Multi-modal Assessment for Dental Diagnosis Assistant: Self-supervised Integration of Radiographic and Clinical Data

Kim, Inseok

Department of Dentistry

Graduate School

Yonsei University

Multi-modal Assessment for Dental Diagnosis Assistant: Self-supervised Integration of Radiographic and Clinical Data

Advisor Park, Wonse

A Dissertation Submitted to the Department of Dentistry and the Committee on Graduate School of Yonsei University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Dental Science

Kim, Inseok

June 2025

Multi-modal Assessment for Dental Diagnosis Assistant: Self-supervised Integration of Radiographic and Clinical Data

This Certifies that the Dissertation of Kim, Inseok is Approved

`

Committee Chair	**Park, Wonse**
Committee Member	**Kim, Kee-Deog**
Committee Member	**Cheong, Jieun**
Committee Member	**Lee, Changmin**
Committee Member	**Shin, Yooseok**

Department of Dentistry
Graduate School
Yonsei University

June 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Multi-modal Assessment for Dental Diagnosis Assistant: Self-supervised Integration of Radiographic and Clinical Data

Introduction

Accurate dental diagnosis is achieved by synthesizing various data including patient history, clinical examinations, and radiographic images. Therefore, actual diagnosis heavily relies on the clinician's experience in synthesizing these data, resulting in variations in diagnostic accuracy. Dental diagnostic assistance utilizing artificial intelligence (AI) learning is expected to contribute to improving this accuracy. While various AI applications are currently emerging in dentistry, research in the diagnostic field is still limited to single-modal learning that uses only radiographic images. This study aims to overcome these limitations by developing a multi-modal AI model that utilizes various types of data necessary for diagnosis through self-supervised learning methods, which are pre-training techniques. The objectives of this study are as follows: first, to develop a single-modal AI diagnostic model using clinical examination data; second, to develop a single-modal AI diagnostic model using periapical radiographic images; and third, to develop a multi-modal AI model combining both models and compare the diagnostic performance among the three models utilizing self-supervised learning techniques.

Methods

For AI model development, 3,341 clinical datasets from 1,344 patients who visited Yonsei Dental Hospital were utilized. Through a screening process, 705 clinical datasets with matching periapical radiographs suitable for training were selected. To develop a single-modal AI diagnostic model using clinical examinations, data were extracted from medical records. The data included categorized patient complaints, gender, and age as basic information, along with seven clinical examinations commonly used for single tooth diagnosis: percussion, mobility, bite, air, cold, hot, and electric pulp test. Self-supervised learning techniques were applied to induce efficient learning, and during the accuracy improvement process, clinical examination types were selected and missing data were imputed with the most frequent values.

To create an AI diagnostic model using periapical radiographic images, 705 periapical radiographs were used. These radiographs corresponded to the diagnostic timepoint of the clinical examinations used in the previous model. To induce efficient learning, a masked autoencoder, which is a self-supervised learning technique for images, was applied. To improve accuracy, lesions and feature points in periapical radiographs were labeled in a detection format, and optimization was

performed to reduce errors.

After maximizing the performance of single-modal AI models using clinical examination data and periapical radiographic images respectively, a multi-modal AI model was constructed by combining the two single-modal models. Subsequently, optimization processes were conducted to reduce overall errors. Model performance was evaluated through target metrics such as accuracy and precision, confusion matrices, and receiver operating characteristic (ROC) curve analysis, and compared through ablation studies that modularized each component.

When developing a multimodal AI diagnostic model, training based on image classification of radiographic data did not fully leverage the advantages of multimodality due to model complexity. In contrast, the detection-based approach demonstrated superior diagnostic performance in the multimodal setting compared to the single modality, particularly in the diagnosis of dental caries, tooth fractures, and pulpitis. Notably, a significant improvement was observed in the diagnosis of tooth fractures.

Conclusion

This study evaluated and compared the diagnostic performance of multi-modal and single-modal approaches in AI-based diagnosis. It also confirmed that the consistency of clinical examination standards, precision of radiographic image labeling, and the quantity and quality of data significantly impact the diagnostic performance of AI models. This research presents the possibilities and limitations of multi-modal approaches in AI-based dental diagnosis and emphasizes the importance of retrospective AI research and the standardization and integration of clinical data. Future research plans to analyze modal data with complementary characteristics according to diagnostic categories and explore the possibilities for improving multi-modal performance in diagnosis.

Key words : Artificial Intelligence, Dental Diagnosis, Radiographic Image, Clinical Examination, Single-modal, Multimodal, Self-Supervised Learning, Masked Autoencoder

1. Introduction

1.1. Advancement of Artificial Intelligence

In recent years, artificial intelligence (AI) has been driving innovation across various industries. In the past, rule-based machine learning and probabilistic models created by humans from data were predominant. However, to create robust models that adaptively respond to the infinite diversity and variability of data, it became necessary to learn the rules themselves by training on large volumes of data, which became the cornerstone for the development of deep learning. The types of data utilized have expanded to include tabular data, numbers, text, and audio data. Subsequently, the emergence of Convolutional Neural Networks (CNNs) brought revolutionary changes to the field of image analysis. By mimicking human visual perception, CNNs can automatically extract and learn features from images, providing much higher accuracy and efficiency than traditional manual image analysis methods, and are now widely used across various fields including manufacturing, quality control, autonomous driving, and the medical industry.

With the emergence of new image analysis models centered on deep learning, accuracy and data processing capabilities continue to improve. These advancements are driving the optimization of model performance, increased learning speed, enhanced data interpretation capabilities, and improved generalization performance, thereby enabling expansion into automation, precision analysis, and predictive systems. In deep learning particularly, various research is being conducted on data processing and augmentation for effective learning, robust neural network models, training techniques such as self-supervised learning, transfer learning, and multi-task learning, as well as optimization techniques like hyperparameter search and the generation of explanatory information.

A representative deep learning model used in recent image analysis is the CNN-based VGGNet model, which demonstrated that neural network depth directly contributes to improved image recognition performance by effectively combining and pooling surrounding information for all pixels using convolution kernels in a bottom-up approach (Simonyan & Zisserman, 2014). The subsequently introduced Vision Transformer (ViT) proposed a top-down approach unlike traditional CNN models, dividing images into multiple patches and learning the relationships between these patches. This showed superior performance compared to existing CNN-based models on large-scale datasets (Dosovitskiy et al., 2020) .

Additionally, various research is actively being conducted to optimize AI model performance and improve generalization. The Decoupled Weight Decay Regularization (AdamW) technique, proposed to solve the generalization problem of the Adam optimizer, improved optimization by separately handling weight decay, thereby proving enhanced model performance (Loshchilov & Hutter, 2017). Furthermore, EfficientNetV2 maximized the efficiency of CNN structures by

introducing new Fused-MBConv operations to reduce model size while increasing learning speed (Tan & Le, 2021). Along with this, ERFNet (Efficient Residual Factorized ConvNet), a model developed for real-time semantic segmentation, successfully reduced computational costs while maintaining accuracy by utilizing residual connections and factorized convolutions (Romera et al., 2017a, 2017b). This improvement in segmentation significantly enhanced identification performance across various images.

Transfer learning and multi-task learning play crucial roles in further improving AI model performance. Transfer learning techniques have proven to significantly enhance CNN-based image classification and object detection performance by applying pre-trained models to new problems (Torrey & Shavlik, 2010). Meanwhile, object detection and data augmentation techniques also serve as important factors in improving AI model performance. For example, the Hybrid Task Cascade (HTC) model showed high performance in image analysis by sequentially combining object detection and mask prediction (Chen et al., 2019), while Fast and Flexible Image Augmentations techniques contributed to improving model generalization performance by applying various data augmentation methods (Buslaev et al., 2020). Multi-task learning (MTL) has shown effective results in improving model generalization performance by simultaneously learning multiple related tasks (Caruana, 1997).

Furthermore, SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique that complements minority class data in imbalanced datasets for tabular data, with reported cases of improved data analysis performance (Chawla et al., 2002). Masked Autoencoders (MAE) are gaining attention as effective unsupervised learning models in image analysis and general image processing by utilizing self-supervised learning methods (He et al., 2022).

Beyond model performance, research on xAI (eXplainable AI) to interpret and enhance the explainability of AI decision processes is also actively progressing. The Layer-CAM (Layer-wise Class Activation Mapping) technique enables more precise object detection than the existing Grad-CAM and presents an improved method to visually explain what CNN-based models are learning from images (Jiang et al., 2021) . In particular, Layer-CAM can analyze activation maps not only in deep layers but also in shallow layers of the network, helping to understand the model's prediction process more intuitively.

1.2. Applications of Artificial intelligence in Medicine

The rapid advancement of image analysis using artificial intelligence (AI) is quickly expanding its applications in the medical field. Particularly, AI has become an essential tool in areas such as early disease diagnosis, medical image analysis, bio signal interpretation, and treatment planning. This transformation has accelerated through both the evolution of deep learning technologies and

the digitization of medical data, inaugurating an era of highly accurate diagnostics and automated analysis.

A prime example is image-based disease analysis. AI has demonstrated various achievements in the field of image-based disease analysis. In pediatric bone age assessment research, CNN-based regression models showed accuracy comparable to expert interpretations, reducing the mean absolute error (MAE) to approximately 5 months, contributing to pediatric growth assessment and orthopedic treatment planning. AI has also excelled in gender and age prediction research. Gender prediction models using Stanford CheXPert and NIH Chest XRay14 datasets, gender determination through spine X-ray analysis using DenseNet (with 99% accuracy for cervical spine and 98% for lumbar spine), and forensic age estimation models using CNN trained on 1,875 pelvic X-ray images from patients aged 10-25 years (recording lower MAE than traditional cubic regression) demonstrate that the fusion of medical big data and artificial intelligence is expanding beyond diagnostic assistance into forensic and personalized medicine fields. (Li et al., 2022; Ren et al., 2018) (Xue et al., 2018) (Li et al., 2019) (Raghu et al., 2021)

Notably, research on gender determination through spine X-ray analysis using the DenseNet model achieved impressive accuracy rates 99% for cervical spine and 98% for lumbar spine—expanding the potential applications in forensic identification and medical AI technology (Xue et al., 2018). Thus, AI-based image analysis is now expanding beyond basic diagnostic assistance, making significant inroads into forensic identification and personalized medicine applications.

With advancements in diagnostic imaging assistance, multi-modal data integration approaches are gaining increasing attention in AI-based medical analysis, transcending single-modal data analysis. Multi-modal AI models that integrate and analyze different types of medical data demonstrate higher accuracy and reliability than conventional single data-based models. This methodology proves particularly valuable for disease diagnosis and treatment planning that require complex and precise analysis.

A significant example is research on early Alzheimer's disease diagnosis, which recorded enhanced diagnostic accuracy by integrating diverse bio-signal data including MRI, EEG (electroencephalogram), and PET (positron emission tomography). This research highlighted the possibility of non-invasively monitoring patients' conditions continuously and validated the effectiveness of multi-modal approaches in diagnosing neurodegenerative diseases (Alberdi et al., 2016).

Furthermore, in osteoporosis prediction research, a multi-modal AI model combining MRI and CT data was developed, achieving a high accuracy of 98.90% and enhancing the reliability of quantitative diagnosis compared to existing methods (Küçükçiloğlu et al., 2024). This represents an important advancement in overcoming the limitations of traditional methods that rely on single imaging techniques, introducing a new diagnostic paradigm that utilizes multi-modal data.

AI-based analysis utilizing chest X-rays is also advancing rapidly. The BIO-CXRNET model, developed to predict mortality risk in COVID-19 patients, demonstrated high accuracy at 89.03%, validating the strengths of multi-modal analysis (Rahman et al., 2023).

Multimodal deep learning has been actively applied across various medical fields. In ophthalmology, a model combining fundus photographs and OCT images has improved the diagnostic accuracy for retinal diseases (He et al., 2021). In the field of neuropsychiatry, multimodal approaches have also shown superior performance compared to unimodal models (Wang et al., 2022). Additionally, for thoracic disease diagnosis, a model integrating chest X-ray images and cough sound data achieved high recognition accuracy (Kumar et al., 2022), demonstrating the potential effectiveness of multimodal learning for early detection and accurate classification of diverse diseases.

AI models that combine multi-modal data offer significant potential beyond basic disease diagnosis, extending to treatment planning, disease progression prediction, and personalized medicine. Additionally, multi-modal AI-based medical analysis technology is expected to play a crucial role in predicting treatment outcomes and developing personalized treatment plans. Current research is progressing toward integrating not only medical images (X-ray, MRI, CT, PET, etc.) but also genetic data, bio-signal data, and patient history information potentially becoming a cornerstone of precision medicine and personalized treatment.

In conclusion, the evolution of AI-based multi-modal fusion techniques is anticipated to further enhance the accuracy and reliability of medical image analysis and disease diagnosis. As medical AI continues to advance, it promises to deliver increasingly innovative and sophisticated diagnostic and treatment solutions in clinical settings.

1.3. Applications of Artificial intelligence in Dentistry

1.3.1 Background of AI Growth in Dentistry

Interest in artificial intelligence (AI) in dentistry has been rising more steeply than in any other field over the past five years. This spread of interest and research stems from the unique environment and conditions of dentistry. There are three main reasons why dentistry is particularly suitable for AI application.

First, dental imaging has a high level of digitization. Dental clinics have adopted digital radiography equipment relatively early, and in most clinical settings, panoramic images, periapical images, CBCT, and oral scans are all stored and managed in digital format. This richly accumulated digital imaging data can be utilized as high-quality learning material necessary for AI training, providing a favorable environment for effectively training artificial intelligence models. Considering

that 'sufficient amount of structured data' is the condition where AI develops fastest in the medical imaging field, dentistry can be considered a highly suitable field for AI applications.

Second, the relative ease and clarity of image interpretation is also a strength of dentistry. Anatomical structures of the jaw and oral cavity, dental caries, alveolar bone loss due to periodontal disease, and lesions such as cysts or tumors appear relatively clearly in periapical radiographs. These lesions have specific patterns, making them easier for image-based AI to identify.

Third, the anatomical features and structures of teeth have relatively little variation between individuals, which is also a favorable condition for AI application. While there are individual differences in the size, arrangement, and shape of teeth, they often show standardized structures in terms of position, shape, and symmetrical structure compared to other organs or anatomical structures. This allows for high performance even with less data and can contribute to producing stable and consistent results in actual clinical applications.

Based on these characteristics, AI research that automates diagnostic and analytical tasks is actively being conducted in the field of dentistry, with many studies showing higher accuracy and efficiency than conventional methods.

1.3.2 Application Areas of AI in Dentistry

Currently, AI technology is being applied to various areas in dentistry and showing remarkable achievements in each area. The main application areas can be broadly divided into tooth detection and segmentation, dental caries and disease detection, dental prosthesis and implant analysis, gender and age estimation, and diagnosis.

AI technology shows high accuracy in tooth detection and segmentation tasks, which are fundamental to dental diagnosis. YOLOv7-based CNN models recorded F1-scores of 0.99 and 0.979 for tooth detection and numbering in bitewing radiographs, respectively (Ayhan et al., 2024), while automated tooth segmentation methods using Mask R-CNN in panoramic radiographs achieved an F1-score of 87.5% and an IoU of 87.7% (Lee et al., 2020). Additionally, algorithms for automatically assigning tooth numbers in panoramic radiographs (Karaoglu et al., 2023) and Panoptic segmentation techniques for automatically segmenting various structures such as maxillary sinuses and mandibular canals have been developed(Cha et al., 2021).

In the area of dental caries and periodontal disease detection, AI models detecting periodontal bone loss show performance similar to that of experts (Krois et al., 2019), and U-Net-based deep learning models that automatically segment tooth features in periapical radiographs recorded 82% sensitivity and precision in dental caries detection (Khan et al., 2021). Furthermore, DenseNet121-

based classification models for oral cancer detection achieved 99% precision and 100% recall (Warin et al., 2021).

In research developing AI models to automatically detect dental prostheses, the Faster R-CNN RegNetX model recorded the highest performance with 97.3% mAP and 77.1% AR (Çelik & Çelik, 2022). Additionally, in implant analysis and automatic classification system research, deep learning methods using the VGG16 model recorded an AUC-ROC value of 0.975, accurately classifying implant sizes (Park et al., 2023).

CNN-based deep learning models for age estimation using panoramic dental X-ray images recorded a mean absolute error (MAE) of 2.95 years, achieving higher accuracy than conventional manual methods (Milošević et al., 2022), and a multi-task learning based AI model called ForensicNet recorded an MAE of 2.93 years for age prediction and 99.2% accuracy for gender classification (Park et al., 2024). These technologies show great potential as forensic identification and pediatric dental diagnostic tools.

The application of AI technology in the field of dental diagnosis has shown particularly noteworthy achievements. The Faster R-CNN Inception v2 model, which automatically diagnoses dental conditions in panoramic radiographs, showed high performance in specific diagnoses such as implants (sensitivity 96.15%) and crowns (sensitivity 96.74%), but relatively low performance in detecting dental caries (sensitivity 30.26%) and calculus (sensitivity 9.34%) (Başaran et al., 2022). AI models for diagnosing temporomandibular joint osteoarthritis (TMJOA) using panoramic radiographs showed expert-level performance with a sensitivity of 73% and specificity of 82% (Choi et al., 2021), and CNN-based AI models for detecting periapical lesions achieved higher diagnostic performance than 14 out of 24 oral and maxillofacial surgeons (Endres et al., 2020).

Despite these achievements, current AI research in dental diagnosis shows several limitations. The most prominent limitation is that most models are based on a single modality, particularly radiographic images. In actual clinical settings, accurate diagnosis is made by comprehensively considering various information such as patient symptoms, clinical examination results, and medical records, not just radiographic images. Models based on single modalities demonstrate inherent limitations in capturing the complexity of clinical situations, resulting in notable performance decline particularly for diagnoses that require rich clinical context, such as dental caries or calculus detection.

Additionally, the size and heterogeneity of datasets are also cited as major limitations (Albano et al., 2024). Most current research uses limited-scale data, and issues with research method heterogeneity and reporting quality make it difficult to compare between studies (Mohammad-Rahimi et al., 2022). Several studies commonly emphasize that data standardization and dataset expansion are necessary to overcome these limitations.

To address these inherent limitations and develop dental diagnostic AI models with enhanced accuracy and reliability, we propose a multimodal approach that integrates diverse data sources. By combining complementary data types, this approach offers richer information and contextual understanding than single-modality methods, leveraging the unique strengths of each modality while mitigating their individual weaknesses

In dental diagnosis, the multimodal approach can be implemented in the form of combining radiographic image data and clinical examination data. Clinical examination data includes the patient's symptoms and direct clinical observation results such as percussion, mobility, cold and hot sensation, and electric pulp tests, providing important clinical information that is difficult to identify in radiographic images. The combination of this clinical examination data and radiographic image data is expected to enable more accurate and comprehensive diagnosis.

However, systematic integration of clinical examination data and radiographic image data in the field of dental diagnosis is still limited, and research applying the latest AI techniques such as self-supervised learning is even more scarce.

1.3.3 Scope and Objectives of the Study

The purpose of this research is to develop a multimodal artificial intelligence diagnostic model that combines clinical examination data and periapical radiographic images using self-supervised learning techniques, and to compare and evaluate its performance with single-modal models. Based on the background previously discussed, this study aims to develop single-modal and multimodal artificial intelligence models to improve dental diagnostic accuracy by utilizing the latest self-supervised learning techniques and Masked Autoencoder.

The specific research objectives are as follows: 1) Develop a single-modal artificial intelligence diagnostic model using clinical examination data, questionnaire results, and patient history to derive diagnostic information that is difficult to obtain from imaging data.

2) Develop a single-modal artificial intelligence diagnostic model specialized in lesion detection and dental structure analysis using dental radiographic images to complement the limitations of image-based diagnosis.

3) Develop a multimodal artificial intelligence model by combining these two single-modal models and evaluate the impact of integrated data utilization on diagnostic performance. Through this, we aim to clearly demonstrate the advantages of multimodal fusion and present its applicability in actual clinical diagnostic processes.

This research aims to present a new methodology that maximizes the advantages of each modality while overcoming the limitations of single-modal approaches. In particular, by applying

self-supervised learning techniques, effective learning is possible even with limited labeled data, and we aim to explore the complementary relationship between clinical examination data and imaging data.

By developing more accurate and reliable dental diagnostic AI models through this study, we ultimately aim to contribute to improving the diagnostic accuracy and work efficiency of dentists. Additionally, the methodology of this research is expected to provide a foundation that can be extensively applied to the diagnosis of various dental diseases in the future.

This study is expected to support clinicians in utilizing more accurate and reliable diagnostic tools, ultimately contributing to improving the quality of dental care and enhancing patient safety. In the following sections, 2. MATERIALS AND METHODS will demonstrate the stages and methods of data collection and refinement for AI learning, and the design process for diagnostic models based on two different types of data: radiographic images and clinical examinations, ultimately explaining the learning process of two single models and one multimodal model. 3. RESULTS will compare the performance of the single AI diagnostic model using radiographic images with the single AI diagnostic model using clinical examinations and present the performance results of the multimodal model that integrates these two approaches. 4. DISCUSSION will address the challenges and iterative improvements encountered during the model development process and outline future development directions, and finally, 5. CONCLUSION will present the concluding remarks.

2. MATERIALS AND METHODS

2.1. Data Processing

2.1.1 Dataset Design

This study utilizes medical records and periapical radiographic images of patients who visited the hospital with tooth-related diseases. The research aims to develop an advanced artificial intelligence model that can achieve accurate diagnoses by learning from patient symptoms recorded in medical charts, clinical tests conducted to diagnose the affected teeth, and periapical radiographic images used to identify lesions.

2.1.2 Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of Yonsei Dental Hospital, Yonsei University (IRB approval number: 4-2018-0561). All data were anonymized to protect participant confidentiality.

2.1.3 Data Collection

The data used in this study targets 1,344 patients who visited the Department of Advanced General Dentistry at Yonsei Dental Hospital from 2013 to 2018. These patients were extracted based on the diagnostic codes of dental caries, tooth fracture, and pulpitis, utilizing a total of 3,341 patient records. The number of patients by diagnosis is summarized in **Table 1**. The collected data can be broadly divided into clinical data and imaging data, which underwent stage-by-stage processing to ensure successful research through research ethics compliance and effective data management. The patient screening process can be seen in **Figure 1**.

2.1.4 Data Preprocessing

Data preprocessing is divided into two processes: transforming patient records and clinical test records (tabular) from electronic medical records, and periapical radiographic images into tensor formats that can be processed by artificial intelligence models.

Table 1. Number of dental clinical records extracted with primary diagnoses of dental caries, pulpitis, and tooth fracture.

Focus Group	Classification	Characteristics	Number of dental records (*n*)
Group 1	K.02	Caries	362
Group 2	K.04	Pulpitis	1914
Group 3	S.2	Fracture	1065

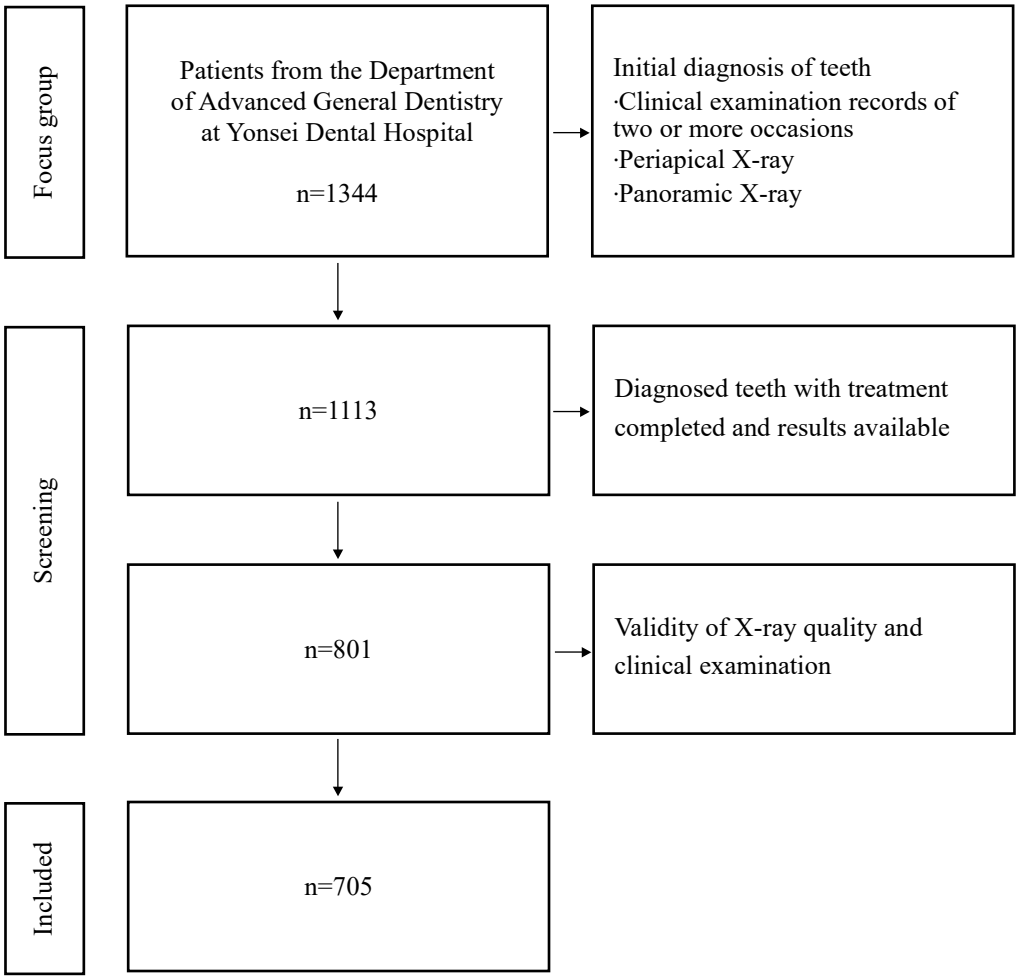

Figure 1. Flow diagram of patient selection process.

Patients who visited the Department of Advanced General Dentistry at Yonsei Dental Hospital(2013-2018) were initially screened based on having at least two clinical examination records at initial diagnosis and both periapical and panoramic X-rays. After additional screening based on teeth with clearly verified treatment results and qualitative validity of X-rays and clinical examination records, a total of 705 patients were ultimately included in the study.

To utilize data from electronic medical records, patient records and clinical test records were transformed into Excel format. The extracted data were stored with anonymized numbers and limited to basic data such as age, gender, chief complaint (C.C.), and clinical tests conducted on the tooth targeted by the C.C.

In the clinical examinations, seven fundamental tests were conducted to assess pulp vitality and sensitivity, as well as the condition of the periodontal ligament and alveolar bone. The percussion test evaluates the response of the periodontal ligament to mechanical stimulation. Mobility is assessed by manually checking the movement of the tooth to evaluate the integrity of periodontal support structures. The bite test applies functional force to detect localized pain or cracks in the tooth structure. The air stimulus test evaluates dentin hypersensitivity by applying a blast of cold air, while the cold test assesses the response of pulpal sensory nerves using thermal stimuli. The heat test helps identify irreversible pulpitis through abnormal or prolonged pain responses. Lastly, the electric pulp test (EPT) determines the vitality of the pulp by evaluating sensory nerve response to electrical stimulation. Each of these tests was recorded using a five-level scale: +++, ++, +, −, and “no test,” as referenced in (Mainkar & Kim, 2018). The extraction process from the patient records is illustrated in **Figure 2.**

Periapical radiographic image data were initially acquired in DICOM file format, which contains subject information. For anonymization purposes, only the image information was extracted and saved as PNG files without information loss. Simultaneously, these were checked against panoramic radiograph to verify that there were no errors in the periapical radiographic images. During anonymization, each dataset was assigned a new number for file naming.

All patient names were anonymized, and to connect the two types of data, arbitrary numeric codes (auto-increment) were assigned to each subject using the numbers given instead of names as indices, linking them to the corresponding image filenames. The analysis of the stored data is shown in **Table 2**.

Subsequently, all personal information, other information unrelated to this study, and DICOM files containing personal information were deleted for conducting the experiment.

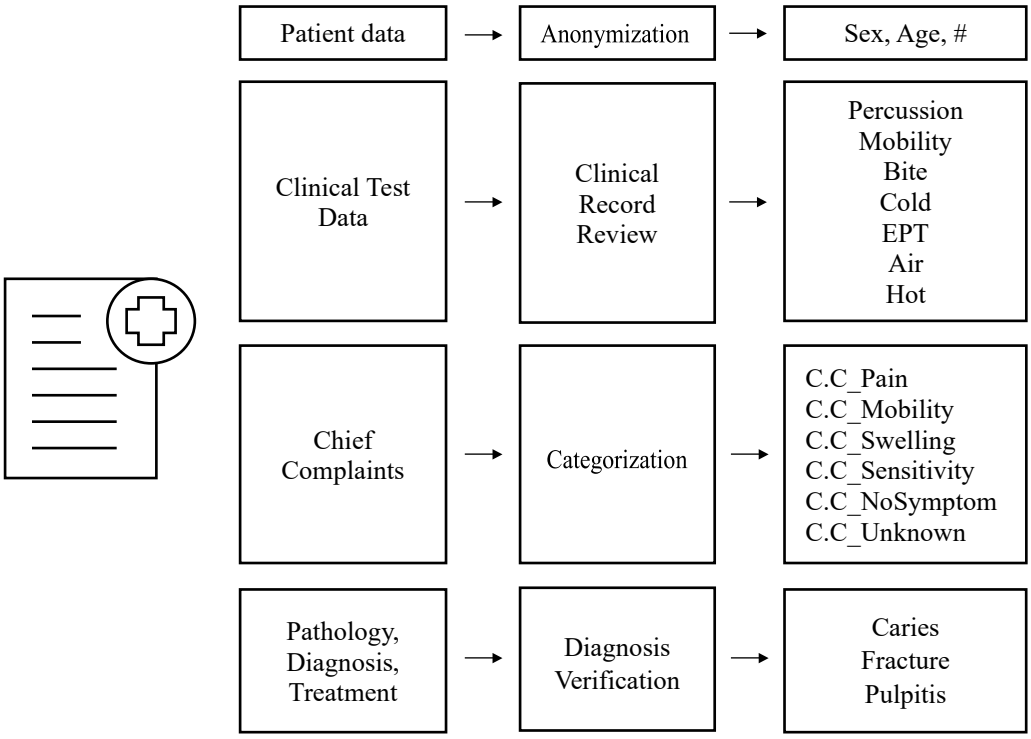

Figure 2. Data extraction and processing workflow from electronic dental records

This figure illustrates the systematic extraction and categorization of patient data from electronic dental records. Patient demographics are anonymized, clinical test parameters are standardized into diagnostic groups, chief complaints are categorized into six symptom types, and diagnostic information is processed to verify physician diagnoses across three dental pathologies: caries, fracture, and pulpitis. This structured workflow enables creation of labeled datasets for multimodal AI diagnostic system training.

Table 2. Number and data types of extracted clinical information from dental records.

Clinical info	**Caries**	**Fracture**	**Pulpitis**	**TOTAL**
N	244	282	179	705
SEX				
Male	105	133	97	335
Female	139	149	85	370
Age				
M	48.7	53.7	56.0	52.6
Percussion	213(138)	258(89)	167(76)	
Mobility	190(155)	243(125)	150(103)	
Bite	46(21)	57(16)	44(11)	
Cold	174(40)	131(71)	105(27)	
EPT	33(11)	45(29)	16(11)	
Air	22(8)	8(4)	16(5)	
Hot	1(0)	4(1)	0	

clinical data exist(negative)

2.2. Single-Modal Model Based on Clinical Test Data

2.2.1 Variables Processing

The clinical data stored for this study consists of categorical, numerical, and ordinal variables, each requiring appropriate preprocessing methods to convert them into a format suitable for artificial intelligence learning. This process includes various procedures such as handling missing values, encoding, and normalization.

Categorical Variables In this study, categorical variables included 'gender_m' (gender), 'CC1_Gum Swelling' (presence of gum swelling), 'CC1_Pain' (presence of pain), 'CC1_Sensitivity' (presence of sensitivity), and 'CC1_No Symptoms' (absence of symptoms). As these are nominal data without concepts of size or order, missing values were replaced with the mode, and One-Hot Encoding was then applied to convert them into binary vector format. This helps the model perform operations between variables and is effective in maintaining the unique characteristics of each category.

Numerical Variables The numerical variable 'age' showed a skewed distribution rather than a normal distribution. Accordingly, a Robust Scaler was applied to normalize based on the median and quartiles to reduce the influence of outliers. Additionally, Power Transformer and Quantile Transformer were also comparatively analyzed, but as they did not show significant performance differences in actual experimental results, the Robust Scaler was ultimately adopted.

Ordinal Variables Clinical examination items such as 'Percussion', 'Mobility', 'Cold', 'EPT', and 'Air' were treated as ordinal variables with a clear order. Like categorical variables, missing values were replaced with the mode, and Ordinal Encoding was used to convert them into numerical values while maintaining order information. Subsequently, a Min-Max Scaler was applied to normalize values between 0 and 1. For variables showing asymmetric distributions, a Quantile Transformer was used supplementarily.

Through these preprocessing strategies, the unique characteristics of variables were preserved while effectively utilizing them for model learning. In particular, scaling and encoding methods appropriate to variable types played an important role in improving model performance and ensuring consistency with clinical interpretation.

2.2.2 Feature Processing

This study selected optimal preprocessing methods considering the statistical characteristics and data distributions according to various variable types in clinical data. Comparative experiments were

conducted on various methodologies, and appropriate strategies were adopted based on model performance and processing efficiency.

First, for missing value imputation, KNN Imputer and Iterative Imputer (MICE) were compared. KNN Imputer has the advantage of simple computation by using the average of adjacent samples to replace missing values but is sensitive to outliers. On the other hand, Iterative Imputer allows for more sophisticated replacement by performing repeated predictions based on regression but requires more computation and may assume multivariate normality. Both methods did not show significant differences in actual model performance, and the KNN Imputer was ultimately chosen considering efficiency and simplicity.

In comparing data scaling methods, Min-Max Scaler, Standard Scaler, and Robust Scaler were tested. Min-Max Scaler normalizes all data between 0 and 1, which is intuitive but very sensitive to extreme values. Standard Scaler adjusts the mean to 0 and standard deviation to 1, making it suitable for normal distributions, but also vulnerable to outliers. In contrast, Robust Scaler has the advantage of minimizing the influence of outliers by basing on the median and quartiles. In this study, the Robust Scaler, which is robust against outliers, was finally selected for 'age,' the only numerical variable.

For normalizing the distribution of numerical variables, PowerTransformer and QuantileTransformer were compared. PowerTransformer is a technique that transforms non-normal distributions closer to normal distributions through Yeo-Johnson or Box-Cox methods, with the advantage of maintaining linearity between variables. QuantileTransformer is a quantile normalization method based on cumulative distribution, which can be particularly useful for variables with extremely large skewness. In this study, the QuantileTransformer was selectively and supplementarily used in parallel, considering the distribution characteristics of each variable.

These preprocessing strategies based on variable characteristics played an important role not only in stability and performance improvement of model learning but also in ensuring consistency and reproducibility of clinical interpretation. Furthermore, the preprocessing design of this study can be said to consider both statistical reliability and practicality, as it verified the validity of the adopted methods through comparative experiments between various approaches.

2.2.3 Data Filtering

In this study, a data filtering process was performed on a total of eight clinical examination items (Percussion, Mobility, Bite, Air, Cold, Hot, EPT, CC). This process aimed to ensure data reliability and quality, and to increase analytical precision and efficiency by removing variables that were unnecessary for model learning or difficult to interpret.

Among the clinical items, the Bite Test was distinguished simply as '+' or '-', making it difficult to view as an ordinal variable, and limited in terms of meaningful information from an analytical perspective. Accordingly, such variables were considered categorical variables and ultimately excluded from the analysis. Additionally, the Hot test, which had an extremely low frequency of occurrence at only 2 out of 776 cases in the entire dataset, was also excluded as it was determined not to make a substantial contribution to learning.

In the case of the CC (Chief Complaint) item, due to the characteristics of university hospital electronic records, there were difficulties in data analysis using the original sentences as they contained referral notes and administrative phrases. Therefore, while a meaning-based approach was attempted, there were limitations in natural language processing-based interpretation due to the characteristics of unstructured sentences and diversity of expressions. Consequently, in this study, the expressions frequently used by patients were refined and simplified into five representative categories (pain, mobility, gum swelling, hypersensitivity, no symptoms), and the CC items were categorized based on these expressions (Brunsvold et al., 1999).

This data filtering procedure was a strategic measure to more clearly define the data to be used for learning, reduce noise, and simultaneously ensure clinical validity and interpretability. As a result, a refined dataset that could enhance the stability and generalization performance of the model was constructed.

2.2.4 Imbalance Handling

During the learning and validation phases, appropriate cross-validation methods and sampling strategies were applied considering the class imbalance problem in the data. The schematic diagram of this process is presented in **Figure 3**. In particular, since the classification problem for patient conditions had an imbalanced distribution among classes, Stratified Cross Validation was used instead of simple K-fold cross-validation. This allowed for maintaining balanced class ratios in each fold, reducing bias in evaluation metrics and enabling more consistent model performance measurement.

In addition, to complement the learning performance of minority classes in a situation where the data size itself was not large, oversampling using SMOTE (Synthetic Minority Oversampling Technique) was performed. SMOTE is a technique that generates new synthetic samples based on existing minority class data, and unlike simple replication, it has the effect of increasing data diversity while preserving the spatial characteristics of surrounding data. This approach created a balanced learning environment that enabled more equitable representation of all classes within the model.

Furthermore, data augmentation was attempted by applying Gaussian Noise Augmentation. This was a strategy to improve generalization performance by randomly adding small noise to the input

values of numerical variables, preventing the model from reacting excessively to minor changes in input values. Considering that measurement deviations can exist in actual clinical data, this noise-based augmentation had validity in reflecting realistic input conditions.

2.2.5 Self-supervised learning

The influence of data quantity in artificial intelligence learning is absolute and acts as a key factor in the generalization ability and performance improvement of models. However, in the medical field, especially in the case of dental clinical data, there are practical constraints in securing large-scale, high-quality data due to high labeling costs and sensitivity in terms of personal information protection. This frequently leads to problems such as data imbalance, lack of labels, and limitations in sample size, and the potential application of Self-Supervised Learning (SSL) is gaining attention as a way to overcome these limitations.

Self-Supervised Learning is a learning method that learns potential representations through pretraining from data without explicit labels, and based on this, produces excellent performance with only a small amount of label information for downstream tasks. SSL primarily sets up pretext tasks and proceeds with learning by generating supervision from the data itself, such as masking and restoring parts of an image, matching temporal order, or comparing similar/dissimilar pairs.

The biggest advantage of these SSL techniques is the reduction of labeling costs and mitigation of imbalanced data problems. In medical settings, pathological state data is often significantly less compared to normal data, and SSL provides the possibility to increase learning efficiency for such rare classes and maximize performance with a relatively small amount of annotated data.

Furthermore, SSL has high complementarity with Multimodal Learning in that it is suitable for processing various forms of data together. In integrating data located in different representation spaces such as radiographs, clinical tabular data, or text-based medical history information, SSL is evaluated as an effective approach for constructing integrated representations by independently learning the expressiveness of each modality and then combining or making them interact.

Therefore, in this study, in the process of developing an artificial intelligence diagnostic model based on dental clinical data, Self-Supervised Learning methods were introduced as a core strategy to solve the problems of limited label numbers and imbalanced data structure, and to effectively implement a multimodal structure.

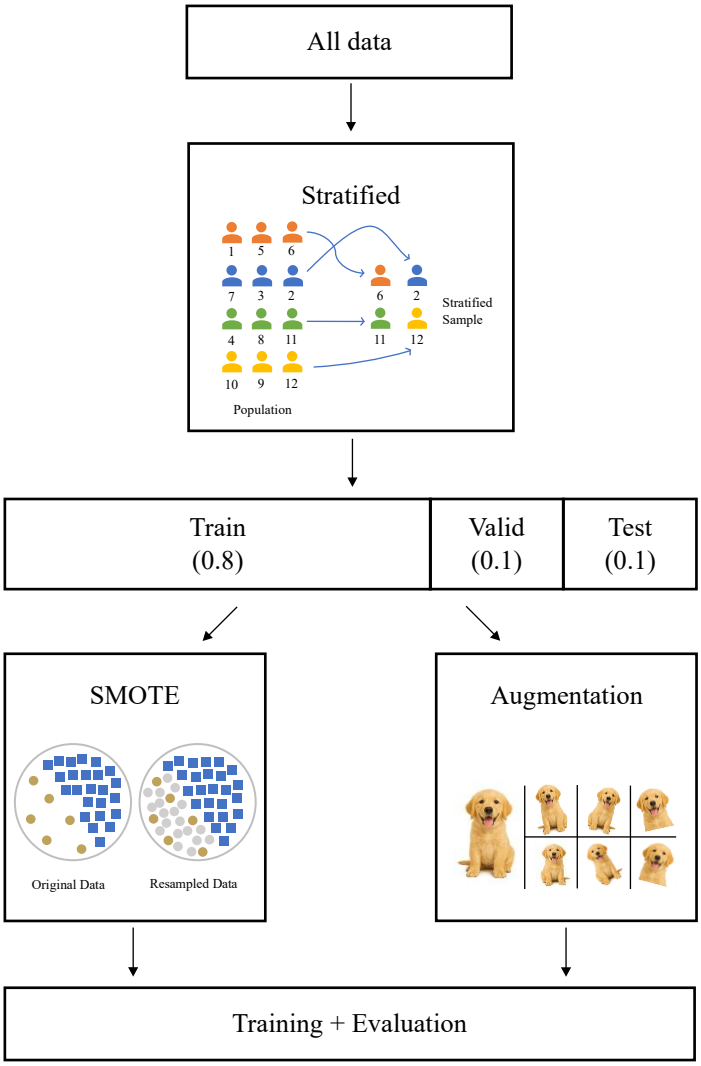

Figure 3. Overview of the data preprocessing and training pipeline.
All data were stratified and split into training (80%), validation (10%), and test (10%) sets. Clinical test data were processed using SMOTE to address class imbalance, while periapical radiographic images underwent data augmentation. Both modalities were used for training and evaluation of the diagnostic models.

2.2.6 Clinical Test Based Deep Learning Model

When starting the research, the problem was initially perceived as distinguishing only one specific lesion from each patient's dataset. However, based on the actual presence of lesions in patients, the problem required applying independent Sigmoid functions to each class node rather than Softmax. Therefore, this research is defined as a multi-label classification problem considering the possibility of simultaneous expression of multiple lesions, rather than a multi-class classification. Consequently, considering cases where two or more lesions could appear simultaneously in a single image or clinical data sample, the model was designed to predict independent probability values for each lesion class in the output layer.

For the actual learning, we utilized eight types of clinical examination data: Percussion, Mobility, Bite (occlusion test), Air (air stimulation), Cold (cold stimulation), Hot (heat stimulation), EPT (electric pulp test), and C.C. (categorized subjective symptoms). For this purpose, in the MLP-based output layer, each node predicts the presence of the corresponding lesion as a probability value, which can be binarized independently as 0 or 1 for all classes.

To ensure stable learning of the model, we applied Batch Normalization (BN) to adjust the data distribution and improve the learning speed of the neural network. The model structure consisted of input layers, 2-3 hidden layers, and as many output nodes as the number of classes. In the hidden layers, nonlinear activation functions such as ReLU and dropout were applied to prevent overfitting. As shown in **Figure 4**, according to the architecture of the designed neural network, the first transformation of the network was applied using the structure: Linear (17 → 1024) - Batch Normalization (BN) - LeakyReLU activation function. Subsequently, the structure of Linear (1024 → 512) - BN - LeakyReLU was repeatedly applied to progressively enhance the feature learning capability of the network. The initial weights of the neural network were set using the He initialization method.

In this study, to maximize the effect of representation learning and improve classification performance with limited supervised learning data, we trained the model by combining Autoencoder-based unsupervised pretraining and transfer learning strategies. In the initial stage, we designed an Autoencoder structure to restore input data, and through this, trained the encoder part to effectively extract latent representations of the input space. The Autoencoder consists of an encoder that compresses input data and a decoder that restores it to its original form. In this study, we applied a multilayer perceptron (MLP) structure to the encoder.

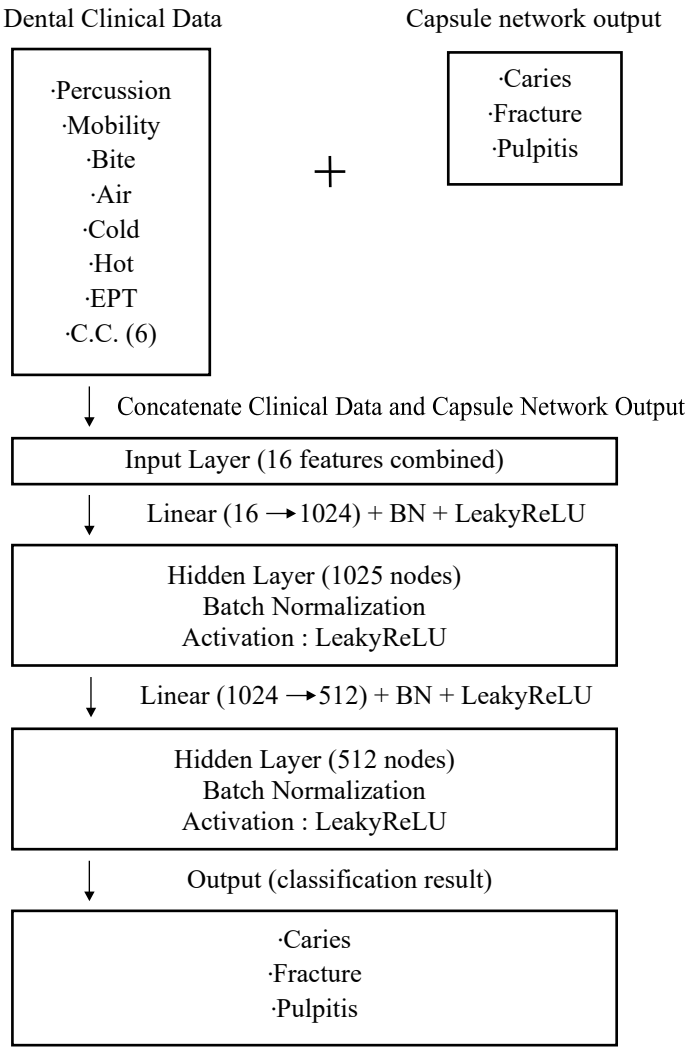

Figure 4. Architecture of clinical data and capsule network outputs.

In this multi-binary classification structure, the loss function was also applied as an extension of Binary Cross Entropy. Initially, BCEWithLogitsLoss (including sigmoid) was used to calculate the binary cross-entropy between the predicted probability and the correct answer for each class. Later, Focal Loss was applied to more effectively correct class imbalances. In a multi-label structure, Focal Loss calculates losses individually for each class and provides the effect of focusing on lesion classes that are not well predicted in the early stages of learning by assigning higher weights to predictions with high difficulty.

Autoencoder training was conducted in an unsupervised manner, using Mean Squared Error (MSE) as the reconstruction loss to minimize the difference between input and restoration. Through this process, the encoder was trained to have the ability to summarize only the essential characteristics from the original input, which provided a favorable foundation for generating generalized representations suitable for subsequent classification tasks.

The model concatenates clinical examination features with capsule network outputs derived from radiographic images. The fused input passes through multiple fully connected layers with Batch Normalization and LeakyReLU activations. The model outputs independent predictions for dental caries, tooth fracture, and pulpitis using a multi-label classification structure. essential characteristics from the original input, which provided a favorable foundation for generating generalized representations suitable for subsequent classification tasks.

After completing Autoencoder training, we separated the decoder and extracted only the pre-trained encoder to transfer it as a feature extractor for the classification model. This transferred encoder showed faster convergence and higher initial performance than randomly initializing weights and could more effectively reflect the structural relationships between complex clinical and image-based variables. The classifier consisted of new fully connected layers connected after this encoder, and in the supervised fine-tuning stage, the entire network was retrained in an end-to-end manner.

Finally, we utilized a Capsule Network (CapsNet) based model to train the distinction of three diseases: Dental Caries, Tooth fracture, Pulpitis,. In this process, the Capsule Network is designed to learn spatial relationships better than traditional CNNs, enabling precise diagnosis based on clinical examination data. Through this process, we effectively extracted patterns from clinical data and enabled the model to learn key information necessary for dental diagnosis.

Each output node generates independent probability outputs through sigmoid functions, which are interpreted as the probability of presence for each lesion. Training was performed based on the AdamW optimization algorithm, applying warm-up (starting with a very low learning rate and gradually increasing it) and a scheduler that reduces the learning rate under certain conditions.

Furthermore, to evaluate the impact of data augmentation on model performance with tabular data, in addition to data oversampling techniques such as SMOTE, we analyzed the performance differences according to the application of Gaussian Noise. The results showed slight performance improvements for classes with fewer data. However, overall performance changes were minimal, which seems to be due to the stability of the model structure and variables used. Through these learning methods, we optimized the model to comprehensively analyze various clinical examination results and classify dental diseases with high accuracy. To enhance the reproducibility and reliability of the model, each experiment was repeated multiple times under identical conditions

2.3. Single-Modal Model Based on periapical radiographic images

2.3.1 Radiological Image Processing

The initially stored radiographic image data underwent various preprocessing and augmentation techniques to improve image quality and ensure data diversity.

In this study, we primarily used PA (periapical) images converted to PNG format. To minimize the impact of image quality on analysis results, we implemented a procedure to quantify image sharpness through contrast and density adjustments and filter out low-quality images in advance based on these measurements..

The degree of blur in an image can generally be determined by the extent of high-frequency component loss. In this research, we quantified the sharpness of each image using the Variance of Laplacian based on the Laplacian operator. This method is based on edge information in images; sharper images have higher Laplacian variance values, while blurrier images have lower values. After calculating the sharpness index for all images, we removed images in the bottom 1% of the distribution from the dataset. This measure was taken to filter out quality-degraded images caused by camera shake or focus deviation during capture, thereby minimizing noise that could negatively impact learning. This process was performed according to quantitative criteria and effectively contributed to removing samples that were likely unable to be analyzed clinically.

This image quality-based filtering played a crucial role in ensuring the reliability and consistency of data that the model would learn, without drastically reducing the overall amount of data. Furthermore, it allowed the subsequent data augmentation and normalization processes to be reflected without distortion, ultimately contributing to improved model performance stability and interpretability.

Subsequently, we configured a preprocessing pipeline using the Albumentations library and torchvision to enhance generalization performance necessary for model learning while preserving image features.

The entire pipeline is structured in the following order. First, we applied CLAHE (Contrast Limited Adaptive Histogram Equalization), a local contrast correction technique for enhancing image contrast. CLAHE equalizes histograms within a limited range in each part of the image, enhancing local contrast without distorting the overall brightness distribution.

Afterwards, images were resized to a fixed resolution to match the model input size, and in case some images were too small, padding was performed through PadIfNeeded to ensure a minimum size (900x900 pixels). This measure was taken to maintain consistent input forms even with images of various sizes.

This series of preprocessing steps contributed to maximizing the model's generalization performance while maintaining image quality and enhancing the reliability and interpretability of the overall analysis results.

2.3.2 Data Augmentation

Various data augmentation techniques were applied to the preprocessing pipeline to improve the generalization performance of the image data. This process was designed to enable the model to respond effectively to diverse imaging conditions that might occur in real clinical environments.

First, horizontal and vertical flipping were applied through HorizontalFlip and VerticalFlip, and image rotation, position shifting, magnification, and reduction were performed using RandomRotate90 and ShiftScaleRotate techniques. In particular, ShiftScaleRotate includes rotation within a maximum range of 20 degrees, helping the model become robust to variations in orientation and placement.

Additionally, to reflect more nonlinear forms of transformation, ElasticTransform was used to locally deform pixels within the image, and GridDistortion was applied to allow learning of structural distortions. Furthermore, RandomGamma was used to randomly adjust gamma values to simulate brightness changes, and GaussianBlur was applied with a low probability to enable learning even in image conditions including blur.

These augmentation techniques help the model adapt to various types of images while emphasizing fine structures such as periapical lesions, ultimately enabling the neural network to learn more robust and generalized features.

After all augmentations and transformations were applied, pixel values were normalized based on the mean and standard deviation of ImageNet (IMAGE_MEAN, IMAGE_STD) to consistently adjust the distribution of data, and ToTensorV2() was applied to convert the images into tensor format that can be input to PyTorch models.

Meanwhile, this pipeline was configured to process bounding box information as well, considering object detection model training. Through BboxParams settings, the bounding box format was designated as pascal_voc, and only objects that satisfied a certain area (min_area=1) and minimum visibility (min_visibility=0.1) were maintained. Additionally, to prevent bounding boxes from going outside the image during the image augmentation process, the clip=True option was applied to ensure that box coordinates always remain within the image.

This augmentation processing pipeline was designed considering the characteristics of radiographic images, with the aim of maximizing the model's generalization performance and robustness while maintaining data quality.

2.3.3 Data Labeling

In this study, we performed direct bounding box labeling work on image data. This was a strategic approach to clearly recognize the differences in interpretation methods between clinical data and image data, and to overcome the specificity of image-based analysis. While clinical data is based on indirect information such as patient subjective symptoms or examination results, image data has the characteristic of allowing direct observation of the structural form and location of lesions. Accordingly, by directly marking and quantifying visually clearly identifiable lesions, we aimed to provide the model with more specific information about the lesions themselves.

The labeling work in this study was performed using the SLAM (playidea lab) platform, targeting a total of 776 periapical radiographs. Labeling was based on BOX unit labeling, and was performed according to consistent criteria by fixing the types of lesions into four classes (dental caries, tooth fracture, bone loss, PDL space widening).

The criteria for interpreting lesions were based on existing radiographic diagnostic guidelines for identifying dental caries and periapical lesions in the jawbone. In particular, dental caries existing on the occlusal surface but not visible in the image were excluded from labeling targets, and dental caries identifiable in the image were all included according to radiographic interpretation criteria (Pitts, 2001). The distinction between dental caries and tooth fractures was based on the presence or absence of straight lines appearing at the boundary of radiolucency, and in cases where it was difficult to differentiate between periapical lesions and bone loss due to periodontitis, electronic medical records were used as supplementary material (Petersson et al., 2012).

Additionally, although thickening of the periodontal ligament space (PDL space widening) is often difficult to definitively categorize as a clear lesion, it was included in the labeling target as it was judged to be an indirect indicator reflecting percussion response or pulp sensitivity in clinical examinations. In this case as well, images with ambiguous or unclear interpretations were interpreted by referring to medical records.

The labeling process did not take place as a one-time event but included repetitive refinement processes. After the first manual labeling was completed, label cleaning was performed using the detection results of an artificial intelligence model trained based on that data. An example of label cleansing can be seen in **Figure 5**. This was done by having humans recheck and modify the label positions generated by AI, and the precision of labels was increased by conducting repeated labeling and inspection work three times.

2.3.4 Self-Supervised Learning with Masked Autoencoder

In the field of medical image analysis, acquiring labeled data is often difficult, which serves as one of the biggest constraints in AI model development. Self-Supervised Learning (SSL) is a powerful learning method to overcome these limitations, allowing data to be learned without labels. A schematic diagram of SSL can be seen in **Figure 6**. Masked Autoencoder (MAE) is a representative technique used in the image domain.

The Masked Autoencoder (MAE) is structured to randomly mask portions of an image and then learn to restore those regions. As shown in **Figure 7**, the model develops the ability to understand and reconstruct the structural, morphological, and semantic features of the image on its own. MAE is trained to infer the whole based on the remaining parts by hiding some portions rather than directly learning the entire input image, making it more sensitive to the core patterns of the image or surrounding structures of lesions. Especially in cases like radiographic images with high resolution but ambiguous lesion boundaries, MAE has the advantage of learning more sophisticated visual representations compared to conventional supervised learning-based models. In this study, we applied MAE-based SSL to radiographic image data, establishing a learning foundation that can effectively grasp the structural information the data possesses. This can play an important role in improving the performance of subsequent classification and detection models.

Providing additional labeled data to a pre-trained model using a self-supervised approach plays a crucial role in precisely adjusting the model's learning direction and enhancing its ability to perform specific diagnostic tasks. In particular, in this study, we conducted box-based labeling for major lesions such as dental caries, tooth fracture, bone loss, and PDL space widening on radiographic images, enabling the model to learn clinical judgment abilities beyond simple restoration. This brings about the effect of improving the performance of lesion detection and classification by assigning diagnostic-centered objectives to the restoration-centered MAE. Consequently, this strategy allows for securing both model accuracy and practicality with only a limited number of labels, significantly increasing the potential for application in actual clinical settings.

To learn the complex structural representations of X-ray images, we first performed pre-training using a Vision Transformer (ViT)-based Masked AutoEncoder (MAE). MAE is a self-supervised learning method based on the Vision Transformer (ViT) structure that masks a certain percentage of the input image and then learns to restore the entire image from the remaining patches.

In this study, we divided the original X-ray images into fixed-size patches, randomly masked some of them (10% was used in this experiment), and learned latent representations by inputting the remaining patches into the encoder. The decoder was then configured to restore the entire image and Mean Squared Error (MSE) was used as the loss function at the pixel level.

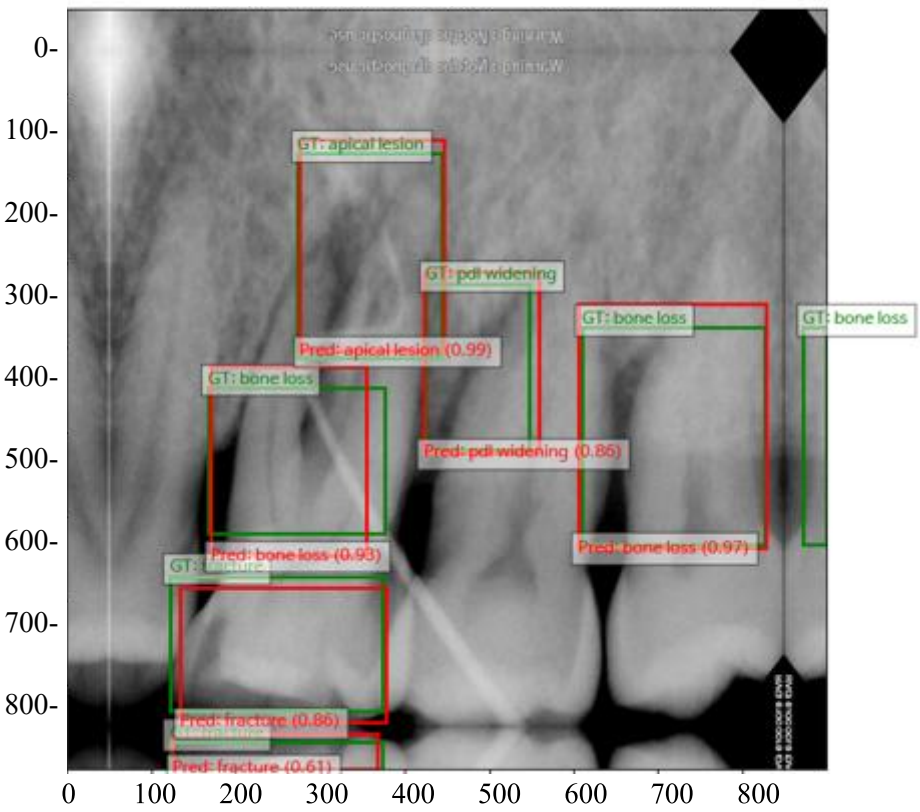

Figure 5. Example of label cleansing process using periapical radiograph.

Ground truth labels (green boxes) and AI-predicted labels (red boxes) are shown for apical lesion, PDL space widening, bone loss, and tooth fracture. The image illustrates the manual correction process of AI-generated annotations to improve labeling accuracy during the development of the object detection model.

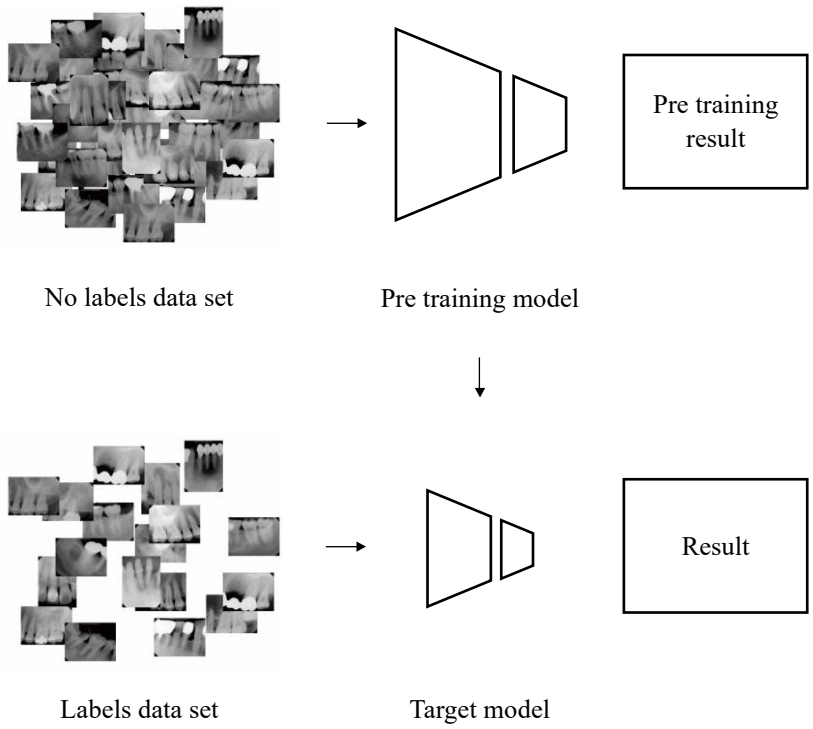

Figure 6. Self-supervised pretraining and transfer learning workflow.

An unlabeled dataset is first used to pretrain a model through a self-supervised learning approach, enabling the model to extract generalizable visual representations. The pretrained model is then fine-tuned on a labeled dataset (target dataset) to perform supervised learning for final diagnostic classification tasks.

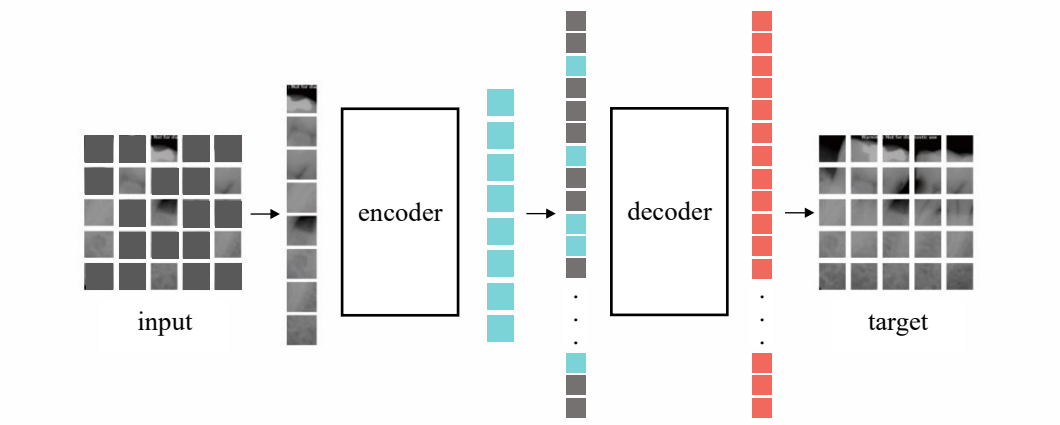

Figure 7. Architecture of the Masked Autoencoder (MAE) for self-supervised learning.

The input image is divided into patches, and a subset is randomly masked. The visible patches are passed through an encoder to generate latent representations, which are then reconstructed by a decoder. The model learns to predict the original image from the masked input, enabling effective feature learning without labels.

In the MAE structure, since the encoder performs computations only on unmasked patches, it has the advantages of high computational efficiency, faster learning, and effective learning of global representations. The encoder pre-trained in this manner was later transferred to lesion detection or classification tasks, particularly serving as a robust feature extractor in X-ray images containing various anatomical structures and disease patterns.

As another pre-training approach, we applied an image restoration (autoencoding) method using a U-Net-based CNN structure. U-Net consists of an encoder-decoder structure and can learn meaningful visual representations through the process of compressing the input image into a low-dimensional latent space and then restoring it to its original resolution Transformer-based techniques and CNN-based approaches employ fundamentally different processing strategies: the former utilizes a top-down mechanism to analyze relationships between distinct image regions, while the latter implements a bottom-up approach that prioritizes local pixel neighborhood connections . In our methodology, we enhanced the input X-ray images through various data augmentation techniques including Gaussian noise application and rotational transformations before strategically applying masking patterns. The U-Net architecture was then trained to accurately reconstruct the original unaltered images from these modified inputs. This process is similar to the Denoising Autoencoder approach and was designed to guide the encoder to focus on extracting important structural features from the input images. The encoder part of U-Net performs progressive abstraction of the image through multiple convolutional blocks and pooling layers, and shares information with the decoder through skip connections, allowing it to learn global representations while maintaining detailed information. After pre-training was completed, the decoder was removed, and the encoder was transferred as a feature extractor for downstream tasks. This encoder was later used as the backbone for detection or classification models, and showed improvements in both convergence speed and final performance through fine-tuning from the pre-trained state.

These two approaches utilize the advantages of Transformer-based and CNN-based visual representation learning, respectively, and in this study, we comparatively applied both models depending on experimental conditions. MAE showed strengths in global structure representation, while U-Net excelled in local detail representation, and there were performance differences depending on specific lesion types. This comparison provides meaningful implications for designing pre-training strategies suitable for the characteristics of X-ray images.

2.3.5 Radiological Model Training

Based on the architecture shown in **Figure 8**, full-scale supervised fine-tuning was conducted across various backbone network structures as described in the main text, with controlled levels of image augmentation. In this process, various types of intensity-adjustable augmentations such as contrast enhancement, rotation, and elastic deformation were designed and applied to the input

images. These augmentations enabled the model to learn robust representations that could accommodate different types of image distortions, inter-patient differences, and even intra-patient variability. The generalization performance of the model varied depending on the extent of augmentation, even within the same backbone structure. After each convolution and linear layer of the network, Batch Normalization was applied to reduce internal covariate shift and secure both learning speed and convergence stability. This particularly helped maintain smooth gradient flow in deep model structures and contributed to alleviating instability that could occur in the early stages of learning.

Additionally, several optimization strategies were implemented in parallel to improve the model's learning stability and generalization performance. First, to prevent overfitting and suppress the growth of unnecessary parameters, an L2 regularization-based Weight Decay term was added to the common Binary Cross Entropy or Focal Loss functions. For this, AdamW (Adam with decoupled weight decay) was adopted as the optimization algorithm. Unlike the original Adam, AdamW applies the weight decay term separately from learning rate updates, which can more accurately reflect the meaning of L2 regularization and is generally known to show better generalization performance in various deep learning benchmarks.

For learning rate settings, warm-up scheduling and learning rate decay strategies were combined. Initially, starting with a low learning rate and linearly increasing it up to a certain step through a warm-up phase, the learning rate was then gradually decreased based on validation loss or epoch count. This allowed the model to be free from unstable parameter updates in initial learning and induced stable convergence.

Furthermore, systematic exploration (hyperparameter optimization) was performed on key hyperparameters such as model structure, learning rate, batch size, dropout ratio, and weight decay intensity. In this process, Random Search and Grid Search were conducted in parallel, and multiple combinations were evaluated based on performance metrics (AUC or macro F1). In some experiments, Bayesian Optimization techniques were also introduced to increase search efficiency, and based on the experimental results, optimal learning settings were determined for each backbone structure.

Through this advanced learning pipeline, we were able to systematically compare and analyze the differences in classification performance of models according to the type of backbone network and input augmentation level. While increasing augmentation intensity contributed to performance improvement in certain structures, excessive transformation sometimes caused performance degradation. This was confirmed to act as an important design element that interacts with the depth and complexity of the model, and whether it was pre-trained.

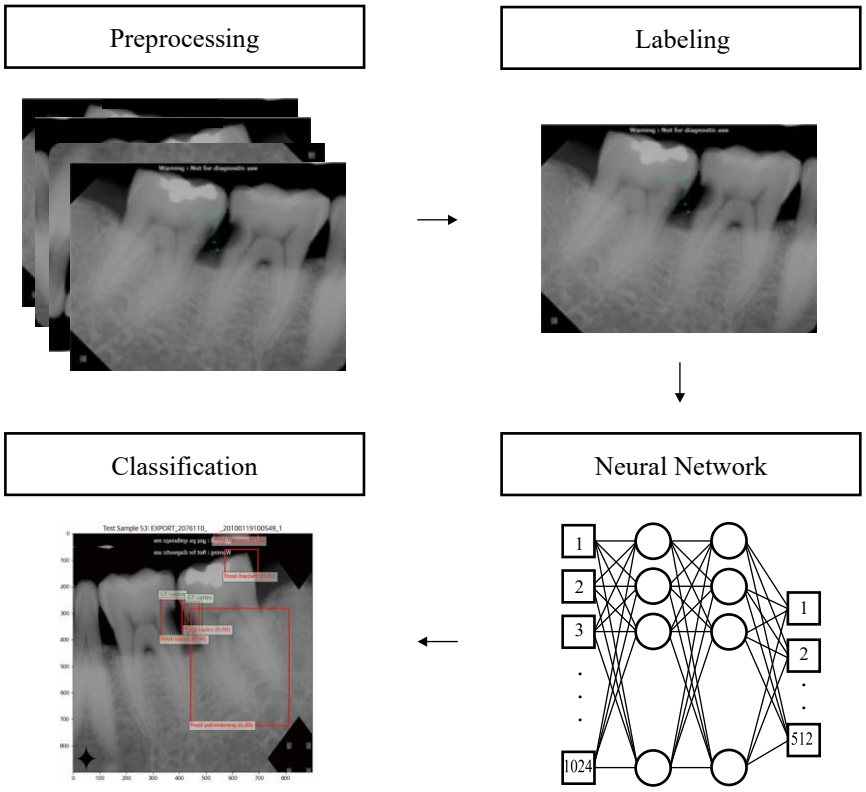

Figure 8. Workflow of radiographic image classification using a neural network.

The process consists of four main stages: (1) Preprocessing of periapical radiographic images, (2) Labeling of dental lesions, (3) Neural network training using labeled data, and (4) Classification output showing predicted lesion regions. This pipeline enables automated diagnosis through supervised learning.

2.4. Multi-modal model

2.4.1 Multi-modal model Structure

The multimodal artificial intelligence model designed in this study adopted a strategy that simultaneously utilizes two different modalities, namely clinical data and radiographic images, to overcome the limitations in representational capacity that may appear in conventional approaches using only single-modality data.

As can be seen in the provided **Figure 9**, clinical examination data consists of numerical and categorical features, and this clinical information first undergoes an encoding process to be converted into a high-dimensional vector representation. Through this, various clinical characteristics are transformed into meaningful numerical representations, and feature extraction is ultimately performed through a Multi-Layer Perceptron (MLP).

Meanwhile, radiographic images are data in the form of images containing spatial characteristics and pathological visual information, which were learned using a CNN-based image encoder. In this study, we applied the latest deep learning architectures such as EfficientNet V2 and Vision Transformer (ViT) as image encoders to extract complex image features more elaborately and effectively. In this process, periapical radiographic images are standardized to an input data size of 640×640, maintaining consistent quality and resolution when input to the encoder.

Clinical information and image information that have gone through two independent encoding processes then move on to the fusion stage. In this model, feature vectors formed in different representation spaces were integrated using an Attention mechanism.

The Attention mechanism selectively emphasizes only important information from data of different modalities, allowing effective integration of complementary information between the two modalities. Through this, the interaction between spatial pathological information obtained from images and numerical and categorical patient characteristics extracted from clinical data is more clearly modeled, enabling more sophisticated prediction of the patient's condition.

After Attention-based fusion, final prediction ($\hat{y}$) is performed through an additional MLP, and in the model's learning process, a Focal Loss function was applied to effectively handle the class imbalance problem. Focal Loss helps perform model learning more efficiently by assigning higher weights to diagnostic classes that appear relatively less frequently.

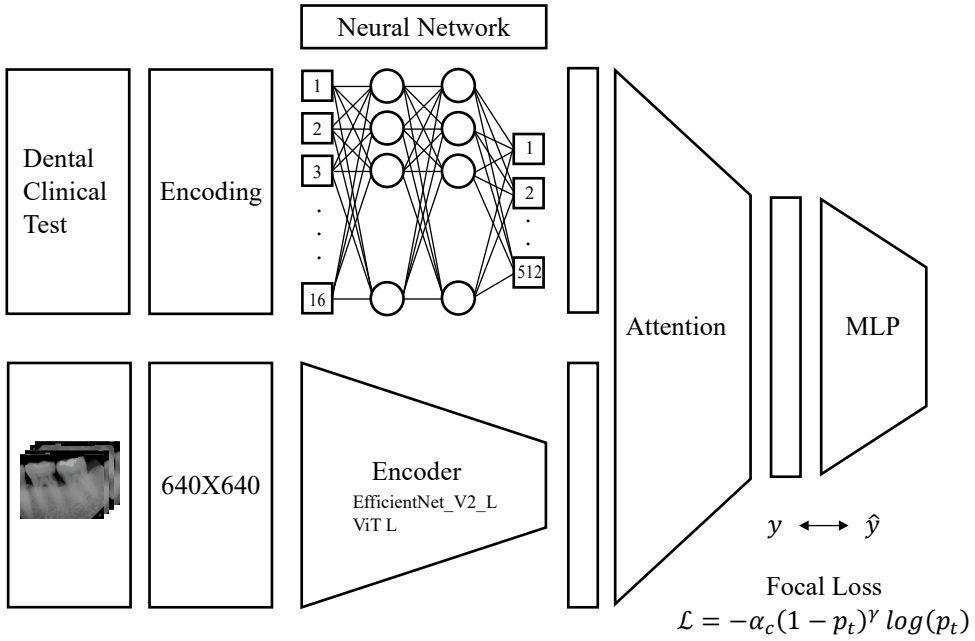

Figure 9. Architecture of the proposed multimodal deep learning model.

The model integrates dental clinical test data (top path) and periapical radiographic images (bottom path). Clinical data is encoded and passed through a multilayer perceptron (MLP), while radiographic images are processed using a convolutional or transformer-based encoder (e.g., EfficientNet_V2_L or ViT_L). The outputs from both modalities are fused via an attention mechanism, followed by classification through an MLP. Focal loss is used to address class imbalance in multi-label classification.

2.4.2 Multi-Modal Modeling

The first approach is a structure that individually encodes each modality, simply combines (concatenates) them into a single vector, and then performs the final prediction through a multi-layer perceptron (MLP). Clinical data consists of standardized numerical and categorical inputs, which were transformed into latent representations by inputting them into an MLP-based clinical encoder. Image data was converted into abstracted feature maps or flattened embeddings through a pre-trained CNN or ViT-based encoder. Subsequently, the two encoding results were simply concatenated to create a merged vector, which was then passed through additional MLP layers to perform the final classification.

This method enables effective fusion under the assumption that the two data types are stably represented individually, and it is simple to implement with fast learning. In this study, this structure was set as the baseline multimodal model to serve as a reference point for comparing performance with other fusion methods.

The second approach was designed to more elaborately reflect the information flow and correlation structure between the two modalities by applying cross-attention or self-attention-based interaction techniques rather than simple combination. In this structure, after independently encoding both clinical data and image data, cross-attention operations were performed using the two representations as queries, keys, and values for each other. For example, by using the clinical encoding result as the query and the image encoding result as the key-value, the model can learn which patterns in the image should be focused on based on the clinical information, and the reverse direction can be designed similarly.

This method provides flexibility to dynamically adjust the relationship between modalities during the learning process by weighting the interaction between each modality based on attention rather than linear combination. Especially in medical image analysis problems where the areas to focus on in images may vary depending on the clinical context, such attention-based structures provide powerful expressiveness.

These two fusion strategies were compared under the same dataset and learning conditions, and while the attention-based structure showed increased model complexity and longer learning times in terms of performance, the simple combination method demonstrated structurally simple yet stable convergence characteristics, showing strengths in terms of deployment potential. However, neither method showed significant performance improvements in integrated representation learning between modalities compared to single-modality-based models. This is interpreted as indicating that clinical data and image information contain differences that are not mutually complementary.

2.4.3 Multi-modal Detection model

In previous multimodal classification experiments, the approach of integrating clinical information and image data to predict lesions showed limited performance improvement relative to structural complexity and did not demonstrate significant improvement compared to single-modality-based models. Accordingly, this study attempted to analyze the causes of this performance stagnation and take a more fundamental approach.

A detailed review of the image data revealed that two or more lesions frequently coexist in a single PA image, clearly demonstrating the limitation that binary classification or multi-label classification approaches cannot sufficiently reflect the spatial characteristics and boundary distinctions between these lesions. In particular, since the location, shape, and range of lesions differ from each other, the necessity arose for the model to explicitly learn where lesions exist within the image, beyond simply determining "whether lesions exist."

Accordingly, this study departed from the existing classification-centered design of single diagnostic models using periapical radiographic images and redesigned multimodal fusion by newly configuring an image-based lesion detection model focused on object detection. For each X-ray image, experts directly labeled the locations of lesions with bounding boxes, and each lesion such as dental caries, tooth fracture, and apical lesions was defined with separate class IDs. Subsequently, training proceeded with a detection structure that could simultaneously perform bounding box regression and multi-class classification based on the labeled data.

For the deep learning model for lesion detection, the DETR (DEtection TRansformer) structure was adopted, and experiments were conducted based on an end-to-end learning method differentiated from existing convolution-based detection models. DETR takes the entire image as input and uses object queries to directly predict lesion objects. Since accurate detection is possible without separate anchor settings or NMS (Non-Maximum Suppression) processes, it was deemed suitable for detecting objects with irregular shapes and sizes, such as lesions.

3. RESULTS

3.1. Single-modal model Evaluation and Analysis

3.1.1 Clinical Test Data Model

In this study, a total of 776 clinical data samples were preprocessed, and the distributions of patient demographics such as sex and age were visualized in **Figure 10**. Additionally, the distribution of each clinical test result and the categorized chief complaints (C.C.) were also visualized for further analysis.

Data preprocessing included variable transformation, categorical variable encoding, scaling, feature selection, and reconstruction, resulting in a high-dimensional dataset. Based on this, various machine learning techniques were experimentally applied. Initial model development focused on dental caries, the most prevalent condition in the dataset, and subsequent experiments were extended to other types of lesions.

The models were developed using the PyCaret AutoML library, which allowed efficient comparison across multiple well-established classification algorithms, including Logistic Regression, Decision Tree, SVM, Gradient Boosting, and Random Forest. In the initial stage, 14 models were trained without considering class imbalance, and their performances were summarized in **Table 3** with AUC as the primary evaluation metric.

Considering the impact of class imbalance on model performance, data rebalancing techniques such as SMOTE were applied, and the models were retrained accordingly. The results presented in **Table 4** demonstrated that the performance of most models improved significantly with Random Forest showing the best results and ultimately being selected for further analysis. As an ensemble learning algorithm based on multiple decision trees, Random Forest effectively captures interactions among high-dimensional clinical variables, offers strong resistance to overfitting, and allows intuitive interpretation of feature importance-aligning well with the goals of this study.

All experiments were conducted using 5-fold cross-validation, and performance was evaluated using metrics such as F1-score, ROC AUC, Precision, and Recall.

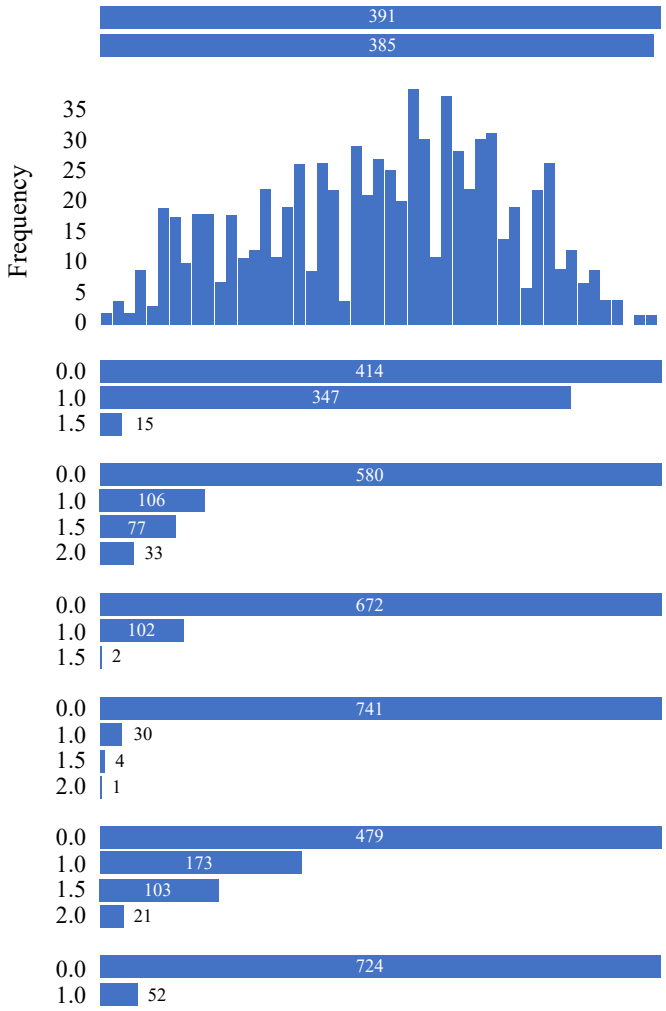

Figure 10. Visualization of demographic distribution and clinical test frequencies.

This figure illustrates the distribution of sex, age, and frequencies of seven clinical test results. Percussion, Mobility, Bite, Air, Cold, and EPT (Electric Pulp Test) used for model training. Each clinical test is visualized with the number of cases per score level, highlighting class imbalance across categories

Table 3. Model training results for dental caries detection without considering class imbalance in clinical data

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT(Sec)
Extreme Gradient Boosting	**0.7847**	**0.8066**	0.5828	0.6570	**0.6115**	**0.4640**	**0.4699**	1.2220
K Neighbors Classifier	0.7230	0.7579	0.7012	0.5181	0.5946	0.3919	0.4033	0.6280
Decision Tree Classifier	0.7572	0.7151	0.6111	0.5801	0.5924	0.4203	0.4227	0.6740
Random Forest Classifier	**0.7847**	0.8061	0.5401	**0.6620**	0.5919	0.4480	0.4542	0.9380
Light Gradient Boosting Machine	0.7765	0.8003	0.5590	0.6461	0.5918	0.4401	0.4474	0.7740
Gradient Boosting Classifier	0.7599	0.7799	0.5691	0.5900	0.5772	0.4102	0.4119	1.2600
Extra Trees Classifier	0.7710	0.7941	0.5399	0.6318	0.5763	0.4217	0.4279	0.8120
Logistic Regression	0.7311	0.7672	0.6346	0.5309	0.5727	0.3801	0.3874	0.7460
Ada Boost Classifier	0.7407	0.7501	0.5970	0.5485	0.5695	0.3850	0.3874	0.8380
Quadratic Discriminant Analysis	0.6844	0.7590	0.7056	0.4808	0.5651	0.3354	0.3559	0.7300
Ridge Classifier	0.7147	0.7485	0.6252	0.5047	0.5560	0.3498	0.3563	0.7520
Linear Discriminant Analysis	0.7119	0.7292	0.6112	0.5005	0.5479	0.3400	0.3456	0.7740
SVM-Linear Kernel	0.7132	0.7395	0.5878	0.5111	0.5413	0.3361	0.3414	0.6800
Naïve Bayes	0.4101	0.5640	**0.9004**	0.3184	0.4700	0.0731	0.1245	0.6300

Table 4. Model training results for dental caries detection considering class imbalance in clinical data.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT(Sec)
Gradient Boosting Classifier	**0.8820**	0.8781	0.6727	**0.6150**	**0.6344**	**0.5651**	**0.5711**	1.4740
Random Forest Classifier	0.8807	**0.8971**	0.6296	0.6137	0.6201	0.5495	0.5505	0.8980
Extreme Gradient Boosting	0.8780	0.8648	0.5933	0.6063	0.5947	0.5236	0.5265	1.1460
Light Gradient Boosting Machine	0.8766	0.8721	0.5933	0.6027	0.5940	0.5218	0.5241	1.3800
Quadratic Discriminant Analysis	0.8326	0.8321	0.7791	0.4823	0.5935	0.4966	0.5203	0.7600
K Neighbors Classifier	0.8354	0.8720	0.7522	0.4805	0.5850	0.4886	0.5089	0.6780
Extra Trees Classifier	0.8683	0.8955	0.5660	0.5757	0.5688	0.4914	0.4926	0.8340
Ada Boost Classifier	0.8504	0.8278	0.6379	0.5211	0.5665	0.4788	0.4868	0.9240
Decision Tree Classifier	0.8505	0.7526	0.6016	0.5184	0.5543	0.4655	0.4690	0.6780
Ridge Classifier	0.8203	0.8644	0.6723	0.4495	0.5379	0.4322	0.4461	0.8020
Linear Discriminant Analysis	0.8230	0.8405	0.6644	0.4526	0.5372	0.4329	0.4460	0.8480
Logistic Regression	0.8244	0.8447	0.6372	0.4596	0.5300	0.4267	0.4377	0.7520
SVM-Linear Kernel	0.8134	0.7986	0.5411	0.4325	0.4713	0.3625	0.3712	0.7140
Naïve Bayes	0.5158	0.6813	**0.9202**	0.2322	0.3707	0.1638	0.2692	0.6440

Throughout the study, a variety of preprocessing strategies, oversampling methods, loss functions, and classification algorithms were compared. Among them, Random Forest consistently demonstrated high accuracy and reliability. SMOTE-based oversampling and stratified cross-validation notably contributed to improved model stability. These findings also provide foundational insights that can be applied to the development of more complex deep learning models in the future.

Following the experiments on single-lesion classification for dental caries, the models were extended to include tooth fracture and pulpitis. Their respective performance results are illustrated in **Figure 11**. Hyperparameter tuning was performed using AUC as the primary criterion, which is particularly suitable for evaluating classifier performance in imbalanced datasets.

Subsequently, a multi-label Random Forest model was trained to simultaneously classify all three conditions (dental caries, tooth fracture, and pulpitis). Although this multi-label model showed a slight decrease in AUC compared to the individual models, it still achieved robust performance with AUC values above 0.78 for all conditions, as shown in **Figure 12**. Interestingly, improvements in pulpitis classification appeared to influence the performance of other conditions as well. Feature importance plots were analyzed to identify key variables contributing to the prediction of each condition

Despite the stable performance of traditional machine learning algorithms such as Random Forest, their inherent structural limitations were also recognized—particularly in capturing complex, nonlinear relationships within high-dimensional clinical and imaging data. Therefore, this study aimed to extend the model to a neural network architecture with enhanced representational capabilities by designing a Multi-Layer Perceptron (MLP).

The MLP-based diagnostic model incorporated pretraining and data augmentation techniques to maximize performance, with the results summarized in **Figure 13**. Evaluation was tailored to the multi-label classification task, with per-class precision, recall, and F1-scores calculated. Additionally, threshold tuning of sigmoid outputs beyond the default 0.5 was performed to optimize the balance between sensitivity and specificity. The model was validated using stratified k-fold cross-validation, with independent training and evaluation in each fold. To ensure prediction stability across multiple lesions, standard deviations across folds and class-level performance variances were also analyzed. The 5-fold validation of the final model is shown in **Table 5**.

As a result, the MLP-based multi-label classification model demonstrated improved diagnostic performance compared to Random Forest. Notably, combining the model with focal loss and class imbalance handling strategies led to consistent performance improvements across all target lesions.

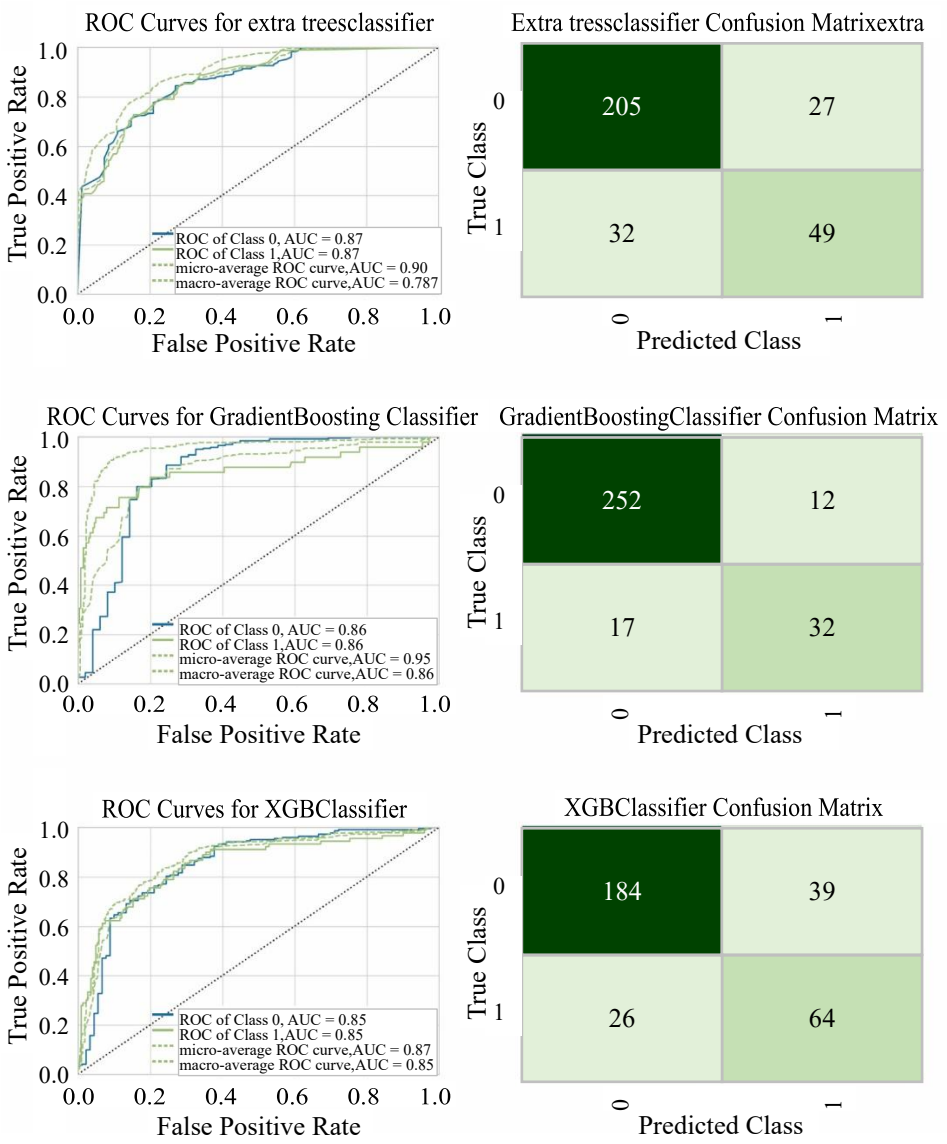

Figure 11. Performance comparison of Random Forest-based models for single-lesion diagnosis (dental caries, tooth fracture, and pulpitis).

The ROC curves and confusion matrices show the classification performance of three Random Forest variants for diagnosing each lesion separately. (a) Extra Trees Classifier for dental caries diagnosis: AUC of 0.87 for both class 0 and class 1. (b) Gradient Boosting Classifier for tooth fracture diagnosis: AUC of 0.86 for both classes, with the highest micro-average AUC of 0.95. (c) XGB Classifier for pulpitis diagnosis: AUC of 0.85 for both classes, with balanced classification performance.

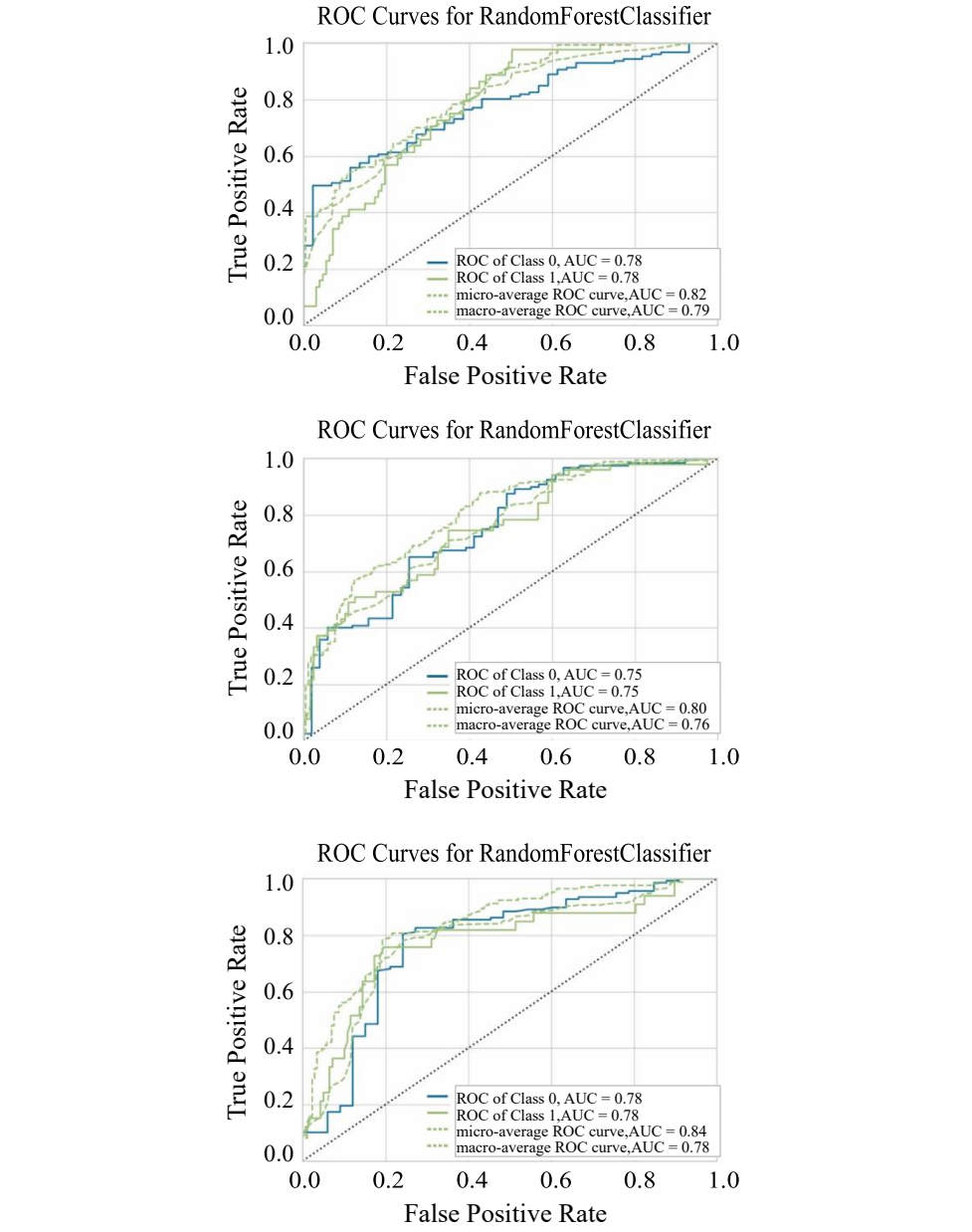

Figure 12. ROC curve performance of the Random Forest classifier for each disease in the unified classification model.

(a) Dental caries classification: AUC = 0.82 , (b) Tooth fracture classification: AUC = 0.80, (c) Pulpitis classification: AUC = 0.84.

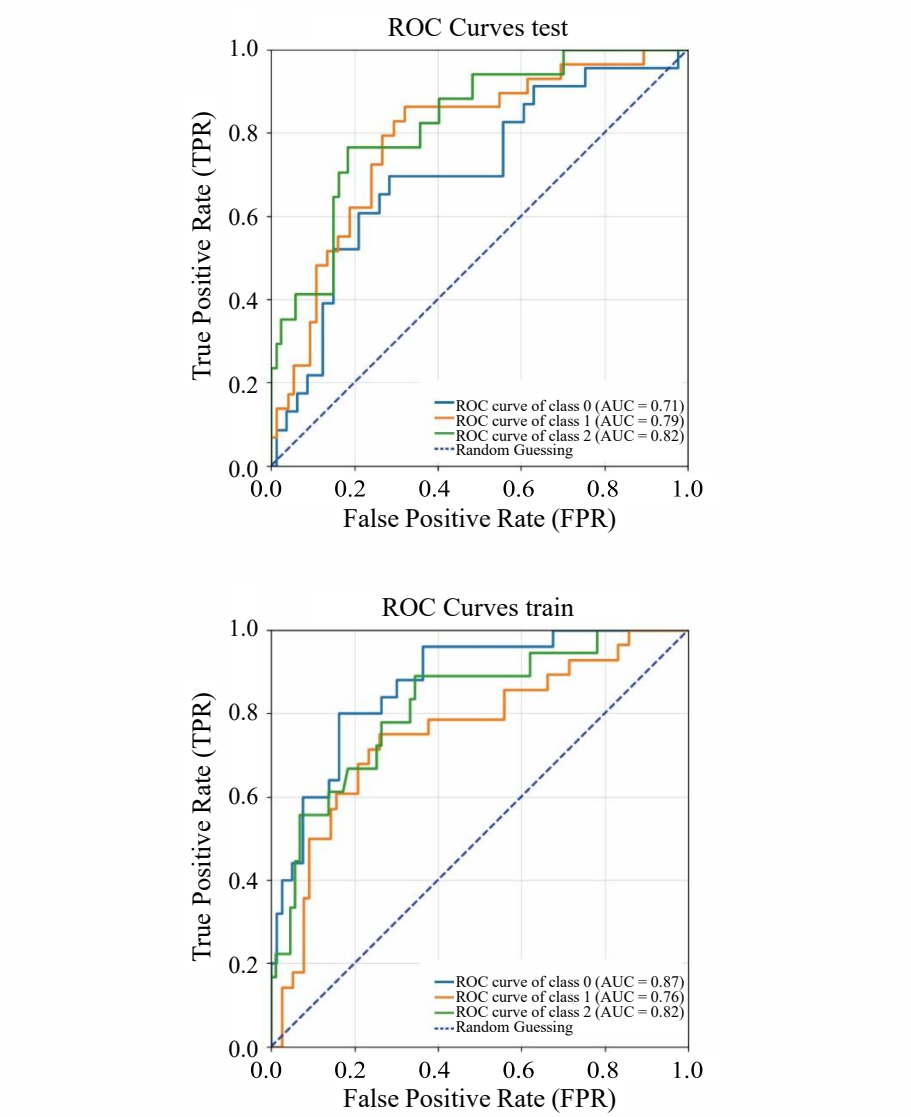

Figure 13. ROC curve comparison for clinical data-based diagnostic model.

(a) Performance of the model without pre-training. The AUC values for class 0 (dental caries), class 1 (tooth fracture), and class 2 (pulpitis) are 0.71, 0.79, and 0.82, respectively, on the test dataset.(b) Performance of the model with pre-training and additive Gaussian noise-based data augmentation. The corresponding AUC values on the training dataset improved to 0.87, 0.76, and 0.82, showing enhanced stability and generalization

Table 5. Results of 5-fold cross-validation based on the random forest model.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8630	0.8920	0.4348	0.5882	0.5000	0.4227	0.4292
1	0.8699	0.8579	0.6087	0.5833	0.5957	0.5182	0.5184
2	0.8973	0.9113	0.7391	0.6538	0.6939	0.6324	0.6341
3	0.8836	0.8022	0.5909	0.6190	0.6047	0.5364	0.5366
4	0.8897	0.8688	0.7273	0.6154	0.6667	0.6011	0.6041
Mean	**0.8807**	**0.8664**	**0.6202**	**0.6120**	**0.6122**	**0.5422**	**0.5445**
Std	0.0126	0.0371	0.1104	0.0253	0.0672	0.0728	0.0716

3.1.2 Radiological Model

To effectively learn the complex structural characteristics of periapical radiographs, two distinct pretraining strategies were employed and compared in this study: (1) a Vision Transformer (ViT)-based Masked Autoencoder (MAE) approach, and (2) a CNN-based U-Net architecture for image reconstruction.

The MAE framework is a self-supervised learning method in which a portion of the input X-ray image is randomly masked, and the remaining visible patches are used to reconstruct the full image. This approach enables the model to learn global visual representations efficiently. In this study, 10% of the image patches were masked, and the model was trained to minimize pixel-wise Mean Squared Error (MSE) between the reconstructed and original images. The encoder, pretrained through this process, was later transferred as a feature extractor for downstream classification and lesion detection tasks.

For the CNN-based pretraining, a denoising autoencoding scheme using the U-Net architecture was implemented. The input X-ray images were augmented with Gaussian noise and geometric transformations (e.g., rotation), followed by masking of specific regions. The model was trained to reconstruct the original image from the corrupted input, guiding the encoder to focus on meaningful local structures. The encoder progressively abstracted the image through multiple convolutional and pooling layers, while skip connections allowed the decoder to recover fine details. After pretraining, the decoder was removed, and the encoder was fine-tuned as the backbone for the downstream classification tasks.

Supervised fine-tuning was conducted on each pretrained encoder. Various levels of image augmentation, such as contrast enhancement, rotation, and elastic deformation, were applied to improve robustness against image variability and inter-patient anatomical differences. Batch Normalization was used after each convolutional and linear layer to reduce internal covariate shift, thereby improving training stability and convergence speed, particularly in deeper architectures.

To enhance training stability and generalization performance, several optimization strategies were employed. The loss functions included Binary Cross Entropy or Focal Loss, both combined with L2-based weight decay to prevent overfitting. The AdamW optimizer was adopted as it decouples weight decay from gradient updates. This improves the effectiveness of L2 regularization and often results in better generalization across deep learning benchmarks.

The learning rate was adjusted using a combined warm-up and decay schedule. Training started with a low learning rate, which was linearly increased over the initial steps, and then gradually decreased based on validation loss or the number of epochs. A systematic hyperparameter search, including random search, grid search, and Bayesian optimization, was performed to tune key parameters such as learning rate, batch size, dropout rate, and weight decay. Final configurations were selected based on model performance metrics, including AUC and macro F1-score.

Through this structured training pipeline, we systematically compared the classification performance of various backbone architectures across different levels of augmentation. The results showed that increasing augmentation strength generally improved model performance. However, overly aggressive transformations sometimes degraded performance, depending on the model's depth, complexity, and pretraining method.

Overall, models that incorporated pretrained encoders outperformed their non-pretrained counterparts. These models exhibited both superior classification accuracy and faster convergence. Specifically, the MAE-based transformer encoder proved effective in capturing global anatomical structures in X-ray images. In contrast, the CNN-based U-Net encoder demonstrated strength in representing localized structural details, leading to higher F1-scores for specific lesion types such as dental caries. The result graphs are shown in **Figure 14.**

These findings underscore the importance of selecting pretraining strategies that are tailored to the unique characteristics of periapical radiographs. Performance variations observed across different lesion types suggest that the anatomical and radiographic nature of the pathology may determine whether global or local visual representations are more effective. This study confirms that transformer- and CNN-based visual representation learning methods offer complementary advantages. It also emphasizes the need to align pretraining strategies with the clinical goals of the diagnostic task.

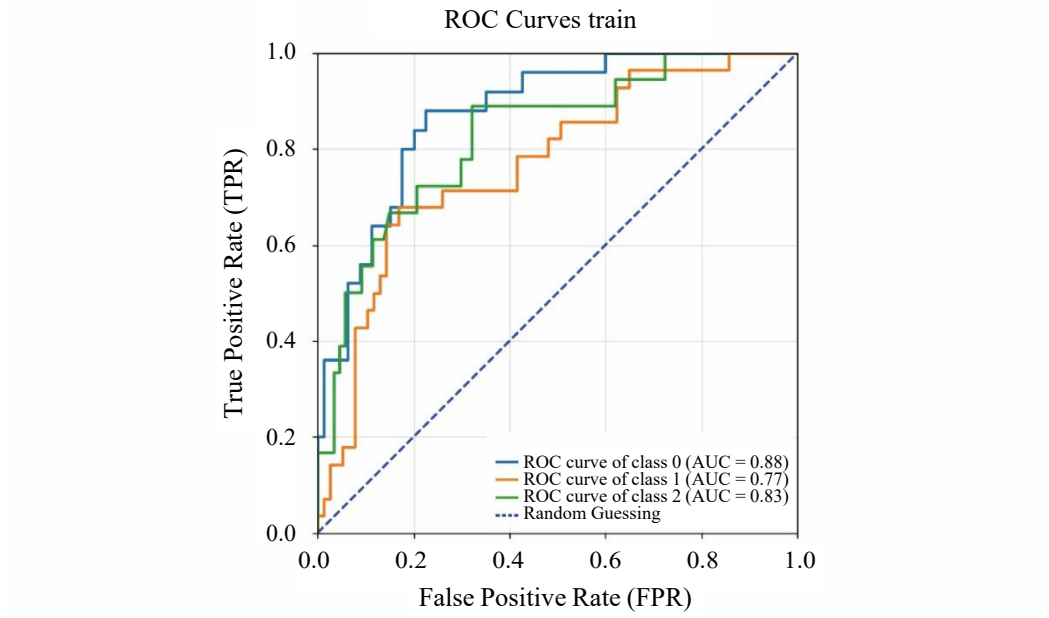

Figure 14. ROC curve of the final radiographic image-based AI model for dental caries, tooth fracture, and pulpitis classification.

This graph represents the final diagnostic performance of the AI model trained using periapical radiographs, incorporating both data augmentation through lesion labeling and self-supervised pretraining. AUC values: Dental caries = 0.88, Tooth fracture = 0.77, Pulpitis = 0.83.

3.2. Multi-modal model Evaluation and Analysis

This study aimed to overcome the representational limitations of single-modality learning and achieve more precise predictions of patient conditions by designing and analyzing a multimodal diagnostic model that integrates clinical data and periapical radiographic images. By effectively fusing two distinct data modalities residing in different feature spaces, the goal was to combine their complementary information and enhance overall predictive performance.

The first fusion approach involved independently encoding clinical and image data, then merging the outputs through simple concatenation into a single feature vector. This merged vector was subsequently passed through a Multi-Layer Perceptron (MLP) for final classification. The clinical data, comprising both numerical and categorical variables, were standardized and input into an MLP-based encoder to produce a latent representation. Meanwhile, the radiographic images were processed using a pretrained CNN or Vision Transformer (ViT) encoder, which generated flattened feature vectors. These two outputs were concatenated and passed through MLP layers for final prediction. This method offers simplicity and fast training, assuming that each modality's representation is independently robust. For this reason, it was established as the baseline multimodal structure for performance comparison with more complex fusion models.

The second fusion approach incorporated cross-attention or self-attention mechanisms to capture more sophisticated interactions between modalities. In this design, clinical and image data were first encoded separately, then used as queries, keys, and values in a cross-attention operation. For example, clinical features served as queries while image features acted as keys and values, enabling the model to learn which visual patterns were most relevant to the clinical context. The inverse direction was also implemented. Unlike simple linear concatenation, this attention-based approach dynamically adjusts the information flow between modalities during training, offering enhanced flexibility. This is particularly useful in medical imaging, where the regions of interest within an image may vary depending on the patient's clinical context.

These two multimodal fusion strategies were evaluated under identical datasets and training conditions. While the attention-based model offered greater expressiveness and adaptability, it also introduced increased model complexity and longer training times. Conversely, the simple concatenation approach provided structural simplicity and stable convergence, making it more practical for clinical deployment. The best performance results using this attention-based model can be seen in **Figure 15**.

However, neither fusion method showed a clear performance advantage over single-modality models. This suggests that the clinical and image data may not exhibit strong complementarity, or that some degree of information redundancy may exist. The detailed results of each experiment are summarized in **Table 6**, with Area Under the Curve (AUC) serving as the primary evaluation metric.

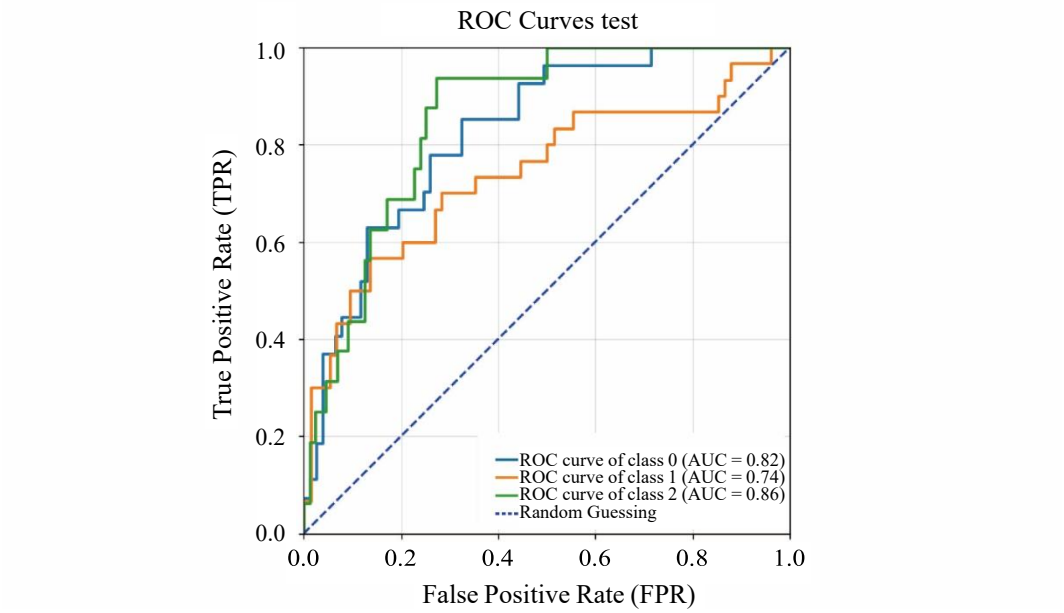

Figure 15. ROC curve performance of the final multimodal model combining clinical examination and periapical radiographs.

The ROC curve illustrates the diagnostic performance for each class using the fused model. AUC values: Dental caries = 0.82, Tooth fracture = 0.74, Pulpitis = 0.86. The model demonstrates improved diagnostic capability through integration of complementary features from both modalities

Table 6. Comparison of diagnostic performance (AUC) across different modality, pretraining, and fusion strategies.

Modality	Pretrained	Fusion	AUC_ Carries	AUC_ Fracture	AUC_ Pulpitis
Clinical only	Yes	N/A	**0.87**	0.76	0.82
X-ray only	Yes	N/A	**0.88**	0.77	0.83
Multimodal	Yes	Attention	**0.82**	0.74	0.86
Clinical only	No	N/A	0.71	0.79	0.82
Clinical only	Yes	N/A	0.82	0.7	0.82
Clinical only	No	N/A	0.74	0.77	0.85
X-ray only	Yes	N/A	0.87	0.71	0.83
X-ray only	No	N/A	0.71	0.72	0.84
X-ray only	MAE	N/A	0.85	0.77	0.84
X-ray only	U-Net	N/A	0.88	0.74	0.84
Multimodal	Yes	Concat	0.71	0.72	0.85
Multimodal	Yes	Attention	0.8	0.72	0.84
Multimodal	No	Concat	0.85	0.7	0.86
Multimodal	No	Attention	0.85	0.73	0.81
Multimodal	Mixed	Concat	0.83	0.73	0.81
Multimodal	Mixed	Attention	0.72	0.78	0.84

Based on these findings, this study sought to address the structural limitations and stagnating performance of multimodal classification models by transitioning to a fundamentally different modeling strategy.

Upon a detailed review of the periapical radiographs, it was frequently observed that multiple types of lesions coexisted within a single image. This complex scenario posed challenges for binary or multi-label classification models, which are not designed to capture spatial relationships or distinguish overlapping lesions. The varied location, shape, and extent of lesions highlighted the need for models that could explicitly learn where lesions are located in an image, not just whether they exist.

To meet this requirement, the research shifted from a classification-focused approach to a lesion detection model. Expert annotators manually labeled the location of lesions within each radiograph using bounding boxes, and each lesion type (e.g., dental caries, tooth fracture, apical lesion) was assigned a distinct class ID. The detection model was trained to perform both bounding box regression and multi-class classification simultaneously.

The architecture chosen for this task was DEtection TRansformer (DETR). Unlike conventional CNN-based detectors, DETR accepts the entire image as input and predicts lesion objects using object queries, without relying on anchor boxes or non-maximum suppression (NMS). This anchor-free, end-to-end framework is particularly suitable for detecting lesions that vary greatly in shape and size and offers a streamlined architecture.

This modeling shift is not only a technical advancement aimed at improving performance, but also holds significant value from a clinical application perspective. In real-world diagnostics, clinicians rely on lesion location information within the image. Therefore, detection-based models provide more intuitive and interpretable outputs than classification-based ones. This study demonstrates that detection models may form the foundation for future integration of image and clinical data, representing a paradigm shift from classification-based to spatially aware, detection-centered multimodal analysis.

Currently, the research team is extending this detection framework by incorporating clinical data, with the aim of building a multimodal lesion detection model that enhances spatial prediction performance through the integration of patient-level clinical information. The extension and experiments of this detection framework can be seen in **Figure 16**.

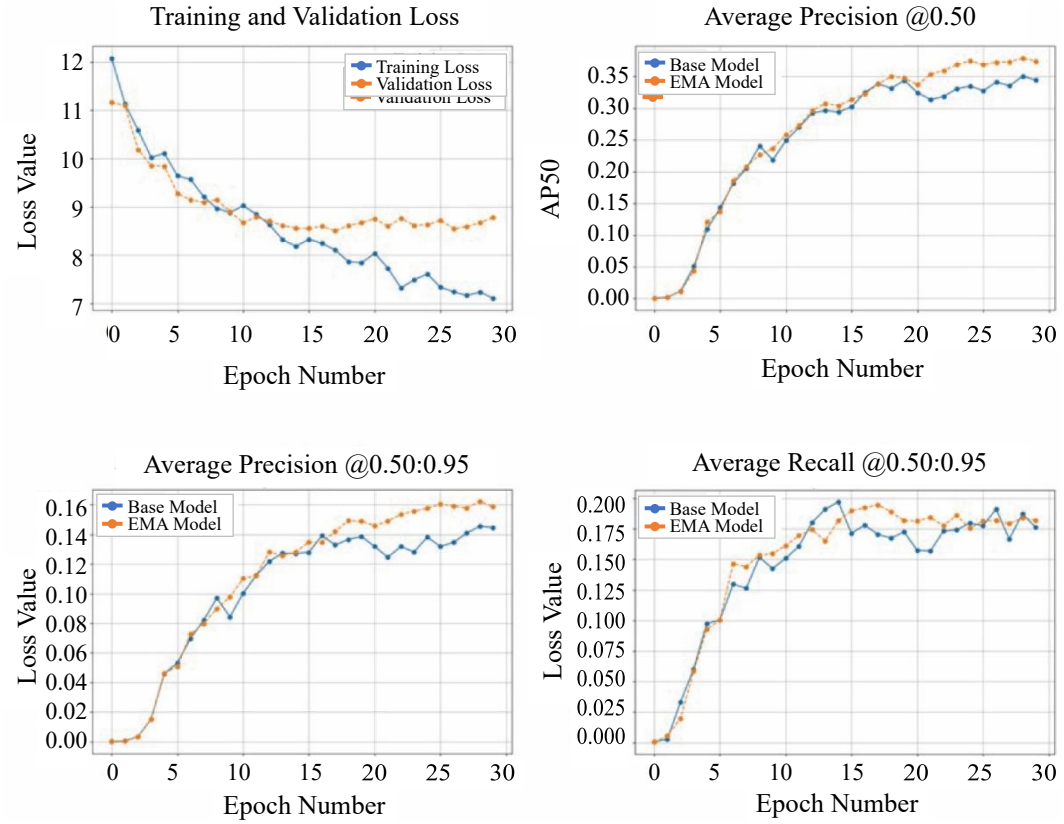

Figure 16. DETR training and performance evaluation of the multimodal detection model.

The graphs illustrate the training and validation loss curves, along with evaluation metrics for the detection model enhanced with Exponential Moving Average (EMA). Compared to the base model, the EMA-enhanced model demonstrates improved stability and accuracy across Average Precision (AP) at 0.50, Average Precision from 0.50 to 0.95, and Average Recall from 0.50 to 0.95 over 30 epochs

3.3. Detection-based Multimodal Performance Results

To overcome the limitations observed in the previous classification-based multimodal approach, this study maintained the single diagnostic model using clinical examination while redesigning the single diagnostic model component using periapical radiographic images from a classification-based diagnostic model to a detection-based model, followed by fusion analysis.

The detection-based radiographic single diagnostic model showed no significant performance differences compared to the existing classification-based model when evaluated independently. However, the multimodal approach fusing with clinical examination results demonstrated performance improvements across all diagnostic areas compared to previous multimodal results: dental caries (AUC: 0.82 → 0.88), tooth fracture (AUC: 0.74 → 0.84), and pulpitis (AUC: 0.86 → 0.90). Notably, tooth fracture diagnosis showed the greatest performance improvement (△AUC = 0.10), representing a relatively higher improvement margin compared to other diseases. The AUC performance graph results of this detection-based multimodal diagnostic model can be seen in **Figure 17.**

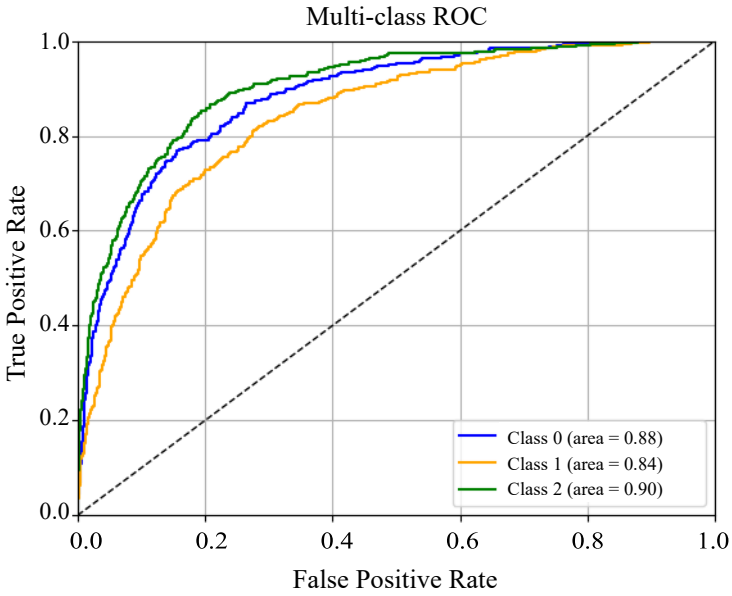

Figure 17. Multi-class ROC curve performance of the multimodal detection model for dental disease diagnosis.

The ROC curve illustrates the diagnostic performance for each class using the fused model. AUC values: Dental caries = 0.88, Tooth fracture = 0.84, Pulpitis = 0.90. The detection model demonstrates consistently high diagnostic capabilities across all dental pathologies, with superior performance compared to classification approaches through enhanced spatial feature integration of clinical examination and radiographic data.

4. DISCUSSION

4.1. Summary of Results

In this study, we developed and analyzed the performance of a multimodal artificial intelligence-based diagnostic model using clinical examination data and periapical radiographic images. The multimodal AI diagnostic model showed results that exceeded the performance of each single-modal AI, and we confirmed the tendency that as the performance of each single-modal improved, the multimodal performance also improved accordingly. This suggests that multimodal approaches can be usefully applied in dental diagnostic assistant systems.

In the single-modal model analysis, the clinical examination model showed rapid performance improvement as data processing became more sophisticated, while the periapical radiographic image model showed relatively limited performance improvement. This is analyzed to be due to the lack of complementarity between periapical radiographic image information and clinical examination information, and the increased complexity in the classification process, which limited performance improvement in multimodal fusion.

To overcome the limitations of the previous classification-based multimodal approach, this study maintained the clinical examination model while redesigning the radiographic image component from a classification-based to a detection-based model. The detection-based approach enabled more accurate spatial mapping between clinical findings and radiographic lesions, significantly improving diagnostic accuracy especially when structural anatomical information and clinical symptoms needed to be correlated.

The detection-based radiographic single diagnostic model showed no significant performance differences compared to the existing classification-based model when evaluated independently. However, in the multimodal approach fused with clinical examination results, performance improvements were observed across all diagnostic areas including dental caries, tooth fracture, and pulpitis compared to previous results. This suggests that the detection-based framework enabled more effective feature extraction and integration in multimodal fusion.

Particularly noteworthy was the substantial performance improvement in tooth fracture diagnosis. This is attributed to the stronger correlation between patient symptoms and lesions compared to other diseases, and the fact that primary disease factors more distinctly overwhelm secondary disease factors. Tooth fracture requires simultaneous confirmation of both patient clinical symptoms and structural features in radiographic images for accurate diagnosis, making it a case where modalities are truly complementary.

Clinical examination data is primarily used to diagnose pulp vitality, while radiographic images represent structural features, making them advantageous for dental caries and fracture detection. These findings confirmed that the types of modalities required differ for specific diagnoses, and that optimal modalities vary by disease type. This suggests that future diagnostic AI systems should adopt disease-specific multimodal strategies rather than applying uniform approaches across all conditions.

4.2. Clinical Examination Single Model

4.2.1 Correlation Between Variables

In this study, a correlation matrix based on Pearson correlation coefficients was computed and visualized in Figure 18 to systematically evaluate the relationships between variables. Pearson correlation coefficients numerically represent the strength and direction of linear associations between continuous variables, and statistical validity was ensured by securing a sufficiently large sample size of over 700.

The visualization of the correlation matrix in **Figure 18** revealed that most variable pairs exhibited weak correlations, with absolute values of the coefficients falling below 0.3 ($|r| < 0.3$), indicating that multicollinearity among variables was not a significant concern. However, moderate correlations were observed between several items that likely reflect similar clinical conditions. Notably, positive correlations were found between 'Mobility' and 'Percussion', as well as between 'CC1_Pain' and 'CC1_Sensitivity', suggesting that these variables may be capturing overlapping clinical features such as periodontal ligament responses or pain sensitivity.

To examine these patterns in more detail, a separate correlation analysis was conducted among clinical symptom variables (Air, Bite, Cold, EPT, Hot, Mobility, Percussion), and the results were visualized in **Figure 19(a)**. In this focused matrix, the strongest correlation was observed between 'Mobility' and 'Percussion', both of which are known to reflect periodontal ligament responses. Additionally, 'Cold' exhibited moderate correlations with these variables, implying a potential overlap between pulpal and periodontal pain responses, and suggesting a clinically meaningful transition zone between these types of stimuli.

Figure 19(b) presents the correlation matrix among CC1 annotation variables (CC1_Pain, CC1_No Symptoms, CC1_Sensitivity, CC1_Gum Swelling, CC1_Mobility, CC1_Unknown). Interestingly, a relatively high positive correlation ($r = 0.58$) was observed between 'CC1_Pain' and 'CC1_No Symptoms', despite their conceptually opposing meanings. This result implies potential inconsistencies in the questionnaire design or data entry, warranting a review of the annotation scheme. Furthermore, 'CC1_Pain', 'CC1_Sensitivity', and 'CC1_Gum Swelling' were found to share moderately high correlations, indicating a likelihood of redundant information. For subsequent modeling, it is recommended to consider variable consolidation or dimensionality reduction techniques to mitigate redundancy.

Additionally, the age variable showed no notable correlation with most other variables, suggesting that it possesses independent explanatory power. Regarding the gender variable, the dummy-encoded 'gender_f' and 'gender_m' displayed perfect negative correlation, and were therefore merged and treated as a single binary variable ('gender'). Collectively, these findings suggest that the variables used in the analysis maintain a generally independent structure, minimizing the risk of interpretive distortion due to overlapping information and thus enhancing the reliability of the predictive modeling.

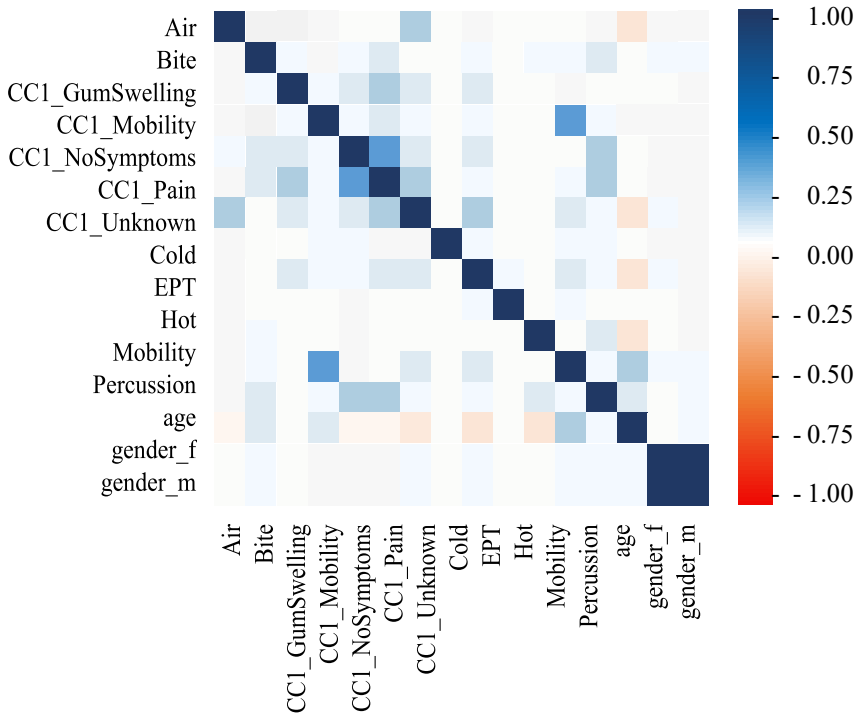

Figure 18. Pearson correlation matrix of clinical examination features.

The figure presents the correlation coefficients among various clinical test variables, chief complaints (CC1), demographic features (age, gender), and diagnostic tests (e.g., Air, Bite, Cold, EPT, Hot, Mobility, Percussion). Positive correlations are shown in blue and negative correlations in red, with stronger relationships appearing closer to ±1 on the color scale. Gender and age variables were one-hot encoded for inclusion in the analysis.

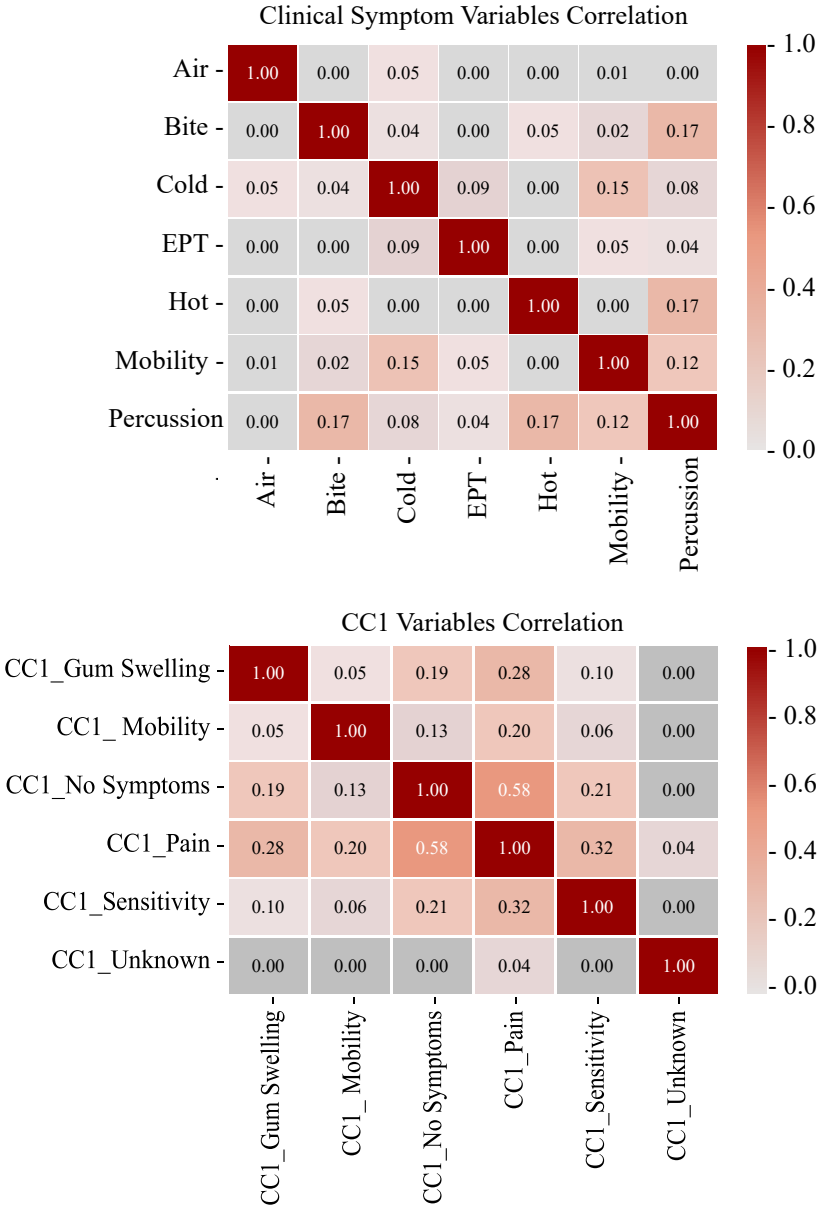

Figure 19. Subgroup Pearson correlation matrix (|r|) of clinical examination features.

(a) Correlation heatmap of clinical examination tests (b) Correlation heatmap of CC1 annotation variables, including pain, sensitivity, gum swelling, mobility, and absence of symptoms.

4.2.2 Feature Processing

The handling of missing values according to the retrospective data collection method was a factor that significantly affected model performance. We compared and analyzed various methodologies for handling missing values in categorical variables, including 'creating a separate category for missing values', 'random substitution', and 'mode substitution'. The method of setting missing values as an independent category called 'missing' had advantages in terms of preserving the information of the original data, but there was a possibility that the model might learn unnecessary patterns. The random substitution method contributed to maintaining the overall data distribution but showed limitations in terms of result reproducibility.

Missing data recorded as 'no examination' was generally marked as "-" or left blank, and based on repeated experiments with various substitution methodologies, the mode substitution method was found to produce the most stable model performance. Therefore, in this study, we consistently applied the mode substitution method to all categorical variables.

The complexity of missing value handling and methodological constraints demonstrate the fundamental limitations of retrospective study design. Since data quality is paramount for effective artificial intelligence model training, future research requires prospective study designs that are well-designed from the outset to maintain controllable conditions for various factors and ensure reliability. This approach would minimize missing value occurrence and secure more accurate and consistent data, thereby enabling further improvement in the performance of artificial intelligence diagnostic models.

4.2.3 Handling Data Imbalance

To solve the class imbalance problem commonly encountered in medical research, we systematically applied Stratified K-Fold Cross Validation during the model evaluation and tuning process. This methodology minimizes learning bias due to class imbalance and enables more reliable evaluation of the generalizability of tuned parameter combinations by dividing samples so that the class distribution within each fold is maintained the same as the distribution of the entire dataset.

During the model tuning process, we maintained consistent data split conditions and random seeds to perform repeated experiments, securing the stability of optimal combination selection and consistency of performance. Each experiment was repeated several times under fixed random state conditions, and performance was recorded in terms of mean and standard deviation, followed by additional verification through seed changes.

As a result of the optimization process, the Random Forest model recorded the highest AUC figures compared to other algorithms for three major lesions, showing superior classification

performance. This is interpreted as the ensemble structure of Random Forest effectively capturing complex interactions between various features, and the optimized hyperparameter combination through the tuning process producing synergistic effects. The tuning strategy and evaluation methodology established in this study are expected to be effectively applied to similar medical data analyses or complex multiclass classification problems in the future.

4.3. Radiographic Single Model

4.3.1 Introduction to Existing Papers

Previous research on lesion detection and diagnosis in periapical radiographic images has already been extensively accumulated. Since research on dental caries is overwhelmingly prevalent, we will compare the diagnostic performance with previous studies in this area.

Artificial intelligence technology is showing particularly prominent achievements in the field of dental caries detection. CNN-based models have shown high performance in detecting dental caries in periapical radiographic images, and models using GoogLeNet Inception v3 have presented the possibility of effectively assisting clinical diagnosis in molars (AUC 0.89) and premolars (AUC 0.92) (Lee et al., 2018). The utilization of artificial intelligence is also expanding in the areas of dental caries and periodontal disease detection. Research applying explainable AI techniques such as Grad-CAM to visually clearly present dental caries areas (Oztekin et al., 2023) and the CariesNet model that segments and detects multistage dental caries with high accuracy (Zhu et al., 2022) have been developed.

Additionally, research using YOLO-based CNN models has developed technology to detect interproximal dental caries in digital bitewing images, recording performance similar to expert readings (Bayraktar & Ayan, 2022). A systematic review of deep learning research for dental caries detection found that CNN-based models achieved up to 86% accuracy and 76% sensitivity (Szabó et al., 2024) and research using YOLOv7-based object detection technology showed that the EfficientNet-B0 classification model achieved improved performance in dental caries identification with AUC 98.31% (Chen et al., 2023).

In this study, the periapical radiographic single diagnostic model for detecting dental caries achieved a performance of AUC 0.88, and the multimodal approach also reached the same performance of AUC 0.88. This represents a competitive level compared to previous studies, but shows relatively lower performance compared to the results of Chen et al. (2023).

This performance difference can be attributed to several factors. Unlike previous studies, this research had limitations inherent to multi-diagnostic models that are not specialized for single diseases, and most importantly, image quality issues and non-uniformity of radiographic conditions

in periapical radiographic images are considered to have acted as major constraining factors. Additionally, inconsistency in image quality due to retrospective data collection and potential labeling accuracy issues may have also affected performance.

4.3.2 The Problem of Processing as Diseases

The biggest challenge faced during the development of models using radiographic images was mapping lesions appearing in the images to accurate diseases. Dental diseases inherently show complex patterns, and multiple lesions frequently coexist within a single image. Due to this complexity, there are many situations where it is difficult to derive an accurate diagnosis based solely on radiographic images.

Moreover, even the same lesion can be assigned different diagnoses depending on the patient's subjective symptoms, the progression of the lesion, and the professional judgment of the clinician. This inherent complexity of the diagnostic process led to difficulties in clear label assignment during artificial intelligence learning, affecting the learning accuracy of the model.

In this study, to solve this problem, we initially attempted an approach of classifying the entire image into a specific disease category but faced fundamental limitations due to the complexity factors mentioned earlier. Therefore, we changed the research direction and redesigned the approach methodology to an object recognition (detection) method that detects individual lesions themselves, enabling more precise lesion recognition and model development including location information.

4.3.3 Labeling

The labeling process acted as a key element directly affecting model performance. In the labeling process of this study, it was observed that as artificial intelligence models became more sophisticated, even micro lesions that are difficult for humans to visually identify were recognized. Among these, cases that were impossible to read with the naked eye of human experts were classified as 'no label' using a conservative approach (Gliga et al., 2023).

The reason for choosing bounding box labeling was that the focus was on discovering lesions with diagnostic value rather than precise segmentation of lesions within periapical images. This approach enabled labeling with relatively relaxed criteria.

To increase the reliability of labeling, all labels underwent a cross-check process by two specialists, and in cases where interpretations did not match, they were classified as 'no label' according to the conservative principle to maintain the strictness of judgment.

To analyze the size distribution of actual labeled lesions, the width and height of each lesion's bounding box were extracted and visualized as a histogram. The analysis revealed that most lesions were distributed in the range of 200-300 pixels, with some lesions corresponding to an even wider range. Additionally, the aspect ratio of lesions (width versus height) was also deeply analyzed in the form of a cumulative distribution.

This statistical analysis provided important criteria for practical modeling parameter decisions such as optimizing anchor box sizes and setting augmentation ranges in designing detection models. Furthermore, by systematically understanding the statistical characteristics of the size and shape of lesions, we established a model learning foundation that could encompass various clinical conditions, which ultimately made a crucial contribution to improving the model's interpretability and generalization performance.

4.4. Technical Issues

During this research process, we faced various technical challenges and gained several valuable insights in the process of solving them.

The Random Forest model showed the best performance in terms of consistency and accuracy. In particular, we confirmed not only the improvement in classification accuracy for a single lesion type of dental caries but also the possibility of extension to other lesions such as tooth fracture and pulpitis. These results suggest that the ensemble structure of Random Forest can effectively capture complex patterns and characteristics of clinical data.

MLP (Multi-Layer Perceptron) based models, especially when applied with Focal Loss, showed balanced performance improvement across all classes and effective response to multiple lesion predictions. This empirically proves the effectiveness of Focal Loss in medical data analysis where class imbalance is prominent.

In the Autoencoder-based approach, we adopted a strategy of learning latent representations through unsupervised learning and transferring them as feature extractors to be used as classification models. By combining Autoencoder-based pre-training and transfer learning, we aimed to improve performance, and this approach was particularly effective in situations with limited labels.

The generalization performance of the radiographic model showed significant differences depending on the intensity of augmentation, backbone structure, and whether pre-training was applied. Interestingly, excessive augmentation was observed to lead to performance degradation, a paradoxical phenomenon. This result emphasizes the importance of setting appropriate intensity and range when establishing a data augmentation strategy.

In multimodal model experiments, two fusion models (concatenation-based and attention-based) showed similar or slightly improved performance compared to single modality models under the same dataset and experimental conditions. Particularly noteworthy is that the structurally complex Attention-based model, contrary to expectations, did not show significant performance improvement. This can be interpreted as the complementarity between clinical information and image data being lower than expected.

Additionally, we confirmed that in complex situations where multiple lesions simultaneously exist within an X-ray image, there is a fundamental limitation with just single or multi-label classification methods. To overcome this limitation, we shifted the paradigm to an object detection-based approach, through which the model not only showed high detection performance in the validation dataset but also impressively identified lesions that were missed in the initial labeling process.

4.5. Multi-modal Model

4.5.1 Comparison with Existing Research

While multimodal approaches in the field of dentistry are still in their early stages, some previous studies have suggested their potential. In particular, research focusing on tooth identification is advancing rapidly, with automatic tooth numbering technologies powered by deep learning showing substantial progression. Previous studies have also reported that when accurate tooth position information is provided in periapical radiographic images, the performance of AI models significantly improves. This study also confirmed a trend consistent with this, and it is anticipated that the development of more precise diagnostic models will be possible through the combination of tooth number automatic recognition technology and lesion detection technology in the future.

In a previous study on a multimodal deep learning model for dental caries prediction (MMDCP), a hybrid model based on CNN and artificial neural networks (ANN) was constructed by integrating radiographic images and clinical data, achieving high accuracy (accuracy 90%, F1-score 89%) compared to single-modal models (Ngnamsie Njimbouom et al., 2022). Additionally, in a study on multimodal deep learning models that automatically combine periapical radiographic images, a ResNet-based model utilizing 4,707 radiographic images and time information recorded superior accuracy compared to single-modal approaches (Pfänder et al., 2023).

However, there are several important obstacles to multimodal performance improvement. First, in this study, the matching process of clinical examination and periapical radiographic image data was based on the specific tooth's notation number. However, in periapical radiographs, lesions are frequently distributed across multiple teeth rather than being confined to a single tooth. Therefore,

if tooth identification is not accurate, confusion can arise in model learning. To solve this problem, we systematically performed lesion and tooth labeling, resulting in significantly improved model performance, suggesting that accurate identification of tooth numbers has a significant impact on model accuracy.

Second, during the AI model training process, we encountered challenges with the inconsistent classification of diagnosis labels, disease terminology, and treatment outcomes, further complicating class differentiation. Dental diagnoses inherently involve multiple concurrent conditions rather than singular causes in the majority of cases. For instance, patients frequently present with simultaneous dental caries and tooth fractures, or concurrent tooth fractures and pulpitis. To accurately model these complex clinical scenarios, we developed an approach that precisely differentiates between primary and secondary conditions while systematically weighting the contribution of each diagnosis to the overall assessment.

4.5.2 Limits of Classification-based Multimodal Fusion

In this study, we designed our experiments based on the hypothesis that a multimodal diagnostic model utilizing both clinical examination information and periapical radiographic images would demonstrate superior performance compared to models based on single modalities alone. However, as can be observed in Table 6, the actual experimental results showed patterns that differed somewhat from these initial expectations.

First, among the single-modal models using clinical data, the pretrained model recorded high performance with an AUC of 0.87 for dental caries, which was either higher than or comparable to most results from multimodal models. Similarly, X-ray image single-modal models also showed performance reaching an AUC of 0.88 when applying U-Net or MAE structures, suggesting that sufficient diagnostic accuracy could be achieved using independent modalities alone.

In contrast, multimodal models utilizing both clinical information and radiographic images showed improved performance under certain conditions but did not consistently demonstrate significant performance improvements compared to single-modal approaches. For example, the attention-based multimodal model recorded an AUC of 0.82 for dental caries, which was actually lower than the results of the single-modal pretrained model. Additionally, there were substantial performance variations depending on the fusion method, and whether pretraining was applied also served as an important factor influencing the results.

These findings suggest that multimodal structures do not always yield superior results compared to single-modal models. One major cause is the lack of representational alignment between the two modalities. In this study, clinical information and image data were matched based on specific tooth numbers; however, due to the nature of periapical radiographic images where lesions often span

multiple teeth, there were limitations in securing accurate tooth-by-tooth matching. For example, even when clinical symptoms existed for tooth #16, radiographic images showed lesions spanning the #15-17 tooth region, making it difficult to establish accurate correspondence relationships.

Furthermore, classification-based multimodal structures had structural limitations in that they only considered global features and failed to effectively utilize local lesion information. The classification approach resulted in the loss of spatial information and precise lesion location data during the process of classifying the entire image into a single category, making precise mapping between clinical findings and radiographic lesions difficult. Additionally, the information obtained from clinical examinations and radiographic images showed redundancy or low correlation for certain diseases, limiting the benefits of multimodal fusion..

4.5.3 Breakthrough with a Detection-based Approach

To overcome the limitations of the classification-based approach, this study introduced a paradigm shift by adopting an object detection-based model. This was a strategic approach to address the problems that became more pronounced limitations in complex clinical situations where multiple lesions exist simultaneously. The detection-based approach could fundamentally resolve spatial information loss issues by simultaneously identifying the location and type of individual lesions instead of classifying the entire image into a single category.

To overcome the aforementioned obstacles to performance improvement and enhance the practical utility of multimodal diagnosis, this study transitioned the artificial intelligence learning goal to the form of object detection. Initial model learning was conducted based on an X-ray image dataset manually labeled by experts, and annotations including the location and class information of each lesion were configured in COCO format. The trained DETR (DEtection TRansformer) model not only showed excellent detection performance in the validation dataset but also demonstrated notable results by accurately predicting lesions that were missed in the initial labeling process in some images. This was an important discovery suggesting the model's potential to complement the limitations of existing labeling.

Based on these results, we designed a cyclical validation system where experts review the prediction results of the DETR model. Specifically, experts conducted additional reviews of high confidence score detection results among the bounding boxes detected by the model, and in this process, numerous lesions overlooked in the initial labeling were identified. Based on this, we introduced an iterative improvement strategy of complementing and modifying existing labels and retraining (fine-tuning) the model with datasets including these improved labels.

The object detection approach significantly improved spatial mapping between clinical findings and radiographic lesions by providing precise location information of lesions. Direct correlation analysis between 'lesions at specific locations' and 'clinical symptoms in corresponding areas'

became possible, which was impossible with previous classification-based models, greatly enhancing the effectiveness of multimodal fusion. As the matching between lesion areas defined by bounding boxes and tooth number-based clinical information became much more precise, complementarity between the two modalities could be practically implemented.

Particularly noteworthy is that the detection-based approach substantially resolved the tooth number matching problem. By enabling precise localization of lesions through bounding boxes, matching errors such as '16th tooth symptoms vs. lesions in 15-17th tooth region' that were problematic previously could be significantly reduced. By implementing an algorithm that automatically identifies corresponding teeth based on the center coordinates and range of detected lesions, the alignment accuracy between clinical data and image data was markedly improved.

The performance improvement in tooth fracture diagnosis was particularly prominent in the detection-based approach, as fractures involve clear structural changes compared to other diseases, making location information crucial for diagnosis. The direct correlation between fracture sites specified by bounding boxes and clinical symptoms such as pain or sensitivity in corresponding areas significantly contributed to model performance improvement. Additionally, in the case of fractures, the concordance between patients' subjective symptoms and objective lesion locations was high, allowing the two information sources to truly complement each other in multimodal fusion.

These results demonstrate that it is possible to analyze the contribution of radiographic and clinical information for each disease and design multimodal structures optimized for each lesion type. The spatial information obtained through the detection-based approach can serve as the foundation for developing more sophisticated multimodal fusion strategies in the future, which is expected to further enhance the practicality and accuracy of dental diagnostic AI systems.

4.5.4. Clinical Implications, Study Contributions and Future Directions

This study presents several important contributions that significantly expand the applicability of artificial intelligence technology in the field of dental diagnosis.

From a methodological perspective, we empirically demonstrated the effectiveness of self-supervised learning and transfer learning, showing that high-performance diagnostic models can be developed even in environments with limited labeled data. In particular, the approach combining Autoencoder-based pre-training with transfer learning provided an effective solution in medical data environments where labels are limited, suggesting the potential for expansion to other medical imaging diagnostic fields.

From a clinical standpoint, the most noteworthy finding is that substantial diagnostic accuracy can be achieved using clinical examination data alone. This presents the possibility of being utilized as an effective diagnostic assistance tool even in environments where radiographic imaging is difficult or in emergency situations. Additionally, the discovery that optimal modality combinations

differ by disease type can serve as the foundation for developing personalized diagnostic strategies in the future.

The paradigm shift to a detection-based approach significantly improved diagnostic accuracy in complex clinical situations and enabled the true effectiveness of multimodal fusion, particularly for diseases where location information is critical, such as tooth fractures. This approach established a solid foundation for developing accurate and efficient diagnostic tools and enhanced the potential for practical application in clinical settings.

However, the retrospective design of this study still poses limitations, and future research needs to ensure data quality and consistency through well-designed prospective studies from the outset. This would minimize the occurrence of missing values and enable the development of more reliable diagnostic models.

Future research directions should include the development of more sophisticated fusion architectures such as cross-modal transformers or modality alignment techniques, advancement of tooth identification-based alignment algorithms, and the design of multimodal architectures optimized for each disease type. In particular, the development of lightweight models capable of real-time diagnosis and the design of user interfaces that can be seamlessly integrated into clinical workflows are important challenges.

Ultimately, this study has established a technical foundation that can further enhance the practicality and accuracy of dental diagnostic AI systems, and it is expected to contribute to complementing the diagnostic capabilities of dental professionals and improving patient treatment outcomes.

5. CONCLUSION

The utilization of artificial intelligence (AI) is rapidly expanding across all industrial sectors, including healthcare, and the field of dentistry is no exception to this transformative trend. AI technology particularly excels in processing vast amounts of data and learning patterns, raising expectations for its supportive role in diagnostic processes. This is emerging as a key factor that can enhance the efficiency and accuracy of dental practice and is considered one of the important technologies shaping the future direction of dentistry. However, while AI technology is advancing at a remarkable pace, its application and adaptation in dentistry are relatively slow, necessitating strategic implementation tailored to the dental context rather than simply following technological trends.

Diagnosis and treatment form the core of clinical decision-making in dentistry, with 'diagnosis' being a representative area where AI technology can make substantial contributions. In particular, AI-based diagnostic assistance systems that integrate various patient data to enable more precise diagnoses are expected to become a fundamental component of digital dentistry in the future. However, to build AI systems applicable in clinical settings, using single-modal data alone is insufficient to adequately reflect complex diagnostic situations. Consequently, there is growing interest in multimodal approaches that comprehensively utilize different types of data such as images, clinical records, and patient interview information.

Developing multimodal AI systems presents various technical and practical challenges. Since different types of data must be learned in parallel to derive the same diagnostic results, the learning conditions are much more demanding than single-modal models, and data composition becomes more complex. In this study, to address these challenges, we first developed individual AI models based on different single-modal data and then implemented a multimodal diagnostic assistance system by effectively combining them. As a result, multimodal models achieved performance improvements across all diseases through detection-based approaches, but did not show consistent performance improvements compared to single-modal models in classification-based approaches. While the true effectiveness of multimodal fusion was confirmed particularly in tooth fracture diagnosis, the complementarity between modalities was limited in dental caries and pulpitis. Through these results, we confirmed the necessity of disease-specific optimal modality strategies while specifically identifying key challenges of multimodal approaches, including data alignment issues and limitations of fusion architectures.

Several important prerequisites must be met for successful clinical utilization of multimodal AI systems. First, to overcome the limitations of retrospective data collection, it is essential to construct high-quality multimodal datasets through systematic and purpose-oriented prospective research designs. Second, optimization of each single-modal AI must be prioritized, as this directly affects the overall performance of the final fusion model. Third, as confirmed in the paradigm shift from classification-based to detection-based approaches, appropriate AI architecture selection matching disease characteristics is crucial. Fourth, improvement in the accuracy of tooth number-based data

matching and development of precise modality alignment techniques utilizing spatial information are necessary.

This study has simultaneously illuminated both the potential and realistic limitations of multimodal AI approaches in the field of dental diagnosis. We demonstrated that effective multimodal fusion is possible even in complex clinical situations through the introduction of detection-based models, achieving substantial performance improvements particularly in diseases where location information is critical. However, we also confirmed that the complementarity between modalities varies by disease type, necessitating customized strategies rather than uniform approaches. Based on these research findings, continuous research should be conducted on establishing prospective data collection frameworks, developing advanced modality alignment algorithms, and designing disease-optimized multimodal architectures. Through these efforts, we ultimately expect to complete practical AI-based dental diagnostic assistance systems that can substantially complement the diagnostic capabilities of dental professionals and improve patient treatment outcomes.

References

Albano, D., Galiano, V., Basile, M., Di Luca, F., Gitto, S., Messina, C., Cagetti, M. G., Del Fabbro, M., Tartaglia, G. M., & Sconfienza, L. M. (2024). Artificial intelligence for radiographic imaging detection of caries lesions: a systematic review. *BMC Oral Health*, *24*(1), Article 274. https://doi.org/10.1186/s12903-024-04046-7

Alberdi, A., Aztiria, A., & Basarab, A. (2016). On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey. *Artificial intelligence in medicine*, *71*, 1-29. https://doi.org/10.1016/j.artmed.2016.06.003

Ayhan, B., Ayan, E., & Bayraktar, Y. (2024). A novel deep learning-based perspective for tooth numbering and caries detection. *Clinical Oral Investigations*, *28*(3), Article 178. https://doi.org/10.1007/s00784-024-05566-w

Başaran, M., Çelik, Ö., Bayrakdar, I. S., Bilgir, E., Orhan, K., Odabaş, A., Aslan, A. F., & Jagtap, R. (2022). Diagnostic charting of panoramic radiography using deep-learning artificial intelligence system. *Oral Radiology*, *38*(3), 363-369. https://doi.org/10.1007/s11282-021-00572-0

Bayraktar, Y., & Ayan, E. (2022). Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. *Clinical Oral Investigations*, *26*(1), 623-632. https://doi.org/10.1007/s00784-021-04040-1

Brunsvold, M. A., Nair, P., & Oates, T. W., Jr. (1999). Chief complaints of patients seeking treatment for periodontitis. *Journal of the American Dental Association*, *130*(3), 359-364. https://doi.org/10.14219/jada.archive.1999.0205

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: fast and flexible image augmentations. *Information*, *11*(2), 125.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*, 41-75. https://doi.org/10.48550/arXiv.1707.08114

Çelik, B., & Çelik, M. E. (2022). Automated detection of dental restorations using deep learning on panoramic radiographs. *Dentomaxillofacial Radiology*, *51*(8), Article 20220244.

Cha, J. Y., Yoon, H. I., Yeo, I. S., Huh, K. H., & Han, J. S. (2021). Panoptic Segmentation on Panoramic Radiographs: Deep Learning-Based Segmentation of Various Structures Including Maxillary Sinus and Mandibular Canal. *Journal of Clinical Medicine*, *10*(12), Article 2577. https://doi.org/10.3390/jcm10122577

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, I. D. S., Yang, C.-M., Chen, M.-J., Chen, M.-C., Weng, R.-M., & Yeh, C.-H. (2023). Deep learning-based recognition of periodontitis and dental caries in dental x-ray images. *Bioengineering*, *10*(8), Article 911.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., & Ouyang, W. (2019). Hybrid task cascade for instance segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Choi, E., Kim, D., Lee, J. Y., & Park, H. K. (2021). Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram. *Scientific Reports*, *11*(1), Article 10246. https://doi.org/10.1038/s41598-021-89742-y

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words:

Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

Endres, M. G., Hillen, F., Salloumis, M., Sedaghat, A. R., Niehues, S. M., Quatela, O., Hanken, H., Smeets, R., Beck-Broichsitter, B., Rendenbach, C., Lakhani, K., Heiland, M., & Gaudin, R. A. (2020). Development of a Deep Learning Algorithm for Periapical Disease Detection in Dental Radiographs. *Diagnostics (Basel)*, *10*(6), Article 430. https://doi.org/10.3390/diagnostics10060430

Gliga, A., Imre, M., Grandini, S., Marruganti, C., Gaeta, C., Bodnar, D., Dimitriu, B. A., & Foschi, F. (2023). The Limitations of Periapical X-ray Assessment in Endodontic Diagnosis-A Systematic Review. *Journal of Clinical Medicine*, *12*(14). https://doi.org/10.3390/jcm12144647

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. IEEE/CVF Conference on Computer Vision and Pattern Recognition,

He, X., Deng, Y., Fang, L., & Peng, Q. (2021). Multi-modal retinal image classification with modality-specific attention network. *IEEE transactions on medical imaging*, *40*(6), 1591-1602.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., & Wei, Y. (2021). Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, *30*, 5875-5888.

Karaoglu, A., Ozcan, C., Pekince, A., & Yasa, Y. (2023). Numbering teeth in panoramic images: A novel method based on deep learning and heuristic algorithm. *Engineering Science and Technology, an International Journal*, *37*, 101316.

Khan, H. A., Haider, M. A., Ansari, H. A., Ishaq, H., Kiyani, A., Sohail, K., Muhammad, M., & Khurram, S. A. (2021). Automated feature detection in dental periapical radiographs by using deep learning. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology*, *131*(6), 711-720. https://doi.org/10.1016/j.oooo.2020.08.024

Krois, J., Ekert, T., Meinhold, L., Golla, T., Kharbot, B., Wittemeier, A., Dörfer, C., & Schwendicke, F. (2019). Deep Learning for the Radiographic Detection of Periodontal Bone Loss. *Scientific Reports*, *9*(1), 8495. https://doi.org/10.1038/s41598-019-44839-3

Küçükçiloğlu, Y., Şekeroğlu, B., Adalı, T., & Şentürk, N. (2024). Prediction of osteoporosis using MRI and CT scans with unimodal and multimodal deep-learning models. *Diagnostic and Interventional Radiology*, *30*(1), 9. https://doi.org/10.4274/dir.2023.232116

Kumar, S., Chaube, M. K., Alsamhi, S. H., Gupta, S. K., Guizani, M., Gravina, R., & Fortino, G. (2022). A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques. *Computer methods and programs in biomedicine*, *226*, 107109.

Lee, J. H., Han, S. S., Kim, Y. H., Lee, C., & Kim, I. (2020). Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology*, *129*(6), 635-642. https://doi.org/10.1016/j.oooo.2019.11.007

Lee, J. H., Kim, D. H., Jeong, S. N., & Choi, S. H. (2018). Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *Journal of dentistry*, *77*, 106-111. https://doi.org/10.1016/j.jdent.2018.07.015

Li, D., Lin, C. T., Sulam, J., & Yi, P. H. (2022). Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms. *Emergency Radiology*, *29*(2), 365-370.

Li, Y., Huang, Z., Dong, X., Liang, W., Xue, H., Zhang, L., Zhang, Y., & Deng, Z. (2019). Forensic age estimation for pelvic X-ray images using deep learning. *European radiology*, *29*, 2322-

2329.
Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
Mainkar, A., & Kim, S. G. (2018). Diagnostic Accuracy of 5 Dental Pulp Tests: A Systematic Review and Meta-analysis. *Journal of Endodontics*, *44*(5), 694-702. https://doi.org/10.1016/j.joen.2018.01.021
Milošević, D., Vodanović, M., Galić, I., & Subašić, M. (2022). Automated estimation of chronological age from panoramic dental X-ray images using deep learning. *Expert systems with applications*, *189*, 116038.
Mohammad-Rahimi, H., Motamedian, S. R., Pirayesh, Z., Haiat, A., Zahedrozegar, S., Mahmoudinia, E., Rohban, M. H., Krois, J., Lee, J. H., & Schwendicke, F. (2022). Deep learning in periodontology and oral implantology: A scoping review. *Journal of Periodontal Research*, *57*(5), 942-951. https://doi.org/10.1111/jre.13037
Ngnamsie Njimbouom, S., Lee, K., & Kim, J.-D. (2022). MMDCP: Multi-modal dental caries prediction for decision support system using deep learning. *International Journal of Environmental Research and Public Health*, *19*(17), 10928.
Oztekin, F., Katar, O., Sadak, F., Yildirim, M., Cakar, H., Aydogan, M., Ozpolat, Z., Talo Yildirim, T., Yildirim, O., Faust, O., & Acharya, U. R. (2023). An Explainable Deep Learning Model to Prediction Dental Caries Using Panoramic Radiograph Images. *Diagnostics (Basel)*, *13*(2). https://doi.org/10.3390/diagnostics13020226
Park, J. H., Moon, H. S., Jung, H. I., Hwang, J., Choi, Y. H., & Kim, J. E. (2023). Deep learning and clustering approaches for dental implant size classification based on periapical radiographs. *Scientific Reports*, *13*(1), 16856. https://doi.org/10.1038/s41598-023-42385-7
Park, S.-J., Yang, S., Kim, J.-M., Kang, J.-H., Kim, J.-E., Huh, K.-H., Lee, S.-S., Yi, W.-J., & Heo, M.-S. (2024). Automatic and robust estimation of sex and chronological age from panoramic radiographs using a multi-task deep learning network: a study on a South Korean population. *International Journal of Legal Medicine*, *138*(4), 1741-1757.
Petersson, A., Axelsson, S., Davidson, T., Frisk, F., Hakeberg, M., Kvist, T., Norlund, A., Mejàre, I., Portenier, I., Sandberg, H., Tranaeus, S., & Bergenholtz, G. (2012). Radiological diagnosis of periapical bone tissue lesions in endodontics: a systematic review. *International Endodontic Journal*, *45*(9), 783-801. https://doi.org/10.1111/j.1365-2591.2012.02034.x
Pfänder, L., Schneider, L., Büttner, M., Krois, J., Meyer-Lückel, H., & Schwendicke, F. (2023). Multi-modal deep learning for automated assembly of periapical radiographs. *Journal of dentistry*, *135*, 104588.
Pitts, N. B. (2001). Clinical diagnosis of dental caries: a European perspective. *Journal of Dental Education*, *65*(10), 972-978. https://doi.org/10.1002/j.0022-0337.2001.65.10.tb03441.x
Raghu, V. K., Weiss, J., Hoffmann, U., Aerts, H. J., & Lu, M. T. (2021). Deep learning to estimate biological age from chest radiographs. *Cardiovascular Imaging*, *14*(11), 2226-2236.
Rahman, T., Chowdhury, M. E., Khandakar, A., Mahbub, Z. B., Hossain, M. S. A., Alhatou, A., Abdalla, E., Muthiyal, S., Islam, K. F., & Kashem, S. B. A. (2023). BIO-CXRNET: a robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest X-ray images and clinical data. *Neural Computing and Applications*, *35*(24), 17461-17483.
Ren, X., Li, T., Yang, X., Wang, S., Ahmad, S., Xiang, L., Stone, S. R., Li, L., Zhan, Y., & Shen, D. (2018). Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. *IEEE journal of biomedical and health informatics*, *23*(5), 2030-2038.

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017a). Efficient convnet for real-time semantic segmentation. 2017 IEEE Intelligent Vehicles Symposium (IV),

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017b). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, *19*(1), 263-272.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szabó, V., Szabó, B. T., Orhan, K., Veres, D. S., Manulis, D., Ezhov, M., & Sanders, A. (2024). Validation of artificial intelligence application for dental caries diagnosis on intraoral bitewing and periapical radiographs. *Journal of dentistry*, *147*, 105105. https://doi.org/10.1016/j.jdent.2024.105105

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. International conference on machine learning,

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242-264). IGI global.

Wang, Y., Tang, S., Ma, R., Zamit, I., Wei, Y., & Pan, Y. (2022). Multi-modal intermediate integrative methods in neuropsychiatric disorders: A review. *Computational and Structural Biotechnology Journal*, *20*, 6149-6162.

Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., & Jantana, P. (2021). Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *Journal of Oral Pathology & Medicine*, *50*(9), 911-918.

Xue, Z., Rajaraman, S., Long, R., Antani, S., & Thoma, G. (2018). Gender detection from spine x-ray images using deep learning. 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS),

Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., & Wu, J. (2022). CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image. *Neural Computing and Applications*, 1-9. https://doi.org/10.1007/s00521-021-06684-2

Abstract in Korean

다중 모달 치과 진단 보조를 위한 방사선 및 임상 데이터의 자가지도 학습 기반 융합 평가

서론

정확한 치과 진단은 환자 병력, 임상 검사, 방사선 영상을 포함한 다양한 데이터를 종합하여 이루어집니다. 때문에 실제 진단은 이러한 데이터를 종합하는 임상의의 경험에 크게 의존하게 되어 진단 정확도에 편차가 발생합니다. 인공지능 학습을 활용한 치과 진단 보조는 이러한 정확도 향상에 기여할 것으로 기대됩니다. 현재 치과에서도 다양한 인공지능의 응용이 나타나고 있으나, 아직 진단 분야의 인공지능 연구는 방사선 영상만을 사용하는 단일 양상 학습에 제한되어 있습니다. 본 연구는 사전 학습의 방법 중 하나인 자가지도학습을 통해 진단에 필요한 다양한 유형의 데이터를 활용하는 다중 양상 인공지능 모델을 개발함으로써 이러한 한계를 극복하는 것을 목표로 합니다. 본 연구의 목적은 다음과 같습니다. 첫째, 임상 검사 데이터를 사용한 단일 양상 인공지능 진단 모델을 개발하고, 둘째, 치근단 방사선 영상을 사용한 단일 양상 인공지능 진단 모델을 개발하며, 셋째, 두 모델을 결합한 다중 양상 인공지능 모델을 개발하여 자가지도학습 기법을 활용해 세 모델 간의 진단 성능을 비교하는 것입니다.

본론

인공지능 모델 개발을 위해 연세대학교 치과대학병원을 방문한 1,344 명 환자의 3,341 개 임상 데이터를 활용했습니다. 선별 과정을 통해 훈련에 적합한 치근단 방사선 사진과 일치하는 705 개의 임상 데이터를 선별했습니다. 임상 검사를 사용한 단일 양상 인공지능 진단 모델을 개발하기 위해 진료 기록부에서 데이터를 추출했습니다. 데이터에는 범주화 된 환자의 주소, 성별, 연령이 기본 정보로 포함되었고, 일반적으로 단일 치아의 진단에 사용되는 7 가지 임상 검사인 타진, 동요도, 교합, 공기자극, 냉자극, 온자극, 전기 치수 검사가 포함되었습니다. 효율적인 학습을 유도하기 위해 자가지도학습 기법을 적용했으며, 정확도 향상 과정에서 임상 검사 유형을 선별하고 변수 처리시 결측 데이터는 최빈값으로 대체했습니다.

치근단 방사선 영상을 사용한 인공지능 진단 모델을 만들기 위해 705 장의 치근단 방사선 사진을 사용했습니다. 이 방사선 사진들은 이전 모델에서 사용된 임상 검사의 진단 시점과 일치합니다. 효율적인 학습을 유도하기 위해 영상에서의

자가지도학습 기법인 마스킹 오토인코더 (Masked Autoencoder, MAE)를 적용했습니다. 정확도 향상을 위해 치근단 방사선 사진의 병변과 특징점을 탐지 형식으로 주석을 달고 오류를 줄이기 위한 최적화를 수행했습니다.

임상 검사 데이터와 치근단 방사선 영상을 각각 사용한 단일 양상 인공지능 모델의 성능을 최대화한 후, 두 단일 양상 모델을 결합한 다중 양상 인공지능 모델을 구축했습니다. 이후 전체 오류를 줄이기 위한 최적화 과정을 진행했습니다. 모델 성능은 정확도와 정밀도, 혼동 행렬, 수신자 조작 특성 곡선(ROC) 분석 등의 목표 지표를 통해 평가되었으며, 각 구성 요소를 요소 제거 실험(Ablation Study)을 통해 비교했습니다.

다중 양상 인공지능 진단 모델 개발 시, 방사선 영상 이미지를 분류 기반으로 학습한 경우에는 모델의 복잡성으로 인해 다중 양상의 이점이 충분히 발휘되지 않았습니다. 반면, 탐지 기반 학습에서는 단일 양상보다 다중 양상 모델이 치아우식증, 치아파절, 치수염 진단에서 더 우수한 성능을 보였으며, 특히 치아파절 진단에서 뚜렷한 성능 향상이 관찰되었습니다.

결론

본 연구는 인공지능 기반 진단에서 다중 양상과 단일 양상 접근법의 진단 성능을 평가하고 비교했습니다. 또한 임상 검사 표준의 일관성, 방사선 영상 표기의 정밀도, 데이터의 양과 질이 인공지능 모델의 진단 성능에 크게 영향을 미친다는 것을 확인했습니다. 본 연구는 인공지능 기반 치과 진단에서 다중 양상 접근법의 가능성과 한계를 제시하며, 후향적 인공지능 연구와 임상 데이터의 표준화 및 통합의 중요성을 강조합니다. 향후 연구에서는 진단명에 따라 상호 보완성을 가지는 양상 데이터를 분석하고, 이를 통해 진단에서의 다중 양상 성능 향상 가능성을 탐구할 계획입니다.

핵심되는 말 : 인공지능, 인공지능 치과 진단, 치근단 방사선 영상, 임상 검사, 단일 양상 인공지능, 다중 양상 인공지능, 자가지도 학습, 마스킹 오토인코더