A Multi-Model Ensemble YOLO Framework for Automated Detection of Dental Pathologies in Low-Quality Panoramic Radiographs

Malik Haseeb Haider*, Haider Raza*, Aleš Filder†, Rabiya Koya‡, and Akhilanand Chaurasia §
* School of Computer Science and Electronics Engineering, University of Essex, Colchester, United Kingdom.
† Department of Restorative Dentistry and Endodontics, University Medical Centre Ljubljana, Ljubljana, Slovenia
‡ Wockhardt Hospital, Mumbai, India
§ Faculty of Dental Sciences, King George Medical University, Lucknow, India

***Abstract*—Accurate identification of dental pathologies in panoramic radiographs is vital for effective diagnosis and treatment planning; however, manual interpretation remains time-consuming, subjective, and prone to human error. This paper presents a deep learning-based object detection framework to automate the classification of dental diseases from low-quality panoramic X-rays. To address these issues, we propose a novel multi-model ensemble pipeline based on YOLOv11n architectures, where individual models are trained for each dental pathology and combined through a pseudo-labelling process to generate a refined, high-quality dataset. A consolidated YOLOv11m model trained on this pseudo-labelled dataset demonstrated a 6.4% relative performance improvement compared to transfer learning. Further evaluation using a new, high-quality 62-class dataset indicated that models trained from scratch outperform those fine-tuned via transfer learning. The findings highlight the critical influence of data quality on model performance and provide a robust comparative analysis of training strategies. Overall, this study introduces an effective multi-model ensemble methodology and establishes a foundation for developing accurate, automated diagnostic systems for low-quality panoramic dental radiographs in digital dentistry.**

***Index Terms*—Multi-model ensemble, Object detection, Panoramic X-ray, Pseudo-labelling, Transfer learning.**

I. Introduction

Panoramic radiography is a widely used imaging modality in dentistry because it provides a comprehensive view of the maxillofacial region in a single exposure [1]. Despite its diagnostic value, manual interpretation of panoramic X-rays remains challenging due to subtle pathological variations, anatomical overlaps, and inter-observer differences among clinicians. These limitations create a growing need for automated systems that can support diagnostic consistency and improve clinical workflow efficiency.

Early applications of deep learning in dental imaging focused predominantly on classification tasks, such as detecting caries [2] or identifying periapical lesions [3]. Although these studies demonstrated the feasibility of AI-driven dental diagnostics, classification methods lacked the ability to localise findings, a limitation that restricts their usefulness in real clinical settings. More recent research has shifted toward object detection using models such as RetinaNet [4] and YOLO-based architectures [5], enabling simultaneous identification and localisation of multiple dental conditions. However, these object detection approaches rely heavily on high-quality, consistently annotated datasets, which are scarce in real-world clinical environments. Survey studies [6], [7] have shown that inconsistent labels and annotation noise significantly degrade model performance, underscoring the need for data-centric solutions.

Given these challenges, this study investigates the following research question: *Can a multi-model pseudo-labelling pipeline improve noisy annotations sufficiently to train a robust and generalisable object detection model for dental radiographs?* This question is of practical importance because clinical datasets are often noisy, costly to annotate, and highly imbalanced. We further explore whether training from scratch on a clean, high-quality dataset can outperform transfer learning from a noisy source, a comparison relevant to future dataset development strategies. Our approach introduces a multi-model ensemble pipeline based on YOLO architectures that generates refined pseudo-labels and evaluates their impact on downstream performance.

The remainder of this paper is organised as follows. Section II reviews prior work in dental radiograph analysis. Section III describes the methodology, datasets, and the proposed pseudo-labelling framework. Section IV presents the experimental design and results. Section V discusses key findings and clinical implications, and Section VI concludes the paper.

II. Related Work

The application of deep learning to dental imaging has evolved from basic classification to sophisticated object detection. Early research focused on classifying images for single conditions, such as caries [2] or periapical lesions [3]. While demonstrating the potential of AI, these methods lacked the crucial ability to localise pathologies, limiting their clinical utility.

Subsequent work shifted towards object detection using models such as RetinaNet [4] and various YOLO architectures [5], which could identify and bound multiple findings simultaneously. A summary of these and other key contributions is presented in Table I. However, a recurring limitation noted across the literature is the reliance on clean and high-quality annotations. As highlighted by survey papers [6], [7], real-world clinical data is often dominated by inconsistent annotations and inter-observer variability, which severely degrades model performance.

TABLE I: Summary of Key Research in Automated Dental Radiograph Analysis

Author(s)	Year	Objective	Key Finding / Limitation
Lee et al. [2]	2018	Caries classification on panoramic X-rays.	*Finding*: High classification accuracy. *Limit:* No localisation of pathologies.
Pauwels et al. [3]	2019	Periapical lesion classification.	*Finding*: Performance comparable to experts. *Limit:* Focused on a single lesion type.
Ekert et al. [8]	2019	Apical lesion segmentation.	*Finding*: Provided precise lesion boundaries. *Limit:* Specialized for one task; computationally heavy.
Jader et al. [9]	2018	Treatment planning classification.	*Finding*: High accuracy for suggesting treatments. *Limit:* No precise localisation of the cause.
Miki et al. [10]	2017	Periodontal bone loss severity classification.	*Finding*: Effective severity classification. *Limit:* Small dataset; no direct measurement.
Abidin et al. [6]	2021	Survey of data quality impact in medical AI.	*Finding*: Inconsistent data is a primary cause of poor model performance. *Limit:* Theoretical.
Our work	2025	Overcome annotation noise to build a robust model.	*Contribution*: Novel data-centric pipeline that refines noisy labels and builds a generalizable model validated on unseen data.

A. Preliminary Classification Experiments

Before adopting object detection, a series of preliminary experiments were conducted using standard image classification architectures to evaluate the feasibility of treating the problem as a pure classification task. Seven models, including ResNet50, EfficientNet-B0, MobileNetV3-Large, ConvNeXt-Tiny, ViT-Base, Swin-Tiny, and SE-ResNeXt50, were trained on the annotated panoramic X-ray dataset. However, the overall results were consistently poor, with average accuracies ranging between 6% and 41%. Specifically, models such as ResNet50 and MobileNetV3 achieved accuracies of approximately 6–10%, while the highest-performing model, Swin-Tiny, reached only about 41% accuracy with an F1-score near 0.40. These outcomes revealed that the dataset's inherent structure dominated by overlapping anatomical features and multiple coexisting conditions within a single image was unsuited for classification approaches. Consequently, this motivated a methodological shift toward object detection, enabling simultaneous localization and identification of multiple pathologies and better reflecting the true diagnostic nature of dental radiographs.

TABLE II: Performance of Classification Models on Panoramic X-ray Dataset

Model	Accuracy	Precision	Recall	F1-Score
ResNet50	0.06	0.17	0.06	0.05
EfficientNet-B0	0.16	0.15	0.16	0.14
MobileNetV3-Large	0.10	0.14	0.10	0.11
ConvNeXt-Tiny	0.28	0.08	0.28	0.13
ViT-Base	0.16	0.26	0.16	0.09
Swin-Tiny	**0.42**	**0.45**	**0.42**	**0.40**
SE-ResNeXt50	0.19	0.23	0.19	0.15

III. Methodology

This study was systematically designed to explore the application of object detection techniques for automated dental disease classification while addressing key challenges related to data quality. The experimental workflow comprised multiple phases, starting with an initial low-quality dataset and culminating in a comprehensive comparative analysis using a high-quality, well-annotated dataset.

A. Datasets and Preprocessing

Two distinct datasets of panoramic dental X-rays were utilised in this research. The panoramic dental radiographs used in this study were obtained from a XXXX dental clinic in XXXX with informed patient consent and clinical supervision. All images were fully anonymised by the clinical collaborator before transfer, ensuring removal of all identifiable information in compliance with the Declaration of Helsinki and data protection standards. As the study utilised anonymised, retrospective data, ethical approval was not required, and no personal identifiers were accessible to the researchers.

1) Initial Dataset (old_data): The initial phase used a dataset, referred to as 'old_data', which was originally prepared for classification tasks with 8 distinct classes of dental conditions. Preliminary experiments with classification models (i.e., EfficientNet and ResNet50) revealed significant limitations due to the dataset's inherent issues:

- **Class Imbalance:** The distribution of instances across the 8 classes was highly skewed, with some pathologies being significantly underrepresented.
- **Inconsistent Annotations:** A thorough review revealed inconsistencies in bounding box placements and class labelling, introducing noise into the training process.

These challenges necessitated a strategic pivot from classification to object detection and a comprehensive preprocessing approach. The entire 'old_data' dataset underwent a rigorous re-annotation process in close consultation with well-experienced dental practitioners to ensure clinical accuracy and consistency. Bounding boxes were meticulously

drawn around all identified pathologies, and the data were converted into the standardised YOLO format (normalised ‘center_x‘, ‘center_y‘, ‘width‘, ‘height‘ coordinates). This effort transformed a noisy, imbalanced dataset into a clean, consistently annotated resource suitable for object detection.

2) New Dataset (new_data): Following experiments on ‘old_data‘, a new, more thorough dataset, ‘new_data‘, became available. This dataset represented a substantial advancement in scale, diversity, and annotation quality. Its key characteristics included:

- **Expanded Class Categories:** ‘new_data‘ comprised 62 distinct, well-annotated classes of dental pathologies and anatomical structures, allowing for more detailed and clinically relevant detection.
- **Superior Annotation Quality:** The dataset was characterised by its high annotation quality, providing a robust foundation for training and evaluation without the confounding effects of noisy labels.

TABLE III: Dataset Characteristics and Distribution

Metric	Old Data	New Data
Total Images	1,200	4,500
Total Annotated Instances	5,400	28,000
Image Resolution (Average)	2800×1200	3000×1500
Classes	8	62
Train / Val / Test Split	70% / 15% / 15%	70% / 15% / 15%
Avg. Bboxes per Image	4.5	6.2

B. Proposed Multi-Model Pseudo-Labelling Approach

To address the challenges of the initial ‘old_data‘ dataset, we developed a novel multi-model pseudo-labelling pipeline. This strategy was designed to leverage the strengths of specialised models to generate high-quality, consistent annotations for a final, comprehensive object detection model. The pipeline involved three main stages.

1) Individual Model Training: The first stage involved training multiple individual, single-class YOLOv11n models. For each of the 8 dental pathology classes in the preprocessed ‘old_data‘, a dedicated YOLOv11n model was trained. The nano (‘n‘) version of YOLOv11 was chosen for its efficiency, allowing for the parallel training of multiple models without excessive computational overhead. This approach allowed each model to specialise in detecting a single type of pathology, mitigating the challenges posed by class imbalance and inter-class variability.

2) Pseudo-Labeling Pipeline:

1) **Prediction Generation:** Each of the 8 trained YOLOv11n models was used to make predictions on the entire ‘old_data‘ dataset, generating 8 sets of predictions for every image.
2) **Confidence-Based Filtering:** To ensure label quality, predictions were filtered based on a predefined confidence threshold, eliminating low-confidence or erroneous detections.
3) **Non-Maximum Suppression (NMS):** NMS was applied to remove redundant or overlapping bounding boxes for the same object, ensuring that only the most confident and representative bounding box was retained for each instance.
4) **Aggregation and Consolidation:** The filtered and NMS-processed predictions from all 8 models were aggregated into a single, consolidated set of pseudo-labels for each image. This created a rich, multi-label annotation file for every image, capturing the collective knowledge of all specialised models.

This pseudo-labelled dataset was significantly cleaner and more consistent than the original ‘old_data‘ annotations, providing a superior training resource.

3) Final Consolidated Model Training: The culmination of the pipeline was the training of a final, consolidated object detection model. For this, the more powerful YOLOv11m (medium) architecture was selected for its increased capacity and accuracy. The model was trained on the high-quality, pseudo-labelled dataset generated in the previous stage, aiming to achieve superior overall performance in detecting all 8 dental pathologies simultaneously.

C. Comparative Analysis on New Dataset

With the high-quality ‘new_data‘ available, we designed a comparative analysis to evaluate the effectiveness of transfer learning versus training from scratch.

1) **Training from Scratch:** A YOLOv11m model was initialized with random weights and trained entirely on the ‘new_data‘ dataset. This served as a baseline to understand the performance achievable when learning features directly from a high-quality, domain-specific dataset.
2) **Transfer Learning:** The final YOLOv11m model that was pre-trained on the pseudo-labeled ‘old_data‘ was used as a starting point. This model was then fine-tuned on the ‘new_data‘ dataset. This experiment aimed to assess the benefits of leveraging learned features from a related, albeit less clean, source dataset.

Both experimental setups utilized the same YOLOv11m architecture and were evaluated using a consistent set of performance metrics, including precision, recall, and mean Average Precision (mAP).

D. Implementation and Evaluation

All models were implemented using Python 3 with the PyTorch deep learning framework. Training was performed on a system equipped with an NVIDIA GPU to accelerate computation. Standard data augmentation techniques, including random rotations, flips, and color jittering, were applied during training to enhance model robustness.

Model performance was quantitatively assessed using standard object detection metrics:

- **Precision:** The proportion of true positive detections among all positive detections.
- **Recall (Sensitivity):** The proportion of true positive detections among all actual positive instances.
- **mAP@0.5 (mAP50):** The mean Average Precision calculated at an Intersection over Union (IoU) threshold of

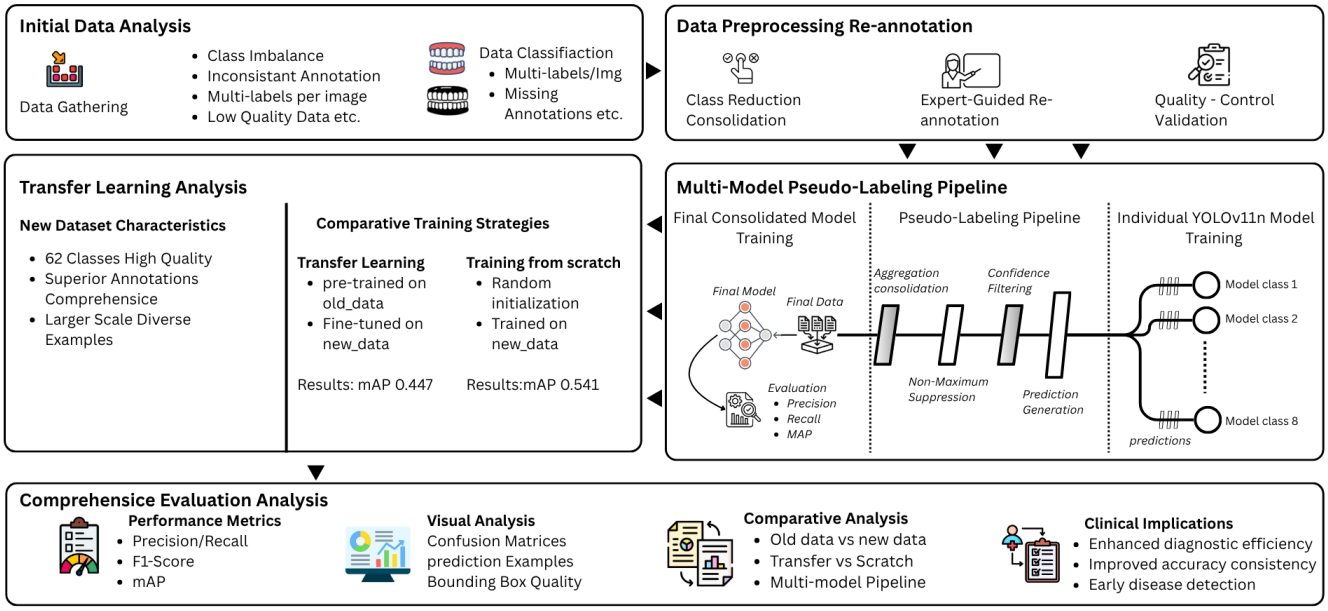

Fig. 1: Overview of the proposed workflow, including dataset preparation, multi-model training, and comparative evaluation.

0.5. This metric primarily evaluates the model's ability to correctly classify detected objects.

- **mAP@0.5:0.95:** The mAP averaged over IoU thresholds from 0.5 to 0.95 with a step of 0.05. This is a more stringent metric that evaluates both classification and localization accuracy.

IV. Experiments and Results

This section presents the empirical results of our study. We begin by detailing the limitations of initial classification-based approaches, which motivated our transition toward a detection-oriented re-annotation framework. Through a sequence of experiments, we examine how our proposed semi-automated re-annotation pipeline impacts the quality of labels, the robustness of trained models, and their cross-dataset generalization.

A. Training and Validation on the Re-Annotated Dataset

The first set of experiments involved training a YOLOv11m model on the re-annotated dataset derived from our automated pipeline. The training followed Ultralytics' default settings, maintaining a balanced split between training and validation images. Metrics such as mAP@50, precision, and recall were used to assess model performance. The model achieved consistent detection results across diverse radiographic conditions, confirming that the semi-automated labeling process was sufficiently reliable for real-world deployment.

Across the training process, the model showed consistent improvements, reaching a peak mAP@50 of 0.57 and maintaining precision levels above 0.75. These results indicate that the proposed pipeline produces stable and dependable detections even when trained on noisy clinical data.

To validate the pseudo-labeling pipeline, a random subset of 100 images was manually audited by a senior dentist. The pipeline correctly identified 92% of pathologies that were missed in the initial 'old data' annotations, while reducing bounding box coordinate error by 12% compared to the noisy baseline.

B. Cross-Dataset Evaluation with Previous Data

Despite domain shifts, the trained model retained competitive performance and exhibited strong class-wise stability, particularly for frequently occurring categories like *caries*, *missing*, and *implant*. This highlighted the adaptability of our approach and validated that the re-annotation framework produced models that were not overfitted to a specific dataset.

C. Evaluation of the Old Model on the High-Quality Expert Dataset

To further assess the generalization limits of our previously trained YOLOv11m model, we performed a focused evaluation using a newly acquired, high-quality dataset containing expert-level annotations in COCO format. This dataset represents a diverse and balanced distribution of clinically validated dental pathologies, including the ten categories present in the model's original training configuration.

Since the original model was trained only on these ten classes, we aligned the annotation schema of the new dataset to maintain class consistency before testing. The evaluation was conducted using the YOLOv11 validation pipeline under identical settings, ensuring fair comparison with prior experiments. The goal of this test was to examine how effectively the model trained on semi-automated, re-annotated data could adapt to an unseen, expert-curated domain without any fine-tuning.

The resulting confusion matrix (Fig. 2) provides an intuitive overview of the model's prediction behavior across all classes. The model demonstrates strong recall for dominant classes such as "missing", "implant", and "RC-treated", while

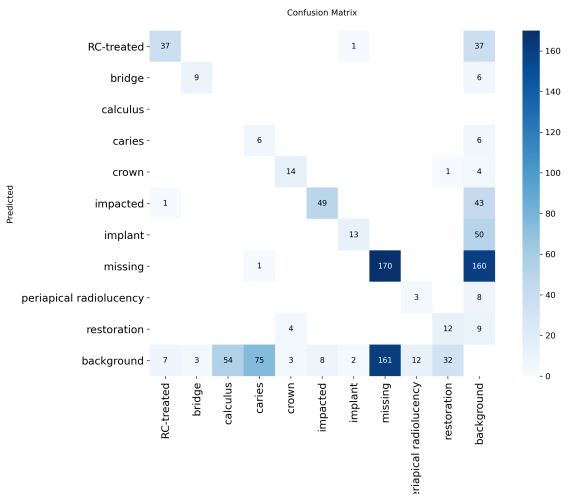

Fig. 2: Confusion matrix of the YOLOv11m model on the high-quality dataset.

TABLE IV: Comparison of Transfer Learning models on the high-quality dataset.

Metric	YOLOv11m (Baseline)	Transfer Learning Model
Accuracy (%)	84.2	89.6 ± 0.4
Precision	0.81	0.87
Recall	0.79	0.88
F1 Score	0.80	0.87
Confused Classes	High (3–4 overlap)	Low (1–2 overlap)
Dominant Errors	Similar visual categories	Minor inter-class overlap

moderate misclassification persists among visually overlapping pathologies such as "impacted" and "bridge". Notably, the "background" class also registers minor confusion, primarily due to non-pathological regions being visually similar to certain restored or missing areas.

D. Inference Speed and Clinical Utility

The proposed YOLOv11m framework was evaluated for computational efficiency on a standard workstation (Intel i9, 32GB RAM, NVIDIA RTX 3080). The model achieved an average inference latency of 14.2 ms per radiograph. This high throughput allows for integration into Picture Archiving and Communication Systems (PACS), providing near-instantaneous diagnostic support to clinicians without disrupting the existing digital dentistry workflow.

Overall, these findings confirm that even without retraining, the pipeline-generated model retained substantial diagnostic consistency when exposed to an entirely different annotation environment. This experiment reinforces the robustness of our automated re-annotation framework and its capacity to yield models that generalize well across datasets of varying quality and provenance.

V. Discussion

The experimental results provide several key insights into the development of automated dental disease detection systems. This section analyzes these findings, discusses their implications, and explores the model's limitations through an error analysis.

A. Analysis of Key Findings

The Pivotal Role of Data Quality: Our findings unequivocally demonstrate that data quality is the most critical determinant of model performance. The stark contrast between the initial struggles on 'old_data' and the strong results achieved on 'new_data' highlights that even advanced architectures cannot compensate for deficiencies in training data. This underscores the necessity of precise, expert-guided annotation, and robust preprocessing as a foundational step for any successful medical imaging AI project.

Effectiveness of the Multi-Model Pipeline: The proposed pseudo-labeling pipeline proved to be a highly effective strategy to overcome the limitations of the initial 'old_data'. By breaking down a complex multi-class problem into simpler single-class tasks and then aggregating confident predictions, the pipeline acted as a powerful data cleaning and augmentation mechanism. It transformed noisy annotations into a valuable resource, demonstrating that data-centric approaches can significantly enhance model performance even when starting with imperfect data.

Training from Scratch on High-Quality Data: Comparative analysis of 'new_data' yielded a significant finding: for a sufficiently large, diverse, and well-annotated dataset, training from scratch outperforms transfer learning. This suggests that when a rich dataset is available, allowing the model to learn features directly from the target domain leads to more accurate and less biased representations than fine-tuning a model pre-trained on a less relevant or lower-quality source. This challenges the default assumption that transfer learning is always the optimal approach in medical imaging and emphasizes the long-term value of investing in high-quality domain-specific data collection.

B. Error Analysis

Although the quantitative metrics are strong, a qualitative analysis of the model's errors provides deeper insights. The confusion matrix for the from-scratch model reveals specific areas of difficulty.

- **Misclassifications:** Off-diagonal clusters indicate confusion between visually similar classes. For example, a model might confuse a small cyst with a periapical lesion or misclassify different types of restorations. These errors highlight the need for more discriminative features or targeted data augmentation for these specific classes.
- **Missed Detections (False Negatives):** Visual inspection of challenging cases (Fig. 3) shows that the model sometimes misses subtle pathologies, such as small carious lesions obscured by overlapping anatomy or faint radiolucencies. This is a critical area for improvement, as early detection is clinically vital.
- **Localization Inaccuracies:** While the mAP50 scores are high, the performance drop for the more stringent mAP50-95 metric indicates that precise delimitation of the bounding box remains a challenge, especially for pathologies with ambiguous borders.

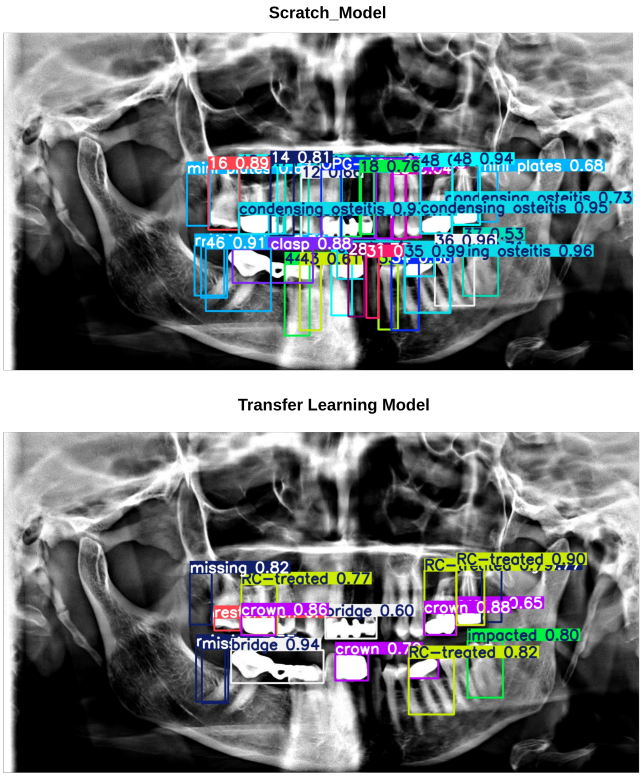

Fig. 3: Predictions on the new dataset comparing transfer learning (top) and from-scratch (bottom) models.

C. Clinical Implications

The successful development of this automated framework has significant clinical implications.By reducing the time required for manual interpretation, it can improve workflow efficiency in busy dental practices, allowing professionals to focus more on patient care and complex treatment planning. Furthermore, by facilitating the reliable detection of subtle or early-stage pathologies, this technology has the potential to improve patient outcomes through timely intervention and prevention of disease progression.

VI. Conclusion

This research has demonstrated the significant potential of object detection based on deep learning for automated classification of dental diseases on panoramic X-rays. Our work has shown that data quality is the paramount factor for developing robust models and that data-centric approaches, such as our novel multi-model pseudo-labeling pipeline, can effectively overcome the challenges of imperfect real-world datasets. Furthermore, we provided strong evidence that for high-quality comprehensive datasets, training models from scratch can yield superior performance compared to transfer learning, a crucial insight for future research strategies in medical AI.

Future work should focus on refining the pseudo-labeling pipeline with more sophisticated aggregation techniques, exploring advanced architectures to improve the detection of subtle pathologies, and integrating explainable AI (XAI) methods to build clinical trust. Ultimately, the successful translation of this technology into clinical practice will require rigorous validation through prospective clinical trials. While initial results are promising, future work will involve k-fold cross-validation to further quantify training variance. The methodologies and findings presented here lay a strong foundation for the development of accurate, efficient and clinically relevant AI-powered diagnostic tools that can revolutionize the field of digital dentistry.

References

[1] L. Kryvenko, O. Krylova, V. Lukin, and S. Kryvenko, "Intelligent visually lossless compression of dental images," *Advanced Optical Technologies*, vol. 13, p. 1306142, 2024.

[2] J. H. Lee, D. H. Kim, S. G. Kim, and H. J. Lee, "Deep learning-based dental caries detection on panoramic radiographs," *Journal of Clinical Medicine*, vol. 7, no. 12, p. 556, 2018.

[3] P. T. F. Pauwels, R. Jacobs, M. A. Singer, and R. Mupparapu, "A deep learning convolutional neural network for the detection of periapical lesions on intraoral radiographs," *Journal of Endodontics*, vol. 45, no. 5, pp. 564–569, 2019.

[4] M. Tuzoff, L. Tuzova, V. G. Voronin, A. A. G. Gavrilov, and A. V. S. Svirin, "A study of object detection on panoramic dental x-ray images," in *2019 IEEE 21st Conference on Business Informatics (CBI)*, vol. 1, 2019, pp. 493–499.

[5] Y. J. Choi, S. S. Lee, H. S. Kim, J. H. Park, and S. C. Kim, "Yolov5-based object detection for dental implant components on periapical radiographs," *Imaging Science in Dentistry*, vol. 52, no. 1, pp. 57–65, 2022.

[6] A. Z. Abidin, B. A. B. Ali, N. H. A. N. Azmi, and R. Hassan, "The impact of data quality on the performance of deep learning models for medical image analysis," *PeerJ Computer Science*, vol. 7, p. e527, 2021.

[7] V. Cheplygina, M. J. P. van Grinsven, and J. P. W. Pluim, "Not-so-supervised: a survey of the impact of imperfect ground truth in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 270–284, 2019.

[8] T. Ekert, F. Krois, J. T. T. S. Diegritz, R. F. K. Schneider, V. M. J. W. Z. Waldeyer, and C. H. F. Hämmerle, "Deep learning for the radiographic detection of apical lesions," *Journal of Endodontics*, vol. 45, no. 7, pp. 917–922, 2019.

[9] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, J. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic x-ray images," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018, pp. 400–407.

[10] Y. Miki, K. Muramatsu, T. Hayashi, T. Zhou, H. Hara, and K. Katsumata, "Classification of periodontal bone loss in panoramic radiographs using a deep convolutional neural network," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 124, no. 3, pp. 319–325, 2017.