

國立臺灣大學醫學院口腔生物科學研究所



碩士論文

Graduate Institute of Oral Biology

College of Medicine

National Taiwan University

Master's Thesis

透過人工智慧模型判讀牙科環口影像

Assessing Dental Panoramic Images with Artificial
Intelligence

王盈智

Yin-Chih Wang

指導教授: 周涵怡 博士
陳敏慧 博士

Advisors: Han-Yi E. Chou, Ph.D.
Min-Huey Chen, Ph.D.

中華民國 114 年 6 月

June, 2025

國立臺灣大學碩士學位論文

口試委員會審定書

透過人工智慧模型判讀牙科環口影像

Assessing Dental Panoramic Images with Artificial
Intelligence

本論文係王盈智君 (R11450008) 在國立臺灣大學口腔生物科學所完成之碩士學位論文，於民國 113 年 7 月 18 日承下列考試委員審查通過及口試及格，特此證明

口試委員： 周涵任 陳政慧

(指導教授)

王偉仲 蔡孝誠

所 長： 鄭世榮





Acknowledgements

首先要感謝我的指導教授周涵怡老師以及陳敏慧老師，周涵怡老師在研究所期間提供我各種機會和幫助，不僅在研究上幫助我，也在精神上鼓勵我，跟著老師能獲得滿滿的愛和關心。陳敏慧老師在我還沒進入研究所之前就願意了解我的研究計劃，讓我能順利拿到珍貴的研究資料，更鼓勵我來推甄，也在入學後持續關心我的各種需求，我很幸運能在碩士期間遇見您們。

感謝蔡孝誠院長，從我剛牙醫系畢業以來，就在臨床上教我很多知識以及技巧，並發想現在的研究主題，和鼓勵我來台大念研究所，才能順利有現在的研究跟結果。

謝謝我爸媽從小在我身上投入的資源還有對我的教育，即使我喜歡走跟別人不同的路，又有很多奇奇怪怪的想法，也從來沒有反對，讓我能自由自在的做所有我想做的事。也謝謝我的弟弟把家裡兩隻貓照顧得很好。還有我的好朋友們，聽我分享各種心情大小事，和我一起吃飯到處玩耍，是你們豐富了我的時光。

最後要謝謝我的人生伴侶 Harry，在研究上提供技術支持也一起合作之外，還在我心情不好的時候想辦法逗我開心，晚診下班時煮好喝的湯等我回家，參與我大大小小的活動跟興趣，一起照顧愛護貓貓們。沒有你，不僅無法產出這份研究，也沒有這個幸福快樂的我。I like me better when I am with you.





摘要

近年來智慧醫療領域發展非常迅速，包含醫學影像的電腦輔助分析，並且也擴展到各種的牙科影像，其中牙科環口影像，或稱為全口 X 光，是包含口腔區域軟硬組織的一張 X 光片圖像。在醫院或牙科診所中，它經常被用作評估患者牙齒狀況和促進溝通的前線工具，使牙醫能夠迅速了解患者的牙齒健康狀況。它有助於早期檢測口腔疾病，從而導致更好的預後，並幫助診斷和規劃治療。此外，全口 X 光還能夠幫助牙醫在治療之前更好地與患者溝通，解釋治療方案和預期結果。本研究主要使用人工智慧模型來自動判讀牙科環口影像，幫助牙醫師在日常臨床工作中有更高的效率。

雖然目前有關牙科環口影像結合人工智慧的研究正蓬勃發展，但大多數的研究目標都集中在單一問題上，例如偵測牙齒、牙齒編號或特定單一發現。有些研究試圖判讀整個影像，但未能有效地融合牙科知識，因此難以正確結合牙齒編號和影像中的牙科相關發現。此外，儘管有幾項研究與牙醫師進行了比較，但也是專注在單一的項目上，並未解決像此研究這樣複雜的問題。最後，大多數研究使用的是相對較小的單一來源資料。

本研究中的人工智慧系統除了偵測牙齒以及其對應之編號外，也檢測臨床發現，如植體、殘根、牙冠/牙橋、根管充填物、補牙、齲齒和根尖囊腫。並使用來自多個國家以及機構的牙科全口 X 光資料，使模型能夠達到高精度並融入牙科知

識。結果顯示，我們的模型經過訓練和測試，能夠在不同的資料集中都準確地識別和分類各種牙科病變及狀況，從而提供詳細的診斷信息。除此之外，透過比較我們的深度學習模型與執業牙醫師的表現，模型的表現與執業牙醫不相上下，並在某些項目中優於牙醫師。

此人工智慧系統可以在日常臨床工作中協助牙醫，減少診斷時間並提高診斷準確性。此外，該模型還具有擴展潛力，可以加入更多牙科臨床發現項目在影像上，或甚至應用於其他醫學影像，進一步提升醫療診斷和治療的效率和效果。總體來說，我們的研究展示了人工智慧在牙科影像分析中的巨大潛力，並為未來的臨床應用提供了堅實的基礎。

關鍵字：牙科環口影像、牙科全景 X 光片、人工智慧、深度學習、智慧醫療



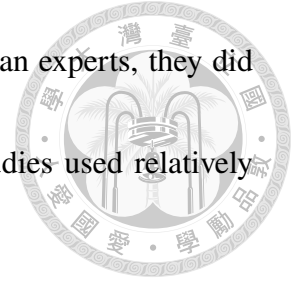
Abstract

Computer-assisted analysis of medical images, particularly through deep learning and artificial intelligence, has rapidly advanced in recent years. These methods have expanded to dental imaging, including dental panoramic images, periapical films, and cone-beam computed tomography, making deep learning a promising approach for dental image analysis.

A dental panoramic radiograph, or orthopantomogram, is an X-ray scan of the oral area. In dental clinics, it is often a frontline tool for assessing patients' dental status and facilitating communication. It helps detect oral diseases early, leading to better prognoses, and aids in diagnosing and planning treatments.

While there has been booming research involving dental panoramic radiographs and artificial intelligence, most of them focus on single issues like tooth detection, numbering, or specific findings. Some attempts to summarize the entire image failed to incorporate dental knowledge effectively, struggling to combine the correct tooth numbers with find-

ings. Additionally, while several studies compared results with human experts, they did not address complex problems like ours. Lastly, many of these studies used relatively small, single-source datasets.



In this work, we provide an end-to-end solution for summarizing dental panoramic images. This involves identifying all teeth and their indices and detecting clinical findings such as tooth number, implants, residual roots, crowns/bridges, root canal fillings, fillings, caries, and periapical radiolucency. Our dataset, sourced from multiple institutions across different countries and ethnicities, allowed our model to achieve high accuracy and incorporate dental knowledge. Comparing our deep learning model's performance with practicing dentists showed comparable generalizability and accuracy. The AI system we developed can assist and pre-warn dentists in their daily clinical work and has the potential to be expanded to include more findings on dental panoramic radiographs or other medical images in the future.

Keywords: Dental panoramic radiographs, Dental Panoramic images, Dental findings, Artificial intelligence, Deep learning, Intelligence medicine



Contents

	Page
Verification Letter from the Oral Examination Committee	i
Acknowledgements	iii
摘要	v
Abstract	vii
Contents	ix
List of Figures	xiii
List of Tables	xv
Chapter 1 Introduction	1
1.1 Background	1
1.2 Dental Panoramic Radiographs	2
1.3 Modeling in Medical Imaging	3
1.4 Modeling and Bottleneck in Dental Panoramic Radiographs	3
1.5 A Proper Dental AI Model	4
Chapter 2 Related Works	5
2.1 Tooth Detection	5
2.2 Dental Knowledge Incorporation and Specific Applications	6
2.2.1 Tooth Numbering	6

2.2.2	Binary Labeling	6
2.2.3	Specialized Dental Applications	7
2.3	Dental Finding Detection	7
2.4	Comparison Against Human	8
2.5	Objective of Work	9
Chapter 3	Materials and Methods	11
3.1	Datasets	11
3.1.1	Promaton Dataset	11
3.1.2	National Taiwan University Hospital Dataset	12
3.2	Data Characteristics	13
3.3	Data Preprocessing	13
3.3.1	Data Split	13
3.3.2	Data Annotation	14
3.3.2.1	Promaton Dataset Annotation	14
3.3.2.2	National Taiwan University Hospital Dataset Annotation	14
3.3.2.3	Human Benchmark for AI Comparison	15
3.4	AI System for Finding Summarization	16
3.4.1	Semantic Segmentation Stage	16
3.4.2	Instance Segmentation Stage	17
3.4.3	Combination Stage	17
3.5	Evaluation Metrics	18
3.6	Hardware	21
Chapter 4	Results	23
4.1	Performance of AI Model across Multinational Data	23



4.2	Reader Performance against AI Model	25
4.3	AI model and Readers Performance Consistency	27
Chapter 5	Discussion	31
5.1	Generalizability	31
5.2	Applicability	32
5.3	Explainability	33
5.4	Performance Variation of Different Dental Findings	34
Chapter 6	Conclusion	35
	References	37







List of Figures

Figure 3.1	Workflow for Datasets	14
Figure 3.2	Annotation Process of Promaton Dataset	15
Figure 3.3	Annotation Process of National Taiwan University Hospital Dataset	15
Figure 3.4	AI System Workflow This figure shows the workflow of our AI system. Starting with a dental panoramic radiograph, the process involves a semantic segmentation stage to classify the tooth number, followed by an instance detection stage to identify dental findings, ultimately providing the final finding summary map.	18
Figure 4.1	AI Model Performance on Dental Findings of Different Datasets .	24
Figure 4.2	Comparison of AI System Performance with Human Readers on Dental Findings in Taiwan This figure presents the receiver operating characteristic (ROC) curves for the AI system alongside the performance of four human readers. Each plot displays a specific dental finding, with the shaded vertical areas representing the 95% confidence intervals (CI) for sensitivity along the ROC curve.	26





List of Tables

Table 3.1	Characteristics of datasets	13
Table 4.1	Dental Findings Summary Performance This table presents the sample counts of eight dental findings across three datasets. The two numbers following the AUC-ROC represent the 95% confidence interval. . . .	25
Table 4.2	Comparison of AI and Reader Performance Metrics Symbols indicate statistical significance: * denotes superiority of AI over the mean of readers, † denotes non-inferiority (at a margin of 0.05) of AI to the mean of readers.	28
Table 4.3	Cohen's Kappa of the AI Model and Readers	29





Chapter 1 Introduction

1.1 Background

Images play a crucial role in the modern medical field, where various types of images have been utilized for diagnosis, treatment planning, and structural analysis for decades. Medical images are often categorized by the different parts of the human body they depict, such as chest X-rays, mammographies, brain scans, and others. Additionally, diverse imaging techniques serve specific purposes, including X-ray radiography, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), diffusion tensor imaging (DTI), and ultrasound imaging.

In dentistry, a range of imaging techniques is employed, from two-dimensional (2D) images such as X-rays and intraoral photos to three-dimensional (3D) images like computed tomography scans and cone beam computed tomography scans. Although 3D images often provide more vivid details and may offer greater accuracy, 2D images like panoramic radiographs, periapical films, and cephalometric images remain widely used in dental clinics due to their convenience and lower cost.

Among these various dental 2D X-rays, a dental panoramic radiograph provides a comprehensive overview of all teeth and their conditions, delivering essential information

for dentists to make preliminary diagnoses. Therefore, due to the abundant information they provide, we chose dental panoramic X-rays as the focus of our research.



1.2 Dental Panoramic Radiographs

A dental panoramic radiograph, or an orthopantomogram, provides a comprehensive view of the oral region, encompassing not only the oral area but also the upper and lower jaws, teeth, nasal area, and sinuses. This type of radiograph is commonly used to identify structures and hard tissues within the oral cavity. The interpretation of these radiographs typically focuses on findings related to individual teeth, such as dental caries and root canal fillings. It offers several advantages, including low radiation exposure, short acquisition time, examination convenience, and broad coverage of the oral and maxillofacial regions, making it a primary dental imaging tool ([Perschbacher, 2012](#)).

However, due to the panoramic radiograph's nature of compressing three-dimensional structures into two-dimensional planes, dental findings are complicated by overlapping structures and features, which often makes it difficult for dentists to reach consensus ([Kweon, Lee, Youk, Lee, & Kim, 2018](#)). Furthermore, the vast amount of information contained in a dental panoramic image can be challenging for dentists to explain to patients ([Henzler, Rasche, Ropinski, & Ritschel, 2018](#)). While there are many tools and mature artificial intelligence (AI) models available to assist doctors in the medical imaging field, the application of AI in dental imaging lags behind that in other medical areas. These challenges inspired us to develop an artificial intelligence-assisted system to interpret and summarize dental panoramic radiographs more effectively.

1.3 Modeling in Medical Imaging

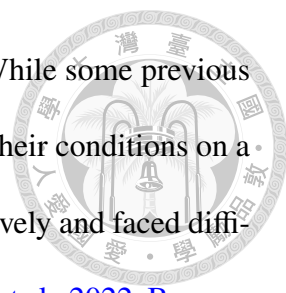


Computer-assisted analysis of medical images, particularly through deep learning and artificial intelligence, has seen significant advancements in recent years ([Shen, Wu, & Suk, 2017](#); [Zhou, Greenspan, & Shen, 2023](#)). Various tasks, such as body composition measurement ([Hsu et al., 2021](#)), cancer detection ([Wang, Khosla, Gargeya, Irshad, & Beck, 2016](#)), diagnosis, and prognosis prediction ([W. Zhu, Xie, Han, & Guo, 2020](#)), have been successfully applied using these technologies. Several deep learning methods and algorithms have also extended to the field of dental imaging, including panoramic images, periapical films, and cone-beam computed tomography. This expansion has made deep learning a promising approach in the analysis of dental images ([Chauhan et al., 2023](#)).

1.4 Modeling and Bottleneck in Dental Panoramic Radiographs

Despite the surge in research involving dental images and deep learning, I have chosen dental panoramic images as my research focus. Consequently, I am keen to identify potential bottlenecks in recent studies concerning artificial intelligence and dental panoramic radiographs. Upon reviewing related works, I have observed several issues and limitations in these studies.

Firstly, most research focuses on isolated issues, such as tooth detection ([Jader et al., 2018](#); [Koch, Perslev, Igel, & Brandt, 2019](#); [Oktay, 2017](#); [Yilmaz, Tasyurek, Amuk, Celik, & Canger, 2023](#)), tooth numbering ([Kim, Kim, Jeong, Yoon, & Youm, 2020](#); [Mahdi, Motoki, & Kobashi, 2020](#); [Tuzoff et al., 2019](#); [Xu et al., 2023](#)), or the detection of individual



findings (Ha, Jeon, Kim, Kim, & Han, 2021; Oztekin et al., 2023). While some previous studies attempted to provide a comprehensive summary of teeth and their conditions on a panoramic image, they struggled to integrate dental knowledge effectively and faced difficulties in accurately associating tooth numbers with findings (Almalki et al., 2022; Başaran et al., 2022; Gardiyanoglu, Ünsal, Akkaya, Aksoy, & Orhan, 2023; Muresan, Barbura, & Nedevschi, 2020; Rohrer et al., 2022). Additionally, although there are several studies comparing the performance of AI with human readers (Endres et al., 2020; Lian, Zhu, Zhu, & Zhu, 2021; Yang et al., 2020), these have not addressed complex problems to the extent our work has. Lastly, most of these studies rely on relatively small datasets from single sources.

1.5 A Proper Dental AI Model

After identifying the bottlenecks in recent research on dental panoramic images and artificial intelligence, and reviewing related works, developing an ideal AI model is undoubtedly our most crucial objective. However, defining what constitutes an ideal dental AI model for this project is essential. Beyond basic factors like robust model performance, there are several other criteria I aim to meet. Firstly, an excellent model should demonstrate generalizability (Krois et al., 2021), meaning it should perform well across diverse datasets, regardless of the originating institutions or the machines used. Secondly, the model we develop should be practically applicable in real-world settings, effectively bridging theory and practice. Lastly, a key distinction of our project from standard computer vision tasks is the integration of dental knowledge, making our model not only functional but also explainable. This incorporation of professional dental insights is vital for achieving our goals.



Chapter 2 Related Works

In recent years, technology based on deep learning and artificial intelligence has garnered significant attention from researchers in the field of medical imaging. Dental imaging, specifically, adopted this trend later than other imaging modalities. In this section, I review research that focuses on artificial intelligence and dental panoramic radiographs, summarizing it from various perspectives.

2.1 Tooth Detection

Previous studies have attempted to utilize artificial intelligence models to perform various tasks on dental images, such as tooth detection and tooth type numbering based on image classification networks. In the work of [Jader et al. \(2018\)](#), the existence and shape of teeth in panoramic images were detected. The article of [Koch et al. \(2019\)](#) improved tooth detection by employing network ensembling techniques, while [Oktay \(2017\)](#) focused on classifying teeth by their types, including incisors, premolars, and molars, without detecting their full shape. Compared the performance of two deep learning-based models on tooth segmentation, [Yilmaz et al. \(2023\)](#) using the bounding box of each tooth without marking the precise tooth contour. Additionally, [Cha, Yoon, Yeo, Huh, and Han \(2021\)](#) segmented not only teeth but also various anatomical structures in dental panoramic im-

ages.



2.2 Dental Knowledge Incorporation and Specific Applications

2.2.1 Tooth Numbering

It is crucial to integrate professional knowledge into tasks that involve specialized fields. For instance, in the context of dental panoramic radiographs, studies such as [Tuzoff et al. \(2019\)](#) and [Kim et al. \(2020\)](#) not only detect the tooth contour but also number the teeth using the dental numbering system. [Mahdi et al. \(2020\)](#) applied various deep learning models to the tasks of tooth recognition and numbering. [Xu et al. \(2023\)](#) extended the numbering to include not only permanent but also primary teeth. By employing accurate methods for tooth numbering, [S.-L. Chen et al. \(2022\)](#) and [Karaoglu, Ozcan, Pekince, and Yasa \(2023\)](#) developed innovative approaches to accomplish these tasks, respectively. These efforts demonstrate the initial steps toward incorporating dental knowledge into AI applications.

2.2.2 Binary Labeling

Additionally, clinical abnormalities in dental panoramic images have often been studied using deep learning techniques. However, the vast majority of past research on dental panoramic images focuses on the simpler task of image classification, which typically involves only binary labeling of the images into healthy or diseased categories. [Verma, Puri, Prabhu, and Smriti \(2020\)](#) detected several abnormalities in dental panoramic radiographs

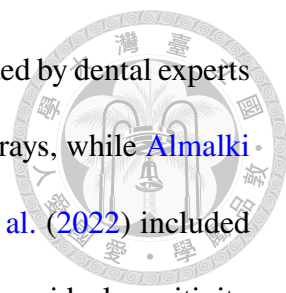
but ultimately classified them into binary labels as either normal or anomalous. [Ha et al. \(2021\)](#) focused solely on the presence of mesiodens in the X-ray. [Oztekin et al. \(2023\)](#) cropped a bounding box for each tooth and marked it as caries or non-caries. [BAYRAK-DAR et al. \(2020\)](#) analyzed alveolar bone loss in the images, categorizing them into classes of bone loss and non-bone loss.

2.2.3 Specialized Dental Applications

Dental researchers have increasingly focused on specific applications of artificial intelligence models for panoramic X-rays. For instance, [Takahashi et al. \(2020\)](#) and [Sukegawa et al. \(2021\)](#) utilized deep learning to classify implant brands in panoramic X-rays, while [Milošević, Vodanović, Galić, and Subašić \(2022\)](#) employed it to estimate the age of patients. To enhance the granularity of research, one significant development is finer classifications. Both [Dayı, Üzen, Çiçek, and Duman \(2023\)](#) and [H. Zhu et al. \(2022\)](#) focused on caries detection, with the authors employing different criteria for classification: Dayı categorized caries based on location, while Zhu assessed them by severity. Periodontitis is another area of interest for dental researchers; [Shon et al. \(2022\)](#) used the relative height of the cemento-enamel junction and alveolar bone in panoramic images to determine the stages of periodontal disease. Additionally, [Kurt-Bayrakdar et al. \(2024\)](#) identified patterns of alveolar bone loss and included furcation defects in their analysis.

2.3 Dental Finding Detection

Given the wealth of information provided in a dental panoramic image, the detection of multiple dental findings is a major research focus. The number and types of findings



selected in these studies vary. [Rohrer et al. \(2022\)](#) utilized data annotated by dental experts to categorize four types of dental restorations in dental panoramic X-rays, while [Almalki et al. \(2022\)](#) predicted four dental findings in their study. [Başaran et al. \(2022\)](#) included ten dental findings, such as dental calculus, although they did not achieve ideal sensitivity for them. [Muresan et al. \(2020\)](#) detected multiple findings in the dental panoramic images, aiming to compare the performance of various models. [Gardiynoğlu et al. \(2023\)](#) had access to abundant data and the same dental findings as our study, except for periapical radiolucencies. However, none of these studies combined the findings with tooth numbering. [Yüksel et al. \(2021\)](#) integrated the FDI tooth number with findings, but they only evaluated four findings related to the teeth.

2.4 Comparison Against Human

With the burgeoning interest in artificial intelligence within the dental imaging field, experts are increasingly curious about the real-world applicability of these models. As a result, comparing the performance of human experts and AI models has become a crucial aspect of current research. Various studies have been conducted on this comparison across different domains. [Yang et al. \(2020\)](#) automated the detection of jaw cysts and tumors and further compared the performance of the model with that of maxillofacial surgery specialists and general practitioners. [Endres et al. \(2020\)](#) detected tooth periapical disease and compared the performance between oral and maxillofacial surgeons. [Lian et al. \(2021\)](#) focused on caries detection and grading, which also included a comparative analysis with dentists.

Finally, recognizing both the merits and limitations of previous work, our study aims

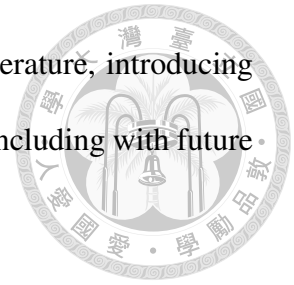
to address the task across three dimensions—generalizability, applicability, and explainability. By utilizing a diverse dataset comprised of X-rays produced by different machines and originating from various countries, comparing the performance between human readers and the artificial intelligence model, and incorporating dental knowledge in the determination of each finding, we strive to enhance the robustness and utility of our model.

2.5 Objective of Work

In this work, we aim to provide an end-to-end solution for summarizing dental panoramic images. This includes detailing all teeth and their indices found in the image, as well as all clinical findings such as implant, residual root, crowns/bridge, root canal filling, filling, caries, and periapical radiolucency associated with each tooth. We utilized a dataset from multiple institutions across different countries and ethnicities, which, to the best of our knowledge, is the most varied and comprehensive compared to previous studies.

Our data were meticulously labeled pixel-wise by human dental experts. We developed a CNN-based system designed to accurately summarize the findings presented in a panoramic image. Our approach combined object detection and semantic segmentation to produce an interpretable output report, and we developed metrics to appropriately evaluate the efficacy of this report. The performance of our deep learning model will be measured using three test sets from different countries. Additionally, we compared the performance of the AI model with that of human readers to demonstrate the practical applicability of the AI model. To our best knowledge, this is the first study to compare the performance between an AI model and human experts in the dental imaging field. This cross-domain and international project brought together dental experts and computer scientists to achieve its

goals. In this thesis, I will discuss our work by reviewing related literature, introducing our data and methodology, presenting results and discussions, and concluding with future applications of our findings.





Chapter 3 Materials and Methods

3.1 Datasets

Over 11,000 dental panoramic radiographs from various countries and institutions. Our data originated from three different countries: the Netherlands, Brazil, and Taiwan. The dental panoramic images from the Netherlands and Brazil were sourced from the Dutch company “Promaton,” while the X-rays from Taiwan were obtained from National Taiwan University Hospital. The collection, annotation procedures, and characteristics of the datasets vary, which will be detailed in the following sections. We excluded radiographs that displayed mixed dentition, the presence of metal accessories including fixed orthodontic devices, or structural bone anomalies (e.g., mandibular fractures or surgery). Additionally, images of poor quality, such as those that were blurred, disproportioned, or typically would require a retake in a clinical setting, were also excluded.

3.1.1 Promaton Dataset

The Promaton dataset, a proprietary collection curated by a Dutch company, “Promaton,” in collaboration with Radboud UMC, includes dental panoramic images from the Netherlands and Brazil. The number of dental panoramic radiographs from these two



countries totaled approximately 8,100 and 1,800, respectively. Data from the Netherlands originated from two different clinical sites within Promaton's own repository, and the Brazilian data were derived from a publicly available dataset collected at the Diagnostic Imaging Center of the Southwest State University of Bahia ([Silva, Oliveira, & Pithon, 2018](#)) and a Brazilian company, "Proradis." These images were partially annotated at the pixel level for findings including tooth number, tooth nerve, inferior alveolar nerve, filling (both metal and non-metal), crown (both metal and non-metal), root canal filling material, caries, periapical radiolucency, implant, residual root, maxillary sinus, and the mandibular and maxillary arches. Moreover, each radiograph could be annotated with one or more of these findings, with variations across each image. We selected dental panoramic X-rays that were fully annotated at the pixel level for specific findings such as tooth number, filling, crown, root canal filling material, caries, periapical radiolucency, implant, and residual root as the materials used in our research. Furthermore, we consolidated findings of metal and non-metal fillings into a single category labeled "fillings," and similarly for crowns, into "crowns/bridges" in our subsequent work.

3.1.2 National Taiwan University Hospital Dataset

The National Taiwan University Hospital dataset was compiled under NTUH IRB #202102018RINB. We gathered approximately 1,300 dental panoramic images from the National Taiwan University Hospital and ended up annotating 251 images. These dental panoramic images are fully annotated at the pixel level for findings including FDI tooth number, implant, residual root, crowns/bridge, root canal filling, filling, caries, and periapical radiolucency.



3.2 Data Characteristics

Table 3.1 presents the characteristics of our different datasets. However, we were unable to obtain the sex and age information for the Netherlands dataset and the age information for the Brazil dataset from the original DICOMs. Despite this, the National Taiwan University dataset (shown as Taiwan in the table) has a female-to-male ratio of 1:0.76, and the mean age of patients is 58.

Characteristic	Netherlands, test ($n = 1,201$)	Brazil ($n = 1,173$)	Taiwan ($n = 251$)
Years	2019 – 2020	[TBA: ?]	2021
Sex, n (%)			
Female	–	[TBA: ?]	143 (56.7%)
Male	–	[TBA: ?]	109 (43.3%)
Age, years			
Mean (SD)	–	–	56.4 (19.4)
Median (IQR)	–	–	58.5 (45.0 – 71.8)
Range	–	–	11 – 95

Table 3.1: **Characteristics of datasets**

3.3 Data Preprocessing

3.3.1 Data Split

We segmented our data into several sets based on nationality and intended use. The radiographs from the Netherlands were divided into a training set ($n = 4,044$), which was randomly split into 70% for training and 30% for validation, and a Netherlands test set ($n = 1,201$) as an internal test set. The images from Brazil ($n = 1,173$) and Taiwan ($n = 251$) were used as separate test sets. Additionally, for performance comparison between the dentists and the AI model, we randomly included 118 images from the Taiwan test set to form the Taiwan test set*. The workflow for our data is shown at figure 3.1.

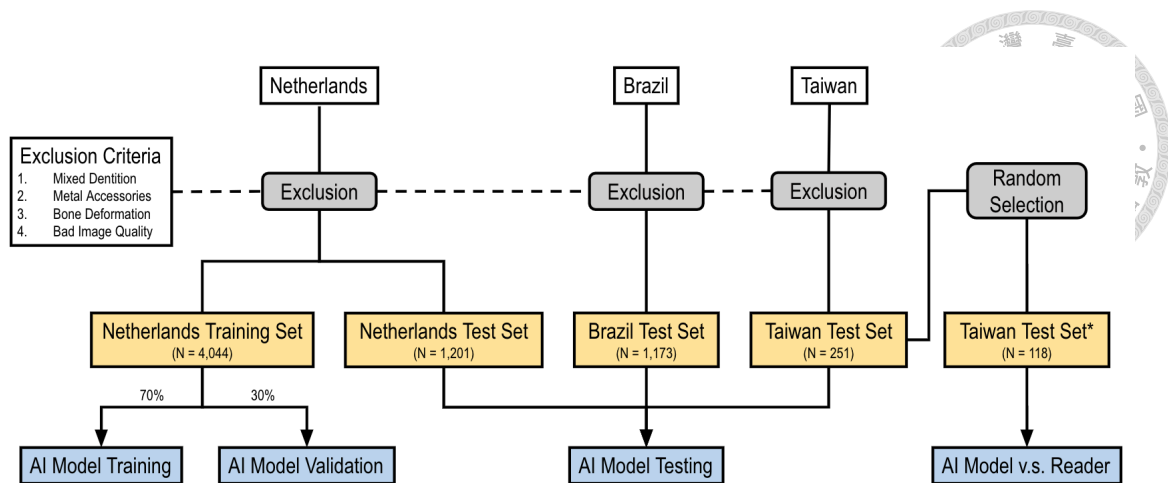


Figure 3.1: **Workflow for Datasets**

3.3.2 Data Annotation

All the dental panoramic radiographs utilized in this study were fully annotated for eight types of findings by dentists. These annotations served as references for our AI model. The annotation methodologies varied slightly between the datasets.

3.3.2.1 Promaton Dataset Annotation

The Promaton dataset, encompassing data from the Netherlands and Brazil, was annotated by three practicing dentists. A dental panoramic radiograph was initially labeled by the first dentist and reviewed by the second to create a reference instance map. This map was then used to generate a finding summary. The third dentist reviewed both the reference instance map and the finding summary to ensure consistency, resulting in the final reference finding summary. (fig.3.2)

3.3.2.2 National Taiwan University Hospital Dataset Annotation

Four practicing dentists were involved in the annotation process for the National Taiwan University Hospital dataset. They directly labeled each dental panoramic radiograph.

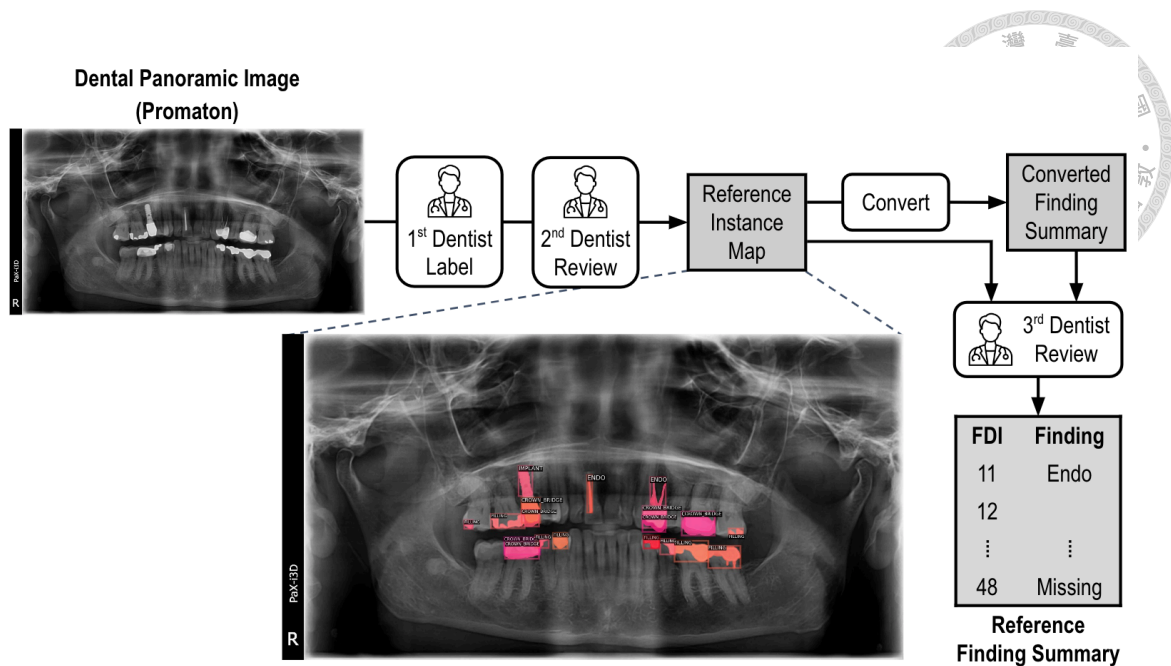


Figure 3.2: Annotation Process of Promaton Dataset

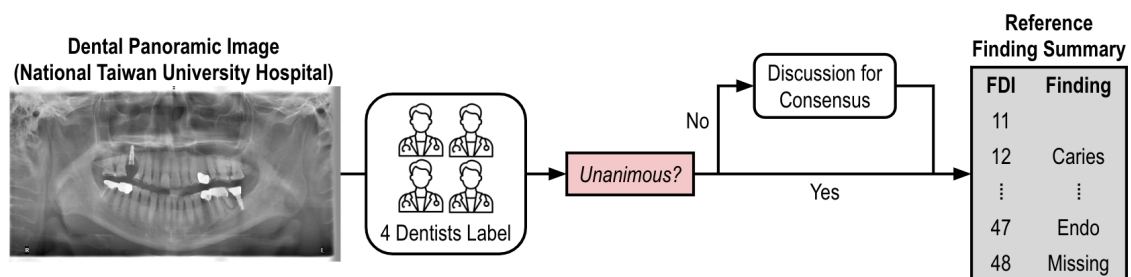


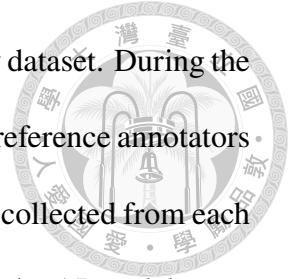
Figure 3.3: Annotation Process of National Taiwan University Hospital Dataset

Agreement among all four dentists was required for each finding; any initial discrepancies were discussed until consensus was achieved, thereby producing the final reference finding summary for this dataset. (fig.3.3)

3.3.2.3 Human Benchmark for AI Comparison

We recruited four dental clinicians including two general dental practitioners (G1, G2) each with 2 and 3 years of experience and two specialized dentists (S1, S2) each with 11 years (prosthodontics/orthodontics) and 15 years (endodontics) of experience to complete the same questionnaire of interpreting 118 dental panoramic images from Taiwan

test set*, which were randomly chose from National Taiwan University dataset. During the evaluation, all readers followed the same protocols established for the reference annotators for the National Taiwan University Hospital dataset. The answers that collected from each dentist were the benchmarks for the comparison between dentists and the AI model.



3.4 AI System for Finding Summarization

The AI system utilized in this study comprises a three-stage pipeline. We adopted two image segmentation methods in our task including semantic segmentation and instance segmentation by employing two models separately. The outcomes from both models were integrated through a computational process to yield our final results.

3.4.1 Semantic Segmentation Stage

Semantic segmentation operates at the pixel level, aiming to classify each pixel into a specific category. This method assigns a unique class to each pixel but does not differentiate between separate objects of the same class. For example, in an image, three cats would all be labeled as "cat," but semantic segmentation would not count or distinguish them as individual entities.

For semantic segmentation in our workflow, we adopted the DeepLab model (L.-C. Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017), using it to classify tooth indices at the pixel level. The primary objective was to delineate the region of each tooth, as the FDI number of each tooth closely correlates with its location. The model segmented 32 regions corresponding to the 32 FDI numbers of teeth, thereby creating the semantic segmentation map.

3.4.2 Instance Segmentation Stage



Building on the principles of semantic segmentation, instance segmentation not only identifies the class to which each pixel belongs but also distinguishes individual objects within the same class. This method often utilizes masks or bounding boxes to localize specific objects. For instance, instance segmentation can individually identify and segment multiple cats in a single image, treating each cat as a unique object.

For the instance segmentation in our work, the model incorporated two functionalities: instance detection and mask generation of instances. We selected the YOLOv8 model (Redmon, Divvala, Girshick, & Farhadi, 2016) for this purpose, aiming to identify and locate teeth and seven types of dental findings (implant, residual root, crowns/bridge, root canal filling, filling, caries, and periapical radiolucency). The instance detection component identified the presence of teeth and findings, while the masking feature outlined their shapes.

3.4.3 Combination Stage

The integration of results from these two stages occurs during the post-processing phase. We utilized a probabilistic algorithm to associate detected objects with classified indices, thereby producing a summary of findings for each image. Initially, the semantic segmentation model identified 32 regions, and the instance segmentation model detected objects, each associated with a confidence score ranging from 0 to 1 for eight possible findings, including tooth FDI numbers. These objects were then matched to the regions indicated by the semantic segmentation model, with each object assigned 32 probabilities corresponding to each region. This setup resulted in an intricate array of 256 scores per

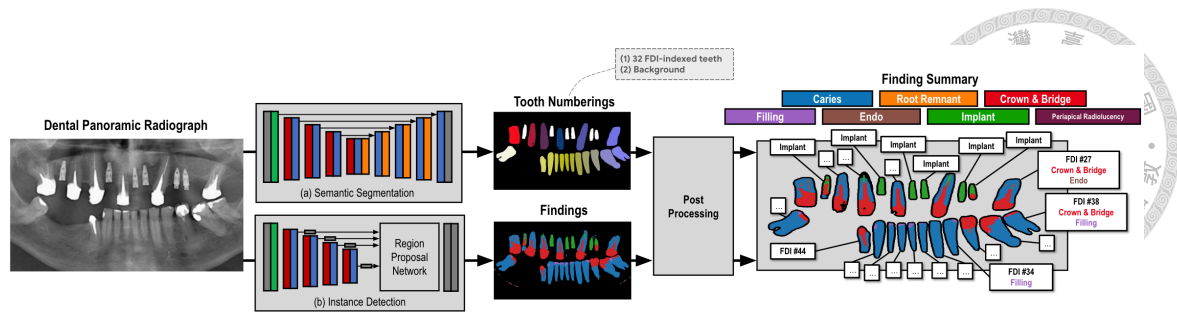


Figure 3.4: AI System Workflow This figure shows the workflow of our AI system. Starting with a dental panoramic radiograph, the process involves a semantic segmentation stage to classify the tooth number, followed by an instance detection stage to identify dental findings, ultimately providing the final finding summary map.

panoramic image, which were analyzed for further evaluation. The workflow is illustrated in figure 3.4.

3.5 Evaluation Metrics

This section would explain the primary metrics for evaluating model and readers performance. Sensitivity and specificity are crucial metrics in evaluating the performance of diagnostic tests and models, particularly in medical applications. Sensitivity, also known as the true positive rate or recall, measures the proportion of actual positives correctly identified by the model. It is calculated using the formula :

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

TP (True Positive) represents the number of correctly identified positive cases, and FN (False Negative) represents the number of actual positive cases incorrectly identified as negative. High sensitivity indicates the model's effectiveness at identifying most actual positive cases, thereby minimizing false negatives.

On the other hand, specificity, or the true negative rate, measures the proportion of actual negatives correctly identified by the model. It is calculated using the formula:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

TN (True Negative) represents the number of correctly identified negative cases, and FP (False Positive) represents the number of actual negative cases incorrectly identified as positive. High specificity indicates the model's effectiveness at identifying most actual negative cases, thereby minimizing false positives. Together, these metrics help ensure the balance between correctly identifying positive cases and avoiding false positives, which is essential for reliable medical diagnostics.

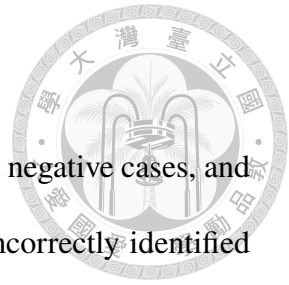
In addition to sensitivity and specificity, precision and the F1 score are important metrics for evaluating model performance. Precision, also known as positive predictive value, measures the proportion of true positive results among the predicted positive cases and is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

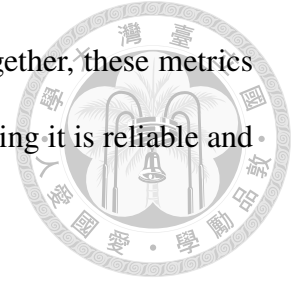
High precision indicates that when the model predicts a positive case, it is likely to be correct. The F1 score, which is the harmonic mean of precision and recall, provides a balanced measure when both false positives and false negatives are important, is calculated using the formula:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Another crucial metric is the AUC-ROC (Area Under the Receiver Operating Characteristic Curve), which evaluates the overall performance of a classification model across all thresholds. The ROC curve plots the true positive rate against the false positive rate at various threshold levels, and the AUC represents the area under this curve. A higher AUC indicates better model performance, with an AUC of 1 representing a perfect model and an



AUC of 0.5 representing a model with no discriminative ability. Together, these metrics provide a comprehensive evaluation of a model's performance, ensuring it is reliable and effective for medical diagnostics.



Additionally, we used Cohen's kappa in order to measure the agreements between readers and the model. Cohen's kappa is a statistical measure used to assess the agreement between two raters or classifiers who classify items into mutually exclusive categories. It is considered more robust than simple percent agreement calculation, as it accounts for the agreement occurring by chance. Mathematically, Cohen's kappa (κ) is calculated using the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

In the formula, p_o is the observed agreement between the raters, and p_e is the expected agreement by chance.

The value of Cohen's kappa ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and values less than 0 suggest systematic disagreement. The standards for interpreting kappa values are: 0.81 - 1.00 for almost perfect agreement, 0.61 - 0.80 for substantial agreement, 0.41 - 0.60 for moderate agreement, 0.21 - 0.40 for fair agreement, 0.01 - 0.20 for slight agreement, and less than 0.00 for poor agreement. Cohen's kappa is widely used in fields such as medicine, psychology, and social sciences to measure inter-rater reliability and ensure the validity of categorical data analysis.

3.6 Hardware

AI system training and testing (case reading) was done on a workstation running on Ubuntu 22.04 with an Intel Core i9-10900 CPU, an NVIDIA GeForce RTX 3080 GPU, and 64GB of system memory.







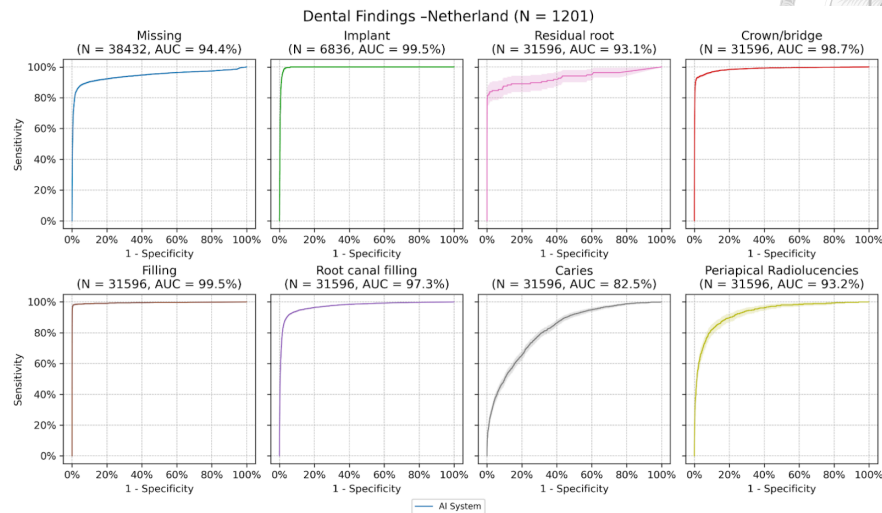
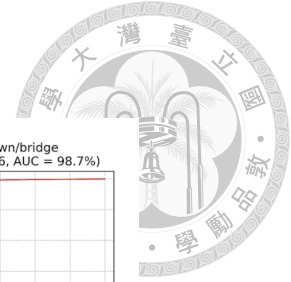
Chapter 4 Results

4.1 Performance of AI Model across Multinational Data

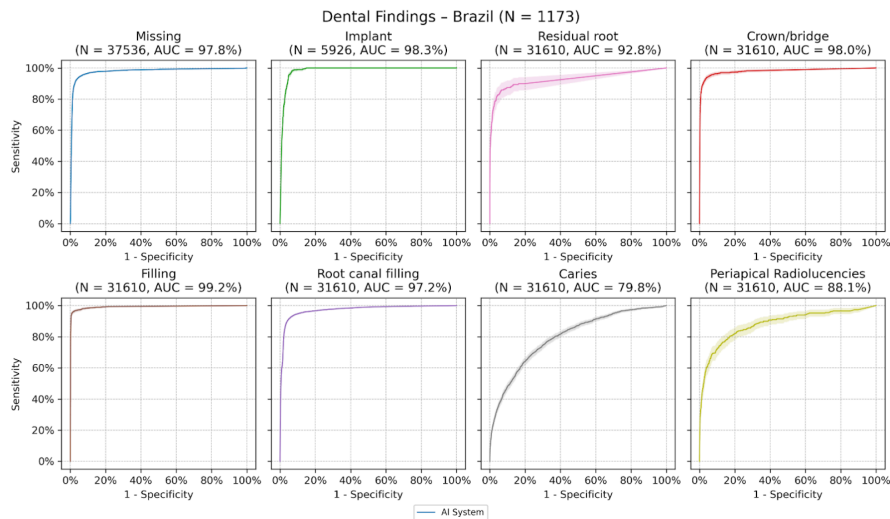
The figures (4.1 (a), 4.1 (b), 4.1 (c)) illustrates the performance of our AI model on the three test sets (Netherlands, Brazil, and Taiwan) separately. Noted that these figures present the receiver operating characteristic (ROC) curves for the performance of the AI system on Netherlands dataset (N=1201), Brazil dataset (N=1173) and National Taiwan University Hospital (shown as Taiwan in the figure) dataset (N=251) separately. Each plot displays a specific dental finding, with the shaded vertical areas representing the 95% confidence intervals (CI) for sensitivity along the ROC curve. Additionally, the table ?? represents the number, positive count, negative count, and AUC-ROC (Area Under the Receiver Operating Characteristic curve) with a 95% confidence interval for each finding across different datasets.

The results demonstrate that our model exhibited consistent trends across all three test sets and achieved high accuracies. Notably, the AUC-ROC exceeded 95% for three findings: implants, crowns/bridges, and root canal fillings. For the remaining findings, both missing teeth and fillings achieved AUC-ROCs of over 95% in both the Netherlands and Brazil sets. Residual roots reached over 90% AUC-ROC across all sets. Periapical radiolucency scored over 90% AUC-ROC in all sets except for the Brazil test set, whereas

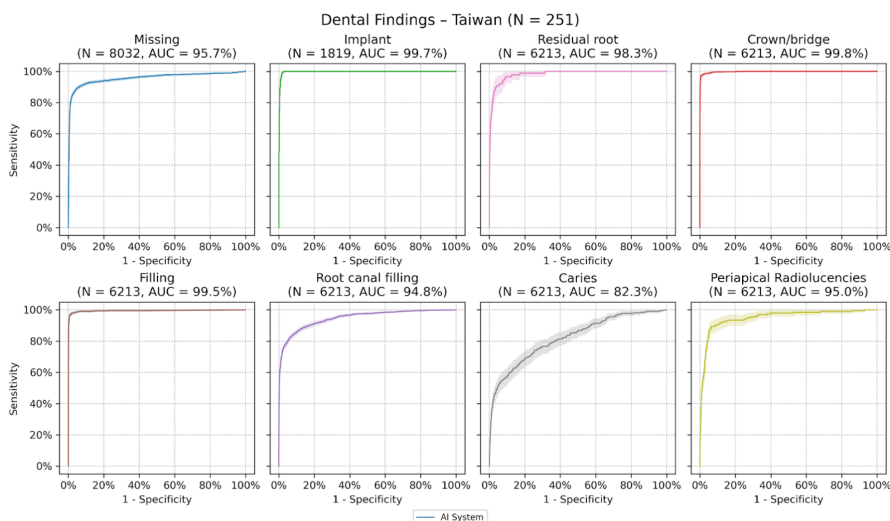
caries attained an AUC-ROC of around 80%.



(a) AI Model Performance on Dental Findings of Netherland Dataset



(b) AI Model Performance on Dental Findings of Brazil Dataset



(c) AI Model Performance on Dental Findings of Taiwan Dataset

Figure 4.1: AI Model Performance on Dental Findings of Different Datasets

Finding	Dataset	Sample #	Positive #	Negative #	AUC-ROC
Missingness	Netherlands	38,432	6,836	31,596	94.4% (94.0%, 94.9%)
	Brazil	38,656	6,166	32,490	97.8% (97.6%, 98.1%)
	Taiwan	8,096	1,715	6,381	99.8% (99.6%, 99.9%)
Implant	Netherlands	6,836	619	6,217	99.5% (99.4%, 99.6%)
	Brazil	6,166	635	5,531	98.3% (98.1%, 98.6%)
	Taiwan	1,715	191	1,524	99.8% (99.6%, 99.9%)
Residual root	Netherlands	31,596	137	31,459	93.1% (89.8%, 96.4%)
	Brazil	32,490	209	32,281	92.8% (90.3%, 95.3%)
	Taiwan	6,381	87	6,294	97.6% (95.7%, 99.5%)
Crown/bridge	Netherlands	31,596	3,038	28,558	98.7% (98.5%, 98.9%)
	Brazil	32,490	995	31,495	98.0% (97.4%, 98.6%)
	Taiwan	6,381	783	5,598	99.8% (99.7%, 99.9%)
Filling	Netherlands	31,596	7,412	24,184	99.5% (99.3%, 99.7%)
	Brazil	32,490	8,352	24,138	99.2% (99.0%, 99.5%)
	Taiwan	6,381	1,319	5,062	99.4% (98.8%, 100.0%)
Root canal filling	Netherlands	31,596	1,867	29,729	97.3% (97.0%, 97.5%)
	Brazil	32,490	1,438	31,052	97.2% (97.0%, 97.4%)
	Taiwan	6,381	580	5,801	95.9% (95.0%, 96.8%)
Caries	Netherlands	31,596	1,063	30,533	82.5% (81.3%, 83.7%)
	Brazil	32,490	1,491	30,999	79.8% (78.6%, 81.0%)
	Taiwan	6,381	317	6,064	83.0% (79.3%, 86.6%)
Periapical radiolucency	Netherlands	31,596	458	31,138	93.2% (92.0%, 94.5%)
	Brazil	32,490	389	32,101	88.1% (85.9%, 90.3%)
	Taiwan	6,381	105	6,276	97.4% (95.9%, 99.0%)

Table 4.1: **Dental Findings Summary Performance** This table presents the sample counts of eight dental findings across three datasets. The two numbers following the AUC-ROC represent the 95% confidence interval.

4.2 Reader Performance against AI Model

In the Taiwan test set* of 118 images, we assessed the performance of the AI system and four readers (G1, G2, S1, and S2) against the reference finding summary. The AI system's receiver operating characteristic (ROC) curve and the set operating points are displayed in figure 4.2. Table 4.2 compares the performance of an AI model and human readers (S1, S2, G1, and G2) in detecting various dental conditions using metrics such as sensitivity, specificity, precision, F1 score, and AUC-ROC. For missing teeth, the AI model achieved a sensitivity of 92.0%, outperforming all human readers, with a specificity of 95.9% and an F1 score of 89.4%. In detecting implants, the AI model had perfect sensitivity (100%) and a specificity of 98.4%, with an F1 score of 92.5% and an AUC-

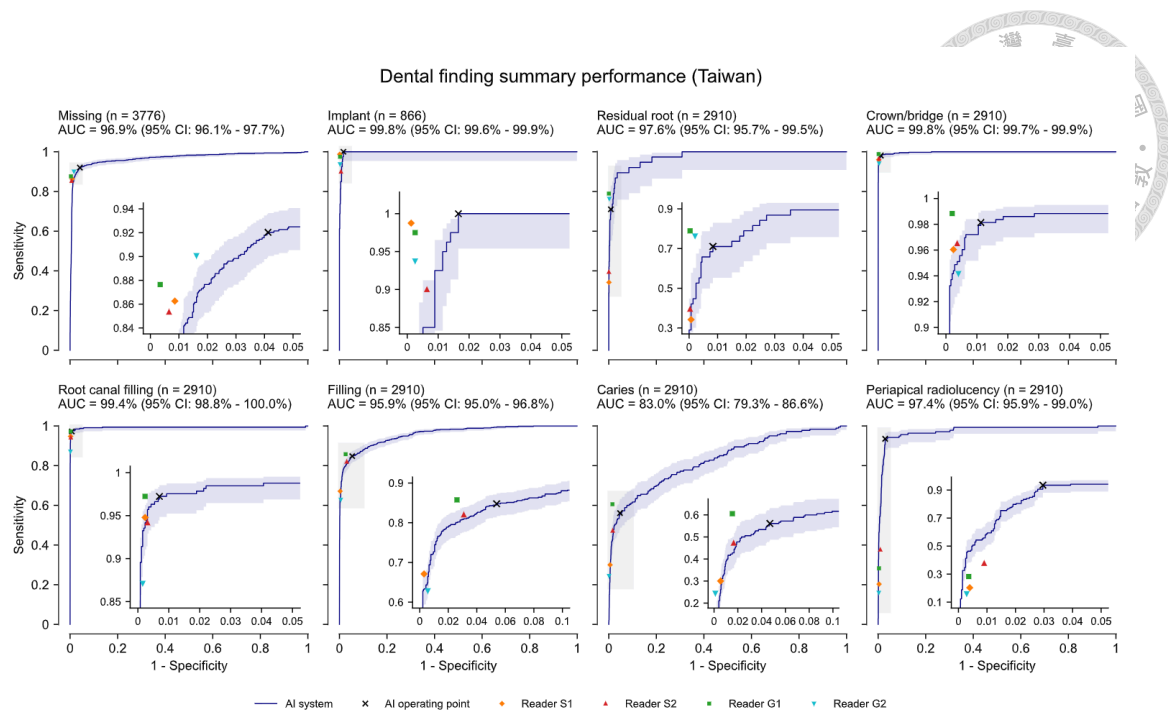


Figure 4.2: Comparison of AI System Performance with Human Readers on Dental Findings in Taiwan This figure presents the receiver operating characteristic (ROC) curves for the AI system alongside the performance of four human readers. Each plot displays a specific dental finding, with the shaded vertical areas representing the 95% confidence intervals (CI) for sensitivity along the ROC curve.

ROC of 99.8%, indicating high accuracy and precision. For residual roots, the AI model showed lower sensitivity at 71.1% but high specificity at 99.2%, comparable to human readers. In the crown/bridge and root canal filling categories, the model's performance in sensitivity, specificity, and F1 score was on par with human readers. However, for caries, the AI model achieved a sensitivity of 56.1% and a specificity of 95.3%, with an F1 score of 49.4% and an AUC-ROC of 83.0%, highlighting challenges in detection compared to other conditions while still being comparable to readers. In detecting periapical radiolucency, the AI model showed a sensitivity of 93.5%, a specificity of 97.0%, an F1 score of 73.9%, and an AUC-ROC of 97.4%, outperforming human readers in sensitivity, precision, and F1 score. Overall, the AI model demonstrated high performance in detecting various dental conditions, even surpassing human readers in some categories.

4.3 AI model and Readers Performance Consistency



Table 4.3 provides a comparison of Cohen's kappa values for various dental conditions detected by the AI model and different readers (G1, G2, S1, S2). In the crown/bridge category, our model achieved high agreement with all four readers, with Cohen's kappa values over 90%. Additionally, the inter-agreement rates between readers were high as well. For detecting implants and root canal fillings, the model also showed high Cohen's kappa values, ranging from around 84% to over 90% with all readers. The finding of missing teeth maintained around 80% Cohen's kappa between the model and all four readers. Interestingly, the Cohen's kappa between readers was higher than between the readers and the model in the category of missing teeth. When it came to detecting fillings, the agreement rate between the model and human readers was slightly lower but still around 70% Cohen's kappa values. The AI model showed over 50% Cohen's kappa with G1 (50.6%) and G2 (57.5%) but only 29.7% with S1 and 23.2% with S2. The agreement rates presented by Cohen's kappa scores were higher between human readers for this finding. Caries and periapical radiolucencies had the lowest Cohen's kappa values among all dental findings detected. Moreover, we observed that the agreement scores between each reader were higher than those between the readers and the AI in these two categories.

Table 4.2: Comparison of AI and Reader Performance Metrics Symbols indicate statistical significance: * denotes superiority of AI over the mean of readers, † denotes non-inferiority (at a margin of 0.05) of AI to the mean of readers.

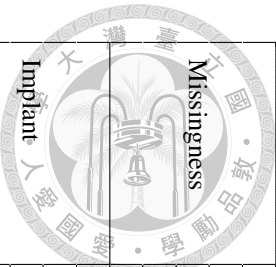
Finding	Metric	AI	Reader S1	Reader S2	Reader G1	Reader G2
 Missingness	Sensitivity	92.0% (90.0%, 93.7%)*†	86.1% (83.7%, 88.3%)	85.2% (82.7%, 87.4%)	87.5% (85.2%, 89.6%)	90.0% (87.8%, 91.8%)
	Specificity	95.9% (95.1%, 96.5%)*†	99.1% (98.7%, 99.4%)	99.4% (99.0%, 99.6%)	99.7% (99.4%, 99.8%)	98.4% (97.9%, 98.8%)
	Precision	86.9% (84.6%, 88.9%)	96.8% (95.3%, 97.8%)	97.5% (96.1%, 98.4%)	98.7% (97.6%, 99.3%)	94.3% (92.5%, 95.7%)
	F1 score	89.4% (87.7%, 91.1%)*†	91.1% (89.5%, 92.8%)	90.9% (89.3%, 92.6%)	92.8% (91.3%, 94.3%)	92.1% (90.6%, 93.6%)
	AUC-ROC	96.9% (96.1%, 97.7%)	–	–	–	–
Implant	Sensitivity	100.0% (95.4%, 100.0%)*†	97.5% (91.3%, 99.3%)	88.8% (80.0%, 94.0%)	96.3% (89.6%, 98.7%)	92.5% (84.6%, 96.5%)
	Specificity	98.3% (97.2%, 99.0%)*†	99.9% (99.3%, 100.0%)	99.4% (98.5%, 99.7%)	99.8% (99.1%, 99.9%)	99.8% (99.1%, 99.9%)
	Precision	86.0% (77.5%, 91.6%)	98.7% (93.2%, 99.8%)	93.4% (85.5%, 97.2%)	97.5% (91.2%, 99.3%)	97.4% (90.9%, 99.3%)
	F1 score	92.5% (88.2%, 96.8%)	98.1% (95.8%, 100.0%)	91.0% (86.1%, 96.0%)	96.9% (93.9%, 99.8%)	94.9% (91.1%, 98.7%)
	AUC-ROC	99.8% (99.6%, 99.9%)	–	–	–	–
Residual root	Sensitivity	71.1% (55.2%, 83.0%)	31.6% (19.1%, 47.5%)	36.8% (23.4%, 52.7%)	76.3% (60.8%, 87.0%)	73.7% (58.0%, 85.0%)
	Specificity	99.2% (98.8%, 99.4%)*†	99.9% (99.8%, 100.0%)	100.0% (99.8%, 100.0%)	100.0% (99.8%, 100.0%)	99.8% (99.5%, 99.9%)
	Precision	52.9% (39.5%, 65.9%)*	85.7% (60.1%, 96.0%)	93.3% (70.2%, 98.8%)	96.7% (83.3%, 99.4%)	82.4% (66.5%, 91.7%)
	F1 score	60.7% (48.7%, 72.7%)	46.2% (29.3%, 63.0%)	52.8% (36.5%, 69.1%)	85.3% (76.2%, 94.3%)	77.8% (67.1%, 88.4%)
	AUC-ROC	97.6% (95.7%, 99.5%)*	–	–	–	–
Crown/bridge	Sensitivity	98.1% (96.4%, 99.1%)*†	95.8% (93.5%, 97.3%)	96.3% (94.0%, 97.7%)	98.6% (97.0%, 99.4%)	93.9% (91.3%, 95.8%)
	Specificity	98.9% (98.4%, 99.2%)*†	99.8% (99.5%, 99.9%)	99.6% (99.3%, 99.8%)	99.8% (99.5%, 99.9%)	99.6% (99.3%, 99.8%)
	Precision	93.8% (91.1%, 95.7%)*	98.6% (96.9%, 99.3%)	97.9% (96.0%, 98.9%)	98.8% (97.3%, 99.5%)	97.6% (95.6%, 98.7%)
	F1 score	95.9% (94.4%, 97.4%)*†	97.2% (95.9%, 98.4%)	97.1% (95.8%, 98.4%)	98.7% (97.9%, 99.6%)	95.7% (94.2%, 97.3%)*
	AUC-ROC	99.8% (99.7%, 99.9%)*	–	–	–	–
Root canal filling	Sensitivity	97.2% (94.8%, 98.5%)*†	94.5% (91.4%, 96.5%)	93.9% (90.7%, 96.0%)	96.9% (94.5%, 98.3%)	86.8% (82.7%, 90.1%)
	Specificity	99.3% (98.9%, 99.6%)*†	99.8% (99.5%, 99.9%)	99.7% (99.4%, 99.8%)	99.8% (99.5%, 99.9%)	99.9% (99.6%, 99.9%)
	Precision	94.6% (91.7%, 96.6%)*	98.1% (95.9%, 99.1%)	97.5% (95.1%, 98.7%)	98.1% (96.0%, 99.1%)	98.6% (96.5%, 99.5%)
	F1 score	95.9% (94.2%, 97.6%)*†	96.3% (94.6%, 97.9%)	95.6% (93.9%, 97.4%)	97.5% (96.2%, 98.9%)	92.3% (90.0%, 94.7%)
	AUC-ROC	99.4% (98.8%, 100.0%)*	–	–	–	–
Filling	Sensitivity	84.8% (81.7%, 87.4%)*†	66.9% (63.1%, 70.6%)	81.8% (78.6%, 84.7%)	85.6% (82.6%, 88.2%)	62.7% (58.8%, 66.4%)
	Specificity	94.6% (93.6%, 95.5%)*	99.7% (99.4%, 99.9%)	96.9% (96.1%, 97.5%)	97.4% (96.7%, 98.0%)	99.4% (99.0%, 99.7%)
	Precision	80.7% (77.5%, 83.6%)*	98.3% (96.6%, 99.2%)	87.6% (84.6%, 90.0%)	89.7% (87.0%, 91.9%)	96.7% (94.5%, 98.1%)
	F1 score	82.7% (80.2%, 85.1%)*	79.7% (76.9%, 82.4%)	84.6% (82.2%, 87.0%)	87.6% (85.4%, 89.8%)	76.1% (73.1%, 79.1%)
	AUC-ROC	95.9% (95.0%, 96.8%)*	–	–	–	–
Caries	Sensitivity	56.1% (48.8%, 63.2%)*†	29.4% (23.3%, 36.5%)	46.7% (39.5%, 54.0%)	60.0% (52.7%, 66.9%)	23.9% (18.3%, 30.6%)
	Specificity	95.3% (94.5%, 96.0%)*†	99.5% (99.1%, 99.7%)	98.4% (97.8%, 98.8%)	98.5% (97.9%, 98.9%)	99.9% (99.7%, 100.0%)
	Precision	44.1% (37.8%, 50.6%)*	77.9% (66.7%, 86.2%)	65.1% (56.6%, 72.8%)	72.0% (64.3%, 78.6%)	93.5% (82.5%, 97.8%)
	F1 score	49.4% (43.4%, 55.3%)*	42.7% (35.0%, 50.5%)*	54.4% (47.7%, 61.1%)*	65.4% (59.5%, 71.4%)*	38.1% (30.0%, 46.2%)*
	AUC-ROC	83.0% (79.3%, 86.6%)*	–	–	–	–
Periapical radiolucency	Sensitivity	93.5% (88.1%, 96.5%)*†	19.6% (13.8%, 27.0%)*	37.0% (29.4%, 45.3%)*	27.5% (20.8%, 35.5%)*	15.2% (10.2%, 22.2%)*
	Specificity	97.0% (96.3%, 97.6%)*†	99.6% (99.3%, 99.8%)*	99.1% (98.7%, 99.4%)*	99.6% (99.3%, 99.8%)*	99.7% (99.4%, 99.9%)*
	Precision	61.1% (54.4%, 67.5%)*	71.1% (55.2%, 83.0%)*	67.1% (55.9%, 76.6%)*	79.2% (65.7%, 88.3%)*	72.4% (54.3%, 85.3%)*
	F1 score	73.9% (68.7%, 79.1%)*†	30.7% (21.8%, 39.6%)*	47.7% (39.4%, 55.9%)*	40.9% (31.9%, 49.8%)*	25.2% (16.4%, 33.9%)*
	AUC-ROC	97.4% (95.9%, 99.0%)*	–	–	–	–



Table 4.3: Cohen's Kappa of the AI Model and Readers

Finding	Reader	v.s. Reader			
		G1	G2	S1	S2
Missingness	AI	80.5% (78.2%, 82.8%)	81.4% (79.2%, 83.6%)	79.5% (77.2%, 81.9%)	78.2% (75.8%, 80.6%)
	G1	—	86.5% (84.5%, 88.5%)	95.2% (94.0%, 96.4%)	94.3% (93.0%, 95.7%)
	G2	—	—	86.1% (84.1%, 88.1%)	85.7% (83.7%, 87.8%)
	S1	—	—	—	92.6% (91.1%, 94.1%)
Implants	AI	90.4% (85.5%, 95.2%)	88.3% (82.9%, 93.6%)	90.4% (85.5%, 95.2%)	84.4% (78.2%, 90.5%)
	G1	—	92.3% (87.8%, 96.8%)	98.6% (96.7%, 100.0%)	90.9% (86.0%, 95.8%)
	G2	—	—	93.7% (89.6%, 97.8%)	85.8% (79.6%, 91.9%)
	S1	—	—	—	92.3% (87.8%, 96.8%)
Residual root	AI	50.6% (37.0%, 64.1%)	57.5% (44.8%, 70.3%)	29.7% (14.9%, 44.6%)	23.2% (9.3%, 37.2%)
	G1	—	75.5% (63.6%, 87.4%)	47.5% (28.9%, 66.0%)	50.7% (32.6%, 68.9%)
	G2	—	—	43.6% (25.8%, 61.4%)	42.7% (25.0%, 60.4%)
	S1	—	—	—	77.3% (60.2%, 94.4%)
Crown/ bridge	AI	94.9% (93.3%, 96.5%)	93.1% (91.2%, 95.0%)	94.2% (92.4%, 95.9%)	94.3% (92.6%, 96.0%)
	G1	—	94.9% (93.2%, 96.5%)	96.3% (94.9%, 97.7%)	96.4% (95.1%, 97.8%)
	G2	—	—	94.4% (92.7%, 96.1%)	94.6% (92.9%, 96.3%)
	S1	—	—	—	95.7% (94.2%, 97.2%)
Root canal filling	AI	93.8% (91.8%, 95.8%)	88.3% (85.5%, 91.1%)	93.1% (90.9%, 95.2%)	92.0% (89.8%, 94.3%)
	G1	—	91.4% (89.0%, 93.8%)	95.8% (94.1%, 97.5%)	93.7% (91.6%, 95.7%)
	G2	—	—	90.2% (87.6%, 92.8%)	88.7% (85.9%, 91.5%)
	S1	—	—	—	92.9% (90.7%, 95.1%)
Filling	AI	74.6% (71.6%, 77.6%)	64.9% (61.4%, 68.5%)	69.7% (66.4%, 73.1%)	72.8% (69.7%, 75.9%)
	G1	—	72.4% (69.0%, 75.7%)	74.9% (71.7%, 78.1%)	79.5% (76.7%, 82.3%)
	G2	—	—	77.7% (74.4%, 81.1%)	69.9% (66.4%, 73.3%)
	S1	—	—	—	74.9% (71.7%, 78.1%)
Caries	AI	33.2% (26.7%, 39.7%)	21.1% (14.6%, 27.6%)	21.3% (14.9%, 27.7%)	28.2% (21.8%, 34.7%)
	G1	—	41.0% (32.3%, 49.7%)	42.7% (34.3%, 51.0%)	51.8% (44.5%, 59.1%)
	G2	—	—	54.3% (43.1%, 65.5%)	42.7% (33.5%, 52.0%)
	S1	—	—	—	46.6% (37.8%, 55.3%)
Periapical radiolucency	AI	24.9% (17.9%, 31.9%)	15.1% (8.9%, 21.2%)	19.0% (12.4%, 25.6%)	34.3% (27.1%, 41.4%)
	G1	—	34.6% (20.7%, 48.5%)	53.9% (40.9%, 66.8%)	41.7% (30.6%, 52.8%)
	G2	—	—	34.0% (19.1%, 48.9%)	23.2% (12.2%, 34.1%)
	S1	—	—	—	38.6% (27.0%, 50.1%)





Chapter 5 Discussion

5.1 Generalizability

The generalizability of a medical image AI model refers to its ability to perform accurately across diverse and independent datasets, ensuring effectiveness in various clinical settings and patient populations. This is crucial for broad applicability, robustness, and consistent diagnostic accuracy, ultimately improving patient care. In medical imaging, where image quality can vary due to different hardware and parameters, and characteristics may differ across races, generalizability is especially important.

Our datasets contained dental panoramic radiographs from three different countries and four clinical sites, making them suitable for measuring the generalizability of our AI model. By analyzing the performance, our model maintained stable results across the three datasets. Through graphical analysis, we observed similar trends for similar types of dental findings across all sets, showcasing the model's potential for broader application in dental diagnostics. Demonstrating this performance not only showed the high generalizability of the model but also built trust among healthcare providers and patients, facilitating the model's acceptance and integration into daily clinical practice. Additionally, it enhances efficiency and scalability, allowing the model to be deployed across multiple healthcare facilities without extensive retraining or customization.

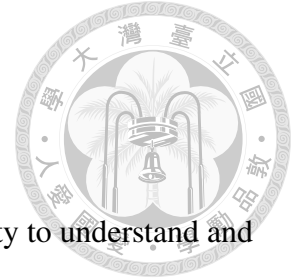
5.2 Applicability



One important goal of all AI models is to transition into real-world applications, and our work is no exception. While AI models can standardize the diagnostic process by reducing variability between practitioners and ensuring consistent care regardless of individual experience, their applicability remains crucial. This determines their capability to be effectively used in real-world dental practices. To demonstrate the applicability of our model, we recruited four human dental experts and compared their performances to our model. The results showed that our AI model achieved performance comparable to the human experts and even exceeded them in some categories.

For instance, we observed that the model performed comparably to human readers in the specificity of seven out of eight categories, effectively identifying patients who do not have the condition and minimizing false positives. Additionally, the model achieved non-inferiority to dentists in the sensitivity of all dental findings except for residual roots. High specificity can reduce unnecessary treatments, conserve medical resources, alleviate patient anxiety, and support accurate clinical decision-making, while good sensitivity ensures that positive cases are accurately identified, reducing the likelihood of missed diagnoses and improving patient outcomes. Moreover, the model significantly improved efficiency by quickly analyzing images, alleviating the workload of medical professionals, and reducing patient wait times. Considering these advantages, we can confidently say that our AI model could be effectively applied in clinical settings and assist humans efficiently.

5.3 Explainability



The explainability of a dental image AI model refers to the ability to understand and interpret how the model makes its decisions. This is crucial because it builds trust and confidence among dentists and patients, ensuring they understand and rely on the AI's conclusions. Explainability also helps in identifying and correcting errors, ensuring regulatory compliance, and supporting clinical decision-making by providing clear insights into complex cases. Additionally, it improves patient communication, as dentists can explain the AI's reasoning, enhancing patient acceptance and adherence to treatment plans.

For our model, we applied the soft-OR method into the modeling process. soft-OR is a logical operation used in computing and AI to combine multiple conditions flexibly by assigning different weights to each. Soft-OR calculates an aggregated score based on the weighted influence of each condition. This score is then compared to a threshold to determine the final outcome.

In the instance segmentation stage, the model detected dental findings as objects. Each object had different probabilities corresponding to 32 teeth. Through the soft-OR calculation, we obtained the final probability of the finding on each tooth. For example, if the model detected four fillings in the image, and for tooth 17, the probabilities of each filling appearing on it are 0.8, 0.6, 0.4, and 0.2, the probability of a filling on tooth 17 is:

$$1 - [(1 - 0.8) * (1 - 0.6) * (1 - 0.4) * (1 - 0.2)] = 0.9616$$

With eight dental findings and 32 teeth, we calculated $8 \times 32 = 256$ scores to build the final summarization. This method allows for nuanced and accurate decision-making,

accommodating varying degrees of significance among conditions, making it ideal for complex real-world applications. In summary, rather than functioning as a black box, soft-OR not only improves the flexibility and accuracy of AI models but also significantly enhances their explainability by providing clear, interpretable, and transparent decision-making processes that align with expert knowledge and real-world scenarios.

5.4 Performance Variation of Different Dental Findings

Although our model maintained high accuracy and AUC-ROC, the performance and agreement rates varied across different findings. We found that well-defined and high-contrast findings such as implants, crowns/bridges, and root canal fillings tended to achieve higher agreement rates, both among human readers and between the AI model and readers. Additionally, the model's performance on these findings was better than on others.

Conversely, in the categories of residual roots, caries, and periapical radiolucencies, these findings had lower Cohen's kappa values compared to other dental findings. We believe this is partly due to the unclear standards for these dental findings, making differentiation difficult. For instance, there is no specific crown/root ratio to definitively classify a tooth as normal or a residual root. Another reason is that dental panoramic radiographs display the entire oral area, which can reduce detail. Findings such as caries and periapical radiolucencies lack high contrast or clear boundaries, affecting the accuracy of both the readers and the model.



Chapter 6 Conclusion

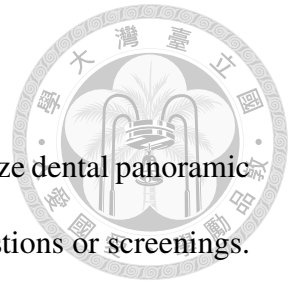
Applying artificial intelligence in medical image analysis has flourished in recent years. Our work focused on dental images, specifically dental panoramic radiographs. We identified eight dental findings: missing teeth, implants, residual roots, crowns/bridges, root canal fillings, fillings, caries, and periapical radiolucencies. The AI model we developed in this work aimed to interpret these dental findings accurately and automatically.

In this study, we collected over 6,000 dental panoramic images from various countries, forming a multinational dataset. The first experiment we presented in the article evaluated the model's performance on three datasets from different countries. The results showed that our AI model maintained high performance across all three sets and exhibited similar trends for the same findings. These results, along with the characteristics of our data, effectively demonstrate the generalizability of our model.

To address the question of whether artificial intelligence can replace human beings, we compared the performance of the AI model to that of human readers. We recruited two general dentists and two specialists for this experiment. All four readers completed the same task of interpreting 118 dental panoramic images from the Taiwan test set*. Despite the comparable performance between the AI model and human experts in most dental findings, diagnosing caries and residual roots remained challenging for both the model

and the readers.

We believe that our model can efficiently and accurately summarize dental panoramic images, providing valuable assistance to dentists for treatment suggestions or screenings. The advancements in deep learning and artificial intelligence have significantly impacted the field of medical imaging, including dental images. These AI-assisted applications have the potential to streamline clinical processes, enhancing efficiency and possibly optimizing the entire clinical procedure. By integrating AI into dental diagnostics, we can support dentists in making more informed decisions, ultimately improving patient outcomes and the overall quality of care. Additionally, the ability of AI to handle large volumes of data quickly and accurately can reduce the burden on dental professionals, allowing them to focus more on patient care and less on routine image analysis. This synergy between human expertise and artificial intelligence can lead to a more efficient, effective, and patient-centered approach to dental healthcare.

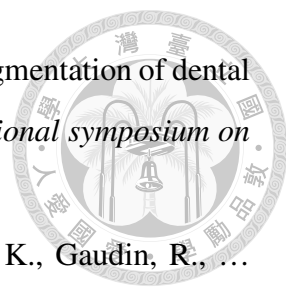




References

- Almalki, Y. E., Din, A. I., Ramzan, M., Irfan, M., Aamir, K. M., Almalki, A., ... Rahman, S. (2022). Deep learning models for classification of dental diseases using orthopantomography x-ray opg images. *Sensors*, 22(19), 7370.
- Başaran, M., Çelik, Ö., Bayrakdar, I. S., Bilgir, E., Orhan, K., Odabaş, A., ... Jagtap, R. (2022). Diagnostic charting of panoramic radiography using deep-learning artificial intelligence system. *Oral radiology*, 1–7.
- BAYRAKDAR, İ., elik, Ö., ORHAN, K., BİLGİR, E., ODABAŞ, A., & ASLAN, A. (2020). Success of artificial intelligence system in determining alveolar bone loss from dental panoramic radiography images. *Cumhuriyet Dental Journal*, 23(4).
- Cha, J.-Y., Yoon, H.-I., Yeo, I.-S., Huh, K.-H., & Han, J.-S. (2021). Panoptic segmentation on panoramic radiographs: Deep learning-based segmentation of various structures including maxillary sinus and mandibular canal. *Journal of Clinical Medicine*, 10(12), 2577.
- Chauhan, R. B., Shah, T. V., Shah, D. H., Gohil, T. J., Oza, A. D., Jajal, B., & Saxena, K. K. (2023). An overview of image processing for dental diagnosis. *Innovation and Emerging Technologies*, 10, 2330001.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.
- Chen, S.-L., Chen, T.-Y., Huang, Y.-C., Chen, C.-A., Chou, H.-S., Huang, Y.-Y., ... others (2022). Missing teeth and restoration detection using dental panoramic radiography

- based on transfer learning with cnns. *IEEE Access*, 10, 118654–118664.
- Dayı, B., Üzen, H., Çiçek, İ. B., & Duman, Ş. B. (2023). A novel deep learning-based approach for segmentation of different type caries lesions on panoramic radiographs. *Diagnostics*, 13(2), 202.
- Endres, M. G., Hillen, F., Salloumis, M., Sedaghat, A. R., Niehues, S. M., Quatela, O., ... others (2020). Development of a deep learning algorithm for periapical disease detection in dental radiographs. *Diagnostics*, 10(6), 430.
- Gardiyanoglu, E., Ünsal, G., Akkaya, N., Aksoy, S., & Orhan, K. (2023). Automatic segmentation of teeth, crown–bridge restorations, dental implants, restorative fillings, dental caries, residual roots, and root canal fillings on orthopantomographs: Convenience and pitfalls. *Diagnostics*, 13(8), 1487.
- Ha, E.-G., Jeon, K. J., Kim, Y. H., Kim, J.-Y., & Han, S.-S. (2021). Automatic detection of mesiodens on panoramic radiographs using artificial intelligence. *Scientific reports*, 11(1), 23061.
- Henzler, P., Rasche, V., Ropinski, T., & Ritschel, T. (2018). Single-image tomography: 3d volumes from 2d cranial x-rays. In *Computer graphics forum* (Vol. 37, pp. 377–388).
- Hsu, T.-M. H., Schawkat, K., Berkowitz, S. J., Wei, J. L., Makoyeva, A., Legare, K., ... others (2021). Artificial intelligence to assess body composition on routine abdominal ct scans and predict mortality in pancreatic cancer—a recipe for your local application. *European journal of radiology*, 142, 109834.
- Jader, G., Fontineli, J., Ruiz, M., Abdalla, K., Pithon, M., & Oliveira, L. (2018). Deep instance segmentation of teeth in panoramic x-ray images. In *2018 31st sibgrapi conference on graphics, patterns and images (sibgrapi)* (pp. 400–407).
- Karaoglu, A., Ozcan, C., Pekince, A., & Yasa, Y. (2023). Numbering teeth in panoramic images: A novel method based on deep learning and heuristic algorithm. *Engineering Science and Technology, an International Journal*, 37, 101316.
- Kim, C., Kim, D., Jeong, H., Yoon, S.-J., & Youm, S. (2020). Automatic tooth detection and numbering using a combination of a cnn and heuristic algorithm. *Applied Sciences*, 10(16), 5624.

- 
- Koch, T. L., Perslev, M., Igel, C., & Brandt, S. S. (2019). Accurate segmentation of dental panoramic radiographs with u-nets. In *2019 ieee 16th international symposium on biomedical imaging (isbi 2019)* (pp. 15–19).
- Krois, J., Garcia Cantu, A., Chaurasia, A., Patil, R., Chaudhari, P. K., Gaudin, R., ... Schwendicke, F. (2021). Generalizability of deep learning models for dental image analysis. *Scientific reports*, *11*(1), 6102.
- Kurt-Bayrakdar, S., Bayrakdar, İ. Ş., Yavuz, M. B., Sali, N., Çelik, Ö., Köse, O., ... Orhan, K. (2024). Detection of periodontal bone loss patterns and furcation defects from panoramic radiographs using deep learning algorithm: a retrospective study. *BMC Oral Health*, *24*(1), 155.
- Kweon, H. H.-I., Lee, J.-H., Youk, T.-m., Lee, B.-A., & Kim, Y.-T. (2018). Panoramic radiography can be an effective diagnostic tool adjunctive to oral examinations in the national health checkup program. *Journal of periodontal & implant science*, *48*(5), 317.
- Lian, L., Zhu, T., Zhu, F., & Zhu, H. (2021). Deep learning for caries detection and classification. *Diagnostics*, *11*(9), 1672.
- Mahdi, F. P., Motoki, K., & Kobashi, S. (2020). Optimization technique combined with deep learning method for teeth recognition in dental panoramic radiographs. *Scientific reports*, *10*(1), 19261.
- Milošević, D., Vodanović, M., Galić, I., & Subašić, M. (2022). Automated estimation of chronological age from panoramic dental x-ray images using deep learning. *Expert systems with applications*, *189*, 116038.
- Muresan, M. P., Barbura, A. R., & Nedevschi, S. (2020). Teeth detection and dental problem classification in panoramic x-ray images using deep learning and image processing techniques. In *2020 ieee 16th international conference on intelligent computer communication and processing (iccp)* (pp. 457–463).
- Oktay, A. B. (2017). Tooth detection with convolutional neural networks. In *2017 medical technologies national congress (tiptekno)* (pp. 1–4).
- Oztekin, F., Katar, O., Sadak, F., Yildirim, M., Cakar, H., Aydogan, M., ... others (2023). An explainable deep learning model to prediction dental caries using panoramic

- 
- radiograph images. *Diagnostics*, 13(2), 226.
- Perschbacher, S. (2012). Interpretation of panoramic radiographs. *Australian dental journal*, 57, 40–45.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).
- Rohrer, C., Krois, J., Patel, J., Meyer-Lueckel, H., Rodrigues, J. A., & Schwendicke, F. (2022). Segmentation of dental restorations on panoramic radiographs using deep learning. *Diagnostics*, 12(6), 1316.
- Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19, 221–248.
- Shon, H. S., Kong, V., Park, J. S., Jang, W., Cha, E. J., Kim, S.-Y., ... Kim, K. A. (2022). Deep learning model for classifying periodontitis stages on dental panoramic radiography. *Applied Sciences*, 12(17), 8500.
- Silva, G., Oliveira, L., & Pithon, M. (2018). Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Systems with Applications*, 107, 15–31.
- Sukegawa, S., Yoshii, K., Hara, T., Matsuyama, T., Yamashita, K., Nakano, K., ... Furuki, Y. (2021). Multi-task deep learning model for classification of dental implant brand and treatment stage using dental panoramic radiograph images. *Biomolecules*, 11(6), 815.
- Takahashi, T., Nozaki, K., Gonda, T., Mameno, T., Wada, M., & Ikebe, K. (2020). Identification of dental implants using deep learning—pilot study. *International journal of implant dentistry*, 6, 1–6.
- Tuzoff, D. V., Tuzova, L. N., Bornstein, M. M., Krasnov, A. S., Kharchenko, M. A., Nikolenko, S. I., ... Bednenko, G. B. (2019). Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, 48(4), 20180051.
- Verma, D., Puri, S., Prabhu, S., & Smriti, K. (2020). Anomaly detection in panoramic dental x-rays using a hybrid deep learning and machine learning approach. In 2020

ieee region 10 conference (tencon) (pp. 263–268).

- Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.
- Xu, M., Wu, Y., Xu, Z., Ding, P., Bai, H., & Deng, X. (2023). Robust automated teeth identification from dental radiographs using deep learning. *Journal of Dentistry*, 136, 104607.
- Yang, H., Jo, E., Kim, H. J., Cha, I.-h., Jung, Y.-S., Nam, W., ... others (2020). Deep learning for automated detection of cyst and tumors of the jaw in panoramic radiographs. *Journal of clinical medicine*, 9(6), 1839.
- Yilmaz, S., Tasyurek, M., Amuk, M., Celik, M., & Canger, E. M. (2023). Developing deep learning methods for classification of teeth in dental panoramic radiography. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*.
- Yüksel, A. E., Gültekin, S., Simsar, E., Özdemir, Ş. D., Gündoğar, M., Tokgöz, S. B., & Hamamcı, İ. E. (2021). Dental enumeration and multiple treatment detection on panoramic x-rays using deep learning. *Scientific reports*, 11(1), 12342.
- Zhou, S. K., Greenspan, H., & Shen, D. (2023). *Deep learning for medical image analysis*. Academic Press.
- Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., & Wu, J. (2022). Cariesnet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic x-ray image. *Neural Computing and Applications*, 1–9.
- Zhu, W., Xie, L., Han, J., & Guo, X. (2020). The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3), 603.

