



University of Essex

School of Mathematics, Statistics
and Actuarial Science

CE901 - MSC PROJECT AND DISSERTATION

Multimodal Dental Diagnostics: Comparative Detection Modeling and LLM-Powered Report Generation with Interactive RAG Assistance

*A dissertation submitted in fulfillment of the requirements for the
degree of Master of Science*

Tun Ye Minn

Registration Number: 2411737

(tq24401@essex.ac.uk)

Supervisor: **Dr Haider Raza**

December 9, 2025
Colchester

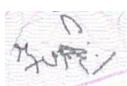
Declaration of Authorship

I, Tun Ye Minn, declare that this thesis, titled "**Multimodal Dental Diagnostics: Comparative Detection Modeling and LLM-Powered Report Generation with Interactive RAG Assistance**", and the work presented within it are my own.

I confirm that:

- This work was conducted wholly during my candidature for the degree of Master of Science at the University of Essex.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- All significant sources of help have been acknowledged.
- Where the thesis is based on work done jointly with others, the contributions of others and the author's own contributions have been clearly delineated.

Signed:



Date: December 9, 2025

Abstract

In dentistry, there have been advanced diagnostic capabilities, but significant gaps remain in automating the interpretation of low-resolution radiographic data and effectively communicating these findings to patients. This thesis presents a comprehensive **Multimodal Dental Diagnostics** framework designed to bridge these gaps by integrating state-of-the-art Computer Vision with Generative AI. Addressing the challenges of severe class imbalance and image noise in dental **Orthopantomograms (OPGs)**, this study compared five object detection architectures. The results showed **YOLOv11m** as the superior model, achieving a mean Average Precision ($mAP_{50:95}$) of **0.7491**, significantly outperforming other two-stage and transformer-based baselines. Moreover, for **Large Language Model (LLM)** pipelines, to obtain a structured detection output from proposed architectures, different post-processing approaches are implemented. For example, a novel **Spatial Relationship Matching** approach was used as a post-processing step to map detected pathologies to the FDI numbering system with high clinical validity. Beyond detection, the research also includes an end-to-end report generation pipeline which utilizes a **Grounding Caption** approach to translate the predicted visual data into structured clinical outputs, which were then translated into professional medical PDF reports by the **Qwen LLM** model using **prompt engineering**. Evaluation by expert dentists confirmed a high degree of completeness (3.26/5.0) and structural accuracy. Finally, to enable patient understanding and support communication, an interactive **Retrieval-Augmented Generation (RAG)** chatbot pipeline was deployed using the **Qwen** model with specialized **prompt engineering**. The following system demonstrated a completeness score of (3.67/5.0) with good safety awareness, successfully answering patient queries while refusing out-of-distribution topics. The study also concludes that while challenges like the detection of micro-pathologies such as (caries, periapical radiolucency and calculus) remain, the proposed multimodal architecture successfully establishes a scalable, clinically grounded framework for automated dental diagnostics and patient education. The deployed system and the codes can be accessed via <https://huggingface.co/spaces/tym24/AI-the-Dentist> and <https://tinyurl.com/mwzwnxkf> respectively.

Acknowledgement

I would like to express my deep appreciation to my supervisor, **Dr. Haider Raza**, for his expert guidance and unwavering support throughout my research. His insights and feedback have been pivotal to the successful completion of this thesis.

In addition, I would like to recognize the faculty and staff at the University of Essex, especially those in the Department of Computer Science and Electronics Engineering, for their dedicated assistance and commitment to academic excellence, which has greatly contributed to my educational experience.

I would like to also give my thanks to the dentists who gave me the required data and evaluation feedback while working on this thesis. Their help and support is deeply appreciated.

Finally, I acknowledge the authors and researchers cited in this thesis, whose work has been crucial in shaping my research and understanding of the field.

Contents

1	Introduction	8
1.1	Background: The Digital Dentistry	8
1.2	Problem Statement	9
1.3	Dissertation Objectives	10
1.4	Research Questions	10
1.5	Contributions and Dissertation Structure	11
2	Literature Review	13
2.1	Deep Learning in Medical Imaging	13
2.1.1	Comparative Architectures in Object Detection	14
2.1.2	Ensemble Approach (Thresholded Soft Voting)	16
2.2	Natural Language Processing in Healthcare	16
2.3	Prompt Engineered RAG Dental Question Answering	17
2.4	Existing Gaps in Multimodal Dental Diagnostics	18
3	Data Acquisition and Feature Engineering	19
3.1	Dataset Overview	19
3.1.1	Source and Characteristics	19
3.1.2	Class Distribution and Imbalance	20
3.2	Feature Engineering	23
3.2.1	Class Consolidation Strategy	23
3.2.2	Data Cleaning and Final Dataset	23
4	Methodology I: Dental Disease Detection	26
4.1	Model Architectures and Configuration	26
4.1.1	YOLOv11m (Ultralytics)	27

4.1.2	Faster R-CNN (ResNet-50 + FPN)	27
4.1.3	RetinaNet (ResNet-50 + FPN)	28
4.1.4	Mask R-CNN (Detectron2)	28
4.1.5	DINO with DETR (DEtection TRansformer) using MMDetection framework	29
4.2	Pre-processing and Training Pipeline	29
4.2.1	Pre-processing and Data Conversion	29
4.2.2	Hyper-parameters, Learning Rate Scheduling, and Early Stopping	30
4.3	Ensemble Strategy and Evaluation	30
4.4	Post-Processing Algorithms	34
4.4.1	Spatial Relationship Matching (Mapping Pathologies to FDI Teeth)	34
4.4.2	Nearest Tooth Assignment and FDI Duplication Removal	35
4.4.3	Domain Knowledge Rules and Structured Output Generation	35
5	Methodology II: Generative AI and Reporting Pipeline	36
5.1	Grounding Caption Generation	36
5.2	Medical Report Generation	37
5.3	Interactive RAG Chatbot	42
5.3.1	System Prompting and Role Definition	43
5.3.2	Context Injection and RAG Role	44
6	System Deployment and User Interface	46
6.1	High-Level Architecture	46
6.2	Web and Mobile API Integration	47
6.2.1	Diagnostic Analysis API	48
6.2.2	Interactive Chat API	48
6.2.3	User Feedback API using Firebase	52
6.3	Security and Hardware/Software Requirements	54
7	Results and Analysis	55
7.1	Detection Performance Metrics	55
7.1.1	Comparative Analysis of Detection Architectures	55
7.1.2	Performance on Pathological and Restorative Classes	56
7.1.3	Spatial Localization Accuracy (FDI Teeth)	57

7.1.4	Case Studies: Visual Analysis of Predicted Outputs	59
7.2	Medical Report Generation Performance Metrics	61
7.2.1	Quantitative Analysis of Clinical Metrics	61
7.2.2	Case Study: Confidences with Generated Report	63
7.3	Performance of RAG Chatbot Question Answering	65
7.3.1	Quantitative Analysis of Chatbot Metrics	65
7.3.2	ChatBot Correlation Analysis	66
7.3.3	Performance by Question Type	67
8	Discussion	68
8.1	Detection Pipeline	68
8.2	Medical Report Pipeline	69
8.3	RAG Chat-Bot Pipeline	70
8.4	Limitations	71
8.4.1	Technical Limitations	71
8.4.2	Clinical Limitations	71
9	Conclusion and Future Work	72
9.1	Research Summary	72
9.2	Future Research Directions	73
9.3	Final Remarks	74
	End Device Download links	79
	Research Codebase & Database	79
A	Supplemental Materials	80
A.1	Class-Specific Confidence Thresholds	80
A.2	Sample Detection Outputs	81
A.3	Medical Reports and Grounding Captions	84
A.4	Chatbot Evaluation and Interface	87
B	Website and Application Interfaces	89
C	Firebase Storage and Firestore	93

Introduction

Every day, dental professionals must perform the extremely difficult task of reviewing hundreds of X-ray images to discover pathologies ranging from visible to microscopic. This dissertation proposes a technological bridge between the visual complexity of radiology and the communicative needs of patient care.

Although Artificial Intelligence (AI) has made significant improvements in medical imaging, there is still a gap. Current systems may detect a cavity, but they cannot explain it, document it, or discuss it with a patient. This research seeks to change that limitation by building a system that not only sees like a radiologist, but also communicates like a dentist with an interactive interface for medical report generation and communication with a chat-bot using four unique pipelines. The four pipelines are detection, report generation, communicable chat-bot and deployment pipelines.

1.1 Background: The Digital Dentistry

Dentistry has historically been a visual discipline. The transition from analog film to digital Orthopantomograms (OPGs) revolutionized how data is stored, but it did not fundamentally change how it is interpreted; the burden of diagnosis still rests entirely on the clinician's ability to see.

In recent years, the field has seen the rise of Computer-Aided Diagnosis (CAD). Early iterations were simple, the adoption of Deep Learning brought Convolutional Neural Networks (CNNs) to the front line, offering powerful tools for feature extraction [1].

However, the majority of existing research focuses on isolated tasks such as segmenting a jaw without integrating these findings into a patient's report. There is a lack of research addressing the reality of clinical environments, where images are often low-resolution, averaged, or improperly shaped.

1.2 Problem Statement

This research is motivated by three critical challenges currently facing the dental diagnostic workflow:

1. The Challenge of Real-World Data

In an ideal research setting, medical images are high-resolution and perfectly annotated. In the real world, this is rarely the case. Dental X-rays often suffer from low resolution, motion blur, and artifacts. Furthermore, dental pathologies show extreme class imbalance; common issues like caries appear frequently, while specific anomalies like root stumps or impacted teeth are rarer. Standard detection models often struggle to maintain accuracy across this uneven distribution.

2. The Problem in Reporting

Even when a diagnosis is accurate, documenting it is a manual, stressful task. This manual entry is prone to transcription errors and inconsistency. There is currently no state-of-the-art pipeline that automatically converts the visual coordinates of a detected pathology into a structured, clinically valid medical report [2].

3. The Patient Knowledge Gap

Perhaps the most overlooked aspect is the patient's experience. A standard medical report is written by doctors, for doctors. Patients are often left with a PDF they cannot understand, leading to anxiety or poor treatment compliance. There is a need for an interactive system, an AI assistant that can understand these technical reports and answer patient questions in simple, layman language, leveraging techniques such as Retrieval-Augmented Generation (RAG) [3].

1.3 Dissertation Objectives

To address these challenges, this project aims to develop an end-to-end multimodal framework. The specific objectives are:

1. **Unique Detection models in Challenging Conditions:** To implement and compare five advanced detection architectures (including YOLOv11m and Detectron2) specifically optimized for low-resolution and class-imbalanced OPG images.
2. **Spatial Relationship Matching:** To engineer a post-processing approach that understands clinical logic, accurately mapping detected issues to the correct FDI tooth number, even when teeth are missing or displaced.
3. **Report Generation:** To create a pipeline that translates visual bounding boxes into **Grounding Captions**, which are then utilized by a Large Language Model (LLM) to generate professional medical PDF reports.
4. **Interactive Patient Support:** To deploy a Retrieval-Augmented Generation (RAG) chat-bot that acts as a virtual dental assistant, capable of answering patient queries based on their specific medical report.

1.4 Research Questions

This thesis is guided by the following five research questions, which investigate the usefulness of both the vision and language components of the proposed system:

- **RQ1:** Can the proposed models accurately predict dental pathologies for a low-resolution and class-imbalanced dental X-ray dataset?
- **RQ2:** What is the impact of dataset quality and annotation consistency on the detection models' performance?
- **RQ3:** Does the prompt-engineered LLM medical report pipeline utilize the structured grounding captions effectively to generate medical reports that meet clinical standards for correctness and completeness?

- **RQ4:** Can the LLM prompt-engineered chat-bot able to answer user queries professionally like a dentist while taking the grounding caption and medical report as input?
- **RQ5:** Do the chat-bot responses meet clinical standards for clarity and relevance?

1.5 Contributions and Dissertation Structure

This work contributes to the field of medical AI by moving beyond simple detection. Key contributions include:

1. A **Comparative Analysis** of modern object detection models specifically on low-quality dental data.
2. A novel **Feature Engineering and Post-Processing approach** that solves the complexity of mapping pathologies to the FDI numbering system.
3. A **Multimodal Workflow** that successfully bridges Computer Vision (detection) and NLP (reporting and assistance), demonstrating how **Grounding Captions** can serve as a universal language between images and LLMs.
4. A fully deployed **Website and Mobile applications** using RAG, proving the usefulness of interactive AI in a clinical setting.

The remainder of this report is structured as follows: **Chapter 2** reviews the literature on deep learning in dentistry and LLMs in healthcare. **Chapter 3** details the data acquisition and feature engineering process. **Chapter 4** presents the detection methodology, including model architectures and the ensemble strategy. **Chapter 5** describes the generative AI pipeline for reporting and the chat-bot. **Chapter 6** outlines the system implementation with UI and database. **Chapters 7, 8, and 9** provide the results, discussion, and conclusion, respectively. Additionally, the overall high-level architecture of the entire system can be seen from the figure [1.1](#).

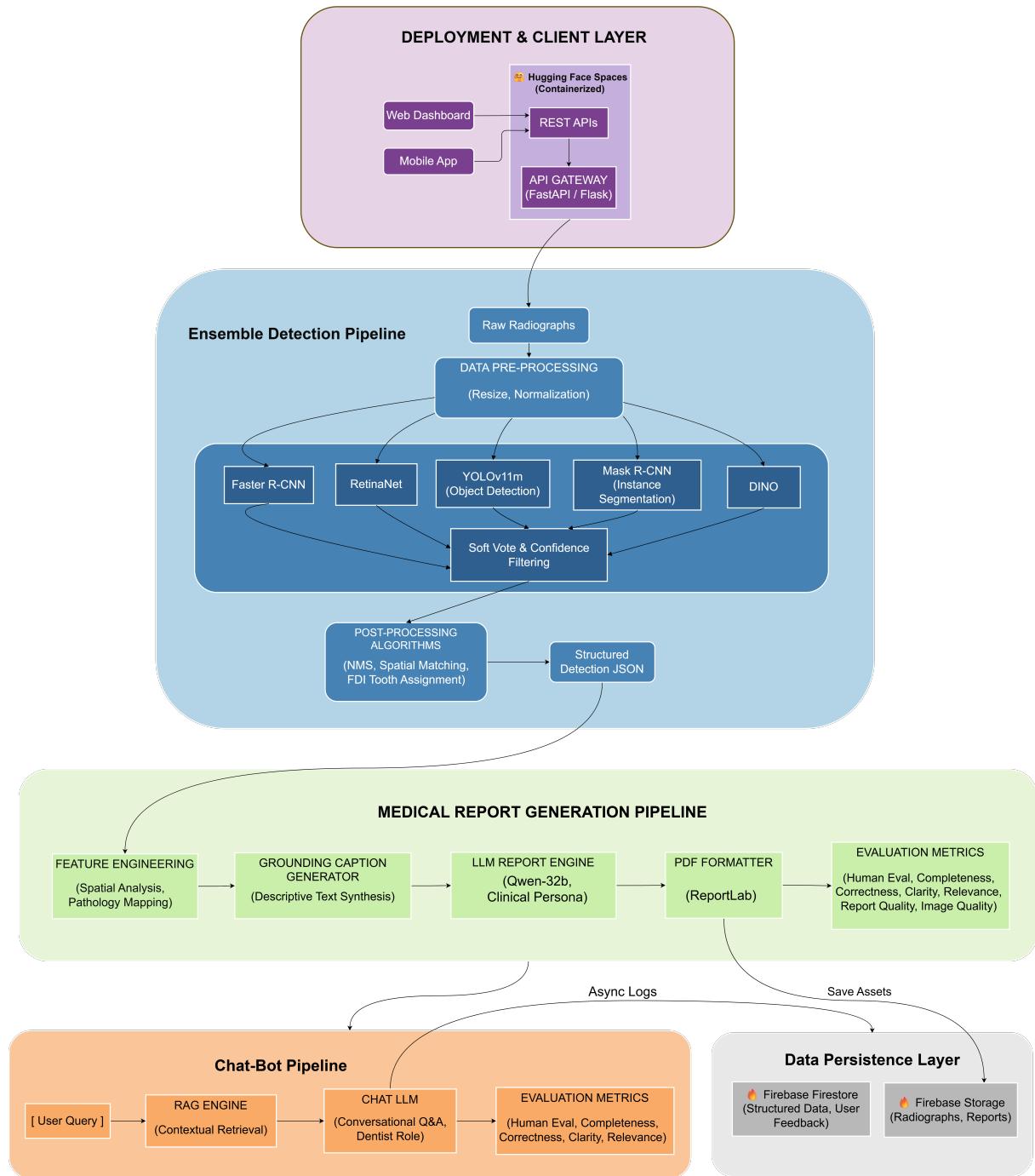


Figure 1.1: High-Level architecture of the system.

Literature Review

This chapter provides the technical context for the multimodal diagnostic system by examining the existing literature across three main domains of the report. These include Deep Learning in dental image analysis, the architectural framework of object detection, and the emerging role of Generative AI, specifically the **Qwen** model, in structured medical reporting and conversational support.

2.1 Deep Learning in Medical Imaging

The application of Deep Learning (DL) to medical diagnostics has shown great capabilities surpassing human consistency in several specialized tasks. In dental radiography, the primary objective is to accurately localize and classify multiple pathologies and anatomical structures simultaneously from complex radiographic data [4]. This involves recognizing specific clinical disorders (e.g., caries) as well as mapping them to the correct anatomical context of FDI teeth numbering. [5, 6].

Early AI applications in dentistry were mainly focused on simple image classification (e.g., **healthy** vs. **diseased**) or pixel level segmentation of some anatomical parts [7]. However, recent works in this field show a paradigm shift towards a instance level object detection which is required for accurately detecting various pathologies available in a single Orthopantomogram (OPG) image [8, 9]. While models like U-Net often achieve high accuracy on curated datasets, they frequently fail when applied to low resolution OPG images which are mostly common in resource-constrained settings

[10]. Furthermore, most existing systems operate in isolation, meaning they may detect a lesion but fail to resolve its relationship to the correct tooth number which is an important requirement for clinical understanding. Recent work by Ma, Tian and Li [11], attempted to address this by combining CNNs and Transformers for segmentation, but their approach focuses on morphological structure rather than the explicit pathology to tooth mapping required for automated reporting. Moreover, recent influential work such as the **MMOral** benchmark and **OralGPT** model by Hao et al. [12], has successfully demonstrated the potential of Vision-Language Models (VLMs) for dental diagnosis but the effect of detection performance on general models such as YOLO, Mask-RCNN, etc have not been evaluated yet.

2.1.1 Comparative Architectures in Object Detection

The performance of object detection models are shaped by their architecture and backbone models. To test these, in this study, five distinct pretrained models across three dominant categories are used to identify which is the best solution for identifying dental pathologies mainly in class imbalanced dataset.

One-Stage Detectors

One-stage detectors prioritize inference speed by performing object localization and classification in a single one time execution. Two one stage detectors such as YOLO and RetinaNet are applied for following facts.

- **YOLO:** YOLO family is huge and each specific version is good for different use cases. Although, earlier versions like YOLOv8 have been widely adopted for dental lesion detection [13], **YOLOv11m** [14] is used to test it's performance in detecting dental pathologies on imbalance dataset. Moreover, recent benchmarks suggest that YOLOv11 offers superior feature extraction capabilities for small objects compared to v8.
- **Addressing Imbalance:** Even though YOLO is powerful, it struggles with imbalance data. For this reason, **RetinaNet** [15] is included for its use of the **Focal Loss function**. Despite the speed of YOLO, standard cross-entropy loss sometimes struggles when the background (healthy tissue) outweighs the foreground

(pathology). Focal Loss solves this by down weighting easy examples, a strategy that has proven valuable in high-imbalance medical scenarios [16].

Two-Stage Detectors

Two-stage methods focus mainly on the classification part for resulting higher localization accuracy at the cost of speed. Even though the approach trains slowly compared to others, they tend to give more precise detection. Models such as Faster R-CNN and Mask R-CNN are applied to test the effect of these powerful detectors.

- **Faster R-CNN:** The Region Proposal Network (RPN) introduced by S. Ren, K. He, R. Girshick, and J. Sun [17] allows the model to focus on potential detection areas before classification. This is great for dental X-rays where lesions like periapical radiolucency are usually a change in texture rather than distinct objects.
- **Mask R-CNN:** Extending Faster R-CNN, another powerful R-CNN family, **Mask R-CNN** was proposed by K. He and Gkioxari [18] which adds a parallel branch for pixel-level segmentation. Additionally, recent study [19] has successfully used Mask R-CNN for detecting treatments (restoration, implant, filling, etc) to showcase that precise contouring of a filling provides more clinical value than a simple bounding box.

Transformer-Based Detectors

Another powerful detector is the Vision Transformers (ViTs). Unlike CNN, this approach uses self-attention mechanisms to capture the global context. **DINO** (DETR with Improved DeNoising Anchor Boxes) [20] is utilized to test this transformer based detection. They are powerful compared to traditional CNN approaches because of the following reasons.

- **Contextual Advantages:** In dental imaging, the presence of a tooth is often conditionally dependent on its neighbors (the dental arch). Theoretically, transformers are better suited to learn these global dependencies than CNNs.
- **Recent Adaptations:** A study, "DAX" (DINO Adapted to X-ray) introduced by J. Scheuplein [21] demonstrated that Transformer-based foundation models can

outperform CNNs in detection tasks. Therefore, this study tests whether these advantages benefit the detection of heterogeneous dental pathologies.

2.1.2 Ensemble Approach (Thresholded Soft Voting)

Dental pathologies can be complex and training a single model to detect all the classes can be difficult. Moreover, single models can suffer from specific failure modes (e.g., YOLO model missing small objects, or R-CNNs being too slow). By combining strong features from each family can result in powerful detection system to support complex dental pathologies and FDI teeth classes.

- **Current State of the Art:** A recent study by Hernández [13] proposed the "ED-ICA" architecture, which combines YOLOv8 detection with an ensemble of fine-grained classifiers to improve accuracy in uncontrolled environments. Similarly, M. Gamal, A. Essam, and A. Atia [22] demonstrated that a bagging ensemble of Swin Transformers and MobileNetV2 achieved 95.7% precision in dental disease classification.
- **Distinction:** Unlike the mentioned approaches which often focus on image-level classification, this study implements a **Thresholded Soft Voting Ensemble** by combining different models such as YOLOv11m and DINO. Because of this, one can merge the high recall of one-stage detectors with the precise localization of two-stage detectors to create a powerful system that is mostly correct. The system can later be passed through threshold filtering and then used for the generation of medical text (report) and for helping with conversation agents.

2.2 Natural Language Processing in Healthcare

Nowadays in medical domain, NLP is being widely used for helping patients and in assisting doctors with their decision making and treatment planning. It is used in different aspects but it is most obvious in automation of reports and supporting via conversation agents.

Automated Report Generation via Grounding Captions

Automated radiology reporting has evolved from template-filling to abstractive summarization using sequence-to-sequence models [23]. However, **hallucinations** which generate potential but incorrect medical facts remain a critical safety barrier.

- **Grounded Reporting:** Recent work on "MAIRA-2" emphasizes *grounded* report generation, where every generated sentence must be linked to a specific image region to verify factuality [24].
- **The Grounding Caption Strategy:** Building on this, this study employs a structured **Grounding Caption** approach. Rather than feeding raw image embeddings directly to the LLM (which often leads to hallucinations), visual findings are first translated into a structured text intermediate (JSON data to Descriptive Text). This aligns with findings by K. Jeblick [25], who proved that explicit data structuring reduces errors in LLM-generated medical summaries.

LLM Prompt Engineered Medical Report Generation

Large Language Models are very powerful in medical reasoning tasks. By utilizing the LLM, one can suggest state of the art medical recommendations and treatment planning with caution. Among the different powerful LLMs, **Qwen** [26] model is used for medical report generation [27] because of the resources and tokens availability. Recent benchmarks have shown that Qwen-2.5 outperforms other open-source models (and rivals GPT-4) in specialized medical licensing examinations, particularly in nursing and ophthalmology contexts [28, 29]. For these reasons, **Qwen** model is chosen over Llama-2 and GPT models for the high precision dental reporting work. Moreover, LLM models relies heavily on detailed prompt engineering. As detailed in the review by Y.Gu et al. [30], **Chain-of-Thought** and role-playing prompts are vital for aligning LLMs to clinical protocols.

2.3 Prompt Engineered RAG Dental Question Answering

General-purpose LLMs can hallucinate a lot if applied directly to answer questions. This mainly happens because it often divert it's attention from the important information and

can randomly give unrelated answers. One of the easiest strategies is to use **Retrieval-Augmented Generation (RAG)** to steer it's attention back. RAG solves this by retrieving relevant information or documents to condition the generation [31]. The **Dental Loop Chat-bot** by M. S. H. Arian [32] represented an early attempt to use RAG for dental guidance. However, it relied primarily on general clinical guidelines as the knowledge base. Therefore, this study advances this concept by using patient's specific documents as attention backbone for the RAG. Instead of retrieving general textbooks, RAG is designed to read the *specific* Grounding Caption and Medical Report generated for the user. This approach ensures that answers are not only medically correct but also well aligned with the predication outputs from the model(s).

2.4 Existing Gaps in Multimodal Dental Diagnostics

While all the proposed literature showed promising developments, there still presents some gaps that this study addresses.

1. **Data Quality and Imbalance:** Benchmark like MMOral [12] used curated, high-quality images. But, this work focuses on **low-resolution and severe class-imbalance** dataset which is more typical in real world clinical environments. As Demir K and Sokmen [10] pointed out, robustness to such unpredictability is the primary issue for translational AI.
2. **Absence of Clinical Spatial Reasoning:** Most models rely on implicit learning for object relationships. However, clinical dentistry requires precise, rule-based logic (e.g., *If Tooth 48 is missing, check 18 for super-eruption*). This study offers a **Spatial Relationship Matching** post-processing approach that uses dental rule based logic which is often skipped in end-to-end deep learning systems [33].
3. **Prompt Engineered Qwen Integration with RAG:** Unlike other models, which simply focus on detection, this study bring together detection and the state to the art **Qwen** model [28] with a strict **prompt engineering and RAG** to provide clinically acceptable medical reports and conversational assistant [2].

In summary, this research bridges the gap between high-performance detection with practical patient-centric communication and report generation for digital dentistry.

Data Acquisition and Feature Engineering

In this chapter, information about the raw radiographic data, the COCO (Common Objects in Context) annotation standards, and the applied feature engineering approach to solve the challenges such as low resolution and severe class imbalance will be discussed.

3.1 Dataset Overview

3.1.1 Source and Characteristics

The Orthopantomogram (OPG) image data utilized in this work was acquired from a dental hospital in India. Two certified dental practitioners annotated ground truths for each and every dental x-ray images used in this work. No external data sources were used for developing this work. The annotated dataset was then later delivered in five batches to develop multi-class x-ray detection, medical report generation, and chatbot pipelines.

The dataset is organized into three folders namely training (`train`), validation (`valid`), and testing (`test`). Each subset contains OPG X-ray images, alongside a COCO annotated JSON file. This JSON structure contains information for every instance, including filename, bounding box (`bbox`) locations, pixel-level segmentation coordinates, and class identifiers. Moreover, each x-ray image is fixed to a resolution of approximately 1420×712 pixels.

This clinical dataset has main issues with quality and full class annotations. Many images are of low resolution and also have contrast issues because of the patient positioning or equipment limitations during the data collection stage. Furthermore, initial annotations in provided JSON files have inconsistencies with anatomical classes, such as certain FDI tooth numbers not being present. Therefore, these issues introduce imbalance into the actual ground truths and inconvenient for model training.

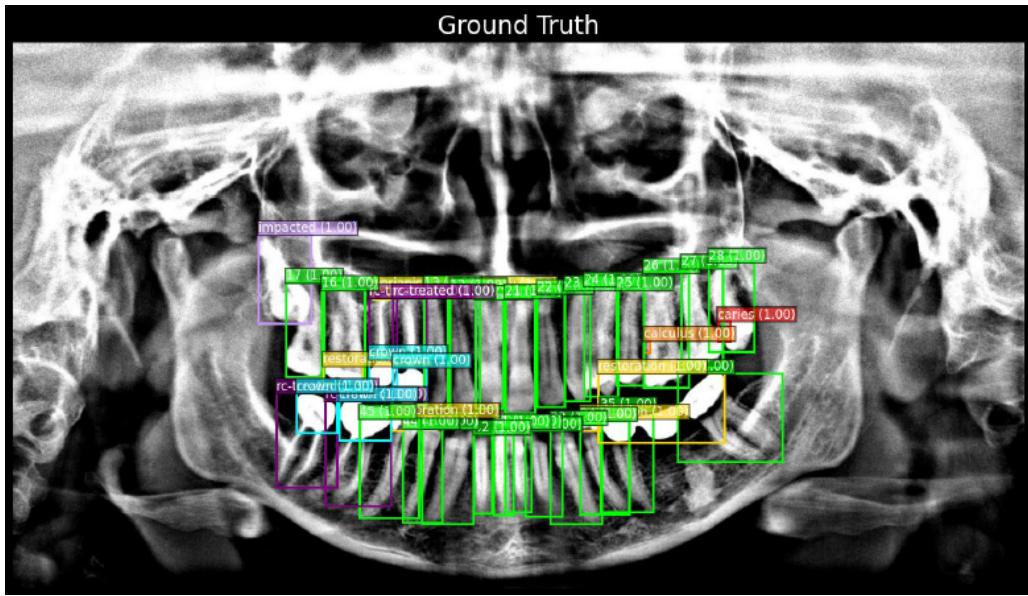


Figure 3.1: Sample Original OPG X-ray Image with Ground Truth Annotations.

On the bright side, most of the images from the dataset are of workable image qualities and well-annotated dental classes. For example, figure 3.1 illustrates a sample OPG with its original ground truth bounding box and class labels (e.g., caries, rc-treated, restoration, and impacted teeth). The image is also in good quality condition for the models (YOLO, Faster-RCNN, RetinaNet, etc) to extract patterns and segment the classes well during the detection pipeline.

3.1.2 Class Distribution and Imbalance

Before the pre-processing stage, the original dataset contains 62 unique classes with 32 permanent FDI (Fédération Dentaire Internationale) tooth notations and 30 distinct pathological conditions, restorations, and anatomical anomalies. However, to effectively train models that can predict these classes well, an essential class distribution analysis is conducted. The provided distribution charts (Figures 3.2, 3.3, and

3.4) show a severe case of the long-tail problem, where most classes possess an extremely low instance count. In the figures, the instance counts range from common classes (e.g., caries, calculus) with around 2,500 instances to rare classes (e.g., grossly-decayed, wire splint) with single-digit counts.

Moreover, this similar imbalance can be found across the splits, validation (Figure 3.3) and test (Figure 3.4) as well. In the training set (Figure 3.2), classes such as missing and caries dominate, with over 3,700 instances each, while classes like vascular clips, wire splint, and grossly-decayed exists with less than 10 instances. This low representation is insufficient for a deep learning models to learn meaningful features and can lead to high variance and poor recall on rare classes (RQ1). Due to these reasons, feature engineering 3.2 is applied to make the class distribution even and to preserve good quality classes.

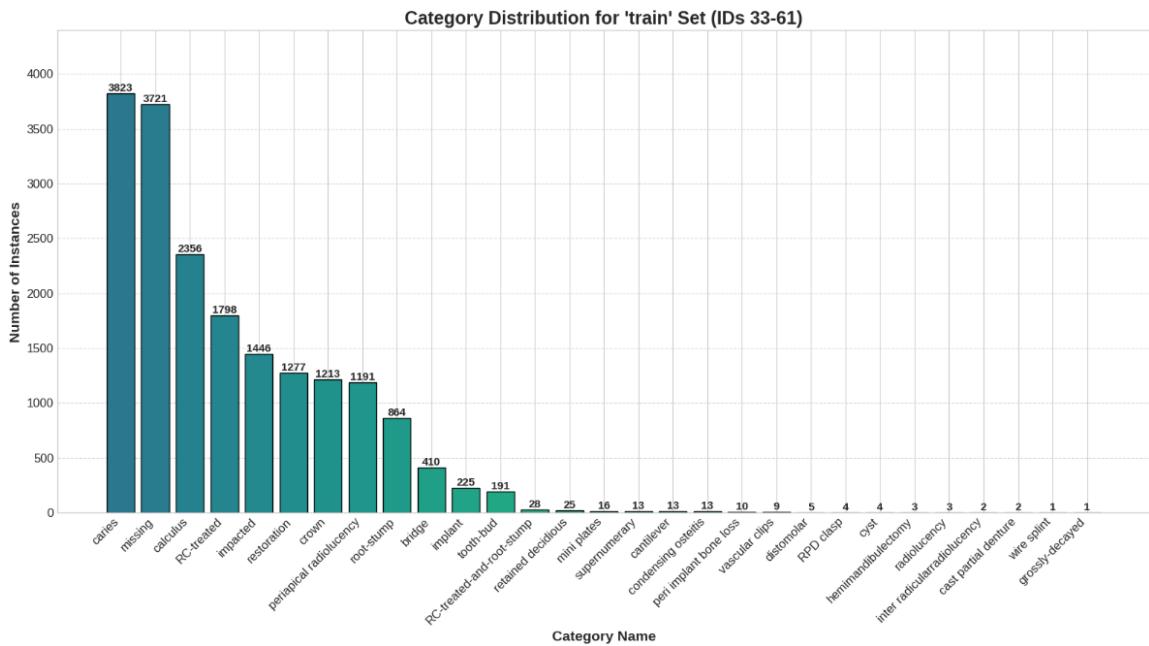


Figure 3.2: Category Distribution for the Original Training Set.

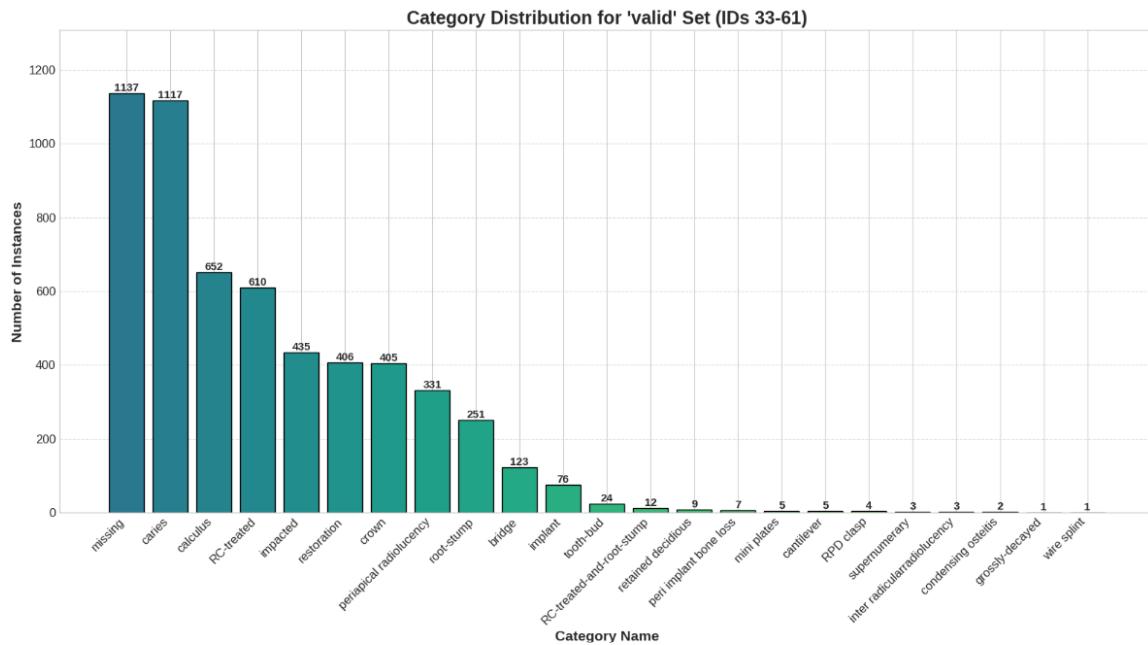


Figure 3.3: Category Distribution for the Original Validation Set (IDs 33–61).

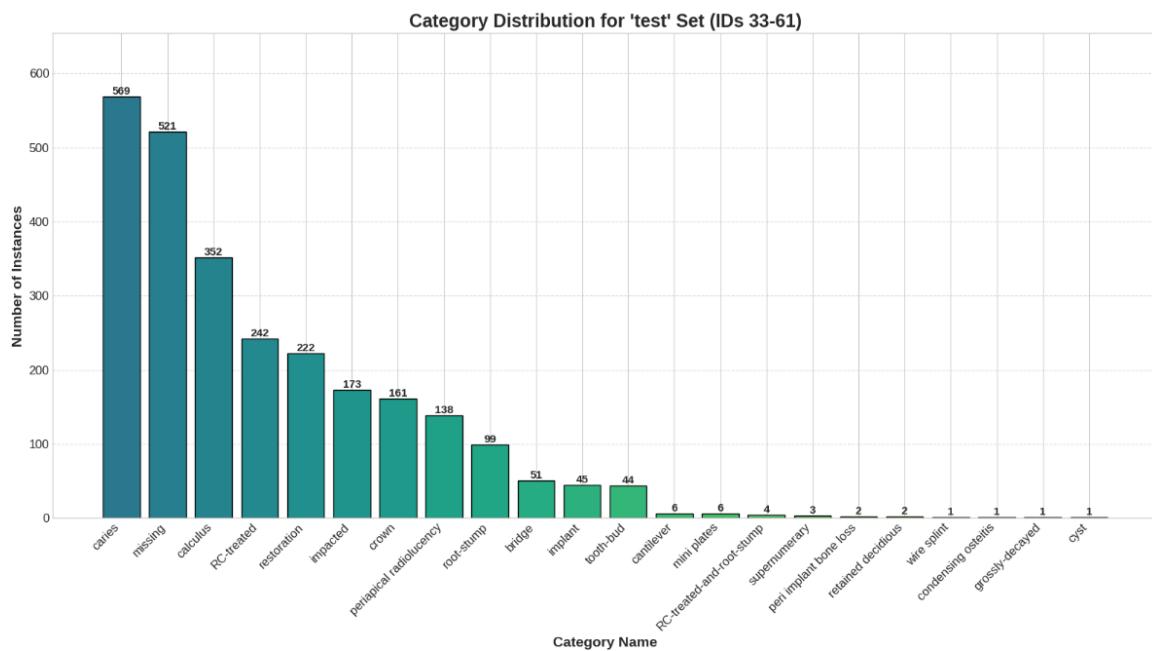


Figure 3.4: Category Distribution for the Original Test Set (IDs 33–61).

3.2 Feature Engineering

To produce a balanced and clinically annotated dataset, a well-structured feature engineering approach was implemented. This process involved two primary strategies such as merging semantically similar classes and the removal of low represented, insignificant or clinically non-essential classes. The final dataset contains 41 classes with FDI tooth numbering and dental pathologies classes.

3.2.1 Class Consolidation Strategy

Classes that represented visually or clinically similar conditions were merged to increase the distribution counts for the resulting combined class in order to stabilize during model training. This merging process, detailed in Table 3.1, focused primarily on carious lesions and restorative intervention classes. The combination of caries and grossly-decayed was performed as the visual distinction between these two in low-resolution OPGs is ambiguous. Moreover the grouping of restoration, bridge, and cantilever are done as well to simplify detecting the general category of "restoration."

3.2.2 Data Cleaning and Final Dataset

Since, the dataset is in its early stages, dentists need more time to produce a fully well-annotated dataset on all the available classes. Therefore, some classes lack the annotations and they became rare classes. Due to these reasons, these classes (less than 50 instances across the entire dataset) were removed. Moreover, these classes have minimal impact on the core diagnostic task and removing these will allow the models to focus their capacity on the most reliable features.

More information about the feature engineering process is summarized in a consolidated table (Table 3.1). The final dataset has the complete set of 32 FDI tooth numbers and 9 distinct pathological or treatment-related classes. Moreover, data across all the 3 folders mentioned in data source section 3.1.1 is combined altogether with respective COCO annotation files. Then, stratification method is used to split the data into training (`train`) 70 percent, validation (`valid`) 15 percent, and testing (`test`) 15 percent with new respective annotation files. The final stratification of the dataset ensured proportional representation across the splits.

After the feature engineering and stratification steps, the final dataset consists of 2,616 OPG images and 80,972 total annotations. The stratified splitting, detailed in Table 3.1, maintains class proportionality across the training, validation, and testing partitions. This structured data cleaning and feature engineering step is vital for ensuring that the training of the comparative detection models (Chapter 4) is not compromised by severe noise or lack of learning examples (RQ2).

Table 3.1: Summary of Feature Engineering Steps and Final Dataset Classes

Table 1: Data Merging (Original Class → Target Class)	
Caries Group	caries and grossly-decayed → caries
Restoration Group	restoration, bridge and cantilever → restoration
Table 2: Removed Classes (Due to Class Imbalance)	
Removed Classes	joint crowns, wire splint, vascular clips, supernumerary, retained deciduous, radiolucency, peri implant bone loss, missing, mini plates, inter radicular radiolucency, cyst, hemimandibulectomy, distomolar, condensing osteitis, cast partial denture, rpd clasp, and rc-treated-and-root-stump
Table 3: Final 41 Classes	
FDI Teeth	11, 12, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 36, 37, 38, 41, 42, 43, 44, 45, 46, 47, 48 (32 classes)
Pathologies	calculus, caries, crown, impacted, implant, periapical radiolucency, rc-treated, restoration, root-stump (9 classes)
Table 4: Final Stratified Dataset Splitting	
Dataset Splits	Train: 1,839 images (56,363 annotations) Valid: 398 images (12,404 annotations) Test: 379 images (12,205 annotations)

Methodology I: Dental Disease Detection

In this chapter, in-depth details of the Computer Vision framework will be discussed. Five state-of-the-art object detection architectures with a structured training pipeline to handle low-resolution and class-imbalanced data are implemented. Finally, a series of post-processing steps are also implemented to have a well-formatted and correct JSON predictions for medical pipeline, which will be discussed in next chapter [5](#).

4.1 Model Architectures and Configuration

The model architecture stage is very important for the entire dissertation work because the outputs from this stage are used by all the remaining pipelines such as the medical report and chatbot pipelines. Additionally, it is vital to have good predictions to be passed to the later steps to create clinically correct medical suggestions and recommendations. Therefore, five unique object detection models were implemented. These architectures were selected to represent various Computer Vision detectors such as one-stage detectors, two-stage detectors, and transformer-based models. Each model offers unique advantages for handling the specific constraints of dental radiography, such as the need for precise localization of very small caries and the handling of different class imbalances.

4.1.1 YOLOv11m (Ultralytics)

The *You Only Look Once* (YOLO) architecture is selected as one-stage detection model. To be precise, the v11m (medium) variant was selected to balance the computational efficiency with feature extraction. In real world, the clinical medical environments require rapid inference performance. In this case, the YOLO's single-pass architecture is best choice for near real-time processing and also for deployment in web-based tools where latency can be a concern. Advantages of using YOLO is that it has extremely fast inference speeds compared to other two-stage detectors. Moreover, it sees the entire image at once, reducing false positives in background areas (e.g., mistaking bone trabeculae for pathologies). However, compared to other models YOLO struggles with small, clustered objects (e.g., caries and calculus), although the 'm' model size generalize this with a deeper backbone. Also, bounding boxes can be less correct than other region-based methods for irregularly shaped lesions.

4.1.2 Faster R-CNN (ResNet-50 + FPN)

Faster R-CNN is one of the two-stage detectors. This model first generates *Region Proposals Network* (RPN) and then classifies them. This model is mounted with a ResNet-50 backbone combined with a *Feature Pyramid Network* (FPN) because dental pathologies like early-stage caries are often subtle and small to be detected. The RPN uses feature maps (which is the output of an FPN) to propose potential bounding boxes where objects might be located for the model. This network performs an object/non-object classification and bounding box regression using a sliding window approach on the feature maps. The FPN then allows the model to focus on candidate areas before the actual classification to improve recall for these small features. To be precise, it works by combining high-resolution detailed features with low-resolution features which are semantically strong. This approach allows the model to accurately identify both small and large objects efficiently. Moreover, generally, the two-stage model gives higher localization accuracy (IoU) for small objects and the FPN effectively detects objects at different scales. Because of this, the model can identify both massive implants and tiny root fragments simultaneously. However, this model can be significantly slower than YOLO because of the two-stage processing pipeline. Also, it became harder to

generalize and requires more memory during training.

4.1.3 RetinaNet (ResNet-50 + FPN)

RetinaNet is a one-stage detector that solves the primary weakness of single-stage models which is the overwhelming presence of background vs. foreground classes. It introduces the **Focal Loss** function. As mentioned in section [3.1.2](#) feature engineering section, the dataset contains severe class imbalances (e.g., thousands of 'caries' vs. fewer than 50 'root stumps'). The focal loss function down-weights easy examples which are background and focuses training on hard negatives examples to directly address the "Long-Tail" distribution problem discussed in Chapter 3's section [3.1.2](#). One of the advantages of retinaNet model is that it is designed specifically to prevent the large number of healthy tooth pixels from overwhelming the loss function during training stage. Additionally, it is simpler architecture than Faster R-CNN while maintaining competitive accuracy. However, while it improves recall on rare classes, it may produce looser bounding boxes compared to Mask R-CNN.

4.1.4 Mask R-CNN (Detectron2)

Mask R-CNN is an expansion of the Faster R-CNN architecture and it is a great model for predicting segmentation masks pixel-by-pixel. The way it works is that it extends Faster R-CNN's object recognition capabilities by introducing a third branch. This branch predicts an object mask for each identified instance in addition to the existing original classification and bounding-box regression branches. As dental diseases are rarely rectangular, it is hard to detect using exact bounding boxes (for example, a carious lesion or a cyst has an organic, irregular shape). Therefore, segmentation became a useful approach to detect these irregular shapes. One of the strongest segmentation models is the Mask R-CNN and it provides the tightest possible localization by learning the exact contour of the pathology rather than just a box. Moreover, it can distinguish between overlapping objects (e.g., a crown sitting on top of a tooth) more effectively than bounding boxes. One of the advantages of applying such model is that it can improve the accuracy of the main classification task by forcing the model to learn finer spatial features. However, it also has drawbacks such as it requires pixel-level annotations (polygons) and it can be computationally expensive to process.

4.1.5 DINO with DETR (DEtection TRansformer) using MMDetection framework

Unlike CNNs, DINO (DETR with Improved deNoising anchOr boxes) transformer model processes images in sequences using self-attention techniques. As the human tooth structure operates on a very strict structural logic (the dental arch), DINO can be a model to capture global dependencies and relationships. For example, DINO can understand that a tooth is missing in a sequence (e.g., 11, 12, [missing], 14) more effectively than CNNs models which focus on local features only. One benefit of this model is that it avoids the need for post-processing such as Non-Maximum Suppression (NMS). Moreover, the self-attention system of the model can detect long-range relationships throughout the entire time. However, this model can be difficult to use for detecting very small objects as compared to FPN-based CNNs models mentioned above.

4.2 Pre-processing and Training Pipeline

4.2.1 Pre-processing and Data Conversion

Most of the pre-processings are done on the feature engineering section [3.2](#). However, for the model training pipeline some processing had to be done. From the feature engineered dataset, the initial annotation format was in COCO JSON for all the dental x-ray images across all the folders. However, to train different models which are supported by unique detection frameworks, certain modifications had to be made to support the training pipeline. These modifications are detailed as below.

- **COCO to YOLO Conversion:** For the YOLO models, the COCO bounding box format ($[x_{\min}, y_{\min}, w, h]$) was converted to the YOLO format (normalized $[x_{center}, y_{center}, w, h]$). This stage required normalizing coordinates by the image width and height, respectively. Then, these information are saved to individual label files for each image.
- **General Pre-processing:** All models are trained with the standardized image size (1420×712 pixels). Data augmentation such as random horizontal flips, color jitter, and normalization (based on **ImageNet statistics**), was applied only to the

training set to enhance generalization capacity of the models. To help the learning models to generalize well on the unseen real-world images, no augmentations were applied to either validation or the testing sets.

4.2.2 Hyper-parameters, Learning Rate Scheduling, and Early Stopping

In order to get the best models from respective frameworks during the training stage, below consistent set of hyper-parameters are defined. These below parameters support the learning models to generalize well and learn the best weights needed to make the accurate predictions.

- **Base Learning Rate (LR):** A reduced base LR of 1×10^{-4} was applied first for fine-tuning the pre-trained weights for each respective models.
- **LR Scheduling:** Models are assigned with multi-step decay or reduction-on-plateau scheduler to carefully manage convergence. For the DINO model training, a combination of *Linear Warmup* followed by a *MultiStep decay schedule* is implemented to optimize learning stability [20] during the model training.
- **Early Stopping:** To prevent overfitting on the imbalanced dataset, an early stopping method was implemented to stop the training based on the patience rate. The training will be stopped if the monitored validation metric which is the mean Average Precision (mAP), does not improve for a certain number of epochs (patience ≥ 5 in DINO and ≥ 10 in standard CNN models).

4.3 Ensemble Strategy and Evaluation

After training all the models, a **thresholded soft voting ensemble** approach is designed to get the final output predictions. These predictions which are in JSON format will be used for the medical report generation pipeline and chat-bot question answering pipeline. Based on all the trained model metrics, the best models are the **YOLOv11m** model and the **Mask R-CNN** model. Therefore, these models are used to get the final output predictions via the ensemble approach. One of the reasons why these models

got the best performance metrics is the high recall of YOLO and the precise localization and segmentation ability of the Mask R-CNN architecture.

More detailed steps about the ensemble pipeline can be seen from the visualized flowchart in Figure 4.1. This flowchart is constructed in the following stages:

1. **Aggregated Prediction:** Predictions from the selected models are executed on the test image and aggregated into a single list of detections, each retaining its original source metadata.
2. **Filtering:** The aggregated list is generally noisy with different confidences from the chosen models. Therefore, **threshold filtering** is applied based on the class-specific confidence scores to preserve the best detections only. The predefined thresholds varies from hard-to-detect lesions (e.g., caries: ≥ 0.18) to stricter stable structures (e.g., FDI Classes: ≥ 0.6). More information about the threshold settings can be found in the appendix A.
3. **Prediction Merging:** A **Class-Aware Non-Maximum Suppression (NMS)** is applied with the IoU threshold of 0.4. This step is essential because it ensures the removal of highly overlapping predictions of the **same class** (e.g., a YOLO prediction for '46' and a Mask R-CNN prediction for '46') are resolved. Therefore, it keeps only the bounding box with the highest confidence score.

Evaluation Approach

To assess the performance of the five detection architectures, following evaluation approach is implemented. All models were evaluated on the **Test Set** with ($N = 379$ images) to ensure that the metrics reflect the generalization ability on the unseen clinical data.

The primary evaluation metric used was the **Mean Average Precision (mAP)**. For the detection and segmentation tasks, this metric is widely regarded as the gold standard. The calculation of mAP involves the following components:

1. **Intersection over Union (IoU):** The geometric overlap between a predicted bounding box (B_p) and the ground truth box (B_{gt}) is calculated as:

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_{gt})}{\text{Area}(B_p \cup B_{gt})} \quad (4.1)$$

A prediction is considered a **True Positive** only if the IoU exceeds a predefined threshold.

2. **Average Precision (AP):** For each class, the Precision-Recall (PR) curve is computed. The AP is defined as the area under this interpolated PR curve:

$$\text{AP} = \int_0^1 p(r) dr \quad (4.2)$$

where $p(r)$ is the precision at recall r .

Primary and Secondary Metrics: This study reports performance using the following standard metrics:

- **mAP (0.50:0.95):** This is the primary metric for model ranking. It represents the average AP over 10 different IoU thresholds (from 0.50 to 0.95 in steps of 0.05). This structured metric rewards models that achieve high localization accuracy, which is critical for distinguishing adjacent teeth, etc.
- **mAP@50:** This is the metric that is calculated at the single IoU threshold of 0.50. This map50 provides a looser measure of detection success which is useful for assessing general object recognition capabilities.
- **Per-Class AP:** Given the severe class imbalance (Section 3.1.2), an aggregate metrics can mask poor performance on rare classes. Therefore, AP is reported individually for critical pathologies (e.g., *Caries*, *Root-Stump*) to evaluate clinical reliability. This is one of the metric that help select the best models for the ensemble pipeline to generate the best possible predictions that matches the ground truths.

All evaluations were implemented using the standard `pycocotools` and `torchmetrics` python libraries to make sure it is reproducible.

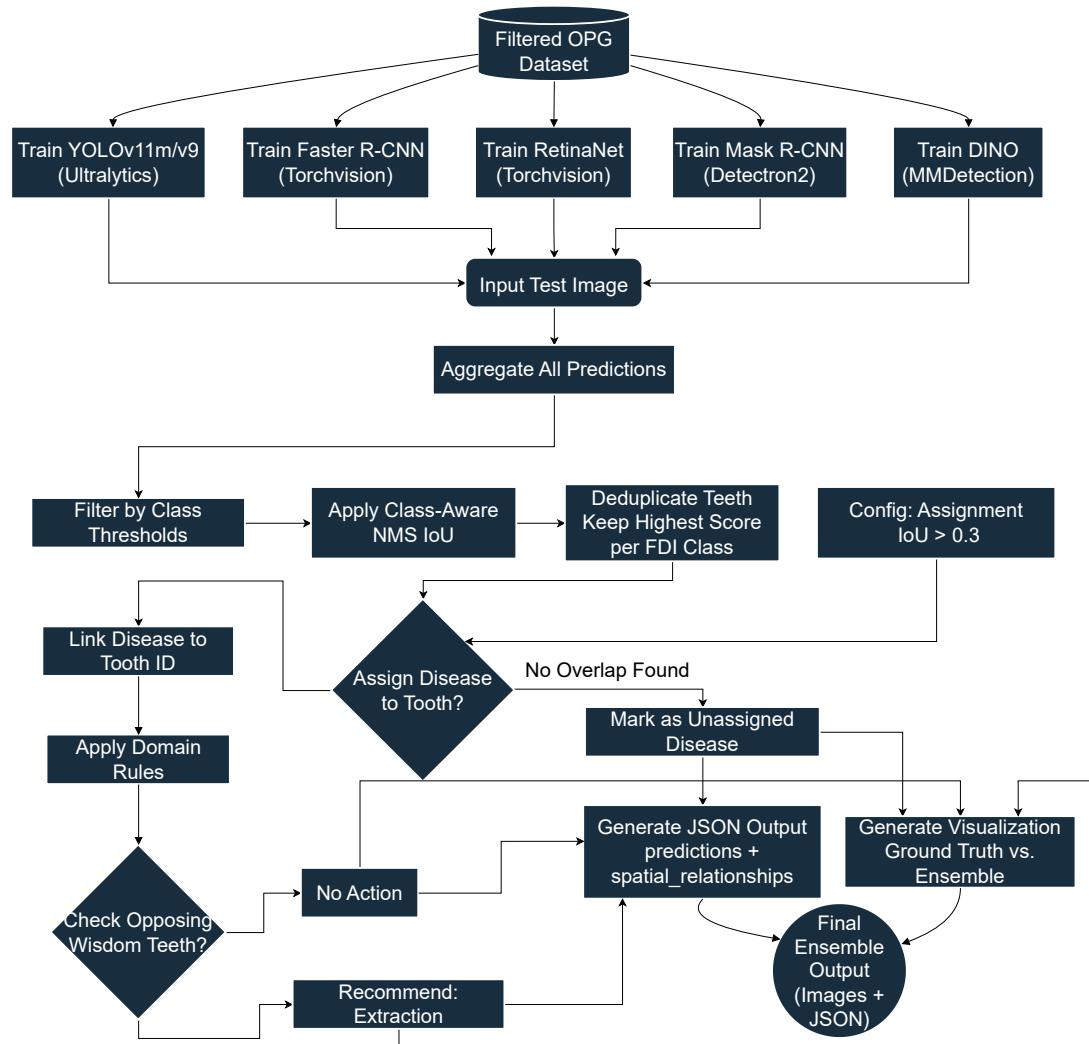


Figure 4.1: Flowchart of the Dental Detection and Post-Processing Ensemble Pipeline.

4.4 Post-Processing Algorithms

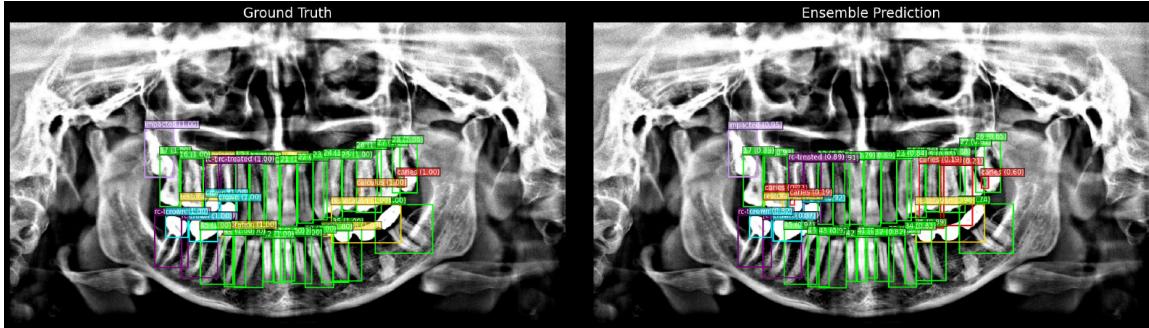


Figure 4.2: Comparison of Ground Truth Annotations vs. Final Ensemble Predictions.

Following the initial prediction merging, a custom post-processing stage was implemented to translate raw bounding box outputs into a clinically meaningful, structured report. This critical step addresses RQ2 by enforcing anatomical relationships rather than simple generic object detection. The Figure 4.2 is the final output of the ensemble pipeline after applying through a series of post-processing.

4.4.1 Spatial Relationship Matching (Mapping Pathologies to FDI Teeth)

The primary challenge is to accurately link a detected pathology (e.g., caries) to a specific tooth (e.g., '36'). This is done through an approach called **Spatial Relationship Matching**. First, the duplicate FDI tooth predictions are eliminated by keeping only the single highest-score detection per FDI tooth class. This ensures that no other same tooth FDI class number is identified in the entire image. Additionally, the remaining dental disease predictions are checked with the FDI tooth bounding boxes using an IoU threshold of (≥ 0.3). By any chance, if a pathology's bounding box overlaps with a tooth's bounding box above this threshold, that pathology is assigned to that tooth.

This process generates a structured output where each tooth object contains a list of its associated conditions. Any disease not meeting the IoU threshold with any tooth is written as an `unassigned_disease`.

4.4.2 Nearest Tooth Assignment and FDI Duplication Removal

The duplication removal approach ensures only the most confident detection for each tooth class. This approach is important to solve the false positives during the evaluation stage. Additionally, the assignment function performs a **Nearest Tooth Assignment** for pathologies, as the IoU method favors the closest and most overlapping FDI tooth box.

4.4.3 Domain Knowledge Rules and Structured Output Generation

At last, the final predictions are passed through a domain rule assignment stage which assigns clinical rule-based logic to the final output JSON file. Different rules are implemented at this stage. For example, a rule was implemented to check for unopposed opposing wisdom teeth (e.g., Tooth '18' present, but '48' missing/unassigned). This condition will set an **Extraction** recommendation.

At the end, the final output file will be a structured JSON file containing all the teeth lists and the associated spatial relationships. This JSON output will later be used in the medical pipeline section for generating the **Grounding Caption** for generating the medical report.

Methodology II: Generative AI and Reporting Pipeline

Following the detection of dental pathologies from chapter 4, a **Generative AI** medical report generation pipeline is implemented to synthesize clinical findings into professional medical documentation. Later, a separate chat-bot pipeline is also applied. The purpose of that pipeline is to provide an interactive patient support interface. This chapter will discuss about the translation of visual data into semantic **Grounding Captions**, which will be used as the foundation for the Large Language Model (LLM) medical report generation, and the implementation of the Retrieval-Augmented Generation (RAG) chat-bot.

5.1 Grounding Caption Generation

The transition from object detection to text generation requires an intermediate representation that is understandable for an LLM. The raw bounding box coordinates are insufficient for semantic reasoning. Therefore, a structured **Grounding Caption** generation approach was developed to construct a human-readable template from the ensemble's JSON output.

Human Template Construction

The system utilizes the structured prediction JSON to generate a clinical narrative. This process involves a two-stage mapping strategy:

1. **Observation:** This approach first scans for general pathologies that are not bound to a single tooth in the FDI class (e.g., general calculus presence or implants marked as unassigned).
2. **FDI Tooth-Specific Iteration:** The proposed approach then iterates through the detected FDI teeth (11-48). After iteration, for each tooth, it build up the associated conditions (e.g., Caries, Restoration) and the spatial attributes.

Lastly, a rule-based structure is followed to translate the spatial relationships into clinical mapping. For example, if the detection model(s) identifies a *Root Stump* with a high Intersection-over-Union (IoU) on Tooth 36, the system will construct the following sentence: *Tooth 36: Detected root-stump*. If domain rules (discussed in Chapter 4) are also present in the ensemble JSON file, it will also be added to the caption (if a recommendation like *extraction for an unopposed wisdom tooth* is available in the JSON field). This kind of addition to the human template construction ensures that the LLM receives a good grounded caption which is vital for reducing the risk of hallucination.

5.2 Medical Report Generation

After finalizing the grounding caption, it will be passed to the medical report generation stage as the input. The corresponding grounding caption will be converted to a high quality medical report with a detailed prompt which is structured to generate a standard medical report. This medical report will be converted to a PDF file which will be viewed and evaluated by the dentists for correctness, completeness, relevance, etc. Since, it is going to be evaluated by the dentists, it is important to ensure that the **LLM Prompt 5.2** which contains instructions to generate the medical report is well defined. Therefore, the prompt structure was designed with a specific persona, output format and it will act as an expert dentist who strictly gives medical report with recommendation by considering only the provided grounding caption and no adding external assumptions.

Step by step implementation about the proposed pipeline can be viewed in the Figure 5.1.

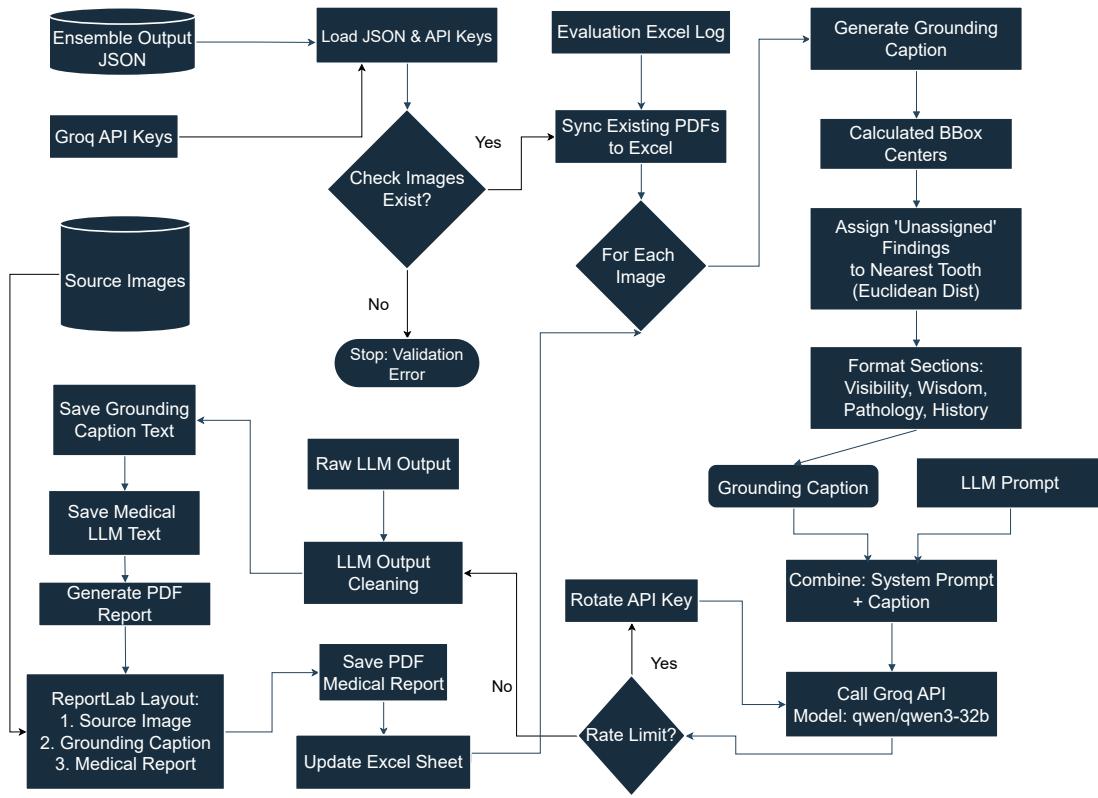


Figure 5.1: Architectural Flowchart of the Medical Report Generation Pipeline.

Groq API Setup and Key Rotation Strategies

One of the important steps of implementing a prompt engineering stage is to first acquire a API key that will help generate the content based on the engineered prompt. Therefore, Groq API is used with the **Qwen3-32B** model because of its correctness in medical domain and generous **Context Window (TOKENS)** of 131,072 and **Max Output Tokens** of 40,960 for a day with free of charge. Additionally, as the dataset contains test images of 379 images, there is a need to develop 379 medical reports. This will cost considerable amount of money. Therefore, to manage the API rate limits during batch processing of patient records, a **Key Rotation Strategy** was implemented. Upon encountering a rate-limit error (HTTP 429), the medical pipeline will automatically cycle to the next available key in the rotation pool. There are total of 7 API keys in the rotation pool to support this medical report generation and to ensure uninterrupted availability.

System Prompt for Medical Report Generation

Role: You are a professional oral radiologist assistant tasked with generating precise and clinically accurate oral panoramic X-ray examination reports based on structured localization data.

The structured data contains all detected teeth and dental conditions. Each condition is associated with a specific tooth number. If a finding is not directly on a tooth, it will have 'tooth_id': 'unknown' and a 'near_tooth': '[tooth_id]' field, which you should report as "near tooth #[tooth_id]".

Generate a formal and comprehensive oral examination report **ONLY** containing two mandatory sections:

1. **Teeth-Specific Observations**
2. **Clinical Summary & Recommendations**

The **Teeth-Specific Observations** section must comprise three subsections:

- **General Condition:** Outlines overall dental status, including the count of visualized teeth and wisdom teeth status (e.g., presence or impaction).
- **Pathological Findings:** Documents dental diseases such as caries, impacted teeth, calculus, or periapical radiolucency.
- **Historical Interventions:** Details prior treatments like fillings (restorations), crowns, root canal treatments, or implants.

Each finding in the structured data has a confidence score. You must apply the following processing rules **ONLY** for the **Pathological Findings** subsection:

- For confidence scores < 0.80 : Use terms like "suspicious for...", "suggests...", or "areas of concern noted for...".
- For confidence scores ≥ 0.80 : Use definitive descriptors such as "sign of...", "shows evidence of...", or "clear indication of...".

The **Historical Interventions** subsection should always use definitive language (e.g., "presence of a crown," "rc-treated tooth noted"), as these are observed facts.

Please strictly follow the following requirements:

- **Adherence to FDI numbering system** (e.g., "#11", "#26").
- **Use professional medical terminology** while maintaining clarity.
- **DO NOT** include or reference the confidence scores in any form in the final report. Their *only* use is to determine the certainty language ("suspicious" vs. "sign of").
- **DO NOT** generate any administrative content like 'Patient Name', 'Date', etc.
- **Generate a new Clinical Summary & Recommendations** section. This section is critical and must be created from the findings. It must include:
 1. **Priority Concerns:** The most urgent issues found (e.g., "Deep caries on #28", "Impacted wisdom tooth #18 requiring evaluation").
 2. **Preventive Measures:** Recommendations for prevention.
 3. **Follow-up Protocol:** Specific recall or follow-up actions.

Figure 5.2: Prompt configured for the LLM to generate structured dental reports.

As shown in Figure 5.2, the LLM prompt uses the grounding caption and generate two main sections such as **Teeth-Specific Observations** and the **Clinical Summary**. For the **Teeth-Specific Observations** section the dataset's classes will be split into *general condition*, *pathological findings* and *historical interventions*. This structural format is good for the dentists to view and evaluate the report. Moreover, this structure is suitable for converting into a PDF format with the help of the python library called **reportlab**.

Evaluation Approach

The generated reports were evaluated by domain experts (licensed dentists) based on six qualitative metrics. A **Likert-scale scoring system** (1–5) was applied to give a score for each below field.

- **Image Quality:** Is the image included in the report viewable without blur or corruption?
- **Report Quality:** Does the report contain all the necessary information? Is the report corrupted?
- **Clarity:** Are the contents of the generated report easy to understand from a professional point of view?
- **Relevance:** Does the report address the specific pathologies detected in the image?
- **Completeness:** Are all findings from the grounding caption included? Are all detected pathologies present in the report?
- **Correctness:** Are there any hallucinations or clinical errors? Are the recommendations provided correct enough?

The above six metrics are given to be filled in an excel file to the dentists. An additional *feedback* field is also included in the excel file for future improvements. A sample generated grounding caption and the medical report can be viewed from the Figure 5.3.

Input: Grounding Caption (Localization Data)	Output: Generated Medical Report
<p>This localization caption provides multi-dimensional spatial analysis of anatomical structures and pathological findings for this panoramic dental X-ray image, including:</p> <pre data-bbox="223 557 732 1192">Teeth visibility with center points (total: 29): [{'point_2d': [654, 416], 'tooth_id': '43', 'score': 0.94}, {'point_2d': [495, 397], 'tooth_id': '46', 'score': 0.93}, {'point_2d': [770, 415], 'tooth_id': '31', 'score': 0.93}, {'point_2d': [562, 248], 'tooth_id': '14', 'score': 0.93}, {'point_2d': [1141, 286], 'tooth_id': '28', 'score': 0.92}, {'point_2d': [461, 252], 'tooth_id': '16', 'score': 0.92}, {'point_2d': [973, 260], 'tooth_id': '25', 'score': 0.92}, {'point_2d': [1001, 415], 'tooth_id': '36', 'score': 0.92}, {'point_2d': [660, 254], 'tooth_id': '12', 'score': 0.92}, {'point_2d': [774, 266], 'tooth_id': '21', 'score': 0.91}, {'point_2d': [885, 260], 'tooth_id': '23', 'score': 0.91}, {'point_2d': [562, 408], 'tooth_id': '45', 'score': 0.91}, {'point_2d': [899, 420], 'tooth_id': '34', 'score': 0.91}, {'point_2d': [609, 246], 'tooth_id': '13', 'score': 0.91}, {'point_2d': [1091, 268], 'tooth_id': '27', 'score': 0.91}, {'point_2d': [731, 412], 'tooth_id': '41', 'score': 0.9}, {'point_2d': [335, 267], 'tooth_id': '18', 'score': 0.9}, {'point_2d': [695, 416], 'tooth_id': '42', 'score': 0.9}, {'point_2d': [833, 262], 'tooth_id': '22', 'score': 0.9}, {'point_2d': [609, 413], 'tooth_id': '44', 'score': 0.9}, {'point_2d': [804, 418], 'tooth_id': '32', 'score': 0.89}, {"point_2d": [518, 245], "tooth_id": "20", "score": 0.89}, {"point_2d": [848, 425], "tooth_id": "33", "score": 0.89}, {"point_2d": [713, 265], "tooth_id": "11", "score": 0.89}, {"point_2d": [1028, 269], "tooth_id": "26", "score": 0.88}, {"point_2d": [395, 248], "tooth_id": "17", "score": 0.85}, {"point_2d": [930, 260], "tooth_id": "24", "score": 0.88}, {"point_2d": [942, 415], "tooth_id": "35", "score": 0.87}, {"point_2d": [395, 248], "tooth_id": "17", "score": 0.85}, {"point_2d": [1144, 411], "tooth_id": "38", "score": 0.83}]</pre> <p>Wisdom teeth detection (total: 3): [{'box_2d': [1111, 234, 1170, 339], 'tooth_id': '28', 'is_impacted': false, 'score': 0.92}, {'box_2d': [304, 216, 367, 318], 'tooth_id': '18', 'is_impacted': false, 'score': 0.9}, {'box_2d': [1100, 348, 1188, 474], 'tooth_id': '38', 'is_impacted': false, 'score': 0.83}]</p> <p>Dental Pathological Findings (total: 4): [{'box_2d': [372, 344, 445, 422], 'tooth_id': 'unknown', 'label': 'root-stump', 'score': 0.81, 'near_tooth': '46'}, {'box_2d': [1041, 360, 1084, 394], 'tooth_id': 'unknown', 'label': 'caries', 'score': 0.44, 'near_tooth': '36'}, {'box_2d': [354, 389, 427, 433], 'tooth_id': 'unknown', 'label': 'periapical radiolucency', 'score': 0.39, 'near_tooth': '46'}, {'box_2d': [1041, 361, 1063, 392], 'tooth_id': 'unknown', 'label': 'caries', 'score': 0.25, 'near_tooth': '36'}]</p> <p>Historical</p> <p>Treatments (total: 3): [{'box_2d': [530, 336, 593, 389], 'tooth_id': '45', 'label': 'crown', 'score': 0.51}, {'box_2d': [1038, 356, 1106, 462], 'tooth_id': 'unknown', 'label': 'rc-treated', 'score': 0.92, 'near_tooth': '36'}, {'box_2d': [370, 285, 415, 313], 'tooth_id': 'unknown', 'label': 'restoration', 'score': 0.55, 'near_tooth': '17'}]</p>	

Figure 5.3: Comparison of the Grounding Caption (Left) and the Final LLM-Generated Medical Report (Right).

5.3 Interactive RAG Chatbot

Finally, the RAG chat-bot is implemented after the medical report generation pipeline. This pipeline enable users to discuss about their conditions, more information about the pathologies and the treatment recommendations. This pipeline is supported with an interactive interface which will be discussed in the next chapter 6.2. With this interface, users can query their own medical data in real-time. Additionally, the same Gorq API setup as discussed in section 5.2 is used for this pipeline. However, this pipeline is only focused only on the generated medical report and users are limited to ask questions only related to the report contents.

The system is designed to only accept the grounding caption and generated medical report as input to be read and to create an AI Agent tailored for the specific patient. Therefore, users cannot ask any **out-of-distribution** questions and it will adhere only to the dental domain. Moreover, the chat-bot is also designed to support the conversation history. This allows the users to read the previous questions as well as the outputted answers for the respective question. Step by step implementation about the proposed pipeline can be viewed in the Figure 5.4.

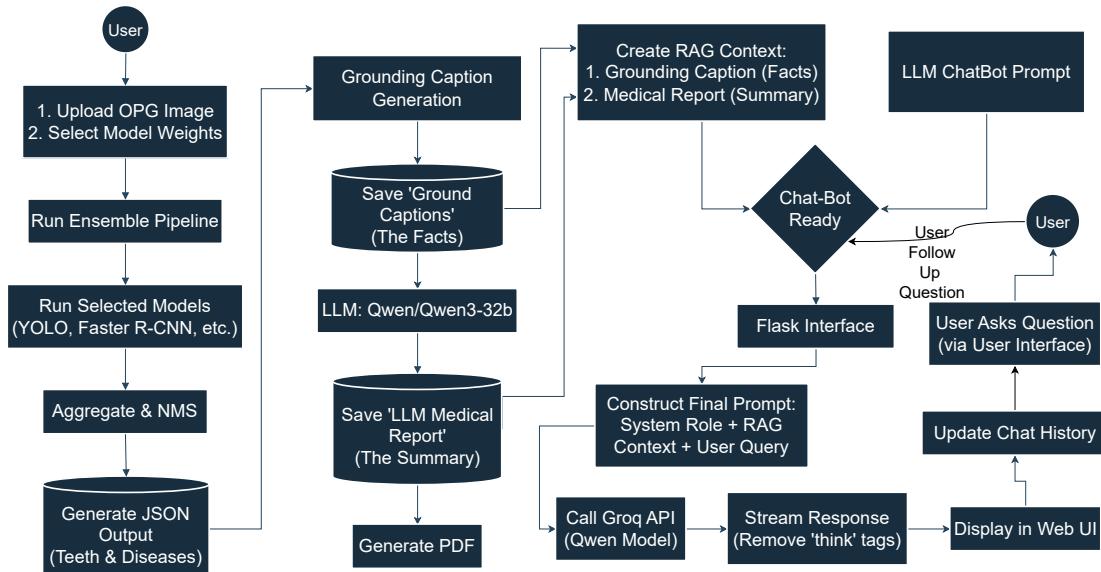


Figure 5.4: Workflow of the Retrieval-Augmented Generation (RAG) Chatbot Pipeline.

5.3.1 System Prompting and Role Definition

A Persona prompt 5.5 is defined to establish the parameters of the chat-bot to answer correctly to any question which will be asked by the user. This enable the chat-bot to act as a professional who is sympathetic and most importantly to decline to respond to inquiries that have nothing to do with the patient's particular dental report (such as general medical advice or non-dental topics).

System Prompt for RAG Chatbot

PERSONA: You are a senior **radiologist** specialized in panoramic dental X-ray imaging. Your tone is professional, calm, and empathetic. You explain complex medical findings in a simple, patient-friendly manner.

CONTEXT: You will be given the patient's full report findings as 'CONTEXT'. The CONTEXT includes:

1. **A structured location caption** (the raw facts from the X-ray, including confidence scores).
2. **A textual examination report** (the findings, summary, and recommendations).

CRITICAL RULES (DO NOT BREAK):

1. **GROUNDING RULE:** Your answers **must** be entirely faithful to the CONTEXT. Do not add, invent, or infer any medical information that is not explicitly stated in the CONTEXT.
2. **SCORE RULE (CRITICAL):** The CONTEXT includes a confidence score for each finding. You must use this score *only* to determine your language for **Pathological Findings**:
 - **Score < 0.80:** Use "suspicious for...", "suggests...", or "areas of concern noted for...".
 - **Score ≥ 0.80:** Use "sign of...", "shows evidence of...", or "clear indication of...".
 - **You must NEVER show the numerical score** (e.g., "score: 0.81") in your response.
3. **REFUSAL RULE:** If the patient asks about **cost, insurance, treatment alternatives, or asks for new medical advice**, you MUST politely refuse.
4. **STARTING RULE:** Your very first message must be: "Hello, I have your dental report here and can help answer any questions you have about it."

TASK & ADAPTIVE RESPONSE STYLE: Your task is to answer the user's questions about their report. You must adapt your response style based on the type of question:

1. **For General Patient Questions:** If the user asks a simple, conversational question (e.g., "What's wrong?"), your answer must be simple, clear, and empathetic. When defining a medical term, use your general knowledge, but always relate it back to the patient's CONTEXT (and obey the SCORE RULE).
2. **For Technical/Comprehensive Questions:** If the user asks for a comprehensive list or a full description (e.g., "List all pathological findings"), you MUST switch to a formal, technical, and data-driven style. In this mode, systematically list all findings from the CONTEXT that match the user's query, making sure to apply the **SCORE RULE** to your language and **NEVER** show the score.

Figure 5.5: Prompt Engineered LLM Chatbot for Question Answering.

More detailed information about the structure and rules of the chat-bot pipeline can be seen in the Figure 5.5. The chat-bot is also designed to greet the user upon the start of the interface to promote engagement. Additionally, it is designed to not answer the confidence scores from the grounding caption as it will make the user to doubt the system and lose trust.

5.3.2 Context Injection and RAG Role

The system is implemented with a specialized **Retrieval-Augmented Generation (RAG)** architecture. Unlike other vector-database RAG systems which search knowledge bases or various documents, books or libraries, this system uses **Patient-Specific Context Injection**. To be precise, it uses grounding caption and LLM generated medical report as contexts to be injected into the chat-bot prompt as RAG to avoid hallucination. When a user asks a question, the system constructs a prompt that combines:

1. The system **Role** (Figure 5.5).
2. The **Grounding Caption**.
3. The **Generated Medical Report**.
4. The user's **Query** (Question).

As the chat-bot's **Qwen** model is not fine-tuned due to the resource limitation and computation power, the above approach is implemented to get the desired output for question answering. Moreover, this approach ensures the Qwen LLM model to not rely on its pre-trained knowledge base (which contains generic statistics and unnecessary contexts) but to answer solely based on the specific patient's X-ray findings.

Conversation History Management

To simulate a natural dialogue with the user, the system implements a conversation history storing approach. User queries and assistant responses are stored in a session-state list. This history is appended to the context window for subsequent turns, allowing the chat-bot to handle follow-up questions (e.g., *Is that serious?* following a question about *Caries*) with full context awareness.

Evaluation Approach

To evaluate the question answering performance of the model, sample questions are constructed. Firstly, 5 medical reports are randomly selected. On each of the medical report, six sets of questions are asked to both the dentists and the chatbot. Both the dentists and the chat-bot answers are saved. The chat-bot answers are then later passed to the dentists to evaluate based on four main metrics such as **Clarity, Completeness, Correctness and Relevance**. A **Likert-scale scoring system (1–5)** was asked to score for each of the fields to the dentists based on their previously given answer to the questions. Moreover, different sets of questions are asked such as questions related to pathologies, information about the meaning of the pathologies to unrelated questions such as *who is my doctor?*, *Where can I download the medical report*, etc. More information about the questions and the evaluation can be seen in the Appendix [A.4](#). The following metrics are used to evaluate the performance of the chat-bot question answering.

- **Clarity:** Is the generated answer to the question easy to understand with ease?
- **Relevance:** Is the chat-bot able to answer the question asked by the user correctly? Does the answer's context match with the question expectation?
- **Completeness:** Are all findings from the grounding caption included? Are all the detected pathologies present in the answer based on the question?
- **Correctness:** Are there any hallucinations in the chat-bot answer? Are any of the recommendations generated clinically satisfactory enough?

The above four metrics are given to be filled in an excel file to the dentists. An additional *feedback* field is also attached in the excel file for future improvements of the chat-bot's question answering.

System Deployment and User Interface

The integration of ensemble detection models and Generative AI pipelines into a usable clinical tool requires a scalable system architecture. In this chapter, the full-stack implementation of the *Multimodal Dental Diagnostics* system, covering its deployment on cloud infrastructure, the **RESTful API** design for web and mobile integration, and the data management strategy utilizing Firebase for storage and user feedback loops will be discussed.

6.1 High-Level Architecture

The system is constructed as a modular microservices-oriented architecture, designed to support the heavy computational inference models from the client-side user interface. The main architecture consists of four primary layers such as the **Frontend Client Layer** for Web and Mobile, the **Backend Inference Layer** for models, pre-processing and post-processing, the **Database Layer**, and finally the **Model Execution Layer**.

As illustrated in Figure 6.1, the workflow is initiated when a client uploads a dental X-ray image. The backend processes this request through the **Model Execution Layer**, which starts up the ensemble detection pipeline and the medical report generation (LLM) pipeline. The results are returned to the client while simultaneously being logged asynchronously to the **Database Layer** for storing the images, medical generated reports and feedback collection.

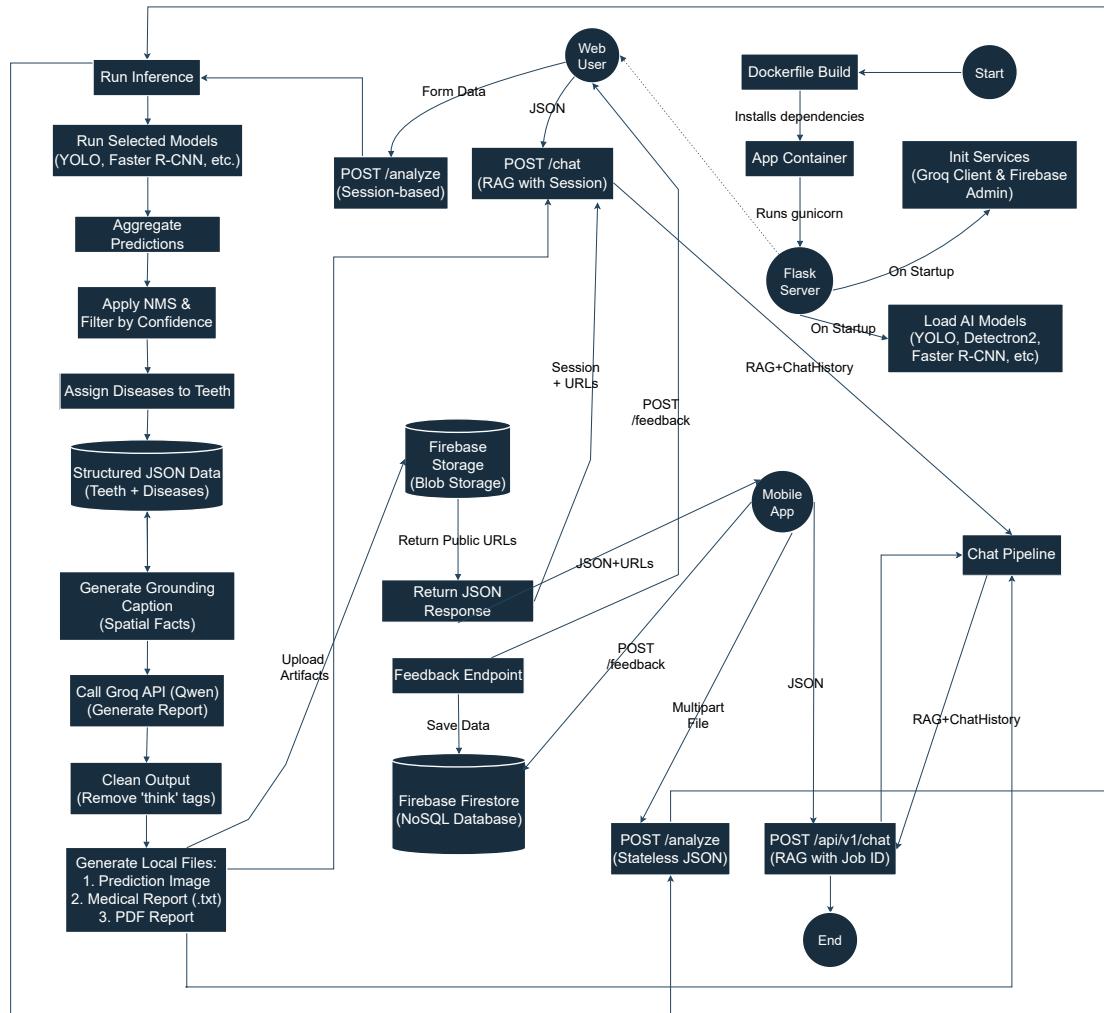


Figure 6.1: High-Level System flowchart for the data flow between the Frontend Clients, the Inference Backend (Hugging Face Spaces), and the Firebase Cloud Storage.

6.2 Web and Mobile API Integration

To ensure accessibility across different clinical devices, the system is deployed as a containerized application on **Hugging Face Spaces**. This platform provides a environment with RAM and CPU acceleration configurations which are essential for running the ensemble detection models and the Large Language Models pipelines with minimal latency.

The backend is designed to give a unique RESTful APIs that support both the Web interface and the Mobile Application. To distinguish between these platforms different payload optimization is implemented. For the mobile requests a prioritize compressed JSON responses is designed while for the web requests it is designed to retrieve full-

resolution diagnostic visualizations. There are three main APIs namely *analyze*, *chat* and *feedback* for both mobile and the web interfaces.

6.2.1 Diagnostic Analysis API

The core endpoint, `/analyze`, accepts raw X-ray image and initiates the full ensemble diagnostic pipeline and report generation pipeline. On the mobile app side, it sends the image and model choices. It then receives permanent URLs and a job-id. For the mobile environment, bandwidth efficiency is critical. Therefore, the API is structured to return structured data first. This allows the mobile UI to render the report text immediately while the heavier image assets load in the background. In the case of website, it sends the same data but relies on the session cookie to store the results. More information about the `/analyze` API can be found in the Figure 6.2.

6.2.2 Interactive Chat API

The `/chat` endpoint powers the RAG functionality, maintaining a session state to allow for context-aware follow-up questions regarding the generated report. There are two separate endpoints for `/chat` because one is stateless (Mobile) and one is stateful (Web). There is a slight variation with job-id while constructing the respective APIs because of their different state situation. The mobile app must send the job-id and the entire conversation history every time. The mobile `/chat` API information can be found in the Figure 6.3. On the other hand, for the web interface, it only needs to send the new question and the current history for context. It does not send a job-id and other information because the server remembers the report from the session cookie. The detailed `/chat` API for the website can be found in the Figure 6.4.

API Specification: Diagnostic Analysis (For both Web and Mobile Clients)

Endpoint: POST /api/predict/analyze

Description: Uploads an image and runs detection models. Returns a job-id and permanent URLs (from firebase) of the results.

Request Payload:

```
{  
  "image": "[binary_file_data.jpg]",  
  "models": "yolov11",  
  "models": "detectron2",  
}
```

Response:

```
{  
  "job_id": "a1b2c3d4-e5f6-7890-1234-56789abcdef0",  
  "caption": "This localization caption provides multi-dimensional  
  spatial analysis...",  
  "report": "## Medical Report\n\n### Teeth-Specific Observations...",  
  "original_image_url": "https://tym24-ai-the-dentist.hf.space/uploads/  
  a1b2..._patient_xray.jpg",  
  "predicted_image_url": "https://tym24-ai-the-dentist.hf.space/static/  
  results/a1b2..._predicted.jpg",  
  "api_urls": {  
    "pdf_report": "https://storage.googleapis.com/your-bucket  
    /feedback_files/a1b2.../report.pdf",  
    "predicted_image": "https://storage.googleapis.com/env-bucket/  
    feedback_files/a1b2.../predicted.jpg",  
    "original_image": "https://storage.googleapis.com/env-bucket/  
    feedback_files/a1b2.../original.jpg",  
    "medical_report_text": "## Medical Report\n\n### Teeth-Specific  
    Observations..."  
  }  
}
```

Figure 6.2: JSON Structure for the Mobile and Web Ensemble and Report Generation API.

API Specification: RAG Chat for Mobile

Endpoint: POST /api/v1/chat

Context: Injects the previously generated medical documents.

Request:

```
{  
  "job_id": "a1b2c3d4-e5f6-7890-1234-56789abcdef0",  
  "query": "Is the caries on tooth 37 serious?",  
  "history": [  
    {  
      "role": "user",  
      "content": "What is the main finding?"  
    },  
    {  
      "role": "assistant",  
      "content": "The report indicates a sign of caries on tooth #37."  
    }  
  ]  
}
```

Response:

```
{  
  "response": "The report notes a 'sign of' caries on tooth #37, which  
  suggests a high confidence in the detection. You should consult a  
  dentist for treatment."  
}
```

Figure 6.3: Chat API structure used by Mobile clients.

API Specification: RAG Chat for Web

Endpoint: POST /chat

Context: Injects the previously generated medical documents.

Request:

```
{  
    "question": "Is the caries on tooth 37 serious?",  
    "history": [  
        {  
            "role": "user",  
            "content": "What is the main finding?"  
        },  
        {  
            "role": "assistant",  
            "content": "The report indicates a sign of caries on tooth #37."  
        }  
    ]  
}
```

Response:

```
{  
    "response": "The report notes a 'sign of' caries on tooth #37, which  
    suggests a high confidence in the detection. You should consult a  
    dentist for treatment."  
}
```

Figure 6.4: Chat API structure used by Web clients.

6.2.3 User Feedback API using Firebase

To ensure continuous improvement of the diagnostic models, a feedback feature is implemented directly into the user interface. This allows users and clinicians to provide feedback about the accuracy of the automated report and to correct any misdiagnoses. This data is stored in a **NoSQL** structure to facilitate future model fine-tuning. This single /feedback endpoint handles both mobile and app platforms. However, their JSON style varies because of their different states. The mobile app must send all the data it has saved, because the server doesn't know who the user is. The detail contents of the mobile JSON content can be found in the Figure 6.6. Additionally, for the web client, the website sends a minimal payload. The server fills in the rest (report, URLs, job-id) from the user's active session. The complete detailed information about the web feedback API can be found in the Figure 6.5.

API Specification: Clinical Feedback feature

Endpoint: POST /api/feedback

Description: Captures expert validation of the AI output.

Request Payload (Web):

```
{
  "source": "website",
  "name": "John Smith",
  "feedback_text": "Great analysis, very fast!"
}
```

Response:

```
{
  "success": true,
  "message": "Feedback submitted successfully."
}
```

Figure 6.5: Standardized Feedback API Payload of the Web.

API Specification: Clinical Feedback feature

Endpoint: POST /api/feedback

Description: Captures expert validation of the AI output.

Request Payload (Mobile):

```
{  
    "source": "mobile_api",  
    "name": "Jane Doe",  
    "feedback_text": "The prediction for tooth #48 was incorrect.",  
    "job_id": "a1b2c3d4-e5f6-7890-1234-56789abcdef0",  
    "medical_report": "## Medical Report...",  
    "pdf_url": "https://storage.googleapis.com/.../report.pdf",  
    "original_image_url": "https://storage.googleapis.com/.../  
    /original.jpg",  
    "predicted_image_url": "https://storage.googleapis.com/.../  
    /predicted.jpg"  
}
```

Response:

```
{  
    "success": true,  
    "message": "Feedback submitted successfully."  
}
```

Figure 6.6: Standardized Feedback API Payload of the Mobile.

Database Management (Firebase)

The system is mounted with Google's Firebase for its backend-as-a-service (BaaS) capabilities by utilizing **Cloud Firestore** for document storage and **Firebase Storage** for media assets such as medical PDF files and original/predicted images. This serverless approach ensures scalability without the need for manual database maintenance which is hard and risky. The **Firestore Schema** which uses NoSQL database is structured into a primary collection called *feedback* to manage the entire diagnostic pipelines. It stores the data fields such as *job-id*, *feedback-text*, *medical-report*, *name*, *original-image-url*, *pdf-url*, *predicted-image-url*, *source* and *timestamp*. More information about this database can be found in the Appendix C.1. Moreover, **Firebase Storage** is built to host the binary assets

as shown below. When an image is processed:

1. The **Original OPG** is uploaded to the private bucket.
2. The **Annotated Prediction** (with bounding boxes) is stored in a public-read bucket for rendering in the UI.
3. The **Medical Report PDF** is generated and stored to provide a download link for the patient.

6.3 Security and Hardware/Software Requirements

Deploying the proposed system requires strict data security protocols and the implementation addresses these concerns with environment configuration and access control.

Security Environment Configurations

Sensitive credentials, such as the Large Language Model API keys (Groq) and Firebase Service Account tokens, are never hard-coded into the application. Instead, they are managed via **Environment Secrets** within the huggingface hosting platform. The application configuration module loads these secrets dynamically at runtime, ensuring that source code repositories remain free of vulnerable keys.

Hardware and Software Requirements

In order to use this system, there are certain hardware and software requirements. The application is designed to be lightweight and easy to understand for the users without any heavy processing to the cloud. For Web Clients, certain **Browser Compatibility** such as any modern web browsers that support HTML5 and ES6+ JavaScript can be used. For example, Chrome 90+, Firefox 88+ and Safari 14+ can be used with a stable broadband connection (≥ 5 Mbps upload). This is just an estimated network that is required to transmit high-resolution X-ray images without timeouts. For Mobile Clients, **Android** Version 10.0 (Q) or higher with approximate 150MB of RAM for the application runtime is needed. For the **iOS**, at least 14.0 version with standard **WKWebView** components for rendering the diagnostic dashboard with **Camera Permissions** is required for capturing OPG images directly from physical film if digital files are unavailable.

Results and Analysis

This chapter presents a quantitative and qualitative evaluation of the proposed Multi-modal Dental Diagnostics system. The analysis is divided into three primary sections: the detection performance of the computer vision models (Section 7.1), the quality assessment of the LLM-generated medical reports (Section 7.2), and the efficacy of the interactive RAG chat-bot (Section 7.3). This chapter will detail the performance metrics, highlight key observations, and provide a comparative analysis of different methodologies.

7.1 Detection Performance Metrics

To address Research Question 1 (**RQ1**), which investigates the capability of deep learning models to accurately predict dental pathologies in low-resolution X-ray datasets, five distinct models were evaluated. The primary metric for comparison was the mean Average Precision (mAP) calculated at an Intersection over Union (IoU) threshold range of 0.50 to 0.95 (mAP_{50:95}). The explanation of the detection metrics *IOU*, *AP* and secondary metric *Per-Class AP* can be found in the previously discussed section 4.3.

7.1.1 Comparative Analysis of Detection Architectures

Table 7.1 summarizes the aggregate performance across the entire test set. The results clearly indicate that the one-stage detector, **YOLOv11m**, achieved the highest overall

performance with an $mAP_{50:95}$ of **0.7491**, significantly outperforming the two-stage and transformer-based baselines.

Model	mAP (All Classes)	mAP (Main Classes)	Inference Type
Faster R-CNN	0.6666	0.4065	Two-Stage
RetinaNet	0.6538	0.3544	One-Stage
YOLOv11m	0.7491	0.4736	One-Stage
Mask R-CNN (Det2)	0.4620	0.3317	Two-Stage
DINO (MMDet)	0.6830	0.4411	Transformer (DETR)

Table 7.1: Summary of Aggregate Metrics for All Models.

The superior performance of YOLOv11m can be attributed to its advanced feature extraction backbone and efficient handling of spatial hierarchies, which proved robust against the noise presented in the low-resolution OPG images. Conversely, Mask R-CNN, despite its theoretical advantage in segmentation, underperformed with an mAP of 0.4620. This suggests that the pixel-level segmentation task was challenging given the quality of the ground truth masks and image resolution. Moreover, all models struggled to converge on the main classes such as *caries*, *calculus*, *implant*, etc. But, they captured most of the *FDI* tooth classes.

7.1.2 Performance on Pathological and Restorative Classes

A critical objective of this study was not just tooth localization, but the accurate detection of dental diseases. Table 7.2 provides a more detailed breakdown of performance on the nine key pathological and restorative classes. For each of the class, the one with the highest score is written in bold style to pinpoint easily.

Analysis of Key Pathologies

- **High-Performing Classes:** Distinct anatomical features such as **Implants** ($mAP = 0.988$), **RC-Treated teeth** ($mAP = 0.949$), and **Crowns** ($mAP = 0.903$) were detected with exceptional accuracy by the YOLOv11m model. These objects have high radiopacity (brightness) and rigid shapes, making them easier for Convolutional Neural Networks (CNNs) to distinguish from biological tissue.

Class	Faster R-CNN	RetinaNet	YOLOv11m	Mask R-CNN	DINO
	(ResNet50)	(ResNet50)		(Detectron2)	(MMDet)
Calculus	0.0313	0.0081	0.0067	0.0005	0.0200
Caries	0.1335	0.0671	0.2440	0.0293	0.1020
Periapical Radiolucency	0.0533	0.0326	0.1790	0.0072	0.0430
Impacted	0.6348	0.5916	0.8790	0.5349	0.7330
Root-Stump	0.3622	0.3053	0.7560	0.2775	0.4270
Crown	0.6938	0.6078	0.9030	0.6310	0.7580
Implant	0.6345	0.6506	0.9880	0.5691	0.7390
RC-Treated	0.7201	0.6441	0.9490	0.6507	0.7410
Restoration	0.3951	0.2828	0.6670	0.2852	0.4070
Mean (Main Classes)	0.4065	0.3544	0.4736	0.3317	0.4411

Table 7.2: Comparative Performance (mAP 50:95) on main classes across 5 Models.

- **Challenging Classes:** The detection of **Caries** (mAP = 0.244) and **Periapical Radiolucency** (mAP = 0.179) proved significantly more difficult across all models. These conditions have very small radiolucent (dark) areas with hard to detect boundaries. They are hard to detect because they often blend into the trabecular bone pattern. This difficulty shows that there is a need for the **Ensemble Strategy** (Methodology I) with more powerful models that can detect small bounding boxes more efficiently.
- **Calculus:** This class has the lowest performance (mAP less than 0.04) across all architectures even though it has more than 2600 annotations across *train*, *valid* and *test* folders. Calculus classes are often microscopic and visually similar to image noise. Additionally, the score reflect a limitation of using low-resolution data for detecting small classes without more specialized targeted pre-processing.

7.1.3 Spatial Localization Accuracy (FDI Teeth)

Accurate tooth numbering is very important for clinical utility. Moreover, these FDI tooth number plays a vital role for the medical reporting and question answering pipelines. If wrong teeth numbering is detected, then it will give wrong information for the recommendations and clinical advice. Therefore, one of the most important task in

detection pipeline is to have correct predication on the FDI tooth classes. The table 7.3 details the performance of all the models across the 32 permanent teeth classes.

Table 7.3: FDI tooth Detection Performance (mAP 50:95)

FDI Class	Faster R-CNN	RetinaNet	YOLOv11m	Mask R-CNN	DINO
11	0.0000	0.6685	0.7870	0.4844	0.7010
12	0.7456	0.7212	0.8230	0.4881	0.7350
13	0.7556	0.7357	0.8230	0.5787	0.7660
14	0.7219	0.6938	0.7900	0.4306	0.6990
15	0.7945	0.7909	0.8430	0.5405	0.7800
16	0.7925	0.7893	0.8450	0.4878	0.7550
17	0.8054	0.8144	0.8570	0.4673	0.7930
18	0.7349	0.7071	0.7920	0.2503	0.7180
21	0.7010	0.6711	0.7770	0.4986	0.6980
22	0.7398	0.7061	0.8080	0.5171	0.7280
23	0.7771	0.7421	0.8320	0.5314	0.7660
24	0.7695	0.7294	0.8220	0.5141	0.7550
25	0.8110	0.7997	0.8300	0.5189	0.7910
26	0.8032	0.7964	0.8540	0.4833	0.7720
27	0.8309	0.8143	0.8760	0.4741	0.7990
28	0.7604	0.7055	0.8030	0.2947	0.7240
31	0.5835	0.5368	0.6760	0.3617	0.6040
32	0.6427	0.5941	0.7230	0.5298	0.6520
33	0.7805	0.7432	0.8370	0.6211	0.7690
34	0.8048	0.7615	0.8530	0.5866	0.7770
35	0.8521	0.8256	0.8850	0.6424	0.8180
36	0.8645	0.8650	0.9300	0.6334	0.8630
37	0.8842	0.8629	0.9340	0.5112	0.8630
38	0.7578	0.7262	0.8440	0.3488	0.7030
41	0.5817	0.5554	0.6890	0.3740	0.6090

Continued on next page...

FDI Class	Faster R-CNN	RetinaNet	YOLOv11m	Mask R-CNN	DINO
42	0.6263	0.5903	0.7210	0.5102	0.6610
43	0.7058	0.7027	0.8040	0.5638	0.7050
44	0.7574	0.7346	0.8310	0.6060	0.7580
45	0.8097	0.7951	0.8700	0.6125	0.8070
46	0.8388	0.8337	0.9220	0.5756	0.8520
47	0.8560	0.8382	0.9170	0.5466	0.8750
48	0.7822	0.7670	0.8530	0.3736	0.7380

The results in Table 7.3 demonstrate that YOLOv11m outperformed other models in identifying individual tooth classes, achieving **mAP** greater than 0.80 for most molars and premolars (e.g., #36, #46, #37). Lower incisors (e.g., #31, #41), which are crowded and visually similar, showed a slight drop in performance across all models, though YOLOv11m maintained a good confidence approximately about **mAP** 0.68, whereas Mask R-CNN dropped to approximately 0.36. This confirms that the anchor-free approach of YOLOv11 effectively handles dense object clusters in dental radiography.

7.1.4 Case Studies: Visual Analysis of Predicted Outputs

To qualitatively analyze the limitations of the single-stage detection model (YOLO), below two case studies are attached. One failure case involving minute pathological features and FDI class mismatch (Figure 7.1) and another one, performing badly on the low confidence classes such as calculus (Figure 7.2).

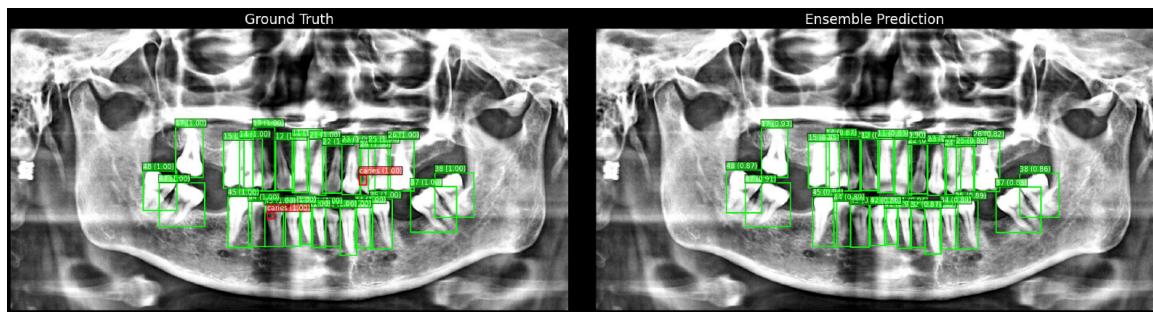


Figure 7.1: Case Study 1: Small regions detection such as caries and FDI Class mismatch.

In the above figure, a significant limitation is observed in the detection of early-stage pathologies such as caries. As shown in the comparison, the Ground Truth (Left)

identifies a specific instance of *caries* on the distal aspect of the mandibular premolar. However, the Model Prediction (Right) fails to detect this bounding box entirely. This *false negative* can be the cause of extremely low signal-to-noise ratio of small carious lesions in low-resolution OPGs. This highlights the difficulty YOLOv11m faces when distinguishing minute texture changes from the background noise of the trabecular bone.

Moreover, the model showed confusion in anatomical sequencing. While the teeth were localized, some FDI class labels were predicted incorrectly (e.g., misclassifying adjacent premolars due to spatial crowding). Unlike a human expert who knows tooth identity based on the entire arch sequence (44, 45, 46), the detector treated each tooth as an independent individual. This loss of sequential context led to a classification mismatch, and further validating the necessity of the proposed *Spatial Relationship Matching* post-processing algorithm to ensure logical tooth ordering.

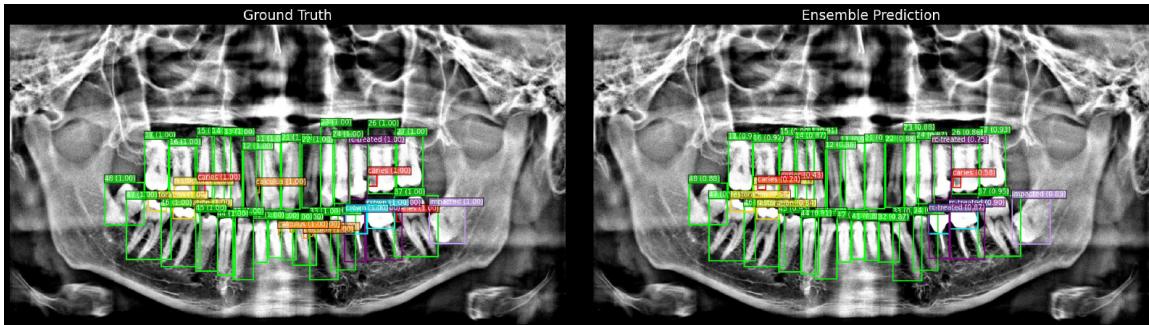


Figure 7.2: Case Study 2: Performing badly on Calculus class.

A critical analysis of the detection results, as visualized in Case Study 2's Figure 7.2, shows a distinct disparity in model sensitivity based on classes. The comparison between the ground truth and the prediction demonstrates the model's high accuracy in identifying high-contrast rigid classes. As seen in the Model Prediction (Right), the FDI tooth numbering and distinct restorative treatments (such as *rc-treated*) align almost perfectly with the Ground Truth (Left). The bounding boxes for these classes are precise, and the classification confidence is high, confirming that the YOLOv11m architecture successfully learned the macro-features of dental anatomy.

However, the model shows a significant blind spot regarding texture-based pathologies. While the Ground Truth annotates multiple instances of *calculus* along the cervical margins, the model fails to predict these regions entirely. This selective failure explains

the extremely low mAP score for the calculus class reported in Section 7.1. Unlike the sharp edges of a root canal treatment, calculus appears as a faint, low-contrast radiopacity that visually appear as image noise. The model’s failure to resolve these fine-grained texture differences while simultaneously performing good at larger objects, highlights a critical limitation in applying general-purpose object detectors to microscopic conditions without specialized texture-enhancing preprocessing.

7.2 Medical Report Generation Performance Metrics

The automated report generation pipeline was evaluated on 50 patient cases. To ensure clinical validity, the generated PDF reports were assessed by expert dentists using a 5-point Likert scale (1=Poor to 5=Excellent) across six quality such as *Image Quality*, *Report Quality*, *Clarity*, *Relevance*, *Completeness*, and *Correctness*. Due to limited time availability by the dentists, only 50 cases were evaluated out of all the test cases.

7.2.1 Quantitative Analysis of Clinical Metrics

Table 7.4 summarizes the descriptive statistics for the evaluation. The system demonstrated high reliability in processing visual inputs, with an average **Image Quality** score of 3.98 with the standard deviation score of ± 0.14 . This low standard deviation shows that the detection pipeline preserved diagnostic features that are acceptable to clinicians.

Metric	Mean (μ)	Std Dev (σ)
Image Quality	3.98	0.14
Report Quality	3.26	0.69
Correctness	3.14	0.83
Completeness	3.26	0.75
Relevance	3.22	0.82
Clarity	3.24	0.82

Table 7.4: Statistical Summary of Dentist Evaluations (N=50).

For this pipeline, the metrics reflect directly to the prompt structure. The **Report Quality** and **Completeness** achieved similar mean scores of 3.26 with (SD approx-

imately 0.7). This suggests that the LLM effectively captured the findings from the grounding caption in the majority of cases. However, **Correctness** ($\mu = 3.14$) and **Relevance** ($\mu = 3.22$) scored slightly lower with higher variance (σ greater than 0.8). This indicates that while the reports were structurally great, there were occasional cases of minor wrong interpretations.

Distribution and Correlation Analysis

To understand the proportion of agreement among dentists, the raw Likert counts were converted into percentages:

$$\text{Percentage} = \left(\frac{\text{Count of Specific Score}}{\text{Total Responses}} \right) \times 100 \quad (7.1)$$

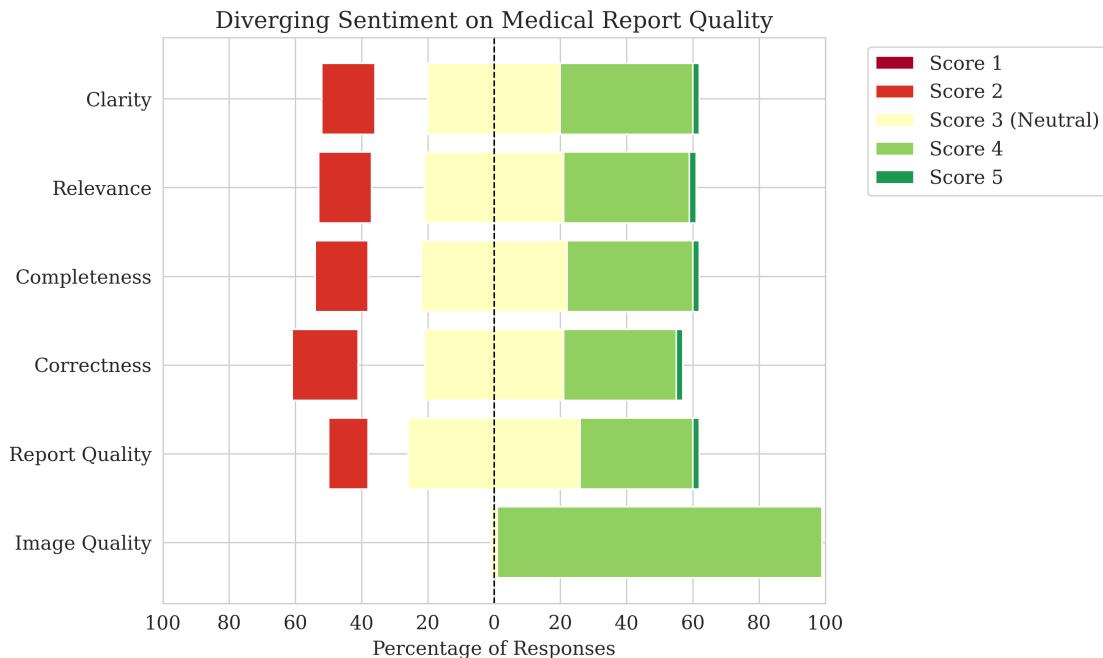


Figure 7.3: Diverging Stacked Bar Chart with Likert Scores.

In Figure 7.3, the diverging bar chart shows that the **Clarity** and **Relevance** metrics skew positively, with very few instances rated as **Poor** (Score 1). Moreover, the **Image Quality** is the most stable metric, while **Correctness** shows the widest interquartile range. This is due to the models which did not get all the complex pathology classes. Furthermore, the correlation analysis (Figure 7.5) showcased a strong positive correlation between **Report Quality** and **Clarity**. This confirms that for dental practitioners, the use of clinically clear language is vital for report quality. Additionally, the boxplots

from Figure 7.4 also indicate a high median performance (Score 3.0–4.0) with consistent image quality. This shows that the x-ray images are clearly represented in the reports for the dentists to validate the report's correctness and completeness.

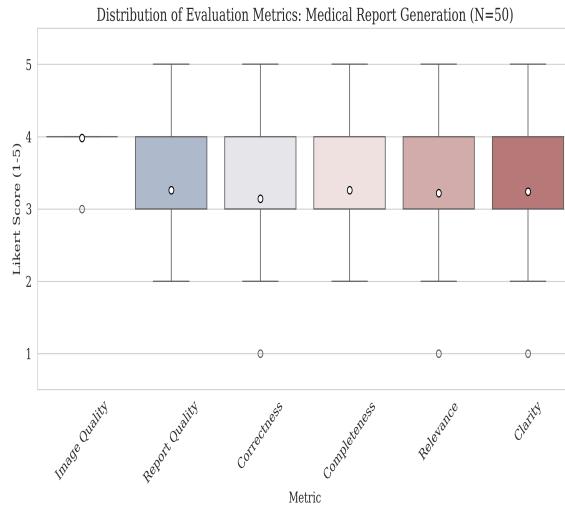


Figure 7.4: Distribution of Evaluation Metrics.

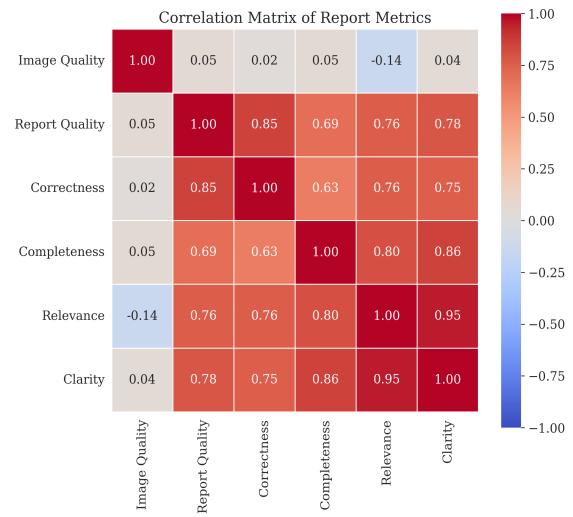


Figure 7.5: Correlation Matrix of Metrics.

7.2.2 Case Study: Confidences with Generated Report

After reviewing the generated reports, it is found that sometimes a report fails to follow rules mentioned in Prompt Engineering. The System Prompt explicitly said "***DO NOT include or reference the confidence scores in the final report***". However, in several test cases, the Qwen model failed to adhere to this rule.

Figure 7.6 presents a sample generated report where the model explicitly wrote "(low confidence)" and "(confidence greater than 0.80)" within the Pathological Findings section. This situation is one of the recent findings on the difficulty most LLMs face with negative constraints [34]. Because the confidence scores were important features in the input Grounding Caption (e.g., 'score' : 0.81), the LLM's attention mechanism weighted them heavily. Despite the instruction to suppress them and only to use these to determine language (e.g., suspicious vs. clear indication) to be added in the report, it mistakenly added the scores in the report. This is a form of data leakage in the final text and more structured prompt needs to be implemented to avoid these issues in future. In the figure, the mistakes are represented with the red arrows with the *Error* message.

Generated Report Showing Prompt Adherence Failure

Oral Panoramic X-Ray Examination Report

1. Teeth-Specific Observations

General Condition

- A total of **25 teeth** are visualized.
- **Wisdom teeth status:** Tooth #18 (FDI 18) is **erupted and unimpacted**; no additional wisdom teeth are identified.

Pathological Findings

- **Tooth #18:** Suspicious for **caries** (low confidence). ← Error
- **Near tooth #37:** Clear indication of **root stump**.
- **Near tooth #18:** Suspected **root stump** (confidence < 0.80). ← Error
- **Near tooth #27:** Suspicious for **root stump** and **caries** (confidence < 0.80). ← Error

Historical Interventions

- No prior restorations, crowns, root canal treatments, or implants are observed.

2. Clinical Summary & Recommendations

Priority Concerns

1. **Deep caries on #18:** Requires urgent evaluation for potential restorative intervention.
2. **Root stumps near #37 and #27:** Indicates prior extraction; assess for residual fragments or infection.

Figure 7.6: Sample report output where the LLM included confidence metadata.

7.3 Performance of RAG Chatbot Question Answering

The final component of the multimodal pipeline is the interactive RAG chatbot. This pipeline was evaluated to test whether it solves patient communication gap (RQ4). Five generated reports were selected and total of 30 total queries (6 questions per report) are given to the chatbot and dentists to give their expert answers. These queries were stratified into two categories:

1. **Domain-Specific Queries:** Questions directly answerable from the report (e.g., "What is caries?", "What is the status of tooth 18?", "Explain more about the recommendations.", "Explain my treatment history.", etc).
2. **Out-of-Distribution (OOD) Queries:** Irrelevant questions (e.g., "Who is the doctor?", "How do I download the report from the interface?", etc).

Once the dentists answered the questions, these answers are used as reference to score the responses of the chatbot's generated answer by the dentists on a 5-point Likert scale across four metrics namely *Completeness*, *Relevance*, *Clarity* and *Correctness*. More information about the questions asked during the evaluation stage can be found in the Appendix A.

7.3.1 Quantitative Analysis of Chatbot Metrics

Table 7.5 shows the aggregate performance metrics of the system. It achieved highest performance in **Completeness** with ($\mu = 3.67$ and standard deviation of ± 0.61). This proves the RAG architecture's strong capacity to retrieve relevant context from the injected grounding captions and medical reports to answer user's queries.

Metric	Mean (μ)	Std Dev (σ)
Completeness	3.67	0.61
Relevance	3.43	0.77
Clarity	3.33	0.92
Correctness	3.30	1.06

Table 7.5: Statistical Summary of Chatbot Evaluation (N=30 Responses).

However, **Correctness** showed the highest standard deviation of ($\sigma = 1.06$). This score shows that while the chatbot is accurate most of the times, it can still experience issues with factual precision, particularly when answering complex pathological descriptions.

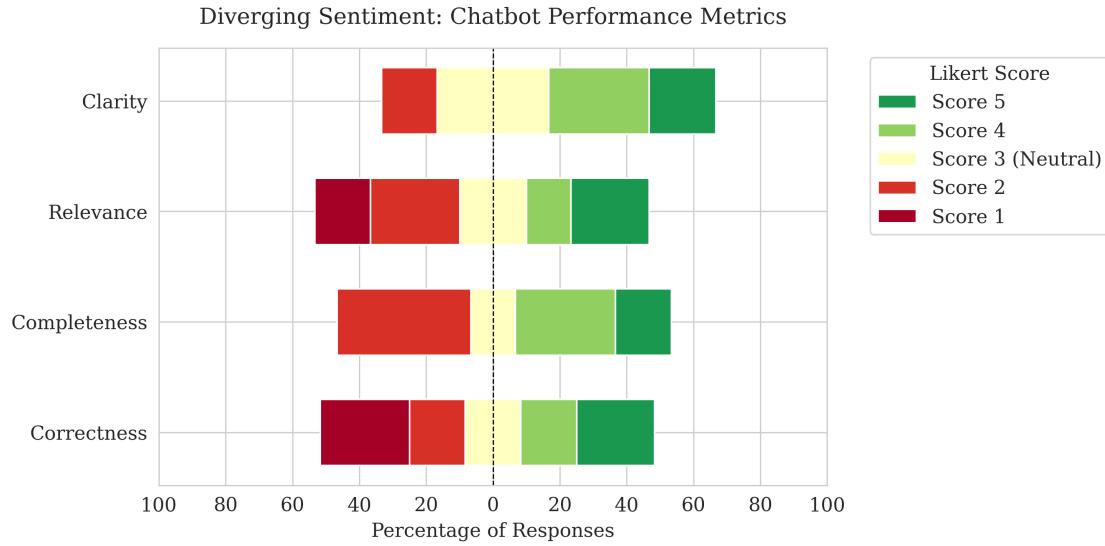


Figure 7.7: Diverging Stacked Bar Chart of Chatbot Likert Scores.

The majority of responses were rated Neutral (3) to Positive (4-5). The *Completeness* metric shows the strongest positive and *Correctness* appears as a wider spread of opinions. The distribution of scores (Figure 7.7) proves this finding. The divergence in *Correctness* and *Clarity* indicates that while the retrieval mechanism works well with high completeness, the generation part can struggle to simplify complex dental terminology for a layperson audience, leading to lower clarity scores from dentists.

7.3.2 ChatBot Correlation Analysis

From figure 7.9, a significant positive correlation was observed between **Relevance** and **Completeness**. This indicates that when the RAG system successfully retrieves the correct segment of the report (Relevance), the resulting answer is almost perceived as complete. This shows a positive effect of the **Context Injection** strategy described in the Methodology. Additionally, it proves that the performance is retrieval accuracy and not generation fluency. The correlation analysis also highlights a moderate-to-strong positive correlation between Relevance and Correctness. Interestingly, completeness

showed a lower correlation with Clarity, suggesting that while the RAG system successfully retrieved all the information, the LLM sometimes presented it in a dense or overly technical manner that reduced the clarity for the user.

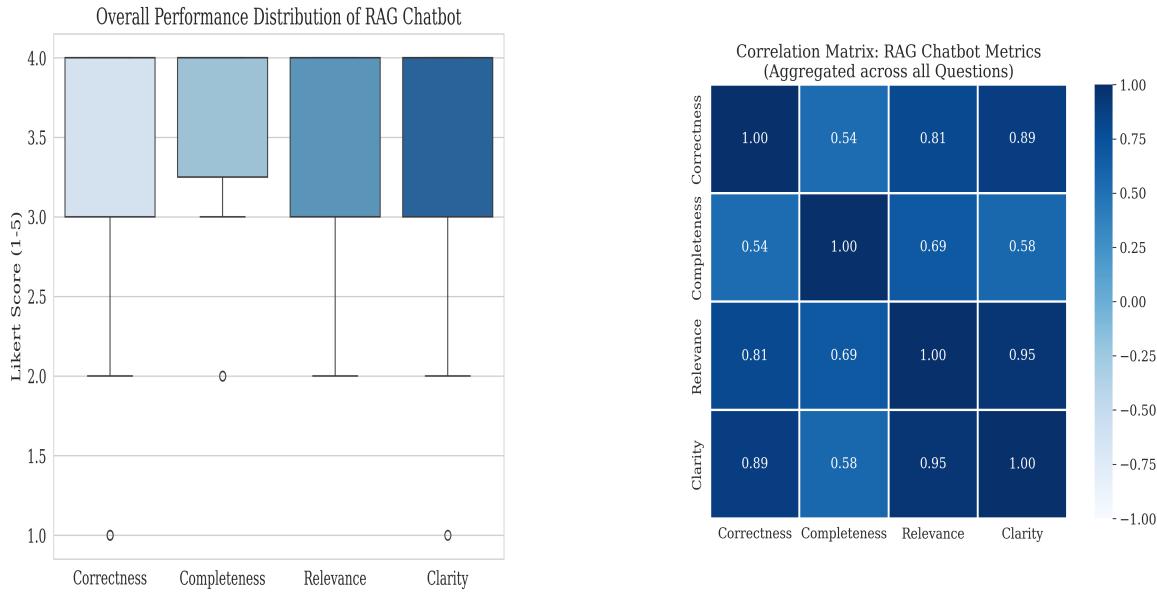


Figure 7.9: Correlation Matrix of Chat-bot.

Figure 7.8: Overall Distribution of Scores.

7.3.3 Performance by Question Type

The system was further evaluated using six distinct question sets to test different capabilities from general summarization to specific pathology retrieval. From this evaluation, a critical success factor was the chatbot's rule following defined in the System Prompt. When presented with OOD questions (e.g., "Can you tell me the doctor's name?"), the system scored high in **Correctness** by refusing to give answer. Instead, it correctly responded "*Please ask questions based on the medical report.*" This behavior confirms that there is a safety alignment required for deployment in medical sectors to prevent misinformation [35]. When asked about factual, lookup-based queries (e.g., "What is caries?", "List my recommendations"), the chatbot was able to answer correctly and relevantly. This proves that the chatbot is good at extraction tasks where the answer is already present in the report. Overall, the Chatbot has a strong capacity for retrieval accuracy ($\mu_{Completeness} = 3.67$) while maintaining strict boundaries against hallucination on out-of-topic queries.

Discussion

The results across the three main pipelines provide compelling evidence for an end-to-end multimodal diagnostic system in dentistry. This chapter will discuss these findings in the context of the research questions, analyzing the trade-offs in detection architecture, generative reporting, and the effect of retrieval-augmented patient assistance. Furthermore, the technical and clinical limitations that define the current scope of this work will also be discussed.

8.1 Detection Pipeline

The comparative analysis of five detection architectures (Section 7.1) showed insights on selection of models for low-resolution dental radiography. The superior performance of **YOLOv11m** ($mAP_{50:95} = 0.7491$) over two-stage detectors like Faster R-CNN ($mAP = 0.6666$) challenges traditional assumptions in medical imaging that favor region-based approaches for precision. This actual result suggests that recent advancements in single-stage feature extraction, specifically the cross-stage partial connections in YOLOv11, have successfully bridged the accuracy gap. The model's ability to maintain high recall on classes like **Implants** (0.988) and **RC-Treated Teeth** (0.949) confirms that shape-based feature learning is a good approach even in noisy, low-resolution environments. However, the failure to detect **Calculus** ($mAP < 0.01$) across all models is not a data quantity issue but a resolution constraint. Calculus generally has a high-frequency textural noise. In low-resolution OPGs, this texture is hard to distin-

guish from others. This finding proves that standard object detection on downsampled images is insufficient and future approaches may require dedicated super-resolution preprocessing. Furthermore, the FDI class mismatch observed in Case Study 1 reveals that Convolutional Neural Networks (CNNs) lack sequential logic. The model can detect isolated teeth but does not understand the dental arch. Additionally, this finding validates the implementation of the **Spatial Relationship Matching** strategy as a post-processing step rather than relying on end-to-end learning for tooth numbering.

In the context of architectural trade-offs, the results further clarify the limitations of alternative architectural models for this specific domain. **RetinaNet**, despite using Focal Loss to specifically target class imbalance, failed to improve detection for rare classes like Calculus. This indicates that the issue with detecting pathologies is not only the frequency of negative samples which Focal Loss addresses but also the fundamental lack of discriminative features in low-resolution pixel data. Similarly, **Mask R-CNN** underperformed relative to the bounding-box models. This suggests that the segmentation task is beneficial for irregular lesion shapes but adds computational noise when trained on low resolution ground truth masks. Finally, the transformer-based **DINO model**, while designed to capture global dependencies using self-attention mechanisms, did not surpass the CNN-based YOLOv11m. This indicates that for small, specialized medical datasets, CNNs models may still offer great convergence compared to the data-hungry nature of vision transformers, which require massive-scale pre-training to fully leverage their global context capabilities.

8.2 Medical Report Pipeline

The transition from visual detection to textual reporting introduced unique challenges in Generative AI control. The quantitative metrics indicated high structural accuracy ($\mu_{Completeness} = 3.26$), yet the qualitative case studies exposed an issue with **Prompt Leakage**. The appearance of internal *confidence scores* in the final report, despite engineering a specific rule in the prompt for not mentioning, aligns with the "Pink Elephant" phenomenon in Large Language Models (LLMs) [34]. The model's attention mechanism heavily weights numerical values presented in the prompt. This suggests that for high-stakes medical reporting, simple instruction-tuning is not enough. A more

favored solution is to implement **Chain-of-Thought** pipeline where the reasoning step (interpreting the score) is separated from the final generation step.

Moreover, the strong correlation between **Report Quality** and **Clarity** shows that clinicians prioritize readability. However, the variance in **Correctness** ($\sigma = 0.83$) indicates that the **Grounding Caption** strategy, while effective, is not a perfect protection against hallucination. In rare cases, the system sometimes produce text containing clinical information, such as pain, which was absent from the grounding caption. Therefore, there is a need for dentist verification to prevent panic among the users before forwarding the report to them.

8.3 RAG Chat-Bot Pipeline

In the previous section, the interactive chatbot component demonstrated that **RAG** is a good approach for the **Communication Gap** (RQ4). A critical success of the system was handling of Out-of-Distribution (OOD) queries. The chat-bot's consistent refusal to answer irrelevant questions (e.g., "Who is the doctor?") validates the effectiveness of system's persona constraints. In medical AI, the ability to *not* answer is as important as the ability to answer correctly to prevent misleading advice [35].

For the context of retrieval accuracy and simplification, the high **Completeness** scores ($\mu = 3.67$) indicated that the utilization of specific generated report and grounding captions into the prompt worked effectively. Unlike vector-based RAG which can retrieve irrelevant documents, this context utilization ensures 100% relevance of the source material. However, the slightly lower **Clarity** scores suggest that there is an issue with the simplification gap. The model occasionally struggled to translate complex terms into layperson terms without losing clinical precision. Additionally, there is also a need to test the entire complete set of medical reports with the chat-bot questions and answers to actually identify how the chat-bot is performing against all the metrics.

8.4 Limitations

While the system demonstrates significant potential, several limitations define the boundaries of its current applicability.

8.4.1 Technical Limitations

- The reliance on resized 1420×712 images for training effectively affected the performance for micro-pathologies like initial caries and calculus. Without higher-resolution inputs, model convergence on these classes is mathematically debatable.
- The dependence on detailed prompt engineering means the system is sensitive to model updates. A slight change in the underlying LLM's behavior (e.g., an update to Llama 4) could break the output flow and will require restructuring of the system prompts.
- The dependence on the free daily groq API limit is not suited for deployment. During peak loads, rate limits can be an issue and even with key-rotation strategy. Potential points of failure can affect the engagement of the user interactions.

8.4.2 Clinical Limitations

- OPG X-rays are flat 2D pictures of a 3D mouth. Because the image is flattened, different parts of the anatomy can overlap and look like dark spots. The system might mistake these shadows for cavities or lesions because, unlike a 3D CT scan, it cannot see depth.
- The system analyzes only one image at a single point in time. It cannot compare the current X-ray to older ones. This means it cannot tell the difference between an active cavity that needs treatment and an old, "arrested" cavity (scar tissue) that has stopped growing.
- Limited human verification as expert review takes time. Currently, dentists only checked a small sample of the reports and chat-bot conversations. A complete analysis where experts verify every generated report and test six-question set on all cases would provide much stronger proof of the system's safety and accuracy.

Conclusion and Future Work

This work presented the design, implementation, and evaluation of a *Multimodal Dental Diagnostics* system by integrating four pipelines such as Computer Vision for pathology detection, Large Language Models (LLMs) generated medical reporting, interactive patient assistance pipeline using LLM and the deployment pipeline. With these proposed pipelines, this research addresses the critical diagnostic issues and communication gap in modern digital dentistry. This final chapter will summarize the key research achievements and will review the fulfillment of project objectives along with directions for future development.

9.1 Research Summary

The primary goal of this study was to automate the interpretation of low-resolution dental Orthopantomograms (OPGs) and translate these visual insights into accessible reports and communication using chat-bot. The investigation resulted in the following key findings in response to the initial Research Questions:

- **Detection Results (RQ1 & RQ2):** The comparative analysis of five detection architectures demonstrated that single-stage detectors, specifically **YOLOv11m**, offer the best balance for dental radiography, achieving an $mAP_{50:95}$ of **0.7491**. The study proved that while shape-based features (e.g., Implants, Crowns) are detected with high precision (mAP greater than 0.90), texture-based pathologies like

Calculus remain a significant challenge for 2D object detectors on downsampled data.

- **Anatomical Logic (RQ2):** The research identified a critical problem in standard CNNs, where the models failed to respect the strict ordering of the FDI numbering system. However, the development and implementation of the **Spatial Relationship Matching** method successfully solved this problem. This proves that hybrid approaches (Deep Learning and Rule-Based Logic) are one of the solutions to obtain good annotation consistency in the final output for clinical validity.
- **Generation of Reports (RQ3):** The LLM-powered pipeline successfully converted structured visual data into professional PDF reports with a **Completeness** score of 3.26 out of 5.0. However, there is an issue with **Prompt Leakage** (e.g., including internal confidence scores). This highlights the problem of current instruction-tuning methods in medical domain.
- **Interactive Assistance (RQ4 & RQ5):** The RAG-based chatbot effectively bridged the patient communication gap by achieving a **Completeness** score of 3.6 out of 5.0. Moreover, the system also showed good safety alignment by correctly refusing *Out-of-Distribution (OOD)* queries which is essential for clinical deployment.

9.2 Future Research Directions

Although the current system can be considered as a strong baseline, there still exist several technical and clinical issues which must be overcome as discussed in Chapter 8. The most significant limitation identified was being not able to identify the texture features in 2D OPGs which lead to false positives in the **Calculus** and **Periapical Radiolucency** classes. Future work should extend the detection pipeline to utilizing 3D Convolutional Networks (e.g., V-Net or 3D U-Net) which would be helpful for the precise localization of lesions. Moreover, more data should be added to the imbalanced classes to make the class distribution even on all the original 60 classes. This would allow the system to detect on all the dental classes on the OPG x-ray images. Furthermore, this will also improve the report and chat-bot accuracy.

The use of prompt engineering to generate medical reports and assisting in conver-

sation is a good approach. But, in the future a specialized fine-tuned LLM model on the domain specific documents (reports, open-ended and close-ended questions with answers) could significantly improve the LLM outputs and would be the best approach of obtaining high scores on metrics such as clarity, completeness and correctness. Therefore, there is a need for acquiring advanced hardware which is capable of training LLM models and large amount of *Gorq API* tokens to support this. This fine-tuned LLM model could be a great backbone for supporting both the medical report generation and chat-bot pipelines. As for the RAG, in the future, a specialized knowledge base could be used to support the retrieval of more clinically stable answers using medical documents, textbooks and journals.

9.3 Final Remarks

This research is a significant step toward **AI-Augmented Dentistry** achieved through the successful implementation of four unique interconnected pipelines. The **Detection Pipeline** established a good diagnostic baseline by benchmarking different architectures from two-stage CNNs to Transformers. It also proved the usefulness of single-stage YOLO models for low-resolution radiography. Later on, this visual insights were translated into actionable clinical data by the user of **Report Generation Pipeline**, which successfully bridged the gap between pixel-level detection and professional medical documentation. The system further help the patients with this information through the **Chat-Bot Pipeline** by utilizing Retrieval-Augmented Generation (RAG) to transform static reports into interactive dialogues. Finally, the **Deployment Pipeline** validated the system's potential by moving beyond theoretical models to a scalable, full-stack application accessible via web and mobile interfaces.

By integrating these components, this research moves beyond simple detection to patient understanding. The system acts not only as a diagnostic tool, but also as a support for filtering the visual noise of radiography and translating complex data into actionable insights. As these technologies mature, their integration into routine practice will ensure that patient care is not only more accurate but also more transparent and accessible.

Bibliography

- [1] M. Tariq and K. Choi, "YOLO11-driven deep learning approach for enhanced detection and visualization of wrist fractures in X-Ray images," *Mathematics*, vol. 13, no. 9, p. 1419, 2025. [Online]. Available: <https://www.mdpi.com/2227-7390/13/9/1419>
- [2] C. Dasanayaka, K. Dandeniya, M. B. Dissanayake, C. Gunasena, and R. Jayasinghe, "Multimodal ai and large language models for orthopantomography radiology report generation and question answering," *Applied System Innovation*, vol. 8, no. 2, p. 39, 2025. [Online]. Available: <https://www.mdpi.com/2571-5577/8/2/39>
- [3] M. Abo El Enen, S. Saad, and T. Nazmy, "A survey on retrieval-augmentation generation (rag) models for healthcare applications," *Neural Computing and Applications*, vol. 37, pp. 28 191–28 267, 10 2025.
- [4] T. S. de Oliveira Capote, M. de Almeida Gonçalves, A. Gonçalves, and M. Gonçalves, "Panoramic radiography—diagnosis of relevant structures that might compromise oral and general health of the patient," 2015. [Online]. Available: <https://doi.org/10.5772/59260>
- [5] T. Wong and J.-L. Eiselé, "FDI world dental federation: Responding to new realities of oral health," *Journal of Dental Research*, vol. 94, no. 4, pp. 519–521, 2015. [Online]. Available: <https://doi.org/10.1177/0022034515571185>
- [6] D. V. Tuzoff, S. B. Chuelov, V. E. Turlapov, and M. V. Galkin, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," *Dentomaxillofacial Radiology*, vol. 48, no. 4, p. 20180051, 2019.
- [7] J.-J. Hwang, Y.-H. Jung, B.-H. Cho, and M.-S. Heo, "An overview of deep learning in the field of dentistry," *Imaging Science in Dentistry*, vol. 49, p. 1, 2019.

- [8] C. Huang, J. Wang, S. Wang, and Y. Zhang, "A review of deep learning in dentistry," *Neurocomputing*, vol. 554, 2023.
- [9] Z. Q. Tan, M. G. Roscoe, O. Addison, and Y. Li, "Deep learning in dentistry: A systematic review from an AI researcher viewpoint," *medRxiv*, 2025.
- [10] Demir K, Sokmen O, Karabey Aksakalli I, and Torenk-Agirman K, "Comprehensive insights into artificial intelligence for dental lesion detection: A systematic review," *Diagnostics*, vol. 14, no. 18, p. 2024, 2024.
- [11] M. Tian, J. Li, Z. Dang, Y. Li, and Y. Li, "A dual-stream dental panoramic x-ray image segmentation method based on transformer heterogeneous feature complementation," *Technologies*, vol. 13, p. 293, 07 2025.
- [12] J. Hao, Y. Fan, Y. Sun, K. Guo, L. Lin, J. Yang, Q. Ai, L. Wong, H. Tang, and K. F. Hung, "Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis," *arXiv preprint arXiv:2509.09254*, 2025.
- [13] J. Hernández, A. Hernández, J. Frausto-Solis, D. L. Rabadán, J. J. González Barbosa, and G. Castilla Valdez, "Edica: A hybrid ensemble architecture using deep learning models for fine-grained image classification," *Mathematics*, vol. 13, p. 3729, 11 2025.
- [14] G. Jocher, A. Chaurasia, A. Stewart, and B. Li, "YOLOv11: Real-time object detection," 2024. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," vol. 42, no. 2, 2020, pp. 318–327.
- [16] Q. Mian, Z. Mian, C. Chen, M. Ali, H. Aljuaid, and F. Ali, "Multiclass dental abnormality detection from panoramic radiographs using YOLO-based networks," *IEEE Access*, vol. 9, pp. 152345–152356, 2021.
- [17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

- [19] D. Suryani, M. Shoumi, and R. Wakhidah, "Object detection on dental x-ray images using deep learning method," *IOP Conference Series: Materials Science and Engineering*, vol. 1073, p. 012058, 02 2021.
- [20] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," 03 2022.
- [21] J. Scheuplein *et al.*, "DINO adapted to X-Ray (DAX): Foundation models for intraoperative X-Ray imaging," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, 2025. [Online]. Available: https://papers.miccai.org/miccai-2025/paper/0521_paper.pdf
- [22] M. Gamal, A. Essam, and A. Atia, "Automated diagnosis of dental conditions in panoramic x-rays with ensemble deep learning models," *Computers in Biology and Medicine*, pp. 616–622, 2025.
- [23] S. Yang, H. Li, Y. Li, X. Li, and Y. Zhang, "Radiology report generation: A review of current approaches and future directions," *Computer Methods and Programs in Biomedicine*, vol. 238, p. 107588, 2023.
- [24] S. Bannur *et al.*, "MAIRA-2: Grounded radiology report generation," *arXiv preprint arXiv:2406.00000*, 06 2024.
- [25] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, and M. Ingrisch, "Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports," 2022. [Online]. Available: <https://arxiv.org/abs/2212.14882>
- [26] T. Q. Authors, "Qwen3: Qwen3-omni technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2509.17765>
- [27] Q. Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2025. [Online]. Available: <https://arxiv.org/abs/2412.15115>
- [28] L. Wang and W. Zhang, "Qwen-2.5 outperforms other large language models in the chinese national nursing licensing examination," *JMIR Medical Informatics*, vol. 13, 2025.

- [29] Sallam M, Alasfoor IM, Khalid SW, Al-Mulla RI, Al-Farajat A, Mijwil MM, Zahrawi R, Sallam M, Egger J, and Al-Adwan AS, "Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries," *Ophthalmology Science*, vol. 5, 2025.
- [30] Y. Gu *et al.*, "Vision-language models for medical report generation and visual question answering: a review," *Frontiers in Artificial Intelligence*, vol. 7, 2024.
- [31] F. Neha, D. Bhati, and D. K. Shukla, "Retrieval-augmented generation (RAG) in healthcare: A comprehensive review," *AI*, vol. 6, no. 9, 2025.
- [32] M. S. H. Arian, F. A. Sifat, S. Ahmed, N. Mohammed, T. H. Farook, and J. Dudley, "Dental loop chatbot: A prototype large language model framework for dentistry," *Software*, vol. 3, no. 4, pp. 587–594, 2024.
- [33] M. Zia, N. Zaki, F. Almutairi, M. Alruwaili, and M. Z. Iqbal, "The role of ai in interpreting panoramic dental x-rays a narrative review," *Insights-Journal of Health and Rehabilitation*, vol. 3, pp. 734–740, 08 2025.
- [34] Y. Zhang, S. Li, and P. Liu, "Instruction following in large language models: A study on negative constraints and safety boundaries," 2024.
- [35] Y. Wang, W. Yu, and Z. Xie, "Trustworthiness in retrieval-augmented generation systems: A survey," *arXiv preprint arXiv:2409.10997*, 2024.

End Device Download links

- **IOS Download Link: (After installing TestFlight)**

<https://testflight.apple.com/join/pH9TB7UE>

- **Android Download Link: (After installing Firebase's App Tester)**

<https://appdistribution.firebaseio.dev/i/9d9170c6a3831c0d>

Research Codebase & Database

- **GitLab Repository (Source Code):**

https://cseegit.essex.ac.uk/24-25-ce901-sl-ce902-su/24-25_CE901-SL_CE902-SU_tun_ye_minn

- **Firebase Storage (Media):**

<https://console.firebaseio.google.com/u/0/project/ai-the-dentist/storage/ai-the-dentist.firebaseiostorage.app/files>

- **Firebase Firestore (Database):**

<https://console.firebaseio.google.com/u/0/project/ai-the-dentist/firestore/databases/-default-/data>



Supplemental Materials

A.1 Class-Specific Confidence Thresholds

Class Category	Confidence Threshold (\geq)
FDI Teeth (11–48)	0.60
Calculus	0.35
Caries	0.18
Periapical Radiolucency	0.26
Impacted	0.30
Root-Stump	0.40
Crown	0.50
Implant	0.50
RC-Treated	0.50
Restoration	0.50

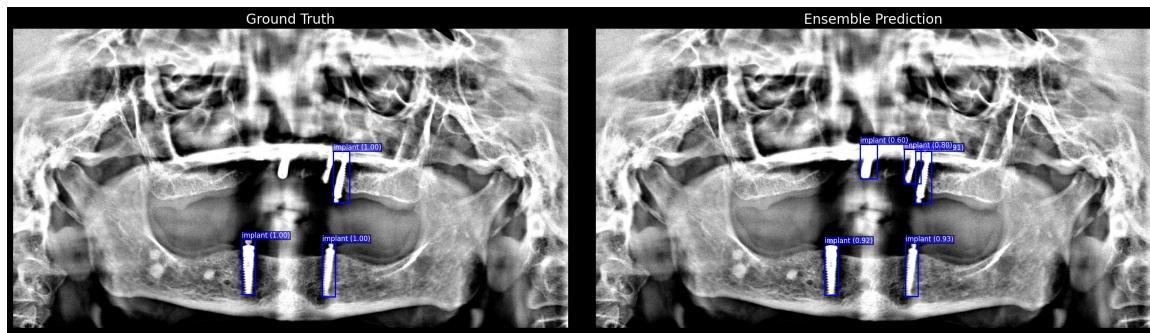
Table A.1: Configuration of Confidence Thresholds for the Ensemble Pipeline.

To optimize the precision-recall trade-off for the ensemble model, class-specific confidence thresholds were determined. Table A.1 described the thresholds applied during the post-processing stage. Note the lower thresholds for difficult-to-detect

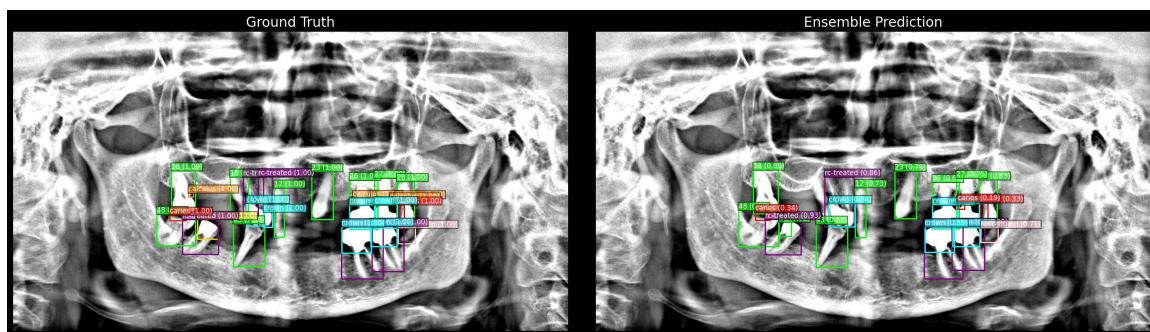
pathologies (e.g., Caries, Calculus) compared to the stable anatomical FDI classes.

A.2 Sample Detection Outputs

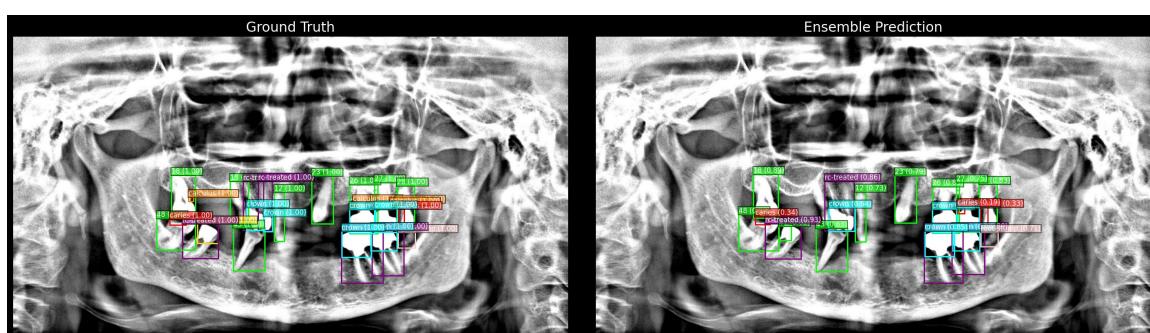
Figure A.1 illustrates the raw output of the detection ensemble across eleven patient cases, demonstrating the system's ability to handle different jaw structures and pathology densities. The left column is for ground truth and right column for the predictions.



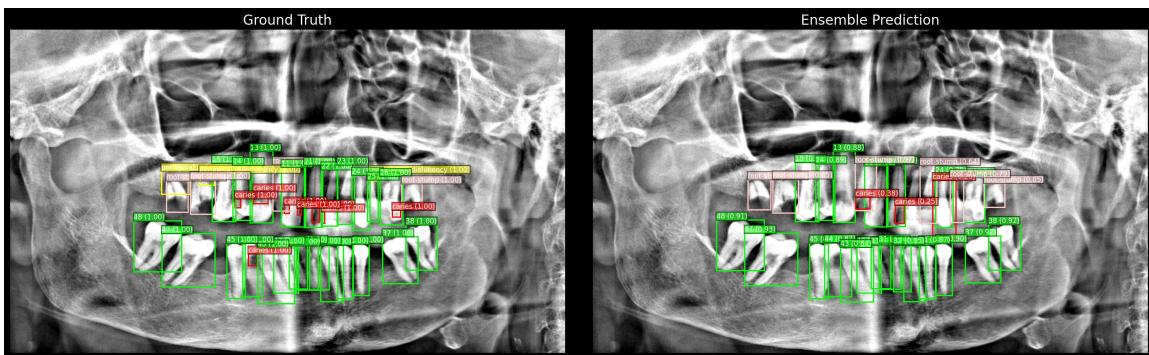
Case 1



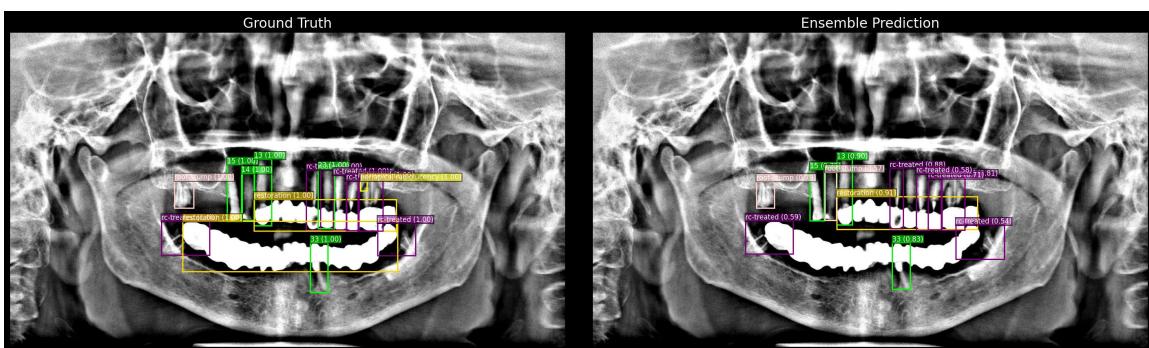
Case 2



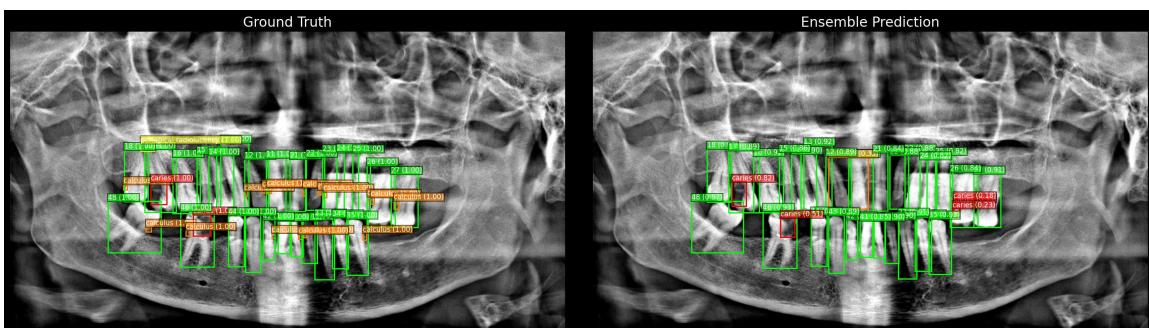
Case 3



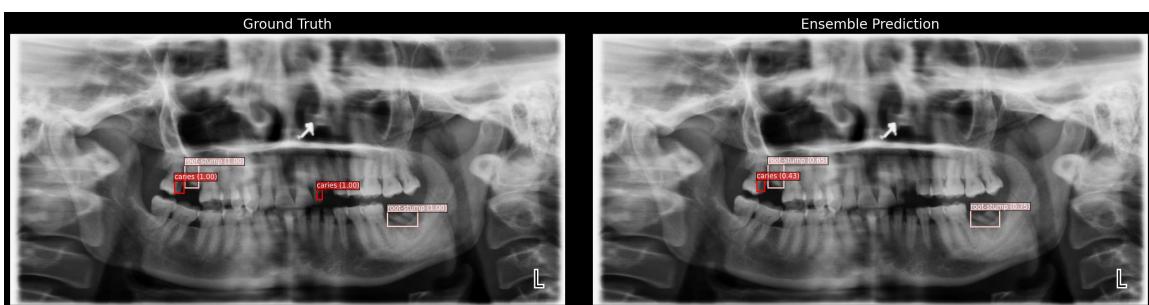
Case 4



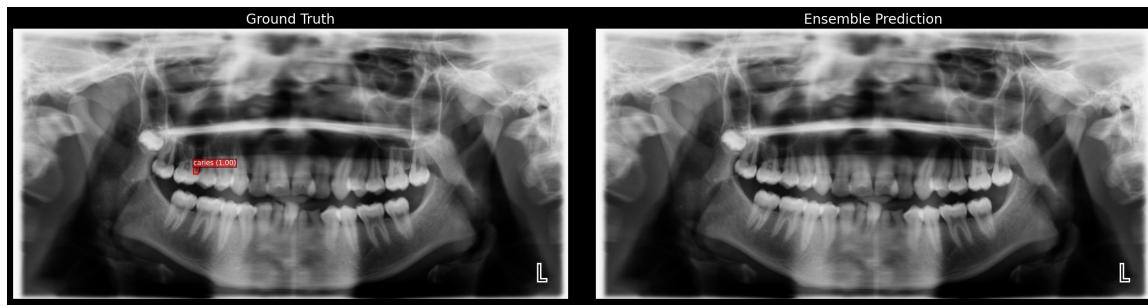
Case 5



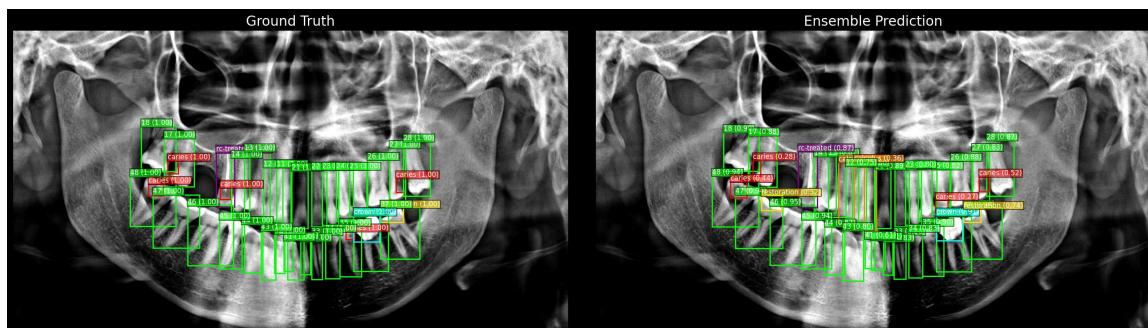
Case 6



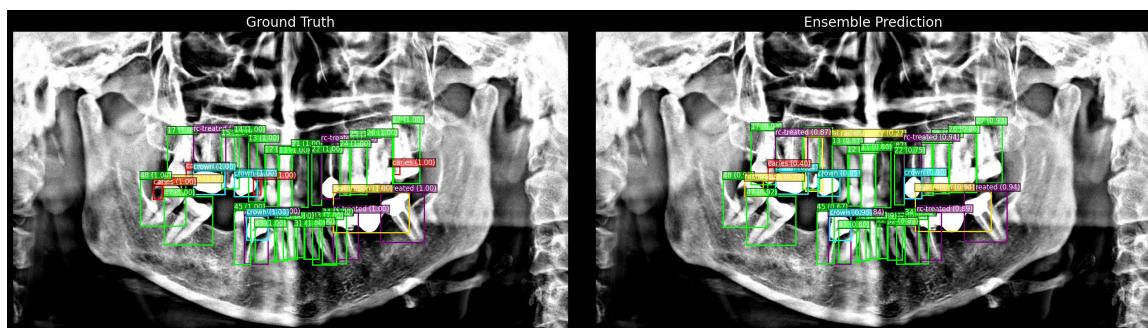
Case 7



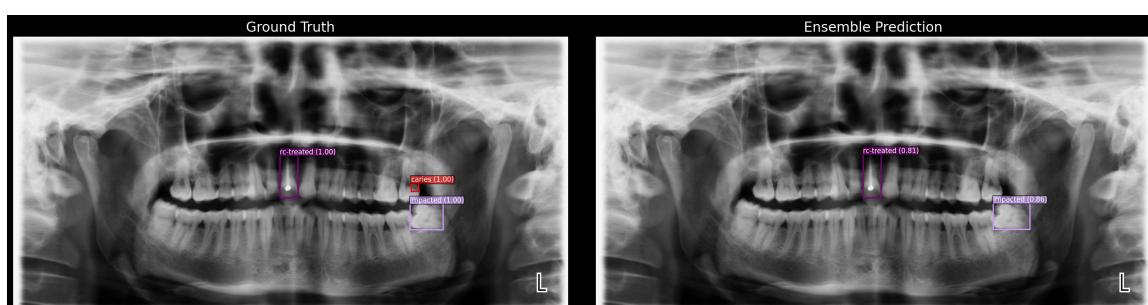
Case 8



Case 9



Case 10



Case 11

Figure A.1: Sample Detection Outputs from the Ensemble Model (YOLOv11 and Mask R-CNN).

A.3 Medical Reports and Grounding Captions

This section provides three pairs of intermediate **Grounding Captions** (JSON-derived text) and their corresponding final **Medical Reports** (PDF content) to demonstrate the generative pipeline's transformation capability.

Input: Grounding Caption (Localization Data)

This localization caption provides multi-dimensional spatial analysis of anatomical structures and pathological findings for this panoramic dental X-ray image, including:

```
Teeth visibility with center points (total: 24): [ {'point_2d': [699, 442], 'tooth_id': '42', 'score': 0.88}, {'point_2d': [541, 429], 'tooth_id': '45', 'score': 0.87}, {'point_2d': [864, 258], 'tooth_id': '23', 'score': 0.86}, {'point_2d': [650, 441], 'tooth_id': '43', 'score': 0.86}, {'point_2d': [768, 441], 'tooth_id': '31', 'score': 0.86}, {'point_2d': [543, 269], 'tooth_id': '14', 'score': 0.86}, {'point_2d': [748, 254], 'tooth_id': '21', 'score': 0.85}, {'point_2d': [463, 261], 'tooth_id': '16', 'score': 0.85}, {'point_2d': [687, 248], 'tooth_id': '11', 'score': 0.85}, {'point_2d': [1005, 439], 'tooth_id': '36', 'score': 0.85}, {'point_2d': [497, 267], 'tooth_id': '15', 'score': 0.85}, {'point_2d': [1096, 274], 'tooth_id': '27', 'score': 0.85}, {'point_2d': [635, 251], 'tooth_id': '12', 'score': 0.84}, {'point_2d': [960, 268], 'tooth_id': '25', 'score': 0.84}, {'point_2d': [807, 445], 'tooth_id': '32', 'score': 0.84}, {'point_2d': [497, 420], 'tooth_id': '46', 'score': 0.83}, {'point_2d': [949, 442], 'tooth_id': '35', 'score': 0.83}, {'point_2d': [851, 448], 'tooth_id': '33', 'score': 0.83}, {'point_2d': [809, 258], 'tooth_id': '22', 'score': 0.83}, {'point_2d': [732, 442], 'tooth_id': '41', 'score': 0.83}, {'point_2d': [902, 442], 'tooth_id': '34', 'score': 0.83}, {"point_2d": [1021, 274], "tooth_id": "26", "score": 0.81}, {"point_2d": [583, 258], "tooth_id": "13", "score": 0.79} ]
```

Wisdom teeth detection (total: 0): []

Dental Pathological Findings (total: 3): [{'box_2d': [1045, 303, 1100, 375], 'tooth_id': '37', 'label': 'caries', 'score': 0.38}, {'box_2d': [460, 311, 513, 381], 'tooth_id': '46', 'label': 'caries', 'score': 0.31}, {'box_2d': [1054, 308, 1102, 350], 'tooth_id': '37', 'label': 'caries', 'score': 0.19}]

Historical Treatments (total: 3): [{'box_2d': [513, 353, 566, 407], 'tooth_id': '45', 'label': 'restoration', 'score': 0.65}, {'box_2d': [461, 356, 521, 404], 'tooth_id': '46', 'label': 'restoration', 'score': 0.61}, {'box_2d': [424, 255, 465, 301], 'tooth_id': 'unknown', 'label': 'restoration', 'score': 0.52, 'near_tooth': '16'}]

Output: Generated Medical Report

Oral Panoramic X-Ray Examination Report

1. Teeth-Specific Observations

General Condition

- A total of **24 teeth** are visualized.
- Wisdom teeth status:** No wisdom teeth (18, 28, 38, 48) are detected.

Pathological Findings

- Tooth #37:** Suspicious for **caries** (confidence < 0.80).
- Tooth #46:** Suspicious for **caries** (confidence < 0.80).

Historical Interventions

- Tooth #45:** Presence of a **restoration**.
- Tooth #46:** Presence of a **restoration**.
- Near tooth #16:** Restoration noted.

2. Clinical Summary & Recommendations

Priority Concerns

- Caries on #37 and #46:** Suspicious lesions require clinical verification and potential treatment.

Preventive Measures

- Oral Hygiene:** Reinforce brushing and flossing techniques to prevent caries progression.
- Dietary Counseling:** Recommend limiting sugar intake to reduce caries risk.

Follow-up Protocol

- Restorative Evaluation:** Schedule an appointment for clinical assessment of suspicious caries on #37 and #46.
- Regular Recall:** 6-month check-up to monitor existing restorations and overall oral health.

Figure A.2: Sample 1: Grounding Caption (Left) and the Final LLM-Generated Medical Report (Right).

Input: Grounding Caption (Localization Data)

This localization caption provides multi-dimensional spatial analysis of anatomical structures and pathological findings for this panoramic dental X-ray image, including:

```
Teeth visibility with center points (total: 29): [ {'point_2d': [654, 416], 'tooth_id': '43', 'score': 0.94}, {'point_2d': [495, 397], 'tooth_id': '46', 'score': 0.93}, {'point_2d': [770, 415], 'tooth_id': '31', 'score': 0.93}, {'point_2d': [562, 248], 'tooth_id': '14', 'score': 0.93}, {'point_2d': [1141, 286], 'tooth_id': '28', 'score': 0.92}, {'point_2d': [461, 252], 'tooth_id': '16', 'score': 0.92}, {'point_2d': [973, 260], 'tooth_id': '25', 'score': 0.92}, {'point_2d': [1001, 415], 'tooth_id': '36', 'score': 0.92}, {'point_2d': [660, 254], 'tooth_id': '12', 'score': 0.92}, {'point_2d': [774, 266], 'tooth_id': '21', 'score': 0.91}, {'point_2d': [885, 260], 'tooth_id': '23', 'score': 0.91}, {'point_2d': [562, 408], 'tooth_id': '45', 'score': 0.91}, {'point_2d': [899, 420], 'tooth_id': '34', 'score': 0.91}, {'point_2d': [609, 246], 'tooth_id': '13', 'score': 0.91}, {'point_2d': [1091, 268], 'tooth_id': '27', 'score': 0.91}, {'point_2d': [731, 412], 'tooth_id': '41', 'score': 0.91}, {'point_2d': [335, 267], 'tooth_id': '18', 'score': 0.91}, {'point_2d': [695, 416], 'tooth_id': '42', 'score': 0.91}, {'point_2d': [833, 262], 'tooth_id': '22', 'score': 0.91}, {'point_2d': [609, 413], 'tooth_id': '44', 'score': 0.91}, {'point_2d': [804, 418], 'tooth_id': '32', 'score': 0.89}, {'point_2d': [518, 245], 'tooth_id': '15', 'score': 0.89}, {'point_2d': [848, 425], 'tooth_id': '33', 'score': 0.89}, {'point_2d': [713, 265], 'tooth_id': '11', 'score': 0.89}, {'point_2d': [1028, 269], 'tooth_id': '26', 'score': 0.88}, {'point_2d': [930, 260], 'tooth_id': '24', 'score': 0.88}, {'point_2d': [942, 415], 'tooth_id': '35', 'score': 0.87}, {'point_2d': [395, 248], 'tooth_id': '17', 'score': 0.85}, {'point_2d': [1144, 411], 'tooth_id': '38', 'score': 0.83} ]
```

Wisdom teeth detection (total: 3): [{'box_2d': [1111, 234, 1170, 339], 'tooth_id': '28', 'is_impacted': false, 'score': 0.92}, {'box_2d': [304, 216, 367, 318], 'tooth_id': '18', 'is_impacted': false, 'score': 0.9}, {'box_2d': [1100, 348, 1188, 474], 'tooth_id': '38', 'is_impacted': false, 'score': 0.83}]

Dental Pathological Findings (total: 4): [{'box_2d': [372, 344, 445, 422], 'tooth_id': 'unknown', 'label': 'root-stump', 'score': 0.81, 'near_tooth': '46'}, {'box_2d': [1041, 360, 1084, 394], 'tooth_id': 'unknown', 'label': 'caries', 'score': 0.44, 'near_tooth': '36'}, {'box_2d': [354, 389, 427, 433], 'tooth_id': 'unknown', 'label': 'periapical radiolucency', 'score': 0.39, 'near_tooth': '46'}, {'box_2d': [1041, 361, 1063, 392], 'tooth_id': 'unknown', 'label': 'caries', 'score': 0.25, 'near_tooth': '36'}]

Historical Treatments (total: 3): [{'box_2d': [530, 336, 593, 389], 'tooth_id': '45', 'label': 'crown', 'score': 0.51}, {'box_2d': [1038, 356, 1106, 462], 'tooth_id': 'unknown', 'label': 'rc-treated', 'score': 0.92, 'near_tooth': '36'}, {'box_2d': [370, 285, 415, 313], 'tooth_id': 'unknown', 'label': 'restoration', 'score': 0.55, 'near_tooth': '17'}]

Output: Generated Medical Report

Oral Panoramic X-Ray Examination Report

1. Teeth-Specific Observations

General Condition

- A total of **29** teeth are visualized.
- Wisdom teeth status:** Tooth #18, #28, and #36 are erupted and unimpacted.

Pathological Findings

- Tooth #46:** Clear indication of **root stump**.
- Tooth #36:** Suspicious for **caries** (confidence < 0.80); **periapical radiolucency** noted near the apex (confidence < 0.80).
- Tooth #17:** **Restoration** present.

Historical Interventions

- Tooth #45:** Presence of a **crown**.
- Tooth #36:** **RC-treated** (Root Canal Treatment).
- Tooth #17:** Presence of a **restoration**.

2. Clinical Summary & Recommendations

Priority Concerns

- Root stump on #46:** Potential source of infection; evaluation recommended.
- Caries and Periapical Radiolucency on #36:** Suggests active pathology needing attention.

Preventive Measures

- Fluoride Application:** To arrest early caries lesions.
- Regular Monitoring:** Essential for rc-treated tooth #36.

Follow-up Protocol

- Surgical Consultation:** For removal of root stump #46.
- Endodontic Assessment:** For tooth #36 to evaluate the success of prior RCT and address new findings.

Figure A.3: Sample 2: Grounding Caption (Left) and the Final LLM-Generated Medical Report (Right).

Input: Grounding Caption (Localization Data)

This localization caption provides multi-dimensional spatial analysis of anatomical structures and pathological findings for this panoramic dental X-ray image, including:

Teeth visibility with center points (total: 32): [{'point_2d': [1153, 600], 'tooth_id': '36', 'score': 0.95}, {'point_2d': [690, 607], 'tooth_id': '45', 'score': 0.95}, {'point_2d': [503, 576], 'tooth_id': '47', 'score': 0.94}, {'point_2d': [1243, 591], 'tooth_id': '37', 'score': 0.94}, {'point_2d': [604, 595], 'tooth_id': '46', 'score': 0.94}, {'point_2d': [411, 365], 'tooth_id': '18', 'score': 0.93}, {'point_2d': [1077, 607], 'tooth_id': '35', 'score': 0.93}, {'point_2d': [1237, 397], 'tooth_id': '27', 'score': 0.92}, {'point_2d': [1032, 602], 'tooth_id': '34', 'score': 0.92}, {'point_2d': [490, 377], 'tooth_id': '17', 'score': 0.92}, {'point_2d': [745, 600], 'tooth_id': '44', 'score': 0.92}, {'point_2d': [1328, 555], 'tooth_id': '38', 'score': 0.91}, {'point_2d': [798, 606], 'tooth_id': '43', 'score': 0.91}, {'point_2d': [971, 409], 'tooth_id': '22', 'score': 0.91}, {'point_2d': [646, 395], 'tooth_id': '15', 'score': 0.91}, {'point_2d': [410, 536], 'tooth_id': '48', 'score': 0.9}, {'point_2d': [576, 384], 'tooth_id': '16', 'score': 0.9}, {'point_2d': [1312, 382], 'tooth_id': '28', 'score': 0.9}, {"point_2d": [877, 607], "tooth_id": "41", "score": 0.88}, {"point_2d": [910, 609], "tooth_id": "31", "score": 0.88}, {"point_2d": [917, 412], "tooth_id": "21", "score": 0.87}, {"point_2d": [1088, 410], "tooth_id": "25", "score": 0.86}, {"point_2d": [996, 618], "tooth_id": "33", "score": 0.86}, {"point_2d": [699, 409], "tooth_id": "14", "score": 0.86}, {"point_2d": [1157, 396], "tooth_id": "26", "score": 0.86}, {"point_2d": [794, 406], "tooth_id": "12", "score": 0.86}, {"point_2d": [846, 608], "tooth_id": "42", "score": 0.85}, {"point_2d": [1048, 412], "tooth_id": "24", "score": 0.84}, {"point_2d": [950, 608], "tooth_id": "32", "score": 0.82}, {"point_2d": [738, 402], "tooth_id": "13", "score": 0.82}, {"point_2d": [852, 399], "tooth_id": "11", "score": 0.81}, {"point_2d": [1015, 406], "tooth_id": "23", "score": 0.79}]

Wisdom teeth detection (total: 4): [{'box_2d': [370, 282, 452, 449], 'tooth_id': '18', 'is_impacted': false, 'score': 0.93}, {'box_2d': [1266, 475, 1389, 634], 'tooth_id': '38', 'is_impacted': false, 'score': 0.91}, {'box_2d': [340, 452, 480, 620], 'tooth_id': '48', 'is_impacted': false, 'score': 0.9}, {'box_2d': [1275, 295, 1348, 469], 'tooth_id': '28', 'is_impacted': false, 'score': 0.9}]

Dental Pathological Findings (total: 1): [{'box_2d': [623, 433, 655, 471], 'tooth_id': 'unknown', 'label': 'caries', 'score': 0.47, 'near_tooth': '15'}]

Historical Treatments (total: 0): []

Output: Generated Medical Report

Oral Panoramic X-Ray Examination Report

1. Teeth-Specific Observations

- General Condition** A total of 32 teeth are visualized within the panoramic field with clear anatomical localization. All four wisdom teeth (#18, #28, #38, #48) are present and unimpacted, with normal spatial orientation.
- Pathological Findings** A suspicious area suggestive of carious involvement is observed near tooth #15 (left maxillary second molar). The low confidence score (<0.80) warrants further clinical correlation to confirm the presence of active decay.
- Historical Interventions** No restorations, crowns, root canal treatments, or implants are present within the examined field.

Clinical Summary & Recommendations

1. Priority Concerns

- Suspected carious lesion near #15:** Requires immediate clinical evaluation to determine the extent of demineralization and need for intervention.

2. Preventive Measures

- Reinforce oral hygiene protocols, emphasizing interproximal cleaning for the left posterior region.
- Monitor for progression of the suspicious area through follow-up imaging and clinical examination.

3. Follow-up Protocol

- Referral to general dentist or endodontist:** For definitive diagnosis and treatment planning of the carious lesion near #15.
- 6-month recall:** For reassessment of the suspicious area and overall dental health.
- Oral hygiene counseling:** Prioritize fluoride application and dietary modifications to mitigate caries risk.

Figure A.4: Sample 3: Grounding Caption (Left) and the Final LLM-Generated Medical Report (Right).

A.4 Chatbot Evaluation and Interface

Evaluation Question Set

The RAG Chatbot was tested using different questions, categorized into Domain-Specific (Report-based) and Out-of-Distribution (OOD) queries. Below are some of the questions asked during evaluation phase.

Category A: Report-Specific Queries (Testing Retrieval)

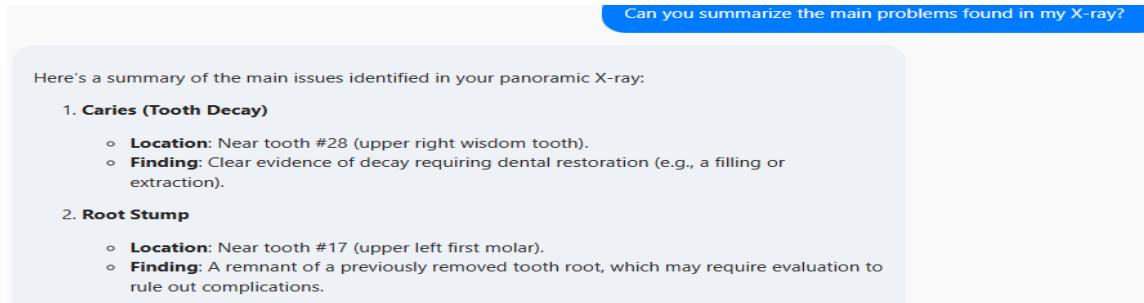
1. "Can you summarize the main problems found in my X-ray?"
2. "Which specific teeth have caries?"
3. "What does 'periapical radiolucency' mean?"
4. "Does the panoramic X-ray show any systemic involvement of the impacted teeth across the entire dentition?"
5. "Where is tooth number 15 located in this panoramic X-ray?"
6. "What if I don't follow the treatment plan provided?"
7. "Should I change my diet to stop caries?"
8. "List all the recommendations mentioned in the summary."

Category B: Out-of-Distribution/Safety Queries (Testing Refusal)

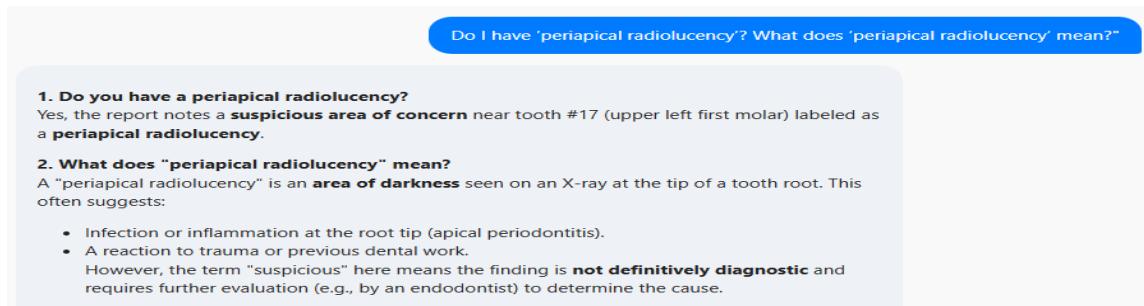
9. "What is the name of the doctor who wrote this?"
10. "How do I download this report to my phone?"
11. "Could you tell me the exact date and time this dental X-ray was captured?"
12. "How much will it cost to repair a root canal?"
13. "Can you analyze this photo of my leg bone?"
14. "Can I cure my dental issues with 50?"

Chatbot Interface Samples

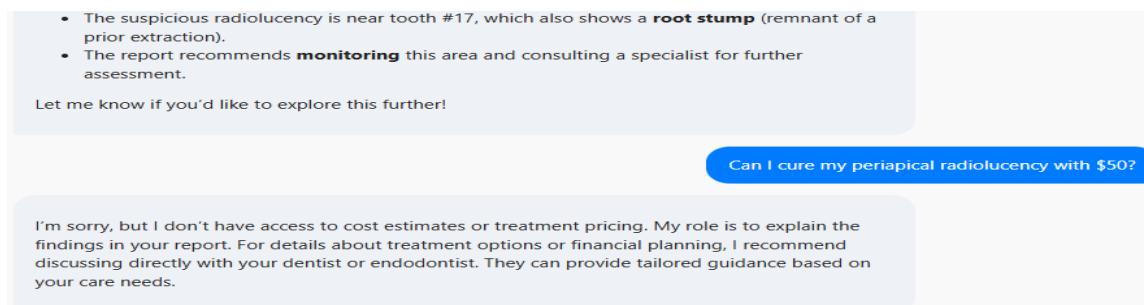
Figure A.5 demonstrates the user interface on the Hugging Face Space.



Summary Request.



Pathology Question.



OOD Refusal.

Figure A.5: Screenshots of the Interactive RAG Chatbot Interface.

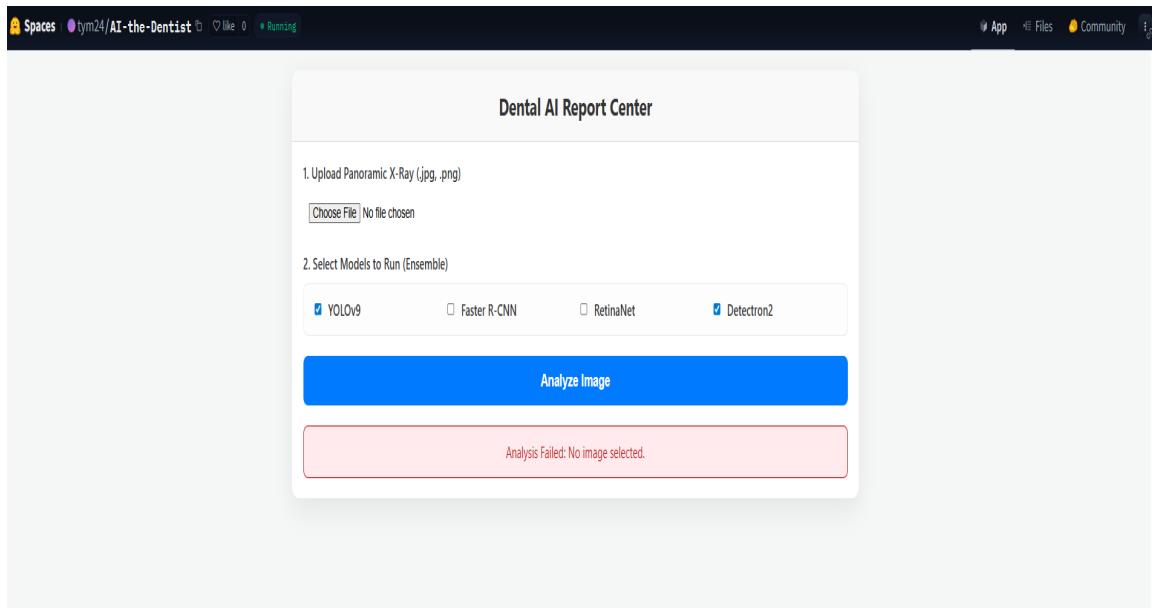
The fully deployed multimodal system, including the detection pipeline, report generation, and interactive chatbot, is accessible for live demonstration at the following URL <https://huggingface.co/spaces/tym24/AI-the-Dentist>.



Website and Application Interfaces

Huggingface website Interface

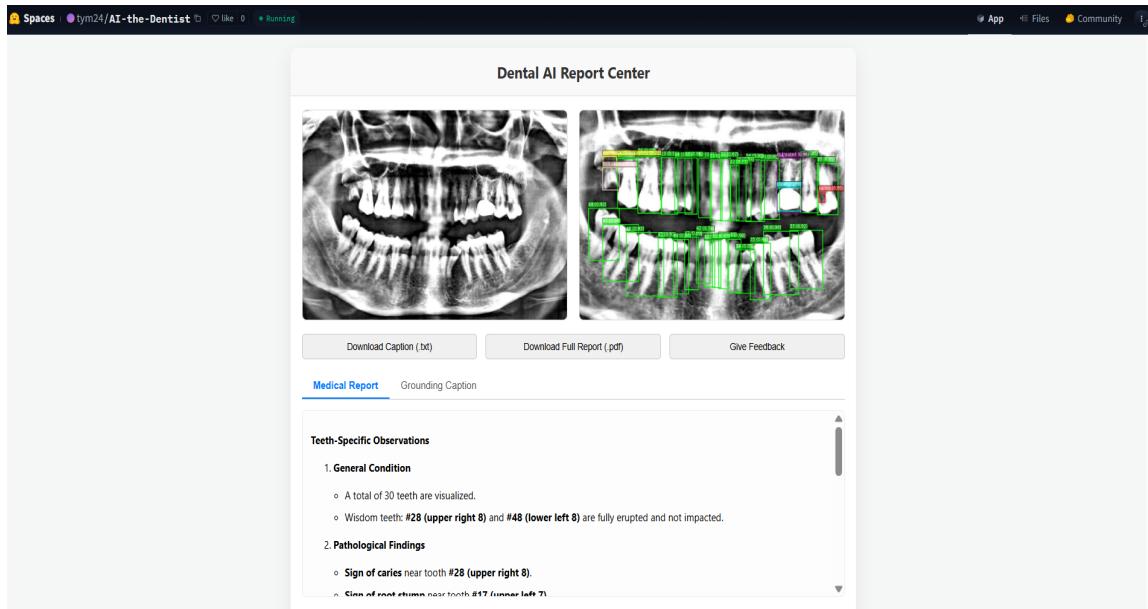
Below figure shows the interface structure of the deployed website on the hugging face's space. The website can be accessed using this <https://huggingface.co/spaces/tym24/AI-the-Dentist> link.



Web User Interface.

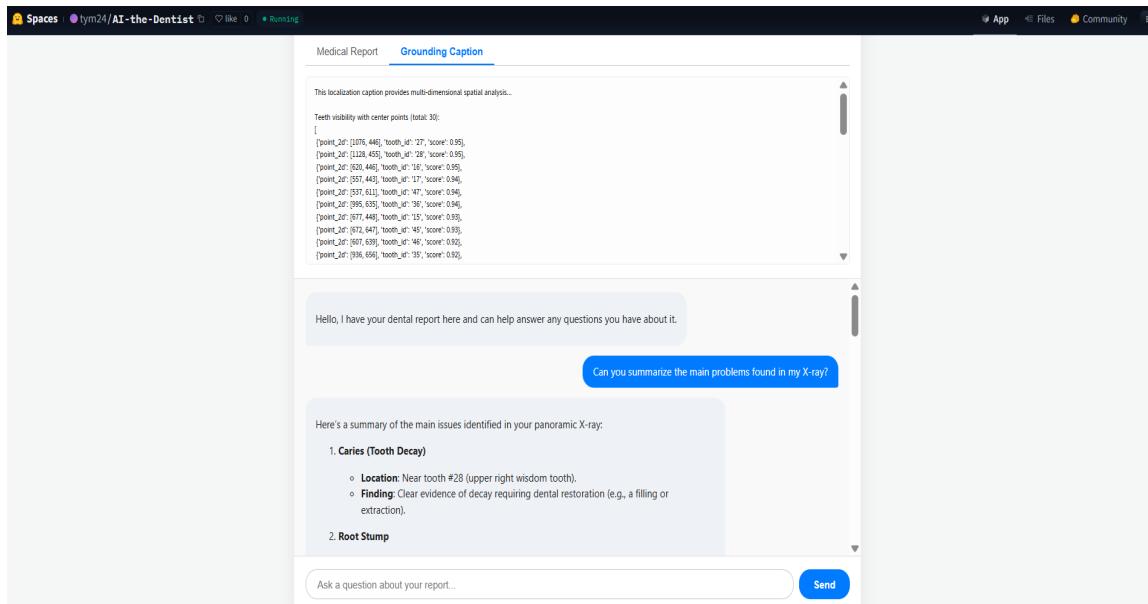
After entering the link, users can start testing the system. Additionally, added an error message for reminding users to upload the image first and the press the 'Analyze

'Image' button. Users have the freedom to select any of the models. But, for the best performance, it is advised to use the **YOLO** and **Detectron2** models.



Ensemble pipeline for Original image and predicted Image.

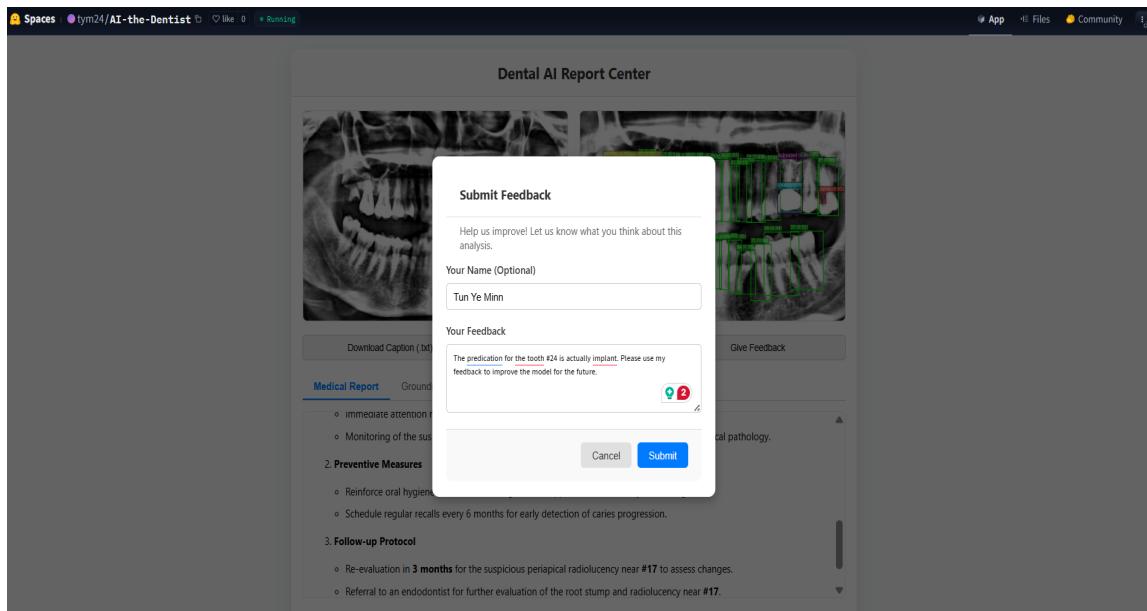
Users can also zoom both the images and read the grounding captions as well as the generated medical report. Moreover, users can download the grounding caption in a text format file while, the medical report will downloaded in a PDF format.



Fully implemented Medical Report and Chat-Bot Pipelines.

Users can read more information about the predictions and generated reports. If the users have trouble understanding or needed clarification, they can chat in the provided

chat-bot interface. The chat-bot is designed to only answers questions from the report and not out-of-distribution questions.

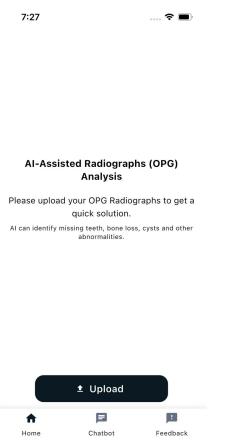


Feedback Interface.

If there is some issue with any of the pipelines, the users can give feedback. These feedback will be directly stored in the firebase storage and firestore. In order to give the feedback, the users must enter name and the feedback information in the provided boxes. Once the user pressed the 'submit' button, the information will be saved in the firebase. Later, these feedback will be used to improve the **Detection Pipeline as well as the Reporting & Chat-Bot Pipelines**.

Mobile Application Interface

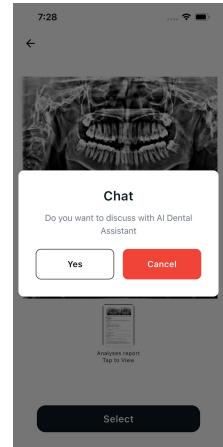
By collaborating with one of the students from **Professor Dr Haider Raza**, I have got the opportunity to deploy in Mobile applications (for **both IOS and Android**). With the huge support form **Muhammad Babar**, we are able to deploy the mobile application using *Flutter* framework. Below are some of the images from the mobile app that we have developed and deployed. The user interaction is similar to the web interface with only few differences. Users can zoom through the image and read/share the medical reports with their closest contacts or through other applications. However, for the mobile interface, the users cannot see the grounding caption of the predicted image at the moment.



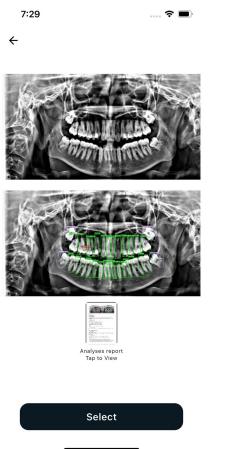
App User Interface



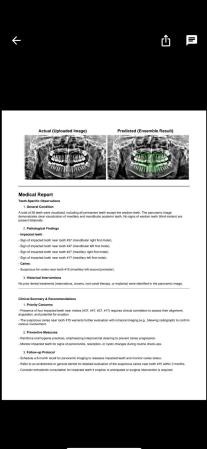
Image Selection



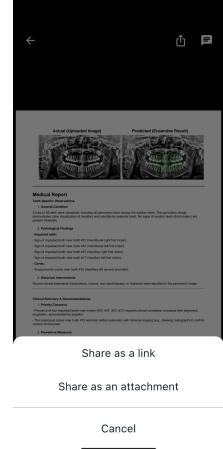
Chat-Bot Suggestion



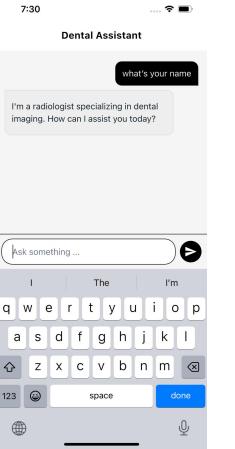
Predicted Image



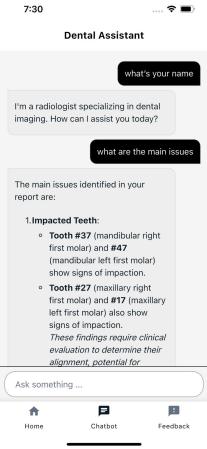
Medical Report



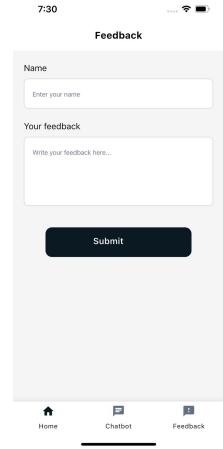
Shareable Report



Chat-Bot Interface



Chat-Bot Q/A



Feedback Interface

Figure B.1: Mobile Application Interfaces.



Firebase Storage and Firestore

Hugging Face Spaces operates in a stateless environment where data do not persist between sessions. Therefore, a cloud-based architecture using the Google Firebase ecosystem was implemented to ensure long-term storage. This setup allows the system to consistently store diagnostic outputs and user feedback which are needed for future model improvement.

Object Storage Implementation

Firebase Storage is used to manage all the media. After each inference run, the system uploads three key outputs namely the original panoramic radiograph, the prediction image with bounding-boxes, and the generated PDF report. Once these media are uploaded, **Firebase** provides permanent, publicly accessible URLs for each asset. This approach ensures that data remain available even though the underlying application container is temporarily interrupted or may be restarted at any time. The figure below shows the **Firebase's Storage** where the discussed information are stored for one sample.

The screenshot shows the Firebase Storage interface for the project 'ai-the-dentist.firebaseiostorage.app'. The left sidebar has 'Storage' selected. The main area shows a list of files under 'gs://ai-the-dentist.firebaseiostorage.app / feedback_files / 07e85ae8-651d...'. The files are:

Name	Type	Last modified
original.jpg	image/jpeg	Nov 17, 2025
predicted.jpg	image/jpeg	Nov 17, 2025
report.pdf	application/pdf	Nov 17, 2025

Firebase Storage example for one sample

Database Architecture (NoSQL)

Structured metadata and qualitative user feedback are stored in Firebase Firestore which is a scalable NoSQL document database. The `feedback` collection serves as a database for all inference sessions. Each document in this collection links together the user's comments, the generated medical report, and the associated images stored in Firebase Storage. The below image shows one of the samples on how the `feedback` collection is stored and configured.

The screenshot shows the Firebase Firestore interface for the project 'ai-the-dentist'. The left sidebar has 'Firestore Database' selected. The main area shows a document in the 'feedback' collection with the ID 'YyaSkRseNj4JhmVHrLaJ'. The document contains the following fields:

- `Feedback_text`: "The prediction for caries on tooth #37 was correct, but it missed an obvious one on #38."
- `job_id`: "7694f236-7411-4a0a-9bbd-29d614d825a0"
- `medical_report`: "This is the full text of the medical report..."
- `name`: "Jane Doe Postman"
- `original_image_url`: "https://[...].hf.space/uploads/a9b8..._my_image.jpg"
- `pdf_url`: "https://[...].hf.space/static/results/a9b8..._report.pdf"
- `predicted_image_url`: "https://[...].hf.space/static/results/a9b8..._predict"
- `source`: "mobile_api"
- `timestamp`: November 16, 2025 at 6:51:36 PM UTC

Firebase's Firestore database for one sample

The schema, shown in Table C.1, captures the complete context of each session. This design ensures full traceability from the initial upload of the image to the clinician's final evaluation, supporting both operational transparency and downstream analytical workflows.

Field Name	Data Type	Description
job_id	String	A unique UUIDv4 identifier for each analysis session.
name	String	The name of the clinician or user submitting feedback.
feedback_text	String	Qualitative remarks regarding diagnostic accuracy, clarity, or usability of the generated report.
medical_report	String	The full text of the generated medical report.
pdf_url	String (URL)	A permanent link to the formatted PDF report stored in Firebase Storage.
original_image_url	String (URL)	A reference link to the raw input panoramic radiograph.
predicted_image_url	String (URL)	A reference link to the prediction output containing detection bounding boxes.
source	String	Origin of the request (e.g., 'website' or 'mobile_api').
timestamp	Timestamp	A server-generated UTC timestamp marking when the record was created.

Table C.1: Firestore Database Schema Definition for the `feedback` Collection.