Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence–Assisted Dental Implant Brand Identification for Prosthodontic Decision Support

Aliye İpek Kuşçu

a.ipek.kuscu@bozok.edu.tr

Yozgat Bozok University

Gökalp Çınarer

Yozgat Bozok University

Fatma Nur Ortataş

Yozgat Bozok University

Ayça Dolar

Yozgat Bozok University

Süha Kuşçu

Yozgat Bozok University

Research Article

Keywords: Dental implant identification, Radiographic diagnosis, Prosthodontic decision support, Feature-level fusion, Deep learning-assisted imaging, Hybrid classification model

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11636559/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Artificial Intelligence–Assisted Dental Implant Brand Identification for Prosthodontic Decision Support

Aliye İpek Kuşçu[1*], Gökalp Çınarer[2], Fatma Nur Ortataş[3], Ayça Dolar[4], Süha Kuşçu[5]

[1] Department of Prosthodontics, Faculty of Dentistry, Yozgat Bozok University, 66100 Yozgat, Türkiye

[2,3,4] Department of Computer Engineering, Faculty of Engineering-Architecture, Yozgat Bozok University, 66100 Yozgat, Türkiye

[5] Department of Prosthodontics, Faculty of Dentistry, Yozgat Bozok University, 66100 Yozgat, Türkiye

*Corresponding author: a.ipek.kuscu@bozok.edu.tr

Co-author e-mails: gokalp.cinarer@bozok.edu.tr; f.nur.ortatas@bozok.edu.tr; ayca.dolar@yobu.edu.tr; suha.kuscu@yobu.edu.tr

Abstract

Background: Identifying dental implant brands from radiographs is important for prosthetic treatment planning when implant records are unavailable. However, visual similarities between implant systems complicate identification. This study evaluated a hybrid framework combining deep and handcrafted radiographic features for implant brand classification.

Methods: A publicly available radiographic dataset was converted into 2,688 implant region-of-interest images representing BEGO, BICON, and ITI. A group-aware split assigned 2,149 images to training and 539 to testing, keeping images from the same source radiograph together. Deep learning, handcrafted feature-based, and feature fusion approaches were

compared. The proposed model combined ConvNeXt-Tiny embeddings with texture, shape, and grayscale descriptors and classified the fused features using logistic regression. Performance was assessed using the McNemar test and bootstrap confidence intervals based on 1,000 resamples.

Results: The proposed model achieved an accuracy of 98.5%, a macro F1-score of 0.974, a macro recall of 0.979, and a macro precision of 0.969. Recall for the underrepresented BEGO class was 96.2%. The McNemar comparison was statistically significant against Swin Transformer ($p = 0.0118$), but not against ConvNeXt ($p = 0.2891$).

Conclusions: Combining deep and handcrafted features showed potential for identifying implant brands within this three-brand dataset and may support prosthodontic decision-making. External validation on larger, multicenter datasets covering more implant systems is required before clinical deployment.

Keywords: Dental implant identification; Radiographic diagnosis; Prosthodontic decision support; Feature-level fusion; Deep learning-assisted imaging; Hybrid classification model

Background

Dental implant supported prosthetic rehabilitation is widely recognized as a predictable and effective treatment modality for the replacement of missing teeth, demonstrating high long-term survival rates and clinically acceptable functional and esthetic outcomes [1,2]. As the global prevalence of implant therapy continues to rise, an increasing number of patients present to prosthodontic clinics for maintenance, repair, or renewal of implant-supported restorations that were placed in different institutions or healthcare systems [3]. In these clinical scenarios, essential information regarding the implanted system—most notably the implant brand and model—is frequently unavailable, posing a significant challenge to

subsequent prosthetic treatment planning [4]. Accurate identification of the implant system is a prerequisite for prosthodontic intervention, as dental implants are manufactured with brand-specific designs, including proprietary connections, abutments, impression copings, and prosthetic screws [5]. This situation is not uncommon in routine prosthodontic practice, particularly among patients who received implant treatment many years earlier or at different dental clinics, where complete documentation or implant passports may no longer be available. Previous reviews have identified implant recognition as a persistent clinical challenge that frequently necessitates radiographic assessment and specialized identification methods before prosthetic treatment can proceed [4,5]. When implant brand information cannot be retrieved from patient records, clinicians are often unable to proceed with prosthetic rehabilitation, resulting in increased chairside time, repeated radiographic examinations, trial-and-error component selection, or invasive exploratory procedures [6]. Conventional approaches, such as contacting previous clinicians or relying on subjective radiographic interpretation, are often time-consuming, unreliable, and highly dependent on clinician experience, particularly given the continuous evolution and diversification of implant systems available on the market [7]. Recent advances in artificial intelligence (AI), particularly convolutional neural network–based deep learning, have demonstrated substantial potential in medical and dental analysis [8,9]. In dentistry, AI models have been successfully applied to panoramic and periapical radiographs for tasks including tooth detection, anatomical landmark identification, implant detection, and complication assessment [10–12]. More recently, deep learning–based approaches have also been successfully employed for automated tooth identification and dental image classification, highlighting the growing role of AI-assisted decision support systems in dental practice. Furthermore, previous studies have emphasized that robust validation strategies, dataset diversity, and clinically realistic evaluation protocols are essential for ensuring reliable

translation of AI systems into routine clinical workflows [13]. In parallel with these developments, several studies have reported promising results for automated implant classification, indicating that deep learning models are capable of extracting subtle radiographic features related to implant morphology that may not be readily discernible to the human eye [14–16]. Prior investigations using panoramic radiographs have achieved high classification accuracy across a limited number of implant brands, with methodological enhancements such as attention-based architectures, multitask learning, and synthetic data generation further improving performance under controlled conditions [17–22]. These studies collectively demonstrate the technical feasibility of automated implant identification and highlight the growing role of AI in implant-related diagnostics. Despite these advances, several critical limitations remain. Most existing studies have primarily focused on maximizing algorithmic performance under varying experimental conditions; however, substantial heterogeneity exists in dataset preparation, preprocessing strategies, data partitioning protocols, and validation methodologies, making direct comparison across studies difficult and limiting assessment of real-world generalizability [23,24]. Although recent evidence indicates that AI-based systems can achieve excellent performance in dental implant identification — often reporting accuracies above 90% — these investigations remain predominantly based on retrospective, single-center datasets with limited external validation. Furthermore, most studies rely exclusively on end-to-end deep learning architectures, while the potential complementarity between deep feature representations and implant-specific handcrafted radiographic descriptors remains largely unexplored [25,26]. Therefore, developing more robust, reproducible, and clinically interpretable frameworks for implant identification continues to represent an important methodological gap in the literature. More importantly, the clinically common and highly significant scenario where patients seek prosthetic rehabilitation but lack knowledge about implant systems, preventing them from

initiating treatment, is limited in the literature. As a result, the translation of AI-based implant identification into practical decision-support tools for prosthodontists remains insufficient. Furthermore, the continuous introduction of new implant systems poses an additional challenge for supervised deep learning models, which rely heavily on large volumes of clinically acquired radiographic data [27]. Ensuring adequate representation of diverse implant brands across varying clinical contexts remains difficult, potentially limiting the generalizability and scalability of existing approaches [28]. Recent attempts to address this issue through artificial image generation and data augmentation have shown promising results; however, these efforts have largely focused on improving classification accuracy rather than supporting clinically relevant decision-making processes [29,30]. Deep learning–based approaches have also demonstrated promising results across a broad spectrum of dental applications, including automated orthodontic treatment planning, further underscoring the potential of artificial intelligence as a decision-support tool in clinical dentistry [31]. It should be noted that the dataset used in this study was obtained from a publicly available open-access source rather than being generated by the authors. The primary motivation of this research stems from a clinically relevant challenge: even experienced prosthodontists and oral radiologists may encounter difficulties in accurately identifying implant brands solely from radiographic images due to the substantial visual similarities among implant systems. Although previous studies have reported promising classification performances, most investigations have primarily focused on improving algorithmic accuracy. Comparisons across studies remain difficult because of heterogeneous evaluation protocols, differences in preprocessing strategies, data partitioning methods, and limited external validation. These methodological differences complicate the assessment of real-world generalizability and clinical applicability. First, the majority of prior studies employ random image-level data splitting without accounting for the fact that multiple ROI crops may be extracted from the

same panoramic radiograph or the same physical implant instance. When such images are distributed across both training and test subsets, the classifier may encounter near-duplicate samples during evaluation that share visual characteristics with training data — a form of data leakage that artificially inflates performance estimates and undermines the validity of reported generalization. To the best of our knowledge, no prior study on dental implant brand classification has explicitly addressed this source of bias through a group-aware partitioning strategy. Second, existing frameworks rely either exclusively on end-to-end deep learning or on handcrafted features in isolation, without systematically investigating the complementary benefits of combining domain-informed morphological descriptors with deep representations within a unified feature space. Although feature-level fusion has demonstrated advantages in other medical image classification domains, this approach has not been systematically applied to dental implant brand identification from panoramic radiographs. In chest X-ray classification, Malik et al. reported that fusing handcrafted descriptors with deep convolutional features significantly improved diagnostic accuracy across ten disease categories [32]. Similarly, Shankar and Perumal demonstrated that integrating LBP-based handcrafted features with InceptionV3 deep representations enhanced COVID-19 classification performance [33]. In breast tumor classification, Cruz-Ramos et al. combined HOG- and LBP-based handcrafted features with DenseNet201 embeddings, achieving superior discrimination between benign and malignant lesions [34]. Almaraz-Damian et al. further showed that fusing shape, color, and texture handcrafted features with CNN-derived representations improved melanoma detection accuracy [35]. Collectively, these findings indicate that handcrafted descriptors capture interpretable microstructural and textural cues that complement the high-level semantic representations of deep networks — a complementary relationship particularly advantageous in scenarios with high inter-class visual similarity and limited training data. Third, classification performance for

underrepresented implant classes — which are clinically critical precisely because of their rarity — is rarely optimized or even reported separately, limiting the clinical reliability and interpretability of existing models. The present study is specifically designed to address each of these gaps within a unified framework. Therefore, the aim of the present study was to develop and evaluate a deep learning–based framework for automated dental implant brand identification from radiographic images, with a focus on clinically relevant prosthodontic applications. By utilizing an openly available dataset comprising three implant brands, this study seeks to assess the feasibility of AI-assisted implant brand recognition as a practical decision-support tool for prosthodontists when implant system information is unavailable. Unlike conventional implant classification studies that rely exclusively on deep learning representations or handcrafted descriptors, the proposed framework integrates implant-specific morphological features with ConvNeXt-derived deep embeddings to exploit complementary radiographic information. Furthermore, the framework was evaluated using a group-aware data splitting strategy specifically designed to prevent information leakage between ROI samples originating from the same implant instance, thereby providing a more reliable assessment of model generalization. Although feature-fusion approaches have been widely investigated in medical image analysis, the present study focuses on the design and validation of an implant-specific radiographic identification framework that combines complementary handcrafted and deep representations under a leakage-aware evaluation protocol. Ultimately, this approach aims to facilitate timely and appropriate prosthetic rehabilitation while reducing diagnostic uncertainty and clinical inefficiency. The null hypothesis was that there is no significant difference between deep learning–based implant brand identification and chance-level or non-assisted identification in terms of diagnostic performance and clinical usability.

Methods

Dataset

The dental implant images used in this study were obtained from an openly available dataset provided through the Roboflow platform [36]. The original dataset consists of images belonging to three different implant brands: BEGO, BICON, and ITI. The raw dataset contains a total of 2,064 annotated images. Since the dataset was provided in object detection format, each source image may contain more than one implant instance. Therefore, within the scope of this study, separate regions of interest (ROIs) were generated for each annotation in order to transform the data into a structure suitable for a classification problem. As a result of ROI extraction and reorganization procedures, a total of 2,688 ROI images were obtained based on direct scanning of the file system. The class distribution within the constructed ROI dataset is as follows: 1,470 ROI images belong to the Bicon class, 1,047 ROI images belong to the ITI class, and 171 ROI images belong to the BEGO class. Since the original panoramic and periapical radiographs may contain multiple implant instances, separate regions of interest were generated for each annotation in order to construct a dataset structure appropriate for the classification task. Fig. 1 illustrates representative raw images from the three implant brands (BEGO, BICON, and ITI) together with the corresponding ROI samples extracted from these images. Although these three implant brands share a broadly similar radiographic appearance, they exhibit distinctive morphological characteristics that can serve as discriminative cues for classification. Table 1 summarizes the key radiographic differences across brands in terms of overall body profile, neck region, thread appearance, body-to-neck transition, and apical appearance. These structural differences, while identifiable under favorable imaging conditions, exhibit considerable overlap in panoramic radiographs due to variations in imaging angle, resolution, and patient anatomy, which collectively contribute to

the classification challenge addressed by the proposed framework. To further assess the reliability of the implant labels provided in the open-access dataset, an additional verification procedure was conducted. Original implant samples corresponding to the investigated implant systems were randomly procured directly from implant manufacturers in sterile packaging. The packages were opened under the supervision of two experienced prosthodontists, and standardized radiographic images of the implants were acquired. These radiographs were subsequently analyzed using the proposed classification framework and compared with the corresponding implant labels. The obtained results demonstrated consistency between the known implant identities and the model predictions, providing additional support for the correctness of the implant brand annotations contained in the publicly available dataset.

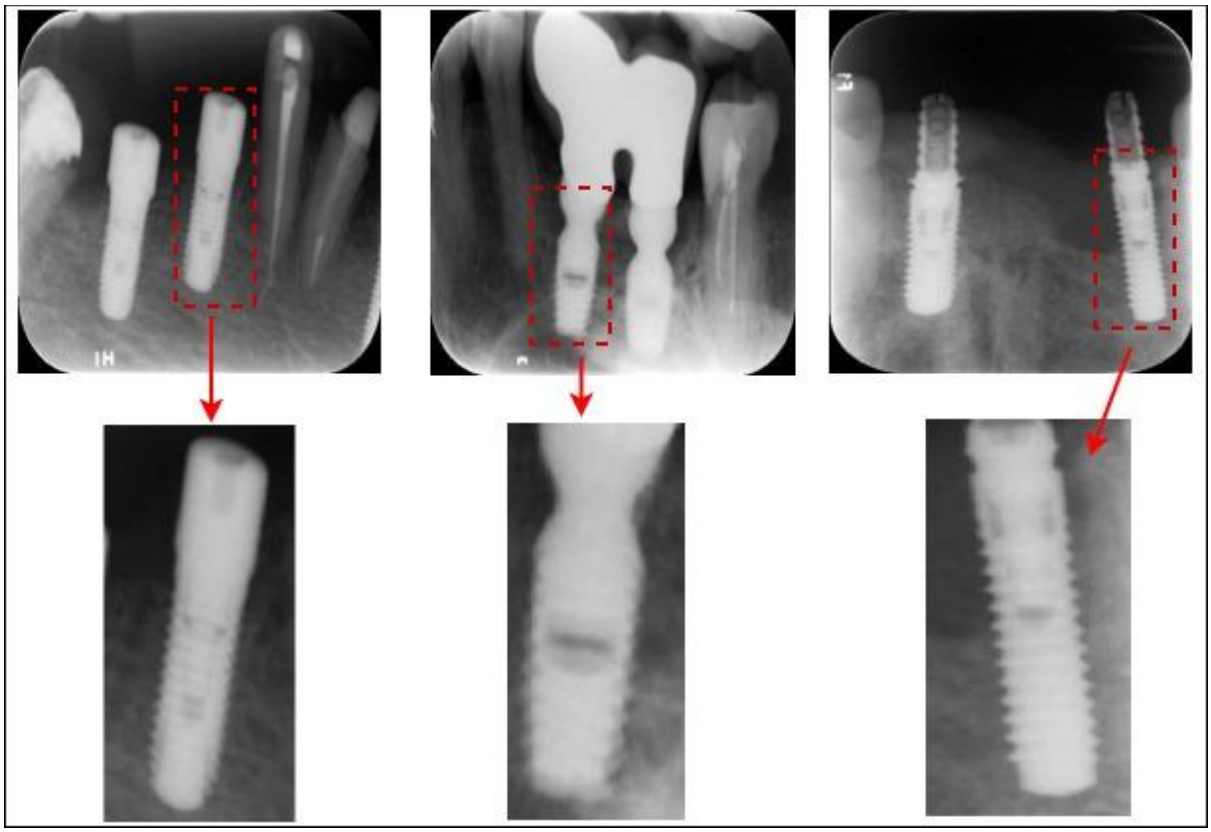

Figure 1. Representative raw dental radiographs and corresponding implant regions of interest.

Table 1. Radiographic morphological characteristics of BEGO, BICON, and ITI implant brands.

Characteristic	BEGO	BICON	ITI
Overall body profile	Predominantly cylindrical	Compact cylindrical profile	Slightly tapered cylindrical profile
Neck region	Relatively wider	Compact transition	Narrower profile
Thread appearance	Regular thread pattern	Less pronounced thread structure	More prominent thread structure
Body-to-neck transition	Abrupt	Smooth	Gradual
Apical appearance	Rounded	Shorter profile	More elongated profile
Potential discriminative cues	Neck geometry, thread spacing	Body contour, neck transition	Thread depth, tapering profile

These observations are based on representative radiographic examples contained in the dataset and are intended to provide qualitative morphological references rather than manufacturer specifications.

Group-based data structure and splitting strategy

It was anticipated that using ROIs derived from the same source image or the same implant instance together in both the training and testing stages could artificially increase model performance due to data leakage. Therefore, a group identity (group_id) derived from the file name was defined for each ROI image. The group identity ensures that ROIs belonging to the same implant instance or originating from the same source image are collected under a common identifier. The dataset was originally provided by the source in the form of training, validation, and test subsets. Within the scope of this study, ROI extraction procedures were performed separately for each subset, and the original splitting structure was preserved. However, during the model training and evaluation stages, ROI samples were reorganized and a group-based evaluation strategy was adopted. For this purpose, ROIs derived from the same source image or the same implant instance were ensured to remain within the same

group, and the dataset was partitioned in a group-based manner. During the splitting process, class labels were used as the target variable, while group identities were used as the grouping variable. The group_id for each ROI was extracted from its filename by removing the Roboflow-assigned augmentation hash suffix (e.g., from bego-0_jpg.rf.3c722e8fe0716ea1814323e54480759c, the group_id bego-0_jpg was derived). This ensures that all ROIs originating from the same source image share a common group_id. Validation confirmed that no group_id appeared in more than one original dataset split (zero cross-split leakage across all 934 unique groups). The resulting 80/20 group-aware split yielded a training set of 2,149 ROIs from 747 unique groups (BEGO: 118, BICON: 1,191, ITI: 840) and a test set of 539 ROIs from 187 unique groups (BEGO: 53, BICON: 279, ITI: 207), with zero group overlap between subsets. A fixed random seed (random_state=42) was applied throughout all splitting and training procedures to ensure full reproducibility.

Image preprocessing and data augmentation

All ROI images were resized to 224 × 224 pixels to be used as inputs for the deep learning models. In order to improve the generalization performance of the model during the training phase, data augmentation techniques were applied. Within this scope, randomly resized cropping, horizontal flipping, low-angle rotation, brightness and contrast adjustment, and low-probability Gaussian blurring operations were employed. Data augmentation procedures were applied only during the training phase in an online (on-the-fly) manner, while the test phase consisted solely of resizing, center cropping, and normalization steps, ensuring reproducibility for each input image. In both the training and testing phases, images were normalized using mean and standard deviation values based on ImageNet dataset statistics and converted into tensor format. The overall dataset preparation and preprocessing workflow followed in this study is illustrated in Fig. 2. Raw panoramic radiographs were first converted into implant-specific regions of interest (ROIs). A group-aware train–test splitting strategy

was applied to prevent data leakage between training and test sets. Following this, images were resized to 224×224 pixels and normalized using ImageNet-based mean and standard deviation values. Data augmentation was applied exclusively to the training subset in an on-the-fly manner, while the test subset underwent only resizing and normalization to ensure deterministic and reproducible evaluation. The processed images were converted into PyTorch tensors and provided as input to the ConvNeXt-Tiny network, which generated 768-dimensional deep feature embeddings. These embeddings were subsequently transferred to the feature-level fusion framework presented in Fig. 3. To improve robustness against variability commonly encountered in clinical radiographs, including differences in implant orientation, image exposure, magnification, and acquisition conditions, a set of data augmentation techniques was applied exclusively during training. These augmentations included random resized cropping, horizontal flipping, small-angle rotation, brightness and contrast adjustment, and low-probability Gaussian blurring. The objective of these transformations was to expose the model to a broader range of image appearances while preserving clinically relevant implant morphology. During testing, no stochastic augmentations were applied; instead, a deterministic preprocessing pipeline consisting of resizing and normalization was used to ensure reproducibility and unbiased performance evaluation.

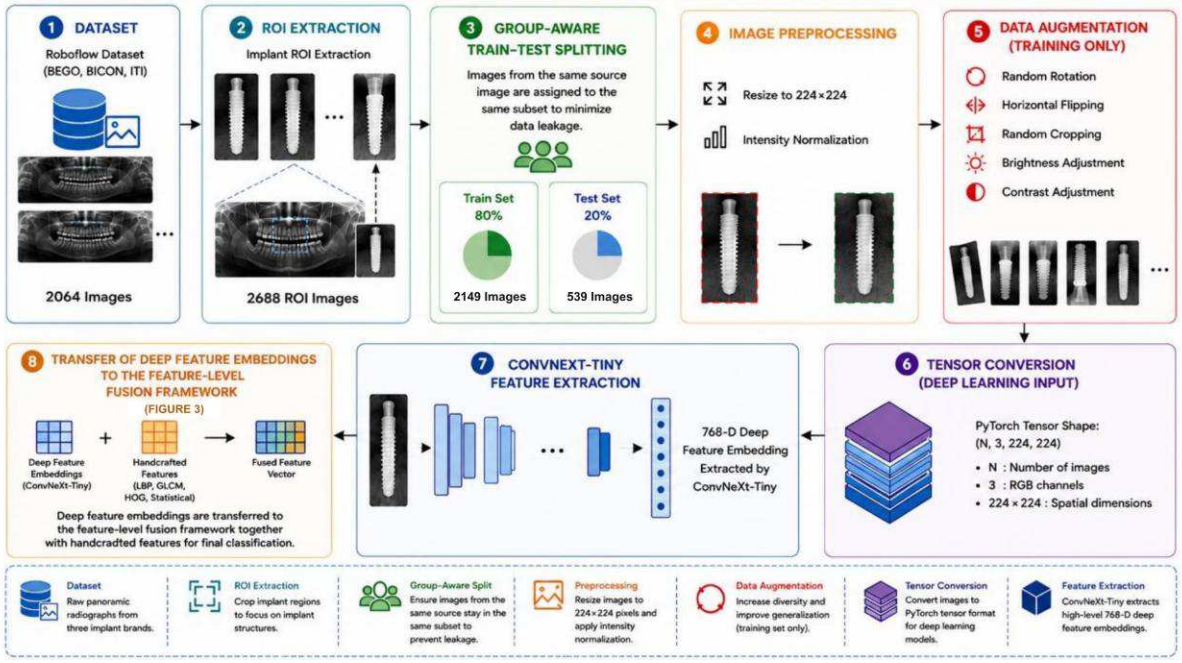

Figure 2. Overview of the dataset preparation and preprocessing pipeline.

Deep learning architectures and hyperparameters

Three different deep learning architectures were employed for the classification of dental implant brands: ResNet-50, ConvNeXt-Tiny, and Swin Transformer (Swin-Tiny). The ResNet-50 model was considered as a baseline approach representing classical convolutional neural networks [37]. ConvNeXt-Tiny and Swin-Tiny models were selected to represent modern convolutional and transformer-based designs, respectively [38,39]. All deep learning models were initialized with weights pre-trained on the ImageNet dataset, and the classification heads were restructured in accordance with the number of classes in the problem. The Adam optimization algorithm was used for model training, and an early stopping mechanism was applied in order to prevent overfitting during the learning process. For early stopping, the training process was terminated when no improvement was observed in validation performance. During the training process, the Automatic Mixed Precision (AMP) approach was utilized to improve computational efficiency and optimize memory usage. All models were trained using the same training protocol and hyperparameter settings to ensure a

fair experimental environment for comparative evaluation. The main training hyperparameters used in the study are summarized in Table 2.

Table 2. Training hyperparameters used for the deep learning models.

Hyperparameter	Value
Number of epochs	50
Batch size	32
Learning rate (LR)	1×10^{-4}
Optimization algorithm	Adam
Early stopping patience	7
Training precision	Automatic Mixed Precision (AMP)

Manual feature extraction and machine learning models

In addition to deep learning approaches, manual features were extracted from ROI-based images. These features were computed to represent implant surface texture, edge structures, and gray-level statistics. The extracted manual features were evaluated independently of the deep learning models. Classification was performed using classical machine learning algorithms based solely on manual features. Within this scope, Support Vector Machines (SVM), Logistic Regression (LR), and XGBoost algorithms were employed. In this stage, images were not directly provided to the models; instead, classification was carried out exclusively on the basis of feature vectors. The manual feature groups extracted from ROI images and the summary information of these features are presented in Table 3.

Table 3. Summary of handcrafted features extracted from ROI images.

Feature group	Description	Number of features
Grayscale intensity statistics	First-order statistical measures describing the pixel intensity distribution, including mean, standard deviation, skewness, kurtosis, minimum, and maximum values	9

Feature group	Description	Number of features
GLCM-based texture features	Texture descriptors derived from the Gray-Level Co-occurrence Matrix, including contrast, correlation, energy, and homogeneity computed across multiple directions	16
LBP features	Histogram-based Local Binary Pattern features capturing local micro-texture patterns and surface irregularities	59
HOG features	Histogram of Oriented Gradients features representing edge orientation distributions and shape-related characteristics	6084
Total		**6168**

Proposed feature fusion model

In the first stage, ROI (Region of Interest) images representing the anatomically meaningful region of the implant are taken as input and subjected to a deep feature extraction process using the ConvNeXt architecture. Through the global pooling operation in the final layers of the ConvNeXt network, deep feature embeddings that summarize spatial information and represent high-level semantic content are obtained. These representations exhibit high discriminative capability for modeling macrostructural differences specific to implant brands. In parallel with deep feature extraction, handcrafted feature extraction is performed on the same ROI images. In this context, local texture variations are captured using Local Binary Pattern (LBP), statistical texture relationships are modeled using the Gray Level Co-occurrence Matrix (GLCM), pixel distribution characteristics are quantified through grayscale intensity statistics, and edge and orientation information are computed using the Histogram of Oriented Gradients (HOG). These handcrafted features represent microstructural, orientational, and textural cues of the implant surface, thereby serving a complementary role to the deep learning representations. In the subsequent stage, the deep feature vector obtained from the ConvNeXt architecture and the handcrafted features are combined using a feature-level fusion approach. This process is carried out through concatenation of the two representations, resulting in a single fused feature vector. In this

way, the high abstraction capability of deep learning and the interpretable, domain-informed nature of handcrafted features are integrated within a shared representation space. In the final stage, the resulting fused feature vectors are fed into a machine learning–based classifier. In this study, Logistic Regression was selected due to its stable performance in high-dimensional feature spaces and its probabilistic interpretability. This design enables the model to be more robust against overfitting while improving generalization performance, particularly in limited data scenarios. For all feature fusion experiments, the dataset was split into training and test sets with an 80%–20% ratio, while preserving the group-based data splitting strategy applied in the previous stages. This approach prevents data leakage between the training and test sets by ensuring that images originating from the same sample or source are not distributed across different subsets, thereby enabling a realistic and generalizable performance evaluation. The overall workflow of the proposed model based on deep learning and handcrafted feature fusion implemented in this study is illustrated in Fig. 3. In the first stage, the ROI images corresponding to the implant region, which are used as input, are fed into a deep learning–based feature extraction process, and high-level feature representations are generated from the global pooling layer of the ConvNeXt architecture. In parallel, handcrafted texture- and shape-based features, including Local Binary Pattern (LBP), Gray Level Co-occurrence Matrix (GLCM), grayscale intensity statistics, and Histogram of Oriented Gradients (HOG), are extracted from the same ROI images. These handcrafted features provide a complementary representation that captures microstructural and orientation-related cues of the implant surface. The high-dimensional feature representations obtained from the deep learning model and the handcrafted features are then combined to form a single unified feature vector (fused feature vector). This vector is designed to simultaneously incorporate the discriminative power of deep representations and the descriptive characteristics of handcrafted features. In the final stage, the resulting fused

feature vectors are fed into a machine learning classifier, for which Logistic Regression was selected in this study.

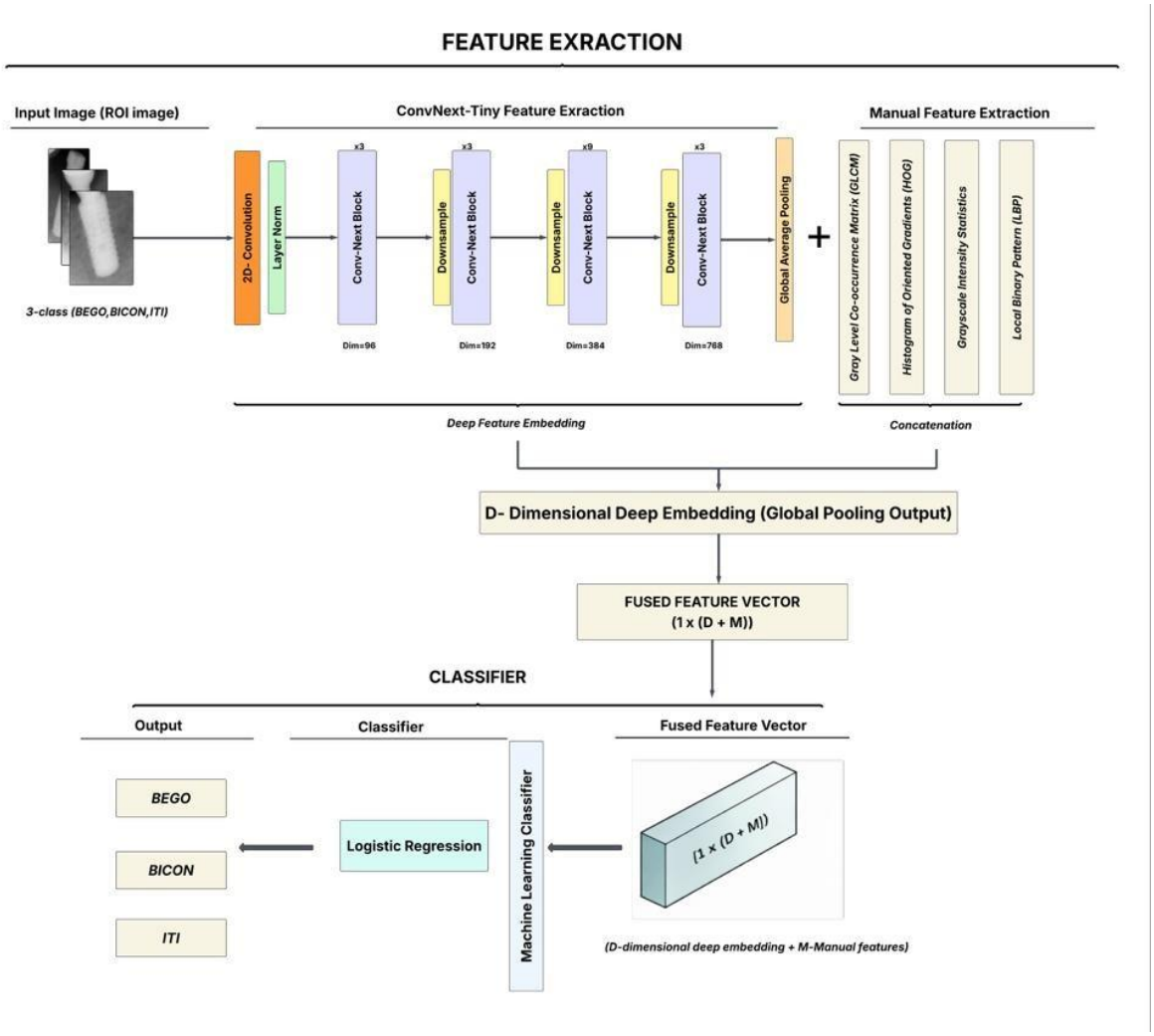

Figure 3. Feature-level fusion framework for dental implant brand classification.

Evaluation metrics

In order to evaluate the performance of the classification models applied in this study, Accuracy, Precision, Recall, and Macro-F1 score metrics were used. The combined use of these metrics aims to provide a fair, balanced, and reliable assessment of model performance in terms of both overall accuracy and class-based sensitivity, particularly in multi-class classification problems and datasets with class imbalance. Accuracy is a fundamental

performance indicator that expresses the ratio of correctly classified samples to the total number of samples and is defined as follows:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Although this metric provides a summary measure of the overall classification success of the model, it may be influenced by dominant classes in cases where there is a high imbalance in the number of samples among classes. For this reason, accuracy was not considered as a standalone decision-making criterion in this study, but rather interpreted together with other class-sensitive metrics. Precision is a metric that measures the proportion of samples predicted as positive by the model that are actually correct. It is defined as the ratio of true positive (TP) predictions to the total number of samples predicted as positive and is mathematically expressed as follows:

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

A high precision value indicates that the model produces a low number of false positive predictions and demonstrates reliability in its positive class assignments. Recall is a metric that measures how many of the actual samples belonging to a given class are correctly identified by the model and is defined as the ratio of true positive (TP) predictions to the total number of actual samples of that class [32]. Mathematically, the Recall metric is expressed as follows:

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

Here, FN represents the cases in which the model incorrectly classifies samples belonging to the relevant class as false negatives. A high recall value indicates that the rate of missed samples for the corresponding class is low and that false negative predictions are minimized. The F1 score is a performance indicator that jointly considers precision and recall metrics and

reflects the balance between these two measures [32]. The F1 score used in this study is calculated based on true positive, false positive (FP), and false negative predictions as follows:

$$F1\ score = TP / (TP + 0.5\ (FP + FN)) \quad (4)$$

This formulation corresponds to the harmonic mean of precision and recall values and aims to evaluate the model in a balanced manner in terms of both classification accuracy and sensitivity. In multi-class classification problems, the Macro-F1 score is obtained by calculating the F1 score separately for each class and then taking the arithmetic mean of these scores, assigning equal weight to all classes regardless of their sample sizes in the dataset. This approach prevents dominant classes from disproportionately influencing the performance evaluation and ensures that the model performance for classes with fewer samples is also fairly reflected. The combined use of Accuracy, Recall, Precision, and Macro-F1 metrics provides a comprehensive evaluation of whether the proposed models exhibit not only high overall classification success but also sensitive, balanced, and generalizable performance across all classes.

Statistical analysis

In this study, statistical hypothesis tests were applied in order to evaluate whether the performance differences observed among different classification approaches were statistically significant. All statistical analyses were performed using only the model predictions obtained on the test set and were not incorporated into the training process in any manner. In the comparison of deep learning–based models and the proposed feature-level fusion model, paired classification results obtained on the same test samples were taken into consideration. Within this context, the McNemar test was employed to assess the difference between the error patterns of two models. Although the chi-square test is often preferred in paired

binomial data analysis, the McNemar test is generally more appropriate but is frequently overlooked [40]. The McNemar test is suitable for comparing categorical predictions produced on the same data samples and is widely used, particularly in classification problems [41]. The test was applied in a two-sided manner, and a p-value of < 0.05 was accepted as the threshold for statistical significance. In addition, the bootstrap resampling method was used in order to more robustly evaluate potential variation in performance metrics and the impact of class imbalance [42]. Bootstrap sampling was performed with 1,000 iterations over the test set, and Macro-F1 scores and class-based recall values were calculated for each iteration. Based on the resulting distributions, 95% confidence intervals were determined, enabling a more reliable comparison of the generalization performances of the models. Through the applied statistical analyses, not only numerical performance differences but also whether these differences were due to random variation were evaluated, and it was aimed to statistically support the performance gains achieved by the proposed model. The McNemar test was specifically selected over alternative tests (e.g., chi-square) because it is designed for paired categorical outcomes—that is, for comparing predictions of two classifiers evaluated on the same test samples under identical conditions—and has been demonstrated to be more appropriate for classifier comparison in this setting (reference [43]). Bootstrap resampling (1,000 iterations) was preferred over parametric confidence intervals because the test set exhibits class imbalance (BEGO: n=53 vs. BICON: n=279), which may violate the normality assumptions of parametric methods; bootstrap-based confidence intervals are distribution-free and provide robust interval estimates under such conditions (reference [44]). In addition, a 5-fold Stratified Group K-Fold cross-validation (StratifiedGroupKFold, k=5, random_state=42) was applied to the training set (n=2,149 ROIs, 747 unique groups) to further assess model stability, ensuring group integrity was preserved across folds. The cross-validation Macro-F1 scores for the proposed model were: Fold 1=0.9922, Fold 2=1.0000, Fold 3=0.9922, Fold

4=0.9919, Fold 5=0.9983 (mean±SD: 0.9949±0.0035). The narrow standard deviation confirms high stability across folds, and the mean cross-validation performance is consistent with the held-out test set result (Macro-F1=0.9738), confirming the absence of overfitting. In addition to bootstrap resampling, k-fold cross-validation was applied to further assess the stability and generalizability of the proposed model. The mean and standard deviation of performance metrics across folds are reported to provide a more comprehensive evaluation of model robustness.

Results

In this section, the performance of the proposed feature-level fusion–based hybrid model is presented in comparison with deep learning and manual feature–based approaches that are widely used in the literature. All experiments were conducted on the ROI-based dataset constructed using the group-based data splitting strategy, and the models were tested under the same preprocessing, data augmentation, and evaluation protocols. Performance evaluations were carried out using Accuracy, F1-score, Macro Precision, and Macro Recall metrics, primarily considering the impact of class imbalance among classes. Since the BEGO class contains a limited number of images, its performance results are also presented separately in the tables. The findings are first presented through the results of deep learning–based models; subsequently, the results of manual feature–based machine learning methods, the comparative analysis of these two approaches, and finally the performance of the proposed feature fusion–based hybrid model are detailed.

Performance of deep learning models

First, in order to evaluate the ability of deep learning–based architectures alone to distinguish implant brands, different architectures were tested under the same experimental conditions.

Within this scope, ResNet-50 was selected as a reference model representing the classical convolutional neural network (CNN) design, which addresses the gradient vanishing problem that emerges with increasing depth through residual connections [37]. ResNet-50 provides a fundamental benchmark, as it is a mature architecture widely used in medical image analysis and frequently preferred for comparison purposes. The ConvNeXt model was evaluated as an architecture that preserves the pure CNN structure while adopting modern design principles inspired by transformer architectures, such as large kernels, layer normalization, and simplified block structures [38]. The purpose of selecting this model was to analyze to what extent representations obtained through the modernization of classical CNN structures can more effectively capture subtle morphological differences between implant brands. In addition, the Swin model was examined as a transformer-based approach capable of modeling both local and global visual context through its hierarchical window-based self-attention mechanism [39]. The inclusion of Swin-Tiny in the evaluation aims to comparatively reveal the performance differences between convolutional architectures and transformer-based approaches, particularly in the context of implant images characterized by limited data and high visual similarity. This analysis aims to isolate the contribution of deep representations to classification performance. All models were trained on the same ROI-based dataset and evaluated on previously unseen test data. In this way, the generalization capabilities of the models were made directly comparable. Table 4 presents a comparative overview of the performances of ResNet-50, Swin Transformer, and ConvNeXt architectures for the classification of dental implant brands. The ResNet-50 model demonstrates that classical convolutional architectures provide a reasonable performance in the general discrimination of implant brands, achieving an accuracy of 92.9%. However, the Macro-F1 score remaining at the level of 89.9% and, in particular, the recall value of 83% for the BEGO class suggest that this architecture is limited in capturing subtle morphological

differences between classes. The Swin Transformer model achieved a significant improvement in overall classification performance compared to ResNet-50, with an accuracy of 96.3%. Nevertheless, the Macro Recall value remaining at 89.8% and the recall value for the BEGO class decreasing to 72% indicate that this architecture produces a high false negative rate for certain classes. This situation may be attributed to the inability of the window-based attention mechanism to adequately capture distinctive local cues belonging to specific classes under limited data scenarios. The ConvNeXt model, on the other hand, achieved the most balanced and highest performance across all metrics, with 97.8% accuracy, a 96.3% Macro-F1 score, and a BEGO recall value of 92%. The fact that both Macro Precision and Macro Recall values are very close to each other reveals that ConvNeXt provides better inter-class balance and is able to represent each implant brand consistently. These results demonstrate that the ConvNeXt architecture, which adopts modern CNN design principles, is able to model both macro- and micro-structural cues related to the implant surface more effectively. When evaluated overall, it is concluded that although the transformer-based Swin architecture exhibits strong performance in terms of overall accuracy, it produces class-based imbalances; whereas the ConvNeXt architecture provides both high overall success and balanced performance across classes, justifying its selection as the feature extractor in subsequent stages.

Table 4. Test performance of the deep learning models.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score	BEGO Recall
ResNet-50	0.929	0.901	0.898	0.899	0.83
Swin Transformer	0.963	0.950	0.898	0.919	0.72
ConvNeXt	0.978	0.963	0.962	0.963	0.92

Performance of manual feature-based machine learning algorithms

In this section, the classification performances of classical machine learning algorithms trained using manual features extracted from ROI images are presented. At this stage, Support Vector Machines (SVM), Logistic Regression (LR), and XGBoost algorithms were evaluated solely on the basis of manually extracted feature vectors, independently of deep learning models. The aim is to reveal to what extent manual features alone provide discriminative information and to establish a basis for comparison with deep learning approaches. The obtained test results are summarized in Table 5. The Logistic Regression model achieved a Macro-F1 score of 0.6485 on the test set, and class-based analysis revealed that it exhibited particularly low recall for the Bego class (recall = 0.30). The XGBoost model, on the other hand, reached a Macro-F1 score of 0.6576 in the test phase and demonstrated higher performance especially in the Bicon (F1 = 0.91) and ITI (F1 = 0.82) classes. However, it was also observed that the recall value for the Bego class remained low in the XGBoost model as well (recall = 0.15). The SVM model trained using manual features exhibited the highest overall performance among the classical methods evaluated, with a Macro-F1 score of 0.6750 on the test set. It is noteworthy that the SVM model provided a more balanced precision–recall distribution for the Bego class despite class imbalance (F1 = 0.44). Nevertheless, in all manual feature–based models, performance remained limited in underrepresented classes, and errors were observed to be concentrated mostly among visually similar implant brands. When evaluated overall, it is observed that manual features provide a certain level of discriminative information between implant classes; however, the achieved performances remain considerably behind the deep learning models reported in previous sections. This finding indicates that the use of manual features alone is limited, but they may play a complementary role when combined with deep learning representations. This situation provides a strong rationale for the feature fusion approach addressed in the following section.

Table 5. Test performance of manual feature-based machine learning models.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score	BEGO Recall
Manual Features + Logistic Regression	0.768	0.660	0.642	0.648	0.30
Manual Features + SVM	0.764	0.676	0.674	0.675	0.43
Manual Features + XGBoost	0.833	0.760	0.651	0.658	0.15

Comparison of deep learning and manual features

In order to evaluate the effectiveness of the feature-level fusion approach that combines ConvNeXt-based deep features with manual texture and shape features, the resulting fused feature vectors were classified using XGBoost and Support Vector Machines (SVM) with an RBF kernel. The results presented in Table 6 indicate that both classifiers achieved high overall classification performance, while exhibiting certain differences in terms of inter-class balance. In this section, the classification performances of deep learning models trained on ROI-based images and classical machine learning approaches based on manual features are also compared with the proposed model. The purpose of this analysis is to quantitatively demonstrate the limitations of the discriminative information provided by manual features alone and the gains achieved through end-to-end learned deep representations. As seen in Table 6, classification of the fused feature vectors constructed using ConvNeXt and manual features with XGBoost resulted in 97.4% accuracy and a 95.4% Macro-F1 score. These results demonstrate that feature fusion significantly improves overall classification performance. However, the recall value of 89% for the BEGO class indicates that the tree-based decision mechanism is not able to sufficiently capture fine discriminative boundaries for certain classes. When SVM was used, accuracy increased to 97.8% and the Macro-F1 score rose to 95.7%. This improvement shows that SVM is able to model the high-dimensional and heterogeneous fused feature space more effectively due to its ability to construct non-linear decision boundaries. Nevertheless, the recall value for the BEGO class

remaining at 89% again reveals that false negatives for certain classes cannot be completely eliminated despite the fusion of manual and deep features. When evaluated overall, although XGBoost- and SVM-based hybrid models exhibit strong performance, they indicate that simpler and more regularized classifiers may be advantageous in terms of providing a more balanced representation across classes. This consideration has been influential in the preference for the proposed Logistic Regression–based hybrid model. The selection of Logistic Regression as the final classifier was not based solely on model simplicity but was supported by empirical evidence obtained during the experimental evaluation. As shown in Table 6, the Logistic Regression–based fusion model outperformed both SVM- and XGBoost-based fusion alternatives in terms of Macro-F1 score, Macro Recall, and BEGO recall. These findings indicate that the fused feature representation exhibited a sufficiently discriminative structure that could be effectively separated using a linear decision boundary. Consequently, the use of more complex classifiers did not provide additional performance gains and, in some cases, resulted in slightly lower class-balanced performance. Therefore, Logistic Regression was selected as the final classifier due to its superior empirical performance, probabilistic interpretability, computational efficiency, and strong generalization characteristics. Feature-level fusion models that combine deep learning representations with manual features outperformed standalone deep learning approaches across all evaluation metrics. In particular, models based on ConvNeXt and manual feature fusion provided consistent improvements in Macro Recall and Macro-F1 scores. Among these models, the proposed model stands out as the most successful structure in terms of both overall performance and class balance, with 0.974 Macro-F1, 0.979 Macro Recall, and 0.96 BEGO recall values. These findings demonstrate that combining the discriminative representations of deep learning with the explanatory power of manual features provides

559 more reliable and generalizable classification performance, particularly for clinically critical
560 and underrepresented classes.

561 **Table 6.** Test performance of machine learning, deep learning, and feature fusion models.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score	BEGO Recall
Manual Features + Logistic Regression	0.768	0.660	0.642	0.648	0.30
Manual Features + SVM	0.764	0.676	0.674	0.675	0.43
Manual Features + XGBoost	0.833	0.760	0.651	0.658	0.15
ResNet-50	0.929	0.901	0.898	0.899	0.83
Swin Transformer	0.963	0.950	0.898	0.919	0.72
ConvNeXt	0.978	0.963	0.962	0.963	0.92
ConvNeXt + Manual Features (Fusion) + XGBoost	0.974	0.958	0.951	0.954	0.89
ConvNeXt + Manual Features (Fusion) + SVM (RBF)	0.978	0.961	0.953	0.957	0.89
Proposed Model	**0.985**	**0.969**	**0.979**	**0.974**	**0.96**

562 The confusion matrix presented in Fig. 4 reveals the class-based error distribution of the
563 proposed feature-level fusion model. In the matrix, it is observed that misclassifications are
564 largely concentrated between specific class pairs, while inter-class confusion remains
565 generally limited. In particular, for samples belonging to the BEGO class, most of the
566 misclassifications are associated with the ITI class, whereas confusion with the BICON class
567 remains quite limited. This situation indicates that the model produces errors in a manner
568 sensitive to the structural similarities between implant brands.

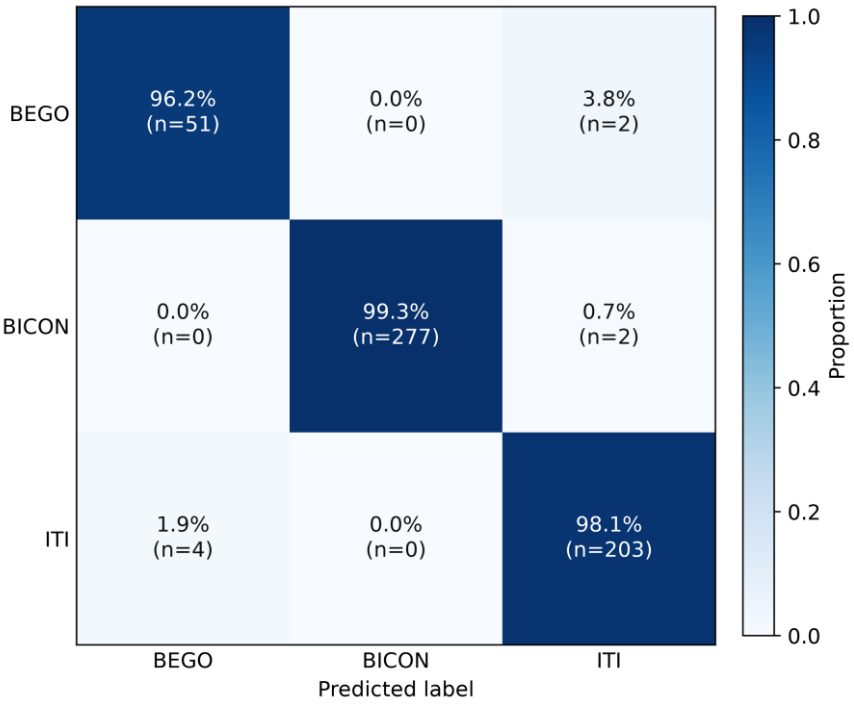

Figure 4. Row-normalized confusion matrix of the proposed logistic regression feature fusion model.

Fig. 5 presents class-specific Receiver Operating Characteristic (ROC) curves for the proposed feature-fusion model, generated using a one-vs-rest strategy based on the probability outputs of the Logistic Regression classifier. Separate curves are shown for the BEGO, BICON, and ITI implant classes, along with the macro-average ROC curve. The fact that most of the curves are located close to the upper left corner indicates that the model achieves a high balance between sensitivity (True Positive Rate, TPR) and specificity (False Positive Rate, FPR) in distinguishing different implant brands. In particular, the ROC curve for the BEGO class, which exhibits consistently high AUC values, shows that the model provides reliable discrimination performance even under limited sample conditions. All AUC values are reported with five-decimal precision to minimize rounding artifacts and improve reproducibility across classes with closely similar performance. The proposed Logistic Regression–based fusion model produced the most balanced and reliable results, with a Macro-F1 score of 0.974, a Macro Recall of 0.98, and a BEGO recall value of 0.96. These

findings indicate that manual features alone possess limited discriminative power; however, they play a supportive and complementary role in enhancing the decision boundaries of deep learning models, particularly in scenarios involving class imbalance. The improvement in sensitivity achieved especially for the underrepresented BEGO class demonstrates that the proposed fusion approach provides a more reliable classification performance from a clinical perspective. When SVM (Linear and RBF) and XGBoost-based fusion models are examined, the Logistic Regression model achieved both the highest Macro-F1 score and the highest BEGO recall value. The experimental results obtained reveal that the feature-level fusion strategy offers higher classification performance and more balanced class performance compared to the use of deep learning or manual features alone.

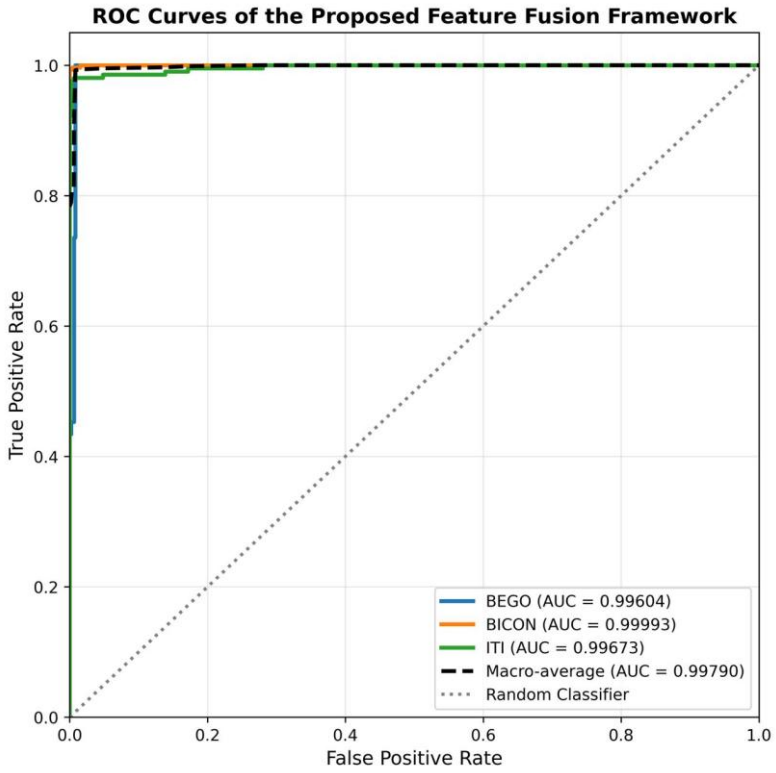

Figure 5. One-vs-rest ROC curves of the proposed feature fusion model.

Ground-truth and predicted label analysis

To provide a qualitative evaluation of the proposed feature-fusion framework, representative correctly classified and misclassified ROI images are presented in Fig. 6. Panels (A)–(C) illustrate correctly classified examples from the BEGO, BICON, and ITI implant classes, respectively, where the predicted labels are in full agreement with the ground-truth annotations. These cases demonstrate that the proposed model is capable of capturing class-discriminative radiographic features under favorable imaging conditions.Panels (D)–(F) present representative misclassification cases. In panel (D), a BEGO implant was incorrectly predicted as ITI; in panel (E), a BICON implant was misclassified as ITI; and in panel (F), an ITI implant was predicted as BEGO. A common pattern observed across these misclassified examples is the involvement of the ITI class, which is consistent with the confusion matrix findings presented in Fig. 4. These errors are attributed to the high degree of structural and morphological similarity between certain implant brands in radiographic appearance, particularly in cases where surface texture and thread geometry features overlap across classes. These findings suggest that while the proposed framework achieves strong overall classification performance, inter-brand visual similarity remains the primary source of residual classification error.

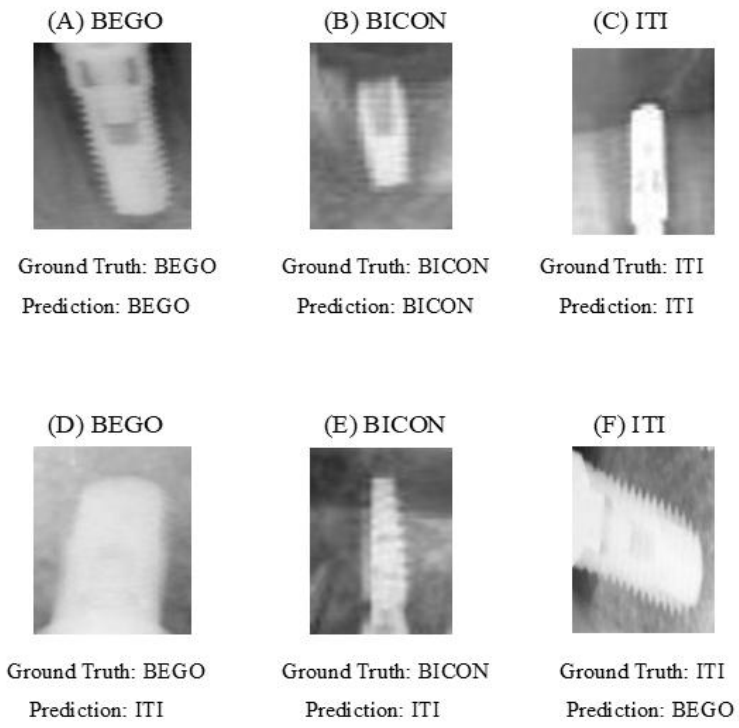

Figure 6. Representative ground-truth and predicted label comparisons for the proposed feature fusion model.

Explainability analysis using Grad-CAM

To assess the interpretability of the proposed feature-fusion framework, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to representative correctly classified and misclassified ROI images. The resulting activation maps are presented in Fig. 7, where warmer colors (red, orange) indicate regions of higher attention and cooler colors (blue) indicate lower activation intensity. For correctly classified examples (panels A–C), the activation maps reveal that the model consistently focuses on clinically relevant and structurally discriminative regions of the implant. In panel (A), high activation is concentrated along the implant body of the BEGO implant. In panel (B), the model attends to the central body region of the BICON implant, which exhibits a characteristic morphology. In panel (C), the activation is distributed across the thread pattern of the ITI implant, which is

consistent with the fine-pitch threading that distinguishes this brand radiographically. These observations indicate that the model has learned to rely on implant-specific structural features rather than background or uninformative regions. In contrast, the misclassified examples (panels D–F) exhibit markedly different activation patterns. In panel (D), where a BEGO implant was predicted as ITI, the activation map shows spatially dispersed attention across two separate regions, suggesting that the model failed to localize a single discriminative structure. In panel (E), the activation is broadly distributed across the upper implant region rather than the body, which may have contributed to the erroneous ITI prediction for a BICON implant. In panel (F), activation is concentrated in the upper corner of the image rather than on the implant body, indicating that the model attended to a less informative region when misclassifying an ITI implant as BEGO. Collectively, these findings suggest that misclassifications are associated with diffuse or mislocalized attention, further supporting the hypothesis that inter-brand structural similarity poses the primary challenge for the proposed classification framework.

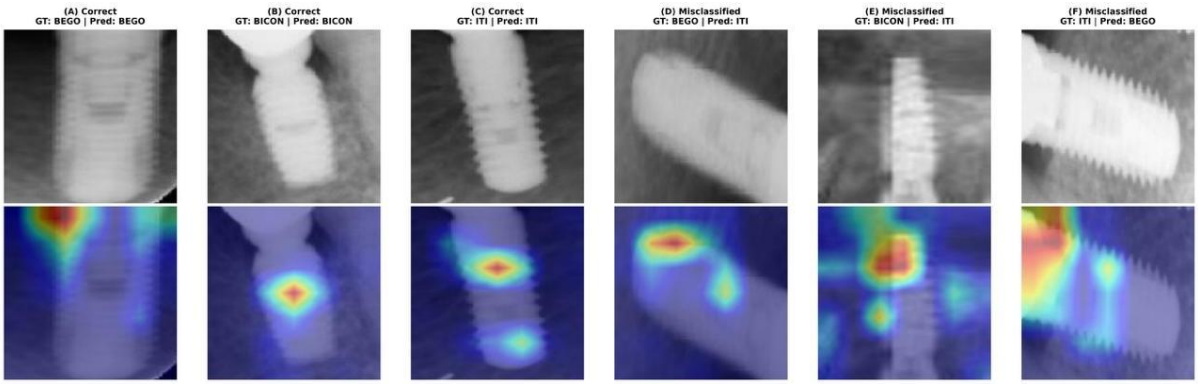

Figure 7. Representative Grad-CAM maps for correctly classified and misclassified implant-brand predictions.

Statistical comparison and robustness analysis

Statistical comparison of models using the McNemar test

In order to statistically compare the proposed feature-level fusion model with deep learning–based baseline approaches, a paired McNemar test was applied using the predictions obtained on the test set [45]. This analysis enables the comparison of error patterns derived from classification results obtained on the same test samples. The results of the McNemar test are presented in Table 7. As shown in Table 7, when the proposed model is compared with the Swin Transformer architecture, a higher number of discordant correct–incorrect classifications in favor of the proposed model (b = 16) is observed, and this difference is found to be statistically significant (p = 0.0118). This finding indicates that the proposed model exhibits a more consistent classification behavior in terms of error patterns compared to the Swin-based approach. In contrast, in the comparison with the ConvNeXt architecture, the error patterns of the two models are largely similar, and the observed difference is not statistically significant (p = 0.2891). This result suggests that both models exhibit comparable decision behaviors on the test set.

Table 7. McNemar test results for pairwise model comparisons.

Comparison	***b***	***c***	***p*-value***
Proposed vs Swin Transformer	16	4	0.0118
Proposed vs ConvNeXt	6	2	0.2891

b: number of samples correctly classified by the proposed model but misclassified by the compared model; *c*: number of samples misclassified by the proposed model but correctly classified by the compared model. *Statistical significance was assessed at $p<0.05$.

Evaluation of performance stability using bootstrap analysis

In order to evaluate the stability of model performance metrics and the impact of class imbalance on the results, a bootstrap resampling analysis with 1,000 iterations was performed on the test set. Within the scope of this analysis, 95% confidence intervals were calculated for the Macro-F1 score and class-based recall values. Bootstrap-based confidence intervals were preferred to assess the stability of performance measures under class imbalance, as single-point estimates such as accuracy may be misleading. The obtained results are presented in Table 8. The results of the bootstrap analysis demonstrate that the proposed feature-level fusion model exhibits high and stable performance across all implant brands. In particular, the narrow confidence intervals observed for the BICON and ITI classes indicate that the classification behavior of the model for these classes is highly stable. In addition, the high mean recall value obtained for the BEGO class and the relatively high lower bound of its confidence interval show that the proposed approach is able to produce reliable results even under class imbalance conditions. These findings reveal that the proposed model offers a strong structure not only in terms of overall classification performance but also in terms of class-based performance balance. The class-wise bootstrap confidence intervals presented in Fig. 8 visually demonstrate the performance stability of the proposed feature-level fusion approach. In particular, for the BEGO class, the fact that the proposed model exhibits a narrower confidence interval compared to other approaches indicates that more consistent classification decisions are produced even under limited and imbalanced sample conditions. This visual finding emphasizes the importance of decision stability in clinically sensitive classes.

Table 8. Bootstrap mean estimates and 95% confidence intervals for Macro-F1 and class-wise recall.

Model	Macro-F1 (95% CI)	BEGO Recall (95% CI)	BICON Recall (95% CI)	ITI Recall (95% CI)
ConvNeXt	0.962 (0.937–0.983)	0.923 (0.849–0.982)	0.996 (0.988–1.000)	0.966 (0.939–0.987)
Swin	0.918 (0.880–0.949)	0.714 (0.585–0.830)	1.000 (1.000–1.000)	0.976 (0.953–0.995)
Proposed feature fusion model	0.973 (0.951–0.990)	0.962 (0.903–1.000)	0.993 (0.981–1.000)	0.980 (0.959–0.996)

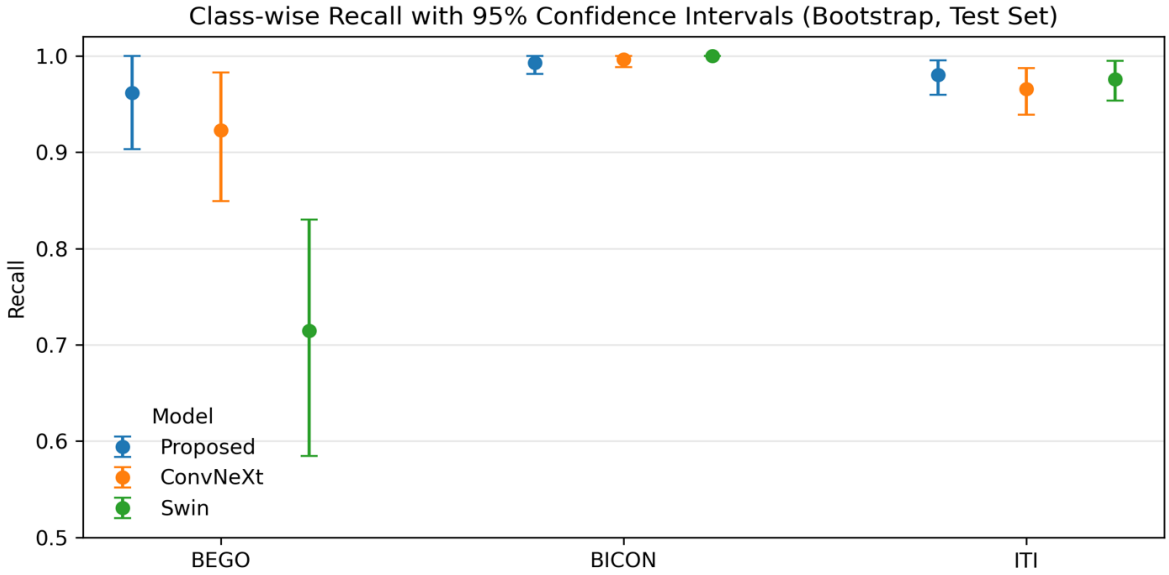

Figure 8. Bootstrap mean class-wise recall and 95% confidence intervals on the test set.

Ablation analysis of handcrafted feature contributions

To quantify the individual contribution of each handcrafted feature group to the overall classification performance of the proposed framework, a systematic ablation study was conducted. Each feature group was evaluated both in isolation using a Logistic Regression classifier and in combination with ConvNeXt-Tiny deep embeddings under the same group-aware 80/20 train–test split used throughout the study. The results are summarized in Table 9.

Table 9. Ablation analysis of handcrafted features combined with ConvNeXt-Tiny embeddings.

Configuration	Macro-F1	BEGO Recall	BICON Recall	ITI Recall
Stats only (9 features)	0.3492	0.0566	0.7634	0.2609
GLCM only (16 features)	0.3343	0.0566	0.8817	0.1594
LBP only (59 features)	0.5145	0.1698	0.8387	0.5072
HOG only (6,084 features)	0.6111	0.2075	0.8495	0.7536
All manual (6,168 features)	0.6317	0.2264	0.8674	0.7826
ConvNeXt embeddings only (768-d)	0.9724	0.9623	0.9928	0.9758
ConvNeXt + Stats	0.9724	0.9623	0.9928	0.9758
ConvNeXt + GLCM	0.9724	0.9623	0.9928	0.9758
ConvNeXt + LBP	0.9724	0.9623	0.9928	0.9758
ConvNeXt + HOG	0.9779	0.9811	0.9928	0.9807
ConvNeXt + All handcrafted (Proposed)	0.9738	0.9623	0.9928	0.9807

Discussion

The present study evaluated the performance of deep learning–based approaches for dental implant classification using panoramic radiographs and demonstrated that the proposed model achieved robust and clinically meaningful classification performance under a controlled experimental framework. Overall, the findings indicate that the developed approach is capable of accurately identifying implant systems from panoramic images, supporting its potential utility as a decision-support tool in prosthodontic practice. Based on these results, the null hypothesis—stating that deep learning–based analysis of panoramic radiographs would not enable reliable dental implant classification—was rejected. Previous studies have reported a wide range of classification performances for dental implant identification using deep learning, with accuracy values varying between approximately 70% and 98%, depending on the imaging modality, dataset size, number of implant systems, and network architecture employed [12,18,19,46–48]. Consistent with this literature, the results of

the present study fall within the upper range of reported performances, confirming that panoramic radiographs contain sufficient discriminative information for implant classification when analyzed using appropriately designed convolutional neural network architectures. However, direct comparison across studies remains challenging due to methodological heterogeneity, including differences in preprocessing strategies, train–test split protocols, and evaluation metrics. Prior investigations have emphasized that high classification performance is often associated with carefully curated datasets and optimized network architectures. Sukegawa et al. demonstrated that attention mechanisms and multitask learning can enhance classification accuracy [18,19]. Similarly, Park et al. reported high performance using a large multicenter dataset, highlighting the importance of data diversity for improving model generalizability [46]. In this context, the performance achieved in the present study is noteworthy, as it was obtained without relying on multicenter data, suggesting effective capture of discriminative implant features within a controlled dataset. The influence of data composition and augmentation strategies has also been emphasized in the literature. Conventional augmentation techniques may provide limited benefit when inter-class variability is low [49] leading recent studies to explore synthetic image generation. Sukegawa et al. reported improved performance using artificially generated X-ray images derived from three-dimensional implant models supporting the notion that synthetic data can enhance feature diversity when visual realism is preserved [29]. These findings align with the broader consensus that synthetic data can enhance feature diversity and mitigate overfitting, provided that artificial images adequately reflect the visual properties of clinical radiographs. Another important factor affecting classification performance is the imaging modality itself. Several studies have demonstrated that periapical radiographs often yield higher diagnostic accuracy than panoramic images due to their superior spatial resolution and reduced geometric distortion [50–52]. Nevertheless, panoramic radiography remains the most widely used

modality in routine clinical practice because of its comprehensive anatomical coverage and lower radiation burden. The results of the present study support the growing evidence that, despite inherent limitations, panoramic images can still serve as a reliable input for deep learning–based implant classification when advanced feature extraction and robust evaluation strategies are employed. Beyond technical performance, the clinical relevance of automated implant identification systems is substantial. Accurate implant identification is essential for appropriate prosthetic rehabilitation when historical records are unavailable. Previous studies have suggested that AI-based systems may reduce chairside time and support clinical decision-making [46,53]. The findings of the present study reinforce the feasibility of integrating AI-driven tools into prosthodontic workflows as adjunctive technologies. The robustness of the proposed approach was further supported by statistical analyses. The paired McNemar test demonstrated that the performance difference between the proposed feature-level fusion model and the Swin Transformer was statistically significant, indicating a consistently improved classification behavior. In contrast, no significant difference was observed between the proposed model and ConvNeXt, suggesting comparable decision patterns. Bootstrap-based confidence interval analysis further confirmed stable performance, particularly for the underrepresented BEGO class, supporting the statistical reliability of the proposed framework. Unlike many previous studies, this research adopted an ROI-based strategy and evaluated deep learning models, manual feature–based approaches, and their combination. Manual features alone were insufficient to capture subtle morphological differences, particularly for the BEGO class, whereas deep learning models achieved consistent improvements across metrics. Notably, ConvNeXt-Tiny outperformed ResNet-50 and Swin-Tiny in terms of Macro-F1 score, indicating the effectiveness of modern convolutional design principles [54]. Furthermore, the study intentionally employed feature-level fusion rather than decision-level fusion [55]. In decision-level fusion strategies, models

761 trained on deep learning and manual features operate independently, and final predictions are
762 combined using majority voting or probability averaging [43,56]. However, because the
763 implant classes in this study exhibited high visual similarity, the complementary nature of
764 deep representations and manual features favored feature-level fusion. The empirical results
765 strongly support this theoretical rationale. The analysis of manual feature–based machine
766 learning methods revealed that while such features provide a certain degree of discriminative
767 information, they remain inadequate for representing complex, high-level visual patterns.
768 Nevertheless, the relatively balanced precision–recall behavior observed in some classes
769 suggests that manual features can serve a complementary role when combined with deep
770 representations [57]. At this stage, the feature fusion strategy, which constitutes the main
771 contribution of the study, significantly improved the classification performance by integrating
772 the complementary strengths of deep learning representations and handcrafted features,
773 achieving a value of 0.973 (0.951–0.990). The hybrid models obtained by embedding-level
774 fusion of deep feature representations extracted from the ConvNeXt-Tiny model with
775 handcrafted feature vectors outperformed both standalone deep learning–based approaches
776 and methods relying solely on handcrafted features used in this study. Moreover, since the
777 deep learning model was already employed as a powerful feature extractor, the use of
778 complex and high-capacity classifiers was not required Logistic Regression (LR) was
779 preferred due to its low computational cost, stable structure, and strong generalization
780 capability, particularly for small- to medium-sized datasets [44]. Notably, the LR-based
781 fusion model achieved the highest macro F1-score on the test set, demonstrating that linear
782 classifiers are capable of learning effective decision boundaries in high-dimensional and
783 enriched feature spaces [58]. Confusion matrix analysis showed that this approach mitigated
784 class imbalance effects and achieved balanced recall performance. These findings indicate
785 that feature fusion–based methods yield higher accuracy and more balanced class-wise

performance, contributing to reliable automated implant identification systems. This study also provides a directly comparable evaluation with previous works using the same publicly available dataset as well as different datasets (Table 10).

Table 10. Methodological comparison of dental implant classification studies.

Ref.	Data source	Model / approach	Data splitting	Data augmentation	Evaluation protocol	Strengths / limitations
[59]	Single-center periapical radiographs (Bicon, Bego, ITI)	CNN, CNN+SVM, CNN+RF (CRF), CNN+DT, VGG16, ResNet50, Xception	5-fold cross-validation	Extensive augmentation (×10 per image)	Accuracy, Precision, Recall, F1-score, ROC AUC	Strong preprocessing and augmentation pipeline; potential overestimation due to augmented samples across folds
[19]	Single-center panoramic radiographs	ResNet (18–152), multi-task learning (brand + treatment stage)	4-fold cross-validation (80/20)	Limited, conventional	Accuracy, Precision, Recall, F1-score	Multi-task learning improves feature sharing; no group-aware splitting
[15]	Multi-center, large-scale dataset	Fine-tuned ResNet-50	Random 80/20 split	Extensive augmentation (rotation, noise, contrast)	Accuracy, Recall, F1-score, AUC	Strong clinical diversity; aggressive augmentation may inflate performance
[29]	Real + synthetic panoramic X-ray images	ResNet-50	Random split	Synthetic data generation + augmentation	Accuracy	Enhanced data diversity; domain gap between synthetic and real images
[60]	Single-center, three implant brands	Instance segmentation-based deep learning	70/20/10 split	None (resizing only)	Accuracy, Recall, F1-score	Clinically simple setup; limited number of brands and samples
[30]	Real + synthetic panoramic X-ray images	CNN-based deep learning (ResNet variants)	Random split	Synthetic data generation + augmentation	Accuracy	Increased data diversity; potential domain gap between synthetic and real images
[61]	Single-center panoramic radiographs (2634 PR images, 11,904 ROIs)	VGG-16/19, ResNet-50/101, GoogLeNet + majority-vote fusion	5-fold cross-validation	Implicit via transfer learning	Accuracy, Precision, Recall, F1-score, Sensitivity, Specificity	High performance with large ROI set; ROI-level CV may cause sample dependency
This study	**Open-source ROI-based dataset**	**ConvNeXt-Tiny, Swin-Tiny + feature-level fusion**	**Group-aware 80/20 split**	**Training-only augmentation**	**Macro-F1, Recall**	**Leakage-free evaluation; metrics robust to class imbalance**

It is observed that the high performance values reported in the literature have mostly been obtained through random data splitting strategies, intensive data augmentation techniques, or the use of synthetic data. In contrast, in the present study, a group-based data splitting strategy was adopted, and data augmentation was limited solely to the training phase, while a deterministic preprocessing pipeline was maintained during the testing process. Despite this more restrictive evaluation protocol, the high Macro-F1 and recall values achieved with the ConvNeXt-Tiny and feature fusion–based models indicate that the observed performance gains are primarily attributable to the quality of the learned representations rather than to aggressive data augmentation. In this respect, the present study provides a more cautious yet clinically more reliable and generalizable performance evaluation compared to the existing literature.

Limitations

Several limitations of this study should be acknowledged. First, the present framework was evaluated using carefully extracted implant ROI images obtained under controlled conditions. In real-world clinical environments, implant identification may become considerably more challenging due to partial implant visibility, overlapping restorations, peri-implant bone loss, metallic artifacts, low-quality radiographs, and variations in image acquisition parameters. Furthermore, uncommon implant systems that are not represented in the current dataset may exhibit morphological characteristics that differ substantially from the studied brands. Therefore, although the proposed framework demonstrated strong performance on the available dataset, its robustness under these challenging clinical conditions requires further investigation through larger and more heterogeneous datasets. From a clinical perspective, the proposed framework may serve as a decision-support tool to assist clinicians in implant identification rather than replacing clinical judgment. Moreover, because the present study included a limited number of implant systems, caution should be exercised when

extrapolating the findings to implant brands not represented in the training dataset. Future studies should incorporate a substantially larger number of implant classes and multicenter datasets to further improve the generalizability and clinical applicability of the proposed framework. Moreover, the current study focused on pre-cropped implant ROI classification rather than end-to-end analysis of complete radiographic images. Future studies should investigate clinically realistic workflows that combine automatic implant localization, ROI extraction, and implant classification within a unified framework. Such evaluations would provide a more comprehensive assessment of the practical applicability of AI-assisted implant identification systems in routine prosthodontic practice. The dataset was derived from a single open-source with a limited number of implant brands. Future studies should focus on multicenter data, broader implant diversity, and advanced architectures integrating segmentation and classification. The incorporation of physics-based simulations and generative models may further enhance dataset diversity and model robustness, facilitating translation into routine clinical practice. This study has several limitations. Most importantly, the key limitation of the present study is the restricted diversity of the dataset. Although the dataset contains 2,688 ROI images, it includes only three implant systems (BEGO, BICON, and ITI) obtained from a single publicly available source. Consequently, the proposed framework was evaluated within a relatively controlled experimental setting and may not fully represent the variability encountered in routine clinical practice, where clinicians must identify a substantially larger number of implant brands with overlapping radiographic characteristics. Therefore, the current findings should be interpreted as a proof-of-concept rather than evidence of immediate clinical deployment. Future studies should focus on expanding the dataset through multicenter collaborations, incorporating additional implant systems with diverse geometrical and radiographic characteristics, and validating the framework on independent external datasets. Such efforts would provide a more

comprehensive assessment of model robustness, generalizability, and real-world clinical applicability.

Conclusions

In this study, a robust and reproducible ROI-based framework for the automatic classification of dental implant brands was presented. A dataset originally provided in object detection format was transformed into a classification problem, and a group-based data splitting strategy was employed to prevent potential data leakage between images derived from the same implant instance. Comparative experiments demonstrated that the ConvNeXt-Tiny architecture provided strong classification performance, while transformer-based models yielded more stable results in scenarios with limited training data. In addition, a feature-level fusion approach was proposed, in which deep learning–based embeddings were combined with manually extracted texture- and intensity-based features. These hybrid models produced higher and more balanced Macro-F1 scores compared to approaches relying solely on deep learning or manual features. The proposed Logistic Regression–based fusion model achieved the highest Macro-F1 (0.974) and BEGO recall (0.96) values, highlighting the effectiveness of hybrid representation learning and the advantages of simple and interpretable classifiers for reliable implant brand identification. The findings underscore the importance of complementary representation learning and rigorous evaluation protocols in the development of dependable implant classification systems. Overall, the proposed framework, with its transparent, reproducible, and generalizable structure, has the potential to contribute to the development of practical automated implant identification solutions for clinical decision-support systems.

List of abbreviations

AI: Artificial intelligence; AMP: Automatic mixed precision; AUC: Area under the curve; CNN: Convolutional neural network; GLCM: Gray-level co-occurrence matrix; Grad-CAM: Gradient-weighted class activation mapping; HOG: Histogram of oriented gradients; LBP: Local binary pattern; LR: Logistic regression; ROC: Receiver operating characteristic; ROI: Region of interest; SVM: Support vector machine.

Declarations

Ethics approval and consent to participate

This study involved a secondary analysis of publicly available, anonymised dental radiographs obtained from Roboflow Universe. No new patient radiographs or clinical data were collected, and no patients were recruited or contacted by the authors. The additional verification used commercially procured, unused dental implants in sterile packaging and did not involve human participants. Institutional ethics committee approval, a formal exemption, and a waiver of informed consent were not obtained for this study. The authors did not obtain additional informed consent for the secondary use of the publicly available images.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used in this study were obtained from publicly available sources hosted on Roboflow Universe: https://universe.roboflow.com/lamiaa/implant-recognition [62] and https://universe.roboflow.com/dental-implant-nyw2h/-dental-implant [36]. The ROI datasets and metadata files generated during preprocessing, including the group identifiers used for

leakage-aware data partitioning, are available from the corresponding author upon reasonable request. The source code supporting the findings of this study, including Jupyter notebooks used for ROI preprocessing, feature extraction, model training, statistical analyses, and figure generation, is publicly available at Zenodo (https://doi.org/10.5281/zenodo.21113740) [63] and through the associated GitHub repository (https://github.com/fnurortatas/implant-brand-identification-feature-fusion_) [64].

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

Aliye İpek Kuşçu (AIK), Gökalp Çınarer (GC), Fatma Nur Ortataş (FNO), Ayça Dolar (AD), and Süha Kuşçu (SK) contributed to the conceptualization and design of the study. AIK, GC, and AD developed the methodology and performed the data analysis. AIK, GC, and AD drafted the manuscript. FNO and SK provided critical revisions and supervised the study. All authors reviewed, revised, and approved the final version of the manuscript.

Acknowledgements

Not applicable.

References

1. Schimmel M, Araujo M, Abou-Ayash S, Buser R, Ebenezer S, Fonseca M, et al. Group 4 ITI Consensus Report: Patient benefits following implant treatment in

907 partially and fully edentulous patients. Clinical oral implants research. 2023;34:257-
908 65.

909 2. Howe M-S, Keys W, Richards D. Long-term (10-year) dental implant survival: A
910 systematic review and sensitivity meta-analysis. Journal of dentistry. 2019;84:9-21.

911 3. Jokstad A, Braegger U, Brunski JB, Carr AB, Naert I, Wennerberg A. Quality of
912 dental implants. International Dental Journal. 2003;53(6):409-43.

913 4. Saghiri MA, Freag P, Fakhrzadeh A, Saghiri AM, Eid J. Current technology for
914 identifying dental implants: a narrative review. Bulletin of the National Research
915 Centre. 2021;45(1):7.

916 5. Al-Wahadni A, Barakat MS, Abu Afifeh K, Khader Y. Dentists' most common
917 practices when selecting an implant system. Journal of Prosthodontics.
918 2018;27(3):250-9.

919 6. Dreyer H, Grischke J, Tiede C, Eberhard J, Schweitzer A, Toikkanen SE, et al.
920 Epidemiology and risk factors of peri-implantitis: a systematic review. Journal of
921 periodontal research. 2018;53(5):657-81.

922 7. Verma A, Singh SV, Arya D, Shivakumar S, Chand P. Mechanical failures of dental
923 implants and supported prostheses: A systematic review. Journal of Oral Biology and
924 Craniofacial Research. 2023;13(2):306-14.

925 8. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.

926 9. Khanagar SB, Al-Ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA,
927 et al. Developments, application, and performance of artificial intelligence in
928 dentistry–A systematic review. Journal of dental sciences. 2021;16(1):508-22.

10. Yüksel AE, Gültekin S, Simsar E, Özdemir ŞD, Gündoğar M, Tokgöz SB, et al. Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. Scientific reports. 2021;11(1):12342.

11. Lee J-H, Jeong S-N. Efficacy of deep convolutional neural network algorithm for the identification and classification of dental implant systems, using panoramic and periapical radiographs: A pilot study. Medicine. 2020;99(26):e20787.

12. Takahashi T, Nozaki K, Gonda T, Mameno T, Wada M, Ikebe K. Identification of dental implants using deep learning—pilot study. International journal of implant dentistry. 2020;6(1):53.

13. Sehar U, Xiong J, Xia Z. Automatic Tooth Labeling After Segmentation Using Prototype-based Meta-learning. Machine Intelligence Research. 2025;22(3):539-52.

14. Park W, Schwendicke F, Krois J, Huh JK, Lee JH. Identification of dental implant systems using a large-scale multicenter data set. Journal of Dental Research. 2023;102(7):727-33.

15. Lee J-H, Kim Y-T, Lee J-B. Identification of dental implant systems from low-quality and distorted dental radiographs using AI trained on a large multi-center dataset. Scientific Reports. 2024;14(1):12606.

16. Kong HJ. Classification of dental implant systems using cloud-based deep learning algorithm: An experimental study. Journal of Yeungnam Medical Science. 2023;40(Suppl):S29-S36.

17. Ahmed WM, Azhari AA, Almufti A, Alsadah ZM, Ahmed AFA, Lahiq A, et al. Development and evaluation of an AI model for dental implant type detection: A

951 comparison of diagnostic accuracy between a deep learning model and dental
952 professionals. Journal of Prosthodontics. 2026;35(1):81-93.

953 18. Sukegawa S, Yoshii K, Hara T, Yamashita K, Nakano K, Yamamoto N, et al. Deep
954 neural networks for dental implant system classification. Biomolecules.
955 2020;10(7):984.

956 19. Sukegawa S, Yoshii K, Hara T, Matsuyama T, Yamashita K, Nakano K, et al. Multi-
957 task deep learning model for classification of dental implant brand and treatment stage
958 using dental panoramic radiograph images. Biomolecules. 2021;11(6):815.

959 20. Tekis H, Zirek T, Tassoker M. Comparative analysis of YOLO variants for dental
960 implant brand identification. Scientific Reports. 2025;15(1):33304.

961 21. Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements.
962 arXiv preprint arXiv:241017725. 2024.

963 22. Ariji Y, Kusano K, Fukuda M, Wakata Y, Nozawa M, Kotaki S, et al. Two-step deep
964 learning models for detection and identification of the manufacturers and types of
965 dental implants on panoramic radiographs. Odontology. 2025;113(2):788-98.

966 23. Lee J-H, Kim Y-T, Lee J-B, Jeong S-N. Deep learning improves implant
967 classification by dental professionals: a multi-center evaluation of accuracy and
968 efficiency. Journal of Periodontal & Implant Science. 2021;52(3):220.

969 24. Bouali R, Mahboub O, Lazaar M. Review of dental diagnosis by deep learning
970 models: trends, applications and challenges. Procedia Computer Science.
971 2024;231:221-8.

25. Ibraheem WI. Accuracy of artificial intelligence models in dental implant fixture identification and classification from radiographs: a systematic review. Diagnostics. 2024;14(8):806.

26. Acerra A, Aliberti A, Amato A, Eccellente A, Santurro A, Giordano F. Artificial Intelligence in Implant Dentistry: Clinical Validity, Diagnostic Performance, Surgical Planning, and Medico-Legal Implications—A Narrative Review. Dentistry Journal. 2026;14(7):389.

27. Jiang X, Yao Y, Tang W, Han D, Zhang L, Zhao K, et al. Design of dental implants at materials level: An overview. Journal of Biomedical Materials Research Part A. 2020;108(8):1634-61.

28. Simonis P, Dufour T, Tenenbaum H. Long-term implant survival and success: a 10–16-year follow-up of non-submerged dental implants. Clinical oral implants research. 2010;21(7):772-7.

29. Sukegawa S, Yoshii K, Hara T, Tanaka F, Taki Y, Inoue Y, et al. Optimizing dental implant identification using deep learning leveraging artificial data. Scientific Reports. 2025;15(1):3724.

30. Kohlakala A, Coetzer J, Bertels J, Vandermeulen D. Deep learning-based dental implant recognition using synthetic X-ray images. Medical & Biological Engineering & Computing. 2022;60(10):2951-68.

31. Sehar U, Xiong J, Zhai J, Xia Z. Automatic orthodontic treatment planning using deep learning. Machine Intelligence Research. 2026:1-17.

32. Malik H, Anees T, Chaudhry MU, Gono R, Jasiński M, Leonowicz Z, et al. A novel fusion model of hand-crafted features with deep convolutional neural networks for

995 classification of several chest diseases using X-ray images. IEEE Access.
996 2023;11:39243-68.

997 33. Shankar K, Perumal E. A novel hand-crafted with deep learning features based fusion
998 model for COVID-19 diagnosis and classification using chest X-ray images. Complex
999 & Intelligent Systems. 2021;7(3):1277-93.

1000 34. Cruz-Ramos C, Garcia-Avila O, Almaraz-Damian J-A, Ponomaryov V, Reyes-Reyes
1001 R, Sadovnychiy S. Benign and malignant breast tumor classification in ultrasound and
1002 mammography images via fusion of deep learning and handcraft features. Entropy.
1003 2023;25(7):991.

1004 35. Almaraz-Damian J-A, Ponomaryov V, Sadovnychiy S, Castillejos-Fernandez H.
1005 Melanoma and nevus skin lesion classification using handcraft and deep learning
1006 feature fusion via mutual information measures. Entropy. 2020;22(4):484.

1007 36. Implant D. Dental Implant Dataset. Roboflow Universe. 2024.
1008 https://universe.roboflow.com/dental-implant-nyw2h/-dental-implant. Accessed Aug
1009 2026.

1010 37. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image
1011 recognition2016.

1012 38. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S, editors. A convnet for the
1013 2020s2022.

1014 39. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., editors. Swin transformer:
1015 Hierarchical vision transformer using shifted windows2021.

40. Pembury Smith MQR, Ruxton GD. Effective use of the McNemar test. Behavioral Ecology and Sociobiology. 2020;74(11):133.

41. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions: john wiley & sons; 2013.

42. Tibshirani RJ, Efron B. An introduction to the bootstrap. Monographs on statistics and applied probability. 1993;57(1):1-436.

43. Kuncheva LI. Combining pattern classifiers: methods and algorithms: John Wiley & Sons; 2014.

44. Hastie T. The elements of statistical learning: data mining, inference, and prediction. Springer; 2009.

45. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-7.

46. Park W, Huh J-K, Lee J-H. Automated deep learning for classification of dental implant radiographs using a large multi-center dataset. Scientific reports. 2023;13(1):4862.

47. Lee J-H, Kim Y-T, Lee J-B, Jeong S-N. A performance comparison between automated deep learning and dental professionals in classification of dental implant systems from dental imaging: a multi-center study. Diagnostics. 2020;10(11):910.

48. Chaurasia A, Namachivayam A, Koca-Ünsal RB, Lee J-H. Deep-learning performance in identifying and classifying dental implant systems from dental imaging: a systematic review and meta-analysis. Journal of Periodontal & Implant Science. 2023;54(1):3.

49. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of big data. 2019;6(1):1-48.

50. Khurshid Z, Faridoon F, Silkosessak OC-UD, Trachoo V, Waqas M, Hasan S, et al. Multi-Regional deep learning models for identifying dental restorations and prosthesis in panoramic radiographs. BMC Oral Health. 2025;25(1):1-17.

51. Sadr S, Mohammad-Rahimi H, Motamedian SR, Zahedrozegar S, Motie P, Vinayahalingam S, et al. Deep learning for detection of periapical radiolucent lesions: a systematic review and meta-analysis of diagnostic test accuracy. Journal of Endodontics. 2023;49(3):248-61.

52. Liu J, Jin C, Wang X, Pan K, Li Z, Yi X, et al. A comparative analysis of deep learning models for assisting in the diagnosis of periapical lesions in periapical radiographs. BMC Oral Health. 2025;25(1):801.

53. Wu Z, Yu X, Wang F, Xu C. Application of artificial intelligence in dental implant prognosis: A scoping review. Journal of Dentistry. 2024;144:104924.

54. Huang C, Wang J, Wang S, Zhang Y. A review of deep learning in dentistry. Neurocomputing. 2023;554:126629.

55. Planet S, Iriondo I. Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. In: 7th Iberian Conference on Information Systems and Technologies (CISTI 2012); 20-23 June 2012; Madrid, Spain. IEEE; 2012.

56. Mi A, Wang L, Qi J. A multiple classifier fusion algorithm using weighted decision templates. Scientific Programming. 2016;2016(1):3943859.

57. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annual review of biomedical engineering. 2017;19(1):221-48.

58. Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceedings of the 21st International Conference on Machine Learning (ICML 2004); Banff, Canada. ACM; 2004. p. 78. doi:10.1145/1015330.1015435.

59. Lashaki RA, Raeisi Z, Razavi N, Goodarzi M, Najafzadeh H. Optimized classification of dental implants using convolutional neural networks and pre-trained models with preprocessed data. BMC Oral Health. 2025;25(1):535.

60. Taşdemir İ, İçen V, Kölüş T. Using Deep Learning Algorithms to Predict Dental Implant Brands from Panoramic Radiographs. Medical Records. 2025;7(1):8-11. doi:10.37990/medr.1524857.

61. Tiryaki B, Ozdogan A, Guller MT, Miloglu O, Oral EA, Ozbek IY. Dental implant brand and angle identification using deep neural networks. The Journal of Prosthetic Dentistry. 2025;133(6):1528-34.

62. lamiaa. IMPLANT RECOGNITION Dataset. Roboflow Universe. Roboflow; 2026. https://universe.roboflow.com/lamiaa/implant-recognition. Accessed Aug 2026.

63. fnurortatas. fnurortatas/implant-brand-identification-feature-fusion_: Initial release for manuscript revision. Zenodo. 2026. https://doi.org/10.5281/zenodo.21113740. Accessed Aug 2026.

64. fnurortatas. implant-brand-identification-feature-fusion_. GitHub. https://github.com/fnurortatas/implant-brand-identification-feature-fusion_. Accessed Aug 2026.