Evaluating AI Models in Dental Education: Potential for Learning and Clinical Training

Katharina Alves Rabelo

A thesis submitted to fulfil the requirements of the degree of Doctor of Philosophy

Sydney Dental School
Faculty of Medicine and Health
The University of Sydney

2025

Statement of Originality

This is to certify that the thesis content is my own work and has not been submitted for any other degree or purpose. I confirm that the intellectual content of this thesis is the result of my own work, and that all assistance received in its preparation and all sources used have been duly acknowledged.

Katharina Alves Rabelo

20 September 2025

Acknowledgements

The journey of pursuing a PhD has been both challenging both challenging and rewarding, but it has also been full of learning and meaningful experiences. This thesis would not have been possible without the guidance and support of my supervisor, Professor Vesna Miletic. I could not have asked for a better role model. Her inspiring approach to teaching, research, and leadership has significantly influenced my academic career. I am deeply grateful for her endless encouragement and mentorship.

I would also like to sincerely thank my co-supervisor, Professor Jinman Kim, for always providing valuable guidance and detailed feedback. The collaborative way of working as a team greatly supported and enriched my progress throughout this project.

I am deeply thankful to Dr. Eduardo Delamare, whose trust, guidance, and encouragement have been essential throughout my journey. His support has opened meaningful paths for my academic and professional growth, and I value the longstanding collaboration that has shaped my career.

The teamwork of computer scientists from the School of Computer Science at the University of Sydney and students from the Master of Digital Health and Data Science at the University of Sydney was invaluable in advancing my progress on this project. A special acknowledgement goes to Zimo Huang who contributed significantly by working on AI models for this thesis. This interdisciplinary collaboration greatly advanced the research, and it would not have been possible without his contributions.

I would like to thank you to my colleagues at Sydney Dental School, Dr Shwetha Hedge and Dr Amelita Simpson for supporting during this journey. Laura Swinckels deserves special mention for sharing her expertise and providing valuable suggestions that greatly supported my project. Thank you, Maha Aman, for encouraging me and offering valuable insights on my thesis. Extend my sincere thanks to dental assistants at the Bligh Building Simulation Clinic, Ashley, Eve, Judith, Megan, Cecille and Alex for their assistance and dedication in facilitating our research.

I am grateful to my friend, Carina Tanaka, for her support since the very beginning and for being an inspiration in the field of research.

A special thanks to my husband, Felipe, for your endless support, companionship and love. This thesis would not have been possible without him. I also thank my child, who has been a constant source of motivation and joy throughout this journey.

Authorship Attribution Statement

Chapter 1: The author compiled and wrote the chapter. No other authors contributed to this chapter.

Chapter 2: Rabelo KA, Huang Z, Delamare E, Kim J, Miletic V. Automated Detection and Classification of Adjacent Tooth Damage Using Deep Learning on Intraoral Images: Enhancing Pre-clinical Education in Restorative Dentistry. *(Submitted - under review)*

Chapter 3: **Rabelo KA**, Huang Z, Delamare E, Hedge S, Kim J, Miletic V. Automated Detection of Positioning Errors in Bitewing Radiographs Using Artificial Intelligence. *(Submitted - under review)*

Chapter 4: **Rabelo KA**, Swinckles L, Delamare E, Zhuoran D, Trinh E, Kim J, Miletic V. AI-Driven Feedback on Bitewing Radiographic Technique: A Comparative Study of Large Language Models. *(Submitted - under review)*

The chapter 2, 3 and 4 are listed as papers. As the first author of these manuscripts, I designed and conducted the studies and wrote the original draft of the manuscripts.

Chapter 5: The author compiled and wrote the chapter. No other authors contributed to this chapter.

Katharina Alves Rabelo

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Vesna Miletic

Artificial Intelligence Statement

I acknowledge the use of ChatGPT (https://chat.openai.com/), using GPT-4, to check grammar of my own work.

Australian Government Support Statement

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Table of Contents

Abstract

Artificial intelligence enhances efficiency and accuracy in clinical dentistry, and deep learning (DL) has achieved significant advancements in the field. However, its application in dental education remains underdeveloped. Dental education involves demanding practical training and assessments. Traditional visual-tactile methods are time-consuming, subjective and prone to inconsistencies.

This thesis investigated the applicability of DL tools for assessing tooth cavity preparation and intraoral radiographic techniques.

Three studies were conducted: (1) intraoral images captured with a commercial intraoral camera were used to assess damage to adjacent teeth during cavity preparations using a DL pipeline with YOLOv5 for detection and DenseNet-169 for classification; (2) convolutional neural network architectures were applied to student-acquired bitewings (BWs) to detect common positioning errors; (3) large language models (LLMs), ChatGPT o1, o3-mini, Gemini 2.0 and Grok 3, were evaluated in providing feedback on radiographic positioning errors using baseline and engineered prompts.

The DL models assessing intraoral images achieved 0.81 accuracy, outperforming clinical educators with excellent performance in detecting damage requiring restoration. The CNN model for identifying positioning errors in BWs achieved high accuracy: 96.3% for cone cutting, 93.4% for interproximal overlap, and 73.2% for incorrect receptor placement. LLMs showed variable but promising performance. Gemini 2.0 and Grok 3 performed best for interproximal overlap with baseline prompts, while prompt-engineered ChatGPT o1 and Gemini 2.0 performed better for incorrect film placement.

DL models demonstrated high performance in detecting adjacent-tooth damage in intraoral images and positioning errors on BWs. Open-source LLMs showed variable performance in analysing BW positioning errors. AI-supported assessment is applicable in training dental procedures involving intraoral and radiographic images.

Conference Presentations

1. ‘Reliability and Consistency of Clinical Educators’ Assessment of Dental Tasks in Tooth Conservation and Radiology: A Pilot Study’ at 2024 Sydney Dental School Research Day held at Westhead Hospital- Learning Studios on 4th of November 2024.

2. ‘AI-Driven Deep Learning in Dental Radiology: Advancing Education, Clinical Training, and Research’ at Oral Research Connect – June Session held at Susan Walkil Health Building on the 4th of June 2025.

3. ‘AI-Driven Deep Learning in Dental Education: Enhancing Education and Clinical Training’ at Sydney Dental School Research Day held at Innovation Centre Westmeath Hospital on 25th of August 2025.

List of figures

Chapter 2: *[Manuscript] Automated Detection and Classification of Adjacent Tooth Damage Using Deep Learning on Intraoral Images: Enhancing Pre-clinical Education in Restorative Dentistry*

Chapter 3: *[Manuscript] Automated Detection of Positioning Errors in Bitewing Radiographs Using Deep Learning*

Chapter 4: *[Manuscript] AI-Driven Feedback on Bitewing Radiographic Technique: A Comparative Study of Large Language Models*

List of tables

Chapter 2: *[Manuscript] Automated Detection and Classification of Adjacent Tooth Damage Using Deep Learning on Intraoral Images: Enhancing Pre-clinical Education in Restorative Dentistry*

Chapter 3: *[Manuscript] Automated Detection of Positioning Errors in Bitewing Radiographs Using Deep Learning*

Chapter 4: *[Manuscript] AI-Driven Feedback on Bitewing Radiographic Technique: A Comparative Study of Large Language Models*

List of Abbreviations

AI	Artificial Intelligence
ANNs	Artificial Neural Networks
API	Application Programming Interface
AR	Augmented reality
AUC	Area under the curve
BW	Bitewing radiograph
BWs	Bitewing radiographs
CAVs	Cloud Based AI-Driven Video Analytics
CBCT	Cone-beam computed tomography
CCE	Cone cutting error
CCWC	Computing and communication workshop and conference
CDSS	Clinical Decision Support System (CDSS)
CEs	Clinical educators
ChatGPT	Chatbot based on a Generative Pre-trained Transformer model
CNN	Convolutional neural network
CNNs	Convolutional Neural Networks
DL	Deep learning
DMFR	Dentomaxillofacial Radiology
DSS	Decision support systems
GANs	Generative adversarial networks’
GPT	Generative Pretrained Transformer
GPUs	Graphics Processing Units
ICCMS	International Caries Classification and Management System
ICDAS	International Caries Detection and Assessment System
ICEMS	Continuous Expertise Monitoring System
IoU	Intersection-over-Union
JPEG	Joint Photographic Experts Group
LLM	Large language model
LLMs	Large language models
mAP	Mean Average Precision
MBW	Molar bitewing
MCQs	Multiple-choice questions
ML	Machine Learning Machine learning
MR	Mixed reality
NLP	Natural language processing techniques
PMBW	Premolar bitewing
PSP	Photostimulable Phosphor plate
ROI	Region of interest
RTP	Research Training Program

SAQs	Short-answer questions
SPSS	Statistical Package for the Social Sciences
SSIM	Structural Similarity Index Measure
USMLE	United States Medical Licensing Exam
VR	Virtual reality
YOLO	You Only Look Once

Chapter 1: Introduction

1.1 Artificial Intelligence technologies

Artificial Intelligence (AI) refers to the capability of programming machines to perform functions traditionally associated with human intelligence. This includes learning, problem-solving, decision-making, and understanding language. AI can be classified in weak and strong AI. Weak or narrow AI includes systems designed and trained to perform specific tasks or solve particular problems, such as image recognition (1), language translation (2), virtual assistants (3) and automated conversational agents (4). These systems operate within a limited scope and do not possess general reasoning abilities. In contrast, strong AI, also called as artificial general intelligence, aims to replicate human intelligence with consciousness, self-awareness, cognitive flexibility, and the ability to reason across a wide range of tasks. A strong AI system would be capable of developing multi-task algorithms and making autonomous decisions across various domains. This type of AI remains a challenge, and to date, there is no AI model capable of meeting all the criteria.

1.1.1 Machine Learning

Machine learning (ML) is a subfield of narrow AI that focuses on developing and applying algorithms to solve pattern recognition problems by learning from data, without the need for explicit coding. It involves the study of algorithms capable of learning from experience and improves its performance through the analysis of patterns in the data (5, 6).

There are three types of ML: supervised, unsupervised and self-supervised or reinforcement learning. In supervised machine learning, a human must provide a substantial dataset containing examples with feature-label pairs to develop a model that predicts labels based on input features. Text classification, also referred to as text categorization or topic identification, is an example of a supervised learning task. In unsupervised machine learning, the objective is to find and understand patterns and then discover results without specific guidance. This type is suited for addressing association and clustering problems. Lastly, self-supervised ML involves utilising aspects of unlabelled data. It learns how to behave in an environment by taking actions and observing the output generated for those actions. It frequently produces representations that are fine-tuned for specific tasks (6, 7).

1.1.2 Deep Learning

The application of artificial intelligence (AI) is rapidly advancing, with deep learning technology, a subset of ML, experiencing notable growth. The deep learning techniques which apply deep neural networks have an architecture of a minimum of 3 layers including input, hidden layers and one output with interconnected nodes (artificial neurons) forming the neural network. This interconnected architecture allows each layer to extract features, pass them to the next layers and generate an output based on the analysis of these combined features. This architecture allows the AI model to process a large number of features in an unstructured data. Deep learning algorithms can be implemented using different learning approaches, supervised learning, unsupervised learning, and hybrid learning. The diverse applications of these algorithms lead to two main types of AI: deductive and generative.

1.1.2.1 Deductive AI

Deductive AI, analytical or discriminative AI, typically employs algorithms designed to analyse data and identify patterns (8). With advancements in deep learning technology, statistical models known as Artificial Neural Networks (ANNs) possess the ability to process raw data and independently extract features, thereby generating an output without the need for human intervention (5, 9). ANNs are effective in pattern recognition and prediction using structured data including speech recognition by converting voice to text and in the text-to-speech by generating human-like speech from a written text input. A specialised type of ANNs, Convolutional Neural Networks (CNNs), is designed to deal with grid-like data, such as image, effectively overcoming the challenge of image-driven pattern recognition. CNNs algorithms include convolutional and pooling layers within their hidden layer which allows them to learn features from raw pixels, making it suitable for image-focused tasks. The majority of CNN models rely on large annotated datasets, however some techniques can be used to enhance training data, such as semi-supervised learning, transfer learning, and data augmentation (10, 11).

The architecture and principles of CNN algorithms created the foundation for computer vision models which excels in tasks including image classification, object detection, and object segmentation. These tasks are mainly performed by models using

supervised or semi-supervised machine learning.

Classification models

The classification models are designed to categorise and label what the entire image represents. The computer can identify the class to which the object in the image belongs. These models are a type of supervised learning which require learning datasets that contain images and a class label for each. It requires less computational power than object detection models (10). Examples of classifiers are ResNet (12) and DenseNet (13) variants.

Object Detection models

The object detection models are designed to localise and classify objects by drawing bounding boxes around them. A larger training data is required than classification model. Annotations should include a bounding box including the object of interest with its corresponding class label (10, 14). Landmark models for this task are YOLO (You Only Look Once) that used the regression-based object detection algorithm (one-stage detectors) (15) and R-CNN family, per instance Faster R-CNN (16)), which has the region-based object detection algorithm (two-stage detectors) as the strategy for object detection (15).

Segmentation models

Segmentation models define the pixel-level boundaries of objects and regions. This type of CNN can outline a specific object in the image. This pixel-wise prediction demands substantial computational power and detailed image annotations. The training often requires sophisticated architectures and high-end Graphics Processing Units (GPUs) (17). The annotation process for segmentation models involves drawing or painting the image at the pixel level to create a 'mask' for each object or region. U-Net (18) and Mask R-CNN Mask R-CNN (19) are segmentation models that demonstrate high performance.

1.1.2.2 Generative AI

Generative AI represents an innovative approach designed to produce synthetic content without human supervision. In this context, one of common deep learning models is the ‘generative adversarial networks’ (GANs), which generates synthetic data by first learning the characteristics of real data using a generator network, and then distinguishing between real and fake data with a discriminator (20).

In computer vision field, style transfer and diffusion models represent two significant advancements in generative image synthesis. The image style transfer algorithm transforms the style of an image while preserving the structure of the content image; consequently, the final output combines the content of the input image with the desired style (21). Diffusion models are representing an emerging topic in computer vision. These models belong to a class of probabilistic generative models that generate images by starting with random visual ‘noise’ and then, step-by-step, refining until a clear image is formed (22).

Other examples of generative AI include natural language processing techniques (NLP) and large language models (LLMs). Natural language processing domain focus on the understanding and processing of human language by computers. At the earlier stages of NLP development, NLP systems mainly relied on rule-based programming or statistical models for tasks like basic translation and sentimental analysis. Nowadays, NLP systems can perform tasks including automatic translation, text summarization and text generation due to the Transformer architecture, a modern neural network. The transformer architecture is based on the attention layers and the ability to process all parts of an input sequence simultaneously, making it significantly more scalable and efficient for handling the massive datasets required for modern natural language processing tasks (23).

The efficiency and scalability of the Transformer architecture enabled the creation of Large Language Models (LLMs), which are highly advanced NLP systems trained on massive collections of text and code, usually including trillions of words sourced from websites, books, and scholarly articles (24).

Large Language Models

After extensive years of development, some LLMs, such as, Google's BARD.3–5 and OpenAI's Generative Pretrained Transformer (GPT), became publicly accessible as chatbots at the end of 2022 and the beginning of 2023, respectively. LLMs are constructed using ML principles for recognising and generating human-like language. They are a type of AI models based on the pre-trained Transformer architecture that was trained on substantial amounts of textual datasets. LLMs are undergoing fine-tuning using reinforcement learning from human feedback. They learn patterns of word usage in language, leveraging this understanding to proficiently perform various NLP tasks. In essence, this process enables LLMs to generate human-like language which can be applied to diverse NLP tasks. These tasks include language translation, text summarization, text classification and question-answering, among others. Additionally, it enhances text generation capabilities and facilitates human-machine interaction (7, 25). There are different LLMs available, Generative Pretrained Transformer (GPT) series developed by OpenAI company, including GPT-3.5, GPT-4, GPT-4.5 and GPT-4o are pre-trained with an autoregressive language modelling objective (26). While Gemini models from Google, in contrast, were fundamentally designed to be natively multimodal, pre-trained from the start on a diverse mix of text, images, video, audio, and code (27).

From 2024, LLMs experienced revolutionary advancements. OpenAI launched the GPT-4o which is a multimodal LLM able to handle text, audio and video as input and outputs. Additionally, the GPT-o1 was introduced, built on a multimodal foundation with enhanced reasoning capabilities and a greater degree of agency (28). Gemini 2.5 represents Google's parallel advancement in reasoning capabilities and agency (29).

A widely adopted application of LLMs is in the form of chatbots, which function as conversational agents capable of interacting with users through natural language, providing responses to questions, simulating conversations, and assisting with tasks. This technology holds significant potential for enhancing the quality of education and research (30). Applying NLP for language generation, understanding, and text classification to LLMs holds significant promise for enhancing dental education. This includes creating clinical scenarios tailored to various dental disciplines and generating feedback during simulated dental procedures, thereby enhancing

assessments and facilitating skill development.

1.2 Applications of AI in Higher Education

Various applications of AI have been studied and implemented across numerous sectors. In the field of education, the majority of AI applications employ deep learning technology, using deductive AI systems (8). An analysis of AI models in higher education through a meta-analysis of literature reveals that the majority of studies have utilised supervised ML algorithms for predicting and monitoring student academic progress. Among these, classification models in deep learning and ML are the most commonly used (31).

Generative AI shows significant potential in academia, using diverse AI technologies like NLP techniques to achieve educational goals (32, 33). NLP is employed in grading processes to enhance efficiency and consistency, eliminating the impact of human fatigue, and contributing to a reduction in subjectivity and bias. This involves analysing semantics and discourse to assess a student's comprehension of a subject or to automate the grading of short answers (30, 34).

LLMs are used in higher education to individualize learning by recommending strategies, developing adaptive learning systems, and identifying students' strengths and weaknesses (35, 36). The capabilities of ChatGPT in learning and teaching in higher education include their use as a search engine, assisting educators in developing curricula or weekly schedules, preparing course materials and assessments and offering personalised study tools for students; enhancing the learning experience and increase efficiency by their capability to summarise long texts, which enables fast access to key information and improve student outcomes (37, 38).

In addition, ChatGPT has been considered as a tool that supports effective self-directed learning and serves as an adjunct to enhance group-based educational activities in nursing school (36, 39). LLM-based chatbots have shown promising effectiveness and applicability in computer science education with positive feedback from students regarding their efficacy as educational tool (40).

Several studies (37, 41-43) have identified limitations of LLMs in the context of higher education, particularly regarding their ability to generate accurate, reliable, and unbiased outputs which require careful verification, especially in an academic setting.

The incorporation of LLMs as AI-driven educational tools has been discussed as a concern regarding cheating, academic dishonesty, data privacy, and unauthorized data collection. Ethical regulation and a balanced integration with human educators are essential (37, 44).

Despite the numerous AI-driven tools for higher education mentioned in the literature, a recent study highlights that only 2% of the studies on AI applications were associated with healthcare education (45).

1.3 AI Applications in Medical Education

In medical education, AI is mostly used in the teaching implementation stage; for instance, deep learning models are used for image recognition to assist medical students in image interpretation (46). Previous research has demonstrated that ML models not only are able to provide immediate and regular feedback for students, enabling them to track their progress, but also to enhance their psychomotor skills in surgery tasks (47, 48).

Three cloud-based artificial intelligence (AI)-driven video analytics platforms were tested: Touch Surgery™ (Medtronic, London, England, UK), Theatre (Palo Alto, California, USA) and C-SATS® (Seattle, Washington, USA). The evaluation aimed to assess their abilities in various aspects, including providing feedback and quantifiable skill-scoring systems for laparoscopic surgery. C-SATS® demonstrated the capability to create written feedback based on cloud video. Even though the study conducted a review of these platforms, relying solely on demonstrations and interviews with the platforms' creators, researchers concluded that with advances in AI, these tools can assess surgical tasks and provide essential feedback to enhance surgical practice (47).

Kayasth et al. (49) evaluated the performance of an AI system that provided

instantaneous feedback on a suturing task based on surgical videos. Authors compared the feedback generated by the AI model and explanations from human experts. The use of TWIX, a module for generating AI-based explanations similar to those provided by an expert, enhances the reliability of AI-based explanations.

Psychomotor performance in simulated subpial tumour resection using a virtual reality simulator was assessed using the Continuous Expertise Monitoring System (ICEMS), a machine learning model created in the study. A Long-Short Term Memory network analysed sixteen performance metrics by considering sequences of movements, using a dataset consisting of video recordings of the performed task. ICEMS effectively distinguished among neurosurgeons, experienced trainees, novice trainees, and medical students (48).

Wang et al. (50) developed a virtual learning system, Alteach, able to create virtual patients, simulating real clinical case scenarios. The system utilises real medical records and employs NLP technology to enhance clinical thinking training in the medical education field. Fifteen students participated in the experiment to evaluate the new system's efficiency, and it was observed that all students improved their critical thinking using virtual cases. The authors highlighted that the platform could serve as a complement in teaching hospitals, enabling them to use their substantial collection of data to improve the quality of learning.

LLMs are emerging as valuable AI-driven educational tools in medical education, with recent studies highlighting their diverse applications. ChatGPT and GPT-4 have demonstrated their ability to generate personalised learning experiences, improved comprehension of medical concepts, and facilitate the effective translation of radiology reports (32). Another study has shown that the integration of ChatGPT in medical education enhanced learning experiences with high levels of satisfaction (51). Additionally, ChatGPT can generate self-assessment quizzes with answer explanations about anatomy (43). Furthermore, ChatGPT-3.5, was able to design a realistic training scenario for breaking bad news in a simulated roleplay between a patient and a physician and provide real-time feedback to the user (52).

LLMs have demonstrated potential in passing medical exams, such as ChatGPT was

able to pass the written part of the United States Medical Licensing Exam (USMLE) (53) and obtained passing score on the Canadian Otolaryngology-Head and Neck Surgery Board exams (54).

1.4 AI Applications in Dental Education

Despite the growing implementation of AI in dentistry, with dental radiology and orthodontics currently leading the way, its application in the specific field of dental education remains largely underdeveloped (55, 56).

A systematic review on AI and virtual teaching models in dental education published in 2021 revealed that immersive tools such as virtual reality (VR), augmented reality (AR), mixed reality (MR), and haptic technology are the most predominant technologies in the field of dental education. Furthermore, studies using AI-driven tools are scarce in dental education (57). To be considered as AI systems, they should involve ML algorithms. While ML principles can be applied to AR or VR applications, these are distinct concepts and technologies.

An example of AI embedded in VR technology is a study done by Collaco et al. (58), who assessed the impact of inferior alveolar dental anaesthesia procedure in a haptic VR simulator using an ML method implemented to give immediate feedback on the student's performance regarding needle insertion point for inferior dental anaesthesia procedure. The ML method was trained with a sample of 50 observations. Classification as 'successful' was determined by drawing a prediction ellipse around the insertion point, while observations outside the prediction ellipse were classified as 'failure.' The model was then tested on 113 observations, resulting in a sensitivity of 83.6%, specificity of 84.5%, and accuracy of 84%. The authors reported that these values indicate the prediction ellipse performed well and validated the ML method.

An ML method for the assessment of Class II amalgam and composite resin restorations done by students was applied and compared to dental educator’s evaluation. The Structural Similarity Index Measure (SSIM), a statistical image similarity metric rather than a deep learning algorithm was used. The compatibility between supervisor evaluations and SSIM analysis was reported as ‘almost perfect’ for amalgam restorations and ‘substantial’ for composite restorations, indicating its potential as an objective tool for educational assessment (59).

Computer vision models employing CNN algorithms have been explored for use in clinical dentistry. They have demonstrated good performance in several applications,

including the assessment of intraoral photographic images for tooth surface and caries detection (60), tooth number recognition and caries detection (61), identification of dental anomalies (62), crowding categorization and extraction diagnosis (63), detection of dental caries and fissure sealants (64), dental plaque detection (65), detection of oral dysplasia area (66) and tooth shade assessment (67).

In the field of DMFR the use of CNNs is extensive. CNNs have demonstrated strong performance in computer vision tasks such as detection, classification, and segmentation, particularly for identifying oral pathologies in radiographs and tomographs. For caries classification, deep learning models trained on radiographs have shown excellent performance, with reported accuracies ranging from 82% to 99% on periapical and panoramic images, and from 68.7% to 94.6% on bitewing radiographs (68). A recent meta-analysis assessed the performance of deep learning models on the assessment of alveolar bone loss and periodontitis on panoramic and periapical radiographs. The overall accuracy across all studies was 0.84 (69). YOLOv8 achieved accuracy of 96.22% on the detection of calculus in bitewings radiographs (70). Deep learning algorithms have been employed for detecting periapical lesions on panoramic radiographs, achieving accuracies above 90% (71). CNN architecture achieved more than 90% of accuracy in detecting and classifying dentigerous and periapical cysts on cone beam computed tomography images using data augmentation (72).

The potential of CNNs for educational purposes, particularly with radiographs, has been investigated. For example, a Siamese neural network was used to detect inconsistencies in radiographs of typodont teeth during root canal treatment preclinical sessions, achieving nearly 90% accuracy. The model has the potential to enhance fairness and reliability of assessment while reducing the administrative burden on educators (73). Another study has demonstrated the ability of object detection CNN to assess the quality of root canal filling performed by dental students from periapical radiographs. YOLOv11m architecture achieved 77.51% precision and 79.03% recall (74). Additionally, a recent study has shown that an AI-augmented radiographic training module improved dental student’s diagnostic accuracy for proximal caries on radiographs by 35% (75).

Several studies have been assessing the performance of LLMs as chatbots for dental education. A recent study has explored the role of Chatbot GPT in undergraduate dental education highlighting its integration in dental education as virtual patients to support communication skills and clinical reasoning. ChatGPT can facilitate realistic patient interactions simulating different clinical scenarios and offering instant feedback on diagnostic and treatment plans. This contributes to inclusivity and fosters a learner-centred approach (76).

GPT-4 Turbo was implemented in an AI platform developed by the University of Sydney, Cogniti, to enhance dental student's history-taking skills. This platform allows the creation AI chatbot agents by uploading relevant resources and customizing prompts. It was implemented in first- and second-year Doctor of Dental Medicine students, who reported that it offered more practice opportunities, although they found it less interesting than in-person teaching (77).

Another study designed an AI-based role-play using Streamlit framework powered by ChatGPT-4. Oral health students reported that the activity was relevant to clinical practice, with majority agreeing that it reflected real dental scenarios. Educators observed that the learning method facilitated peer discussion and promoted peer learning (78).

The accuracy of chatbots in answering different types of questions across various dental subjects has been investigated. ChatGPT3.5 and ChatGPT4 have shown accuracy levels similar to dental surgeons trained in oral medicine and pathology when providing differential diagnoses for oral and maxillofacial diseases. ChatGPT-4 achieved 80.18% accuracy, while oral medicine/pathology specialists reached 86.64% (79).

In the discipline of periodontology, ChatGPT3.5 and ChatGPT4 were assessment on multiple-choice questions and demonstrated accuracy of 57.9% and 73.6%, respectively (80).

Publicly accessible chatbots including Google Bard, ChatGPT4, ChatGPT 3.5, Llama, Sage, Claude 2 100k, Claude-instant, Claude-instant-100k, and Google Palm were

evaluated using True and false questions in the field of paediatric dentistry. Their outputs were assessed by two paediatric dentistry academics and compared with responses from three groups: general dentists, paediatric specialists and students. ChatGPT-4 demonstrated the highest accuracy at approximately 78% and all chatbots, expect ChatGPT-3.5, showed acceptable consistency (81).

The accuracy of outputs from chatbots for prosthodontic education varies according to the question content. A study evaluating the performance of Copilot, Gemini, chatgpt-3.5, Claude Pro and Perplexity found that questions on removable partial dentures had the lowest accuracy, while those related to dental implantology achieved the highest accuracy (75%). This study highlights the limitations of LLMs as reliable education tools in prosthodontics and emphasises the need for further advancements to enable better integration in dental education (82).

In endodontic education, an analysis of current studies revealed limited research on chatbot performance and emphasised the need for customised chatbots, which should be validated for accuracy and relevance prior to implementation (83).

The ability of LLMs in passing dental licensing exams has been assessed. A strong performance of GPT-4 on multiple-choice assessments was observed, with the model achieving marks above the minimum passing score in dental licensing exams in the United States and the United Kingdom (84). Similarly, both GPT-4 and Claude3-Opus attained scores above the required cut-off and showed particularly high proficiency in answering multiple-choice questions in selected subjects of the Korean Dental Licensing Examination. However, all LLMs under evaluation demonstrated lower performance compared to students (85). A recent analysis comparing ChatGPT and Google Bard for dental education purposes highlighted variability in LLM performance, emphasizing the need for targeted training in evidence-based content generation (86).

Specific in the context of Dentomaxillofacial Radiology (DMFR) education, a comparative study assessed LLMs including ChatGPT, ChatGPT Plus, Bard, and Bing Chat against students. All chatbots showed low performance, achieving only 35% accuracy on a descriptive question related to image interpretation. The study concluded that chatbot performance in oral and maxillofacial radiology was

unsatisfactory (87). Another study compared student and ChatGPT versions 3.5 and 4 performances in answering DMFR questions. Chatbots performed better on multiple-choice than on open-ended questions. ChatGPT-4 matched 4th-year dental students, scoring 33.3% and 20% in multiple-choice and open-ended questions, respectively. The authors concluded that ChatGPT's knowledge in DMFR is currently unreliable (88).

Despite the growing implementation of AI in clinical dentistry its application in dental education field remains largely underdeveloped. AI has potential to support solutions in several key dental education challenges.

Dental education involves demanding practical training. Most dental procedures require precise manual dexterity and hand–eye coordination, which students must develop within a short timeframe. Preclinical and clinical procedures performed by students are evaluated to provide feedback during practical sessions or for use in practical assessments. Traditional assessment methods are currently used. They rely on visual and tactile inspections, which is time-consuming and subjective. Considering the limited teaching resources, providing individual guidance throughout the simulation session is a challenge, often leading to students having to wait for assistance. Additionally, there are calibration issues with the clinical educators, potentially impacting the consistency and effectiveness of the training.

AI-driven evaluation tools have the potential to enhance evaluation by providing rapid, objective, standardized feedback and less-costly assessment. Furthermore, studies have shown that AI can make the learning process more effective and engaging (89, 90). In addition, applying criteria consistently, eliminating human subjectivity AI-driven grading tools ensures fairness. The integration of AI in cloud-based education systems can support unsupervised practical sessions and remote learning. This area requires significant research efforts to allow for evidence-based teaching and policy.

The application of deep learning algorithms to support teaching and assessment in dental procedures, such as tooth cavity preparation and intraoral radiograph techniques, has not yet been largely investigated. Moreover, research on the use of

LLM outputs to provide feedback during practical sessions is rare.

1.5 Aim and objectives

The aim of this thesis was to develop AI-driven deep learning tools for dental education.

The thesis contains three objectives.

- **Objective 1** is to compare traditional assessment methods with deep learning models to detect a key critical issue (damage on tooth adjacent to a Class II cavity) in tooth preparation skills using intraoral cameras. Objective 1 is addressed and covered in Chapter 2.

- **Objective 2** is to develop and evaluate CNN architectures for detecting and classifying common positioning errors in bitewings (BWs). Objective 3 is addressed in Chapter 3 of this thesis.

- **Objective 3** is to compare the performance of publicly available LLMs in providing feedback on positioning errors in BW radiographs. Objective 3 is addressed in Chapter 4.

1.6 Significance of the thesis

The outcomes of this thesis are critical for advancing the understanding and applicability of deep learning algorithms in providing feedback during practical clinical sessions and supporting practical assessments. The findings are valuable for guiding the fine-tuning of deep learning models for tooth cavity preparation through the analysis of intraoral images, and for intraoral radiograph technique evaluation. They also support the integration of additional cavity preparation criteria and intraoral radiographic techniques, contributing to the development of an AI-supported assessment system. Regarding publicly available LLMs, the outcomes provide evidence of their capability to identify and generate outputs useful for feedback on radiographic techniques.

1.7 Thesis outline

This thesis contains four chapters and is structured in the following manner:

- **Chapter 1** (current chapter) introduces the thesis by providing a literature review including a concise background on the key aspects of the topics, summarises the gaps in the literature and explains the objectives and significance of the thesis.

- **Chapter 2** investigates the clinical educator's assessments and compares them with CNNs model evaluations of tooth damage adjacent to cavity preparations using intraoral camera images. This study used one of key assessment criteria to guide the selection and training of CNN algorithms for an AI-driven assessment tool.

- **Chapter 3** presents a more advanced training and evaluation of CNN algorithms for detecting and classifying the type, side and common positioning errors in bitewing radiographs.

- **Chapter 4** investigates the performance of publicly available LLMs and compares their ability to provide feedback on positioning errors in bitewing radiographs. Furthermore, it demonstrates and discusses the use of prompt design techniques in LLMs with reasoning capabilities to enhance output.

1.8 References

1. Chen H, Geng L, Zhao H, Zhao C, Liu A. Image recognition algorithm based on artificial intelligence. Neural Computing and Applications. 2022;34(9):6661-72.
2. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:160908144. 2016.
3. Kepuska V, Bohouta G, editors. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). 2018 IEEE 8th annual computing and communication workshop and conference (CCWC); 2018: IEEE.
4. Adamopoulou E, Moussiades L. Chatbots: History, technology, and applications. Machine Learning with applications. 2020;2:100006.
5. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty. 2018;33(8):2358-61.
6. Zhang A, Lipton ZC, Li M, Smola AJ. Dive into deep learning: Cambridge University Press; 2023.
7. Lidströmer N, Aresu F, Ashrafian H. Basic Concepts of Artificial Intelligence: Primed for Clinicians. In: Lidströmer N, Ashrafian H, editors. Artificial Intelligence in Medicine. Cham: Springer International Publishing; 2022. p. 3-20.
8. Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthc J. 2022;9(2):190-3.
9. Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015;61:85-117.
10. Zhao X, Wang L, Zhang Y, Han X, Deveci M, Parmar M. A review of convolutional neural networks in computer vision. Artificial Intelligence Review. 2024;57(4):99.
11. O'shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint arXiv:151108458. 2015.
12. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ, editors. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
14. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep Learning for Computer Vision: A Brief Review. Comput Intell Neurosci. 2018;2018:7068349.
15. Redmon J, Divvala S, Girshick R, Farhadi A, editors. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
16. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137-49.
17. Yu Y, Wang C, Fu Q, Kou R, Huang F, Yang B, et al. Techniques and Challenges of Image Segmentation: A Review. Electronics. 2023;12(5):1199.
18. Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015: Springer.
19. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):386-97.
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Communications of the ACM. 2020;63(11):139-44.
21. Liu L, Xi Z, Ji R, Ma W. Advanced deep learning techniques for image style transfer: A survey. Signal Processing: Image Communication. 2019;78:465-70.
22. Croitoru F-A, Hondru V, Ionescu RT, Shah M. Diffusion models in vision: A survey. IEEE transactions on pattern analysis and machine intelligence. 2023;45(9):10850-69.

23. Gillioz A, Casas J, Mugellini E, Abou Khaled O, editors. Overview of the Transformer-based Models for NLP Tasks. 2020 15th Conference on computer science and information systems (FedCSIS); 2020: IEEE.
24. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
25. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. Meta-Radiology. 2023;1(2):100017.
26. OpenAi. GPT-4.1, GPT-4.5, and GPT-4o Overview. 2025.
27. Team G, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.
28. Wu S, Peng Z, Du X, Zheng T, Liu M, Wu J, et al. A comparative study on reasoning patterns of OpenAI's o1 model. arXiv preprint arXiv:241013639. 2024.
29. Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:250706261. 2025.
30. Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, Bin Saleh K, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Social Adm Pharm. 2023;19(8):1236-42.
31. Fahd K, Venkatraman S, Miah SJ, Ahmed K. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Education and Information Technologies. 2022;27(3):3743-75.
32. Bahroun Z, Anane C, Ahmed V, Zacca A. Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings through Bibliometric and Content Analysis. Sustainability. 2023;15(17):12983.
33. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J. 2023;3(1):e103-e.
34. Bonthu S, Rama Sree S, Krishna Prasad MHM. Improving the performance of automatic short answer grading using transfer learning and augmentation. Engineering Applications of Artificial Intelligence. 2023;123:106292.
35. Baars M, Khare S, Ridderstap L. Exploring students' use of a mobile application to support their self-regulated learning processes. Frontiers in Psychology. 2022;13:793002.
36. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT-Reshaping medical education and clinical management. Pakistan Journal of Medical Sciences. 2023;39(2):605.
37. Yigci D, Eryilmaz M, Yetisen AK, Tasoglu S, Ozcan A. Large language model-based chatbots in higher education. Advanced Intelligent Systems. 2025;7(3):2400429.
38. Hariri W. Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. arXiv preprint arXiv:230402017. 2023.
39. O'Connor S. GPT.(2023). Open artificial intelligence platforms in nursing education: tools for academic progress or abuse. Nurse Educ Pract.66.
40. Neumann AT, Yin Y, Sowe S, Decker S, Jarke M. An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. IEEE Transactions on Education. 2025;68(1):103-16.
41. Salman I, Ameer O, Khanfar M, Hsieh Y-H. Artificial intelligence in healthcare education: evaluating the accuracy of ChatGPT, Copilot, and Google Gemini in cardiovascular pharmacology. Frontiers in Medicine. 2025;12.
42. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 2023;11(6):887.
43. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci

Educ. 2024;17(5):926-31.
44. Guizani S, Mazhar T, Shahzad T, Ahmad W, Bibi A, Hamam H. A systematic literature review to implement large language model in higher education: issues and solutions. Discover Education. 2025;4(1):35.
45. Crompton H, Burke D. Artificial intelligence in higher education: the state of the field. International Journal of Educational Technology in Higher Education. 2023;20(1):22.
46. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in Medical Education: Global situation, effects and challenges. Education and Information Technologies. 2023:1-23.
47. Gendia A. Cloud Based AI-Driven Video Analytics (CAVs) in Laparoscopic Surgery: A Step Closer to a Virtual Portfolio. Cureus. 2022;14(9):e29087.
48. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. npj Digital Medicine. 2022;5(1):54.
49. Kiyasseh D, Laca J, Haque TF, Miles BJ, Wagner C, Donoho DA, et al. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. Commun Med (Lond). 2023;3(1):42.
50. Wang M, Sun Z, Jia M, Wang Y, Wang H, Zhu X, et al. Intelligent virtual case learning system based on real medical records and natural language processing. BMC Med Inform Decis Mak. 2022;22(1):60.
51. Thomae AV, Witt CM, Barth J. Integration of ChatGPT Into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications. JMIR Med Educ. 2024;10:e50545.
52. Webb JJ. Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News. Cureus. 2023;15(5):e38755.
53. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
54. Long C, Lowe K, Santos Ad, Zhang J, Alanazi A, O'Brien D, et al. Evaluating ChatGPT-4 in otolaryngology–head and neck surgery board examination using the CVSA model. MedRxiv. 2023:2023.05. 30.23290758.
55. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of Artificial Intelligence on Dental Education: A Review and Guide for Curriculum Update. Education Sciences. 2023;13(2):150.
56. Arsiwala-Scheppach LT, Chaurasia A, Müller A, Krois J, Schwendicke F. Machine Learning in Dentistry: A Scoping Review. J Clin Med. 2023;12(3).
57. Saghiri MA, Vakhnovetsky J, Nadershahi N. Scoping review of artificial intelligence and immersive digital tools in dental education. J Dent Educ. 2022;86(6):736-50.
58. Collaço E, Kira E, Sallaberry LH, Queiroz AC, Machado MA, Crivello Jr O, et al. Immersion and haptic feedback impacts on dental anesthesia technical skills virtual reality training. Journal of Dental Education. 2021;85(4):589-98.
59. Oguzhan A, Peskersoy C, Devrimci EE, Kemaloglu H, Onder TK. Implementation of machine learning models as a quantitative evaluation tool for preclinical studies in dental education. Journal of Dental Education. 2025;89(3):383-97.
60. Park EY, Cho H, Kang S, Jeong S, Kim E-K. Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. BMC Oral Health. 2022;22(1):573.
61. Yoon K, Jeong H-M, Kim J-W, Park J-H, Choi J. AI-based dental caries and tooth number detection in intraoral photos: Model development and performance evaluation. Journal of Dentistry. 2024;141:104821.
62. Ragodos R, Wang T, Padilla C, Hecht JT, Poletta FA, Orioli IM, et al. Dental anomaly detection using intraoral photos via deep learning. Scientific Reports. 2022;12(1):11577.
63. Ryu J, Kim Y-H, Kim T-W, Jung S-K. Evaluation of artificial intelligence model for crowding categorization and extraction diagnosis using intraoral photographs. Scientific

Reports. 2023;13(1):5177.
64. Xiong Y, Zhang H, Zhou S, Lu M, Huang J, Huang Q, et al. Simultaneous detection of dental caries and fissure sealant in intraoral photos by deep learning: a pilot study. BMC Oral Health. 2024;24(1):553.
65. You W, Hao A, Li S, Wang Y, Xia B. Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments. BMC Oral Health. 2020;20(1):141.
66. Camalan S, Mahmood H, Binol H, Araújo ALD, Santos-Silva AR, Vargas PA, et al. Convolutional Neural Network-Based Clinical Predictors of Oral Dysplasia: Class Activation Map Analysis of Deep Learning Results. Cancers (Basel). 2021;13(6).
67. Karcioglu AA, Efitli E, Simsek E, Ozdogan A, Karatas F, Senocak T. ML-based tooth shade assessment to prevent metamerism in different clinic lights. Lasers in Medical Science. 2025;40(1):39.
68. Liang Y, Li D, Deng D, Chu CH, Mei ML, Li Y, et al. AI-Driven Dental Caries Management Strategies: From Clinical Practice to Professional Education and Public Self Care. International Dental Journal. 2025;75(4):100827.
69. Khubrani YH, Thomas D, Slator PJ, White RD, Farnell DJJ. Detection of periodontal bone loss and periodontitis from 2D dental radiographs via machine learning and deep learning: systematic review employing APPRAISE-AI and meta-analysis. Dentomaxillofacial Radiology. 2025;54(2):89-108.
70. Lin TJ, Lin YT, Lin YJ, Tseng A, Lin CY, Lo LT, et al. Auxiliary Diagnosis of Dental Calculus Based on Deep Learning and Image Enhancement by Bitewing Radiographs. BIOENGINEERING-BASEL. 2024;11(7).
71. Ver Berne J, Saadi SB, Oliveira-Santos N, Marinho-Vieira LE, Jacobs R. Automated classification of panoramic radiographs with inflammatory periapical lesions using a CNN-LSTM architecture. Journal of Dentistry. 2025;156:105688.
72. Esmaeilyfard R, Esmaeeli N, Paknahad M. An artificial intelligence mechanism for detecting cystic lesions on CBCT images using deep learning. Journal of Stomatology, Oral and Maxillofacial Surgery. 2025;126(6):102152.
73. Ibrahim M, Omidi M, Guentsch A, Gaffney J, Talley J. Ensuring integrity in dental education: Developing a novel AI model for consistent and traceable image analysis in preclinical endodontic procedures. Int Endod J. 2025.
74. Ayhan M, Kayadibi İ, Aykanat B. RCFLA-YOLO: a deep learning-driven framework for the automated assessment of root canal filling quality in periapical radiographs. BMC Medical Education. 2025;25(1):894.
75. Parekh DK, Gohel DA. Augmented teaching of the next-generation-dentists in diagnosis of common dental diseases with the overjet artificial intelligence (AI) module. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. 2025;139(3):e71-e2.
76. Thorat VA, Rao P, Joshi N, Talreja P, Shetty A. The Role of Chatbot GPT Technology in Undergraduate Dental Education. Cureus. 2024;16(2):e54193.
77. Or A, Sukumar S, Ma A, Ang D, Liu M, Ritchie HE, et al. Enhancing Dental Students' History-taking Skills with a Generative Artificial Intelligence Chatbot. Journal of Dental Education. 2025;n/a(n/a):e13952.
78. Jones B, Desu A, Honig CDF. Artificial Intelligence Chatbots as Virtual Patients in Dental Education: A Constructivist Approach to Classroom Implementation. Eur J Dent Educ. 2025;n/a(n/a).
79. Tomo S, Lechien JR, Bueno HS, Cantieri-Debortoli DF, Simonato LE. Accuracy and consistency of ChatGPT-3.5 and −4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnostic performance analysis. Clin Oral Investig. 2024;28(10):544.
80. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. Journal of Periodontology. 2024;95(7):682-7.
81. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke

F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. Journal of Dentistry. 2024;144:104938.
82. Eraslan R, Ayata M, Yagci F, Albayrak H. Exploring the potential of artificial intelligence chatbots in prosthodontics education. BMC Medical Education. 2025;25(1):321.
83. Mohammad-Rahimi H, Setzer FC, Aminoshariae A, Dummer PMH, Duncan HF, Nosrat A. Artificial intelligence chatbots in endodontic education—Concepts and potential applications. International Endodontic Journal. 2025;n/a(n/a).
84. Chau RCW, Thu KM, Yu OY, Hsung RT-C, Lo ECM, Lam WYH. Performance of Generative Artificial Intelligence in Dental Licensing Examinations. Int Dent J. 2024;74(3):616-21.
85. Kim W, Kim BC, Yeom HG. Performance of Large Language Models on the Korean Dental Licensing Examination: A Comparative Study. Int Dent J. 2025;75(1):176-84.
86. Aldukhail S. Mapping the Landscape of Generative Language Models in Dental Education: A Comparison Between ChatGPT and Google Bard. European Journal of Dental Education. 2025;29(1):136-48.
87. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol. 2024;53(6):390-5.
88. Öztürk HP, Avsever H, Şenel B, Ayran Ş, Peker MÇ, Özgedik HS, et al. ChatGPT in dentomaxillofacial radiology education. J Health Sci Med. 2024;7(2):224-9.
89. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Academic Medicine. 2021;96(11S):S62-S70.
90. Hu C, Li F, Wang S, Gao Z, Pan S, Qing M. The role of artificial intelligence in enhancing personalized learning pathways and clinical training in dental education. Cogent Education. 2025;12(1):2490425.
1.

Chapter 2: *[Manuscript] Automated Detection and Classification of Adjacent Tooth Damage Using Deep Learning on Intraoral Images: Enhancing Pre-clinical Education in Restorative Dentistry*

Automated Detection and Classification of Adjacent Tooth Damage Using Deep Learning on Intraoral Images: Enhancing Pre-clinical Education in Restorative Dentistry

Katharina Alves Rabelo[1], Zimo Huang[2], Eduardo Delamare[1], Jinman Kim[2], Vesna Miletic[1]

[1] Faculty of Medicine and Health, Sydney Dental School, The University of Sydney, Sydney, Australia

[2] Faculty of Engineering, School of Computer Science, The University of Sydney, Sydney, Australia

Short title: Deep Learning for Adjacent Tooth Damage Detection

*Corresponding author: Katharina Alves Rabelo

Faculty of Medicine and Health, Sydney Dental School, The University of Sydney.

Address: 2 Chalmers Street, Surry Hills NSW 2010, Australia

E-mail: katharina.alvesrabelo@sydney.edu.au

Acknowledgements

We acknowledge the academic expertise of the Tooth Conservation clinical educators at the University of Sydney, Faculty of Medicine and Health, Sydney Dental School. The authors are grateful to Elvis Trinh, a lecturer at Sydney Dental School, for his help with data acquisition and labelling. Additionally, we extend our sincere thanks to the dental assistants at the Bligh Building Simulation Clinic for their assistance and dedication in facilitating data collection.

Abstract

Introduction: Tooth preparation involves intricate high-precision cutting with potential for proximal damage to adjacent teeth, which remains an issue despite protective measures. Traditional educator assessments are often subjective, relying on visual and tactile inspections. In contrast, artificial intelligence (AI) enhances efficiency and accuracy in various clinical applications, yet its use in evaluating dental procedures via intraoral images remains largely unexplored. This study compared assessment methods for adjacent tooth damage and applied deep learning models to automatically detect and classify tooth surface damage using intraoral images.

Materials and Methods: The study evaluated intra- and inter-examiner reliability and the correlation between model- and image-based assessments. Dental students performed Class II cavity preparations on simulation models, and adjacent teeth were imaged using a commercial intraoral camera. Damage was classified as none, requiring restoration, or not requiring restoration. The AI model used YOLOv5 for object detection and DenseNet-169 for classification with cross-validation. Accuracy, sensitivity, precision, and F1 scores were calculated for each method.

Results: Fleiss' kappa showed moderate agreement among clinical educators for both model and image assessments. Inter-examiner reliability was higher for model-based than image-based assessments at corresponding time points. The AI model achieved 0.81 accuracy, significantly outperforming human assessments. Notably, it perfectly detected 'damage requiring restoration' but was less accurate for 'damage not requiring restoration'.

Conclusion: AI-driven image analysis significantly enhanced assessment accuracy and feedback consistency. This advancement holds potential to guide future developments in AI-supported dental training and practice with intraoral images.

Keywords: Artificial intelligence; Deep learning; Convolutional Neural Networks; Decision support system; Restorative dentistry; Dental education

Introduction

Practical dental education is demanding, as it requires highly precise manual dexterity and hand-eye coordination to be developed in a relatively short time. Tooth preparation is particularly challenging for dental students due to milimeter-sized movements using highspeed rotary burs across a small area of the tooth. Furthermore, there is often obscured accessibility and the proximity of other tissues associated with tooth preparation. Proximal damage of adjacent teeth has been reported as an ongoing issue regardless of protection used [1,2].

Practical education is guided by assessment and feedback provided by clinical educators (CEs). This is performed using visual and tactile inspection with dental instruments and a set of predefined criteria. It is widely known that subjective assessment is unreliable, not just in dentistry [3,4] but also in medical education [5,6].

Attempts to provide objective assessments in dental education include digital tools for cavity preparations, crown preparations, and other dental tasks. For instance, systems like Dental Teacher™ and PrepCheck® have been used to evaluate cavity preparations on plastic teeth, demonstrating their potential to improve learning efficiency and individual performance through digital feedback [7]. Similarly, the Preppr software showed significant advantages for self-directed learning, with students achieving acceptable preparation standards more efficiently than traditional methods [8]. Insufficient or inconsistent feedback generally in clinical assessment has significant implications. For example, lack of strict guidelines on whether a damaged adjacent tooth requires a restoration may create plaque-retentive surfaces increasing caries risk. On the other hand, invasively restoring minor damage unnecessarily increases the cost and the duration of treatment.

Artificial intelligence (AI) offers significant benefits enhancing both efficiency and accuracy in numerous clinical applications. AI-powered clinical diagnostic decision support systems (DSS) have made tremendous advancements through deep learning algorithms including Convolutional Neural Networks (CNNs). These networks build deep models by training on large datasets, allowing them to learn useful image features to identify various levels of image representation [9].

These AI advancements have been applied to dental imaging for various DSS applications including caries detection [10], dental plaque detection [11] and oral dysplasia [12]. The results have shown 68-73% accuracy [10,12] and mean intersection-over-union of 0.724 [11] indicating good performance and the potential use of AI technology in oral health applications. Despite advancements in AI in dentistry, the specific integration of AI models for the assessment of dental procedures, such as, cavity tooth preparation, remains an unexplored area. Particularly important application would be the use of commercially available intraoral cameras for image acquisition to be analysed by a deep learning algorithm. This may have implications for clinical practice in which AI-assisted clinical decision-making would use images obtained by intraoral cameras. Our extensive literature review did not find data on integration of commercially available intraoral cameras and AI algorithms in dental practice and education, particularly in the context of cavity preparation assessments.

The aims of the study were to (1) investigate CE's assessment of tooth damage adjacent to a cavity preparation and (2) to apply deep learning models for the automated detection and classification of tooth surface damage using intraoral camera images. The working hypothesis tested in this study was that the AI-based model would demonstrate higher performance metrics compared to CE assessment methods. Additionally, the following null hypotheses were tested: (1) there is no statistically significant difference in inter- and intra-examiner reliability assessments of damage on tooth adjacent to a Class II cavity preparation (H0 #1), (2) there is no statistically significant difference in reliability between two time points at least three weeks apart (H0 #2), and (3) there is no statistically significant difference in reliability between assessment methods (dental model- and image-based) (H0 #3), (4) there is no significant correlation between dental model- and image-based assessment methods (H0 #4).

Materials and Methods

This study was reviewed and approved by the University Human Ethics Committee (2024/HE000118).

Sample preparation and Image acquisition

At a dental simulation clinic, 402 Class II cavities were prepared by dental students on the first molars in dental simulation models during their tooth conservation practical simulation sessions. Each dental model consisted of a pair of jaws (upper and lower) with typodont adult teeth and soft gingivae (Columbia Dentoform, USA), mounted on a mannequin head within a dental simulation unit.

Images of the distal surfaces of the second premolars adjacent to Class II cavities were captured using an intraoral camera SiroCam UAF Plus (Dentsply Sirona, n.d.), which features auto-focus and captures images at a resolution of 1276 x 796 pixels (720p HD resolution). This camera is commercially available and accessible to students in the University of Sydney dental simulation clinic at each dental simulation unit.

The inclusion criteria required that the images show the entire distal surface adjacent to the proximal box of the Class II cavity and that the images be well-focused. If these criteria were not met — for instance, if the image was blurry or did not include the full distal surface — the image was retaken.

Image classification/annotation

The distal surfaces of the second premolars adjacent to the cavity preparations were categorised into three groups: (1) no damage, (2) damage requiring restoration, and (3) damage not requiring restoration. Surfaces with scratches were classified as those that do not require restoration, while small cavities (0.5 mm and deeper) were classified as those that do require restoration (Figure 1).

Figure 1. Image representing the types of tooth damage. (A) no damage, (B) damage not requiring restoration, and (C) damage requiring restoration.

The "gold standard" classification was determined by the consensus of three academics senior CEs with expertise in restorative dentistry. These educators assessed the dental models during the practical session in the same manner they typically do for grading purposes.

Human assessment

A total of 16 Class II cavities created on first molars attached to dental simulation models, were selected. Corresponding intraoral images of the distal surfaces of the premolars were taken using a standardised setup with the intraoral camera available. The same inclusion criteria described above were applied, and any image that did not fulfilled the inclusion criteria was deleted and retaken.

Eighteen CEs from the Tooth Conservation team in the dental simulation clinic assessed the distal surfaces of 16 second premolars and evaluated the images of the same distal surfaces at two different times at least three weeks apart. The images were randomly organised to prevent examiners from recognizing the tooth on the model.

The assessment was performed at the Simulation Clinic at the University of Sydney and recorded in two different surveys created in Redcap: one for dental model-based assessment and one for the image-based assessment. Each survey question asked the examiners to classify the damage ((1) no damage, (2) damage requiring restoration, and (3) damage not requiring restoration)) on the distal area of the second premolars in a specific model and tooth number.

The dental model-based assessment was conducted using visual inspection and tactile sensation with a dental explorer n.6 (Nordent Manufacturing Inc., n.d.) with the model placed on a flat surface. CES had the option to evaluate the distal surfaces with the naked eye or dental loupes, which had to be recorded in the survey and remain consistent between the first and second assessment times. They were allowed to manipulate the model and reposition the operator light for better assessment.

For the image-based assessment, images were projected in full-screen mode on a monitor with a 21.5-inch liquid crystal display (AG Neovo n.d.). Images were organised in the same sequence as in the survey; however, CEs were allowed to navigate through the images independently during the assessment time, could assess the images at their own pace and were permitted to advance or revisit the images as necessary for a thorough evaluation.

Statistical Analysis

Human Assessment Reliability

Data were analysed using SPSS version 22 (SPSS Inc., IL, USA). To analyse the intra- and inter-examiner reliability of damage assessments, Cohen's Kappa and Fleiss's Kappa values were calculated, respectively. The classification of Cohen's kappa values and their corresponding strength of agreement is based on Altman (1990) [13] and adapted from Landis and Koch (1977), as follows: a kappa value less than 0.20 indicates poor agreement, values between 0.21 and 0.40 reflect fair agreement, values ranging from 0.41 to 0.60 represent moderate agreement, and values between 0.61 and 0.80 signify good agreement. Finally, kappa values from 0.81 to 1.00 indicate very good agreement. The correlation between the dental model- and image-based assessments were investigated by means of the Pearson's correlation after confirmation that both variables were normally distributed using the scatterplot. A 0.05 level of significance was adopted.

Convolutional neural network (CNN) algorithm

The automated assessment of the damage on tooth adjacent to a Class II cavity preparation was based on CNN frameworks.

Object detection model

The object detection model was used to automatically detect the region of interest (ROI) and generate bounding boxes to automate the cropping of the ROI, which is the area that the damage most likely will occur during the preparation of a Class II cavity. This process was implemented using CNN-based YOLOv5 (You Only Look Once version 5). YOLOv5 is an advanced object detection algorithm that balances computational efficiency with detection accuracy (Khanam and Hussain, 2024).

Batch 1 images consisted of 250 images, each annotated by a senior academic (KR) using Rectlabel software (Version 2024.11.16, Ryo Kawamura) with rectangular bounding boxes indicating the ROI which was used for the YOLOv5 model training. The YOLOv5 was initially pretrained on the COCO dataset [14] and fine-tuned using our images, which were resized to 640×640 pixels and normalized based on the mean and standard deviation of pixel intensity values across the dataset.

YOLOv5 was implemented using the PyTorch framework with Adam optimizer, learning rate of 1e-02, and trained for 100 epochs. A five-fold cross-validation was employed for training where the dataset was divided into five equally sized subsets (folds), while preserving the class, with each fold containing 20% of the data. Four parts (80% of the data) were used for training, and the remaining part (20% of the data) was used for validation. The model is trained on four folds and validated on the remaining fold, and this process is repeated five times, with each fold serving as the validation set exactly once. Cross-validation is a standard machine learning evaluation technique for model performance while minimising bias and variance [15]. The object detection model performance was evaluated using the mean Average Precision (mAP) at an Intersection-over-Union (IoU) threshold of 0.5. All training and inference processes were conducted using an NVIDIA RTX A6000 GPU.

The pretrained YOLOv5 model was used to obtain bounding boxes for Batch 2, (136 images) and the test dataset (16 images). After all bounding boxes were created and confirmed by the senior academic (KR), the ROI was cropped using these bounding boxes to facilitate the classification model's training in the next step. Finally, each image was exported at a resolution of 512 × 512 pixels.

Classification model

A classification model was used to assess the damage on tooth adjacent to a Class II tooth cavity preparation in intraoral cameras images. Among different CNNs selected and tested for this dataset, the Dense Convolutional Network (DenseNet-169), with a depth of 169 layers, demonstrated the best performance and was chosen for subsequent training and testing.

The classification model, DenseNet-169, was trained using a total of 386 cropped images using both Batch 1 and Batch 2. These cropped images were annotated as described above (Dataset Preparation/Annotation) with one of three damage classification labels: no damage, damage not requiring restoration, and damage requiring restoration.

The training datasets, Batch 1 and Batch 2, contained a diverse representation of the three damage categories. A total of 386 images were used, comprising 134 images (34.7%) labelled as ‘no damage,’ 124 images (32.1%) labelled as ‘damage not requiring restoration,’ and 128 images (33.2%) labelled as ‘damage requiring restoration. For testing, 16 randomly selected images, approximately preserving the class distribution, were used to evaluate the model's performance in classifying tooth damage into the three categories. The workflow of the study is presented in Figure 2.

In order to feed the DenseNet-169**,** all images were first resized to 512x512 pixels and normalized according to the mean and standard deviation of the entire dataset. The DenseNet model was trained using the Adam optimizer with a learning rate of 1e-05 with 200 epochs. The experiment was carried out using a NVIDIA RTX A6000 GPU.

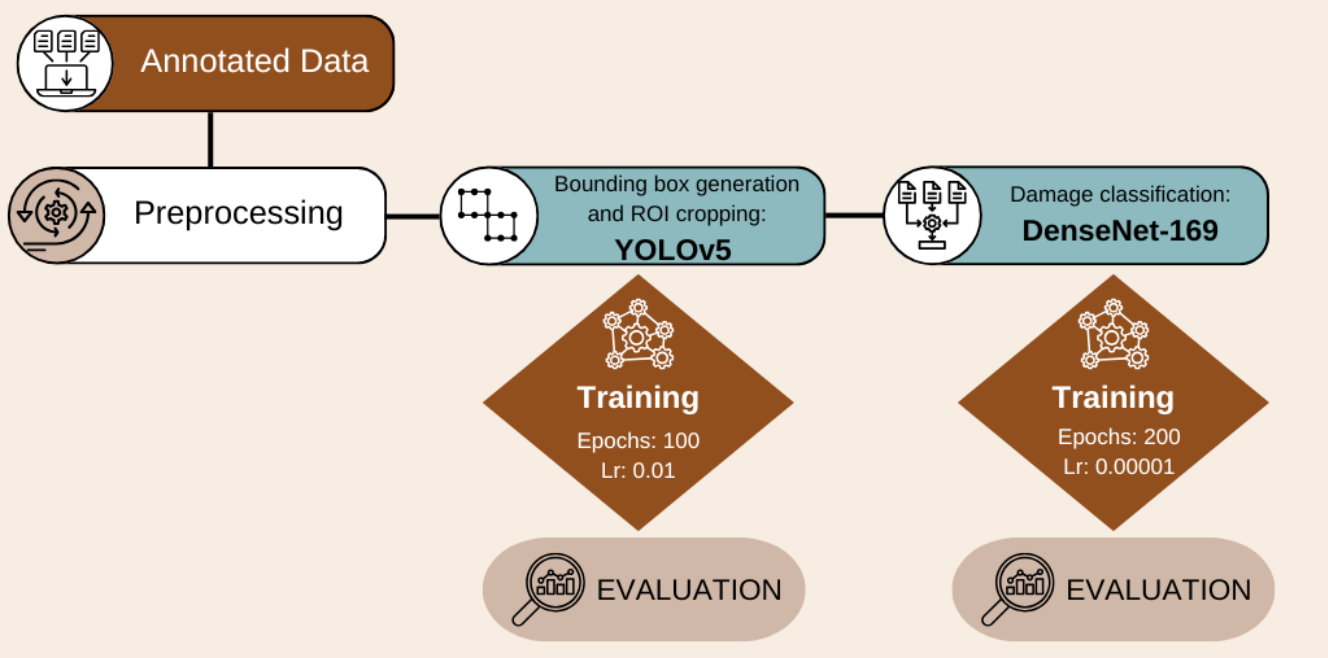

Figure 2. Flowchart for automated tooth damage detection and classification.

Assessment Methods Evaluation Metrics

Several performance metrics were calculated for dental model-based, image-based and AI model-based (DenseNet-169 model) assessments. These metrics included accuracy (the proportion of correctly predicted instances), recall (the proportion of actual positives correctly identified, also known as recall), precision (the proportion of correct positive predictions among all positive predictions), the F1 score (the harmonic mean of precision and sensitivity, providing a balance between the two) and area under the curve (AUC). In addition, a confusion matrix was generated.

Results

Fleiss' kappa demonstrated moderate agreement among clinical educators' assessments across both methods—models and images—at Time 1 (models: κ = 0.51, 95% CI [0.48, 0.53*], p* < 0.001; images: κ = 0.449, 95% CI [0.421, 0.478], *p* < 0.001). The inter-examiner reliability assessing the damage on models was higher (Time 1: k=0.505; Time 2: k=0.422) than looking at images (Time 1: k=0.449; Time 2: k=0.383) when comparing the same assessment times (Times 1 and 2) (Table 1).

Table 1. Summary of inter-examiner reliability for dental model- and image-based assessment methods at Time 1 and 2.

Type	Fleiss Value	Classification	95% Confidence Interval	*p value*
Inter-examiner reliability time 1 (models)	0.51	Moderate	0.48-0.53	< 0.001
Inter-examiner reliability time 2 (models)	0.42	Moderate	0.39-0.45	< 0.001
Inter-examiner reliability time 1 (images)	0.45	Moderate	0.42–0.48	< 0.001
Inter-examiner reliability time 2 (images)	0.38	Fair	0.36-0.41	< 0.001

Table 2 shows the inter-examiner agreement for damage assessment differed across three categories and assessment methods (dental models and images) at different time points. A decreased reliability, ranging from poor to fair, was observed in the category 'Damage not requiring restoration', with inter-examiner agreement values for models recorded at 0.393 (95% CI [0.353, 0.432]) at time 1 and 0.280 (95% CI [0.241, 0.320]) at Time 2. For images, the agreement values were 0.287 (95% CI [0.247, 0.326]) at Time 1 and 0.193 (95% CI [0.153, 0.232]) at time 2. All *p*-values were statistically significant (*p* < 0.00001).

Table 2. Summary of inter-examiner reliability (%) for each damage category at Time 1 and Time 2 using both assessment methods (dental model- and image-based).

Category	Inter-examiner Agreement (%) Time 1 (Models)	Inter-examiner Agreement (%) Time 2 (Models)	Inter-examiner Agreement (%) Time 1 (Images)	Inter-examiner Agreement (%) Time 2 (Images)
No damage	0.595 (0.555-0.634)	0.532 (0.493-0.572)	0.407 (0.368-0.447)	0.324 (0.284-0.364)
Damage not requiring restoration	0.393 (0.353-0.432)	0.280 (0.241-0.320)	0.287 (0.247-0.326)	0.193 (0.153-0.232)
Damage requiring restoration	0.545 (0.505-0.584)	0.476 (0.437-0.516)	0.652 (0.613-0.692)	0.622 (0.582-0.661)
p-value	0.00000	0.00000	0.00000	0.00000

Values in parentheses represent 95% confidence intervals. All reported p-values (p = 0.00000) indicate statistically significant agreement across categories and time points in the same assessment method.

The intra-examiner reliability analysis comparing assessment methods (dental model- and image-based) at the same time point (Time 1 and Time 2) revealed varied levels of agreement among CEs (Table 3). Approximately 22% of CEs (n = 4; CE 4, 8, 11, and 15) demonstrated consistently good agreement across both time points using different methods, with substantial kappa values indicating a consistent performance. Similarly, 22% of CE (n = 4; CE 2, 6, 10, and 18) exhibited low agreement (poor to fair) when comparing models and images at both Time 1 and Time 2. Notably, two CEs (2 and 6) showed consistently low agreement across both time points and methods reflecting variability in performance.

Table 3. Summary of intra-examiner reliability between dental model- and image-based assessment methods at Time 1 and Time 2.

	Time 1		Time 2	
	Kappa Value	**P value**	**Kappa Value**	**P value**
CE 1	0.392	0.032	0.810	<0.001
CE 2	0.205	0.202	0.164	0.404
CE 3	0.590	0.002	0.429	0.12
CE 4	0.621	<.001	0.614	<0.001
CE 5	0.484	0.008	0.282	0.106
CE 6	0.329	0.77	0.284	0.150
CE 7	0.243	0.139	0.706	<0.001
CE 8	0.686	<.001	0.686	<0.001
CE 9	0.527	0.003	0.529	0.002
CE 10	0.367	0.018	0.048	0.776
CE 11	0.610	<.001	0.800	<0.001
CE 12	0.621	<.001	0.535	0.002
CE 13	0.238	0.154	0.487	0.009
CE 14	0.429	0.007	0.392	0.024
CE 15	0.805	<.001	0.682	<0.001
CE 16	0.422	0.014	0.444	0.027
CE 17	0.676	<.001	0.385	0.039
CE 18	0.291	0.118	0.243	0.154

The analysis of intra-examiner reliability for dental model-based assessment between Time 1 and Time 2; and for image-based assessment between Time 1 and Time 2 revealed higher kappa values in 5 assessors (28%) (Table 4). Specifically, these CEs achieved kappa values exceeding 0.8 for models and above 0.7 for images. Additionally, assessments on dental models generally exhibited higher kappa values compared to images.

Table 4. Summary of Intra-Examiner reliability for dental model-based assessment between Time 1 and Time 2; and for image-based assessment between Time 1 and Time 2.

	Models		**Images**	
	Kappa Value	**P value**	**Kappa Value**	**P value**
CE 1	.807	<.001	.400	.026
CE 2	.200	.304	.487	.003
CE 3	.590	.002	.619	<.001
CE 4	.540	.001	.442	.007
CE 5	.800	<.001	.568	.002
CE 6	.452	.011	.452	.002
CE 7	.718	<.001	.590	.002
CE 8	.797	<.001	1.000	<.001
CE 9	.813	<.001	.713	<.001
CE 10	.518	.003	.587	.001
CE 11	.895	<.001	.617	<.001
CE 12	.813	<.001	.721	<.001
CE 13	.385	.041	.700	<.001
CE 14	1.000	<.001	.624	<.001
CE 15	.590	.001	.706	<.001

	Models		Images	
	Kappa Value	**P value**	**Kappa Value**	**P value**
CE 16	.893	<.001	.500	.004
CE 17	.800	<.001	.662	<.001
CE 18	.662	<.001	.512	.004

Kappa values indicate the level of agreement between Time 1 and Time 2 assessments for models and images. *P*-values indicate the statistical significance of the agreement. Values closer to 1 represent stronger agreement, while values closer to 0 represent weaker agreement.

Correlation between dental model- and image-based assessment methods is not statistically significant (r=.243, p > .05) (Figure 3).

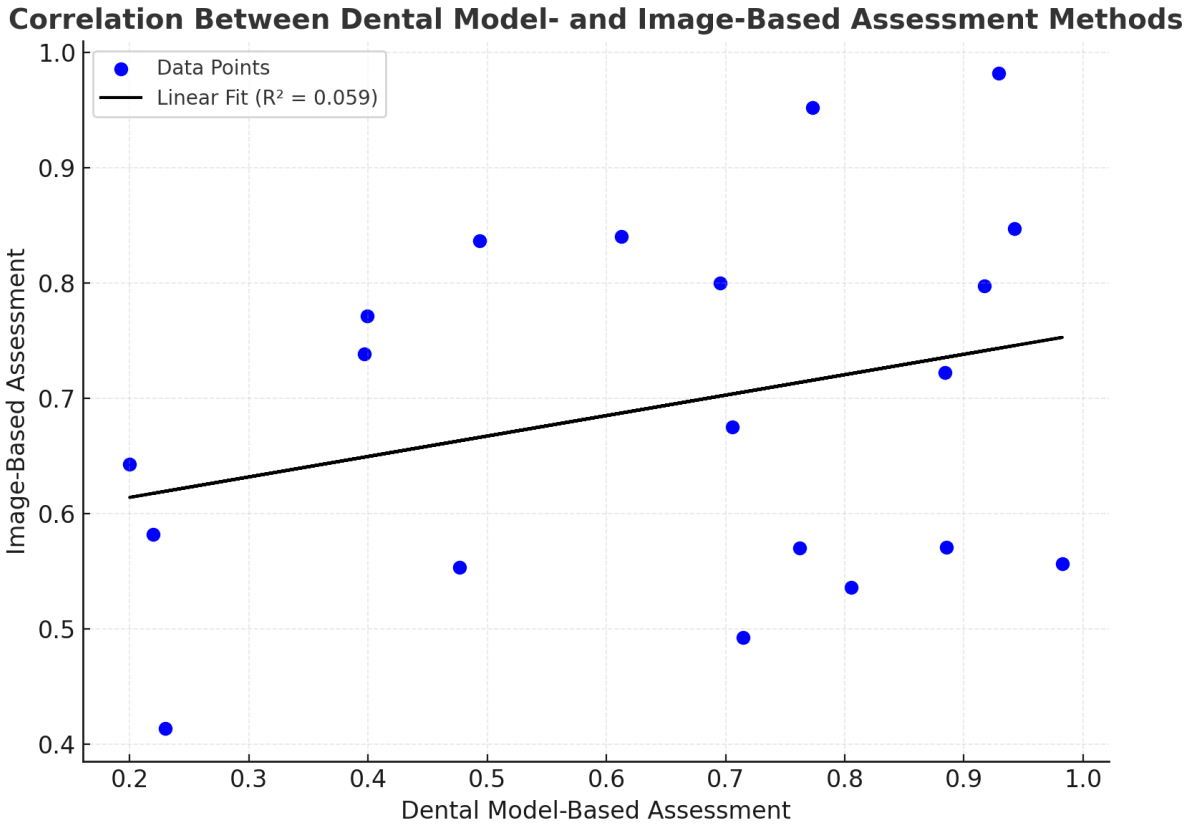

Figure 3. Scatter plot showing correlation between dental model- and image-based assessment methods using kappa values.

Figure 4 shows the bounding boxes created by YOLOv5. The object detection model had the ability to accurately detect the adjacent tooth to a tooth cavity preparation including the proximal surface as the ROI. It should be noted that the images were taken at different angles and distances to reflect the common practice. Even though this resulted in different views of tooth cavities and adjacent teeth, YOLOv5 was able to handle the complexities of the intraoral images and tooth morphology. The ROI was not confined only to the proximal surface and, instead, included the entire adjacent tooth. Nevertheless, the classification model, DenseNet-169, accurately classified the damage that was located only on the proximal area.

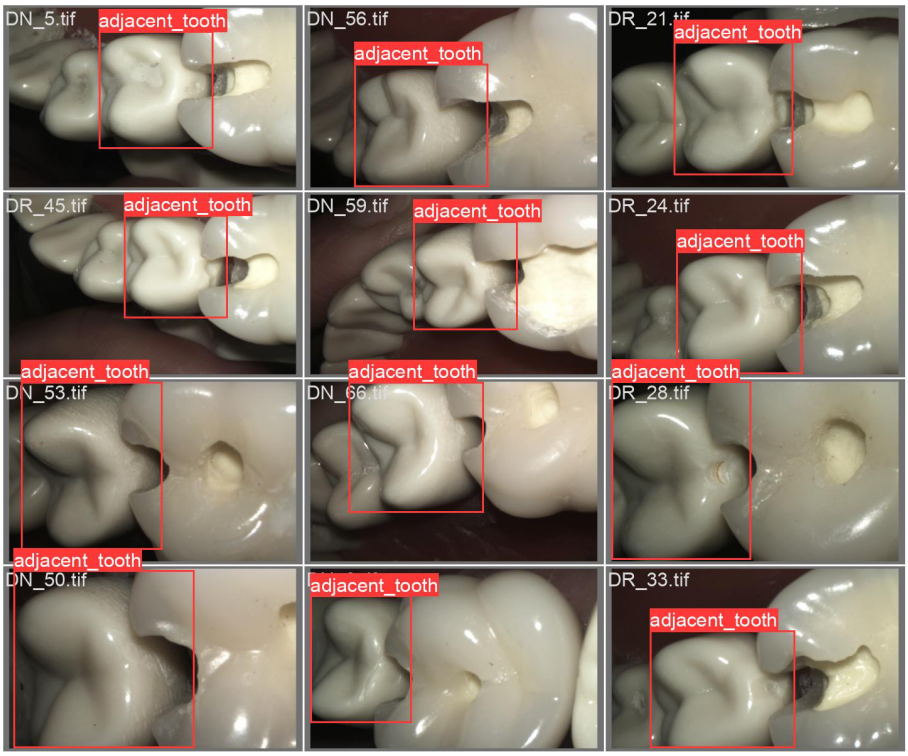

Figure 4. Bounding boxes on the distal surfaces on the tooth adjacent to a Class II cavity preparation created by the YOLOv5 model.

The performance metrics for the assessment methods, dental-model based, image-based and the AI classification model DenseNet-169 are summarised in Table 5. Table 6 highlights the overall evaluation metrics for each class of tooth damage for the DenseNet-169.

Table 5. Assessment methods evaluation metrics.

Assessment Method	Accuracy	Precision	Recall	F1	AUC
AI model-based	0.81	0.83	0.81	0.82	0.89
Image-based Time 1	0.65	0.66	0.65	0.63	0.74
Image-based Time 2	0.67	0.68	0.67	0.65	0.76
Dental model-based Time 1	0.75	0.75	0.75	0.73	0.81
Dental model-based Time 2	0.69	0.72	0.69	0.67	0.77

Table 6. Category-wise evaluation metrics for the DenseNet-169.

Category	Accuracy	Precision	Recall	F1	AUC
No damage	0.75	0.60	0.75	0.67	0.79
Damage not requiring restoration	0.71	0.83	0.71	0.77	0.80
Damage requiring restoration	1.00	1.00	1.00	1.00	1.00

The confusion matrix shown in Figure 5 illustrates the performance of the DenseNet-169 classifier in identifying damage across the three groups. All correct predictions were observed for “damage requiring restoration”. For “no damage” 1 out of 4 predictions was incorrect with the model assigning the one to “damage not requiring restoration” category. The highest number of incorrect predictions (2 out of 7) were found for “damage not requiring restoration” category where the model incorrectly assigned these cases to “no damage”.

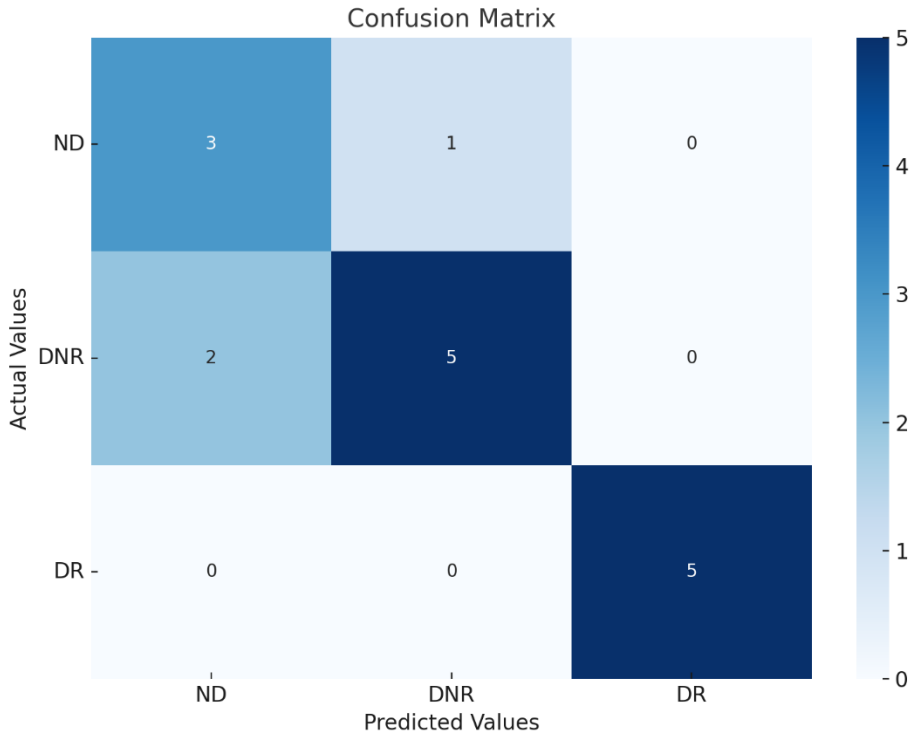

Figure 5. Confusion matrix representing the predicted versus actual classification outcomes for the DenseNet-169 classifier applied to intraoral images. ND: no damage, DNR: damage not requiring restoration, DR: damage requiring restoration.

Discussion

The working hypothesis was confirmed - all performance metrics were higher for the AI-model based assessment method compared to both CE assessment methods, dental model- and image- based. Regarding null hypotheses, significant differences in inter- and intra-examiner reliability were found when assessing the damage on tooth adjacent to a Class II cavity preparation, thus H0 #1 was rejected. Similarly, significant differences were observed between time points (Time 1 and 2) and assessment methods (images and dental models), leading to the rejection of H0 #2 and H0 #3. The null hypothesis related to the correlation between dental model- and image-based assessment methods (H0 #4), was upheld, as no correlation was found.

Iatrogenic damage on the proximal surfaces of adjacent teeth was reported to be between 5% and 50% with and without the use of proximal protection in Class II cavity preparation, respectively [2]. Although the frequency of damage may seem low with proximal protection, it leads to unnecessary permanent damage of the adjacent tooth, potentially requiring further restoration. In dental assessment, damage to the adjacent tooth is often used as a critical barrier determining pass or fail result [16]. This is why, in the present study, damage to the adjacent tooth was chosen as the criterion for evaluating the agreement between CEs' assessment and AI model's training and testing.

Intra- and inter-examiner agreements were higher at Time 1 than at Time 2. This finding can be attributed to CE fatigue and overconfidence in doing the same task again, leading to lack of focus and decreased attentiveness. This finding is highly important and relevant for repeated and extended periods of assessment, such as assessment of dental models of an entire student cohort in preclinical exams. It also reflects a need for more reliable and consistent assessment tools to ensure not only relevant feedback but also fairness in formal examination.

The inter-examiner reliability in assessing damage using dental model-based assessment was higher (Time 1: k=0.505; Time 2: k=0.422) compared to image-based assessment (Time 1: k=0.449; Time 2: k=0.383) at the same assessment times (Times 1 and 2). As shown in Table 2, the 'no damage' classification was more reliable using the dental model-based assessment method, likely because CEs could use a dental

explorer to confirm the absence of damage, particularly in cases where a dirty surface on the image might resembled damage. Conversely, the ‘damage requiring restoration’ category had higher agreement with the image-based method, as this more extensive damage is typically clearer and does not depend on tactile sensation. These results align with previous research demonstrating that tactile sensation, as used in the dental model-based assessment method, enhances the predictive validity of caries detection. The tactile sensation provided by a dental explorer enables better differentiation of rough lesion surfaces from sound areas [17,18].

Image-based assessment was included in this research because images captured with an intraoral camera may be the only available assessment option in the context of teledentistry and tele-education [19]. Literature shows that assessment based in intraoral digital photographs is an affordable adjunct tool for caries detection [20,21]. The intraoral camera used in this study, SiroCam UAF Plus (Dentsply Sirona, n.d.), is a high-quality device that is both relevant to clinical applications and widely accepted in dental practice.

Intra-examiner agreement results, comparing the same assessment method (dental model- or image-based) across two time points, showed inconsistent agreement among CEs. While some CEs demonstrated high levels of intra-examiner agreement, others exhibited relatively low intra-examiner agreement, which could be influenced by their experience and familiarity with the assessment criteria. Notably, higher Kappa values reflect consistency rather than accuracy. These findings highlight the need for calibration and the development of additional tools to improve calibration training and enhance the reliability and consistency of the assessment method.

Intra-examiner agreement at the same time point (Time 1 or Time 2), comparing dental model- and image-based assessment methods, showed that 22% of CEs (n=4) consistently demonstrated good agreement, while an equal proportion (22%) exhibited low agreement (poor to fair). These findings highlight that assessment using different methods tends to result in lower agreement levels. However, the results also confirm that the images possess sufficient quality for assessing damaged teeth, as only approximately 22% of CEs reported low agreement. Additionally, all Fleiss' Kappa coefficients recorded p-values below 0.05, indicating a statistically significant level of

inter-examiner agreement and confirming that the alignment among CE was not due to chance. These findings indicate that intraoral images captured with the SiroCam UAF Plus possess the clarity and quality needed to support their use in dental task assessments. Furthermore, they demonstrate potential for integration into advanced AI-supported systems, particularly in teledentistry and tele-education.

The combination of technology with traditional methods appears essential for achieving the best outcomes. Direct feedback from faculty is valued by students for its clarity and guidance, whereas technology allows for independent practice and self-assessment [22]. A recent literature review emphasised the challenges of ensuring objectivity, validity, and fairness in technical skill assessments, suggesting that combining multiple methods may yield optimal results [23].

Integrating intraoral imaging and AI technologies to bridge the gaps in dental education and enhance dental task assessment and feedback is a promising option, based on the present results. Further study is needed to implement this AI-assisted analysis of intraoral images to other assessment tasks and criteria as well as clinical decision-making. Previous research on the use of intraoral images in automatic caries detection using deep learning technology, such as, CNN frameworks, showed excellent results [11,24].

The YOLOv5 architecture, in object detection mode, enables simultaneous detection of multiple objects in a single forward pass, making it highly suitable for medical imaging tasks like identifying adjacent teeth in intraoral radiograph [25]. Promising results were recently demonstrated using the YOLOv5 model for detecting white spot lesions in post-orthodontic intraoral photographs [26]. The present study demonstrated and validated YOLOv5's ability to accurately detect the correct tooth to be assessed and identify relevant damage area.

For the damaged tooth classification task, DenseNet has shown better results compared to other deep learning models due to its innovative dense connectivity architecture. This design enables connections between all layers, allowing shallow layers to directly transfer information to deeper ones. There are various DenseNet models, such as DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264,

each differing in the number of layers, offering flexibility in depth and computational requirements to suit diverse tasks. DenseNet-201 outperformed other models in recognising prosthodontic scenarios using intraoral images of the maxilla [27]. DenseNet-169 has been effectively applied to various medical imaging tasks, including brain tumour classification through magnetic resonance images [28]; knee osteoarthritis detection [29], and pneumonia image classification using chest X-rays [30].

Among the classification models evaluated in the present study, the DenseNet-169 architecture demonstrated the highest accuracy on the image-based dataset for classifying tooth damage. Overall, the model demonstrates reliable performance, with an accuracy of 0.81 and an AUC of 0.8948 showcasing solid classification capabilities. The precision, recall, and F1 score values suggest that the model is well-optimized for both identifying damage (high recall) and avoiding unnecessary misclassifications (high precision). These results highlight the potential clinical and educational utility of the model in aiding the detection and classification of adjacent tooth damage using intraoral camera images.

The inter-examiner agreement results revealed a lower level of agreement for the ‘Damage Not Requiring Restoration’ category compared to the other classifications. This reduced agreement is likely due to the intermediate nature of this category, which sits between ‘No Damage’ and ‘Damage Requiring Restoration’. Borderline cases can create ambiguity during classification, contributing to overlaps between categories. Additionally, the use of plastic teeth, which do not differentiate between enamel and dentine, may introduce surface irregularities surface irregularities introduced during the manufacturing process, that mimic scratches, potentially leading to misclassification.

Similarly, the AI model confusion matrix analysis reveals the model's strengths and areas for improvement in classifying tooth damage using intraoral camera images. The AI model accurately identified cases of ‘Damage Requiring Restoration’ with no misclassifications, demonstrating its robustness in detecting clinically significant damage, which is crucial for preventing the need for further dental treatment. However, some overlap was observed between the ‘No Damage’ and ‘Damage Not Requiring Restoration’ categories. The confusion matrix indicated that ‘Damage Not Requiring

Restoration' was the most challenging category for the AI model as well as for CEs. These misclassifications suggest the need for further refinement to enhance the AI model's ability to distinguish subtle differences between these categories. Despite these limitations, the AI model's overall performance underscores its potential utility as a diagnostic aid in clinical and educational settings, particularly in reliably identifying cases requiring restorative intervention.

While dental models provide higher reliability for dental damage assessments, image-based assessment methods remain a viable alternative, particularly in situations where images are the only assessment option, such as in teledentistry and tele-education. As an additional tool for CEs, AI-based analysis of intraoral images can reduce the assessment burden and limit the number of criteria assessed by CEs.

Future studies should focus on incorporating the complete Class II cavity tooth preparation criteria and expanding to other tooth cavity preparation classes (e.g., Class III and Class IV) with various image angles. Additionally, integrating automated segmentation and patient-specific intraoral images could further enhance assessment accuracy and reliability, ultimately supporting both clinical practice and dental education.

Conclusions

The inter-examiner agreement with the dental model-based assessment method was higher than that with the image-based assessment method at different time points. Intra-examiner agreement showed inconsistencies between assessment methods and time points. The AI model demonstrated reliable performance, with an accuracy of 0.81 and an AUC of 0.8948, highlighting its strong classification capabilities, balanced by high recall for identifying damage and high precision for minimising misclassifications. The ‘Damage Not Requiring Restoration’ was most challenging both for CEs and the AI model.

References

1. Medeiros VA, Seddon RP. Iatrogenic damage to approximal surfaces in contact with Class II restorations. *J Dent*. Feb 2000;28(2):103-10.
2. Al-Bukhary RA, Mannaa AI. Assessment of Proximal Protection Usage by Dental Students During Class II Cavity Preparations: An In Vivo Pilot Study. *Cureus*. Oct 2023;15(10):e47550.
3. Khalaf K, El-Kishawi M, Mustafa S, Al Kawas S. Effectiveness of technology-enhanced teaching and assessment methods of undergraduate preclinical dental skills: a systematic review of randomized controlled clinical trials. *BMC Medical Education*. 2020/08/28 2020;20(1):286.
4. Sadid-Zadeh R, D'Angelo EH, Gambacorta J. Comparing feedback from faculty interactions and virtual assessment software in the development of psychomotor skills in preclinical fixed prosthodontics. *Clin Exp Dent Res*. Oct 2018;4(5):189-195.
5. Sa B, Ezenwaka C, Singh K, Vuma S, Majumder MAA. Tutor assessment of PBL process: does tutor variability affect objectivity and reliability? *BMC Medical Education*. 2019/03/08 2019;19(1):76.
6. Gittinger FP, Lemos M, Neumann JL, et al. Interrater reliability in the assessment of physiotherapy students. *BMC Medical Education*. 2022/03/16 2022;22(1):186.
7. Nagy ZA, Simon B, Tóth Z, Vág J. Evaluating the efficiency of the Dental Teacher system as a digital preclinical teaching tool. *Eur J Dent Educ*. Aug 2018;22(3):e619-e623.
8. Tiu J, Cheng E, Hung TC, et al. Effectiveness of Crown Preparation Assessment Software As an Educational Tool in Simulation Clinic: A Pilot Study. *J Dent Educ*. Aug 2016;80(8):1004-11.
9. Jogin M, Mohana, Madhulika MS, Divya GD, Meghana RK, Apoorva S. Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. 2018:2319-2323.
10. Jiang H, Zhang P, Che C, Jin B, Zhu Y. CariesFG: A fine-grained RGB image classification framework with attention mechanism for dental caries. *Engineering Applications of Artificial Intelligence*. 2023/08/01/ 2023;123:106306.
11. You W, Hao A, Li S, Wang Y, Xia B. Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments. *BMC Oral Health*. May 13 2020;20(1):141.
12. Camalan S, Mahmood H, Binol H, et al. Convolutional Neural Network-Based Clinical Predictors of Oral Dysplasia: Class Activation Map Analysis of Deep Learning Results. *Cancers (Basel)*. Mar 14 2021;13(6)
13. Altman DG. *Practical statistics for medical research*. Chapman and Hall/CRC; 1990.
14. Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. Springer; 2014:740-755.
15. Stone M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*. 1974;36(2):111-133.
16. Council AD. *Practical examination handbook for dentists*. Vol. 2024. 2018. Accessed 20 November 2024. https://adc.org.au/files/assessment/handbooks/ADC_Practical_Exam_Handbook_General.pdf

17. Nyvad B, Machiulskiene V, Baelum V. Construct and Predictive Validity of Clinical Caries Diagnostic Criteria Assessing Lesion Activity. *Journal of Dental Research*. 2003/02/01 2003;82(2):117-122.
18. Wierichs RJ, Werren TT, Jaruszewski L, Meyer-Lueckel H. Tactile sensation in relation to roughness and reflection of active initial lesions in primary (deciduous) and permanent dentition in vitro. *Journal of Dentistry*. 2024/11/01/ 2024;150:105374.
19. Nascimento da Silva Mulder J, Ramos Pinto M, Aníbal I, et al. Teledentistry Applied to Health and Education Outcomes: Evidence Gap Map. *J Med Internet Res*. 2024/11/27 2024;26:e60590.
20. Boye U, Walsh T, Pretty IA, Tickle M. Comparison of photographic and visual assessment of occlusal caries with histology as the reference standard. *BMC Oral Health*. Apr 27 2012;12:10.
21. Bottenberg P, Jacquet W, Behrens C, Stachniss V, Jablonski-Momeni A. Comparison of occlusal caries detection using the ICDAS criteria on extracted teeth or their photographs. *BMC Oral Health*. 2016/09/07 2016;16(1):93.
22. Stoilov M, Trebess L, Klemmer M, Stark H, Enkling N, Kraus D. Comparison of Digital Self-Assessment Systems and Faculty Feedback for Tooth Preparation in a Preclinical Simulation. *International Journal of Environmental Research and Public Health*. 2021;18(24):13218.
23. Uoshima K, Akiba N, Nagasawa M. Technical skill training and assessment in dental education. *Japanese Dental Science Review*. 2021/11/01/ 2021;57:160-163.
24. Kang S, Shon B, Park E, Jeong S, Kim E-K. Diagnostic accuracy of dental caries detection using ensemble techniques in deep learning with intraoral camera images. *PLOS ONE*. 09/06 2024;19
25. Khanam R, Hussain M. What is YOLOv5: A deep look into the internal features of the popular object detector. *arXiv preprint arXiv:240720892*. 2024;
26. Ozsunkar PS, Özen DÇ, Abdelkarim AZ, et al. Detecting white spot lesions on post-orthodontic oral photographs using deep learning based on the YOLOv5x algorithm: a pilot study. *BMC Oral Health*. 2024/04/24 2024;24(1):490.
27. Islam NM, Laughter L, Sadid-Zadeh R, et al. Adopting artificial intelligence in dental education: A model for academic leadership and innovation. *J Dent Educ*. Nov 2022;86(11):1545-1551.
28. Prakash RM, Kumari RSS, Valarmathi K, Ramalakshmi K. Classification of brain tumours from MR images with an enhanced deep learning approach using densely connected convolutional network. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. 2023;11(2):266-277.
29. Al-rimy BAS, Saeed F, Al-Sarem M, Albarrak AM, Qasem SN. An Adaptive Early Stopping Technique for DenseNet169-Based Knee Osteoarthritis Detection Model. *Diagnostics*. 2023;13(11):1903.
30. Bundea M, Danciu GM. Pneumonia Image Classification Using DenseNet Architecture. *Information*. 2024;15(10):611.

Chapter 3: *[Manuscript] Automated Detection of Positioning Errors in Bitewing Radiographs Using Deep Learning*

Automated Detection of Positioning Errors in Bitewing Radiographs Using Deep Learning

Original investigation

Katharina Alves Rabelo, **BDS, MSc**[1], Zimo Huang, **B.S. (Hons.)**[2], Eduardo Delamare **BDS, MSc**[1], Shwetha Hegde, **BDS, MDS, FHEA**[1], Jinman Kim, **B.S. (Hons.), PhD**[2], Vesna Miletic, **BDS, MSc, PhD**[3]

[1] Dentomaxillofacial Radiology, Sydney Dental School, Faculty of Medicine and Health, The University of Sydney, Camperdown, NSW, 2050, Australia

[2] School of Computer Science, Faculty of Engineering, The University of Sydney, Camperdown, NSW, 2050, Australiaa

[3] Restorative and Reconstructive Dentistry, Sydney Dental School, Faculty of Medicine and Health, The University of Sydney, Camperdown, NSW, 2050, Australia

Acknowledgements

We acknowledge Dr Amelita Simpson, a lecturer at Sydney Dental School, for her help in organising the data. Additionally, we extend our sincere thanks to the dental assistants at the Bligh Building Simulation Clinic for their assistance and dedication in facilitating data collection.

***Corresponding author:** Katharina Alves Rabelo
Sydney Dental School, Faculty of Medicine and Health, The University of Sydney.
Address: Susan Walki Health Building, Level 6, Western Ave, Camperdown, NSW, 2050,
Australia.
E-mail: katharina.alvesrabelo@sydney.edu.au

Abstract

Rationale and Objectives: Bitewing radiographs (BWs) are essential diagnostic tools in dentistry but are often affected by positioning errors, compromising their diagnostic value and increasing retake rates. This study aimed to develop and evaluate a deep learning-based automated system for detecting and classifying BW type (premolar/molar), side (right/left) and common positioning errors.

Materials and Methods: A total of 403 BWs were collected from dental student assessments acquired from dental radiology mannequins. BWs were consensus-labelled by experts for type (premolar/molar), side (right/left), and positioning errors (cone cutting error (CCE), interproximal overlap, and incorrect receptor placement). A balanced dataset supported training and validation. Convolutional neural network (CNN) architectures, including ResNet and DenseNet variants, were employed. Five-fold cross-validation was used to assess performance based on accuracy, precision, recall, F1 score, and area under the curve. A confusion matrix was generated for multi-class CCE classification.

Results: The system achieved high accuracy in classifying BW type (95.8%), side (99.0%), CCE (96.3%), and interproximal overlap (93.4%). Multi-class CCE classification exhibited moderate performance (accuracy: 79.3%), with the model reliably identifying ‘critical’ and ‘minimal’ errors. Accuracy for detecting incorrect receptor placement was 73.2%.

Conclusion: The CNN-based system can effectively detect and classify BW type, side, and most positioning errors in mannequin-acquired radiographs. For BW type, side and the positioning errors: CCE (presence and absence) and interproximal overlap (presence and absence), achieving accuracies in the range of 93%-99%. The classifier demonstrated a moderate level of performance for incorrect receptor placement error, with an accuracy of 73.2%.

Keywords: Deep learning; Convolutional neural network; Dental digital radiography; Bitewing; Imaging errors

Introduction

The bitewing radiograph (BW) is a common radiographic technique in dentistry due to its ability to show crowns of opposing posterior teeth and the alveolar crest in one image. For optimal BW acquisition, the receptor can be oriented horizontally or vertically and must be positioned parallel to the buccal and lingual teeth surfaces. Additionally, the X-ray beam must be directed through the interproximal spaces and perpendicular to the receptor (1). The geometry of the horizontal and vertical angulations, combined with appropriate receptor placement, provides a clear view of the interproximal surfaces without overlapping structures (1, 2). Despite the availability of positioning devices for the BW technique, the receptor position and the accurate beam alignment with the receptor and anatomical landmarks largely rely on the operator's technical skills. Therefore, bitewing radiographs are susceptible to a range of positioning errors, also referred to as technical errors, that may compromise diagnostic interpretation. These errors can increase the frequency of retakes, leading to additional radiation exposure, higher costs, and delays in dental procedures.

Despite advancements in Dentomaxillofacial Radiology (DMFR), including the development of digital receptors, enhanced X-ray units, and positioning devices, positioning errors continue to be reported in the literature as one of the most prevalent causes of unacceptable intraoral radiograph quality. A recent literature review reveals that retake rates for intraoral, extraoral, and cone-beam computed tomography (CBCT) imaging range from 5% to 20% due to imaging errors. In particular, BWs had an average rejection rate of 11.25%, with positioning errors and patient movement frequently identified as the main causes (3). In another study, improper angulation (26.1%) and incorrect receptor placement (11.2%) have been identified as the most prevalent positioning errors attributed to intraoral radiograph retakes (4). Furthermore, in periapical radiographs, overlapping of proximal surfaces (5, 6) has been reported as one of the most frequent imaging errors, further emphasising the impact of suboptimal receptor positioning and X-ray beam angulation inconsistencies on intraoral radiographs.

A recent large number of studies have demonstrated that AI models can significantly enhance radiographic interpretation. Particularly, Deep learning (DL) models using Convolutional Neural Networks (CNNs) have shown high performance when applied

to computer vision tasks such as detection, classification and segmentation (7). When trained on BW datasets, CNNs can effectively detect primary and secondary caries in permanent teeth (8), proximal caries in mixed dentition (9), as well as pulp chamber calcifications (10), with reported accuracies of 68.9%–71.9%, 96.4%, and 86.18%, respectively. A systematic review found DL models trained on BW datasets achieved accuracies ranging from 71.11% to 94.5% for the detection of interproximal caries (11).

The accuracy variability in caries detection models could be attributed to differences in dataset selection. Some studies excluded radiographs with positioning errors, such as improper angulation or image distortion (8, 12), as well as those with cone-cuttings, horizontal overlap, receptor placement errors, motion, or artifacts that could hinder diagnosis (9, 13). Additionally, other studies explicitly stated that only high-quality images were included (14), while others did not mention any form of image quality assessment as part of the inclusion or exclusion criteria (15). The natural, biological variability in human subjects adds the complexity and may reduce accuracy of the models. Standardising anatomical structures and exposure settings, such as in BWs taken from mannequins, contributes to explaining a relationship between the accuracy of the AI model to detect a pathology and specific parameters, such as positioning errors.

A recent study has shown that a CNN (dentalXrai Pro, dentalXrai Ltd.) has improved the diagnostic accuracy of clinicians in detecting enamel carious lesions (16). While these models often achieve high-performance metrics, their results are largely based on radiographs free from imaging errors. However, in daily dental practice, positioning errors are common, posing a significant challenge to the robustness and generalizability of DL models trained on strictly curated datasets. Recent studies highlight that performance of DL models is dependent on complexity of the characteristics of dental imagery dataset used for training and validation stages (17). One study explicitly showed outputs from the CNN model Inception-ResNet-v2, in which areas of interproximal overlap, a common positioning error, were **misclassified** as caries (14). Radiographs with variable levels of positioning errors may introduce similar complexity and negatively affect models trained under ideal conditions.

Based on these findings, the introduction of a DL-based tool for detection of positioning errors has the potential to enhance AI research on BWs by improving management of dataset heterogeneity. Additionally, it may benefit clinical tasks by informing dental practitioners of the diagnostic potential of each image prior to interpretation, as well as assisting clinicians in recognising and correcting positioning errors, ultimately improving the quality of radiographic acquisition. Therefore, the aim of the study was to assess the performance of an automated system of DL models developed for classifying BW type (premolar/molar), side (right/left) and detecting common positioning errors in BWs acquired from dental radiology mannequin.

Methods

Study design

This retrospective study was reviewed and approved by the University Human Ethics Committee (2024/HE000846). The study applies CNN algorithms to automatically classify bitewing according to the type (premolar and molar), side (right and left), and to detect and categorize common positioning errors.

Dataset

BW radiographs were acquired from adult dental radiology mannequin during regular assessment periods in the discipline of DMFR at the University, in 2021 and 2022. These radiographs were obtained by students enrolled in the Bachelor of Oral Health and Doctor of Dental Medicine programs. The BWs were taken using a Heliodent Plus intraoral X-ray device (Dentsply Sirona, Bensheim, Germany) with a focal spot size of 0.4 mm and a pre-programmed BW exposure setting (tube voltage: 70 kV, tube current: 7 mA, exposure time: 0.10 s), using intraoral Photostimulable Phosphor plate (PSP) receptor (Air Techniques, Melville, New York, USA) and cardboard loop film holders (ADM, n.d.).

A total of 403 BWs from 3 different mannequins of the same X-ray Phantom (Nissin, Minami-ku, Kyoto, Japan) were collected. The selected BWs were uploaded to the Research Data Store networked drive in .jpg format and 1980x1486 resolution.

A senior investigator and lecturer in DMFR with over 8 years of experience selected the radiographs and classified the images based on type (premolar or molar), side (right or left), and the presence of specific positioning errors. Two other investigators – also lecturers in DMFR from Sydney Dental School – then assessed the selected images. A consensus method was used to classify and label the BWs according to the following positioning errors: cone cutting error (CCE); interproximal overlap and incorrect receptor placement.

All images were analysed until investigators reached a consensus on the presence of each positioning error. This was then considered the ground truth for training and validating the AI models.

Type and side of bitewing

The type of BW was determined based on the region of interest (ROI) visible in the image. In this study, it was resolved that to be classified as a premolar bitewing (PMBW), the image must show the entire mesial half of the first maxillary premolar. For a molar bitewing (MBW), at least half of the most posterior erupted third molar, or the entire distal surface of the last erupted molar must be visible. Examples of MBW and PMBW are shown in Figure 1. BW images were further classified as right or left based on the orientation of the image and the anatomical structures visible.

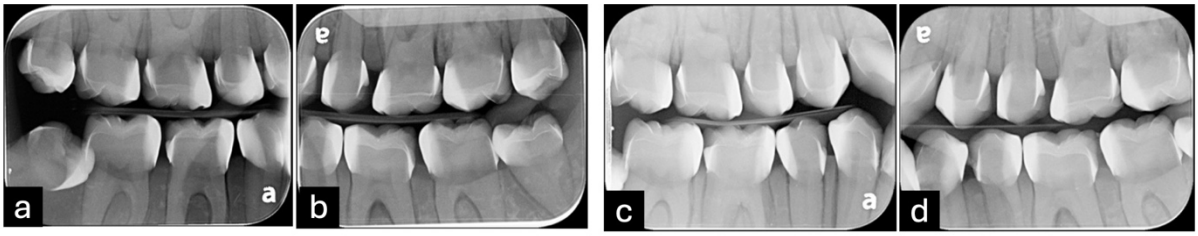

Figure 1. Images representing the types and sides of BWs. (a) Right Molar, (b) Left Molar, (c) Right Premolar, and (d) Left Premolar.

Cone cutting error

Cone cutting error (CCE) in BW radiographs occurs when part of the image appears as a well-defined, radiopaque, non-exposed area due to improper alignment of the X-ray beam with the receptor. A binary classification was performed, categorising BWs into two groups: presence and absence of CCE. Subsequently, a multi-class classification was performed on BWs with the presence of CCE, categorised into three classes based on its extent: (Minimal) CCE that does not affect the ROI; (Significant) CCE partially affecting the ROI, compromising three or fewer interproximal areas; and (Critical) CCE significantly affecting the ROI (resulting in a non-diagnostic radiograph), where more than three interproximal areas are not visible. Examples of BWs demonstrating the three classes of CCE are shown in Figure 2. All BWs classified with a critical CCE (non-diagnostic) have not undergone further assessment for other positioning errors included in this study.

Figure 2. Images representing the types of CCEs. (a) Minimal, (b) Significant, and (c) Critical.

Interproximal overlap

Interproximal overlap in BWs occurs when proximal surfaces of adjacent teeth appear superimposed. BWs were categorised based on the presence or absence of interproximal overlap. For this assessment, interproximal overlaps were defined as those extending beyond the contact point. In PMBWs, interproximal overlap between the mandibular canine and first premolar was not considered a positioning error. Similarly, in MBWs, interproximal overlap around partly erupted third molars was not considered an error as seen on Figure 3.

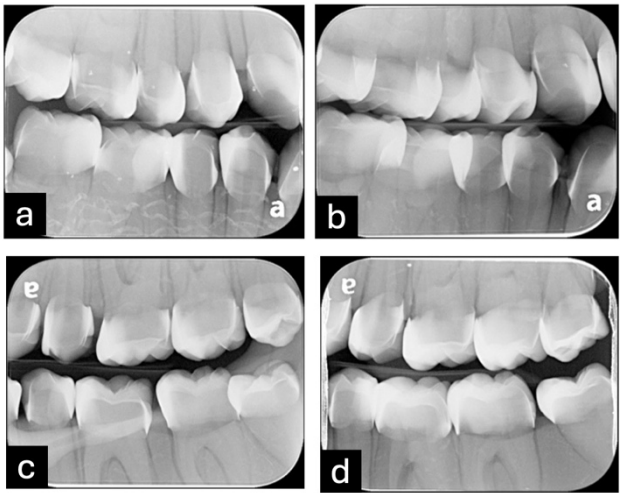

Figure 3. Images representing BWs with interproximal overlap positioning error. (a and b) Right Premolar BWs with interproximal overlaps, and (c and d) Left Molar BWs with interproximal overlaps.

Incorrect Receptor Placement

In BWs, the receptor should be placed so the upper and lower teeth are evenly captured in the image. Accordingly, the position of the occlusal plane can be used to evaluate receptor placement. A template was used for this assessment (Figure 4). The receptor placement was considered correct when the occlusal plane would extend into the green band area of the template (Fig. 5 (A and C)). If the occlusal plane extended into the yellow band area of the template, the receptor placement was considered incorrect (Fig. 5 (B and D)). If the occlusal plane extended outside the yellow band area of the template, the film placement was considered unacceptable.

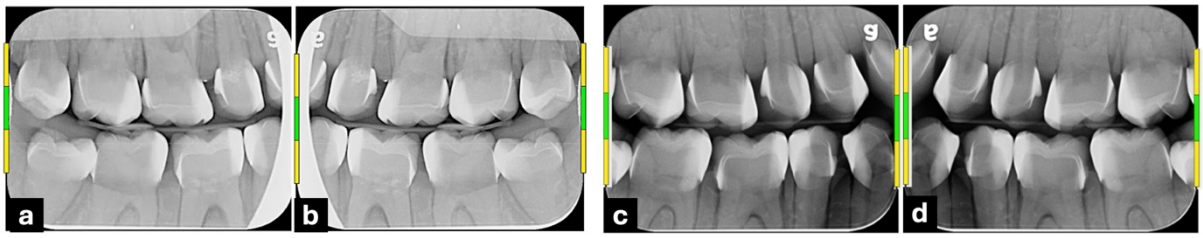

Figure 4. Templates used for assessment of incorrect receptor placement. (a) Right Molar, (b) Left Molar, (c) Right Premolar, and (d) Left Premolar.

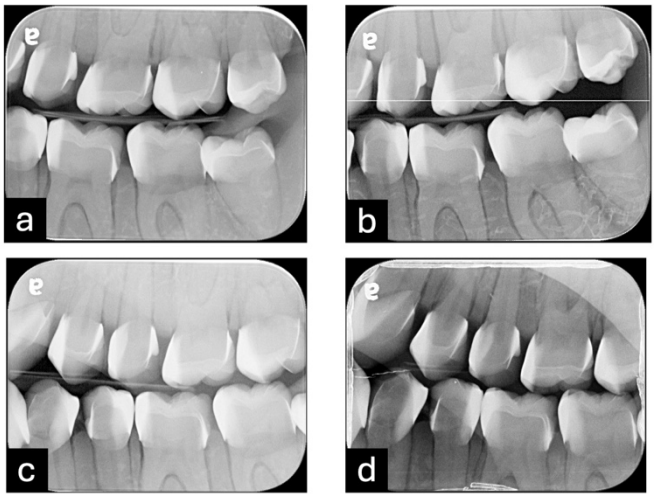

Figure 5. BWs representing the receptor placement classification. (a and c) Correct receptor placement, and (b and d) incorrect receptor placement.

Experimental Setup

For the automated detection of the positioning errors, a structured pipeline of sequential steps was implemented to enhance the AI model's efficiency. This pipeline included six tasks designed to classify BW radiographs based on five criteria: type, side, CCE, interproximal overlap, and incorrect receptor placement. The CCE was divided into two sub-tasks using both binary and multi-class classification methods.

The resultant dataset of 403 BWs was used to classify images according to type (premolar or molar) and side (right or left). The presence and absence of CCE was evaluated across the entire dataset (403 BWs). Additionally, a multi-class classification was conducted on BWs exhibiting CCE (179 BWs), with the cases grouped into three classes according to its extent (minimal, significant and critical CCE). Furthermore, based on the CCE multi-class classification, all critical CCE (non-diagnostic) (70 BWs) images were excluded from the other two positioning error assessments, interproximal overlap and incorrect receptor placement. For the incorrect receptor placement task, one additional BW was excluded as it was considered non-diagnostic due to partial crown cut off caused by improper receptor placement. The BW class distribution for each task is shown in Table 1.

Table 1. Summary of Evaluated Tasks, Class Distributions, and Best-Performing Models with Hyperparameters

Task	Class Distribution			Total	Model	Hyperparameters	
						Learning rate	Batch size
BW Type	Premolar (208)	Molar (195)		403	DenseNet169	5e-05	8
BW Side	Right (173)	Left (230)		403	DenseNet121	5e-04	8
CCE (binary)	Present (179)	Absent (224)		403	DenseNet161	1e-04	4
CCE (multi-class)	Critical (70)	Significant (61)	Minimal (48)	179	DenseNet121	5e-04	4
Interproximal overlap	Present (171)	Absent (162)		333	DenseNet161	5e-05	4
Incorrect receptor placement	Present (155)	Absent (177)		332	ResNet34	5e-04	4

All BWs were pre-processed by resizing the JPEG images to 512 × 512 pixels and normalised across all images by scaling pixel values to a consistent range, ensuring uniform input for the CNNs. The classification models were trained for 200 epochs using a NVIDIA GeForce RTX 2080 Ti. Each task was treated independently and evaluated through a 5-fold cross-validation experiment while ensuring that class distribution remained consistent across folds.

A set of CNNs, including ResNet34, ResNet50, DenseNet121, DenseNet161, and DenseNet169, was used as the baseline models for all classification tasks. Among the CNNs tested, the best-performing model for each task, along with its optimal hyperparameters (training configuration), including learning rate and batch size, was selected. The final selected models, along with their corresponding parameters for each task, are summarized in Table 1.

Performance Metrics

The CNNs' predictions were evaluated against the labelled ground truth using multiple performance metrics. These metrics included **accuracy** (the proportion of correctly predicted instances across all classes), **recall** (the proportion of actual positive cases correctly identified, also known as sensitivity), **precision** (the proportion of true positive predictions among all predicted positives), and the **F1 score** (the harmonic mean of precision and recall, providing a balanced measure of model performance). Additionally, the **area under the curve (AUC)** was calculated to evaluate the model's ability to distinguish between classes. To further evaluate classification performance, a **confusion matrix** was generated for classification model applied to the multi-class classification of CCE.

Results

The performance of the CNN classifier DenseNet169, which distinguishes between MBWs and PMBWs (the BW type classification task), and the performance of the DenseNet121 model, which categorises BWs as right or left (the BW side classification task), is summarised in Table 2.

Table 2. Performance of BW Type (Molar vs. Premolar) and Side (Left vs. Right) Classification Models

		BW Type			**BW side**	
Class distribution	Premolar	Molar	Total	Right	Left	Total
Number of BWs	208	195	403	173	230	403
Metrics			**Average**			**Average**
Precision (%)	95.7	95.9	95.8	98.8	99.1	99.0
Recall (Sensitivity) (%)	96.2	95.4	95.8	98.8	99.1	99.0
F1 Score (%)	95.9	95.6	95.8	98.8	99.1	99.0
AUC (%)	99.3	99.3	99.3	99.3	99.3	99.3
Accuracy (%)			**95.8**			**99.0**

Model used: DenseNet169 Model for type classification and DenseNet121Model for Bitewing Side Classification

The CCE was evaluated through two classification tasks: binary and multi-class. The binary classification task aimed to distinguish between the absence and presence of CCE, while the multi-class classification task further categorized the extent of CCE in the ROI into three classes: Minimal, Severe, and Critical. The CCE binary classification task using DenseNet161 showed excellent performance, achieving an overall accuracy of 96.3%. For the multi-class classification, the DenseNet121 model achieved an overall accuracy of 79.3%, with the highest performance observed in the 'critical' class. The evaluation metrics for the binary and multi-class classifications are shown in Table 3. The confusion matrix shown in Figure 6 illustrates the performance of the DenseNet-121 in classifying the extent of CCE in three classes. The matrix highlights the model's higher accuracy in identifying the 'Critical' and 'Minimal' classes, while a decreased performance for the 'Significant' class.

Table 3. Performance of Binary and Multi-Class Classification Models for CCE.

	Binary Classification (DenseNet161)			Multi-Class Classification (DenseNet121)			
Class distribution	Present	Absent	Total	Critical	Significant	Minimal	Total
Number of BWs	179	224	403	70	61	48	179
Metrics			**Average**				**Average**
Precision (%)	98.8	94.5	96.6	87.3	71.4	76.9	78.6
Recall (Sensitivity) (%)	92.7	99.1	95.9	88.6	65.6	83.3	79.2
F1 Score (%)	95.7	96.7	96.2	87.9	68.4	80.0	78.8
AUC (%)	98.6	98.6	98.6	94.9	72.2	94.5	87.2
Accuracy (%)			**96.3**				**79.3**

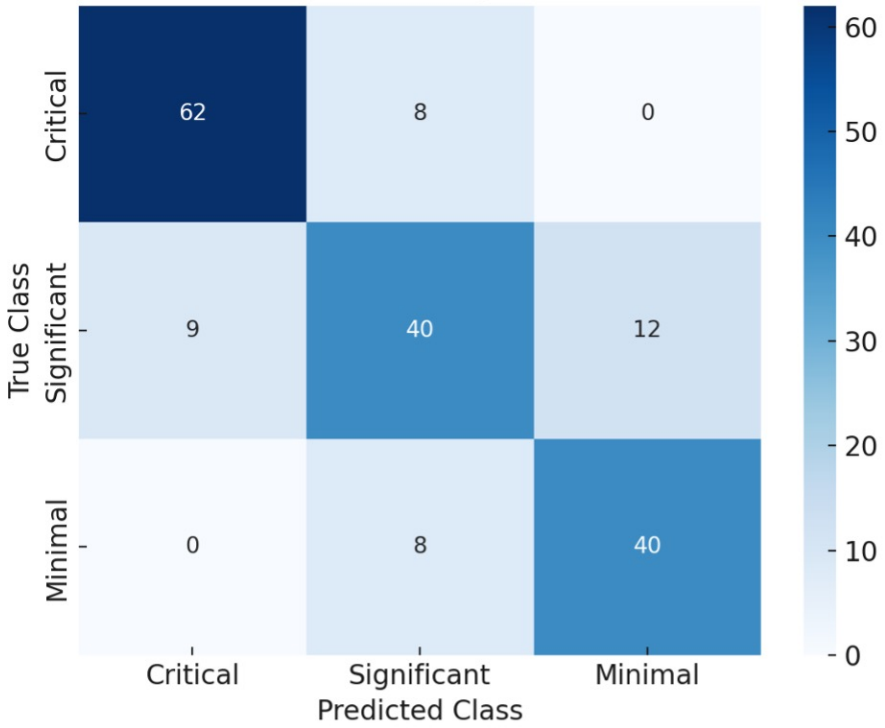

Figure 6. Confusion matrix representing the predicted versus actual classification outcomes for the DenseNet-121 classifier applied to BWs to classify the extent of CCE in three classes. (Minimal) CCE that does not affect the region of interest (ROI);

(Significant) CCE partially affecting the ROI, compromising three or fewer interproximal areas; and (Critical) CCE significantly affecting the ROI (resulting in a non-diagnostic radiograph), where more than three interproximal areas are not visible.

The CNN classifier demonstrated strong performance in detecting interproximal overlap, achieving an overall accuracy of 93.4%. The model maintained consistent performance across key metrics, with average precision, recall, and F1 score all reaching 93.4%. The AUC was 97.8%, indicating a high ability to distinguish between images with and without interproximal overlap. These results are shown in Table 4.

The performance of the ResNet34 model for the receptor placement binary classification task is summarised in Table 4. The model achieved an overall accuracy of 73.2%. The F1 score reflects a balance between precision and recall and is slightly higher for the 'Present' class (75.5) than for the 'Absent' class (70.4). This indicates that the model is more effective at detecting incorrect receptor placement than at correctly identifying cases without issues. The AUC, 79.4%, indicating moderate overall performance.

Table 4. Performance of Interproximal Overlap and Incorrect Receptor Placement Classification Models

	Interproximal Overlap			**Incorrect Receptor Placement**		
Class distribution	Present	Absent	Total	Present	Absent	Total
Number of BWs	171	162	333	155	177	332
Metrics			**Average**			**Average**
Precision (%)	93.6	93.2	93.4	72.6	73.7	73.1
Recall (Sensitivity) (%)	93.6	93.2	93.4	68.4	77.4	72.9
F1 Score (%)	93.6	93.2	93.4	70.4	75.5	72.9
AUC (%)	97.8	97.8	97.8	79.4	79.4	79.4
Accuracy (%)			**93.4**			**73.2**

Model Used: Interproximal overlap: DenseNet161; Incorrect Receptor Placement: ResNet34

Discussion

The CNN classification models performed well in binary classification tasks for BW type (premolar and molar), side (right and left), and the positioning errors: CCE (presence and absence) and interproximal overlap (presence and absence), achieving accuracies of 95.8%, 99%, 96.3%, and 93.4%, respectively.

Recent studies on quality enhancements of intraoral radiographs using AI algorithms have primarily focused on post-processing techniques (18, 19). However, while such models enhance image clarity, they do not address positioning errors, which remain a significant cause of non-diagnostic radiographs. This study focuses on positioning errors such as CCE, interproximal overlap, and improper receptor placement, as they are among the most prevalent causes of unacceptable image quality in BW radiographs, as evidenced by previous research (3, 4).

A recent systematic review found that the complexity of a radiograph is one of the factors influencing errors in dental radiograph interpretation, which depends on the type (BW, periapical, CBCT) and the quality of the radiograph (image contrast, presence of imaging errors) (20). A previous study classified BWs based on diagnostic quality by categorising them as diagnostic or non-diagnostic, based on the presence of interproximal overlaps (21). However, no studies to date have explored the applicability of AI tools for the detection of specific imaging errors for either clinical or research purposes. The present work fills this gap and demonstrates good performance of CNN classification models in BW analysis. Our findings support the hypothesis that AI-based detection tools can effectively identify imaging errors. As suggested by recent review, such tools have the potential to address documented sources of dataset heterogeneity in dental imagery, therefore enhancing the robustness, generalizability, and interpretability of future AI downstream tasks (22).

The selection of mannequin-based radiographs taken by students as the dataset for this study offers a standardised setting that ensures consistent anatomical structures while maintaining variability in execution due to differences in student techniques. This controlled approach is particularly advantageous for assessing technical errors, as it eliminates patient-related variability, allowing for a more focused evaluation of acquisition-related faults.

The present results demonstrate, as a key advantage, the potential of these classifiers to be integrated into dental imaging management software, reducing the time spent correctly mounting images and detecting common positioning errors. An automated classification tool could prevent radiograph mounting errors during image processing, reducing misinterpretations and treatment errors. Integrating the type and side of BW tasks into a Clinical Decision Support System (CDSS) in dentistry helps define ROI. This allows region-specific imaging error detection, improving the accuracy of error identification and analysis. By recognising positioning errors upfront, the CDSS can determine a radiograph's diagnostic usefulness before attempting to detect any oral pathology. Lastly, beyond clinical applications, these distinction tasks can also be integrated into interactive dental learning resources to help students recognise and correct mounting and positioning errors.

Previous studies have explored similar approaches for side classification in BW radiographs. A previous study reported that a ResNet-34 model achieved 97.56% accuracy in recognizing inverted images (upside-down) using a relatively small dataset (23). A recent study employed DenseNet-121, pre-trained on the ImageNet dataset with transfer learning and fine-tuning, achieving 100% accuracy in distinguishing left and right in a 1000 BWs dataset (24). These studies are consistently demonstrating good results. In contrast, our study employed DenseNet-121 without transfer learning or fine-tuning, using a smaller dataset of 403 BWs and considered multiple positioning errors. Despite this raw approach, the model demonstrated excellent classification performance, with an overall accuracy of 99.0%. These results should be viewed in light off BWs taken from mannequins, which may have contributed to this high overall accuracy.

While the consensus in the literature indicates the use of transfer learning and fine-tuning for improved model performance with small datasets (25, 26) , the findings in our study challenge this assumption, showing that a model trained from scratch without relying on pretrained model can still achieve high accuracy in right-left classification in a small dataset of mannequin's BW radiographs BWs. In that respect, our study has used a mannequin radiographs, further investigation in clinical data should be conducted to confirm such findings.

The use of a positioning device with a locator ring can minimise the prevalence of CCEs, it remains one of the most frequently reported causes of diagnostically unacceptable images, leading to re-exposures (3). In this study, a multi-class classification was intentionally used to direct compare with the binary classification performance metrics. CCEs can compromise the ROI to various degrees, e.g. if the CCE affects the area of interest in a premolar/molar BW the image is non-diagnostic. Therefore, we propose an AI-driven workflow that first identifies and classifies the CCE as a first step, excludes all images with critical CCE as the second step, then, evaluates other positioning errors, and lastly detects oral pathologies.

The best classification model employed for the multi-class classification for CCE, DenseNet121, showed lower accuracy than the best model for binary classification, DenseNet161, where the accuracy decreased from 96.2% to 79.3%. It is expected, as we increase the class number and keep the same dataset size the accuracy tends to decrease (27). Based on the images with the wrong predictions, adding premolar and molar sub-class for each type of CCE would improve the model classification performance.

The AI model's confusion matrix analysis reveals its strengths and areas for improvement in classifying CCE by the extent to which they affect the region of interest (ROI). Overall, the model is most accurate at recognising the extremes, 'Minimal' and 'Critical' CCE, while it sometimes confuses 'Significant' CCE with the other two categories. The observation of the radiographs with incorrect predictions suggests that the primary issue lies in the ROI not being defined beforehand, prior to assessing CCE severity across the three classes. This pattern underscores the need for further refinement, such as adjusting decision thresholds, which means changing the cut-off values the model uses when deciding which class a case belongs, to more accurately distinguish 'Significant' CCE from its minimal or critical counterparts. This demonstrates, as suggested by a previous study (28) that the level of reporting for different positioning errors may need to be considered at object level or pixel level, therefore employing other types of computer vision tasks.

One of the most critical positioning errors in BWs is the interproximal overlap, which can compromise the assessment of proximal tooth surfaces, particularly in the diagnosis of caries. This error may arise from incorrect horizontal angulation or variations in tooth positioning. Because interproximal overlap can undermine diagnostic performance, studies have excluded BWs with this error when training/testing AI models (8, 9, 12, 29-31). Unfortunately, the performance of these models cannot be reliable in clinical routine as ideal radiographs, without this positioning error is rare. The classifier model, DenseNet161, trained in this study achieved a 93.4% accuracy in detecting BWs to detect the absence or presence of interproximal overlap. This represents a significant step forward in creating models capable of detecting overlap in a specific interproximal space to alert the pathology detection AI-tools to this positioning error, improving the diagnostic accuracy and a clinical practice routine.

Incorrect receptor placement was labelled using a classification template based on the vertical orientation of the long axis of the image receptor in relation to the floor. The correct receptor placement in BWs, the occlusal plane appears flat or rising upward from the midline, reflecting the natural curve of Spee. Any variation from this expected pattern was used to identify and classify incorrect receptor placement. Despite a modest overall accuracy of 73.2%, the model demonstrated a moderate level of performance, as reflected in the F1 score, which balances both precision and recall. The F1 score was slightly higher for the ‘present’ class (75.5%), indicating a stronger ability of the model to correctly identify actual incorrect receptor placement. Moreover, the relatively balanced precision and recall values across both classes suggest consistent performance, which may support more confident clinical decision-making. However, the findings also indicate that while the model is more effective at detecting placement issues, it may produce some false positives when classifying normal cases, suggesting room for further refinement to improve specificity.

The dataset used was relatively small; however, it was carefully selected and structured to ensure class balance, which is essential for fair model evaluation. This study focused on classification models, which may not be optimal for detecting all types of positioning errors; however, they are commonly used as the initial stage before further optimization. If satisfactory performance is achieved through whole-

image analysis, classification represents a preferable strategy. It minimizes errors introduced by additional preprocessing and requires fewer computational resources than segmentation or object detection CNN tasks (32). Additionally, the study was conducted using radiographs obtained from mannequins, which, although standardised, do not fully capture the variability seen in clinical settings.

Future research should expand on the proposed workflow for automatic identification of positioning errors by incorporating other model architectures or alternative CNN approaches. This study provides a foundation for large clinical trials employing multi-centre datasets with patient-acquired bitewing radiographs, aiming to enhance the robustness and generalizability of the results.

The initial assessment of the radiograph to advise the oral pathology AI detection model about the presence of positioning errors is expected to enhance the clinical applicability of AI systems. Additionally, the findings from this study demonstrate the potential of classification CNN models for the automated identification of BW side, type and positioning errors supporting radiographic technique training in educational and clinical settings.

Conclusions

Overall, our results indicate that an automated system of DL models is capable of satisfactorily detecting and classifying most of the common positioning errors in BW. For BW type (premolar and molar), side (right and left), and the positioning errors: CCE (presence and absence) and interproximal overlap (presence and absence), achieving accuracies in the range of 93%-99%. Regarding the incorrect receptor placement error, the model demonstrated a moderate level of performance, with an overall accuracy of 73.2%.

References

1. Mallya S, Lam E. White and Pharoah's oral radiology: principles and interpretation: Elsevier Health Sci.; 2018.
2. Whaites E, Drage N. Essentials of dental radiography and radiology: Elsevier Health Sci.; 2013.
3. Yeung AWK, Wong NSM. Reject Rates of Radiographic Images in Dentomaxillofacial Radiology: A Literature Review. Int J Environ Res Public Health. 2021;18(15):8076.
4. Acharya S, Pai KM, Acharya S. Repeat film analysis and its implications for quality assurance in dental radiology: An institutional case study. Contemp Clin Dent. 2015;6(3):392-5.
5. Dastgir Bhatti U, Nehra A, Tariq A, Rafique I, Shaikh G. Common Radiographic Errors in Dentistry. Acta Sci Dent Sci. 2020;4:01-4.
6. Peker I, Alkurt MT. Evaluation of radiographic errors made by undergraduate dental students in periapical radiography. N Y State Dent J. 2009;75(5):45-8.
7. Hung KF, Ai QYH, Leung YY, Yeung AWK. Potential and impact of artificial intelligence algorithms in dento-maxillofacial radiology. Clin Oral Investig. 2022;26(9):5535-55.
8. Chaves ET, Vinayahalingam S, van Nistelrooij N, Xi T, Romero VHD, Flügge T, et al. Detection of caries around restorations on bitewings using deep learning. J Dent. 2024;143:104886.
9. Gonzalez C, Badr Z, Güngör HC, Han S, Hamdan MD. Identifying Primary Proximal Caries Lesions in Pediatric Patients From Bitewing Radiographs Using Artificial Intelligence. Pediatr Dent. 2024;46(5):332-6.
10. Yuce F, Öziç MÜ, Tassoker M. Detection of pulpal calcifications on bite-wing radiographs using deep learning. Clin Oral Investig. 2023;27(6):2679-89.
11. Mahizha SI, Annrose J, Mano Christaine Angelo J, Domilin Shyni I, veda Giri Gv. Deep convolutional neural networks for early detection of interproximal caries using bitewing radiographs: A systematic review. Evidence-Based Dentistry. 2025.
12. García-Cañas Á, Bonfanti-Gris M, Paraíso-Medina S, Martínez-Rus F, Pradíes G. Diagnosis of Interproximal Caries Lesions in Bitewing Radiographs Using a Deep Convolutional Neural Network-Based Software. Caries Res. 2023;56(5-6):503-11.
13. Karakus R, Öziç MU, Tassoker M. AI-Assisted Detection of Interproximal, Occlusal, and Secondary Caries on Bite-Wing Radiographs: A Single-Shot Deep Learning Approach. JOURNAL OF IMAGING INFORMATICS IN MEDICINE. 2024;37(6):3146-59.
14. Estai M, Tennant M, Gebauer D, Brostek A, Vignarajan J, Mehdizadeh M, et al. Evaluation of a deep learning system for automatic detection of proximal surface dental caries on bitewing radiographs. Oral Surg Oral Med Oral Pathol Oral Radiol. 2022;134(2):262-70.
15. Bayrakdar IS, Orhan K, Akarsu S, Çelik Ö, Atasoy S, Pekince A, et al. Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. Oral Radiol. 2022;38(4):468-79.
16. Arsiwala-Scheppach LT, Castner NJ, Rohrer C, Mertens S, Kasneci E, Cejudo Grano de Oro JE, et al. Impact of artificial intelligence on dentists' gaze during caries detection: A randomized controlled trial. J Dent. 2024;140:104793.
17. Krois J, Garcia Cantu A, Chaurasia A, Patil R, Chaudhari PK, Gaudin R, et al. Generalizability of deep learning models for dental image analysis. Sci Rep. 2021;11(1):6102.
18. Moran MBH, Faria MDB, Bastos LF, Giraldi GA, Conci A. Combining Image Processing and Artificial Intelligence for Dental Image Analysis: Trends, Challenges, and Applications. EAI/Springer Innov Commun Comput2022. p. 75-105.
19. Latke V, Narawade V. Detection of dental periapical lesions using retinex based image enhancement and lightweight deep learning model. Image Vis Comput. 2024;146:105016.

20. Hegde S, Gao J, Vasa R, Nanayakkara S, Cox S. Australian Dentist's Knowledge and Perceptions of Factors Affecting Radiographic Interpretation. Int Dent J. 2024;74(3):589-96.
21. Barayan MA, Qawas AA, Alghamdi AS, Alkhallagi TS, Al-Dabbagh RA, Aldabbagh GA, et al. Effectiveness of Machine Learning in Assessing the Diagnostic Quality of Bitewing Radiographs. Appl Sci. 2022;12(19):9588.
22. Delamare E, Fu X, Huang Z, Kim J. Panoramic imaging errors in machine learning model development: a systematic review. Dentomaxillofac Radiol. 2024;53(3):165-72.
23. Chen MC, Chen CH, Chen MH, editors. Artificial Intelligence (AI) for Dental Intraoral Film Mounting. IFMBE Proceedings; 2020.
24. Ayhan B, Ayan E, Bayraktar Y. A novel deep learning-based perspective for tooth numbering and caries detection. Clin Oral Investig. 2024;28(3):178.
25. Dalkıran A, Atakan A, Rifaioğlu AS, Martin MJ, Atalay R, Acar AC, et al. Transfer learning for drug-target interaction prediction. Bioinformatics. 2023;39(39 Suppl 1):i103-i10.
26. Iman M, Arabnia HR, Rasheed K. A Review of Deep Transfer Learning and Recent Advancements. Technologies. 2023;11(2):40.
27. Luo C, Li X, Yin J, He J, Li D, Zhou J. How does the data set and the number of categories affect CNN-based image classification performance? J Softw. 2019;14(4):168-81.
28. Büttner M, Schneider L, Krasowski A, Pitchika V, Krois J, Meyer-Lueckel H, et al. Conquering class imbalances in deep learning-based segmentation of dental radiographs with different loss functions. J Dent. 2024;148.
29. Van Nistelrooij N, Chaves ET, Cenci MS, Cao L, Loomans BAC, Xi T, et al. Deep Learning-Based Algorithm for Staging Secondary Caries in Bitewings. Caries Res. 2024.
30. Panyarak W, Suttapak W, Wantanajittikul K, Charuakkra A, Prapayasatok S. Assessment of YOLOv3 for caries detection in bitewing radiographs based on the ICCMS™ radiographic scoring system. Clin Oral Investig. 2023;27(4):1731-42.
31. Bayati M, Savareh BA, Ahmadinejad H, Mosavat F. Advanced AI-driven detection of interproximal caries in bitewing radiographs using YOLOv8. Sci Rep. 2025;15(1).
32. Zhao X, Wang L, Zhang Y, Han X, Deveci M, Parmar M. A review of convolutional neural networks in computer vision. Artificial Intelligence Review. 2024;57(4):99.

Chapter 4: *[Manuscript] AI-Driven Feedback on Bitewing Radiographic Technique: A Comparative Study of Large Language Models*

AI-Driven Feedback on Bitewing Radiographic Technique: A Comparative Study of Large Language Models

Short title: **Large Language Models for Bitewing Technique Feedback**

Katharina Alves Rabelo[1], Laura Swinckles[1,2,3] Eduardo Delamare[1], Zhuoran Duan[4], Elvis Trinh[1], Yongpei Ma[4], Jinman Kim[4], Vesna Miletic[1]

[1] Faculty of Medicine and Health, Sydney Dental School, The University of Sydney, Sydney, Australia

[2] Department of Oral Public Health, Academic Centre for Dentistry Amsterdam, University of Amsterdam and Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[3]Inholland University of Applied Sciences, Amsterdam, Netherlands

[4]Faculty of Engineering, School of Computer Science, The University of Sydney, Sydney, Australia

*Corresponding author: Katharina Alves Rabelo

Faculty of Medicine and Health, Sydney Dental School, The University of Sydney.

Address: 2 Chalmers Street, Surry Hills NSW 2010, Australia

E-mail: katharina.alvesrabelo@sydney.edu.au

Acknowledgements

Data collection and survey management were conducted using REDCap, facilitated by The University of Sydney.

Funding: The authors received no specific funding for this work.

Abstract

Introduction: Large Language Models (LLMs) can enhance education by making learning more engaging and accessible, though their performance in dentistry is inconsistent. Bitewing radiographs are critical for caries detection but prone to diagnostic errors from operator positioning. This study compared the performance of four LLMs with reasoning capabilities, GPT o1, GPT o3-mini, Gemini 2.0 Flash and Grok 3, in providing feedback on positioning errors in bitewing radiographs.

Materials and Methods: Ten positioning error scenarios were created by specialists in dentomaxillofacial radiology (DMFR). Each LLM was tested using both original and engineered prompts via web-based interfaces. Eighty outputs were assessed by two DMFR specialists for detail, content, clarity, and relevance on a 4-point Likert scale. Statistical analysis included Wilcoxon signed-rank and Kruskal–Wallis tests with Dunn's post-hoc comparisons and Bonferroni correction ($p<0.05$).

Results: Most outputs generated with original prompts scored higher than those with engineered prompts. GPT-o1 was the only model to show significant improvement with the engineered prompt (69.38%). LLMs generally performed better when evaluating incorrect receptor placement than interproximal overlap. GPT-o3-mini showed the weakest performance across content assessment category and interproximal overlap scenarios. GPT-o1 and Gemini 2.0 with engineered prompts achieved the highest scores for incorrect film placement, while Gemini 2.0 and Grok 3 with original prompts performed best on interproximal overlap.

Conclusion: The tested LLMs demonstrated variable feedback quality across different categories relevant for bitewing positioning errors but were consistently more accurate in evaluating incorrect receptor placement. Engineered prompts did not consistently enhance output quality across models.

Keywords: artificial intelligence; large language models; chatbots; dental education; dental radiology; bitewings

Introduction

AI-based large language models (LLMs), including family of models that includes OpenAI's ChatGPT, Google Gemini, Grok and Llama demonstrate broad intelligence in understanding instructions and delivering information. Through the application of deep learning algorithms, these AI models have iteratively refined their performance and interaction capabilities (1). AI capabilities, such as reasoning, planning, decision-making, and in-context learning, have enable LLMs to produce human-like output with user-friendly access and quick response times, making them effective as educational tools (1-4)

In dentistry, LLMs have been assessed as a supplementary tool for diagnosis and decision-making across various dental specialties, such as orthodontics (5, 6), prosthodontics (7) and endodontics (8) demonstrating potential to support clinical decision-making, leading to evidence based treatment modalities, consequently improving patient care. The application of LLMs in patient education has also been acknowledged, due to their accessibility and capacity to simulate human-like communication. This application has been investigated in several dental topics including oral and maxillofacial surgery (9), gingival and endodontic health (10), implantology (11), dental trauma (12) and oral cancer (13).

Initially, the use of LLMs in education raised concerns, particularly regarding written assessments, due to the potential for plagiarism and academic dishonesty. Despite early concerns, currently, there is a growing emphasis on advancing educational methods through the integration of AI. ChatGPT, a text-based interface, has been considered as a tool that supports effective self-directed learning and serves as an adjunct to enhance group-based educational activities (14).

In dental education, the development of innovative educational methods using LLMs have been highlighted. This includes generating individualized learning resources based on student preferences (15) and implementing chatbots powered by ChatGPT-4 (16). The use of ChatGPT in an undergraduate dental course revealed that students who used it for a learning assignment achieved better results on knowledge assessments than those who used conventional research methods, suggesting that integrating LLMs into the curriculum may enhance learning outcomes and

engagement (17, 18). However, some concerns were raised about occasional inaccuracies and unreliable citations in dentistry (19).

Research has shown that GPT-4 performed well on multiple-choice questions, scoring above the passing mark in dental licensing examinations in both the US and the UK (20). GPT-4 and Claude3-Opus achieved overall scores exceeding the cut-off scores and performed exceptionally well in specific subjects in the Korean Dental Licensing Examination (21). Additionally, a recent systematic review and meta-analysis concluded that GPT-4 requires more dental training data to achieve accuracy above threshold required for clinical application and education (22).

In Dentomaxillofacial Radiology (DMFR) education, studies compared the performance of dental students and LLMs using text-based communication in answering exam questions. ChatGPT-4 has matched the performance of 4th-year dental students, achieving the same scores of 33.3% for multiple-choice questions (MCQs) and 20% in open-ended questions (23). Another study comparing ChatGPT, ChatGPT Plus, Bard, and Bing Chat revealed that their performance ranged from 50% to 63.5% while students' overall accuracy scored 81.2%. All LLMs have shown better performance in short-answer questions (SAQs) than MCQs and performed poorly in image interpretation questions, achieving accuracies below 35% (24).

Despite these advances, the above studies assessing LLMs in DMFR education use these models sporadically and without validation and clear linkage to learning objectives. Thereby, it cannot be considered as AI-driven learning tools integral to dental curricula. An obvious starting point would be to implement LLMs in the basic learning outcomes in DMFR practical teaching, by providing detailed written feedback to students, for example in radiographic technique sessions

One of the areas of interest is bitewings (BW) as key diagnostic tools for caries detection. Despite the availability of positioning devices for the BW technique, positioning errors may occur due to operator's technical skills. LLMs have the potential to overcome existing limitations in DMFR clinic practical sessions and provide timely and consistent feedback to dental students. Therefore, the aim of the study was to compare the performance of open-source LLMs in providing feedback on positioning errors in BW radiographs.

Materials and Methods

Ethics Approval

This study did not include human participants or any patient data; therefore, approval from a research ethics board was not necessary.

Design

Ten BWs positioning error scenarios, based on common errors and frequent student doubts regarding their causes and corrections, were developed by specialists and lecturers in DMFR with over eight years of experience.

The original prompt (Prompt A) was written to resemble the output of a computer vision model, using dot points and ending with a standard question across all scenarios: 'Why did it happen and how should I correct it?'.

A second, engineered prompt (Prompt B) was a refinement of Prompt A, based a prompt engineering method of systematic prompt design. This study used specific techniques for effective prompt engineering modifying the prompt design (25).

Before the final version of the engineered prompt (Prompt B), testing and optimizations using GPT-4o and Grok 3 were performed. These LLMs have shown to diverge in terms of output lengths during the test phase. The GPT-4o has shown a more desirable output length with reasonable detail. In contrast, Grok 3 (think mode) has generated an output with a significant amount of redundancy. For these reasons, three modifications on the original prompt (Prompt A) were done. The following sentences were added in the engineered prompt: 'The following is a dentistry scenario' at the beginning and 'Assume that you are now a dental radiologist' before following the query. In addition, after many rounds of testing, the query was modified to use more formal and structured wording, and to limit the output length, as follows: 'Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details'. The final engineered prompt (Prompt B) which retained the dot points describing the positioning errors from the original prompt (Prompt A) with three changes mentioned, yielded stable responses from both test LLMs. An example of Prompt A and B can be seen in the Table 1, and all prompts are shown in Supplementary Table 1.

Table 1. Example of two types of prompts: original (Prompt A) and engineered prompt (Prompt B).

Original - Prompt A	**Engineered prompt - Prompt B**
A left molar bitewing shows: • The area of interest is completely visible • No cone cut • Interproximal overlap present, with palatal and lingual cusps mesial to the buccal cusps. • Centralised occlusal plane Why did it happen and how should I correct it?	The following is a dentistry scenario: A left molar bitewing shows: • The area of interest is completely visible • No cone cut • Interproximal overlap present, with palatal and lingual cusps mesial to the buccal cusps. • Centralised occlusal plane Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

Four different LLMs with reasoning capabilities, GPT o1, GPT o3-mini-2025-01-31, Gemini 2.0 Flash Thinking (experimental), and Grok 3 (Think Mode) were assessed in this study (Table 2). The default parameters of each model including temperature and maximum token length were used.

All models were tested using the same set of radiographic scenarios and two standardized prompt designs (Prompt A and B) were used in each LLM for direct comparison of their outputs. Each positioning error scenario with the query prompt, was submitted once to each LLM by one of the authors (L.S. or D.D.), without any follow-up questions, rephrasing, or additional input. Additionally, each prompt, was introduced individually using the “new chat” option to ensure standardized and context-independent responses.

Original prompts (Prompt A) and engineered prompts (Prompt B) were used in the web text-based interfaces, the GPT-series models were accessed using the OpenAI Application Programming Interface (API) to generate responses from GPT o1 and GPT o3-mini-2025-01-31. Gemini 2.0 Flash (experimental) was accessed via the Google's Gemini API through Google AI Studio platform, while Grok 3 (Think mode) was evaluated through its web interface, as no public API was available.

Table 2. Summary of LLMs assessed in the study.

	Model name	**Model parameters (Billion)**	**Organization & Country**
1	GPT o3-mini-2025-01-31 †	Estimate 30 (26)	OpenAI, USA
2	GPT o1	Estimate 200 (27)	OpenAI, USA
3	Gemini 2.0 Flash Thinking (experimental) ‡	Estimate above 20 (28)	Google, USA
4	Grok 3 (Think Mode)	Estimate 10-20 (29)	xAI, USA

† The model parameters of GPT o3-mini were not officially reported. This parameter size was reported by the developer community.

‡ No source mentions any information about the parameters of the Gemini 2.0 Flash, so this speculation is based on estimates of the parameters of the Gemini 1.5 Flash on the web.

Output Analysis

A total of 80 outputs were randomly organised in a file by one investigator (L.S.) and evaluated by consensus between two DMFR specialists (K.R. and E.D.) using Research Electronic Data Capture (REDCap) survey across four categories: (1) extension of detail, (2) content, (3) clarity, and (4) relevance. The assessment rubric used to assess LLMs outputs was adapted from (6) and is available in the Supplementary Table 2. A 4-point Likert scale (0: poor; 1: fair; 2: good; 3: very good; 4: excellent) was used for assessment. The outputs were provided to the examiners in a randomized order, numbered from 1 to 80, to ensure blinding and prevent identification of the LLM and the type of prompt. Outputs are available in Supplementary Table 3.

Statistical Analysis

Wilcoxon signed-rank test was used to assess differences in total scores between the original prompt (Prompt A) and the engineered prompt (Prompt B) within the same LLM. For inter-group comparison within each category, the score achieved in that category was used to evaluate differences. A total of eight groups, four LLMs using two different prompts, were compared using Kruskal–Wallis' test followed by Dunn's post-test pairwise comparisons with Bonferroni correction to control the overall p-value in multiple comparisons. Furthermore, the same non-parametric approach (Kruskal–Wallis followed by Dunn–Bonferroni tests) was applied to evaluate differences on model performance on interproximal overlap scenarios (Scenarios 1–3) and incorrect film placement scenarios (Scenarios 6-10) by LLMs using original (Prompt A) and engineered (Prompt B) prompts. The data were analysed using SPSS software (SPSS Version 29; IBM, New York, NY, USA) with the level of significance set at $p<0.05$.

Results

The overall performance of each LLM using the original prompt (Prompt A) and the engineered prompt (Prompt B) in providing feedback on positioning error scenarios is shown on Table 3. GPT o1 showed a statistically significant improvement in total score when using the engineered prompt (69.38%) compared to the original prompt (55%) ($p = 0.041$; Wilcoxon Signed Ranks Test). In contrast, no significant differences were observed for the other models ($p>0.05$).

Table 3. Model performance using the original and the engineered prompts.

LLM	Original Prompt (Prompt A) - Total Score	Engineered Prompt (Prompt B) – Total Score	p-value
GPT o3-mini	47 (29.38%)	44 (27.5%)	0.953
GPT o1	88 (55%)	111 (69.38%)	0.041*
Gemini 2.0 Flash Thinking	104 (65%)	101 (63.13%)	0.441
Grok 3 (Think Mode)	99 (61.88%)	52 (32.50%)	0.168

Note: Wilcoxon Signed Ranks Test.
Original prompt (Prompt A) and engineered prompt (Prompt B).
Maximum possible total score 160.
*Statistically significant difference ($p<0.05$).

LLMs outputs were evaluated across four categories: Extension of detail, Content, Clarity, and Relevance by Kruskal–Wallis's test. Results demonstrate statistically significant differences between LLMs in Clarity ($\chi^2(7) = 22.68$, $p = .002$) and Relevance ($\chi^2(7) = 22.08$, $p = .002$) categories. No significant differences were observed in Content ($\chi^2(7) = 9.81$, $p = .200$) and Extension of detail ($\chi^2(7) = 7.43$, $p = .386$). These results are shown in Figure 1.

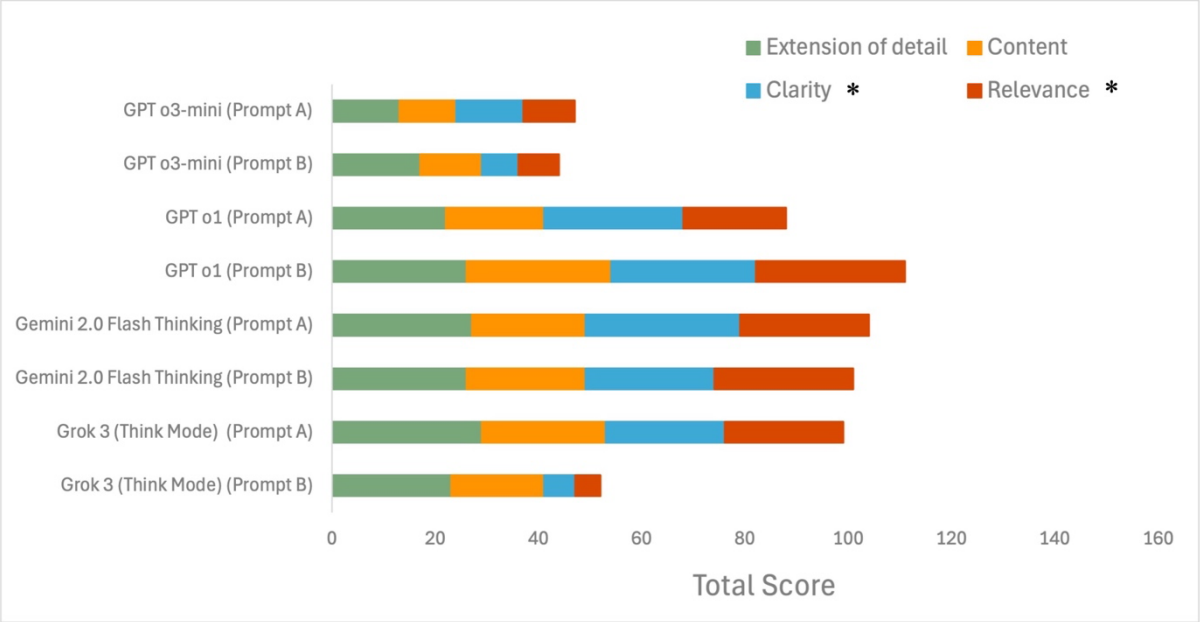

Figure 1. Total score per model per assessment category using two different prompts (Prompt A and Prompt B). Note: Kruskal–Walli's test. Maximum possible score per category 40. Maximum possible total score 160. Original prompt (Prompt A) and engineered prompt (Prompt B). *Statistically significant differences in *Clarity* ($\chi^2(7)=22.68$, $p=0.002$) and *Relevance* ($\chi^2(7)=22.08$, $p=0.002$).

There was a statistically significant difference in clarity observed between Gemini 2.0 (Prompt A) and Grok 3 (Prompt B), with Grok 3 scoring significantly lower (Mean Difference=32.800, adjusted p=0.028; Dunn's post-test with Bonferroni correction). In the relevance category, Grok 3 (Prompt B) was significantly outperformed by GPT o1 (Prompt B) (Mean Difference=32.350, adjusted p=0.034). Other comparisons approached significance (adjusted $p < 0.10$) but did not meet the Bonferroni-corrected threshold. These results are summarized in Table 4.

Table 4. Pairwise comparisons for clarity and relevance assessment categories.

Comparison	Category	Mean Difference	Raw p-value	Adjusted p-value †
Gemini 2.0 (Prompt A) vs Grok 3 (Prompt B)	Clarity	32.800	<0.001	**0.028***
GPT o1 (Prompt B) vs Grok 3 (Prompt B)	Clarity	30.450	0.002	0.062
Gemini 2.0 (Prompt A) vs GPT o3-mini (Prompt A)	Clarity	30.850	0.002	0.055
Gemini 2.0 (Prompt A) vs Grok 3 (Prompt B)	Relevance	28.800	0.004	0.110
Gemini 2.0 (Prompt B) vs Grok 3 (Prompt B)	Relevance	30.350	0.002	0.066
GPT o1 (Prompt B) vs Grok 3 (Prompt B)	Relevance	32.350	0.001	**0.034***

Note: Dunn's post-test with Bonferroni correction. The significance level was 0.050. Original prompt (Prompt A) and engineered prompt (Prompt B)

† Significance values have been adjusted by the Bonferroni correction for multiple tests.

* Statistically significant differences ($p<0.05$).

For the interproximal overlap scenarios (Scenarios 1–3), there was a statistically significant difference in output scores across the eight models ($\chi^2(7) = 57.171$, $p < 0.001$; Kruskal–Wallis test). As shown in Figure 2, Gemini 2.0 and Grok 3 using the original prompt (Prompt A) demonstrated the best performance, achieving higher scores by providing more effective feedback on interproximal overlap. In contrast, GPT o3-mini using both prompts (Prompts A and B) received the lowest scores, indicating the weakest performance in providing feedback on scenario related to this positioning error. In addition, GPT o1 showed a moderate performance, better than o3-mini but lower than the top-performing models, Gemini 2.0 and Grok 3. Lastly, models generally performed better when using the original prompt, for instance, Grok 3 with Prompt A outperformed its engineered prompt (Prompt B) counterpart.

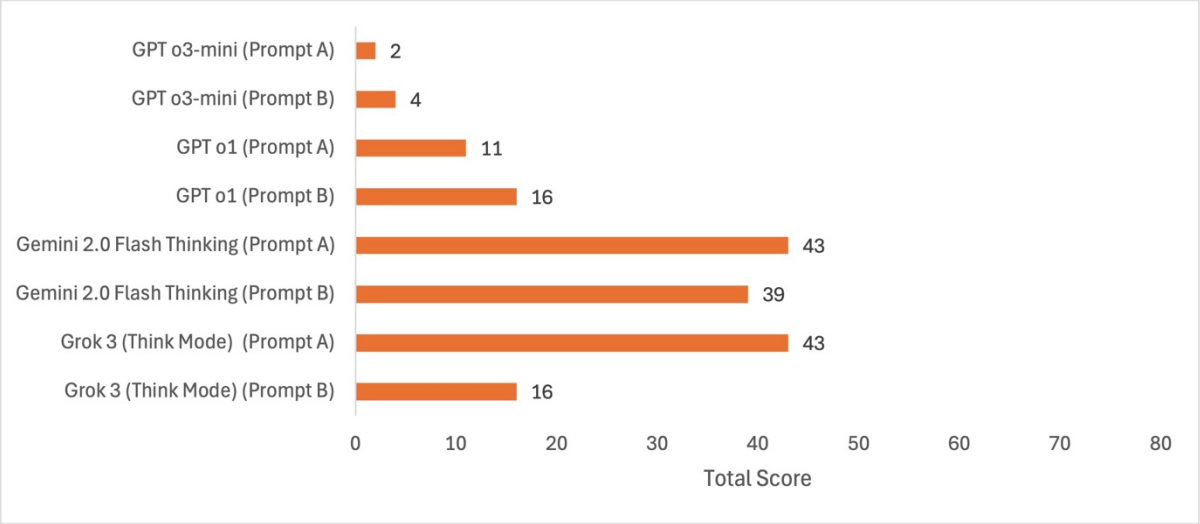

Figure 2. LLMs performance on interproximal overlap scenarios (Scenarios 1–3) by LLMs using original (Prompt A) and engineered (Prompt B) prompts. Note: Maximum possible score 80. Original prompt (Prompt A) and engineered prompt (Prompt B).

Pairwise comparisons using Dunn's test with Bonferroni correction are summarized in Table 5. The results revealed that Gemini 2.0 and Grok 3 using original prompts (Prompt A) performed significantly better than GPT 03-mini using both prompts (Prompts A and B), GPTo1 (Prompt A) and Grok 3 (Prompt B) ($p < 0.05$). Additionally, GPT o3-mini, using both the original and engineered prompts, was significantly outperformed by Gemini 2.0 and Grok 3. Furthermore, original prompt (Prompt A) used in Grok 3 was significantly more effective compared to the engineered prompt (Prompt B) (adjusted p-value=0.039).

Table 5. Summary of Pairwise Comparisons for Interproximal Overlap Scenarios Output scores for interproximal overlap

Comparison	Mean Difference	Raw p-value	Adjusted p-value †
Gemini 2.0 (Prompt A) **vs GPT o3-mini (A)**	52.750	<.001	.000*
Gemini 2.0 (Prompt A) **vs GPT o3-mini** (Prompt B)	50.000	<.001	.000*
Gemini 2.0 (Prompt A) **vs GPT o1** (Prompt A)	40.667	<.001	.005*
Gemini 2.0 (Prompt A) **vs Grok 3** (Prompt B)	34.833	.001	.039*
Gemini 2.0 (Prompt B) **vs GPT o1** (Prompt A)	34.792	.001	.039*
Gemini 2.0 (Prompt B) **vs GPT o3-mini** (Prompt A)	46.875	<.001	.000*
Grok 3 (Prompt A) **vs GPT o3-mini** (Prompt A)	52.750	<.001	.000*
Grok 3 (Prompt A) **vs GPT o3-mini** (Prompt B)	50.000	<.001	.000*
Grok 3 (Prompt A) **vs GPT o1** (Prompt A)	40.667	<.001	.005*
Grok 3 (Prompt A) **vs Grok 3** (Prompt B)	34.833	.001	.039*

Note: Kruskal Wallis' test followed by Dunn's post-test with Bonferroni correction. Original prompt (Prompt A) and engineered prompt (Prompt B).

† Significance values have been adjusted by the Bonferroni correction for multiple tests.

* Statistically significant differences (p<0.05).

The total scores for the incorrect film placement scenarios (Scenarios 6–10) outputs from each LLM using prompt A and prompt B are shown in Figure 3. There was a statistically significant difference in output scores among the eight models ($\chi^2(7) = 33.45$, $p < .001$; Kruskal–Wallis's test). These results indicate that the selected LLMs achieved higher scores when providing feedback on incorrect film placement compared to interproximal overlap. Additionally, GPT o1 using the engineered prompt (Prompt B) demonstrated the highest score, while Grok 3 using the same engineered prompt (Prompt B) showed the lowest performance.

The pairwise comparisons done by Dunn's post-test with Bonferroni correction identified statistically significant differences (Figure 3). Grok 3 (Prompt B) was significantly outperformed by three models in providing feedback on incorrect film placement. GPT o1 (Prompt A) scored significantly higher than Grok 3 (Prompt B) (mean difference=47; raw $p<0.001$; adjusted $p=0.027$), as did GPT o1 (Prompt B) (mean difference=70; raw $p<0.001$; adjusted $p<0.001$), and Gemini 2.0 (Prompt B) (mean difference=63; raw $p<0.001$; adjusted $p<0.001$). These results indicate that Grok 3 (Prompt B) produced the weakest feedback outputs for incorrect film placement scenarios, while GPT o1, particularly when using the engineered prompt (Prompt B), was the most effective.

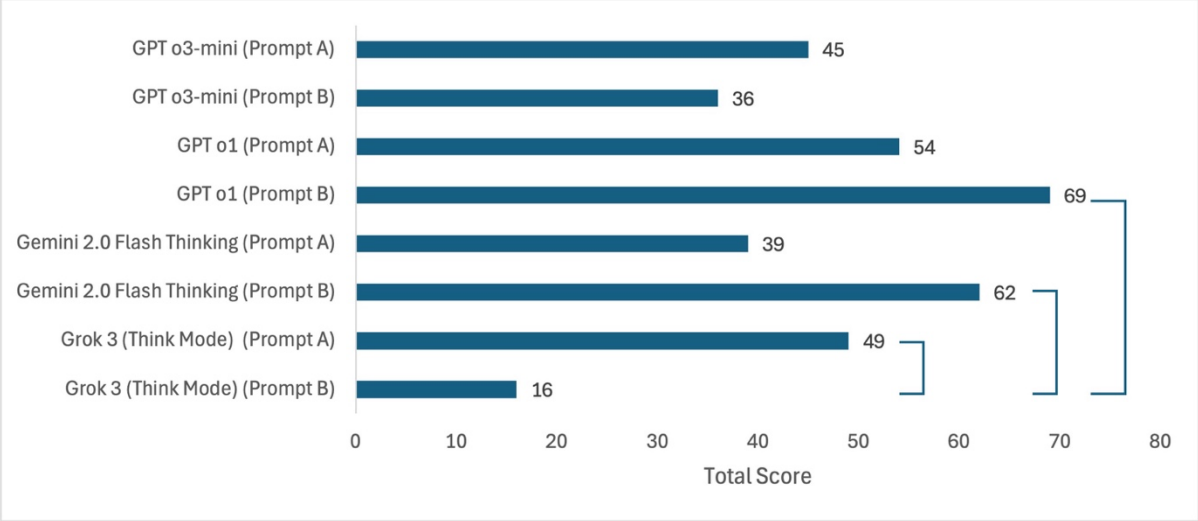

Figure 3. LLMs performance on incorrect film placement scenarios (Scenarios 1–3) by LLMs using original (Prompt A) and engineered (Prompt B) prompts. Note: Kruskal Wallis' test followed by Dunn's post-test with Bonferroni correction. Original prompt (Prompt A) and engineered prompt (Prompt B). Maximum possible score 80. Groups connected by lines are statistically significantly different (adjusted $p<0.05$).

Discussion

This study presents a comprehensive comparison of the performance of multiple LLMs in providing feedback on positioning errors in BW radiographs, considering different prompt designs and output-related characteristics.

Detailed and actionable feedback is crucial for improving learning outcomes and avoiding repeating mistakes. In dental radiography, imaging errors feedback can improve clinician's radiographic acquisition techniques, reduce the number of radiation exposures, improve patient's experience and enhance student's learning experiences. LLMs offer instant responses, personalised interactions through a user-friendly platform with extensive knowledge, making them a potentially effective tool for self-directed learning and an adjunct during clinical sessions to support students in dental schools.

Some studies using LLM to answer open-ended questions in the dental domain have shown promising results. The accuracy of ChatGPT-4 in providing answers to open-ended questions for general dentists in the field of oral surgery reached 71% accuracy (30). In addition, ChatGPT-4 achieved 80% correct diagnosis of oral and maxillofacial diseases in 37 clinical cases (31). These studies show the potential of LLMs to become a supplementary education tool in dentistry; however, careful consideration should be given to inaccuracies as previous highlighted by a recent systematic review (19).

Recently, it has been demonstrated that the prompt has a considerable influence on the model's behaviour (32). In addition, an enhanced prompt can enhance model's performance on the way that special fine-tuning might be unnecessary (33, 34). The process of systematic designing and optimising the input prompt to achieve desirable responses is one of the techniques of prompt engineering (35). Based on this context, the present study assessed the outputs generated by four LLMs with reasoning capabilities using two different prompts: original prompt and an engineered prompt proposed previously (25). The same approach has been adopted in previous studies (36, 37).

Most of the outputs generated by LLMs using the original prompt achieved higher total scores than those from the engineered prompts. The only exception was GPT-o1,

which, having the largest parameter size, showed a statistically significant improvement using engineered prompts compared to its baseline ($p = 0.041$). These findings contrast with studies in the dental field using GPT 4 for implant-supported prostheses related prompts (38) and in medical sciences area using GPT 3.5-turbo (39), where engineered prompts outperformed originals. The prompt engineering method adopted in this study was previously proposed by Ekin (25). This method is specifically focused on ChatGPT and may have influenced these results, as only the complete ChatGPT model, GPT o1, had the outputs improved with the engineered prompt (Prompt B).

Additionally, all models assessed in this study, GPT o3-mini, GPT o1, Gemini 2.0, and Grok3 are general-purpose LLMs advertised as having built-in reasoning capabilities. Therefore, these findings suggest that for modern LLMs with reasoning skills, systematic prompt design method may not always be needed; a well-framed and clear prompt may be sufficient. This prompt engineering method may be more important for LLMs without advanced reasoning capabilities.

Overall, the selected LLMs, GPT o1, GPT o3-mini, Gemini 2.0 flash and Grok3, achieved total scores between 27.5% and 69.38% when generated outputs using two prompts (Prompt A and B), demonstrating substantial performance variability. The performance of LLMs in answering open-ended question varies a lot, ranging from 52% to 71%, showing that some responses can be too general, inaccurate, or lacking evidence-based support (30, 40). Two studies achieved higher performance in open-ended questions in oral surgery and oral and maxillofacial diagnosis using GPT 4 with 71% and 80% accuracy, respectively (30, 31). The comparisons between these studies are limited. The differences may be attributed to model capacity and architecture, domain specific and prompt-model interaction. For tasks that depend on specialised knowledge widely available online, such as disease descriptions or oral-surgery management guidelines, the larger-parameter and longer-trained GPT 4 is generally the most suitable option. By contrast, generating feedback on the causes and corrections positioning errors in radiographs demands robust chain-of-thought reasoning. GPT o1 was selected for this purpose based on its reasoning skills. Therefore, the optimal LLM depends on the task demands to each model architectural strengths.

A more detailed analysis of LLM outputs based on four assessment categories showed that content and the extension of detail did not statistically differ between the eight groups (four models and two different prompts). Similar results were demonstrated in a study assessing LLM outputs in answering open-ended questions about dental avulsion, which found no significant difference between GPT 3.5 and Gemini v1.5 Pro model scores in the open-ended, descriptive questions (12).

Grok3 with engineered prompt was outperformed by Gemini 2.0 with original prompt in clarity and GPT o1 with engineered prompt in relevance. The differences were statistically significant. These results suggest that the optimisation in the prompt design produced only isolated improvements. In another study (38), assessing outputs about implant-supported prostheses, a specific (engineered) prompt enhanced the overall reliability of GPT 4 in clarity, relevance, and key-concept coverage from 70.9 % to 78.8 %. The same engineered prompt can boost or suppress performance depending on how well it aligns with a model's cognitive style (41, 42).

LLMs in this study found it easier to provide feedback on incorrect receptor placement than on interproximal overlap. This difference likely reflects positioning error complexity. The incorrect receptor placement error depends on a single factor, the position of the receptor. By contrast, interproximal overlap requires the assessment of three factors - the receptor placement position, the horizontal angulation of X-ray beam and the spatial relationship between buccal and lingual cusps of teeth, to determine the precise direction for horizontal angulation correction. It aligns with another study where a case-complexity factor was consistently associated with higher interpretation-error rates (43).

Furthermore, the engineered prompt made Grok3 perform worse but enhanced the outputs of GPT o1 for both interproximal overlap and incorrect receptor placement errors. The likely cause is that the prompt engineering using the systematic prompt design technique forced an unnatural chain of thought, which disrupted Grok's retrieval-based answers.

This highlights that the value of prompt optimisation is highly dependent on the model architecture and task characteristics. Therefore, prompt engineering techniques,

especially those related to prompt re-design, must consider the model architecture of the specific task. Additionally, Grok became publicly available in February 2025, and, to date, no peer-reviewed study has documented the use of prompt engineering for this model for comparison purposes.

For the interproximal overlap feedback and content category, GPT o3-mini using both prompts (Prompts A and B) showed the weakest performance. This suggests the small variant of GPT lacks the detailed spatial reasoning needed to explain radiographic positioning errors compared with full parameter GPT models. Additionally, even though the LLM was advertised as having reasoning capabilities, evidence shows that such abilities emerge only when the parameter scale is sufficiently large (44). Although mini versions cost about half the price of full versions, its performance is still inadequate to explain positioning errors in BW, especially when more than one factor is involved and more spatial reasoning is required. Consequently, despite the limitations of this study, small variants or experimental models should not be used for educational purposes.

In this study, the hyperparameter settings of the LLMs, including temperature, top-p, max_tokens, and seed values, were not adjusted and remained at their default configurations. Prompt engineering was kept minimalistic in this study, only giving guidance to the LLM as to who is providing the feedback. By defining the dental radiologist in the prompt, the LLM is expected to provide a more discipline specific feedback and expert advice. In the original prompt, Prompt A, this segment was kept open for interpretation by LLM. Other prompt engineering methods, such as chain-of-thought prompting or the prompt pattern catalog approach (Chen, Zhang et al. 2025), were not explored. However, currently LLMs with high performance in reasoning are likely to be less dependent on prompt engineering methods. Further research is needed to better understand how to enhance outputs for discipline-specific content feedback. Finally, this study employed a single-shot approach only, which prevented an assessment of the consistency or variability of the LLMs's outputs across multiple iterations.

Since the open-source LLMs, such as GPT, Gemini and Grok are not specifically trained on dental topics, future studies on the use of LLMs in DMFR education,

especially in radiographic techniques, should explore techniques such as retrieval-augmented method (45) to incorporate relevant external knowledge into the model's input, potentially improving the accuracy and relevance of their outputs. Additionally, the use of large vision language model should be investigated to enable the analysis of radiographic images allowing their performance in analysing the spatial relationship of structures, interpreting and providing feedback.

Conclusions

The systematic prompt design technique used in this study enhanced the performance of GPT o1 only, while GPT o3-mini consistently showed the lowest total scores with both the original and engineered prompts. No statistically significant differences were observed between LLMs for the Content and Extension of detail categories. However, Grok 3 using the engineered prompt (Prompt B) was outperformed in Clarity and Relevance categories by Gemini 2.0 with the original prompt (adjusted $p = 0.028$) and by GPT o1 with the engineered prompt (adjusted $p = 0.034$), respectively. Finally, higher scores were achieved in the LLMs' outputs when providing feedback on incorrect film placement error scenarios compared to interproximal overlap positioning error scenarios.

References

1. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. arXiv Preprint posted online July 2023. 2023.
2. Moskovich L, Rozani V. Health profession students' perceptions of ChatGPT in healthcare and education: insights from a mixed-methods study. BMC Med Educ. 2025;25(1):98.
3. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare (Basel) [Internet]. 2025; 13(6).
4. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study. JMIR Med Educ. 2024;10:e51344.
5. Albalawi F, Khanagar SB, Iyer K, Alhazmi N, Alayyash A, Alhazmi AS, et al. Evaluating the Performance of Artificial Intelligence-Based Large Language Models in Orthodontics—A Systematic Review and Meta-Analysis. Appl Sci (Basel) [Internet]. 2025; 15(2).
6. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod. 2024.
7. Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131(4):659.e1-.e6.
8. Özbay Y, Erdoğan D, Dinçer GA. Evaluation of the performance of large language models in clinical decision-making in endodontics. BMC Oral Health. 2025;25(1):648.
9. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023;124(5):101471.
10. Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of Large Language Models in Patient Queries on Gingival and Endodontic Health. Int Dent J. 2025;75(1):151-7.
11. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. J Prosthet Dent. 2025.
12. Tokgöz Kaplan T, Cankar M. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini. Dent Traumatol. 2025;41(2):178-86.
13. Ji K, Han J, Zhai G, Liu J. Assessing the Capabilities of Generative Pretrained Transformer-4 in Addressing Open-Ended Inquiries of Oral Cancer. Int Dent J. 2025;75(1):158-65.
14. O'Connor S. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? Nurse Educ Pract. 2023;66:103537.
15. Thorat VA, Rao P, Joshi N, Talreja P, Shetty A. The Role of Chatbot GPT Technology in Undergraduate Dental Education. Cureus. 2024;16(2):e54193.
16. Jones B, Desu A, Honig CDF. Artificial Intelligence Chatbots as Virtual Patients in Dental Education: A Constructivist Approach to Classroom Implementation. Eur J Dent Educ. 2025;n/a(n/a).
17. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study. JMIR Med Educ. 2024;10:e51344.
18. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. Eur J Dent Educ. 2024;28(4):865-76.
19. Alhazmi N, Alshehri A, BaHammam F, Philip M, Nadeem M, Khanagar S. Can Large Language Models Serve as Reliable Tools for Information in Dentistry? A Systematic

Review. Int Dent J. 2025;75(4):100835.
20. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of Generative Artificial Intelligence in Dental Licensing Examinations. Int Dent J. 2024;74(3):616-21.
21. Kim W, Kim BC, Yeom HG. Performance of Large Language Models on the Korean Dental Licensing Examination: A Comparative Study. Int Dent J. 2025;75(1):176-84.
22. Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, et al. Large Language Models in Dental Licensing Examinations: Systematic Review and Meta-Analysis. Int Dent J. 2025;75(1):213-22.
23. Öztürk HP, Avsever H, Şenel B, Ayran Ş, Peker MÇ, Özgedik HS, et al. ChatGPT in dentomaxillofacial radiology education. J Health Sci Med. 2024;7(2):224-9.
24. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol. 2024;53(6):390-5.
25. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. Authorea Preprints. 2023.
26. Vardhan H. OpenAI o3-mini: The cost-efficient genius redefining STEM AI 2025 [Available from: https://medium.com.
27. Thompson AD. o1: Smarter than we think 2024 [Available from: https://lifearchitect.ai.
28. Reddit. Reddit – the heart of the internet 2024 [Available from: https://www.reddit.com.
29. Byteplus. How many parameters will Grok 3 have? 2024 [Available from: https://www.byteplus.com.
30. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, et al. Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. 2024;24:46-52.
31. Tomo S, Lechien JR, Bueno HS, Cantieri-Debortoli DF, Simonato LE. Accuracy and consistency of ChatGPT-3.5 and −4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnostic performance analysis. Clin Oral Investig. 2024;28(10):544.
32. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv Preprint posted online July 2023. 2023.
33. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv Preprint posted online Nov 2023. 2023.
34. Maharjan J, Garikipati A, Singh NP, Cyrus L, Sharma M, Ciobanu M, et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. Sci Rep. 2024;14(1):14156.
35. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y). 2025;6(6):101260.
36. Nguyen D, Swanson D, Newbury A, Kim YH. Evaluation of ChatGPT and Google Bard Using Prompt Engineering in Cancer Screening Algorithms. Acad Radiol. 2024;31(5):1799-804.
37. Wang Y, Liang L, Li R, Wang Y, Hao C. Comparison of the Performance of ChatGPT, Claude and Bard in Support of Myopia Prevention and Control. J Multidiscip Healthc. 2024;17:3917-29.
38. Freire Y, Santamaría Laorden A, Orejas Pérez J, Ortiz Collado I, Gómez Sánchez M, Thuissard Vasallo IJ, et al. Evaluating the influence of prompt formulation on the reliability and repeatability of ChatGPT in implant-supported prostheses. PLoS One. 2025;20(5):e0323086.
39. Kee XLJ, Sng GGR, Lim DYZ, Tung JYM, Abdullah HR, Chowdury AR. Use of a

large language model with instruction-tuning for reliable clinical frailty scoring. J Am Geriatr Soc. 2024;72(12):3849-54.
40. Batool I, Naved N, Kazmi SMR, Umer F. Leveraging Large Language Models in the delivery of post-operative dental care: a comparison between an embedded GPT model and ChatGPT. BDJ Open. 2024;10(1):48.
41. Patel D, Raut G, Zimlichman E, Cheetirala SN, Nadkarni GN, Glicksberg BS, et al. Evaluating prompt engineering on GPT-3.5's performance in USMLE-style medical calculations and clinical scenarios generated by GPT-4. Sci Rep. 2024;14(1):17341.
42. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. 2024;7(1):20.
43. Hegde S, Gao J, Vasa R, Cox S. Factors affecting interpretation of dental radiographs. Dentomaxillofac Radiol. 2023;52(2):20220279.
44. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824-37.
45. Ng KKY, Matsuba I, Zhang PC. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. NEJM AI. 2025;2(1):AIra2400380.

Chapter 5: General Discussion and Conclusions

General Discussion and Conclusions

The aim of this thesis was to develop AI-driven deep learning tools for dental education to enhance the teaching and assessment of tooth cavity preparation and intraoral radiographic techniques.

The literature review highlighted the growing integration of AI in dental education, with potential to enhance student learning. The use of AI to provide feedback to students during practical sessions can enhance consistency, efficiency and support remote learning and teledentistry. In addition, AI technology has the potential to enhance objectivity, strengthen validity, and promote fairness in technical skill assessments. However, studies using CNN algorithms to assess dental procedures remain limited. Furthermore, studies exploring the use of LLM outputs to deliver feedback in practical sessions are rare. Therefore, this thesis investigated DL models to address these gaps and advance the implementation of AI in dental education.

5.1 Summary of key findings

5.1.2 Adjacent tooth damage

- YOLOv5 followed by DenseNet-169 models, employed to assess tooth damage on intraoral camera images, achieved 81.25% accuracy, outperforming clinical educators.

-A moderate agreement among CEs for assessing damage on tooth adjacent to a Class II cavity preparation in both model and image assessments.

-Inter-examiner reliability was higher for model-based than image-based human assessments at corresponding time points.

-The multi-class classification of damage was a challenge for the classification task performed by CNN algorithms. It perfectly detected 'damage requiring restoration' but was less accurate for 'damage not requiring restoration'.

5.1.3 Positioning errors in bitewing radiographs

-Several CNN models, including ResNet and DenseNet variants, were employed for classification.

-Models performed well in binary classification tasks for BW type (premolar and molar), side (right and left), and the positioning errors: CCE (presence and absence) and interproximal overlap (presence and absence), achieving accuracies of 95.8%, 99%, 96.3%, and 93.4%, respectively. Accuracy for detecting incorrect receptor placement was 73.2%.

-Multi-class CCE classification exhibited moderate performance (accuracy: 79.3%), with the model reliably identifying ‘critical’ and ‘minimal’ CCE.

-A workflow for an AI-driven tool to assess bitewings was proposed, beginning with bitewing type and side identification, followed by the assessment of positioning errors in sequence: cone cutting, interproximal overlap, and film placement.

5.1.4 Feedback on bitewing radiographic technique

-Most of the outputs generated by LLMs using the original prompt achieved higher total scores than those from the optimised prompts. The only exception was ChatGPT o1 which showed a statistically significant improvement in total score when using the optimised prompt (69.38%) compared to the original prompt.

-LLMs in this study showed higher scores on outputs providing feedback on incorrect receptor placement than on interproximal overlap.

-For the interproximal overlap feedback and content category, ChatGPT o3-mini using both prompts (Prompts A and B) showed the weakest performance.

-ChatGPT o1 (prompt B) and Gemini 2.0 (prompt B) showed the highest score in providing feedback on incorrect film placement.

-Gemini 2.0 (prompt A) and Grok 3 (prompt A) showed the highest scores in providing feedback on interproximal overlap positioning error.

5.2 General Discussion

This thesis explores the application of AI technology in dental education, specifically focusing on its potential for assessing dental procedures. The findings from this research have significant implications for the future integration of AI in dental education.

The first study employing CNN algorithms to detect and classify tooth damage demonstrated that intraoral camera images acquired from variable angles are suitable for training these DL models. This finding corroborates other studies that used intraoral camera images with CNN algorithms, achieving accuracies above 80% for the detection of tooth and oral diseases (1-3).

Therefore, intraoral camera images acquired by dental students during regular sessions can be used as input for DL models to support the assessment of dental procedures. Commercially available dental cameras represent a more affordable alternative compared to intraoral scanners. Although these cameras provide only two-dimensional images, appropriate angles that clearly depict the area that damage most likely to be present. These images can still provide sufficient information for AI model training and evaluation for dental procedures evaluation. This highlights their potential applicability as practical and cost-effective tools for integrating AI-based assessment and feedback systems into dental education.

In this thesis, Chapter 2, intraoral camera was used to obtain images from of prepared cavities and adjacent teeth in simulation models for AI model training. Images were captured in different angles showing the proximal surface of the adjacent tooth in each case for tooth damage identification. A non-standardised approach in angle selection was chosen to simulate students taking these intraoral camera images in their daily practice. The results have shown that even with different angles, AI models could identify and classify the damage in most of the cases. Using existing intraoral cameras without the need of extra equipment integrates efficiently AI with current simulation clinic setup.

Traditional educational approaches rely mainly on feedback from tutors, which can vary in quality and consistency (4). Corroborating with the literature, the first study demonstrated a higher accuracy of CNNs compared to traditional method of assessment tooth damage. This finding highlights the potential of CNN models to be integrated into the assessment of tooth cavity preparations. This implementation can make part of the practical pre-clinical and clinical assessments and feedback more objective, consistent, reliable, universal and scalable. Moreover, both CNN models used, YOLOv5 and DenseNet-169 are open-source architectures that can be leveraged to train models for evaluating other dental procedures assessable through photographic imaging.

By contrast, this innovative method is associated with additional costs due to the requirement for an intraoral camera. In addition, the specific algorithms applied require further refinement to effectively handle multi-class classification tooth damage.

The CNN-based multi-class classification in first and second studies showed reduced performance in intermediate categories, damage not requiring restoration and significant CCE. Given the fact that different image types were used, this limitation seems to be inherent to the CNN algorithm rather than imaging processing. Similar results were observed on damage tooth traditional assessment method conducted by CEs, where the intermediate category appeared to be the most challenging to evaluate. This tendency suggests that both CEs and CNNs are more reliable at recognising extremes. These results can be due to the overlapping features with both extremes categories. An analysis of the misclassified intraoral camera images indicated that, in some cases, irregularities in the texture of typodont teeth resulting from the manufacturing process contributed to misclassification. In addition, the absence of clear guidelines for determining when damage requires restoration, combined with the difficulty of distinguishing enamel and dentine layers on plastic teeth, made the annotation of these damages challenging even for senior academics. A potential solution to this limitation is the development of a quantitative rubric based on precise depth measurements of the damage**,** enabling consistent annotations for AI model training.

Regarding BWs, a potential solution to improve model performance in CCE multi-class classification is the introduction of premolar and molar subclasses within each CCE type. This approach could be implemented using segmentation or object detection models beforehand to determine the ROI, or alternatively through manual annotation with these subclasses.

The use of an object detection model, YOLOv5, to identify and restrict the ROI before the application of a classification model is an effective approach for detecting and classifying small regions within an image, such as tooth damage. In contrast, the bitewing study employed only a classification model without a preceding object detection step. Classification models are the default step to take prior to further optimisation, such as combining with other model architectures. If the classification task demonstrates satisfactory performance by analysing the whole image, it is the best approach. This strategy decreases the chance of errors that may arise from adding an extra pre-processing step, such as object detection task.

The second study revealed that classification models without the prior use of an object detection model using a binary classification approach, were effective in identifying the type and side of BWs, as well as most positioning errors. AI models that require lower computational demands are advantageous for implementation in educational settings, as it facilitates accessibility and scalability. Therefore, these findings suggest that in dental education, classification models should be prioritised for tasks where they can perform well, as they require less computational power.

The proposed workflow for an AI-driven bitewing assessment tool establishes a critical foundation for the educational and clinical implementation of multi-agent systems. This sequential approach, simulating the step-by-step reasoning of a human expert, enables a series of specialised agents to perform a more reliable analysis. A recent study demonstrated a significant improvement in performance, achieving near-perfect detection of pathologies in chest X-rays using an agentic AI system (5).

Hyperparameters are important for deep learning models training and performance. It is generally accepted that there is a common range of learning rates and epochs for each type of deep learning model, which influences how efficiently the model

learns and converges. In addition, these values depend on the training approach, from scratch or fine-tuned, and the specific characteristics of the task. It is crucial to set suitable hyperparameters to achieve a balance between underfitting and overfitting. In this thesis, learning rates and epochs were selected based on their model convergence.

The third study evaluated outputs generated by LLMs in providing feedback on the most common positioning errors in BWs. The models demonstrated performance variability depending on the LLM, the specific positioning error as well as the type of prompt, baseline or optimise. Prompt design techniques are part of the prompt engineering process that aim to generate more specific outputs (6). In the present study, all LLMs with reasoning capabilities, except GPT-o1, achieved higher total scores with the original prompts compared to the optimized prompts. These results suggest that the prompt optimization did not consistently enhance performance as expected. One possible explanation is that the prompt design method, which was primarily tailored to GPT models, may have contributed to this result. In addition, prompt design techniques may be more relevant for LLMs without advanced reasoning capabilities. Furthermore, the task demands and the architectural strengths of each LLM appear to have a greater influence on performance than prompt design.
LLMs are increasingly integrated into the daily learning resources of dental students and practitioners. Therefore, the findings from the third study, along with the recent literature demonstrating inconsistencies in LLMs performance in dental topics should be communicated to dental community. The high risk of bias and inaccuracies in the outputs, as well as the risk of overreliance, should be highlighted. This is essential to support informed adoption and responsible use in education and clinical practice.

The majority of LLMs, especially open source, are not specifically trained on dental topics. This limitation could be addressed by exploring advanced techniques to develop more specialised and reliable LLMs. Examples include fine-tuning, Retrieval-Augmented Generation, and a hybrid approaches that combine the strengths of both.

The GPT o3-mini showed reduced performance with both original and optimised prompts. This was evident in the content category for the all-positioning-errors scenario and in generating feedback on interproximal overlap scenarios. Consistent

with OpenAI documentation, the smaller o-series model, o3-mini, underperformed the full models (7). This outcome likely reflects differences in model scale and inference-time reasoning budget rather than prompt design. These findings suggest that lightweight variants of LLMs should be used with caution in educational contexts, as they may increase the risk of providing unclear or misleading feedback.

The integration of AI in dental education raises concerns regarding privacy and ethical issues (8-10). Concerns regarding data security and potential student or patient data leakage must be addressed. Sensitive health information can potentially be used to train or fine-tune AI models without a prior consent. Transparency and accountability should be ensured by de-identifying images and obtaining explicit consent from students or patients before using their images or information in AI models. Therefore, it is essential to train students and dental practitioners to manage data securely and to implement stronger encryption and security measures on platforms that use AI. Furthermore, ethical guidelines on the use of AI should be continuously reviewed, strengthened and implemented in dental education and clinical settings.

All three studies highlight the critical need for a careful, multi-step approach when deploying AI models for feedback or as an assessment tool. This thesis sets the foundation for large clinical trials that will use data from several universities in a multi-site study, aiming to make the results robust and universally applicable. Additionally, a future study will include hyperparameter sensitivity analysis experiments for further refinement of AI models and tasks. Therefore, future development must prioritise a methodical process for dataset selection, model training, and technology selection, as well as dedicated fine-tuning to ensure the AI's effectiveness and reliability for specific educational objectives.

5.3 Recommendations for future research

Based on the findings of this research, several key directions are proposed to advance the application of deep learning in dental education in pre-clinical and clinical settings. Future work should focus on improving AI model's generalizability by training on diverse datasets. This can be achieved by collaborations that incorporate patient BW datasets from different institutions, acquired using different X-ray units, receptor types,

technical abilities, and across varied demographic populations. For studies using intraoral camera images, diversity should also be ensured by including images captured with different intraoral cameras, image resolutions and general settings.

Another important direction for future research is to explore the integration of the same AI model across different universities and curricula, including adaptation to diverse disciplines and dental procedures. This should also explore the use of other imaging modalities, such as 3D intraoral scans, periapical and panoramic radiographs. For cavity preparations, AI models could assess criteria like cavity outline and clearance of contact. In restorative dentistry, CNNs could assess restorations using already adopted criteria (11), such as aesthetic and structural integrity criteria. In endodontics, CNNs could be applied to assess root canal preparation, obturation and long-term success of the procedure using radiographs.

5.4 Conclusions

AI is capable of assessing different types of images, intraoral camera images and bitewings, detecting key areas and feature of images, such as damage to the adjacent tooth and positioning errors in radiographs.

AI supported assessment improves accuracy and reliability overcoming limitation such as low inter- and intra-examiner reliability. AI with other technologies, such as intraoral cameras provides support to dental procedures assessment, focussing on most challenging criteria. AI can effectively complement human assessors on dental procedures providing feedback and supporting assessments.

AI models still not perfect for dental procedures evaluation. One of the key challenges is the accuracy of these models detecting and classifying intermediate categories in multi-class classification tasks.

Open-source LLMs, GPT-o1, GPT o3-mini, Gemini 2.0 Flash and Grok 3 Think Mode demonstrated variable performance in providing feedback on bitewing positioning errors. The prompt design technique included (1) assigning the LLM a specific role assuming a set of professional competencies (specialist) and (2) re-wording the query using a more formal and structured wording, and to limit the output length. The final optimised prompt did not enhance outputs for the majority of LLMs with reasoning capabilities.

5.5 References

1. Yoon K, Jeong H-M, Kim J-W, Park J-H, Choi J. AI-based dental caries and tooth number detection in intraoral photos: Model development and performance evaluation. Journal of Dentistry. 2024;141:104821.
2. Liu Y, Cheng Y, Song Y, Cai D, Zhang N. Oral screening of dental calculus, gingivitis and dental caries through segmentation on intraoral photographic images using deep learning. BMC Oral Health. 2024;24(1):1287.
3. Kang S, Shon B, Park EY, Jeong S, Kim E-K. Diagnostic accuracy of dental caries detection using ensemble techniques in deep learning with intraoral camera images. PloS one. 2024;19(9):e0310004.
4. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res. 2020;99(7):769-74.
5. Chen W, Dong Y, Ding Z, Shi Y, Zhou Y, Zeng F, et al. RadFabric: Agentic AI System with Reasoning Capability for Radiology. arXiv preprint arXiv:250614142. 2025.
6. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y). 2025;6(6):101260.
7. OpenAI. OpenAI o3-mini: OpenAI; 2025 [Available from: https://openai.com/index/openai-o3-mini/?utm_source=chatgpt.com.
8. Claman D, Sezgin E. Artificial Intelligence in Dental Education: Opportunities and Challenges of Large Language Models and Multimodal Foundation Models. JMIR Med Educ. 2024;10:e52346.
9. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098-102.
10. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. Eur J Dent Educ. 2024;28(4):865-76.
11. Hickel R, Mesinger S, Opdam N, Loomans B, Frankenberger R, Cadenaro M, et al. Revised FDI criteria for evaluating direct and indirect dental restorations-recommendations for its clinical use, interpretation, and reporting. Clin Oral Investig. 2023;27(6):2573-92.

6. Appendices

6.1 Evidence of manuscript submission (Chapter 2 - Adjacent tooth damage)

My Submissions KATHARINA

European Journal of Dental Education

JOURNAL HOME AUTHOR GUIDELINES EDITORIAL CONTACT

Submission Overview

Initial Submission

This submission is under consideration and cannot be edited

Article Type	Original Article
Title	Automated Detection and Classification of Adjacent Tooth Damage Using Deep Learning on Intra-Oral Images: Enhancing Pre-clinical Education in Restorative Dentistry

Manuscript Files

Name	Type of File	Size
Main text.docx	Anonymized Main Document - MS Word	1.2 MB
Title page.docx	Title Page	15.4 KB
Figure 1.png	Figure	4.7 MB
Figure 2.png	Figure	220.1 KB
Figure 3.png	Figure	111.1 KB
Cover letter.docx	Cover letter / Comments	20.3 KB

6.2 Evidence of manuscript submission (Chapter 3 - Positioning errors in bitewing radiographs)

Academic Radiology

Automated Detection of Positioning Errors in Bitewing Radiographs Using Deep Learning

--Manuscript Draft--

Manuscript Number:	
Article Type:	Original Investigation
Keywords:	Deep learning; Convolutional neural network; Dental digital radiography; Bitewing; Imaging errors
Corresponding Author:	Katharina Alves Rabelo The University of Sydney AUSTRALIA
First Author:	Katharina Alves Rabelo
Order of Authors:	Katharina Alves Rabelo
	Zimo Huang
	Dr Eduardo Delamare
	Dr Shwetha Hegde
	Prof Jinman Kim
	Prof Vesna Miletic
Abstract:	Rationale and Objectives: Bitewing radiographs (BWs) are essential diagnostic tools in dentistry but are often affected by positioning errors, compromising their diagnostic value and increasing retake rates. This study aimed to develop and evaluate a deep learning-based automated system for detecting and classifying BW type (premolar/molar), side (right/left) and common positioning errors. Materials and Methods: A total of 403 BWs were collected from dental student assessments acquired from dental radiology mannequins. BWs were consensus-labelled by experts for type (premolar/molar), side (right/left), and positioning errors (cone cutting error (CCE), interproximal overlap, and incorrect receptor placement). A balanced dataset supported training and validation. Convolutional neural network (CNN) architectures, including ResNet and DenseNet variants, were employed. Five-fold cross-validation was used to assess performance based on accuracy, precision, recall, F1 score, and area under the curve. A confusion matrix was generated for multi-class CCE classification. Results: The system achieved high accuracy in classifying BW type (95.8%), side (99.0%), CCE (96.3%), and interproximal overlap (93.4%). Multi-class CCE classification exhibited moderate performance (accuracy: 79.3%), with the model reliably identifying ‘critical’ and ‘minimal’ errors. Accuracy for detecting incorrect receptor placement was 73.2%. Conclusion: The CNN-based system can effectively detect and classify BW type, side, and most positioning errors in mannequin-acquired radiographs. For BW type, side and the positioning errors: CCE (presence and absence) and interproximal overlap (presence and absence), achieving accuracies in the range of 93%-99%. The classifier demonstrated a moderate level of performance for incorrect receptor placement error, with an accuracy of 73.2%.

6.3 Evidence of manuscript submission (Chapter 4 - Feedback on bitewing radiographic technique

My Submissions KATHARINA

European Journal of Dental Education

JOURNAL HOME AUTHOR GUIDELINES EDITORIAL CONTACT

Submission Overview

Initial Submission This submission is under consideration and cannot be edited

Download Reviewer PDF

Article Type	Original Article
Title	AI-Driven Feedback on Bitewing Radiographic Technique: A Comparative Study of Large Language Models

Manuscript Files

Name	Type of File	Size
Main text.docx	Anonymized Main Document - MS Word	412.9 KB
Title page.docx	Title Page	15.5 KB
Figure_1.pdf	Figure	324.7 KB
Figure_2.pdf	Figure	147.8 KB
Figure_3.pdf	Figure	247.1 KB
Supplementary Table 1-Prompts.docx	Supplementary Material for Review	24.8 KB
Supplementary Table 2 - Assessment Rubric.docx	Supplementary Material for Review	18 KB
Supplementary Table 3-Outputs.docx	Supplementary Material for Review	63.6 KB
Cover letter.docx	Cover letter / Comments	20.3 KB

6.4 Human Ethics approval letter (Chapter 2 - Adjacent tooth damage)

THE UNIVERSITY OF SYDNEY

RESEARCH INTEGRITY
& ETHICS ADMINISTRATION

HUMAN RESEARCH ETHICS APPROVAL

The University of Sydney confirms that this project meets the requirements of the National Statement on Ethical Conduct in Human Research.

Project identifier:	2024/HE000118
Project title:	The Use of Artificial Intelligence in Dental Education
Version:	0.01
Chief Investigator:	Vesna Miletic
Authorised project team:	Eduardo Delamare
	Jinman Kim
	Katharina Alves Rabelo
Date of approval:	Thursday, 9 May, 2024
Project end date:	08 May 2028

Provisos *(if applicable)*

Project summary

Recently, artificial intelligence (AI) is revolutionising education. AI is categorized into two primary fields: deductive and generative. In the field of education, the majority of these applications employ deep learning technology, using deductive AI systems. Generative Adversarial Networks (GANs) represent an innovative approach in machine learning, designed to produce synthetic content. GANs have been effectively used in medical education to create case scenarios and instantaneous feedback to students, their application in dental education remains underexplored. This research project aims to develop and validate new AI models by utilising data from students' dental procedures in simulation clinics and from surveys.

Documents approved

Filename	Document type	Document version	Application version
CE_E-mail to CE_v2_29022024_clean version.pdf	Recruitment	Version 2	0.01 - Initial Application
CE_Questioannarie_v1_29022024_clean version.pdf	Survey or questionnaire	Version 1	0.01 - Initial Application
STU_Questioannarie_v1_29022024_clean version.pdf	Survey or questionnaire	Version 1	0.01 - Initial Application
STU_pcf_students_AI_v2_29022024_cleanversion.pdf	Participant Consent Form (PCF)	Version 2	0.01 - Initial Application
20240206_Announcement on Canvas for students.pdf	Recruitment	Version 1	0.01 - Initial Application
20240206_PIS for clinical educators.pdf	Participant information statement (PIS)	Version 1	0.01 - Initial Application
20240206_PCF for clinical educators.pdf	Participant Consent Form (PCF)	Version 1	0.01 - Initial Application
20240206_PIS for students.pdf	Participant information statement (PIS)	Version 1	0.01 - Initial Application

6.5 Survey (Chapter 2 - Adjacent tooth damage)

Confidential

Models_Cavity Preparation Task Assessment: Damage to Adjacent Teeth

This form is a part of the research project which seeks to develop and validate new AI models with the goal of enhancing the learning experience in dental pre-clinical education.

It requires you to assess teeth adjacent to Class II cavity preparations, using your knowledge and experience in evaluating cavity preparations made by dental students in the dental simulation clinic.

By default, all form responses are collected anonymously.

Participant ID	_____
DATE:	_____
Which method will you use to evaluate the following dental models?	◯ Naked eyes ◯ Loupes
MODEL 1	
MODEL 1 #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 1 #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 1 #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 1 #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 2	

MODEL 2 #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 2 #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 2 #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 2 #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 3	
MODEL 3 #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 3 #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 3 #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 3 #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 4	

MODEL 4 #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 4 #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 4 #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL 4 #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
Comments	_____

Pictures_Cavity Preparation Task Assessment: Damage to Adjacent Teeth

This form is a part of the research project which seeks to develop and validate new AI models with the goal of enhancing the learning experience in dental pre-clinical education.

It requires you to assess teeth adjacent to Class II cavity preparations, using your knowledge and experience in evaluating cavity preparations made by dental students in the dental simulation clinic.

By default, all form responses are collected anonymously.

Participant ID _____

DATE: _____

MODEL A - Picture assessment

Please, evaluate the following teeth displayed on the screen:

MODEL A

#15D

Tooth: 15
Surface: Distal

◯ No damage to adjacent tooth
◯ Damaged not requiring restoration
◯ Damaged requiring restoration

MODEL A

#25D

Tooth: 25
Surface: Distal

◯ No damage to adjacent tooth
◯ Damaged not requiring restoration
◯ Damaged requiring restoration

MODEL A

#35D

Tooth: 35
Surface: Distal

◯ No damage to adjacent tooth
◯ Damaged not requiring restoration
◯ Damaged requiring restoration

MODEL A

#45D

Tooth: 45
Surface: Distal

◯ No damage to adjacent tooth
◯ Damaged not requiring restoration
◯ Damaged requiring restoration

MODEL B - Picture assessment

Please, evaluate the following teeth displayed on the screen:

MODEL B #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL B #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL B #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL B #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL C - Picture assessment Please, evaluate the following teeth displayed on the screen:	
MODEL C #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL C #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL C #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL C #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration

MODEL D - Picture assessment

Please, evaluate the following teeth displayed on the screen:

MODEL D #15D Tooth: 15 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL D #25D Tooth: 25 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL D #35D Tooth: 35 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration
MODEL D #45D Tooth: 45 Surface: Distal	◯ No damage to adjacent tooth ◯ Damaged not requiring restoration ◯ Damaged requiring restoration

6.6 Human Ethics approval letter (Chapter 3 - Positioning errors in bitewing radiographs)

THE UNIVERSITY OF SYDNEY

RESEARCH INTEGRITY
& ETHICS ADMINISTRATION

HUMAN RESEARCH ETHICS APPROVAL

The University of Sydney confirms that this project meets the requirements of the National Statement on Ethical Conduct in Human Research.

Project identifier:	2024/HE000846
Project title:	Developing AI-Powered Systems for Enhanced Feedback in Dental Radiology
Application version:	0.02
Chief Investigator:	Dr Katharina Alves Rabelo
Project team:	Dr Eduardo Delamare
	Dr Shwetha Hegde
Project start date:	23 Aug 2024
Project end date:	22 Aug 2028
Date of issue:	Friday, 23 August, 2024

Project summary

This project aims to use existing assessment data obtained during regular assessments in Dental Radiology at Sydney Dental School between 2020 and 2023. The project will analyse radiographs taken and reports written by students, along with the marking dataset used for assessment, based on criteria such as identification of film faults and corrections needed. Dental radiographs, students' reports and tutors' feedback will be collected, to develop an artificial intelligence (AI) model for detecting film faults. Student's reports will be used to train an AI model by comparing their evaluations with tutor's reports and a gold standard assessment made by dentomaxillofacial radiologists who are investigators in this research. The objective is to train the AI model to assess radiographs and provide detailed feedback for clinicians and dental students. The goal is to enhance radiographic techniques for clinicians and students, and to improve the learning experience in radiology sessions at dental schools.

Documents approved

Document type	File name	Document version	Application version
Application Attachment	project-description_AIradio_V1.docx	2	0.02

Conditions of Approval

- Research must be conducted according to the approved proposal.
- An annual progress report must be submitted on or before the anniversary of approval and a final report on completion of the project.
- You must report as soon as practicable anything that might warrant review of ethical approval of the project including:
  - Serious or unexpected adverse events (which should be reported within 72 hours).
  - Unforeseen events that might affect continued ethical acceptability of the project.

6.7 Prompts used in LLMs (Chapter 4 - Feedback on bitewing radiographic technique)

Original Prompts (Prompt A)
1. A left molar bitewing shows: • The area of interest is completely visible • No cone cut • Interproximal overlap present, with palatal and lingual cusps mesial to the buccal cusps. • Centralised occlusal plane Why did it happen and how should I correct it?
2. A right molar bitewing shows: • The area of interest is completely visible • No cone cut • Interproximal overlap present, with palatal and lingual cusps distal to the buccal cusps. • Centralised occlusal plane Why did it happen and how should I correct it?
3. A left premolar bitewing shows: • The area of interest is completely visible • No cone cut • Interproximal overlap present between first premolars and canines, with palatal and lingual cusps mesial to the buccal cusps. • Centralised occlusal plane Why did it happen and how should I correct it?
4. A right premolar bitewing shows: • The area of interest is partially visible • There is cone cut on the mesial mandibular corner affecting the area of interest • No interproximal overlap • Centralised occlusal plane Why did it happen and how should I correct it?
5. A left molar bitewing shows: • The area of interest is partially visible • There is cone cut on the distal side affecting the area of interest • No interproximal overlap • Centralised occlusal plane Why did it happen and how should I correct it?
6. A left premolar bitewing shows: • The area of interest is partially visible. The mesial aspects of 24 and 34 are cut off • No cone cutting • No interproximal overlap • Centralised occlusal plane Why did it happen and how should I correct it?
7. A right molar bitewing shows: • The area of interest is partially visible. The mesial aspect of 16 is cut off • No cone cutting

- No interproximal overlap
- Centralised occlusal plane

Why did it happen and how should I correct it?

8. A right premolar bitewing shows:
- The area of interest is completely visible
- No cone cut
- No interproximal overlap
- Tilted occlusal plane

Why did it happen and how should I correct it?

9. A right molar bitewing shows:
- The area of interest is completely visible
- No cone cut
- No interproximal overlap
- More of the maxillary molars can be seen than the mandibular molars

Why did it happen and how should I correct it?

10. A left premolar bitewing shows:
- The area of interest is completely visible
- No cone cut
- No interproximal overlap
- More of the maxillary premolars can be seen than the mandibular premolars

Why did it happen and how should I correct it?

Optimized Prompts (Prompt B)
1. The following is a dentistry scenario: A left molar bitewing shows: - The area of interest is completely visible - No cone cut - Interproximal overlap present, with palatal and lingual cusps mesial to the buccal cusps. - Centralised occlusal plane Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.
2. The following is another dentistry scenario: A right molar bitewing shows: - The area of interest is completely visible - No cone cut - Interproximal overlap present, with palatal and lingual cusps distal to the buccal cusps - Centralised occlusal plane Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.
3. Now we have a new dentistry scenario: A left premolar bitewing shows: - The area of interest is completely visible - No cone cut - Interproximal overlap present between first premolars and canines, with palatal and lingual cusps mesial to the buccal cusps. - Centralised occlusal plane Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.
4. Now we introduce a dentistry scenario: - A right premolar bitewing shows: - The area of interest is partially visible - There is cone cut on the mesial mandibular corner affecting the area of interest - No interproximal overlap - Centralised occlusal plane Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

5. How about this scenario:

A left molar bitewing shows:
- The area of interest is partially visible
- There is cone cut on the distal side affecting the area of interest
- No interproximal overlap
- Centralised occlusal plane

Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

6. How about this dentistry scenario:

- The area of interest is partially visible. The mesial aspects of 24 and 34 are cut off
- No cone cutting
- No interproximal overlap
- Centralised occlusal plane

Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

7. How about this dentistry scenario:

A right molar bitewing shows:
- The area of interest is partially visible. The mesial aspect of 16 is cut off
- No cone cutting
- No interproximal overlap
- Centralised occlusal plane

Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

8. How about this dentistry scenario:

A right premolar bitewing shows:
- The area of interest is completely visible
- No cone cut
- No interproximal overlap
- Tilted occlusal plane

Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

9. Consider the following dentistry scenario:

A right molar bitewing shows:
- The area of interest is completely visible
- No cone cut
- No interproximal overlap
- More of the maxillary molars can be seen than the mandibular molars

Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

10. Consider the following dentistry scenario:

A left premolar bitewing shows:
- The area of interest is completely visible
- No cone cut
- No interproximal overlap
- More of the maxillary premolars can be seen than the mandibular premolars

Assume that you are now a dental radiologist. Explain the cause of the error and how it could be corrected, controlling the length of your answer but not leaving out essential details.

6.8 Outputs assessment rubric (Chapter 4 - Feedback on bitewing radiographic technique)

Criteria	0 (Poor)	1 (Fair)	2 (Good)	3 (Very Good)	4 (Excellent)
Extension of Detail	Too general (no factors covered; no direction)	Some factors covered	Most factors covered (lacks objectivity)	Most factors covered; objective	All relevant points covered
Content (Accuracy)	At least two inaccuracies or	One major inaccuracy	One minor inaccuracy (e.g., x-	Mostly correct (minor	Completely correct; all key

Criteria	0 (Poor)	1 (Fair)	2 (Good)	3 (Very Good)	4 (Excellent)
	the main factor incorrect		ray beam should be perpendicular to interproximal surfaces)	information missing)	points presented accurately
Clarity	Response not clear at all; assessor very confused	Response obscures most key points	Response obscures many key points	Response mostly clear; some key points unclear	All key points presented clearly
Relevance	At least two irrelevant points	One irrelevant point	No inaccuracies but some information missing	Minor information missing	Completely relevant

6.9 Survey (Chapter 4 - Feedback on bitewing radiographic technique)

Confidential

Positioning Errors in BWs

This form is a part of the research project which seeks to develop and validate new dental AI models.

By default, all form responses are collected anonymously.

Record ID _____

Record ID _____

DATE: _____

This survey consists of 10 Positioning errors scenarions, each with eight proposed Large Language Model responses. We ask that you evaluate these responses based on the following criteria:

We appreciate your time and effort in completing this survey.

Scenario

◯ 1 = Scenario 1
◯ 2 = Scenario 2
◯ 3 =Scenario 3
◯ 4 = Scenario 4
◯ 5 = Scenario 5
◯ 6 =Scenario 6
◯ 7 = Scenario 7
◯ 8 =Scenario 8

Output 1					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	○	○	○	○	○
Content	○	○	○	○	○
Clarity	○	○	○	○	○
Relevance	○	○	○	○	○

Output 2					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯

Output 3					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯

Output 4					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯

Output 5					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯

Output 6					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯

Output 7					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯

Output 8					
	4 = Excellent	3 = Very good	2 = Good	1 = Fair	0 = Poor
Extension of Details	◯	◯	◯	◯	◯
Content	◯	◯	◯	◯	◯
Clarity	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯