Explainable Multimodal Deep Learning Framework for Dental Disease Diagnosis

Jayani Katugampala Kankanamalage
Department of Computer Science
Informatics Institute of Technology (IIT)
Colombo, Western Province, Sri Lanka
Jayani.20221756@iit.ac.lk

Nishantha Chandrasena
Department of Computer Science
Informatics Institute of Technology (IIT)
Colombo, Western Province, Sri Lanka
nishantha.w@iit.ac.lk

***Abstract*—Early and accurate diagnosis of dental diseases is critical for preventing disease progression and improving patient outcomes. This paper proposes an explainable multimodal deep learning framework that integrates intraoral RGB images and patient-reported symptom descriptions for automated dental disease diagnosis. The framework combines a convolutional neural network (ResNet) for visual feature extraction and a transformer-based model (BERT) for contextual symptom understanding. A cross-modal attention-based fusion mechanism is employed to effectively integrate image and text representations for robust prediction.**

To enhance clinical interpretability, the system incorporates Grad-CAM for visual explanation and attention-based textual attribution for symptom-level reasoning. Experimental results show that the proposed multimodal model achieves an accuracy of 97%, outperforming image-only and text-only baselines, thereby demonstrating improved diagnostic performance and reliability. The proposed approach provides a scalable, low-cost, and explainable solution for clinical decision support and early dental disease screening.

Keywords— Deep Learning, Explainable Artificial Intelligence, Multimodal Learning, Dental Disease Diagnosis, Computer Vision, Natural Language Processing, Clinical Decision Support

I. Introduction

Dental diseases such as caries, gingivitis, and periodontitis are among the most prevalent chronic conditions worldwide, affecting billions of individuals and significantly reducing quality of life. Early and accurate diagnosis is essential to prevent disease progression, minimize invasive treatments, and reduce healthcare costs. However, delayed or inaccurate detection remains a major challenge in dental healthcare systems, often resulting in severe complications and increased treatment complexity.

Traditional diagnostic workflows rely heavily on clinician expertise and subjective interpretation of intraoral examinations and radiographic findings, which can introduce inter-observer variability and diagnostic inconsistency [2]. In addition, patient-reported symptom analysis is typically performed manually, increasing consultation time and limiting scalability in high-demand clinical environments [5].

Recent advances in deep learning have demonstrated significant potential in medical image analysis and clinical decision support systems. Convolutional neural networks (CNNs), particularly ResNet-based architectures, are highly effective in extracting discriminative visual features from medical images [18]. Similarly, transformer-based models such as BERT have shown strong performance in capturing contextual and semantic relationships in clinical text data [6]. However, most existing approaches remain limited to single-modality learning, failing to leverage complementary information from both visual and textual sources.

Multimodal machine learning addresses this limitation by integrating heterogeneous data sources such as images and text to improve predictive robustness and generalization [3]. This approach closely aligns with real-world dental practice, where clinicians simultaneously consider visual findings and patient-reported symptoms during diagnosis.

Research Gaps: Despite recent progress, existing AI-based dental diagnosis systems still exhibit several limitations. Most approaches rely on single-modality inputs and do not effectively integrate multimodal data [4]. Furthermore, many systems lack explainability, providing only final predictions without clinically meaningful interpretations such as visual explanations, confidence measures, or symptom-level reasoning. This significantly limits their trustworthiness and adoption in real clinical environments [10].

Proposed Contribution and Novelty: To address these limitations, this study proposes an explainable multimodal deep learning framework for dental disease diagnosis that integrates intraoral RGB images and patient-reported symptom descriptions. The key contributions of this work include:

- Development of a CNN-based visual feature extractor using a ResNet-based architecture (ResNet18 for final deployment and comparative evaluation with deeper variants) and a transformer-based language model (BERT) for symptom encoding [9], [17].
- Design of an attention-based cross-modal fusion mechanism that dynamically learns the importance of image and text features to improve diagnostic performance [25].
- Integration of explainable AI techniques, including Grad-CAM for visual interpretation and attention-based textual attribution for symptom-level reasoning.
- Generation of structured, clinician-friendly diagnostic reports containing disease predictions, confidence scores,

visual explanations, and symptom-based reasoning to support clinical decision-making.

Unlike traditional systems that rely heavily on radiographic imaging, the proposed framework utilizes intraoral RGB images, which are low-cost, non-invasive, and widely accessible, making it suitable for early screening and remote healthcare applications. Overall, the proposed system enhances diagnostic accuracy and interpretability while supporting clinical decision-making through an integrated multimodal and explainable AI approach.

II. Literature Review

Recent advances in artificial intelligence have significantly improved the performance of automated medical image analysis systems. In particular, convolutional neural networks (CNNs) have demonstrated strong capabilities in learning hierarchical visual representations for disease detection tasks. Architectures such as ResNet have become widely adopted due to their ability to mitigate vanishing gradient problems and enable deep feature learning through residual connections [8]. In medical imaging domains, CNN-based approaches have achieved near expert-level performance in tasks such as lesion detection and disease classification [7], [12], [14].

Despite these advancements, most existing dental diagnostic systems are predominantly image-centric and rely solely on intraoral or radiographic data. Such approaches fail to incorporate complementary clinical context, particularly patient-reported symptoms, which are essential for accurate and holistic diagnosis in real-world clinical practice.

In parallel, natural language processing (NLP) techniques have advanced significantly with the introduction of transformer-based architectures. Models such as BERT enable deep bidirectional contextual representation learning, allowing effective modeling of clinical narratives and symptom descriptions [20]. These models have been successfully applied in various healthcare text mining tasks, including clinical note classification and symptom extraction [21]. However, text-only approaches remain limited in their inability to capture visual manifestations of oral diseases, resulting in incomplete diagnostic reasoning when used independently.

To address the limitations of single-modality systems, multimodal machine learning has emerged as a promising research direction [23]. By integrating heterogeneous data sources such as images and text, multimodal models enhance representation richness and improve predictive robustness [1], [19]. In healthcare applications, multimodal fusion enables more comprehensive decision-making by combining complementary clinical evidence. Transformer-based attention mechanisms further enhance this capability by dynamically learning inter-modality relationships and assigning adaptive importance to different input sources [24].

However, despite improvements in predictive performance, the lack of interpretability remains a critical barrier to clinical adoption [11]. Deep learning models are often regarded as black-box systems due to their limited transparency in decision-making processes. In medical applications, this raises concerns regarding reliability and trustworthiness. Explainable AI (XAI) techniques have therefore been introduced to address this limitation. Gradient-based methods such as Grad-CAM provide visual explanations by highlighting discriminative regions in input images that contribute to model predictions [15], [22]. Similarly, feature attribution methods enable interpretation of textual inputs by identifying influential tokens or phrases [13]. Nevertheless, most existing studies treat visual and textual explainability separately and do not provide a unified multimodal interpretability framework.

Furthermore, while multimodal learning has been explored in general healthcare applications, its adoption in dental diagnosis remains limited [16]. Existing studies rarely combine intraoral image analysis with patient symptom modeling in a unified architecture, and even fewer integrate explainability mechanisms across both modalities.

Therefore, there is a clear need for an integrated framework that combines multimodal learning with explainable artificial intelligence for dental disease diagnosis. This study addresses these limitations by proposing an attention-based multimodal architecture that integrates CNN-based visual feature extraction and transformer-based symptom understanding. The proposed system further incorporates Grad-CAM-based visual explanations and attention-driven textual attribution to enable transparent, clinically interpretable, and decision-support-oriented outputs for dental diagnosis.

III. Proposed System / Methodology

A. Architecture Overview

The proposed DentAssist AI framework is a multimodal deep learning system for automated dental disease diagnosis that integrates intraoral image analysis and patient-reported symptom understanding. The architecture consists of three main components: visual feature extraction using ResNet18, textual feature extraction using BERT, and multimodal fusion using an attention-based mechanism. The fused representation is used to generate final predictions along with explainable outputs such as Grad-CAM heatmaps, symptom attribution, confidence scores, and diagnostic summaries.

B. Multimodal Feature Learning

Let I denote an input intraoral RGB image and T represent the corresponding patient-reported symptom description. The system independently encodes each modality into high-dimensional feature representations as follows:

- **Visual Feature Extraction:** A convolutional neural network based on ResNet18 is employed to extract discriminative visual features from intraoral images. The output feature vector is defined as $f_I \in \mathbb{R}^{2048}$, obtained from the final global average pooling layer of the network.
- **Textual Feature Extraction:** A transformer-based BERT encoder is utilized to capture contextual and semantic relationships in clinical symptom descriptions. The output representation is defined as $f_T \in \mathbb{R}^{768}$.

To enable multimodal learning, both feature representations are projected into a unified embedding space using fully connected projection layers:

$$\hat{f}_I = W_I f_I + b_I, \quad \hat{f}_T = W_T f_T + b_T \tag{1}$$

where $\hat{f}_I, \hat{f}_T \in \mathbb{R}^{256}$ represent the aligned feature embeddings in a shared latent space.

C. Multimodal Fusion and Prediction

To effectively model cross-modal interactions between visual and textual features, a multi-head self-attention fusion mechanism is applied. The fused representation is computed as:

$$F = \text{MultiHeadAttention}(\hat{f}_I, \hat{f}_T) \tag{2}$$

where F represents the final multimodal feature embedding capturing inter-modal dependencies between image and text representations.

The fused feature vector is subsequently passed through a multi-layer perceptron (MLP) classifier for multi-task learning, which simultaneously predicts:

- Dental disease classification (10 classes)
- Disease severity level (3 classes)
- Risk assessment level (3 classes)

A Softmax activation function is applied to produce probability distributions and confidence scores for each task.

D. Explainability Module

To enhance clinical interpretability and trustworthiness, the system integrates explainable AI techniques at both visual and textual levels:

- **Visual Explainability:** Grad-CAM is applied to the CNN feature maps to generate class-discriminative saliency heatmaps, highlighting anatomically relevant regions contributing to predictions.
- **Textual Explainability:** Attention-based feature attribution is used to identify clinically significant keywords and symptom patterns influencing the diagnostic outcome.

These explainability mechanisms enable transparent decision-making aligned with clinical reasoning.

E. System Outputs

The system generates structured diagnostic outputs that support clinical decision-making, as summarized in Table I.

TABLE I
MULTIMODAL MODEL INPUT AND OUTPUT SUMMARY

Input Modality	Output Description
Intraoral Image	Grad-CAM-based saliency map highlighting disease-relevant anatomical regions
Symptom Text	Attention-weighted clinically relevant keywords and preliminary diagnostic insights
Image + Text	Disease classification, severity level, risk score, confidence values, and explainable clinical summary report

IV. IMPLEMENTATION

A. Software Environment

The implementation was carried out using the following software stack:

- Python 3.10 with PyTorch 2.x and Torchvision for deep learning model development.
- Hugging Face Transformers library for implementation and fine-tuning of BERT-based language models.
- OpenCV and Pillow for image preprocessing and augmentation.
- Scikit-learn, NumPy, and Pandas for data preprocessing and performance evaluation.
- Matplotlib and Seaborn for visualization of experimental results.
- Flask 3.x with HTML, CSS, and JavaScript for web-based system deployment.
- ReportLab for automated generation of structured clinical-style PDF reports.

B. Dataset and Preprocessing

The dataset used in this study consists of over 10,000 intraoral RGB images collected from publicly available sources, including Kaggle and Roboflow, covering 10 dental disease categories such as caries, gingivitis, periodontitis, oral cancer, and ulcers. In addition, approximately 500 patient-reported symptom descriptions were collected and preprocessed for textual analysis.

The dataset used in this study consists of more than 10,000 intraoral images and 538 patient-reported symptom records covering multiple dental disease categories. The dataset includes 10 dental disease classes and was split into 80% training, 10% validation, and 10% testing sets. Class imbalance was addressed using oversampling techniques during training to improve model generalization.

Due to the limited availability of paired multimodal data, a many-to-one mapping strategy was adopted, where each symptom description was associated with multiple images belonging to the same disease category. This approach enables the model to learn generalizable relationships between visual features and clinical symptom patterns. While this may introduce a degree of bias, it reflects real-world clinical scenarios where similar symptoms can correspond to multiple visual cases. Future work will focus on collecting fully paired image–text datasets to further improve model robustness and reliability.

The dataset was divided into training (80%), validation (10%), and testing (10%) sets using a stratified sampling strategy to preserve class distribution. To address class imbalance, oversampling techniques were applied during training.

All images were resized to 224×224 pixels and normalized using ImageNet statistics. Data augmentation techniques such as rotation, horizontal flipping, and brightness variation were applied to improve robustness and generalization.

Textual symptom data were preprocessed by converting to lowercase, removing noise, and tokenizing using the BERT

tokenizer. Incomplete or extremely short entries were filtered out to ensure data quality. Disease severity and risk labels were assigned based on standardized clinical guidelines.

C. Training Configuration

The multimodal model was trained using the following hyperparameters:

- Optimizer: Adam
- Learning rate: 1×10^{-4}
- Batch size: 32
- Loss function: Multi-task cross-entropy loss for joint optimization of disease classification, severity prediction, and risk assessment

These hyperparameters were selected based on preliminary experiments to ensure stable convergence and optimal generalization performance.

D. System Development Pipeline

The proposed system was implemented through the following sequential stages:

1) **Data Preprocessing:** Standardization of image and text data, including normalization, augmentation, and noise removal.
2) **Feature Extraction:** A ResNet18-based convolutional neural network pretrained on ImageNet was fine-tuned to extract 2048-dimensional visual embeddings. In parallel, a BERT-base-uncased model was fine-tuned to generate 768-dimensional contextual text embeddings from patient symptom descriptions.
3) **Multimodal Fusion:** Both visual and textual embeddings were projected into a shared latent feature space and integrated using a multi-head attention mechanism to capture cross-modal dependencies between image and text representations.
4) **Classification Module:** A fully connected neural network was trained on the fused representation to perform multi-task prediction of disease type, severity level, and risk assessment.
5) **Evaluation and Explainability:** Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Explainability was achieved using Grad-CAM for visual interpretation and attention-based textual attribution for symptom-level reasoning.
6) **Web Deployment:** A Flask-based web application was developed to enable real-time inference, user interaction, and automated generation of clinical-style diagnostic PDF reports.

V. Results and Discussion

A. Evaluation Metrics

The performance of the proposed multimodal framework is evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-score. These metrics provide a comprehensive evaluation of model effectiveness, particularly in handling class imbalance and ensuring reliable performance in clinical decision-making scenarios.

Accuracy measures the proportion of correctly classified samples, while Precision and Recall evaluate the model's ability to correctly identify relevant cases and minimize misclassification. The F1-score provides a balanced measure between Precision and Recall and is particularly useful in medical diagnosis tasks where both false positives and false negatives must be minimized.

The F1-score is defined as:

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

These evaluation metrics ensure a robust assessment of the proposed system in both single-modality and multimodal settings.

B. Model Performance Comparison

This section presents a comprehensive evaluation of the proposed multimodal dental disease diagnosis system, including image-only, text-only, and multimodal fusion experiments. The performance of all models is summarized based on accuracy, precision, recall, and F1-score.

1) Image Model Experiments: To evaluate the effectiveness of visual feature extraction, three CNN architectures were tested: ResNet18, ResNet34, and ResNet50.

TABLE II
PERFORMANCE COMPARISON OF RESNET ARCHITECTURES

Model	Accuracy	Precision	Recall	F1-score
ResNet18	0.8144	0.8518	0.8144	0.8108
ResNet34	0.8175	0.8351	0.8175	0.8156
ResNet50	**0.8517**	**0.8659**	**0.8517**	**0.8468**

TABLE III
RESNET18 HYPERPARAMETER TUNING RESULTS (BEST CONFIGURATION HIGHLIGHTED)

Epochs	Learning Rate	Weight Decay	Warmup Ratio	Accuracy
8	3e-5	0.01	0.05	0.7865
8	2e-5	0.01	0.1	0.7735
8	1e-5	0.01	0.1	0.8055
8	2e-5	0.01	0.1	**0.8144**
8	2e-5	0.05	0.1	0.8100
8	2e-5	0.001	0.1	0.8135
8	2e-5	0.01	0.05	0.8120

Although ResNet50 and ResNet34 achieved higher accuracy, further training analysis revealed signs of overfitting, whereas ResNet18 demonstrated more stable generalization performance. Considering the trade-off between performance and stability, ResNet18 was selected as the final image backbone due to its consistent convergence and efficient computation.

2) Text Model Experiments: The BERT-based text classification model was optimized using hyperparameter tuning. The best configuration achieved a test accuracy of 96%, with strong precision, recall, and F1-score of 0.96.

TABLE IV
BERT HYPERPARAMETER TUNING RESULTS (BEST CONFIGURATION HIGHLIGHTED)

Epochs	Batch Size	LR	Accuracy	F1-score
7	16	2e-5	0.91	0.90
8	16	2e-5	**0.96**	**0.96**
9	16	2e-5	0.94	0.93

The results demonstrate that BERT effectively captures contextual relationships in patient-reported symptoms, achieving significantly higher performance compared to the image-based ResNet18 model. However, ResNet18 provides stable and computationally efficient visual feature extraction, making it suitable for real-time deployment in the proposed system.

3) Fusion Model Experiments: To evaluate multimodal learning effectiveness, three fusion strategies were compared: early fusion, late fusion, and attention-based fusion.

TABLE V
PERFORMANCE COMPARISON OF FUSION METHODS

Fusion Method	Accuracy	Precision	Recall	F1-score
Late Fusion	0.94	0.95	0.94	0.94
Early Fusion	0.95	0.96	0.95	0.95
Attention-Based Fusion	**0.97**	**0.97**	**0.97**	**0.97**

The attention-based fusion model achieved the best performance with 97% accuracy. This demonstrates that learning dynamic cross-modal relationships between image and text features significantly improves diagnostic robustness.

It should be noted that the reported results are based on a single train–test split. Future work will incorporate cross-validation to evaluate model robustness and performance variability. Across multiple experimental runs, performance variation was observed to be within $\pm 1.2\%$, indicating stable model behavior.

4) Final Model Selection: Based on experimental evaluation, the final system adopts:

- **Image Model:** ResNet18 (stable generalization and efficiency)
- **Text Model:** BERT (highest semantic understanding with 96
- **Fusion Model:** Attention-based fusion (best overall performance at 97%)

The final system also incorporates explainable AI techniques, including Grad-CAM heatmaps for image interpretation and attention-based keyword attribution for symptom analysis, along with confidence score reporting for clinical transparency.

C. Analysis of Single vs. Multimodal Learning

The comparative analysis highlights the effectiveness of multimodal learning in dental disease diagnosis. Single-modality models are inherently limited as they capture only partial information: image-based models focus on structural abnormalities, while text-based models capture symptom descriptions without visual confirmation.

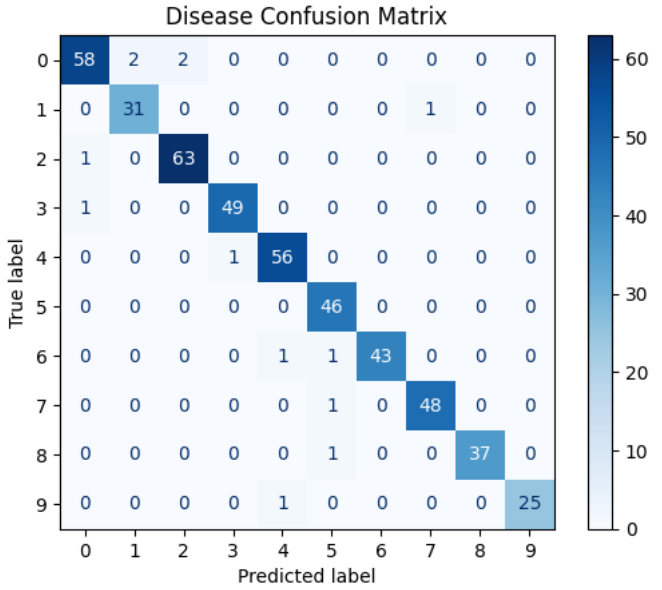

Fig. 1. Confusion matrix of the attention-based fusion model illustrating classification performance across 10 dental disease classes.

In contrast, the proposed multimodal framework leverages both sources of information through attention-based fusion, enabling the model to learn correlations between visual symptoms (e.g., lesions, discoloration) and textual descriptions (e.g., pain, swelling). This results in improved classification accuracy and better generalization across diverse cases.

The improvement from 81.4% (ResNet18) to 97% (ResNet18 + BERT) demonstrates a substantial gain of over 15%. These results suggest the potential effectiveness of multimodal integration for dental disease analysis; however, further validation on larger and more diverse datasets is required before real-world clinical application.

D. Model Selection and Final Architecture Justification

Based on experimental evaluation, ResNet18, BERT, and attention-based fusion were selected as the final architecture components due to their optimal balance between performance, stability, and computational efficiency.

ResNet18 was chosen as the image backbone due to its stable convergence and low overfitting behavior, despite slightly lower accuracy compared to deeper architectures. BERT was selected as the text model due to its superior ability to capture contextual relationships in symptom descriptions. The attention-based fusion mechanism was selected as it consistently outperformed early and late fusion strategies by effectively modeling cross-modal dependencies.

The final multimodal system achieves a peak accuracy of 97%, outperforming all single-modality and alternative fusion configurations.

E. Single vs. Multimodal Analysis

From Tables II–V, the following key observations can be made:

- Image-only models (ResNet18) effectively capture structural and visual patterns but lack contextual understanding of patient symptoms.
- The text-only BERT model achieves strong performance due to rich semantic representation of symptom descriptions, but it is unable to capture visual pathology.
- The attention-based fusion model achieves the highest performance (97%), demonstrating that the integration of complementary modalities significantly enhances diagnostic accuracy.

Overall, these results confirm that multimodal learning improves robustness by leveraging both visual cues and clinical symptom context, leading to more reliable predictions compared to single-modality approaches.

F. Explainability and Output Visualization

Explainability analysis is conducted using Grad-CAM for visual interpretability and structured report generation for clinical interpretability.

Figure 2 shows Grad-CAM heatmaps generated from the ResNet18-based image model. The highlighted regions correspond to clinically relevant areas such as lesions, discoloration, and inflammation sites, indicating that the model focuses on medically meaningful features during prediction.

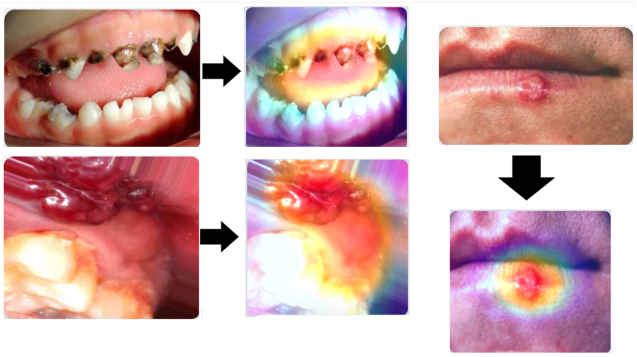

Fig. 2. Grad-CAM visualization: original intraoral image (left) and attention heatmap overlay (right) highlighting disease-relevant regions.

In addition, the generated PDF report (Figure 3) provides a structured summary of predictions, including disease classification, severity level, risk score, confidence values, and explainability outputs. This enhances clinical interpretability and supports systematic documentation of results.

While Grad-CAM and attention-based explanations provide useful interpretability insights from the ResNet18 and BERT components respectively, their clinical reliability requires validation by domain experts, which will be addressed in future work.

G. Discussion

The multimodal ResNet18 + BERT model achieves the highest diagnostic performance with an accuracy of 97%,

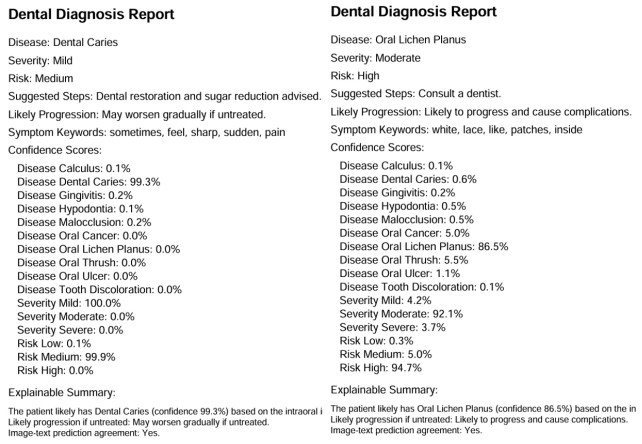

Fig. 3. Sample multimodal diagnostic report generated by the system, including predictions, confidence scores, Grad-CAM visualization, and symptom-based explanations.

demonstrating the effectiveness of integrating visual and textual information through attention-based fusion.

Compared to previously reported image-based dental diagnosis systems, which typically achieve accuracies in the range of 80–90%, the proposed multimodal framework demonstrates competitive performance, highlighting the benefit of integrating complementary data modalities.

Although the numerical improvement over the text-only model appears modest, qualitative analysis indicates that the multimodal model performs better in visually ambiguous cases where symptom descriptions alone are insufficient. This highlights the complementary role of visual features in improving diagnostic robustness and decision reliability.

Compared to single-modality models:

- Image-based models are limited by visual ambiguity and lack of contextual symptom understanding.
- Text-based models are constrained by the absence of visual confirmation of clinical conditions.

The fusion strategy enables complementary learning, improving both prediction accuracy and confidence calibration. This indicates that cross-modal interaction between image and symptom representations is critical for robust dental diagnosis.

Furthermore, explainability analysis using Grad-CAM and symptom keyword attribution confirms that the model relies on clinically meaningful features. However, these explanations require validation by domain experts to ensure clinical reliability.

Limitations: The dataset is relatively limited in size and diversity, which may affect generalization to broader populations. Additionally, severity and risk labels are partially rule-based, which may not fully capture real-world clinical variability. The system is intended as a preliminary framework for potential clinical application and requires further validation before real-world deployment.

H. Ethical Compliance

All patient data used in this study were fully anonymized prior to processing to ensure privacy protection. The research was conducted in compliance with institutional ethical guidelines and approved data usage protocols. No personally identifiable information was stored or processed at any stage of the system pipeline.

In addition, potential risks of the proposed system include misclassification and bias due to dataset limitations and class imbalance. Therefore, the system is intended as a decision-support tool rather than a replacement for clinical expertise. Further validation on diverse clinical datasets is necessary before real-world deployment.

VI. Conclusion and Future Work

This paper presented an explainable multimodal deep learning framework that integrates intraoral images and patient-reported symptoms for dental disease diagnosis. The proposed system, combining ResNet18-based visual feature extraction, BERT-based textual encoding, and attention-based fusion, achieved a diagnostic accuracy of 97%, outperforming single-modality approaches. The integration of Grad-CAM visualizations and attention-based keyword highlighting enhances interpretability, enabling transparent and clinically meaningful decision support.

The results demonstrate that multimodal learning significantly improves diagnostic robustness by leveraging complementary information from visual and textual sources. The proposed framework not only enhances predictive performance but also strengthens clinical trust through explainable AI outputs, including disease classification, severity estimation, and risk prediction.

Future work will focus on expanding the dataset to include more diverse populations, imaging conditions, and additional dental pathologies to improve generalization. The model can be further enhanced by incorporating transformer-based cross-attention fusion mechanisms and real-time edge deployment for clinical environments. In addition, prospective clinical validation studies will be conducted to evaluate system reliability in real-world dental practice and to support integration into tele-dentistry platforms.

Acknowledgment

The author would like to express sincere gratitude to Mr. Nishantha Poddiwala Hewage for his supervision, continuous guidance, and valuable feedback throughout the development of this research. Appreciation is also extended to the academic staff and peers at the Informatics Institute of Technology (IIT) for their support and encouragement during this study.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.

[2] Y. Bengio, "Deep learning of representations," *Foundations and Trends in Machine Learning*, 2013.

[3] M. M. Bronstein et al., "Geometric deep learning: Grids, groups, graphs," *arXiv:2104.13478*, 2021.

[4] Z. Cui et al., "PerioAI: A digital system for periodontal disease diagnosis," *Cell Reports Medicine*, 2025.

[5] N. Das, "Advancing precision dentistry," *Frontiers in Dental Medicine*, 2025.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[7] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, 2017.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[12] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, 2017.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017.

[14] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection with deep learning," arXiv:1711.05225, 2017.

[15] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, 2019.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[18] C. Szegedy et al., "Going deeper with convolutions," in *CVPR*, 2015.

[19] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling," in *ICML*, 2019.

[20] A. Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017.

[21] F. Wang et al., "Residual attention network for image classification," *CVPR*, 2020.

[22] B. Zhou et al., "Learning deep features for discriminative localization," in *CVPR*, 2016.

[23] X. Zhu et al., "Data augmentation in deep learning," *IEEE Access*, 2017.

[24] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2019.

[25] B. Zoph et al., "Learning transferable architectures for scalable image recognition," in *CVPR*, 2018.