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.

Open-Set Recognition for Dental Implant System Identification: A Multicenter Retrospective Study

Sangyeon Lee
The Catholic University of Korea Seoul St. Mary's Hospital
Donghyun Kim
Dongguk University
Changgyun Kim
tiockdrbs@kangwon.ac.kr

Kangwon National University

Research Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11286657/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Deep learning models for dental implant identification are typically trained on a fixed set of systems, although new systems continuously enter the market. Such closed-set models assign unregistered implants to known systems, often with high confidence. We compared open-set recognition methods for detecting unregistered implant systems while identifying registered ones.

Methods

In this retrospective study, 100,105 implant images from panoramic radiographs across 15 institutions were obtained from a public dataset. Thirteen systems with > 2,000 images were defined as known classes, while 29 systems (14,004 images) were withheld as unknown classes. A ConvNeXt classifier was trained using ten implant-level splits. Score-based, distance-based (OpenMax), and training-based (ARPL) methods were compared at a threshold accepting 95% of known validation images.

Results

The classifier achieved 0.901 accuracy for system identification and 0.968 accuracy at the manufacturer level. OpenMax achieved an AUROC of 0.898 and detected 61.4% of unknown images with 0.868 precision, outperforming all other methods across ten splits (AUROC, 0.825–0.844; detection, 29.1–49.4%; adjusted P = 0.012). Its advantage was retained after excluding institutions without known-class images. Unknown systems from manufacturers represented in training were harder to detect than those from new manufacturers (57.5% vs. 64.4%), and undetected implants were often assigned to a small number of known systems, frequently from the same manufacturer.

Conclusions

Closed-set evaluation overlooks a major risk of automated implant identification. A post hoc distance-based method detected most unregistered implants without retraining and with minimal loss of closed-set accuracy, providing a practical safeguard for existing classifiers and supporting expert review of flagged cases.

1. Background

Dental implants have become a standard treatment for replacing missing teeth, with millions of procedures performed worldwide each year [1–3]. Unlike the prosthetic crown and abutment, which require periodic maintenance or replacement, the implant fixture remains in the bone for decades, and patients therefore need lifelong management of their implants [4, 5]. When treatment records are

unavailable because of clinic closure, patient relocation, or treatment abroad, clinicians must identify the implant system from radiographs alone [6, 7]. This task is difficult because implant components are not standardized across manufacturers, and misidentification can lead to mechanical instability, prosthetic failure, and biological complications such as peri-implantitis [1, 8–11]. When the system cannot be identified at all, a functioning implant may be removed and replaced, exposing patients to additional surgical risk and cost. This burden falls especially on older patients, whose insurance coverage for implant treatment is often limited to a fixed number of procedures [12–15]. Implant identification has traditionally relied on visual comparison of radiographs with manufacturer catalogs, a process that is time-consuming, subjective, and dependent on the clinician's experience. To address this, several studies have applied deep learning to classify implant systems from panoramic and periapical radiographs, reporting high accuracy [16–22]. In a multicenter study, deep learning models performed comparably to or better than dental professionals [19]. However, these studies share a common design: models are trained on a fixed set of implant systems, ranging from three to about a dozen, and evaluated on test images drawn from the same systems. Under this closed-set design, every input image is forced into one of the known classes, and the reported accuracy reflects performance only on implants that the model has already seen. This assumption does not hold in clinical practice. Hundreds of implant systems are commercially available, and manufacturers continually introduce new systems and design modifications while discontinuing older ones [23]. Therefore, no training dataset can remain complete, and a deployed model will inevitably encounter implants from systems absent from its training data. A closed-set model has no mechanism to recognize such cases; it assigns the unregistered implant to the most visually similar known system, often with high confidence. This silent misclassification is particularly hazardous because it gives the clinician a confident but incorrect answer, which may lead to the selection of incompatible prosthetic components, the very error that automated identification is intended to prevent. For an identification tool to be safe in a continuously evolving market, the ability to recognize when an implant falls outside its knowledge is a requirement rather than an optional feature.

Open-set recognition addresses this problem by requiring a model both to classify samples from known classes and to reject samples from classes not seen during training [24]. Related approaches for detecting out-of-distribution inputs have been studied extensively [25], and their importance has been increasingly recognized in medical imaging, including COVID-19 lung lesion segmentation [26], digital pathology [27], dermatology [28], and capsule endoscopy [29]. Existing methods fall broadly into three categories. Score-based methods derive a confidence score from the output of a conventionally trained classifier, such as the maximum softmax probability or the maximum logit [30]. Distance-based methods, such as OpenMax, assess how far the activation pattern of an input lies from the patterns typical of each known class [31]. Training-based methods, such as adversarial reciprocal points learning, modify the training objective to constrain the space occupied by known classes [32]. Score-based and distance-based methods are applied post hoc and therefore require no retraining of an existing classifier. Benchmark studies on natural images have shown that the relative performance of these approaches depends on the task, and that a well-trained classifier with a simple score can be competitive with dedicated training-based methods. However, it remains unclear which approach is most effective when

unknown classes closely resemble known ones, as is the case for successive implant systems of the same manufacturer, and open-set recognition has not yet been applied to dental implant identification.

In this study, we formulated dental implant identification as an open-set recognition problem and systematically compared representative methods. An overview of the study is shown in Fig. 1. Using a public multicenter dataset of 100,105 implant images from panoramic radiographs, we trained classifiers on 13 implant systems and withheld 29 additional systems as unregistered classes, simulating the arrival of implants not represented in the training data. The withheld systems included both products from manufacturers represented in the training set and products from manufacturers that were entirely absent, reflecting the two ways in which unfamiliar implants appear in practice. The contributions of this study are threefold. First, we established an evaluation protocol for implant identification under open-set conditions, with data splits that keep all images of the same implant together and a rejection threshold determined from known classes only. Second, we compared score-based, distance-based, and training-based open-set recognition methods under identical conditions and found that a post hoc distance-based method consistently outperformed the others. Third, we analyzed the structure of recognition failures, showing how detection depends on the manufacturer and the institution of origin and to which known systems undetected implants are assigned.

2. Methods

2.1 Data acquisition

This retrospective study used the publicly available ‘Implanted Dental Implant Image Data’ dataset distributed by AI-Hub, a national data platform operated by the National Information Society Agency of Korea [33]. The dataset consists of cropped images of individual implant fixtures extracted from panoramic radiographs collected at 15 dental institutions in Korea. Each image is a single-channel grayscale crop of 256 x 128 pixels (width x height). For each image, the dataset provides the institution, imaging modality, manufacturer, fixture model, fixture diameter and length, tooth position, implant placement date and radiograph acquisition date. The radiographs were acquired between September 2000 and December 2021. Because only de-identified, cropped image data from a public dataset were used, this study was granted an exemption by the Public Institutional Review Board designated by the Ministry of Health and Welfare of Korea (Exemption No. P01-202511-01-038), and the requirement for informed consent was waived.

2.2 Data Curation

Of the 100,122 images in the dataset, one image with an invalid acquisition date and 16 duplicate images were removed during quality control. The final dataset comprised 100,105 images representing 42 implant systems from 16 manufacturers. Because patients were frequently imaged at multiple follow-up visits, many implants appeared in more than one image. Therefore, we defined a physical implant as a unique combination of institution, placement date, tooth position, fixture model, diameter, and length. This yielded 23,931 implants, with a mean of 4.2 images per implant. Implant systems with more than

2,000 images were designated as known classes, and all remaining systems as unknown classes. This resulted in 13 known classes (86,101 images; 20,660 implants) and 29 unknown classes (14,004 images; 3,271 implants) (Table 1, Additional file 1). To reflect the two ways in which unfamiliar implants arise in practice, unknown classes were further divided by manufacturer. The first group comprised systems from manufacturers represented among the known classes (known-manufacturer unknowns; 12 classes, 6,200 images). The second comprised systems from manufacturers absent from the known classes (new-manufacturer unknowns: 17 classes, 7,804 images). In addition, two institutions contributed no images of known classes. Unknown images were therefore labeled by source as originating from shared sites, which also contributed known images (10,572 images), or from unknown-only sites (3,477 images).

Table 1
Composition of the dataset

Class	Manufacturer	Model	Images	Implants	Institutions
Known classes					
ts3	Osstem	TS III	15376	4327	8
su	Dentium	Superline	13386	3447	8
im	Dentium	Implantium	11993	2325	7
ast	Dentsply	Astra	10721	1855	6
is3	Neobiotech	IS III	6074	1749	4
us2	Osstem	US II	5835	1712	4
is	Neobiotech	IS I	5366	953	8
gs3	Osstem	GS III	3935	633	4
ex	Warantec	Explant	3648	1255	3
bra	Nobel Biocare	Brånemark	2640	494	4
in	Warantec	Internal	2562	790	2
lu	Shinhung	Luna	2347	661	3
is2	Neobiotech	IS II	2218	459	4
Subtotal (13 classes)			86101	20660	13
Unknown classes					
Known manufacturer (12 classes)			6200	1263	10
New manufacturer (17 classes)			7804	2008	12
Subtotal (29 classes)			14004	3271	14
Total (42 classes)			100105	23931	15

2.3 Data Splitting

The dataset contained no patient identifiers. To prevent images of the same implant or implants placed in the same patient on the same day from being distributed across different subsets, images were grouped by institution and placement date, resulting in 8,715 groups. Known-class images were partitioned into ten folds using stratified group k-fold splitting, preserving class proportions while keeping each group within a single fold. The data were divided into training (60%), validation (10%), and

test (30%) sets. This procedure was repeated using ten random seeds to generate ten independent splits. All unknown-class images were withheld from training, validation, and threshold determination and were used exclusively as the unknown test set in each split.

2.4 Classification Model and Training

We used ConvNeXt-small pretrained on ImageNet-22K as the backbone network, implemented with the timm library [34]. Images were used as their native resolution without resizing to avoid distortions of fixtures. The grayscale channel was replicated to three channels and normalized with ImageNet statistics. Training-time augmentation consisted of random horizontal flipping, random rotation within ± 15°, and random brightness and contrast adjustment within ± 20%. The backbone was followed by a linear classification layer and trained with cross-entropy loss. We used AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.05 for the backbone. The linear classifier was trained with a learning rate of 1×10^{-3} without weight decay. A cosine annealing learning-rate schedule was used for 10 epochs with a batch size of 32. The checkpoint with the highest known-class validation accuracy was selected. All experiments were run on a single NVIDIA Tesla T4 GPU. One model was trained per split, and all post-hoc open-set methods described below were applied to this same model.

2.5 Open-Set Recognition Methods

Each method assigns a score s(x) to input x, with higher scores indicating a higher likelihood of belonging to a known class. Let ($z \in \mathbb{R}^C$) denote the logit vector for C known classes. OpenMax recalibrates classifier logits based on their distance from class-specific activation patterns [31]. For each class C, the mean activation vector was calculated from correctly classified training samples. A Weibull distribution was fitted to the 20 largest distances for each class, and the logits of the 10 highest-scoring classes were recalibrated accordingly. An unknown activation was then derived from the recalibrated logits, and its softmax probability P_0 was used to define the open-set score as $s(x) = 1\text{-}P_0$. The tail size and number of recalibrated classes were fixed at 20 and 10 without tuning on unknown data. Four post-hoc scores were evaluated using the same classifier: maximum softmax probability (MSP) [30], maximum logit score (MLS) [35], energy score [36], and ODIN [37]. We also evaluated adversarial reciprocal points learning (ARPL) [38], which uses learnable reciprocal points for each known class during training. The original loss formulation was used with lambda = 0.1, and the maximum class logit was used as the open-set score. ARPL models were trained separately for each split using the same backbone, data, augmentation, and optimization settings as the baseline classifier.

2.6 Evaluation

For every method and split, a rejection threshold τ was set as the score at which 95% of known validation images were accepted. Images with $s(x) < \tau$ were classified as unknown, and all others were assigned their predicted known class. Closed-set performance on the known test set was evaluated by accuracy and per-class and macro-averaged precision, recall, and F1 score. Manufacturer-level accuracy was also computed, defined as the proportion of known test images whose predicted class belonged to the correct manufacturer. Threshold-independent open-set performance was evaluated with unknown

images using four metrics, the area under the receiver operating curve (AUROC), the area under the precision-recall curve (AUPR), the false positive rate at a 95% true positive rate on the known test set (FPR95), and the area under the open-set classification rate curve (OSCR). At the threshold τ, we reported the precision, recall, and F1 score for the unknown class, as well as the macro-averaged F1 score over known classes with rejection applied. Open-set performance was evaluated using all unknown images and, separately, using unknown images from shared sites only. To account for repeated imaging of the same implant, we additionally performed implant-level evaluation by averaging the scores across all images of each implant.

2.7 Subgroup and Error Analysis

Unknown detection rates at τ were compared between known-manufacturer and new-manufacturer unknowns, stratified by site type. Detection rates were also computed for each unknown class. Classes with fewer than 30 images or fewer than 10 implants were pooled for reporting. To characterize recognition failures, we analyzed two types of errors. First, we computed the proportion of closed-set misclassifications among known classes in which the predicted class belonged to the same manufacturer as the true class. Second, for unknown images that were not rejected, we recorded the known classes to which they were assigned. Confusion matrices and assignment tables were aggregated over the ten splits.

2.8 Statistical Analysis

All metrics are reported as the mean ± standard deviation over the ten splits. The primary method was compared with each other method using the two-sided Wilcoxon signed-rank test on paired per-split values. We also report the mean paired difference with its 95% confidence interval. P values were adjusted to multiple comparisons using the Holm-Bonferroni method within each metric, with $p < 0.05$ after adjustment considered statistically significant. Analyses were performed with Python 3.13.5, PyTorch 2.11 [39], scikit-learn 1.6.1 [40], and SciPy 1.16.3 [41]. All codes are publicly available at https://github.com/errory230/Dental-Implant-Classification.

3. Result

3.1 Dataset characteristics

The final dataset comprised 100,105 images of 2,931 implants from 15 institutions (Table 1). The 13 known classes accounted for 86,101 images (20,660 implants), ranging from 2,218 images (is2) to 15,376 images (ts3). The 29 unknown classes accounted for 14,004 images (3,271 implants), of which 6,200 images belonged to 12 known-manufacturer classes and 7,804 images to 17 new-manufacturer classes. The number of images per implant ranged from 1 to 36 (mean 4.2, median 3). Across the ten splits, the known images were divided into a mean of 51,779 training (50272–52348), 8,646 validations (8,020 – 9,228), and 25,676 test images (24,857 – 26,601), and all 14,004 unknown images were used for testing.

3.2 Closed-set classification performance

On the known test set without rejection, the base classifier achieved an accuracy of 0.906 ± 0.009. With OpenMax recalibration, accuracy was 0.901 ± 0.009 and the macro-averaged F1 score was 0.854 ± 0.008 (Table 2). Per-class F1 scores exceeded 0.90 for nine of the 13 classes; the lowest were observed for *is2*, *in*, *ex*, and *is*. Misclassifications were concentrated between products of the same manufacturer. 68.0% ± 2.6% of all closed-set errors involved a predicted class from the same manufacturer as the true class, most frequently between *ex* and *in* (Warantec), among *is*, *is2*, and *is3* (Neobiotech), and between *im* and *su* (Dentium) (Fig. 2a). Consequently, manufacturer-level accuracy was 0.968 ± 0.004.

Table 2
Closed-set classification performance for known classes

Class	Precision	Recall	F1 score	Test images
ts3	0.965 ± 0.008	0.982 ± 0.008	0.973 ± 0.004	4568
ast	0.971 ± 0.007	0.962 ± 0.014	0.966 ± 0.008	3149
us2	0.949 ± 0.018	0.956 ± 0.011	0.952 ± 0.006	1838
gs3	0.951 ± 0.021	0.928 ± 0.016	0.939 ± 0.007	1156
su	0.919 ± 0.018	0.911 ± 0.019	0.915 ± 0.015	3950
lu	0.932 ± 0.019	0.899 ± 0.024	0.915 ± 0.019	720
im	0.900 ± 0.023	0.923 ± 0.014	0.911 ± 0.012	3612
bra	0.914 ± 0.029	0.888 ± 0.048	0.900 ± 0.032	794
is3	0.873 ± 0.042	0.885 ± 0.020	0.878 ± 0.023	1811
is	0.845 ± 0.052	0.769 ± 0.048	0.803 ± 0.028	1631
ex	0.742 ± 0.047	0.714 ± 0.058	0.725 ± 0.024	1101
in	0.658 ± 0.047	0.670 ± 0.082	0.659 ± 0.027	736
is2	0.525 ± 0.066	0.616 ± 0.080	0.559 ± 0.033	609
Macro average			0.854 ± 0.008	
Accuracy			0.901 ± 0.009	
Manufacturer-level accuracy			0.968 ± 0.004	

Each cell shows the percentage of images from the true class assigned to each predicted class, with each row summing to 100%. Values below 1% are not shown. For closed-set setting, classes are grouped by manufacturer, as indicated above the matrices, and red boxes denote classes from the same manufacturer.

3.3 Comparison of open-set recognition methods

OpenMax achieved an AUROC of 0.898 ± 0.009, an AUPR of 0.825 ± 0.015, an OSCR of 0.830 ± 0.014, and an FPR95 of 0.389 ± 0.035 (Fig. 2b). Table 3 and Fig. 3a-d summarize the performance of all methods. At the threshold set on the known validation set, the proportion of known test images accepted was 0.949 ± 0.003 for OpenMax, consistent with the target of 0.95. At the threshold, it detected 61.4 ± 4.0% of unknown images with a precision of 0.868 ± 0.008 (unknown F1, 0.718 ± 0.029). The macro-averaged F1 score over known classes with rejection applied was 0.766 ± 0.011. When the scores of all images of an implant were averaged, the implant-level AUROC was 0.926 ± 0.007. The four score-based methods and ARPL performed similarly to one another, with AUROCs between 0.825 and 0.844 and unknown recall between 0.291 and 0.494. MSP had the lowest unknown recall (0.291 ± 0.025). OpenMax outperformed each of these five methods in AUROC, OSCR, and unknown recall in all ten splits (all adjusted p = 0.012 for every comparison; Fig. 3d). Compared with ARPL, OpenMax improved AUROC by 0.054 (95% CI, 0.041–0.067) and unknown recall by 0.120 (95% CI, 0.083–0.156). OpenMax recalibration reduced closed-set accuracy by 0.005 (95% CI, 0.005–0.006) relative to the base classifier.

Table 3
Comparison of open-set recognition methods

Method	Closed-set accuracy	AUROC	OSCR	FPR95	Unknown precision	Unknown recall
All unknown images (n = 14004)						
MSP	0.906 ± 0.009	0.825 ± 0.017	0.778 ± 0.019	0.708 ± 0.033	0.759 ± 0.023	0.291 ± 0.025
MLS	0.906 ± 0.009	0.840 ± 0.023	0.780 ± 0.022	0.516 ± 0.041	0.839 ± 0.013	0.488 ± 0.051
Energy	0.906 ± 0.009	0.840 ± 0.023	0.780 ± 0.023	0.510 ± 0.039	0.841 ± 0.012	0.493 ± 0.050
ODIN	0.906 ± 0.009	0.843 ± 0.024	0.784 ± 0.024	0.516 ± 0.053	0.839 ± 0.015	0.485 ± 0.060
ARPL	0.906 ± 0.009	0.844 ± 0.018	0.785 ± 0.018	0.515 ± 0.054	0.834 ± 0.019	0.494 ± 0.051
MAV distance	0.906 ± 0.009	0.904 ± 0.006	0.834 ± 0.010	0.393 ± 0.023	0.867 ± 0.006	0.611 ± 0.028
OpenMax	**0.901 ± 0.009**	**0.898 ± 0.009**	**0.830 ± 0.014**	**0.389 ± 0.035**	**0.868 ± 0.008**	**0.614 ± 0.040**
Unknown images from shared institutions only (n = 10527)						
MSP	–	0.803 ± 0.018	0.759 ± 0.020	0.739 ± 0.028	0.680 ± 0.026	0.261 ± 0.022
MLS	–	0.809 ± 0.027	0.755 ± 0.026	0.578 ± 0.039	0.773 ± 0.016	0.426 ± 0.051
Energy	–	0.810 ± 0.027	0.754 ± 0.026	0.573 ± 0.038	0.776 ± 0.016	0.430 ± 0.051
ODIN	–	0.816 ± 0.028	0.761 ± 0.027	0.567 ± 0.052	0.778 ± 0.020	0.434 ± 0.060
ARPL	–	0.816 ± 0.024	0.760 ± 0.021	0.573 ± 0.059	0.769 ± 0.029	0.435 ± 0.054
MAV distance	–	0.900 ± 0.008	0.832 ± 0.010	0.428 ± 0.025	0.823 ± 0.008	0.577 ± 0.030
OpenMax	**–**	**0.892 ± 0.010**	**0.825 ± 0.014**	**0.412 ± 0.032**	**0.826 ± 0.009**	**0.591 ± 0.038**

3.4 OpenMax performance and robustness

Using the distance to the mean activation vector without Weibull recalibration yielded an AUROC of 0.904 ± 0.006 and an unknown recall of 0.611 ± 0.028. This approach had a slightly higher AUROC than OpenMax (mean difference, 0.006; 95% CI, 0.001–0.010; Holm-adjusted P = 0.027), whereas unknown

recall did not differ between the two methods (mean difference, 0.003; 95% CI, − 0.018 to 0.023; Holm-adjusted P = 0.70). Across hyperparameter settings, OpenMax AUROC ranged from 0.858 to 0.901 for tail sizes of 10–100 and α values of 1–13 and exceeded that of ARPL in every setting (Fig. 3e). In a two-dimensional t-SNE embedding of the logit vectors, unknown images overlapped partially with known-class clusters but were concentrated at their periphery and in regions between clusters (Fig. 4a).

3.5 Effect of institution on unknown detection

Of the 14,004 unknown images, 3,477 (24.8%) originated from the two institutions that contributed no known-class images. With OpenMax, the detection rate was 68.4% ± 5.7% for unknown images from these unknown-only sites and 59.1 ± 3.8% for those from shared sites, a difference of 9.3 percentage points (%p). This difference was larger for the score-based methods and ARPL, ranging from 12.1%p for MSP to 25.3%p for Energy, and was 23.7%p for ARPL. When restricted to shared sites, the AUROC of OpenMax decreased by only 0.006 (0.898 to 0.892), compared with decreases of 0.022–0.031 for the score-based methods and 0.028 for ARPL. The advantage of OpenMax over these methods therefore increased in the shared-site evaluation (Table 3). To assess whether the observed differences could be explained by class composition, we compared detection rates within the six unknown classes with at least 20 images from each site type. Detection was lower at unknown-only sites than at shared sites for four classes: *ar* (54.5% vs. 81.8%), *aoe* (70.5% vs. 83.3%), *ss3* (81.9% vs. 92.8%), and *tspl* (80.5% vs. 85.2%). It was higher for two classes: *uf* (57.4% vs. 35.4%) and *aoi* (70.8% vs. 56.6%). The *it* class, for which all images originated from an unknown-only site, was almost never detected (0.6% ± 0.6%). Thus, unknown images from unknown-only sites were not consistently more likely to be rejected.

3.6 Unknown detection by manufacturer group and class

OpenMax detected 57.5% ± 3.3% of known-manufacturer unknown images and 64.4% ± 5.1% of new-manufacturer unknown images. Detection was consistently lower for known-manufacturer unknowns at both shared sites (56.9% vs. 61.3%) and unknown-only sites (61.3% vs. 70.5%). Consistent with this difference, the distribution of OpenMax unknown probabilities for known-manufacturer unknowns overlapped more with that of known images (Fig. 4b). In the t-SNE embedding of the logit vectors, unknown images overlapped partially with known-class clusters but were concentrated at their periphery and in regions between clusters, and known-manufacturer unknowns lay closer to the clusters of known systems than new-manufacturer unknown did (Fig. 4c). Compared with ARPL, OpenMax improved detection by 18.6 percentage points for known-manufacturer unknowns (38.9% to 57.5%) and by 6.7 percentage points for new-manufacturer unknowns (57.7% to 64.4%). Detection rates varied substantially across unknown classes (Table 4). Five classes (*ss3*, *xin*, *ts*, *xi*, and *tspl*) had detection rates above 80%, whereas five classes (*eb*, *uf*, *scl*, *co1*, and *it*) had rates below 40%.

Table 4
Unknown detection rate of OpenMax by class

Class	Manufacturer	Model	Images	Implants	Institutions	Detection rate
Known-manufacturer unknown classes						
ss3	Osstem	SS III	87	33	4	0.894 ± 0.045
xi	Dentsply	Xive	533	105	2	0.824 ± 0.064
rep	Nobel Biocare	Replace Select	1513	188	2	0.706 ± 0.049
ss2	Osstem	SS II	571	166	5	0.704 ± 0.038
gs2	Osstem	GS II	1120	206	4	0.549 ± 0.096
us3	Osstem	US III	606	200	4	0.536 ± 0.051
eb	Neobiotech	EB	1512	264	3	0.395 ± 0.045
it	Warantec	IT	190	60	1	0.006 ± 0.006
New-manufacturer unknown classes						
xin	Megagen	Exfeel Internal	362	65	1	0.889 ± 0.065
ts	Straumann	TSstandard	920	142	4	0.830 ± 0.042
tspl	Straumann	TS Standard Plus	592	122	2	0.821 ± 0.062
ar	Megagen	AnyRidge	173	38	3	0.766 ± 0.069
aoe	Megagen	AnyOne External	1032	315	3	0.711 ± 0.068
aoi	Megagen	AnyOne Internal	1312	606	5	0.658 ± 0.063
uf2	DIO	UF II	357	75	2	0.618 ± 0.093
bl	Straumann	Bone Level	1077	241	4	0.616 ± 0.101

Class	Manufacturer	Model	Images	Implants	Institutions	Detection rate
Known-manufacturer unknown classes						
tsv	Zimmer	Tapered Screw-Vent	136	41	3	0.610 ± 0.125
xex	Megagen	Exfeel External	778	103	1	0.514 ± 0.088
3i	Biomet 3i	3i	160	29	2	0.438 ± 0.079
uf	DIO	UF	419	111	3	0.382 ± 0.074
scl	Dentis	S-Clean Tapered	87	18	1	0.299 ± 0.093
co1	Cybermed	Core1	329	88	1	0.158 ± 0.048
Pooled*	7 classes		138	55		0.678 ± 0.055

*Classes with fewer than 30 images or fewer than 10 implants were pooled: *sim*, *iu*, *ank*, and *sl* (known manufacturer) and *ati*, *sv*, and *bio* (new manufacturer).

3.7 Assignment of undetected unknown implants

Undetected unknown images were not distributed uniformly across known classes but were often concentrated in one or two classes (Fig. 5). For known-manufacturer unknowns, undetected images were frequently assigned to another implant system from the same manufacturer: *eb* was assigned to *is* in 94% of cases (Neobiotech), *it* to *in* in 78% and *iu* to *ex* in 91% (Warantec), *us3* and *ss2* to *us2* in 55% and 76%, respectively (Osstem), and *gs2* to *gs3* in 50% (Osstem). For new-manufacturer unknowns, several systems were assigned predominantly to a single known class from a different manufacturer: *xex* and *3i* to *us2* (96% each), *co1* to *ts3* (94%), tsv to *ast* (91%), *bl* to *ts3* (83%), and *scl* to *su* (82%).

For each unknown class (row), the percentage of undetected images – images accepted as known by OpenMax – assigned to each known class (column), pooled across the ten data splits. Each row sums to 100%, and values below 5% are not shown.

4. Discussion

In this study, we evaluated dental implant identification under open-set conditions, in which a model encounters implant systems absent from its training data. Using a large multicenter dataset, we compared score-based, distance-based, and training-based open-set recognition methods. OpenMax achieved the highest unknown-detection performance, identifying 61% of unregistered implant images at

a threshold that accepted 95% of registered implants. Its advantage appeared to arise primarily from measuring the distance between an input logit vector and the class-specific mean activation vectors. OpenMax was also less affected by institution of origin than the other methods. Importantly, undetected unknown implants were not assigned randomly across known classes but were often concentrated in a small number of known systems, including systems from the same manufacturer.

Previous deep learning studies of dental implant identification have reported high classification accuracy under closed-set conditions [15–21]. Our closed-set accuracy of 0.906 falls within the range of previous results, but two important differences in study design should be considered. First, we kept all images of the same physical implant, including follow-up radiographs, within a single subset. This grouping accounted for the repeated imaging of implants in our dataset, with 95% of images belonging to implants with multiple radiographs. Second, a closed-set model assigns every input to one of the known classes, regardless of whether the implant is represented in the training data. In our dataset, this would have forced all unknown images to be assigned to a known class, creating the potential for unrecognized misclassification. Thus, closed-set accuracy alone may provide an incomplete assessment of the reliability and practical utility of an implant identification system in a setting where previously unseen implant systems may be encountered.

The advantage of a distance-based method may be related to the characteristics of the unknown implants in this task. Many unregistered systems appeared sufficiently similar to registered systems that the classifier assigned them to a known class with high confidence. Therefore, methods that rely on a single summary of confidence may have difficulty distinguishing these inputs from known samples. In contrast, the distance to mean activation vector considers the relative pattern of logits across classes, allowing an input to be rejected when its pattern differs from that typical of the predicted class, even when its maximum logit is high. Together with previous results from natural image benchmarks [42], these findings suggest that the relative effectiveness of open-set methods may depend on the characteristics of the task. Fine-grained medical tasks in which unknown classes are close variant of known ones may be particularly suited to distance-based approaches, although this should be evaluated in other domains.

These findings have direct practical implications. Distance score can be applied post hoc and require only the logits of the training set, allowing them to be added to an existing implant classifier without retraining or architectural changes. Applying the rejection threshold reduced closed-set accuracy by only 0.5 percentage points. In clinical use, implants flagged as unknown could be referred to expert review or comparison with catalogs rather than being assigned a potentially incorrect system. Even among flagged implants, the manufacturer-level accuracy of 96.8% suggests that the predicted manufacturer could help narrow the search. In addition, averaging scores across all available radiographs of an implant improved the AUROC from 0.898 to 0.926, suggesting that prior radiographs, if available, can improve the reliability of unknown rejection.

Recognition failures followed systematic patterns. Unregistered products from manufacturers already represented in the training set were harder to detect than products from new manufacturers, and undetected images were often assigned to another system from the same manufacturer, such as *US III* to US *II* or *GS II* to *GS III*. This pattern may reflect shared design features across manufacturers, like successor systems. Among new-manufacturer systems, several were also assigned predominantly to a single known system from a different manufacturer, suggesting substantial visual similarity in panoramic radiographs. Identifying known systems that frequently absorb unknown implants may help clinicians interpret predictions for those classes with appropriate caution.

Multicenter datasets can contain institution-specific imaging characteristics that may confound the model [43, 44]. In our dataset, two institutions contributed only unknown-class images, raising the possibility that some unknown images were detected partly through site-specific characteristics rather than implant morphology. However, several observations suggested that the performance of OpenMax was not primarily driven by site-specific characteristics. Restricting the evaluation to shared sites reduced its AUROC only modestly. Restricting the evaluation to shared sites reduced its AUROC only modestly, whereas the reduction was larger for the other methods. Detection also did not consistently increase at unknown-only sites within the same unknown class, and the it class, which originated entirely from an unknown-only site, was almost never detected. Nevertheless, the greater institution-related variation observed for the score-based methods and ARPL indicates that site-specific characteristics can influence open-set evaluation. Therefore, future multicenter studies should report performance stratified by institution to assess the robustness of unknown detection across imaging settings.

This study has several limitations. First, all data were derived from a single public dataset comprising 15 Korean institutions, and the distribution of implant systems reflects the Korean market. External validation is therefore needed to determine whether the findings generalize to other markets and imaging settings. The absence of patient identifiers also limited patient-level grouping. Although grouping by institution and placement date kept images of the same implant together, implants placed in the same patient on different dates could not be linked. Our definition of unknown classes also has limitations. Unknown systems were defined using an image-frequency threshold rather than implant release date. A temporal design, in which models trained on earlier systems are evaluated on subsequently introduced systems, would more directly simulate the evolution of the implant market. Some unknown classes also contained relatively few images or implants, making their class-specific detection estimates less certain. Finally, the images were fixture crops prepared by the dataset provider, whereas a deployed system would additionally require reliable automated implant detection and localization. Moreover, approximately 40% of unknown implants were still classified as known systems. Open-set recognition should therefore be viewed as a supplementary safeguard that supports, rather than replaces, clinical judgment.

Future work should validate these findings on external and temporally separated datasets and investigate incremental learning, allowing newly encountered systems to be incorporated into the model after expert identification. Further analysis of failure cases in relation to implant design features, such as

connection type and body geometry, may also clarify why certain unregistered systems are difficult to distinguish from registered ones.

Declarations

Ethics approval and consent to participate

The study was exempted from IRB review (Exemption No. P01-202511-01-038), and the requirement for informed consent was waived.

Availability of data and materials

The datasets used and/or analyzed during the current study are available in the ‘Radiographic Image Data of Dental Implants’ provided by AI-Hub (https://aihub.or.kr) and operated by the National Information Society Agency (NIA) of South Korea. The data are publicly available to researchers upon registration and approval through the AI-Hub portal.

Competing interests

The authors declare that they have no competing interests

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2026-25578241)

Acknowledgements

Not applicable

References

1. Kaur M, Abou-Arraj RV, Lin CP, Geisinger ML, Geurs NC. A 5-Year Retrospective Analysis of Biologic and Prosthetic Complications Associated With Single-Tooth Endosseous Dental Implants: Practical Applications. Clin Adv Periodontics. 2021 Dec;11(4):225-232. doi: 10.1002/cap.10155. Epub 2021 Apr 16. PMID: 33829671; PMCID: PMC8928153.
2. Jadhav, L., Madiwal, V. & Rajwade, J.M. The dental implant surface: a review of the past, present and future. J Mater Sci: Mater Med 37, 21 (2026). https://doi.org/10.1007/s10856-025-06997-x
3. Wisdom, C., Chen, C., Yuca, E. et al. Repeatedly Applied Peptide Film Kills Bacteria on Dental Implants. JOM 71, 1271–1280 (2019). https://doi.org/10.1007/s11837-019-03334-w
4. L. Hjalmarsson, M. Gheisarifar, and T. Jemt, “A Systematic Review of Survival of Single Implants as Presented in Longitudinal Studies With a Follow-Up of at Least 10 Years,” European Journal of Oral Implantology 9, no. Suppl 1 (2016): S155–S162.

5. Barkarmo, S. and Kowar, J. (2025), Outcome of Single Dental Implants Over 38–40 Years: A Long-Term Follow-Up Study. Clin Implant Dent Relat Res, 27: e13443. https://doi.org/10.1111/cid.13443
6. Sharma D, Jhingta P, and Bathla S. Challenges in dental implant identification and need of universal dental implant identification, numbering, and nomenclature system. Int J Prosthodont Restor Dent 2020;10(4):146-150
7. Nuzzolese, E., Lusito, S., Solarino, B. & Di Vella, G. Radiographic dental implants recognition for geographic evaluation in human identification. J. Forensic Odontostomatol. 26, 8–11 (2008).
8. Heitz-Mayfield LJA, Salvi GE. Peri-implant mucositis. Proceedings from the 2017 World Workshop for Classification of Periodontal and Peri-implant Diseases and Conditions. J Periodontol. 2018; 89(1): S257-S266.Suppl.
9. Salvi GE, Aglietta M, Eick S, Sculean A, Lang NP, Ramseier CA. Reversibility of experimental peri-implant mucositis compared with experimental gingivitis in humans. Clin Oral Implants Res. 2012; 23: 182-190.
10. Schwarz F, Derks J, Monje A, Wang H-L. Peri-implantitis. Proceedings from the 2017 World Workshop for Classification of Periodontal and Peri-implant Diseases and Conditions. J Periodontol. 2018; 89(Suppl): S267-S290.
11. Oh S-L, Shiau HJ, Reynolds MA. Survival of dental implants at sites after implant failure: a systematic review. J Prosth Dent. 2020; 123: 54-60.
12. Go, H., Ahn, J., Ahn, S. V., Shin, H., Amano, A., & Choi, Y. (2024). Estimated economic costs of dental implants among older South Korean adults, 2015–2018. Clinical Oral Implants Research, 35(12), 1560–1567. https://doi.org/10.1111/clr.14343
13. Saleh, M. H. A., Dias, D. R., & Kumar, P. (2024). The economic and societal impact of periodontal and peri-implant diseases. Periodontology 2000, 98(1), 100–118. https://doi.org/10.1111/prd.12568
14. Saleh MHA, Dias DR, Kumar P. The economic and societal impact of periodontal and peri-implant diseases. Periodontol 2000. 2025 Jun;98(1):100-118. doi: 10.1111/prd.12568. Epub 2024 May 1. PMID: 38693603; PMCID: PMC12842886.
15. Ardakani MSZ, Bayati M. Global situation of oral health coverage toward universal health coverage: A scoping review. Preventive Medicine Reports, vol 58, October 2025, 103227
16. Takahashi, T., Nozaki, K., Gonda, T. et al. Identification of dental implants using deep learning—pilot study. Int J Implant Dent 6, 53 (2020). https://doi.org/10.1186/s40729-020-00250-6
17. Sukegawa S, Yoshii K, Hara T, Yamashita K, Nakano K, Yamamoto N, Nagatsuka H, Furuki Y. Deep Neural Networks for Dental Implant System Classification. Biomolecules. 2020; 10(7):984. https://doi.org/10.3390/biom10070984
18. Kim J-E, Nam N-E, Shim J-S, Jung Y-H, Cho B-H, Hwang JJ. Transfer Learning via Deep Neural Networks for Implant Fixture System Classification Using Periapical Radiographs. Journal of Clinical Medicine. 2020; 9(4):1117. https://doi.org/10.3390/jcm9041117
19. Marta R-L. Miguel G-P, Shantanu V., et al. Artificial intelligence applications in implant dentistry: A systematic review. The Journal of Prosthetic Dentistry. vol 129. Feb 2023.

https://doi.org/10.1016/j.prosdent.2021.05.008

20. Hadj Saïd M, Le Roux MK, Catherine JH, Lan R. Development of an Artificial Intelligence Model to Identify a Dental Implant from a Radiograph. Int J Oral Maxillofac Implants. 2020 Nov/Dec;36(6):1077-1082. doi: 10.11607/jomi.8060. PMID: 33270045.
21. 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. https://doi.org/10.3390/diagnostics10110910
22. Lee, Jae-Hong DDS, PhD∗; Jeong, Seong-Nyum DDS, PhD. 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 99(26):p e20787, June 26, 2020. | DOI: 10.1097/MD.0000000000020787
23. U.S. Food and Drug Administration. 510(k) Premarket Notification: Bioline Dental Implant System (K250872). 2025.
24. SCHEIRER, Walter J., et al. Toward open set recognition. IEEE transactions on pattern analysis and machine intelligence, 2012, 35.7: 1757-1772.
25. Hendrycks, Dan, et al. "The many faces of robustness: A critical analysis of out-of-distribution generalization." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
26. Camila Gonzalez, et al. Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation. Medical Image Analysis. Nov 2022, 102596, vol 82. 10.1016/j.media.2022.102596
27. Milda Pocevičiūtė et al. Out-of-distribution detection in digital pathology: Do foundation models bring the end to reconstruction-based approaches? Computers in Biology and Medicine. Vol 184. Jan 2025. 109327. https://doi.org/10.1016/j.compbiomed.2024.109327
28. A G Roy, et al. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. Medical image analysis. Jan 2022, 102274, vol 75. Doi: 10.1016/j.media.2021.102274
29. Q. Tan, L. Bai, G. Wang, M. Islam and H. Ren, "Endoood: Uncertainty-Aware Out-of-Distribution Detection in Capsule Endoscopy Diagnosis," 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 2024, pp. 1-5, doi: 10.1109/ISBI56570.2024.10635759
30. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations (ICLR). 2017.
31. Bendale A, Boult TE. Towards open set deep networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:1563–1572.
32. Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2021. https://doi.org/10.48550/arXiv.2103.00953
33. National Information Society Agency. 'Radiographic Image Data of Dental Implants', AI-Hub, 2022, https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=536

34. Wightman, R., Raw, N., Soare, A., Arora,. A., Ha, C., Reich, C., Guan, F., & Kaczmarzyk, J. (2023). *PyTorch Image Models* (Version v0.8.10dev0) [Computer software]. Zenodo. doi.org
35. Xixi Liu, Yaroslava Lochman, Christopher Zach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23946-23955
36. Liu W, Wang X, Owens J, Li Y. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems. 2020;33.
37. Liang et al. "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks." International Conference on Learning Representations, 2018.
38. Chen G, Peng P, Wang X, Tian Y. Adversarial reciprocal points learning for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;44(11):8621–8632.
39. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. 2019;32:8024–8035.
40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
41. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.
42. Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2021). Open-Set Recognition: a Good Closed-Set Classifier is All You Need?. *Arxiv Preprint (Cs.cv)*. https://arxiv.org/abs/2110.06207v2
43. Tocilă-Mătășel C, Dudea SM, Iana G. Addressing Multi-Center Variability in Radiomic Analysis: A Comparative Study of Image Acquisition Methods Across Two 3T MRI Scanners. Diagnostics (Basel). 2025 Feb 17;15(4):485. doi: 10.3390/diagnostics15040485. PMID: 40002637; PMCID: PMC11854186.
44. Campello, V.M., Martín-Isla, C., Izquierdo, C. *et al.* Minimising multi-centre radiomics variability through image normalisation: a pilot study. *Sci Rep* **12**, 12532 (2022). https://doi.org/10.1038/s41598-022-16375-0

Figures

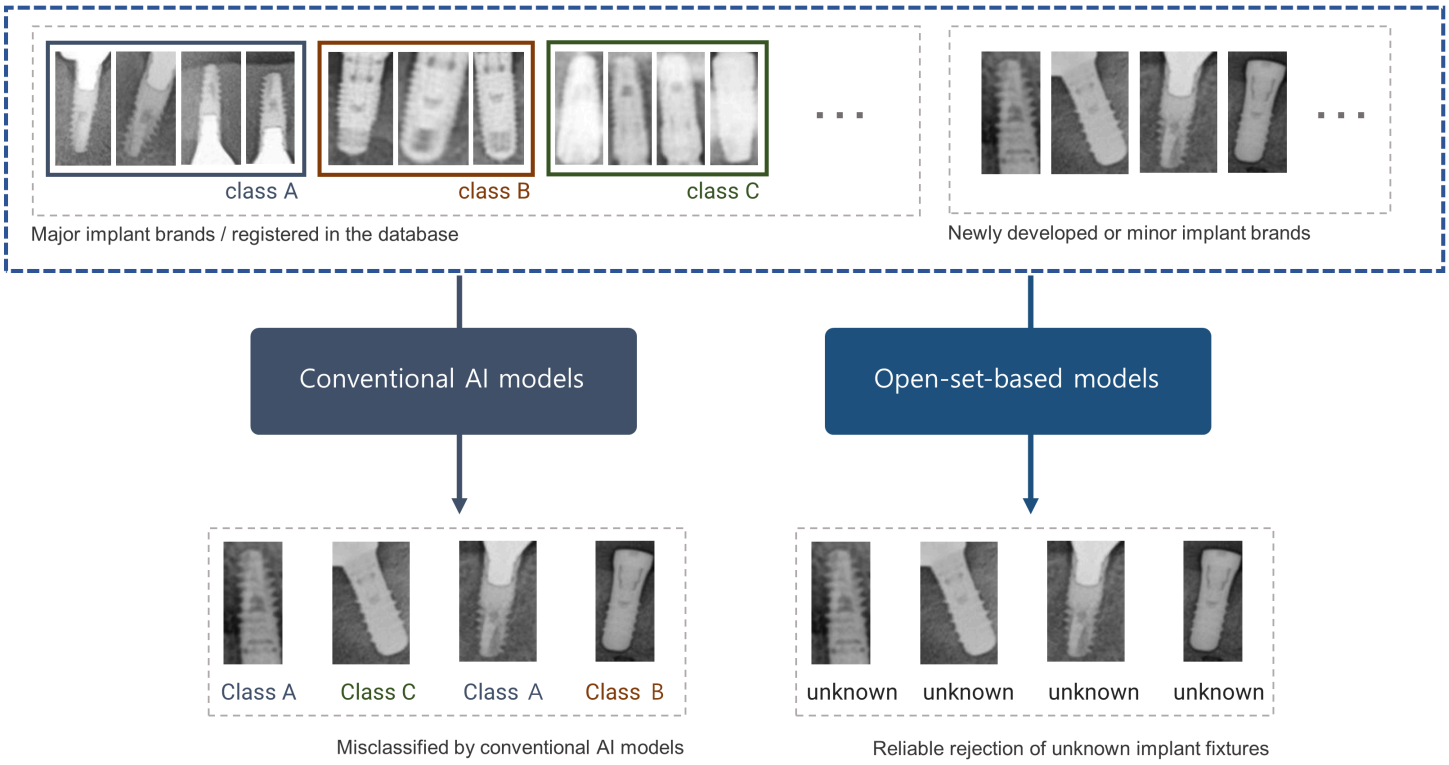


Figure 1

Overview of the study design and open-set recognition pipeline.

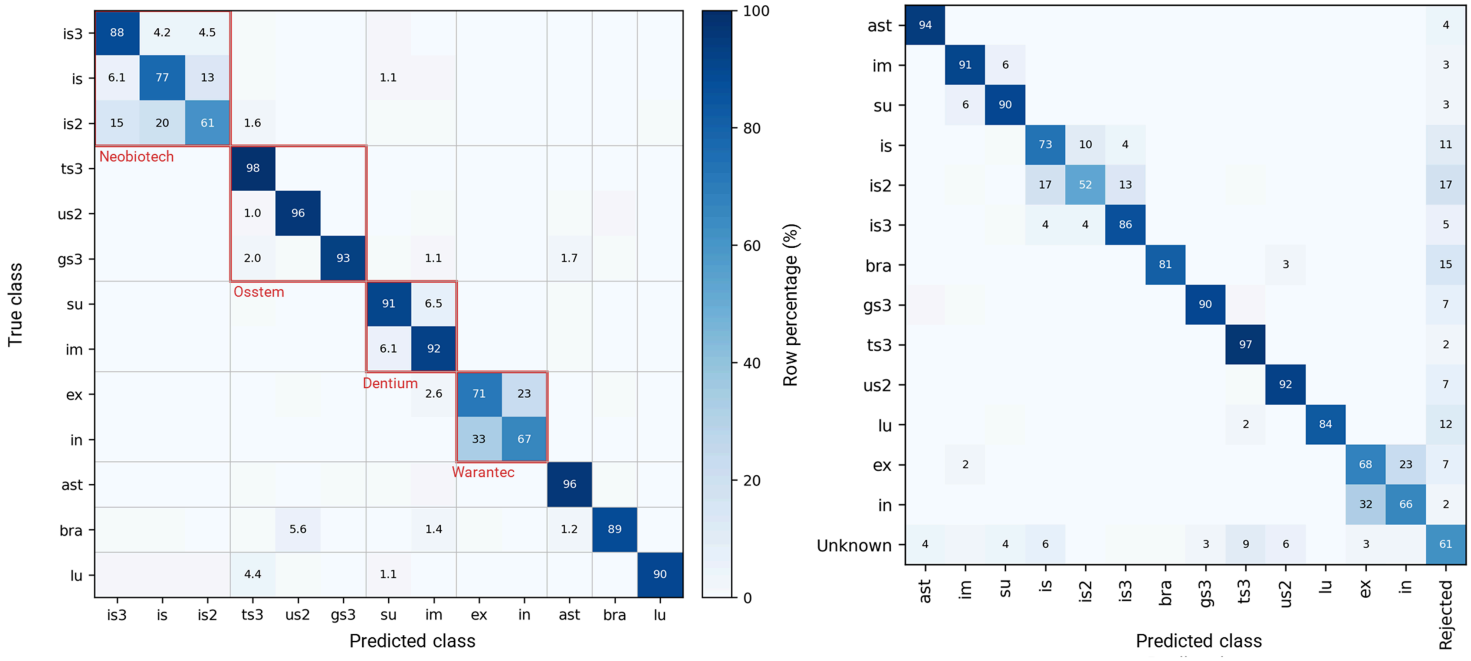


Figure 2

Confusion matrices for closed-set and open-set models.

Each cell shows the percentage of images from the true class assigned to each predicted class, with each row summing to 100%. Values below 1% are not shown. For closed-set setting, classes are grouped by manufacturer, as indicated above the matrices, and red boxes denote classes from the same manufacturer.

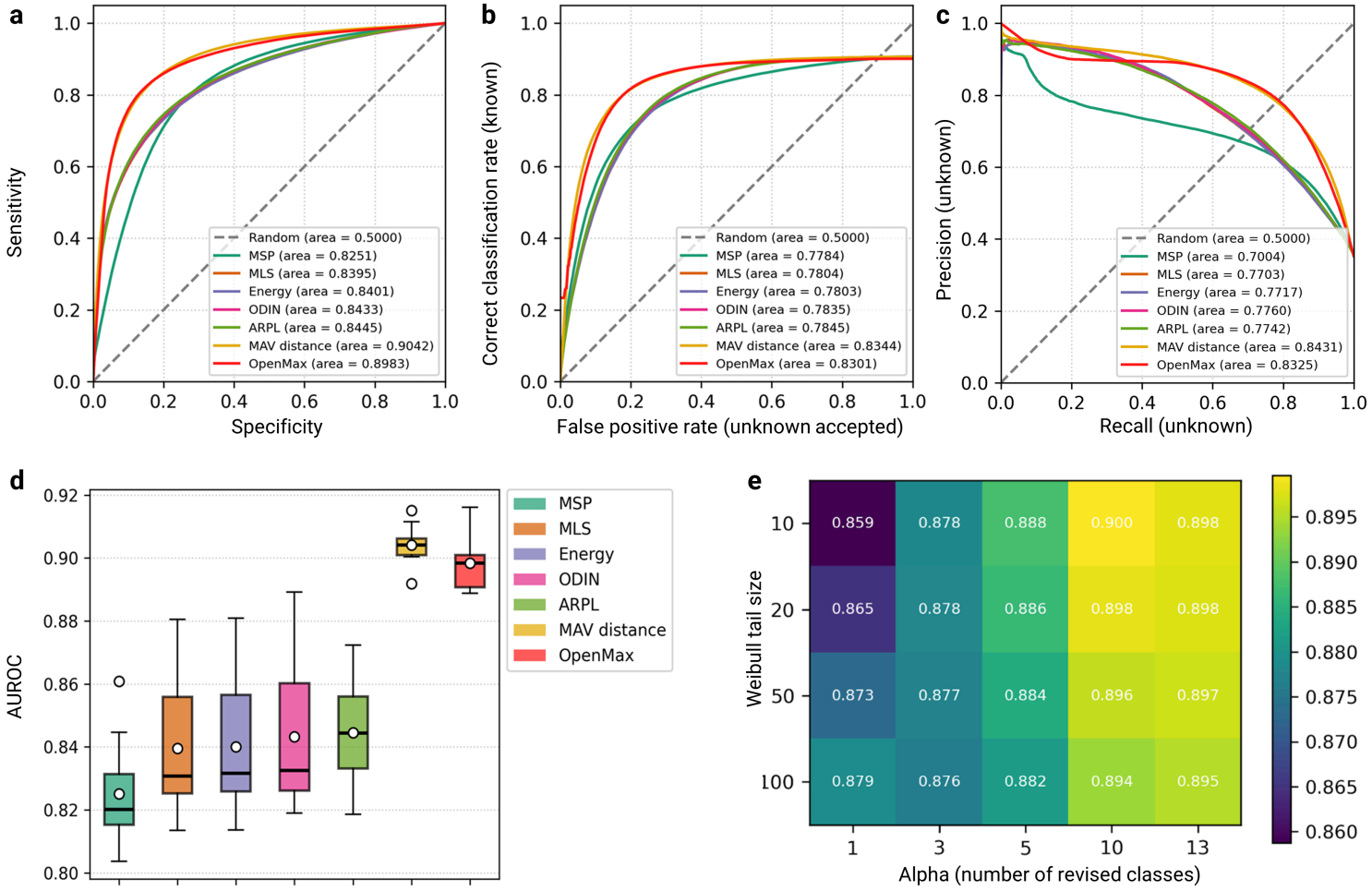


Figure 3

Comparison of open-set recognition methods.

a) ROC curves, b) open-set recognition curves (OSRCs), and c) precision-recall curves for each method. d) Distribution of AUROC values across ten cross-validation splits for each method. e) Effect of training hyperparameters on the AUROC of OpenMax.

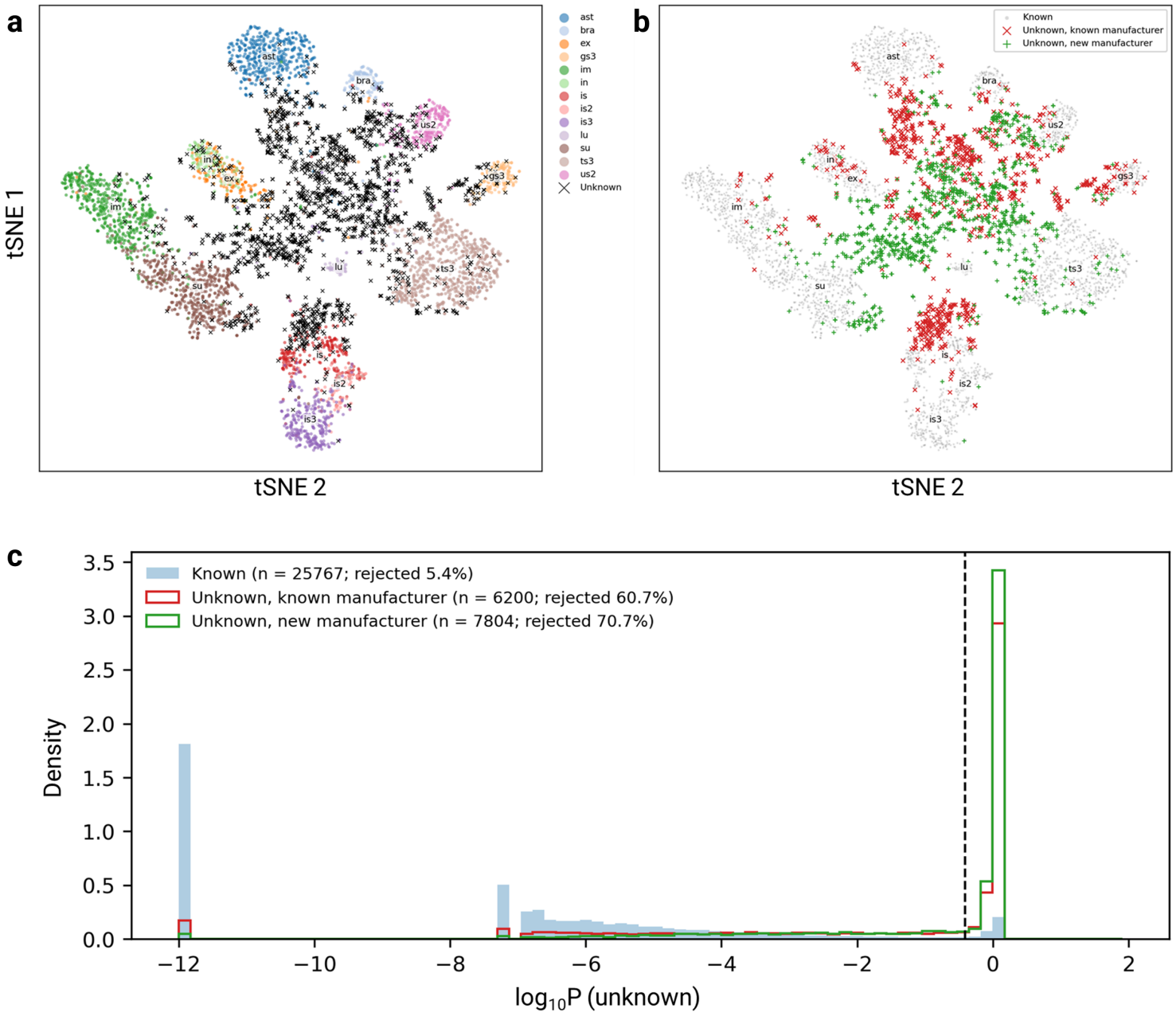


Figure 4

Representative analysis of known and unknown implants.

a) tSNE visualization of the data distribution, with unknown images indicated by black ×. b) tSNE visualization with unknown images separated into known-manufacturer and new-manufacturer groups. c) Distribution of distance scores across known and unknown image groups.

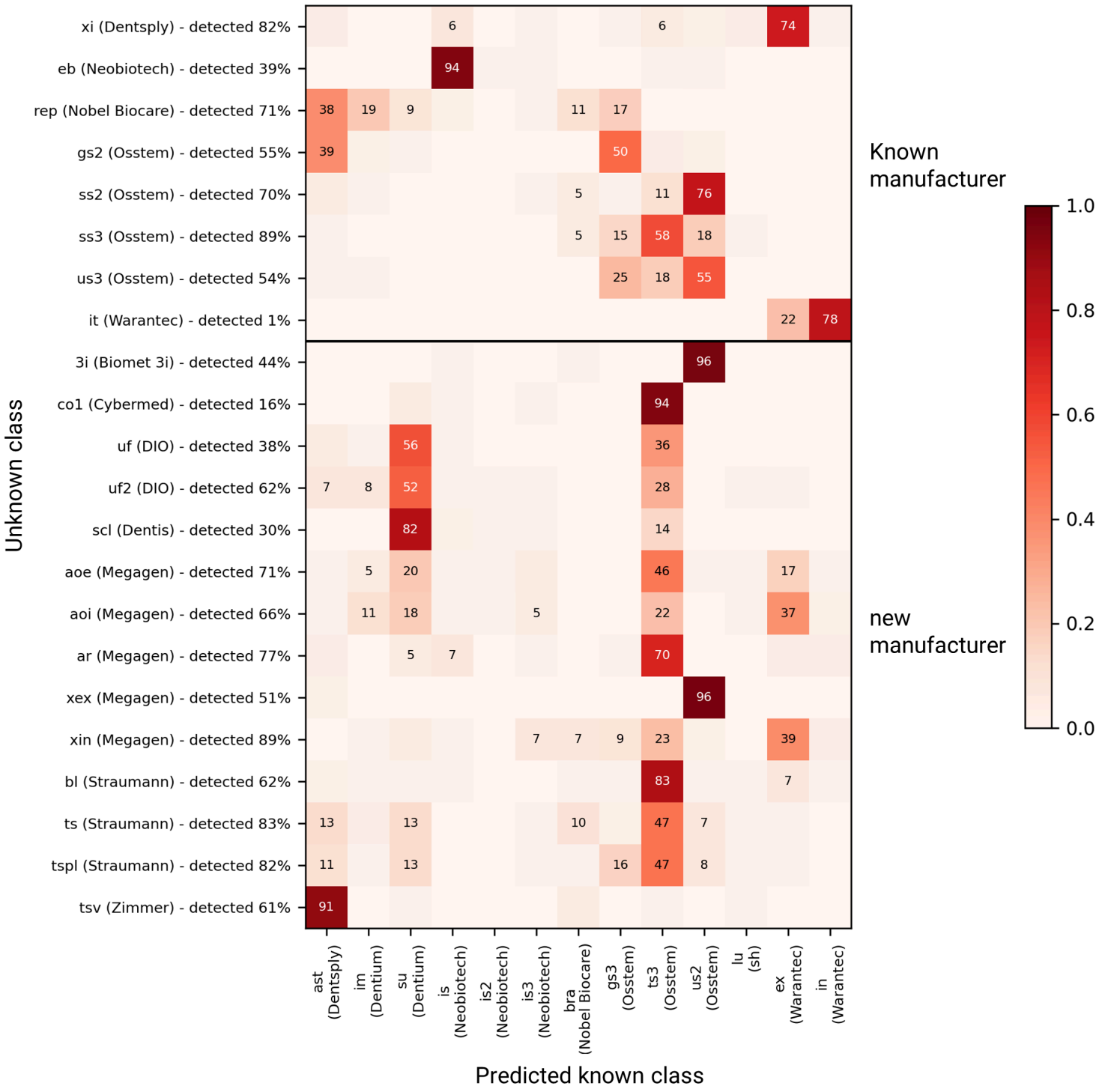


Figure 5

Assignment patterns of undetected unknown implants.

For each unknown class (row), the percentage of undetected images – images accepted as known by OpenMax – assigned to each known class (column), pooled across the ten data splits. Each row sums to 100%, and values below 5% are not shown.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx