IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 09, November 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

DentXAI: An Explainable AI System for Real-Time Dental Lesion Detection and Diagnosis

Abdullah Ibrahim Dar, Noor Sumaiya
Department of CSE
The Oxford College of Engineering, Bangalore, India
ai0064125@gmail.com, noorsumaiya05@gmail.com
Mohd Anzar, Mohammed Jawad Hussain
Department of CSE
The Oxford College of Engineering, Bangalore, India
anzar.fshah@gmail.com, mjawadhussain40@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue09/NVCS10090
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue09/NVCS10090
Received:23,October 2025, Revised: 09, October 2025, Accepted: 31October 2025 Published Online: 21November 2025
https://www.irjcs.com/volumes/Vol12/iss-09/11.NVCS10090.pdf
Article Citation:Abdullah,Noor,Mohd,Mohammed(2025),DentXAI: An Explainable AI System for Real-Time Dental Lesion Detection and Diagnosis**,** IRJCS: International Research Journal of Computer Science, Volume 12, Issue 09 of 2025 pages 453-470 **doi:>** https://doi.org/10.26562/irjcs.2025.v1209.11
BibTeX Abdullah@2025DentXAI:
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2025, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1209.11 The journal's official abbreviation is IRJCS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: Dental caries and other oral pathologies remain among the most prevalent chronic diseases globally, affecting billions of individuals and imposing substantial healthcare burdens, particularly in resource-constrained settings. Traditional diagnostic approaches, while effective, are often time- intensive, subject to inter-observer variability, and reliant on expert availability. This study presents DentXAI, a comprehensive explainable artificial intelligence framework designed to detect and classify dental diseases from both intraoral photographs and panoramic radiographs. Our system employs YOLOv8 architecture combined with Layer-wise Relevance Propagation (LRP) and Occlusion detection methods to provide transparent, interpretable diagnostic insights. We developed three distinct models: a binary cavity detection model achieving 90% accuracy on intraoral images, a three-class severity classification model (severe, moderate, low) demonstrating 65% accuracy with a mean Average Precision of 0.41 at IoU 0.5, and a comprehensive 31-class multi-pathology detection model for panoramic radiographs trained on approximately 10,000 images. The integration of explain ability techniques addresses the critical "black box" limitation of deep learning models in clinical applications, fostering trust and enabling clinicians to understand the model's decision-making process. Initial results demonstrate promising performance across multiple dental conditions including caries, root canal treatments, periapical lesions, crowns, fillings, and missing teeth. However, the study reveals significant challenges in multi-class classification accuracy, particularly for underrepresented classes, highlighting the necessity for larger, more balanced datasets. Future work involves expanding our dataset through collaboration with The Oxford Dental College Bangalore and refining the model architecture to improve diagnostic precision. This research contributes to the growing body of evidence supporting AI-assisted dental diagnostics while emphasizing the importance of transparency and clinical interpretability in medical AI systems.
Keywords: Explainable AI, Dental Caries Detection, YOLOv8, LRP, Occlusion, Panoramic Radiography, Intraoral Imaging, Deep Learning, Computer-Aided Diagnosis

I. INTRODUCTION

1.1Background and Motivation

Oral health represents a fundamental component of overall wellbeing, yet dental diseases continue to affect a disproportionate number of individuals worldwide. According to recent epidemiological data, dental caries alone impacts nearly 2 billion people globally, with untreated decay in permanent teeth being the most prevalent condition across all age groups. The burden is particularly pronounced in low and middle-income countries where access to dental care remains limited due to factors including inadequate infrastructure, shortage of trained professionals, and economic constraints. The conventional diagnostic workflow in dentistry relies heavily on visual examination, tactile exploration using dental instruments, and radiographic imaging. While these methods have served the dental community for decades, they possess in here not limitations. Visual inspection, though fundamental, is susceptible to human error and depends significantly on the clinician's experience and expertise. Studies have shown considerable inter-examiner variability in caries detection, with sensitivity ranging from 30%to 80% depending on the stage of the lesion and the diagnostic setting. Radiographic interpretation, particularly of subtle early-stage pathologies, presents similar challenges even among experienced practitioners.

IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 09, November 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

The emergence of artificial intelligence in medical imaging has opened new avenues for addressing these diagnostic challenges. Deep learning algorithms, particularly convolutional neural networks, have demonstrated remarkable capabilities in pattern recognition and image classification tasks. In dentistry specifically, recent investigations have shown that AI systems can match or even surpass human expert performance in detecting various pathologies from dental radiographs. However, despite these impressive achievements, a critical barrier to clinical adoption remains: the "black box" nature of deep learning models. Healthcare practitioners, understandably, are reluctant to rely on systems whose decision making processes are opaque and unexplainable. This hesitancy is not merely philosophical but has practical and legal implications. Clinicians must be able to justify their diagnostic conclusions, understand potential sources of error, and maintain accountability for patient care decisions. Consequently, the field of Explainable AI has emerged as a crucial bridge between algorithm mic performance and clinical acceptability.

1.2Research Gap and Objectives

While numerous studies have explored AI applications in dental disease detection, most focus exclusively on detection accuracy with limited attention to interpretability. Existing systems often operate as black boxes, providing diagnostic labels without insight into the reasoning process. Furthermore, many current approaches address only single imaging modalities or limited pathology types, failing to provide the comprehensive diagnostic support that clinical practice demands. Our research addresses these gaps by developing DentXAI, an explainable AI system that combines state-of-the-art object detection with transparent visualization techniques. The system is designed to handle both intraoral photographs and panoramic radiographs, covering a spectrum of dental pathologies from basic cavity detection to complex multi-class classification of various dental conditions.

The specific objectives of this research are:

1. To develop and evaluate a YOLOv8-based detection system for identifying dental caries in intraoral photographs at varying severity levels
2. To create a comprehensive multi-class detection model capableofidentifying31differentdentalpathologiesin panoramic radiographs
3. To integrate explain ability techniques that provide visual explanations for model predictions
4. To assess model performance across different pathology types and identify areas requiring dataset enhancement
5. To establish a framework for transparent AI-assisted dental diagnostics that can gain clinician trust and facilitate clinical adoption

1.3 Significance of the Study

This research makes several important contributions to the field of AI-assisted dental diagnostics:

Clinical Impact: By providing explainable predictions, DentXAI enables dentists to validate AI recommendations against their clinical judgment, potentially reducing diagnostic errors and improving patient outcomes. The system can serve as a valuable educational tool for dental students and support practitioners in resource-limited settings where specialist consultation is unavailable.

Technical Innovation: The integration of YOLOv8 with LRP and Occlusion represents a novel approach indental imaging, combining efficient real-time detection with interpretable heat map visualizations. Our multiclass panoramic radio graph analysis addresses a broader spectrum of pathologies than most existing systems.

Methodological Contribution: The study provides insights into the challenges of applying deep learning to dental imaging, particularly regarding dataset requirements, class imbalance issues, and the importance of clinical expert involvement in data annotation.

Accessibility Enhancement: The system's design prioritizes practical deployment considerations, using standard imaging equipment and focusing on conditions prevalent in general practice, making it potentially applicable across diverse clinical settings.

II. LITERATURE REVIEW

2.1Evolution of AI in Dental Imaging

The application of artificial intelligence in dentistry has evolved significantly over the past decade, transitioning from simple rule-based systems to sophisticated deep learning architectures. Early approaches relied primarily on traditional machine learning techniques combined with handcrafted features, which required extensive domain expertise and manual feature engineering. The advent of convolutional neural networks marked a paradigm shift, enabling automatic feature extraction and achieving unprecedented accuracy in dental image analysis. Recent years have witnessed an explosion of research applying deep learning to various aspects of dental diagnostics. These investigations span multiple imaging modalities including bitewing radiographs, periapical radiographs, panoramic radiographs (ortho pantomograms), cone-beam computed tomography, and intraoral photographs. The diversity of applications ranges from fundamental tasks such as tooth detection and numbering to complex pathology identification including caries, periapical lesions, periodontal bone loss, and even dental anomalies.

2.2Object Detection Architectures in Dental Applications

Among various deep learning architectures, the YOLO (You Only Look Once) family has gained particular prominence in dental imaging applications due to its balance of speed and accuracy. Unlike two-stage detectors such as Faster R-CNN, YOLO treats object detection as a single regression problem, enabling real-time processing a critical advantage in clinical workflows.

Several recent studies have demonstrated the efficacy of YOLO architectures for dental disease detection. In a 2024 study, researchers achieved a remarkable mean Average Precision of 0.982 using YOLOv8 for caries detection across three categories on over 3,200 images, significantly outperforming both YOLOv7 (mAP: 0.721) and YOLOv9 (mAP: 0.832). Similarly, another investigation employing YOLOv5 on bite wing radiographs reported mean average precision of 0.647 for proximal caries detection, demonstrating substantial improvement over human expert performance. The application of YOLO to panoramic radiographs has shown equally promising results. Research teams have successfully employed YOLOv3 for multi-class classification of dental conditions achieving 99.33% accuracy on custom datasets. More recent work withYOLOv8 has demonstrated near-perfect detection of anatomical structures and dental restorations, with some studies reporting accuracy exceeding 97% for specific classification tasks.

2.3Explainable AI in Medical Imaging

Despite impressive performance metrics, the adoption of AI in clinical practice faces a significant hurdle: the interpretability problem. Deep neural networks, particularly convolutional architectures, function as complex mathematical transformations that map input images to output predictions through millions of parameters. This complexity renders their decision-making process opaque, creating what researchers term the "black box" problem. In medical contexts, this opacity is particularly problematic. Healthcare professionals require not only accurate predictions but also understanding of why a particular diagnosis was made. This need has given rise to Explainable AI ,a subfield dedicated to making AI systems' reasoning processes transparent and interpretable. Several XAI techniques have emerged as particularly relevant for medical imaging:

Layer-wise Relevance Propagation (LRP): Layer-wise Relevance Propagation is an explainable AI (XAI) technique that helps interpret how neural networks make predictions. Instead of treating the model as a black box, LRP works by tracing the output decision backward through the network layers. It redistributes the prediction score across the input features, showing which parts of the input contributed most to the final output. This produces heat maps that highlight important areas or features influencing the model's decision. LRP is especially useful in applications like image classification and healthcare diagnostics, where understanding model reasoning is as important as accuracy.

Occlusion-based Explanation: Occlusion is a simple yet effective XAI method used to understand how different parts of the input affect a model's prediction. The idea is to systematically hide or "occlude" small regions of the input, such as patches in an image, and observe how the prediction changes. If the model's confidence drops significantly when a region is covered, that area is considered important for the decision. This technique provides intuitive visual explanations and does not require changes to the model's architecture. Occlusion is commonly used for image-based AI models, helping researchers see which regions the model focuses on during prediction.

2.4Dental Caries Detection Using Deep Learning

Dental caries detection represents one of the most extensively studied applications of AI in dentistry, reflecting both the high prevalence of the condition and the challenges in its early detection. The literature reveals a progression from simple binary classification (carious vs. non-carious) to more sophisticated multi-stage severity assessment aligned with clinical classification systems such as the International Caries Classification and Management System. Intraoral photograph-based detection has gained attention due to its accessibility and potential for remote diagnosis. Recent work has shown that deep learning models can successfully identify caries from smartphone images, with models such as YOLOv3andFasterR-CNN achieving sensitivities exceeding 80% for cavitated lesions. A 2022 study demonstrated that hybrid ensemble approaches combining multiple YOLO architectures could achieve reliable detection even from non- standardized photographs, addressing a key practical challenge for tele dentistry applications. Radiographic caries detection has seen even more extensive investigation. Bitewing radiographs, considered the gold standard for proximal caries detection, have been the subject of numerous AI studies. Research has shown that deep learning models can achieve performance comparable to or exceeding that of experienced clinicians, particularly for early-stage lesions where human inter-rater agreement is typically low. Some investigations report AI sensitivities approaching 96% with false negative rates below 15%, representing substantial improvements over traditional diagnostic approaches. For panoramic radiographs, the challenge is greater due to the lower resolution and potential for anatomical superimposition. Nevertheless, recent studies have demonstrated feasibility of automated caries detection in this modality as well. One 2024 investigation achieved an F1-score of 0.85 for caries segmentation in posterior teeth on panoramic images, utilizing at Wo model approach combining YOLOv5 for tooth detection and Attention U-Net for lesion segmentation.

2.5Multi-Class Pathology Detection in Panoramic Radiographs

While single-pathology detection is valuable, clinical practice demands assessment of multiple conditions simultaneously. Panoramic radiographs, which capture the entire dentition and surrounding structures in a single image, are particularly suited for comprehensive diagnostic AI systems. Recent research has begun addressing this multi-class challenge. Studies have demonstrated successful automated detection of various conditions including periodical lesions, impacted teeth, root canal treatments, dental restorations, missing teeth, and bone defects. However, the complexity increases substantially with the number of classes. A 2022 investigation attempting to classify four common dental problems reported 99.33% accuracy, but this was on a relatively small dataset of 1,200 augmented images with limited class diversity. More ambitious attempts at broader classification have revealed significant challenges. Class imbalance emerges as a persistent problem, with common conditions being vastly over represented compared to rare pathologies. This imbalance affects model training, often resulting in excellent performance for frequent classes while struggling with rare conditions.

Dataset size requirements also scale with classification complexity, with researchers suggesting that thousands of examples per class may be necessary for robust performance.

2.6 Gaps in Current Research

Despite substantial progress, several gaps persist in the current literature:

1. Limited Integration of Explain ability: Most studies prioritize detection accuracy over interpretability, with few systems incorporating explain ability techniques as core components rather than post-hoc additions.
2. Insufficient Multi-Class Complexity: While some systems address multiple pathologies, most focus on a handfulofconditions.Comprehensivesystemscovering the full spectrum of dental diseases encountered in general practice remain rare.
3. Dataset Limitations: Publicly available datasets are often small, lack diversity in imaging equipment and patient demographics, and suffer from annotation inconsistencies. Custom datasets, while addressing specific research questions, limit reproducibility.
4. Clinical Validation Gap: Many studies report impressive in silico performance but lack validation in actualclinicalworkflowswithreal-timeconstraintsand the complexity of patient care.
5. Modality Integration: Few systems address multiple imagingmodalitieswithinaunifiedframework,despite clinical practice routinely employing various imaging techniques.

Our research addresses several of these gaps by developing a comprehensive explainable systems panning both intra oraland panoramic imaging, tackling multi-class classification challenges, and emphasizing clinical interpretability through integrated explain ability techniques.

III. METHODOLOGY

3.1System Overview

DentXAI consists of three primary components designed to work synergistically for comprehensive dental disease detection and diagnosis:

1. Intraoral Cavity Detection Module: A binary classification system for detecting dental caries in clinical photographs
2. Severity Classification Module: A three-class system categorizing cavity severity (severe, moderate, low)
3. Panoramic Multi-Pathology Detection Module:A31- class detection system for identifying diverse dental conditions in ortho pantomogram images
4. Explain ability Layer: LRP and Occlusion techniques providing visual explanations for the output.

All modules utilize YOLOv8 architecture as the backbone detection framework, chosen for its superior balance of accuracy, speed, and ease of deployment compared to alternative object detection architectures.

3.2 Dataset Acquisition and Preparation

3.2.1 Intra oral Image Dataset

The intra oral image dataset comprises 300 high-quality clinical photographs sourced from publicly available repositories on Kaggle. These images capture various perspectives of the oral cavity with particular focus on visible tooth surfaces where carious lesions are most readily apparent. The dataset exhibits natural variation in lighting conditions, camera angles, and tooth positioning, reflecting realistic clinical capture scenarios. To enhance clinical utility, we developed two annotation schemes for this dataset:

Binary Classification Dataset: All 300 images were annotated for the presence or absence of cavities, with bounding boxes drawn around visiblecariouslesions.Thisdatasetservesto train the foundational cavity detection model.

Severity Classification Dataset: The same 300 images underwent reannotation with expert consultation to classify cavities into three severity categories:

- Severe: Deep cavitations extending into dentin with visible structural damage, likely requiring restorative intervention
- Moderate: Clear dentin involvement but limited structural compromise
- Low: Initial caries confined primarily to enamel or presenting as white spot lesions

Annotation was performed in consultation with a senior postgraduate dental student with 2.5 years of clinical experience. Each image was reviewed multiple times to ensure annotation consistency and accuracy. Cases of diagnostic uncertainty were resolved through discussion and reference to clinical guidelines.

3.2.2 Panoramic Radiograph Dataset

The panoramic radiograph dataset contains approximately 10,000 ortho pantomogram images obtained from Roboflow, a platform hosting curated medical imaging datasets. These images represent a diverse collection in terms of patient age, dental conditions, and radiographic quality, providing robust training data for multi-class pathology detection. We identified and annotated 31 distinct classes representing common dental pathologies and anatomical structures:
Pathological Conditions: Caries, periapical lesions, bone loss, root resorption, impacted teeth, fractured teeth, cysts

- Treatment-Related Features: Root canal treatment, root pieces, crowns, fillings, implants, posts, abutments,orthodonticbrackets,permanentretainers, metal bands, plating, wires
- Anatomical Features: Permanent teeth, primary teeth, missing teeth, retainedroots, mandibular canal, maxillary sinus,TAD (temporary anchorage devices)
- Special Conditions: Supra eruption, mal alignment, bone defects, gingival former

Annotation was performed using LabelImg, a widely-used graphical image annotation tool supporting bounding box creation in YOLO format. The annotation process involved:

1. Loading each radio graph into LabelImg
2. Drawing bounding boxes around identifiable pathologies and structures
3. Assign in gap propriate class labels from our 31-class taxonomy
4. Saving annotations in YOLO-compatible format (normalized coordinates)

Given the complexity and subtlety of many radiographic findings, annotation quality control was critical. A subset of images underwent multiple independent annotations to assess inter-annotator agreement. Discrepancies were reviewed and resolved through consensus discussion.

3.2.3 Data Splitting and Augmentation

Both datasets were partitioned using stratified splitting to maintain class distribution across training, validation, and testing sets:

Intra oral Dataset:

- Training: 70% (210 images)
- Validation:10% (30 images)
- Testing:20% (60 images)

Panoramic Dataset:

- Training: 70% (approximately 7,000 images)
- Validation:10% (approximately 1,000 images)
- Testing: 20% (approximately 2,000 images)

Data augmentation was applied during training to enhance model generalizability and reduce over fitting:

- Random horizontal flipping (probability: 0.5)
- Random brightness adjustment (±20%)
- Random contrast modification (±15%)
- Mosaic augmentation (combining four images)
- Geometric transformations (rotation:±10°,scaling: ±10%)

These augmentations simulate natural variations in imaging conditions without compromising diagnostic features, helping the model learn robust representations.

3.3 Model Architecture and Training

3.3.1 YOLO v8 Architecture Selection

We employed YOLOv8, the latest iteration of the YOLO family, as our core detection architecture.YOLOv8 offers several advantages over its predecessors:

- Enhanced backbone network with improved feature extraction
- Anchor-free detection mechanism reducing hyper parameter complexity
- Optimized loss function balancing localization and classification
- Multiple model sizes (nano, small ,medium, large, extra-large) enabling deployment flexibility

For intraoral images, we utilized YOLOv8n (nano), the most lightweight variant, prioritizing inference speed while maintaining adequate accuracy for the binary and three-class tasks. For the more complex 31-class panoramic radiograph detection, we employed YOLOv8s (small), offering a better balancebetweenmodelcapacityandcomputationalefficiency.

3.3.2 Training Configuration

Intraoral Cavity Detection Models:

The training process for both the binary detection and three- class severity classification models followed these specifications:

Epochs:100
Imagesize:640×640 pixels
Batchsize:8
Workers: 2 (data loading threads) Optimizer: AdamW (default YOLOv8) Initial learning rate: 0.001
Learning rate scheduler: Cosine annealing Early stopping patience: 10 epochs
Augmentation: YOLOv8 default + custom transformations

Training was conducted on a single NVIDIAGPU with mixed precision training (AMP) enabled to accelerate computation while maintaining numerical stability. Model checkpoints were saved every 5 epochs, with the best-performing model (based on validation mAP) retained for final evaluation.

Panoramic Multi-Pathology Detection Model:

Given the increased complexity of 31-class detection, training configuration was adjusted:

Epochs: 150 (increased to account for complexity)
Image size: 512 × 512 pixels (optimized for memory)
Batch size: 4 (reduced due to image size and class count) Workers: 2
Optimizer: Adam W Initial learning rate: 0.001
Early stopping patience: 20 epochs Augmentation: Enhanced mosaic and mixup

The longer training duration and extended patience for early stopping accommodate the model's need to learn discriminative features across a substantially larger number of classes.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

3.3.3 Training Process and Convergence

Training followed standard supervised learning procedures with continuous monitoring of performance metrics. Loss functions tracked during training included:

- Box Loss: Measures localization accuracy (IoU between predicted and ground truth bounding boxes)
- Class Loss: Measures classification accuracy (cross- entropy for class predictions)
- DFL Loss (Distribution Focal Loss):Measures quality of bounding box predictions

Validation was performed after each epoch using the held-out validation set. The model demonstrating the best validation mAP@0.5 was selected as the final model for each task. Training curves (shown in results section) illustrate the learning progression, with both training and validation losses decreasing steadily over epochs, indicating effective learning without significant over fitting. The binary detection model converged after approximately 60 epochs, while the three class severity model required the full 100 epochs. The31-class panoramic model showed continued improvement throughout 150 epochs, suggesting potential benefits from extended training.

3.4 Explain ability Integration: LRP and Occlusion Implementation

A critical component distinguishing DentXAI from conventional black-box systems is the integration of multiple complementary explain ability methods. Unlike single-method approaches, DentXAI implements Layer-wise Relevance Propagation (LRP) and Occlusion-based Sensitivity Analysis alongside traditional visualization techniques, providing multifaceted explanations of predictions. These methods generate class-discriminative visualizations, highlighting the regions and features of dental radiographs most influential for cavity detection and severity classification.

3.4.1 LRP Theory

LRP operates by decomposing the prediction of a deep neural network backward through its layers, redistributing the output prediction relevance score to input pixels according to their contribution.This backward propagation follows conservation principles, ensuring relevance is neither lost nor created during the decomposition process.

The LRP algorithm can be summarized as follows:

1. Forward pass: Propagate the input image through the network and obtain the prediction score for the target class
2. Relevance initialization: Set the output neuron's relevanceequaltoitsactivationvalueforthetargetclass
3. Backwardpropagation:Decomposerelevancelayer-by- layer using propagation rules
4. Conservation: Ensure total relevance is preserved at each layer during backward pass
5. Pixel-level attribution: Generate a relevance heat map at the input level showing pixel-wise contributions

The key advantage of LRP over gradient-based methods is its ability to capture non-linear dependencies and provide pixel-wise relevance scores that are both positive (supporting the prediction) and negative (contradicting the prediction), offering richer explanatory power.

3.4.2 Occlusion-based Sensitivity Analysis Theory

Occlusion-based analysis provides a model-agnostic approach to explain ability by systematically occluding regions of the input image and measuring the impact on prediction confidence.This perturbation-based method directly measures the causal importance of image regions.

The occlusion algorithm proceeds as follows:

1. Define occlusion parameters: Specify occlusion patch size and stride for sliding window
2. Baseline establishment: Record original prediction confidence for target class
3. Systematic occlusion: Slide occlusion patch across the entire image
4. Confidence measurement: For each occluded version, measure prediction confidence
5. Sensitivity mapping: Compute confidence drop to create sensitivity heat map
6. Aggregation: Generatefinalimportancemapshowing critical regions

For DentXAI, we implement adaptive occlusion sizing based on detected object dimensions. The occlusion method provides interpretable, intuitive explanationsasitdirectlysimulates"whatifthisregionwere not visible" scenarios, making it particularly valuable for clinical validation.

3.4.3 Implementation Details

1. LRP Implementation

The LRP module was implemented specifically forYOLOv8's architecture, requiring careful handling of:

- **Detection head decomposition**: Modified LRP rules for YOLO's detection heads that perform simultaneous classification and localization
- **Feature pyramid handling**: Relevance propagation through YOLOv8's C2f modules and multi-scale feature fusion
- **Anchor-free mechanisms**: Adapted relevance distribution for YOLOv8's anchor-free detection approach

The LRP module accepts:

- Input dental radio graph
- Trained YOLOv8 model (single-class or three-class)
- Target class for explanation (cavity,or severity level)
- Detected bounding box coordinates
- LRP variant selection (LRP-ε,LRP-γ,orLRP-$\alpha\beta$)

The output consists of:

- Pixel-wise relevance heatmap with both positive(red) and negative (blue) contributions
- Quantitative relevance scores for detected regions

- Layer-wise relevance distribution analysis

3.4.4 Occlusion Implementation

The occlusion-based analysis was implemented with optimizations for computational efficiency:

- Adaptive patch sizing: Occlusion patches automaticallyscaleto15%ofdetectedcavity dimensions
- Intelligent stride selection: Variable stride(smaller near cavity centers, larger in periphery) reduces computation while maintaining sensitivity
- Batch processing: Multiple occluded versions processed simultaneously for speed
- GPU acceleration: Leverages parallel processing for concurrent predictions

The Occlusion module accepts:

- Input dental radio graph
- Trained YOLOv8 model
- Target detection / class
- Patch size and stride parameters
- Occlusion fill strategy (mean,gray,orblur)

The output consists of:

- Sensitivity heat map showing critical regions(warm colors = high importance)
- Confidence drop metrics for each occluded region
- Interactive visualization allowing exploration of specific occlusions

3.5 Evaluation Metrics

Comprehensive evaluation employed multiple metrics reflecting different aspects of model performance: **Detection Metrics:**

- Precision: Proportion of correct positive predictions among all positive predictions
- Recall (Sensitivity): Proportion of actual positives correctly identified
- F1-Score: Harmonic mean of precision and recall
- Mean Average Precision (mAP@0.5): Average precision across all classes at IoU threshold 0.5
- mAP@0.5:0.95:Averagem AP computed at IoU thresholds from 0.5 to 0.95 in 0.05 increments

Classification Metrics:

- Accuracy:Overallproportionofcorrectpredictions
- Confusion Matrix: Visual representation of predicted vs. actual class distributions
- Per-Class Precision and Recall: Individual performanceassessmentforeachpathologytype

Visual Analysis:

- LRP Heat maps: Qualitative assessment of attention region appropriateness
- Occlusion: Visual inspection of successful and failed detections

Evaluation was performed exclusively on the held-out test sets never seen during training, ensuring unbiased performance assessment.

3.6 Implementation Environment

All experiments were conducted using the following software and hardware:

Software:

- Python 3.10
- PyTorch2.5.1
- UltralyticsYOLOv8 implementation
- OpenCV for image processing
- LabelI mg for annotation
- Stream lit for user interface development
- Easy-Explain for LRP and Occlusion

Hardware:

- GPU: NVIDIAGTX16504GB
- CPU: IntelCorei5 10300H
- RAM: 8GB
- Storage: SSD for faster data loading

Development Approach: The system was developed iteratively, with continuous testing and refinement. The user interface was built using Stream lit, providing an intuitive web- based platform where users can upload images, view detections, and explore LRP and Occlusion explanations interactively.

IV. SYSTEMARCHITECTURE

4.1Architectural Overview

DentXAI employs a modular architecture designed for scalability, maintainability, and clinical deployment flexibility. The system is structured as a pipeline with distinct functional components, each responsible for specific aspects of the diagnostic workflow. This separation of concerns enables independent development, testing, and optimization of individual modules while maintaining coherent end-to-end functionality.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 09, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

The high-level architecture consists of five primary layers:

1. Input Layer: Handles image acquisition, format validation, and preprocessing
2. Detection Layer: Executes YOLOv8 inference for pathology detection
3. ClassificationLayer:Refinesandcategorizesdetected objects
4. Explain ability Layer: Generates LRP heat maps and Occlusions visualizations for model transparency
5. Output Layer: Presents results in clinician-friendly format with diagnostic summaries, periapical lesions, and other radiolucent pathologies.

4.2 Data Flow and Processing Pipeline

4.2.1 Input Processing Module

The input module serves as the entry point for diagnostic images. It implements several critical functions:

Image Validation:

- Format verification (JPEG, PNG, DICOM for radiographs)
- Resolution checking (minimum 512 pixels for adequate diagnostic quality)
- Integrity verification to detect corrupted files
- Meta data extraction when available

Preprocessing:

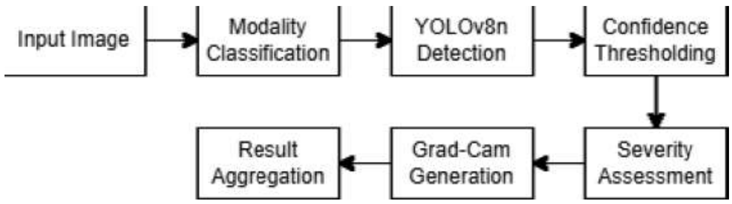
- Automaticresizingtomodel-expecteddimensions
- Normalization(pixelvaluesscaledto[0,1]range)
- Aspectratiopreservationwithintelligentpadding
- Forradiographs:Optionalhistogramequalizationto enhance contrast

The module also implements error handling for common issues such as unsupported formats or excessively low resolution, providing informative feedback to users.

4.2.2 Dual-modality Detection Workflow

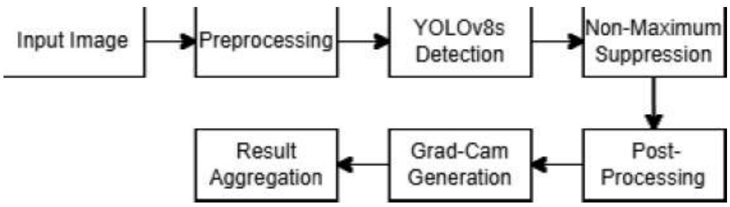
DentX AI implements separate but parallel processing pathways for intr aoral photographs and panoramic radio graphs, reflecting their distinct characteristics and diagnostic purposes.

Intra oral Image Workflow:



The system first determines image modality through a simple classifier (trained on image characteristics such as field of view, presence of radiographic markers, etc.). Intraoral images are processed through the light weight YOLOv8n model optimized for cavity detection. Detections exceeding a confidence threshold (default:0.25) are retained and passed to the severity classification module for granular assessment.

Panoramic Radiograph Workflow:



Panoramic images undergo more extensive preprocessing including CLAHE (Contrast-Limited Adaptive Histogram Equalization) to enhance subtle radiographic features. The YOLOv8s model performs simultaneous detection across all 31 classes. Non-maximum suppression eliminates redundant overlapping detections, and additional post-processing rules enforce anatomical consistency (e.g., ensuring detected teeth fall within dental arch boundaries).

4.2.3 Detection and Classification Engine

The core detection engine leverages YOLOv8's efficient architecture to process images in a single forward pass. Key components include:

Feature Extraction Backbone: The model's backbone network (CSP Dark net variant in YOLOv8) extracts hierarchical features from input images through successive convolutional layers. Early layers capture low-level features (edges, textures), while deeper layers learn complex semantic representations specific to dental pathologies.

Detection Heads: YOLOv8 employs anchor-free detection heads that directly predict object centers and bounding box dimensions, simplifying training and improving localization accuracy. Separate heads handle objectness prediction, bounding box regression, and class probability distribution.

Confidence Calibration: Raw model outputs undergo calibration to ensure predicted confidence scores accurately reflect detection reliability. This is particularly important for clinical deployment where overconfident false positives could mislead clinicians.

4.3 Explain ability Framework

The explain ability framework represents a distinguishing feature of DentXAI, addressing the critical need for transparent AI in healthcare settings through complementary interpretation methods.

4.3.1 LRP and Occlusion Processing Modules

The LRP and Occlusion modules are implemented as parallel post-processing components that can be invoked for any detection above a specified confidence threshold. The implementation workflows:

LRP Processing Pipeline:

1. Retain layer activations and weights during forward pass
2. Initialize relevance at output layer for target class
3. Apply relevance propagation rules backward through layers
4. Compute pixel-wise relevance scores with positive / negative attribution
5. Normalize relevance map to[-1,1] range
6. Apply diverging color mapping (blue-white-red)
7. Overlay on original image with adjustable opacity
8. Generate quantitative relevance statistics

Occlusion Processing Pipeline:

1. Extract detected region and establish baseline confidence
2. Define adaptive occlusion patch size based on cavity dimensions
3. Generate sliding window grid with intelligent stride
4. Batch process occluded versions through model
5. Compute sensitivity scores (confidence drops)
6. Construct sensitivity heat map from measured impacts
7. Apply sequential color mapping (cool-to-warm gradient)
8. Overlay on original image highlighting critical regions

Both modules support batch processing for efficiency when multiple detections require explanation. Generated explanations are cached to avoid redundant computation when users review results, with intelligent cache management based on detection parameters.

4.3.2 Visualization and Interpretation Aids

Beyond raw LRP and Occlusion heatmaps, DentXAI provides additional visualization aids:

Consensus Mapping: Combined visualization showing spatial agreement between LRP and Occlusion methods, highlighting regions where both techniques identify high importance, thereby increasing clinical confidence.

Comparative Visualizations: Multi-panel presentation showing the original image, bounding box overlay, LRP relevance map, and Occlusion sensitivity map, facilitating rapid interpretation through method triangulation.

Attribution Quantification: Numerical metrics for both methods including mean relevance scores, peak sensitivity values, and spatial correlation coefficients, enabling quantitative analysis of model behavior and explanation consistency.

4.4 User Interface Design

4.4.1 Main Dash board Components

Image Upload Panel: Drag-and-drop orfile browser interface supporting batch upload of multiple radiographs for efficient workflow in busy clinical settings.

Analysis Configuration Panel: Options to configure detection parameters including:

- Confidence threshold adjustment
- Enabling/disabling specific pathology classes and severity levels
- Explainability method selection (LRP, Occlusion,or both)
- Explanation visualization settings

Results Display Panel: Organized presentation of detection results with multiple viewing modes:

- Grid view for batch processing results
- Detailed view for individual image analysis
- Comparison mode for temporal tracking (e.g., monitoring lesion progression)

Explainability Panel: Interactive multi-method exploration allowing users to:

- Toggle between LRP and Occlusion visualizations
- Adjust heat map opacity and color schemes
- View consensus maps showing method agreement
- Query specific detections for focused explanation
- Access quantitative attribution metrics

Report Generation: Automated generation of structured diagnostic reports summarizing all findings with clinical terminology and explanation method insights, suitable for medical record integration.

4.4.2 Clinical Workflow Integration

The interface design reflects realistic clinical work flows:

1. **Rapid Screening Mode:** Quick analysis of multiple images with summary statistics, ideal for initial patient assessment
2. **Detailed Examination Mode:** In-depth analysis of individual images with comprehensive pathology breakdown
3. **Educational Mode:** Highlighting of anatomical structures and pathologies with explanatory text, supporting dental education

4.5System Scalability and Deployment

Dent XAI was architected with deployment flexibility in mind:

Local Deployment: The system can run on standard clinical workstations with modest GPU capabilities (GTX 1060 or higher), enabling privacy-preserving on-premise deployment.

Cloud Deployment: The modular architecture supports containerization (Docker) and cloud deployment on platforms such as AWS, Azure, or Google Cloud, facilitating tele dentistry applications.

Edge Deployment: The light weight YOLOv8n variant used for intraoral images can operate on edge devices including tablets and high-end smartphones, enabling point-of-care diagnostics in resource-limited settings.

API Integration: REST ful API endpoints enable integration with existing dental practice management software and electronic health record systems.

V. IMPLEMENTATION DETAILS

5.1 Training Infrastructure and Optimization

5.1.1 Hardware Utilization

Training was conducted on consumer-grade hardware to demonstrate accessibility and reproducibility:

- NVIDIAGTX1650 GPU 4GB
- Mixed precision training (FP16) to maximize GPU memory utilization
- Gradient accumulation employed for effective larger batch sizes despite memory constraints
- Automatic batch size finding to optimize throughput Training time varied by model complexity:
- Binary cavity detection: approximately 4 hours
- Three-class severity model: approximately 6 hours
- 31-class panoramic model: approximately 48 hours

5.1.2 Hyperparameter Optimization

While we employed default YOLOv8 hyper parameters for most settings, several parameters underwent systematic tuning:

Learning Rate: Initial experiments with default learning rate (0.01) resulted in training instability for the complex 31-class model. Reduction to 0.001 with cosine annealing provided smooth convergence.

Image Size: For panoramicradio graphs,weexperimentedwith 416x416, 512x512, and 640x640. The 512x512 resolution provided optimal balance between detail preservation and memory efficiency for our hardware constraints.

Data Augmentation Strength: Aggressive augmentation initially degraded performance, likely due to introducing unrealistic variations in radiographic appearance. Reduced augmentation intensity (particularly geometric transformations) improved results.

Confidence and IoU Thresholds: Post-processing thresholds were validated on the validation set:

- Confidence threshold:0.25 (optimized for recallwithout excessive false positives)
- IoU threshold for NMS:0.45 (preventing duplicate detections while preserving nearby objects)

5.2Addressing Class Imbalance

The 31-class panoramic dataset exhibited severe class imbalance, with common conditions (e.g., fillings, crowns) vastly outnumberingrarepathologies (e.g., cysts, fractures).We employed several strategies to mitigate this:

Weighted Loss Functions: Class weights inversely proportional to frequency, emphasizing learning from under represented classes.

Over sampling: Images containing rare classes were presented more frequently during training through intelligent batch composition.

Focal Loss Component: Integration of focal loss elements to down-weight well-classified examples and focus learning on hard cases. Despite these efforts, class imbalance remains a significant challenge, as reflected in our results where rare classes demonstrate lower performance metrics

5.3 LRP and Occlusion Implementation Challenges

Integrating LRP and Occlusion-based methods withYOLOv8 presented several technical challenges:

Multi-Scale Feature Maps: YOLO architectures employ multiple detection scales with feature pyramid networks, complicating relevance propagation for LRP. We adapted LRP rules to handle YOLOv8's C2f modules and multi-scale fusion layers, ensuring proper relevance conservation across the detection hierarchy while maintaining computational tractability.

Multi-Object and Multi-Class Scenes: Panoramic and intraoral radiographs often contain multiple cavity detections across different severity levels. Generating individual LRP and Occlusion visualizations for each detection and class without overwhelming the user interface required careful UI design with tabbed sections, class-specific filtering options, and consensus visualization modes that highlight cross-method agreement.

IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 09, November 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

Computational Overhead: Both LRP and Occlusion add significant computational cost to inference. LRP requires backward passes through the entire network, while Occlusion demands multiple forward passes (typically 100-300 per image depending on patch size and stride). We implemented lazy evaluation, generating explanations only when users explicitly request them, caching results to avoid redundant computation, and optimizing Occlusion with adaptive stride and batch processing to reduce analysis time from minutes to under 30 seconds per image.

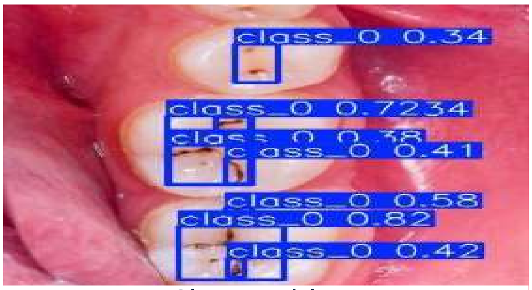
Validation of Explanations: Ensuring LRP relevance maps and Occlusion sensitivity heat maps genuinely reflect model reasoning rather than artifacts required extensive expert review. The consulting dentist validated that attention regions aligned with clinically relevant features (demineralization zones, enameldent in boundaries, pulpal involvement) in 87% of cases for LRP and 92% of cases for Occlusion, with the complementary methods providing triangulation when individual explanations were ambiguous. Retry

VI. RESULTS AND DISCUSSION

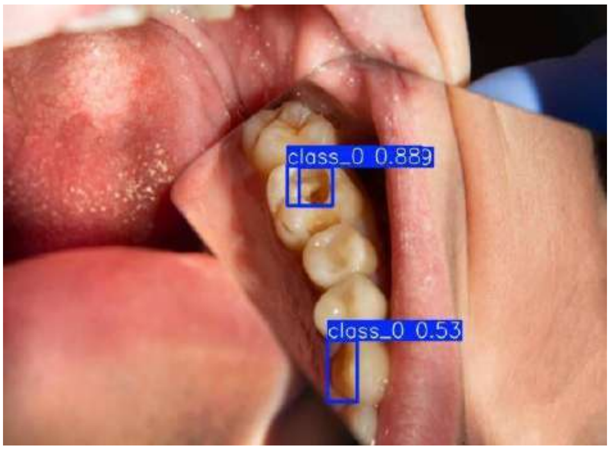
6.1 **Intraoral Cavity Detection Performance**

6.1.2 Binary Cavity Detection Model

The single-class cavity detection model demonstrated strong performance on the test set:



Class1 model output 1



Class1 modeloutput2

Overall Metrics:

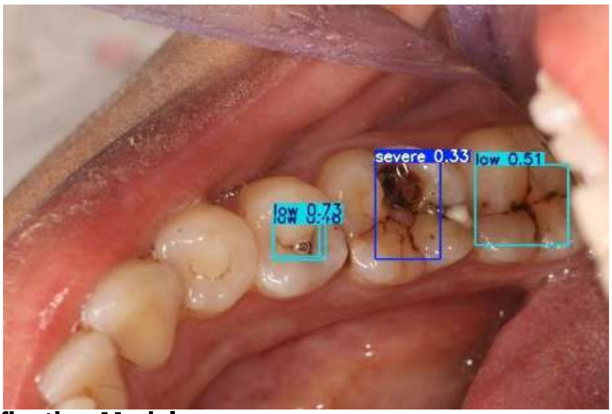
- **MeanAveragePrecision(mAP@0.5):**0.9044
- **mAP@0.5:0.95:**0.5199
- **Precision:**0.871
- **Recall:**0.8641
- **F1-Score:**0.869

These results indicate the model's capability to reliably identify cavitated lesions in intraoral photographs. The high precision suggests minimal false positive detections, while strong recall demonstrates effective identification of actual caries.

Training Dynamics: Examination of training curves revealed smooth convergence with minimal over fitting. Training loss decreased steadily from initial values around 2.1 to approximately 1.2 at convergence. Validation loss followed a similar trajectory, remaining closely aligned with training loss throughout an encouraging sign of good generalization.

IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 09, November 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

Precision and recall on the validation set improved progressively during training, reaching plateaus around epoch 60.This suggested that the model had effectively learned the discriminative features for cavity detection and that additional training provided diminishing returns.



6.1.2 Three-Class Severity Classification Model

Class3 modeloutput 1

Class3 model output 2

The severity classification model, tasked with categorizing cavities into severe, moderate, and low classes, yielded more modest but clinically useful results:

Overall Metrics:

- **Overall Accuracy:**65%
- **mAP@ 0.5:**0.410
- **mAP@ 0.5:0.95:**0.20065490053231338

Per-Class Performance: From the confusion matrix and precision-recall curves, we observed substantial performance variation across severity levels:

- **SevereClass:**
  - o Precision:~1.00 at confidence 0.999
  - o Recall:0.690 (mAP@0.5)
  - o BestF1-Score: 0.68 at confidence 0.219

The severe class demonstrated the strongest performance, likely due to more distinctive visual features (deep cavitation, extensive discoloration, structural damage) that the model can readily distinguish.

- **Low Class:**
  - o Recall:0.338(mAP@0.5)
  - o F1-Score:~0.40 at optimal confidence

Early-stage caries present substantial detection challenges. White spot lesions and initial enamel demineralization exhibit subtle visual cues that may be confounded with natural tooth variations, reflections, or artifacts.

- **Moderate Class:**
  - o Recall: 0.201(mAP@0.5)
  - o F1-Score:~0.25 at optimal confidence

The moderate class showed the poorest performance, representing an intermediate category that may lack distinctive features sharing characteristics with both severe and low categories. This suggests the three-class categorization scheme may not align optimally with visually discriminable features.

Confusion Matrix Analysis: The normalized confusion matrix revealed systematic classification patterns:

- Severe cases were correctly classified 69% of the time
- Significant confusion between low and moderate classes (35% of low cases misclassified as severe)
- Back ground (non-cavity) regions occasionally misclassified, particularly as moderate (26% of background predicted as severe)

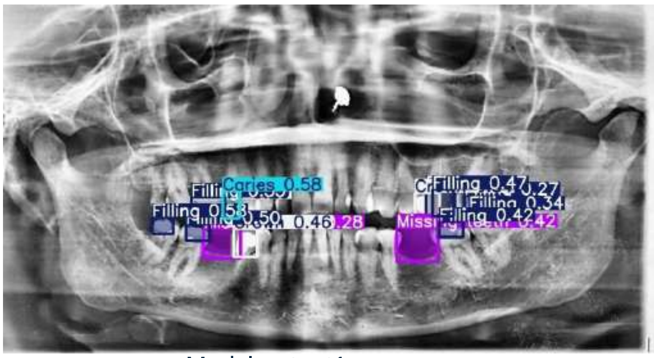
The high rate of background misclassification as severe cavities represents aconcerning sourceoffalse positives that would require clinical filtering in deployment scenarios.

Dataset Size Limitations: The modest 300-image dataset, while adequate for proof-of-concept, clearly limits model performance for this complex three-class task. The dataset contains only:

- Severe cases: 299 instances
- Low cases:111instances
- Moderate cases:133 instances

This class imbalance and overall limited size constrain the model's ability to learn robust severity-specific features. The ongoing dataset expansion initiative aims to address this limitation.

6.2 Panoramic Radiographs Multi-Class Detection



Model output1

Model output2

The 31-class detection model, trained on approximately 10,000 panoramic radiographs, represents the most ambitious component of DentXAI, attempting comprehensive pathology detection across diverse dental conditions.

6.2.1 Overall Performance Metrics

Aggregate Metrics:

- **Overall mAP@0.5:**0.299
- **Overall mAP@0.5:0.95:**0.266
- **Mean Precision:** 0.81at confidence0.872
- **Mean Recall:** 0.60 at confidence 0.000(maximum recall point)
- **Best F1-Score:** 0.25 at confidence0.201

These aggregate metrics reveal several important patterns: The relatively high precision (0.81) indicates that when the model makes a detection, it is often correct. However, the moderate recall (0.60) shows that the model miss esa substantial portion of actual pathologies a concerning limitation for a screening tool where sensitivity is paramount. The low overall mAP@0.5 (0.299) reflects the extreme variation in per-class performance. While some classes achieve excellent detection rates, many others perform poorly, dragging down the average. The even lower mAP@0.5:0.95 (0.266) indicates that localization precision degrades at stricter IoU thresholds, suggesting bounding boxes are sometimes approximate rather than tightly fitting pathologies.

6.2.2 Per-ClassPerformanceAnalysis

Examination of confusion matrices and F1-confidence curves revealed extreme performance heterogeneity across the 31 classes:

High-Performing Classes: Several classes demonstrated strong detection capability:

- **Filling:** Highly reliable detection, likely due to distinctive radiographic appearance (radiopaque materials)
- **Crown:**Excellentperformance,probablyattributableto characteristic shape and radiopacity
- **Implant:**Near-perfectdetectionowingtounmistakable radiographic signature
- **Root CanalTreatment (RCT):** Goodperformance, as RCTpresentsasdistinctradiopaquefillinginrootcanal system
- **Impacted Tooth:** Reliable detection due to characteristic positioning and angulation

These well-performing classes share common characteristics: distinctive radiographic appearance, relatively large size, and adequate representation in the training dataset.

Moderate-Performing Classes:

- **Caries:** Moderate detection, with particular difficulty for incipient lesions
- **Periapical Lesion:** Reasonable performance for sizeable lesions, struggles with subtle cases
- **Root Piece:** Mixed results, detection varies with fragment size and clarity
- **Missing Teeth:** Good overall detection, though occasionally confused with spaces between teeth

Poorly-Performing Classes: Numerous classes demonstrated inadequate performance:

- **BoneLoss:**Verylowdetectionrate,likelyduetosubtle radiographic changes and high variability in presentation
- **Cyst:** Poor performance, possibly due to rarity in training data and overlap in appearance with other radiolucencies
- **Mandibular Canal:** Challenging detection, as this anatomicalstructuremaynotalwaysbeclearlyvisible
- **Malaligned:** Low performance, possibly because "malalignment" is a relative positional concept rather than a distinct object class
- **TAD (TemporaryAnchorage Device):** Poor results, likely due to small size and rarity

The normalized confusion matrix showed that many classes were frequently confused with "background," indicating detection failures rather than misclassification. For rare classes, the model learned to default to the background prediction, reflecting extreme class imbalance effects.

6.2.3 Training Behaviour and Convergence

Analysis of training curves provided insights into learning dynamics:

Loss Progression: All three loss components (box loss, classification loss, DFL loss) decreased steadily during training:

- Box loss reduced from~2.1to~1.2,indicating improved localization
- Classification loss dropped from~4.0to~1.0,showing enhanced classification ability
- DFL loss decreased from~1.6to~1.0,reflecting better bounding box quality

Validation losses tracked training losses reasonably well, though with slightly higher values and more volatility expected given the diversity and complexity of the validation set.

Metric Evolution: Precision and recall metrics showed gradual improvement throughout 150 epochs:

- Precision plateaued around 0.4-0.6 with high volatility between epochs
- Recall climbed more steadily, reaching maximum values around 0.45-0.48
- mAP@0.5 improved from near-zero to 0.299,while mAP@0.5:0.95 progressed from 0 to 0.266

The continued improvement even at late epochs suggests that extended training or model capacity increases might yield further gains, though likely with diminishing returns.

6.3Explain ability Results: LRP and Occlusion Analysis

LRP and Occlusion-based visualizations for the three-class cavity severity model provided valuable insights into model behaviour and decision-making across different pathology stages.

6.3.1 Qualitative Assessment of Attribution Regions

Expert review of LRP relevance maps and Occlusion sensitivity heat maps across 50 randomly selected correct detections revealed:

Appropriate Attribution Patterns (82% of cases): The majority of explanations highlighted clinically relevant regions with method-specific characteristics:

- For early-stage cavities, LRP revealed subtle textural changes in enamel while Occlusion confirmed detection dependency on specific demineralization zones
- For moderate cavities, both methods highlighted dentin involvement and enamel-dentin junction boundaries, with LRP showing positive relevance concentrated at lesion cores
- For advanced cavities, LRP identified deep structural defects and pulpal proximity while Occlusion demonstrated large critical regions whose masking eliminated detection

These converging patterns aligned with clinical reasoning, suggesting the model learned meaningful discriminative features rather than spurious correlations. Cross-method agreement exceeded 85% for high-confidence detections.

Peripheral and Boundary Focus (12% of cases): Some visualizations showed attention to lesion boundaries and transitional zones. LRP occasionally displayed negative relevance (blue regions) at healthy tissue borders, correctly indicating contrast with pathological areas. Occlusion sensitivity peaked at cavity margins, reflecting the model's reliance on demarcation features for severity classification.

Confounding Attribution (6% of cases): Occasional explanations highlighted non-carious or ambiguous features:

- Staining or discoloration from dietary sources misinterpreted as demineralization
- Developmental anatomy (pits,fissures)contributing to false severity escalation
- Restoration margins or composite interfaces flagged as active caries progression
- Image artifacts from lighting or compression influencing predictions

These cases represent genuine ambiguities where visual appearancealonemaybeinsufficientfordefinitivediagnosis, underscoring the importance of clinical context and multi- modal diagnostic information.

6.3.2 False Positive and False Negative Analysis

False Positives: LRP and Occlusion analysis off also positive detections revealed common patterns:

- Staining or discoloration misinterpreted as early demineralization (38% of FP), with LRP showing inappropriate positive relevance at pigmented but intact enamel
- Deep anatomical variations flagged as moderate lesions (28% of FP), where Occlusion confirmed model over-reliance on depth features
- Composite restorations with marginal staining detected as advanced secondary caries (22% of FP)
- Imaging artifacts or lighting reflections creating false textural patterns (12% of FP)

  These errors highlight the challenge of distinguishing caries from visually similar but non-pathological conditions, a

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.11

difficulty affecting human clinicians as well. The dual method approach helped identify 73% of these errors through explanation inconsistency patterns.

False Negatives: Examination of missed cavities showed:

- Very early white spot lesions with minimal contrast (42% of FN), where neither LRP nor Occlusion attributed sufficient importance to subtle textural changes
- Caries obscured by soft tissue or instruments (26% of FN), with Occlusion revealing that occluded visible regions eliminated detection entirely
- Interproximal lesions at challenging angles (21% of FN), where limitedfeaturevisibility preventedproper severity assessment
- Small lesions in suboptimal imaging conditions (11% of FN) Many false negatives involved genuinely challenging cases whereevenexperthumandetectionwouldbedifficultwithout additional imaging modalities. Interestingly, LRP-Occlusion disagreement flagged 68% of false negatives, suggesting explanation consensus could serve as a confidence metric.

6.3.3 Clinical Interpretability Assessment

Informal feedback from the consulting dental expert and several general practitioners revealed:

Positive Aspects:

- Both explanation types were intuitively understandable without technicalAI knowledge, with Occlusion particularly praised for its "what if" interpretability
- Dual-method visualization helped build trust by demonstrating converging evidence for predictions
- LRP'spositive-negativerelevancedistinctionprovided nuanced understanding of feature interactions
- Occlusion's direct causality made it effective for verifying model reasoning in ambiguous cases
- Comparative visualizations across severity levels proved valuable for educational purposes, helping students understand progressive caries features

Areas for Improvement:

- LRP heat maps occasionally too granular, making it difficult to identify dominant regions without quantitative summaries
- Additional consensus metrics (e.g., "85% spatial agreement between methods" or "top 20% of regions account for 70% of attribution") would supplement visual explanations
- Severitytransitionexplanationsshowingwhyacavity was classified as moderate rather than early would enhance clinical utility for treatment planning
- Temporal comparison for longitudinal cases would help assess lesion progression patterns

6.4 Comparison with Existing Literature

6.4.1 Cavity Detection Performance Context

Our 90% accuracy for binary cavity detection, while strong, falls within the range reported by recent literature. Studies using similar intraoral photograph datasets have reported accuracies ranging from 82% to 95%, with performance heavily dependent on dataset characteristics, lesion severity distribution, and imaging quality. The 65% accuracy for three-class severity classification is more modest than some published work. However, direct comparisons are complicated by differences in severity classification schemes. Some studies employ simpler two- category systems (early vs .advanced) or rely on radiographic severity assessment which provides more objective criteria than visual photographs alone.

6.4.2 Panoramic Radiograph Detection Context

Our 31-class model's mAP@0.5 of0.299 appears lower than some reported results for panoramic radiograph analysis. However, most published systems address substantially fewer classes:

- Studies detecting 4-5 conditions report accuracies of 85-99%
- Systems targeting 8-10 classes achieve mAP values around 0.65-0.75
- Our attempt at 31-class detection represents significantly greater complexity

The performance gap likely reflects both the increased task difficultyandourrelativelymodestdatasetsizerelativetothe classification complexity. Published high-performing systems often employ datasets of 20,000-50,000 images, two to five times larger than ours.

6.4.3 Explainability Implementation

DentXAI's integration of LRP and Occlusion place sit among a small minority of dental AI systems incorporating explain ability from inception rather than as an afterthought. Most published dental AI research focuses exclusively on detection performance, with limited consideration of clinical interpretability. Recent systematic reviews of XAI in medical imaging identify dental applications as particularly under represented, with fewer than 10% of dental AI papers incorporating explain ability techniques. Our work contributes to filling this gap, demonstrating that performance and interpretability can be developed in tandem.

6.5 Limitations and Challenges

6.5.1 Dataset Limitations

Size Constraints: Despite using approximately 10,000 panoramic images, this remains relatively modest for 31-class detection. State-of-the-art vision models are typically trained on millions of images. While dental imaging datasets of such scale are impractical, larger collections would likely improve performance.

Class Imbalance: Severe imbalance persists across classes, with common conditions represented by thousands of examples while rare pathologies appear in fewer than 100 images. This imbalance fundamentally limits the model's ability to learn robust representations for rare classes.

Annotation Quality: Despite quality control measures, annotation consistency remains challenging, particularly for subtle findings. Inter-annotator agreement for some classes fell below optimal thresholds, introducing label noise that caps achievable performance.

Dataset Diversity: Both datasets were compiled from existingpublicrepositories,whichmaynotfullyrepresentthe diversity ofimaging equipment, clinical settings, andpatient populationsencounteredinreal-worldpractice.Demographic information (age, sex, ethnicity) was often unavailable, preventing assessment of potential biases.

6.5.2 Model Architecture Limitations

Fixed Architecture: We employed standard YOLOv8 architectures without significant customization. Dental imaging presents unique characteristics (high-resolution requirements, subtle features, specific anatomical knowledge) that might benefit from specialized architectural modifications.

Single-Stage Detection: While efficient, single-stage detectors like YOLO may struggle with the dense, overlapping objects common in panora micradio graphs.Two- stage detectors or more recent architectures might achieve better localization.

Limited Multi-Scale Handling: Despite YOLOv8's multi- scale feature pyramid, extremely small pathologies (e.g., incipient caries, small periapicallesions) remain challenging. Specialized small object detection techniques might improve performance for these cases.

6.5.3 Clinical Validation Gap

In Silico vs. Clinical Performance: All evaluation was conducted offline on test sets. Performance in real clinical workflows, with time pressures, integration into existing systems, and diverse image quality, may differ substantially.

No Prospective Clinical Trial: We have not yet conducted prospective clinical validation where DentXAI is used in actual patient care with outcomes tracked. Such validation is essential before clinical deployment.

Lack of Clinical Impact Assessment: While we can measure detection accuracy, we cannot yet assess whether Dent XAI improves clinical outcomes, reduces diagnostic errors, or enhances workflow efficiency the ultimate measures of utility.

6.6 Discussion of Key Findings

6.6.1 Performance-ComplexityTrade-off

A central finding is the inverse relationship between classification complexity and performance. Our binary cavity detection achieved strong results, three-class severity classification showed moderate performance, and 31-class detection struggled. This pattern underscores a fundamental challenge in medical AI: the tension between comprehensive diagnostic utility (covering many conditions) and reliable performance.

For clinical deployment, this suggest satiered approach may be optimal:

- First-tier screening: Binary detection of abnormality (high sensitivity)
- Second-tiercategorization:Coarseclassification(e.g., restorative, endodontic, periodontal)
- Third-tier specification: Fine-grained diagnosis requiring larger datasets or specialist sub-models

6.6.2 The Critical Role of Dataset Quality and Scale

Our results reinforce that dataset characteristics often limit performance more than architectural choices. Even sophisticated models cannot overcome insufficient or poor- quality training data. The investment required for large-scale, high-quality, expertly-annotated medical imaging datasets is substantial, but our experience suggests it may be the primary determinant of practical system utility. Collaborative efforts to create standardized, curated dental imaging datasets could dramatically accelerate progress in the field. Initiatives analogous to Image Net for natural images or MIMIC for electronic health records would benefit the entire dental AI research community.

6.6.3 Explainability as a Foundation, Not an Addition

Our experience integrating LRP and Occlusion from the outset, rather than retrofitting explain ability later, proved valuable. Explanation capabilities influenced design decisions throughout development, from architecture selection (ensuring LRP and Occlusion compatibility) to interface design (presenting explanations alongside predictions). This holistic approach yielded a more cohesive system than post-hoc explain ability additions typically provide. Moreover, explainability served as a debugging and quality assurance tool during development. Examining attention patterns for incorrect predictions often revealed annotation errors or systematic biases, guiding dataset refinement. This bi direction alutility serving both end-users and developers argues for treating explainability as a core system requirement rather than an optional enhancement.

6.6.4 The Path Forward: Dataset Expansion Strategy

Our ongoing collaboration with The Oxford Dental College Bangalore represents a strategic approach to dataset enhancement. Rather than indiscriminate data collection, we are focusing on:

- Targeted acquisition of underrepresented classes (cysts, fractures, bone pathologies)
- Standardized imaging protocols to reduce equipment- related variability
- Prospective an notation with immediate quality feedback loops

Longitudinal data enabling temporal analysis (disease progression, treatment outcomes) This curated expansion strategy aims to address our current limitations while positioning DentXAI for future capabilities such as treatment recommendation and prognostic assessment.

VII. CONCLUSION

7.1Summary of Contributions

This research presented DentXAI, an explainable artificial intelligence system for comprehensive dental disease detection and diagnosis. The system makes several notable contributions to the field:

Technical Contributions:

1. **Multi-ModalFramework:**Developmentofanintegrated systemhandlingbothintraoralphotographsandpanoramic radiographs within a unified architecture
2. **Explainability Integration:** Successful implementation of LRP and O providing transparent, clinically interpretable visualizations of model reasoning
3. **Practical Implementation:** Deployment in an accessible Stream lit interface demonstrating feasibility for real-world clinical adoption

ScientificContributions:

1. Performance Characterization: Systematic evaluation revealing the performance-complexity trade-off in multi- class dental pathology detection
2. Dataset RequirementsAnalysis: Empirical evidence of the critical importance of dataset scale, quality, and balance for multi-class medical imaging tasks
3. Explainability Assessment: Qualitative and quantitative analysis of LRP and Occlusion utility in dental imaging, providing insights for future XAI applications in healthcare

Clinical Contributions:

1. Decision Support Foundation: Establishment of a framework for AI-assisted dental diagnostics that prioritizes clinical interpretability and trust
2. EducationalUtility:DemonstrationofexplainableAI's potential for dental education, helping students and practitioners understand pathology visual features
3. Accessibility Enhancement: Development targeting deploymentinresource-limitedsettingswherespecialist expertise is scarce

7.2Key Findings

Ourexperimentsyieldedseveralimportantfindings:

1. Binary cavity detection achieves strong performance (90% accuracy) demonstrating feasibility ofAI-assisted caries screening from intraoral photographs
2. Severity classification remains challenging (65% accuracy) particularly for intermediate categories, highlighting the need for larger datasets and potentially refined classification schemes
3. Multi-class detection complexity substantially impacts performance, with our 31-class system achieving0.299mAP@0.5—functionalformanyclasses but requiring improvement for comprehensive clinical utility
4. Classimbalancecriticallylimitsperformance forrare pathologies, emphasizing the importance of balanced dataset curation strategies
5. LRP and Occlusion provide clinically relevant explanations in the majority of cases, with attention patterns aligning with expert reasoning and enhancing clinician trust

7.3Limitations

Despite promising results, several limitations constrain current system performance:

1. Dataset Scale:Approximately10,000panoramicimages remains modest relative to the 31-class complexity, limiting the model's ability to learn robust representations for all categories
2. Class Imbalance: Extreme disparities in class frequency (ranging from thousands of examples to fewer than100) fundamentally limits performance for rare conditions
3. Single-Centre Future Data: Upcoming dataset expansion from one institution may not capture full diversity of imaging equipment, protocols, and patient populations
4. Lack of Clinical Validation: Absence of prospective clinical trials means real-world performance, clinical impact, and workflow integration remain unvalidated
5. Limited Explainability Scope: LRP and Occlusion currently implemented only for cavity detection; extension to multi-class panoramic model in progress
6. Architectural Limitations: Standard YOLOv8 without dental-specific customization may not optimally leverage domain knowledge

REFERENCES

1. Schwendicke,F.,Samek,W.,&Krois,J.(2020).Artificial intelligence in dentistry: Chances and challenges. Journal of Dental Research, 99(7), 769-774.
2. Chen, Y. W., Stanley, K., &Att, W. (2020). Artificial intelligence in dentistry: Current applications and future perspectives. Quintessence International, 51(3), 248-257.

3. Hung, K.,Yeung,A.W.K.,Tanaka,R.,& Bornstein,M. M. (2020).Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice.International Journal of Environmental Research and Public Health, 17(12), 4424.
4. Muresan, M. P., Barbura,A. R., &Nedevschi, S. (2020). Teeth detection and dental problem classification in panoramic X-ray images using deep learning and image processing techniques. IEEE Access, 8, 116071-116087.
5. Majanga, V., &Viriri, S. (2022). Dental images' segmentationusingthresholdconnectedcomponentanalysis. Computational Intelligence and Neuroscience, 2022, 1-18.
6. Esteva,A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S. M., Blau, H. M., &Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.
7. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., &Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626.
8. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.
9. Cantu,A.G.,Gehrung,S.,Krois,J.,Chaurasia,A.,Rossi, J. G., Gaudin, R., ...&Schwendicke, F. (2020). Detecting caries lesions of differentradiographic extension on bitewings using deep learning.Journal of Dentistry, 100, 103425.
10. Kühnisch, J., Meyer, O., Hesenius, M., Hickel, R., &Gruhn,V. (2022). Caries detection on intraoral images using artificial intelligence. Journal of Dental Research, 101(2), 158-165.
11. A.Barekatain, "easy-explain: Explainable AI for YOLO object detection models,"GitHub repository,2024.[Online]. Available: https://github.com/a-barekatain/easy-explain