Reliable Dental Radiograph Diagnosis via Calibrated Hybrid Representation Learning

Sayem Ahmed Shayeed [1] **Md. Jalal Uddin Chowdhury** [1] **Shadman Sakib** [2]

Abstract

Automated diagnosis from intraoral radiographs can enable scalable oral-health screening in settings with limited access to specialist interpretation. We present a calibrated hybrid representation-learning framework for multi-class dental radiograph diagnosis that combines a fine-tuned ConvNeXt-Tiny feature extractor with classical machine-learning classifiers trained on frozen deep embeddings. We evaluate the framework on DentIRO, a four-class single-tooth intraoral radiograph dataset, using patient-level train, validation, and locked-test splits to reduce evaluation leakage. On the locked test set, the proposed model achieves approximately 98.6% macro-F1 and 99.1% accuracy, with low calibration error and Grad-CAM evidence highlighting clinically meaningful tooth structures. These results demonstrate the potential of calibrated hybrid deep-feature pipelines for accurate, reliable, and interpretable dental radiograph analysis.

1. Introduction

Dental diseases, including caries, periodontal conditions, and pulpal pathologies requiring root canal therapy, affect an estimated 3.5 billion people worldwide and remain a major global health burden (World Health Organization, 2023). Intraoral radiography is central to dental diagnosis, supporting the detection of interproximal caries, periapical lesions, and restoration-related abnormalities that are often not visible through clinical inspection alone (Liu et al., 2026; Jones et al., 2025). However, radiographic interpretation is time-intensive, requires specialist expertise, and is subject to inter-observer variability, limiting scalable screening in resource-constrained and underserved clinical settings (Schwendicke et al., 2020). This challenge is especially relevant for low-resource healthcare systems, including many underserved communities in Muslim-majority regions, where access to dental specialists and radiographic interpretation may be limited.

Automated dental radiograph diagnosis presents both clinical and technical challenges. Intraoral radiographs are grayscale, low-contrast, and structurally different from the natural images used to pre-train most modern vision models (Mutawa et al., 2026; Mei et al., 2025). Moreover, many existing dental imaging benchmarks are limited by narrow diagnostic scope, inconsistent annotation protocols, or evaluation settings that may not fully control patient-level leakage (Mine et al., 2025; Jones et al., 2025; Absar et al., 2025). These limitations make it difficult to assess whether high-performing models are learning clinically meaningful diagnostic patterns or exploiting dataset-specific artifacts. Recent approaches commonly fine-tune convolutional networks such as ResNet, EfficientNet, DenseNet, and ConvNeXt, or transformer-based models such as ViT, in an end-to-end manner (He et al., 2016; Tan & Le, 2019; Huang et al., 2017; Liu et al., 2022; Dosovitskiy et al., 2021; Ozdemir et al., 2026). While effective, such models may produce poorly calibrated confidence estimates on small medical datasets, which is problematic for clinical decision support (Guo et al., 2017; Yuan et al., 2025). Beyond discrimination performance, reliable dental AI systems should be evaluated under leakage-aware splits, report calibration quality, and provide interpretable evidence that model attention is aligned with clinically meaningful tooth structures.

We present a calibrated hybrid representation-learning framework for multi-class diagnosis from intraoral dental radiographs. The proposed framework fine-tunes a ConvNeXt-Tiny backbone as a domain-adapted feature extractor and trains classical machine-learning classifiers on frozen deep embeddings, separating representation learning from decision modeling. We evaluate the framework on DentIRO (Shoib et al., 2026), a four-class single-tooth intraoral radiograph dataset covering Healthy, Caries, Crowned, and Root Canal cases, using patient-level train, validation, and locked-test splits.

Our contributions are summarized as follows:

[1]Department of Computer Science and Engineering, Leading University, Sylhet, Bangladesh [2]Department of Information Systems, University of Maryland, Baltimore County, Baltimore, USA. Correspondence to: Shadman Sakib <ssakib1@umbc.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea.

- We provide a leakage-aware benchmark for multi-class single-tooth diagnosis on DentIRO, using patient-level splits and a locked-test evaluation protocol.
- We introduce a hybrid ConvNeXt-Tiny representation-learning pipeline with classical machine-learning heads, achieving approximately 98.6% macro-F1 and 99.1% accuracy on the locked test set.
- We evaluate reliability beyond accuracy through calibration metrics and qualitative Grad-CAM visualizations, showing low calibration error and attention patterns concentrated around clinically meaningful dental structures.

2. Method

We propose a calibrated hybrid representation-learning framework for multi-class diagnosis from intraoral dental radiographs. The framework separates representation learning from decision modeling: a ConvNeXt-Tiny backbone is first fine-tuned to learn domain-adapted radiographic features, and classical machine-learning classifiers are then trained on frozen deep embeddings. This design retains the representational strength of modern deep networks while allowing lightweight decision heads to operate on a stable feature space.

2.1. Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote a dataset of intraoral radiographs, where $\mathbf{x}_i \in \mathbb{R}^{H\times W\times C}$ is an input image and $y_i \in \mathcal{Y}$ is its diagnostic label. In this work, $\mathcal{Y} = \{\textit{Healthy}, \textit{Caries}, \textit{Crowned}, \textit{Root Canal}\}$. Given a pre-processed radiograph $\hat{\mathbf{x}}_i$, the goal is to predict a diagnostic label $\hat{y}_i$ together with a calibrated probability vector $\hat{\mathbf{p}}_i \in \Delta^{|\mathcal{Y}|-1}$.

2.2. Backbone Fine-Tuning

We use ConvNeXt-Tiny (Liu et al., 2022) as the visual backbone due to its strong transfer performance and convolutional inductive bias for local image structure. The ImageNet-pretrained classification layer is replaced with a task-specific linear head, and the network is fine-tuned on the training split using cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\frac{1}{|\mathcal{D}_{\mathrm{train}}|} \sum_{(\hat{\mathbf{x}}_i, y_i)\in\mathcal{D}_{\mathrm{train}}} \log p_\theta(y_i \mid \hat{\mathbf{x}}_i). \quad (1)$$

The best checkpoint is selected according to validation macro-F1:

$$\theta^* = \arg\max_{\theta} F_1^{\mathrm{macro}}(\mathcal{D}_{\mathrm{val}}; \theta). \quad (2)$$

After fine-tuning, the task-specific classification layer is removed and the backbone parameters are frozen.

2.3. Frozen Deep Embeddings

The frozen backbone maps each pre-processed radiograph to a compact 768-dimensional representation:

$$\mathbf{f}_i = \Phi_{\theta^*}(\hat{\mathbf{x}}_i) \in \mathbb{R}^{768}, \quad (3)$$

where Φ_{θ^*} denotes the selected ConvNeXt-Tiny feature extractor. For a split $\mathcal{D}_s$, where $s \in \{\mathrm{train}, \mathrm{val}, \mathrm{test}\}$, the corresponding feature matrix is

$$\mathbf{F}_s = [\Phi_{\theta^*}(\hat{\mathbf{x}}_1); \ldots; \Phi_{\theta^*}(\hat{\mathbf{x}}_{N_s})] \in \mathbb{R}^{N_s\times 768}. \quad (4)$$

The downstream classifiers are trained only on these frozen representations, which decouples feature learning from final decision modeling.

2.4. Classical Decision Heads

Given frozen embeddings $\mathbf{f}_i$, we train a classical classifier $\mathcal{C}_\psi$ to produce class scores $\mathbf{s}_i = \mathcal{C}_\psi(\mathbf{f}_i) \in \mathbb{R}^{|\mathcal{Y}|}$. We evaluate three decision heads: radial-basis-function support vector machines, random forests, and multinomial logistic regression. For the SVM head, the RBF kernel is defined as:

$$k(\mathbf{f}_i, \mathbf{f}_j) = \exp\left(-\gamma\|\mathbf{f}_i - \mathbf{f}_j\|_2^2\right). \quad (5)$$

For the logistic-regression head, class probabilities are modeled as:

$$p(y = k \mid \mathbf{f}_i) = \frac{\exp(\mathbf{w}_k^\top \mathbf{f}_i + b_k)}{\sum_{j=1}^{|\mathcal{Y}|} \exp(\mathbf{w}_j^\top \mathbf{f}_i + b_j)}. \quad (6)$$

The predicted diagnosis is obtained by:

$$\hat{y}_i = \arg\max_{k\in\mathcal{Y}} s_{ik}. \quad (7)$$

All hyperparameters and classifier choices are selected using the validation split only. The selected backbone–classifier pair is fixed before locked-test evaluation.

2.5. Probability Calibration

Reliable confidence estimates are important for clinical decision support. Given class scores $\mathbf{s}_i$, we apply temperature scaling (Guo et al., 2017) on the validation split:

$$\hat{\mathbf{p}}_i = \mathrm{softmax}\left(\frac{\mathbf{s}_i}{T^*}\right), \quad (8)$$

where the scalar temperature $T^* > 0$ is fitted by minimizing validation negative log-likelihood:

$$T^* = \arg\min_{T>0} \sum_{(\hat{\mathbf{x}}_i, y_i)\in\mathcal{D}_{\mathrm{val}}} -\log \mathrm{softmax}\left(\frac{\mathbf{s}_i}{T}\right)_{y_i}. \quad (9)$$

Calibration quality is evaluated using Expected Calibration Error and Brier score in the experiments.

2.6. Visual Explanation

To provide qualitative evidence about the image regions used by the model, we generate Grad-CAM visualizations (Dennis, 2026) from the final convolutional stage of the ConvNeXt-Tiny backbone. Let $\mathbf{A} \in \mathbb{R}^{C' \times H' \times W'}$ denote the final-stage feature maps and let s^c be the class score for target class c. The Grad-CAM channel weights are computed as:

$$\alpha_k^c = \frac{1}{H'W'} \sum_{u=1}^{H'} \sum_{v=1}^{W'} \frac{\partial s^c}{\partial A_{k,u,v}}, \tag{10}$$

and the corresponding heatmap is:

$$L_{\mathrm{GradCAM}}^c = \mathrm{ReLU}\left(\sum_{k=1}^{C'} \alpha_k^c \mathbf{A}_k \right). \tag{11}$$

The heatmap is upsampled to the input resolution and overlaid on the original radiograph. We use these maps to assess whether discriminative regions align with clinically meaningful tooth structures, such as crown, root, and canal regions, while treating Grad-CAM as qualitative evidence of model attention rather than proof of causal clinical reasoning.

3. Experiments

3.1. Dataset and Pre-processing

We evaluate on DentIRO (Shoib et al., 2026), a four-class single-tooth intraoral radiograph dataset containing 5,300 cleaned images from 3,243 patients. The classes are *Healthy*, *Caries*, *Crowned*, and *Root Canal*. To reduce evaluation leakage, images are partitioned at the patient level into fixed train, validation, and locked-test splits using a 70/15/15 ratio, yielding 3,721, 791, and 788 images, respectively. Split integrity is verified before model development using patient-overlap and duplicate-file checks. Each grayscale radiograph is replicated into three channels, resized to 224×224 using aspect-ratio-preserving padding, and normalized using training-split statistics only. During training, we apply mild geometric and intensity augmentations to improve robustness while preserving diagnostic structure.

3.2. Baselines

We compare the proposed hybrid framework against representative supervised and self-supervised baselines under the same data splits and pre-processing protocol. Supervised CNN baselines include ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), EfficientNet-B3 (Tan & Le, 2019), and an end-to-end ConvNeXt-Tiny (Liu et al., 2022). Transformer baselines include ViT-B/16 and ViT-B/32 (Dosovitskiy et al., 2021). We also include a self-supervised ResNet-50 baseline using SimCLR (Chen et al., 2020), where the encoder is pre-trained on the training split without labels and then fine-tuned for supervised classification. For the proposed framework, we evaluate ConvNeXt-Tiny embeddings with three classical decision heads: radial-basis-function support vector machines, random forests, and multinomial logistic regression. All model selection is performed on the validation split before evaluating the selected configuration on the locked test set.

3.3. Implementation Details

All deep models are initialized from ImageNet-pretrained weights unless stated otherwise. ConvNeXt-Tiny is fine-tuned using AdamW with weight decay 0.01, a linear warm-up schedule followed by cosine annealing, and early stopping based on validation macro-F1. For the hybrid framework, embeddings are extracted from the selected ConvNeXt-Tiny checkpoint and used to train the classical decision heads. Hyperparameters for the decision heads are selected using validation macro-F1 only. All experiments are repeated across three random seeds, $\{42, 2026, 3407\}$, where applicable. Validation results are reported as mean $\pm$ standard deviation across seeds. Locked-test results are computed only after training, hyperparameter selection, and calibration are finalized.

3.4. Evaluation Protocol

Macro-F1 is the primary metric because the dataset is imbalanced across diagnostic categories. We also report accuracy and macro-recall, with macro-recall corresponding to balanced accuracy. Reliability is assessed using Expected Calibration Error and Brier score, with temperature scaling fitted only on the validation split. Grad-CAM visualizations are generated for representative locked-test examples to assess whether model attention aligns with clinically meaningful tooth regions. The locked test set is accessed only after all model-selection decisions are completed, providing a leakage-aware estimate of generalization under fixed patient-level splits.

4. Results and Discussion

Table 1 reports validation and locked-test performance under the fixed patient-level split. Results are reported in percentages for readability. The proposed hybrid ConvNeXt-Tiny + ML framework achieves the highest mean validation macro-F1 of 98.53 ± 0.20 and the highest mean locked-test macro-F1 of 98.59 ± 0.13, with locked-test accuracy of 99.07 ± 0.07. The representative ConvNeXt-Tiny + Logistic Regression run used for diagnostic visualization reaches 98.66 macro-F1 and 99.11% accuracy on the locked test set. The results show that ConvNeXt-Tiny provides a strong representation for single-tooth intraoral radiographs. The hybrid decision layer further improves the grouped mean

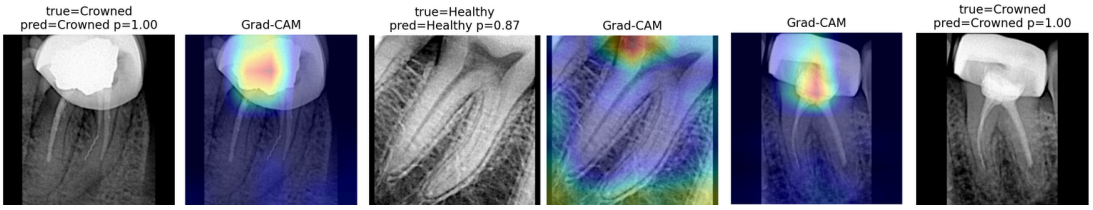


Figure 1. Grad-CAM visualizations for representative locked-test predictions. The heatmaps highlight tooth-associated crown, root, and canal regions, providing qualitative evidence that the model attends to clinically meaningful structures.

Table 1. Performance summary on DentIRO. Values are percentages. Validation results are averaged over three seeds. Locked-test results are reported as mean ± standard deviation where available.

Model	Val. Macro-F1	Test Acc.	Test Macro-F1
ConvNeXt-Tiny CNN	98.44 ± 0.29	99.03 ± 0.07	98.52 ± 0.12
ConvNeXt-Tiny + ML	$\mathbf{98.53 \pm 0.20}$	$\mathbf{99.07 \pm 0.07}$	$\mathbf{98.59 \pm 0.13}$
DenseNet-121	97.91 ± 0.75	—	—
ViT-B/16	97.89 ± 0.19	98.73 ± 0.66	98.06 ± 1.04
ViT-B/32	97.70 ± 0.84	98.73 ± 0.13	98.02 ± 0.24
SimCLR ResNet-50	95.93 ± 0.20	—	—
EfficientNet-B3	94.05 ± 1.63	—	—
ResNet-50	93.32 ± 0.85	—	—

over the end-to-end ConvNeXt-Tiny baseline, while ViT-B/16 and ViT-B/32 remain competitive but show lower mean locked-test macro-F1. SimCLR pre-training improves ResNet-50 over the supervised ResNet-50 baseline, suggesting that dental-domain self-supervision is useful, although it does not match the strongest ConvNeXt-based models.

4.1. Class-wise and Diagnostic Analysis

The representative hybrid ConvNeXt-Tiny + Logistic Regression model shows strong performance across all four classes. *Caries* remains the most challenging category, with F1 of 96.97, compared with 98.86 for *Healthy*, 99.10 for *Crowned*, and 99.73 for *Root Canal*. This pattern is clinically plausible because carious findings can be subtle and variable in radiographs, whereas crowned and root-canal-treated teeth often contain more visually distinctive structures.

4.2. Calibration and Interpretability

Calibration results indicate that the strongest models provide reliable probability estimates in addition to high discrimination performance. The hybrid ConvNeXt-Tiny + ML models show low calibration error, with ECE below 0.009 in the locked-test evaluation. For the hybrid ML heads, the fitted temperature remains $T^* = 1.0$, indicating that the learned decision scores are already well calibrated in this setting. By contrast, temperature scaling improves the direct ConvNeXt-Tiny CNN baseline, reducing its ECE from 0.0087 to 0.0057. As shown in Figure 1, Grad-CAM visualizations provide qualitative evidence that the learned representations emphasize tooth-associated regions, including crown, root, and canal structures. These visual patterns support the clinical plausibility of the learned features, although saliency maps should be interpreted as qualitative attention evidence rather than causal explanations of model reasoning.

4.3. Discussion and Limitations

Overall, the results suggest that high locked-test performance on DentIRO is achievable when evaluation is performed under patient-level splits with duplicate checks. The strongest performance comes from ConvNeXt-based representations, indicating that modern convolutional backbones transfer effectively to single-tooth intraoral radiographs. The hybrid framework provides a strong and reliable configuration by combining deep radiographic features with lightweight classical decision heads. The main limitation is that all results are obtained on a single dataset. Although the locked-test protocol reduces evaluation leakage, external validation on multi-center radiograph collections is required before clinical deployment. In addition, the small performance differences among the strongest models suggest that future work should focus not only on accuracy, but also on cross-site robustness, calibration under distribution shift, annotation efficiency, and clinician-centered evaluation.

5. Conclusion

We presented a calibrated hybrid representation-learning framework for multi-class diagnosis from intraoral dental radiographs. By combining a fine-tuned ConvNeXt-Tiny feature extractor with classical machine-learning decision heads, the proposed framework achieves approximately 98.6% macro-F1 and 99.1% accuracy on the DentIRO locked test set under patient-level evaluation. These results show that pretrained convolutional representations can transfer effectively to single-tooth radiograph classification when evaluated with leakage-aware splits. Beyond accuracy, the framework supports reliability-oriented assessment through calibration analysis and interpretable dental radiograph analysis. The hybrid models show low calibration error and clinically meaningful saliency patterns. Future work will focus on external validation, robustness under acquisition shifts, and clinician-centered deployment.

References

Absar, S., Ahmed, I., Sakib, S., Sharma, N., Bhuiyan, S. T., Alam, S. B., Islam, A., and Rahman, R. Multi-strategy optimization of u-net variants for orthopantomogram segmentation. In *2025 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, pp. 17–22. IEEE, 2025.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Dennis, D. Deep learning-based dental image analysis using grad cam convolutional network. *European Journal of Prosthodontics and Restorative Dentistry*, 34(2):38–47, 2026.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

Jones, B., Lambach, M., Chen, T., Michou, S., Kilpatrick, N., Curtis, N., Burgner, D. P., Vannahme, C., and Silva, M. Dental caries detection in children using intraoral scans and deep learning. *Journal of Dentistry*, 160:105906, 2025.

Liu, H., Gong, Z., Wen, B., Qiu, L., Meng, X., Cai, G., Zeng, P., Chen, S., Shi, M., Zhang, X., et al. Deep learning model development and clinical validation for radiographic surrogate markers of implant esthetic risk. *npj Digital Medicine*, 2026.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Mei, L., Deng, K., Cui, Z., Fang, Y., Li, Y., Lai, H., Tonetti, M. S., and Shen, D. Clinical knowledge-guided hybrid classification network for automatic periodontal disease diagnosis in x-ray image. *Medical Image Analysis*, 99: 103376, 2025.

Mine, Y., Okazaki, S., Taji, T., Kawaguchi, H., Kakimoto, N., and Murayama, T. Benchmarking multimodal large language models on the dental licensing examination: challenges with clinical image interpretation. *Journal of Dental Sciences*, 2025.

Mutawa, A., Altarakemah, Y. Y., and Thirupathy, K. Deep learning applications for dental-disease classification using intraoral photographic images: Current status and future perspectives. *AI*, 7(3):85, 2026.

Ozdemir, D., Ozcan, C., Karaoglu, A., Pekince, A., Yasa, Y., Kazangirler, B. Y., and Meseci, E. An enhanced deep learning model for detection and classification of dental caries in panoramic radiographs. *Neural Computing and Applications*, 38(1):3, 2026.

Schwendicke, F., Samek, W., and Krois, J. Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, 99(7):769–774, 2020.

Shoib, M. M. H., Rahman, A., Bijoy, M. H. I., Durjoy, S. H., Shikder, M. E., Hasan, M. Z., Mazumder, R. C., and Rashid, B. Dentiro: A high-quality multi-class single-tooth intraoral radiograph dataset for automated dental diagnosis, 2026.

Srijiranon, K., Varisthanist, N., and Tanantong, T. A study of simclr-based self-supervised learning for acne severity grading under label-scarce conditions. *Technologies*, 14 (2):116, 2026.

Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Varoquaux, G. and Colliot, O. Evaluating machine learning models and their diagnostic value. In *Machine Learning for Brain Disorders*. Springer, 2023.

World Health Organization. Oral health: Key facts. Technical report, World Health Organization, 2023. URL https://www.who.int/news-room/fact-sheets/detail/oral-health.

Yuan, L., Zhu, L., Jo, K., and Li, Y. Dental identification based on single tooth feature concatenation. *IEEE Access*, 2025.

A. Appendix

A.1. Additional Methodological Details

A.1.1. Baseline Architectures

To provide comparative context, we evaluate supervised CNN, transformer, and self-supervised baselines under the same patient-level splits and pre-processing protocol used in the main paper. The first baseline is a shallow five-layer CNN trained from scratch, with batch normalization and ReLU activations after each convolutional layer. We also evaluate EfficientNet-B0 and EfficientNet-B3 (Tan & Le, 2019), compound-scaled convolutional networks based on MBConv blocks with squeeze-and-excitation modules. ResNet-50 (He et al., 2016) is included as a standard residual-network baseline, while DenseNet-121 (Huang et al., 2017) is included to evaluate dense feature reuse across convolutional layers. All supervised CNN baselines are fine-tuned from ImageNet-pretrained weights unless otherwise stated.

For transformer-based baselines, we evaluate ViT-B/16 and ViT-B/32 (Dosovitskiy et al., 2021). These models divide each input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into non-overlapping patches and process the resulting patch-token sequence using transformer self-attention. These baselines compare convolutional inductive biases with patch-based global attention for single-tooth intraoral radiograph classification.

Finally, we include a self-supervised ResNet-50 baseline using SimCLR (Chen et al., 2020; Srijiranon et al., 2026). The encoder is first pre-trained on the training split without labels using contrastive learning and then fine-tuned for supervised four-class classification. This baseline evaluates whether label-free domain pre-training on dental radiographs provides useful representation learning signal.

A.1.2. Self-Supervised Pre-training with SimCLR

To investigate whether label-free pre-training on domain data provides a complementary signal, we apply SimCLR (Chen et al., 2020) as a standalone self-supervised baseline. Given an unlabeled radiograph $\mathbf{x}$, two stochastic augmented views are generated:

$$\tilde{\mathbf{x}}_i = t_i(\mathbf{x}), \qquad \tilde{\mathbf{x}}_j = t_j(\mathbf{x}), \qquad t_i, t_j \sim \mathcal{T}, \tag{12}$$

where $\mathcal{T}$ denotes the augmentation distribution. In our experiments, $\mathcal{T}$ includes random cropping, mild intensity perturbation, and Gaussian blur.

A ResNet-50 encoder f_θ maps each augmented view to a representation $\mathbf{h} = f_\theta(\tilde{\mathbf{x}})$, and a projection head g_ϕ maps the representation to a normalized contrastive embedding:

$$\mathbf{z} = \frac{g_\phi(\mathbf{h})}{\|g_\phi(\mathbf{h})\|_2}. \tag{13}$$

For a positive pair (i, j), the normalized temperature-scaled contrastive loss is defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}, \tag{14}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature, and N is the batch size. The full NT-Xent loss over a mini-batch of N images, yielding $2N$ augmented views,

$$\mathcal{L}_{\mathrm{SimCLR}} = -\frac{1}{2N}\sum_{k=1}^{N}\left[\log \frac{\exp\left(\mathrm{sim}(\mathbf{z}_{2k-1}, \mathbf{z}_{2k})/\tau\right)}{\sum_{m\neq 2k-1}\exp\left(\mathrm{sim}(\mathbf{z}_{2k-1}, \mathbf{z}_m)/\tau\right)} + \log \frac{\exp\left(\mathrm{sim}(\mathbf{z}_{2k}, \mathbf{z}_{2k-1})/\tau\right)}{\sum_{m\neq 2k}\exp\left(\mathrm{sim}(\mathbf{z}_{2k}, \mathbf{z}_m)/\tau\right)}\right]. \tag{15}$$

We use $\tau = 0.2$, pre-train on $\mathcal{D}_{\mathrm{train}}$ only for 80 epochs with batch size 64, discard the projection head, and then fine-tune the encoder for supervised classification.

A.1.3. Probability Calibration

Deep neural classifiers are known to produce overconfident predictive probabilities (Guo et al., 2017). We therefore evaluate post-hoc temperature scaling as a calibration method. Given class scores $\mathbf{s}_i \in \mathbb{R}^K$ for sample i, calibrated probabilities are computed as:

$$\hat{\mathbf{p}}_i = \mathrm{softmax}\left(\frac{\mathbf{s}_i}{T^*}\right), \tag{16}$$

Algorithm 1 Hybrid ConvNeXt-ML Training and Inference

Require: Training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, locked test set $\mathcal{D}_{\text{test}}$, ImageNet-pretrained ConvNeXt-Tiny backbone Φ_θ, classifier family $\mathcal{C}$, seed set $\mathcal{S}$
Ensure: Test predictions $\hat{\mathbf{y}}_{\text{test}}$ and calibrated probabilities $\hat{\mathbf{P}}_{\text{test}}$
Stage 1: Backbone fine-tuning
for $s \in \mathcal{S}$ **do**
  Initialize $\theta^{(s)}$ from ImageNet weights and set random seed s
  Fine-tune ConvNeXt-Tiny on $\mathcal{D}_{\text{train}}$ using cross-entropy loss
  Evaluate macro-F1 on $\mathcal{D}_{\text{val}}$
  Save checkpoint if validation macro-F1 improves
end for
Select θ^* as the checkpoint with the highest validation macro-F1
Stage 2: Frozen feature extraction
Extract $\mathbf{F}_{\text{train}} = \{\Phi_{\theta^*}(\hat{\mathbf{x}}_i)\}_{i=1}^{N_{\text{train}}}$
Extract $\mathbf{F}_{\text{val}} = \{\Phi_{\theta^*}(\hat{\mathbf{x}}_i)\}_{i=1}^{N_{\text{val}}}$
Stage 3: ML head training and calibration
Train candidate classifiers $\mathcal{C}_\psi$ on $(\mathbf{F}_{\text{train}}, \mathbf{y}_{\text{train}})$
Select classifier hyperparameters using validation macro-F1
Compute validation class scores $\mathbf{S}_{\text{val}} = \mathcal{C}_\psi(\mathbf{F}_{\text{val}})$
Fit calibration temperature T^* on $(\mathbf{S}_{\text{val}}, \mathbf{y}_{\text{val}})$
Stage 4: Locked test evaluation
Extract $\mathbf{F}_{\text{test}} = \{\Phi_{\theta^*}(\hat{\mathbf{x}}_i)\}_{i=1}^{N_{\text{test}}}$
Compute test class scores $\mathbf{S}_{\text{test}} = \mathcal{C}_\psi(\mathbf{F}_{\text{test}})$
Predict labels $\hat{\mathbf{y}}_{\text{test}} = \arg\max_k \mathbf{S}_{\text{test},k}$
Compute calibrated probabilities $\hat{\mathbf{P}}_{\text{test}} = \text{softmax}(\mathbf{S}_{\text{test}}/T^*)$
return $\hat{\mathbf{y}}_{\text{test}}, \hat{\mathbf{P}}_{\text{test}}$

where $T^* > 0$ is a scalar temperature fitted on the validation split by minimizing negative log-likelihood:

$$T^* = \arg\min_{T>0} \sum_{(\hat{\mathbf{x}}_i, y_i) \in \mathcal{D}_{\text{val}}} -\log \text{softmax}\left(\frac{\mathbf{s}_i}{T}\right)_{y_i}. \tag{17}$$

For neural baselines, $\mathbf{s}_i$ denotes logits. For classical ML heads, decision-function scores or class-score proxies are used where applicable.

Calibration quality is evaluated using multiclass Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{18}$$

where $B_1, \ldots, B_M$ are confidence bins, $\text{acc}(B_m)$ is the empirical accuracy in bin B_m, and $\text{conf}(B_m)$ is the mean predicted confidence in the same bin. We use $M = 15$ equal-width bins. We also report the Brier score:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\hat{p}_{ik} - \mathbf{1}[y_i = k]\right)^2. \tag{19}$$

A.1.4. Training and Inference Algorithm

Algorithm 1 summarizes the complete two-stage training and inference procedure. Temperature scaling is applied to class scores, not to final predicted labels.

A.2. Evaluation Protocol

A.2.1. METRICS

Macro-F1 is the primary metric because DentIRO is imbalanced across the four diagnostic categories. For K classes, macro-F1 is computed as:

$$F_1^{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \frac{2P_k R_k}{P_k + R_k}, \tag{20}$$

where P_k and R_k denote precision and recall for class k. We additionally report accuracy, macro-recall, weighted F1, Cohen's κ, per-class precision/recall/F1/specificity, macro one-vs-rest ROC-AUC, macro average precision, ECE, and Brier score.

A.2.2. BOOTSTRAP CONFIDENCE INTERVALS

For test-set metrics, we compute 95% bootstrap confidence intervals using $B = 2000$ stratified bootstrap resamples. For metric M, the confidence interval defined as:

$$\text{CI}_{95\%}(M) = \left[M^{(0.025)},\ M^{(0.975)} \right], \tag{21}$$

where $M^{(\alpha)}$ denotes the α-quantile of the bootstrap metric distribution.

A.2.3. LOCKED-TEST PROTOCOL

Following best practices in clinical machine learning (Varoquaux & Colliot, 2023), the locked test split is accessed only after model selection, hyperparameter tuning, and calibration are completed using the training and validation splits. The test split is not used for choosing the backbone checkpoint, classifier family, classifier hyperparameters, calibration temperature, or ablation settings.

A.2.4. MULTI-SEED REPRODUCIBILITY

All applicable models are trained with three independent random seeds: $\{42, 2026, 3407\}$. Validation results are reported as mean $\pm$ standard deviation across seeds. Locked-test results are reported as mean $\pm$ standard deviation for model families evaluated on the locked test set. The representative hybrid ConvNeXt-Tiny + Logistic Regression run is used only for diagnostic plots and per-class analysis.

A.3. Ablation Study

A.3.1. ABLATION AXES

To isolate the effect of key design choices, we conduct validation-set ablation experiments using macro-F1 as the selection criterion. The locked test set is not used for ablation-based model selection. The ablation axes are:

1. **Backbone Architecture**: ConvNeXt-Tiny vs. EfficientNet-B3 vs. ResNet-50 as feature extractors.
2. **Decision Head**: SVM-RBF vs. Random Forest vs. Logistic Regression vs. a standard fine-tuned linear head.
3. **Pre-training Strategy**: ImageNet supervised pre-training vs. SimCLR domain pre-training vs. random initialization.
4. **Input Modality**: replicated three-channel grayscale input vs. single-channel adapted first layer.
5. **Optimization**: unified learning rate vs. differential learning rates for backbone and classification head.
6. **Calibration**: uncalibrated vs. temperature-scaled predictions, evaluated using ECE and Brier score.

A.3.2. ABLATION RESULTS

The ablation results are summarized in Table 2 and Figure 2. The results support two observations: loss-function changes produce relatively small differences, while label availability has a larger effect on performance. Class-weighted cross-entropy

Table 2. Validation ablation summary for ConvNeXt-Tiny. Values are percentages. Loss-function and label-scarcity results are validation-set sensitivity analyses and are not locked-test comparisons.

Setting	Macro-F1	Caries F1
Cross-entropy	98.3	97.2
Class-weighted CE	98.5	97.2
Focal loss	98.2	97.4
10% labels	92.1	79.6
25% labels	95.6	89.5
50% labels	97.6	94.7
100% labels	98.3	97.2

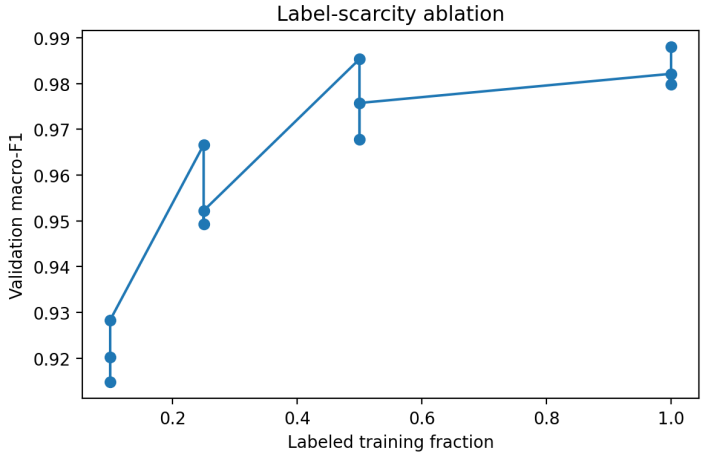


Figure 2. Label-scarcity ablation for ConvNeXt-Tiny. Validation macro-F1 improves substantially as more labeled training data are available, indicating that annotation availability has a stronger effect than small loss-function changes.

gives a small improvement over standard cross-entropy, increasing validation macro-F1 from 98.3% to 98.5%. Focal loss gives the highest *Caries* F1 among the loss variants, but it does not improve overall macro-F1. Label scarcity has a stronger effect: validation macro-F1 increases from 92.1% with 10% of labels to 98.3% with the full training set, while *Caries* F1 increases from 79.6% to 97.2%.

A.4. Extended Results and Analysis

We report extended results under the fixed patient-level split. Macro-F1 is used as the primary metric because DentIRO is imbalanced across *Healthy*, *Caries*, *Crowned*, and *Root Canal*. In addition to aggregate validation and locked-test performance, we discuss model-family behavior, diagnostic plots, calibration, ablation trends, and Grad-CAM-based qualitative explainability. The interpretation of these results is supported by the DentIRO annotation protocol. The dataset labels were independently reviewed by licensed dental experts, audited through internal clinical committees, checked through post-labeling visual review, and finalized through expert consensus before model development. Thus, class-wise findings are interpreted with respect to dentist-verified diagnostic labels rather than unreviewed folder-level labels.

A.4.1. CNN Baseline Analysis

Among the supervised CNN baselines, ConvNeXt-Tiny is the strongest direct CNN model. It achieves approximately 98.4% validation macro-F1 and 98.5% locked-test macro-F1. The low seed-level variance indicates stable optimization and makes ConvNeXt-Tiny the most reliable end-to-end CNN comparator in our experiments. DenseNet-121 is the second strongest CNN baseline, with approximately 97.9% validation macro-F1. Its performance is close to ConvNeXt-Tiny but less stable across seeds. Dense feature reuse may help preserve local radiographic patterns, but the model does not match ConvNeXt-Tiny in either mean performance or stability.

The supervised ResNet-50 baseline obtains approximately 93.3% validation macro-F1, while EfficientNet-B3 obtains

Table 3. Locked-test per-class performance for the representative hybrid ConvNeXt-Tiny + Logistic Regression run. Metric values are percentages; support denotes the number of test images.

Class	Precision	Recall	F1	Support
Healthy	98.6	99.1	98.9	218
Caries	97.6	96.4	97.0	83
Crowned	99.1	99.1	99.1	111
Root Canal	99.7	99.7	99.7	376

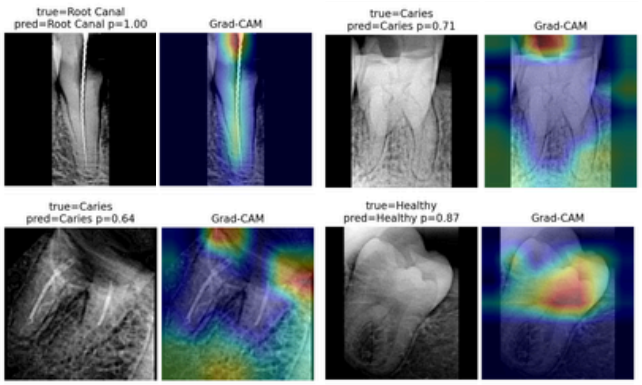


Figure 3. Grad-CAM visualization for the end-to-end ConvNeXt-Tiny CNN baseline.

approximately 94.1%. EfficientNet-B0 performs poorly in the tested configuration, with approximately 40.5% validation macro-F1. This should be interpreted as the behavior of the tested lightweight configuration rather than a broad conclusion about the EfficientNet family. A Grad-CAM visualization for the direct ConvNeXt-Tiny CNN baseline is shown in Figure 3.

A.4.2. Vision Transformer Baseline Analysis

The ViT baselines are competitive but do not consistently outperform ConvNeXt-Tiny. ViT-B/16 obtains approximately 97.9% validation macro-F1 and 98.1% locked-test macro-F1, while ViT-B/32 obtains approximately 97.7% validation macro-F1 and 98.0% locked-test macro-F1. These results show that transformer-based representations can learn useful diagnostic structure from DentIRO, although their aggregate performance remains slightly below the strongest ConvNeXt-based models.

A.4.3. Hybrid ConvNeXt-Tiny with Classical Machine Learning

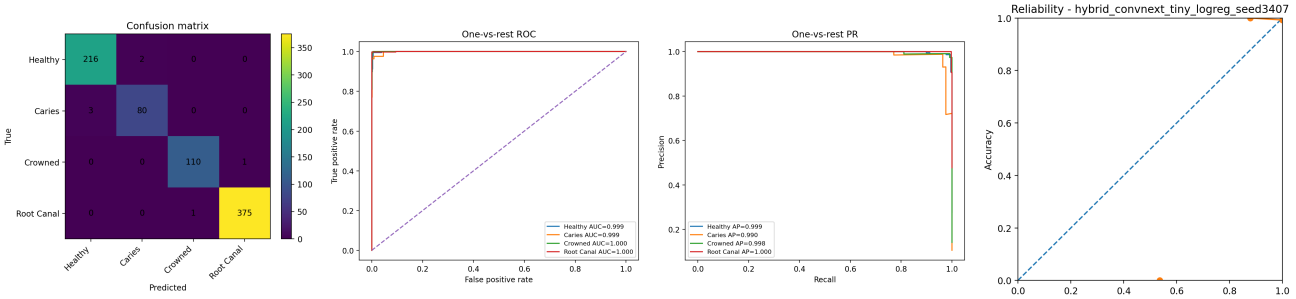
The hybrid ConvNeXt-Tiny pipeline uses ConvNeXt-Tiny as a feature extractor and trains classical machine-learning classifiers on frozen image representations. As shown in Table 1, ConvNeXt-Tiny + ML achieves the strongest grouped performance among the evaluated configurations, reaching approximately 98.6% locked-test macro-F1. The representative hybrid ConvNeXt-Tiny + Logistic Regression run obtains 99.1% accuracy, 98.7% macro-F1, 99.9% macro ROC-AUC, and 99.7% macro PR-AUC. Its class-wise performance is reported in Table 3, and the corresponding diagnostic plots are shown in Figure 4. The margin over direct ConvNeXt-Tiny is small, so the result should be interpreted as a strong and reliable refinement of a competitive ConvNeXt representation rather than evidence of clear architectural dominance. As shown in Table 3, *Caries* remains the most challenging class, while *Root Canal* and *Crowned* achieve the strongest class-wise performance. This is clinically plausible because carious findings may be subtler and more variable than crown or root-canal treatment patterns.

A.4.4. Self-Supervised ResNet-50 Analysis

Self-supervised pre-training improves the ResNet-50 baseline. The supervised ResNet-50 model obtains approximately 93.3% validation macro-F1, while the SimCLR-pretrained and fine-tuned ResNet-50 reaches approximately 95.9%. This is an improvement of about 2.6 macro-F1 points, suggesting that dental-domain self-supervision provides useful representation learning signal. However, the SSL ResNet-50 baseline remains below the strongest ConvNeXt- and ViT-based models.

Table 4. Locked-test calibration summary. ECE and Brier scores are reported as unit-scale values rounded to three decimals. For hybrid ML models, the fitted temperature remains $T^* = 1.0$, so calibrated and uncalibrated values are identical.

Model family	ECE before	ECE after	Brier before	Brier after
ConvNeXt-Tiny CNN	0.009	0.006	0.017	0.016
ViT-B/16	0.007	0.008	0.022	0.021
ViT-B/32	0.010	0.007	0.022	0.021
Hybrid ConvNeXt-Tiny + ML	0.009	0.009	0.016	0.016
Hybrid ConvNeXt-Tiny + LogReg	0.008	0.008	0.016	0.016



(a) Confusion matrix (b) ROC curves (c) PR curves (d) Reliability

Figure 4. Locked-test diagnostic plots for the representative hybrid ConvNeXt-Tiny + Logistic Regression model. The model shows strong discrimination and low calibration error, with remaining errors concentrated in visually similar classes.

A.4.5. Calibration Analysis

Calibration is evaluated using ECE and Brier score, with detailed results reported in Table 4. Temperature scaling improves the direct ConvNeXt-Tiny CNN baseline, reducing mean ECE from approximately 0.009 to 0.006 and mean Brier score from approximately 0.017 to 0.016. ViT-B/32 also improves after calibration, with ECE decreasing from approximately 0.010 to 0.007. ViT-B/16 shows mixed calibration behavior: mean ECE increases slightly, while mean Brier score decreases slightly. For the hybrid ConvNeXt-Tiny + ML candidates, the fitted temperature remains $T^* = 1.0$, so calibrated and uncalibrated values are identical. Thus, the hybrid models should be described as already showing low calibration error in this evaluation, rather than being improved by temperature scaling.

A.4.6. XAI and Grad-CAM Analysis

Grad-CAM visualizations for the hybrid ConvNeXt-Tiny + Logistic Regression model are shown in Figure 1. The heatmaps often highlight tooth-associated structures, including crown, root, and canal-related regions. Because DentIRO labels were established through dentist-led annotation, committee audit, and expert consensus, these saliency patterns can be discussed with respect to clinically defined diagnostic categories. At the same time, Grad-CAM should be interpreted as qualitative explainability evidence rather than proof of causal clinical reasoning. The visualizations are useful for examining whether model attention is anatomically plausible, but they do not establish that the model reasons in the same way as a dental expert.

Overall, the extended results show that high locked-test performance on DentIRO is achievable under patient-level evaluation with clinically validated labels. ConvNeXt-Tiny is the strongest direct CNN baseline, the selected hybrid ConvNeXt-Tiny + ML candidates obtain the highest grouped locked-test macro-F1, and ViT-B/16 remains closely competitive. The differences among the strongest models are small, so the main conclusion is not that a single architecture dominates. Rather, DentIRO appears highly separable for strong pre-trained vision models when leakage is controlled and annotations are dentist-verified. Considering macro-F1, seed stability, class-wise behavior, calibration, clinically validated annotations, and qualitative Grad-CAM evidence together, the hybrid ConvNeXt-Tiny pipeline is a strong representative configuration, while direct ConvNeXt-Tiny remains an important end-to-end comparator.