela-

tionships, leading to highly faithful and sparse explanations. However, this localized inductive bias increases sensitivity towards missing structural information and confines contextual reasoning.

Conversely, ViTUNet uses self-attention operations to model long-range anatomical dependencies, generating explanation maps that are more spatially understandable and structurally aligned with clinically relevant regions. The constantly higher boundary Dice scores observed across multiple explanation methods demonstrate contextual reasoning and improved anatomical consistency.

These findings exemplify that segmentation accuracy alone is inadequate for evaluating medical image segmentation models. Hence, the proposed QEEF provides a methodical procedure for evaluating explanation quality and interpretability via independent assessment of faithfulness, sparsity, and boundary Dice.

Negative faithfulness metrics do not mirror poor segmentation performance; instead, they indicate limitations in the reliability of the explanation method. This suggests that depending on a single explainability method may be inadequate, and that complementary techniques should be considered for well-founded analysis. Moreover, from a clinical perspective, explanation methods displaying low faithfulness scores should be interpreted with caution, as they may not consistently represent proof used by the model during prediction. Therefore, the QEEF is expected to support comparative evaluation of explanation quality instead of validating clinical preparedness. Practically, computational explainability evaluations should be interdependent with expert clinical review before implementation in decision-support systems.

Absolute QEEF metric values should not be directly compared across different imaging modalities; rather, relative patterns and ranking-based comparisons should provide a more meaningful measure of cross-domain interpretability behavior. Observed variations in sparsity scores indicate that certain QEEF metrics are domain-specific and may require normalization for robust cross-domain evaluation.

Notably, the results should be interpreted considering the limited size and diversity of the original dataset. Whilst augmentation improves robustness and helps mitigate overfitting, it cannot entirely replicate genuine clinical variability. Hence, external validation on larger multi-institutional datasets is required before clinical interpretation.

External validation on breast cancer images demonstrated that the proposed framework remains applicable under domain shifts across different medical imaging types. Whilst similar qualitative explanation patterns were observed, the quantitative explainability metrics varied substantially across modalities, suggesting that certain QEEF metrics may be domain-dependent. Moreover, this study presents a computational evaluation of explanation quality only; no formal validation by dental experts has been conducted.

3.4. Limitations

This study encounters several limitations, including but not limited to the following:

- The limited dataset with original images may restrict generalizability; thus, data augmentation cannot substitute actual clinical variability.
- Limited number of segmentation benchmarks.
- Lack of inter-observer reliability analysis.
- Lack of clinical expert validation.
- Lack of multi-center external validation.
- Domain dependence of particular QEEF metrics.
- The present study does not include repeated experiments or formal statistical significance analysis, but it describes explainability results from single trained model cases.

4. Conclusions

This study presents the QEEF, a framework for methodically assessing explainability characteristics in CNN and Vision Transformer-based medical image segmentation models. The framework was validated via two representative model architectures—namely, the Watershed Encoder–Decoder Neural Network (WEDN) and the Vision Transformer U-Net (ViTUNet)—on dental bitewing X-ray images.

Experimental results from this study demonstrated that both models achieved significant segmentation performance in terms of Dice coefficient, IoU, and accuracy. However, the explainability analysis uncovered major differences in their underlying reasoning behavior and feature representation schemes. Qualitative explainability analysis via Grad-CAM, occlusion sensitivity, and ensemble explanations illustrated that WEDN heavily relied on localized edge-based features and short-range spatial dependencies, whereas ViTUNet exhibited broader contextual reasoning and more spatially understandable attention patterns. These observations highlight that models with comparable segmentation performance may display primarily different interpretability characteristics.

Therefore, to neutrally evaluate explanation quality, the proposed QEEF used three independent evaluation dimensions, namely, faithfulness, sparsity, and boundary Dice. The quantitative results demonstrated that WEDN generally achieved higher faithfulness and sparsity scores, exhibiting stronger dependence on localized discriminative features. Conversely, ViTUNet consistently achieved higher boundary Dice scores, suggesting outstanding anatomical alignment and contextual consistency of explanation maps. These findings show how QEEF can differentiate between localized and context-sensitive explanation behavior while providing complementary insights into model interpretability.

Moreover, external validation on breast cancer images exemplified that the framework maintains information under domain shifts across different medical imaging types. Whilst similar qualitative explanation patterns were observed across imaging domains, quantitative explainability metrics varied considerably, suggesting that certain QEEF metrics may display domain dependence.

Ultimately, this study establishes that segmentation accuracy alone is inadequate for evaluating medical image segmentation systems. By presenting an independent quantitative assessment of faithfulness, sparsity, and boundary Dice, the QEEF offers a structured and replicable methodology for evaluating explanation quality, model transparency, and interpretability.

Therefore, the framework represents practicability towards more trustworthy and potentially clinically interpretable deep learning models for medical image segmentation. Instead of proposing new explainability metrics, the QEEF introduces a unified quantitative evaluation approach that integrates interdependent explainability measurements into a structured framework for comparing explanation quality across medical image segmentation models.

Future Directions

Even though the proposed QEEF demonstrated promising results across multiple explainability methods and imaging types, future work will focus on cross-device generalization, multi-center datasets, and assessment of inter-observer variability. Moreover, it will investigate k-fold cross-validation and independent external datasets to enhance the robustness and generalizability of the proposed QEEF. Several future work prospects include the following:

- Developing hybrid CNN–transformer segmentation architectures and utilizing the QEEF to methodically evaluate whether combining local feature extraction and global contextual reasoning improves both segmentation performance and explanation quality.

- Investigating additional explainability methods past Grad-CAM and occlusion sensitivity, such as integrated gradients and Layer-wise Relevance Propagation (LRP). These will explore transformer-aware attribution methods—such as attribution rollout, attention flow, and gradient-based aggregation—to determine their effectiveness within the QEEF.
- Investigating the robustness of explanations under low contrast, domain shifts, and noisy conditions to assess the explanation stability in real-world clinical setups.
- Conducting clinically guided validation studies involving clinical experts to determine whether explanation regions are identified by different XAI methods, and evaluating whether QEEF metrics correspond with clinically meaningful anatomical structures and diagnostic-relevant decision-making steps.
- Investigating automated explainability to compare systems that integrate the QEEF into the evaluation pipeline of segmentation models, facilitating standardized comparison of interpretability across different architectures and datasets.

Eventually, future work will focus on extending the QEEF towards multi-modal imaging systems, real-time clinical segmentation applications, and larger multi-institutional datasets. Therefore, such developments could establish the QEEF as a generalized quantitative benchmark for evaluating explainability and trustworthiness in medical image segmentation systems. Future work could also explore transformer-specific attribution techniques such as attention rollout, attention flow, and gradient-based attention aggregation. Moreover, it will include repeated experiments, statistical hypothesis testing, and confidence interval estimation to assess the statistical robustness of QEEF-specific comparisons.

Author Contributions: Conceptualization, V.M.; Methodology, V.M.; Validation, Y.L. and D.O.; Investigation, V.M.; Resources, E.M.; Writing—Original Draft, V.M.; Writing—Review and Editing, E.M. and D.O.; Visualization, D.O.; Supervision, E.M. and Y.L.; Project Administration, E.M., Y.L. and D.O.; Funding Acquisition, E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This Research was funded by the University of South Africa, grant number 40900.

Data Availability Statement: The data and code used to support the findings of this study will be made publicly available upon request to the corresponding authors. The bitewing dental dataset used in this study is publicly available via the Mendeley Data, V1. https://doi.org/10.17632/4fbdxs7s7w.1.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
2. Shen, D.; Wu, G.; Suk, H.I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221–248. [CrossRef] [PubMed]
3. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
4. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
5. Schwendicke, F.A.; Samek, W.; Krois, J. Artificial intelligence in dentistry: Chances and challenges. *J. Dent. Res.* **2020**, *99*, 769–774. [CrossRef] [PubMed]
6. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
7. Chen, J.; He, Y.; Frey, E.C.; Li, Y.; Du, Y. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv* **2021**, arXiv:2104.06468.
8. Zhou, H.Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; Yu, Y. Nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Trans. Image Process.* **2023**, *32*, 4036–4045. [CrossRef] [PubMed]

9. Tragakis, A.; Kaul, C.; Murray-Smith, R.; Husmeier, D. The fully convolutional transformer for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3660–3669.
10. Xie, Y.; Zhang, J.; Shen, C.; Xia, Y. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2021; pp. 171–180.
11. Li, Z.; Zheng, Y.; Shan, D.; Yang, S.; Li, Q.; Wang, B.; Zhang, Y.; Hong, Q.; Shen, D. Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. *IEEE Trans. Med. Imaging* **2024**, *43*, 2254–2265. [CrossRef] [PubMed]
12. He, A.; Wang, K.; Li, T.; Du, C.; Xia, S.; Fu, H. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Trans. Med. Imaging* **2023**, *42*, 2763–2775. [CrossRef] [PubMed]
13. Yan, Q.; Liu, S.; Xu, S.; Dong, C.; Li, Z.; Shi, J.Q.; Zhang, Y.; Dai, D. 3D medical image segmentation using parallel transformers. *Pattern Recognit.* **2023**, *138*, 109432. [CrossRef]
14. Gao, Y.; Zhou, M.; Metaxas, D.N. UTNet: A hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2021; pp. 61–71.
15. Huang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Tong, R. ScaleFormer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv* **2022**, arXiv:2207.14552.
16. Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; Kong, A.W.K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 55–68.
17. Holzinger, A. The next frontier: AI we can really trust. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing: Cham, Switzerland, 2021; pp. 427–440.
18. Sreeram, A.; Balamurugan, R.; Aditya, M.N. Explainable AI for Panoramic Dental Radiographs Using Contrastive Learning and U-Net Based Segmentation. *J. Soft Comput. Paradig.* **2025**, *7*, 114–123. [CrossRef]
19. Gipiškis, R.; Tsai, C.W.; Kurasova, O. Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey. *ICT Express* **2024**, *10*, 1331–1354. [CrossRef]
20. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
21. Bhati, D.; Neha, F.; Amiruzzaman, M. A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. *J. Imaging* **2024**, *10*, 239. [CrossRef] [PubMed]
22. Saranya, A.; Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis. Anal. J.* **2023**, *7*, 100230. [CrossRef]
23. Chaddad, A.; Hu, Y.; Wu, Y.; Wen, B.; Kateb, R. Generalizable and explainable deep learning for medical image computing: An overview. *Curr. Opin. Biomed. Eng.* **2025**, *33*, 100567. [CrossRef]
24. Borys, K.; Schmitt, Y.A.; Nauta, M.; Seifert, C.; Krämer, N.; Friedrich, C.M.; Nensa, F. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. *Eur. J. Radiol.* **2023**, *162*, 110786. [CrossRef] [PubMed]
25. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef]
26. Majanga, V.; Mnkandla, E.; Koulla Moulla, D.; Thotempudi, S.; Sena, A.D. Watershed encoder–decoder neural network for nuclei segmentation of breast cancer histology images. *Bioengineering* **2026,** *13*, 154. [CrossRef] [PubMed]
27. Majanga, V.; Mnkandla, E.; Sunday, E.O.; Ajala, B.; Sree, T. ViTUNet: Vision Transformer U-Net Hybrid Model for Carious Lesions Segmentation on Bitewing Dental Images. *Appl. Sci.* **2026**, *16*, 3693. [CrossRef]
28. Kybic, J.; Tichý, A.; Kunt, L. Dental Caries in Bitewing Radiographs. Mendeley Data, V1. 2023. Available online: https://data.mendeley.com/datasets/4fbdxs7s7w/1 (accessed on 13 July 2026).
29. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]