



Learning curves in transfer-learning dental radiograph classification with pre-trained and random features

Curvas de aprendizaje en la clasificación de radiografías dentales por transferencia con características preentrenadas y aleatorias

Verónica Karina Arias Murillo¹

Melissa Castillo Jimenez²  

¹ Universidad de Ciencias Médicas de Santiago de Cuba. Santiago de Cuba, Cuba.

² Universidad de Ciencias Médicas de La Habana. La Habana, Cuba

Submitted: 16-12-2025 | Revised: 12-02-2026 | Accepted: 16-04-2026 | Published: 20-04-2026

Corresponding Author: Melissa Castillo Jimenez 

ABSTRACT

Introduction: The number of annotated images required to train a dental radiograph classifier is rarely quantified, and it is unknown whether transfer learning changes how fast performance saturates. **Methods:** Pre-trained and randomly initialized ResNet-18 features were extracted from 182 annotated teeth in the DENTEX dataset, and a logistic regression classifier was trained on stratified subsamples of 5, 10, 20, 40, 80 and 127 teeth over five seeds, with test accuracy recorded at each size. **Results:** With pre-trained features, accuracy rose from 50.2% at 5 teeth to 67.3% at 127 teeth, reaching 95% of full-data accuracy at 80 teeth. With random features, accuracy reached 58.2% at 127 teeth and plateaued earlier, at 40 teeth, but never matched the pre-trained curve. At 20 training teeth, pre-trained features classified caries at 80.7% but deep caries at 6.0% and no periapical lesion correctly. **Conclusions:** Transfer learning with pre-trained features reaches its plateau only around 80 annotated teeth, later than a random-feature baseline, and the small-sample performance is concentrated in the majority class, so the number of annotations a dental task requires depends on the class balance rather than on the overall accuracy alone.

Keywords: learning curve; transfer learning; dental radiograph; sample size; deep learning.

RESUMEN

Introducción: El número de imágenes anotadas necesario para entrenar un clasificador de radiografías dentales rara vez se cuantifica, y se desconoce si la transferencia de aprendizaje cambia la rapidez con que el rendimiento se satura. **Métodos:** Se extrajeron características de ResNet-18 preentrenado y aleatorio de 182 dientes anotados del conjunto DENTEX, y se entrenó un clasificador de regresión logística con submuestras estratificadas de 5, 10, 20, 40, 80 y 127 dientes en cinco semillas, registrando la exactitud de prueba en cada tamaño. **Resultados:** Con características preentrenadas, la exactitud subió de 50,2 % con 5 dientes a 67,3 % con 127, alcanzando el 95 % de la exactitud con datos completos a los 80 dientes. Con características aleatorias, la exactitud alcanzó el 58,2 % con 127 dientes y se estabilizó antes, a los 40 dientes, pero nunca igualó la curva preentrenada. Con 20 dientes de entrenamiento, las características preentrenadas clasificaron caries al 80,7 % pero caries profundas al 6,0 % y ninguna lesión periapical correctamente. **Conclusiones:** La transferencia con características preentrenadas alcanza su meseta solo alrededor de 80 dientes anotados, más tarde que una línea de base aleatoria, y el rendimiento con muestras pequeñas se concentra en la clase mayoritaria, de modo que el número de anotaciones que requiere una tarea dental depende del equilibrio de clases más que de la exactitud global.

Palabras clave: curva de aprendizaje; transferencia de aprendizaje; radiografía dental; tamaño muestral; aprendizaje profundo.

INTRODUCTION

Annotated data is the principal bottleneck in dental artificial intelligence, because every lesion, tooth and restoration in a training set must be labelled by a clinician whose time is scarce, and because the open dental collections that do exist are far smaller than the natural-image corpora on which general-purpose models are trained.(1,2) A research group that wants to build a caries detector or an implant classifier must therefore decide how many radiographs to annotate before the marginal benefit of another label stops justifying its cost, and in practice that decision is usually made by convention, habit or available budget rather than by measurement.(3)

The cost of that decision is asymmetric: annotating too little produces a model that fails silently on the classes it has barely seen, while annotating too much wastes scarce clinician effort that could have been spent on the next task, so a measured answer to the sample-size question is a practical necessity rather than a purely methodological nicety. The question of how many annotated images are needed is a learning-curve question, and it has a concrete answer for a given dataset and pipeline, yet it is seldom answered in the dental imaging literature.(4)

Transfer learning changes the shape of that curve. Instead of training a network from scratch, the standard dental pipeline reuses a backbone pre-trained on ImageNet and trains only a lightweight head, so that the model begins with a representation of visual structure that no amount of dental data could teach it from a small sample.(5,6)

The pre-trained backbone provides generic features, edges, textures and shapes, that transfer to a radiographic target even though ImageNet contains no radiographs, because the low-level statistics of natural images overlap sufficiently with those of medical images.(7-9) Studies of transfer learning in medical imaging have shown that pre-trained features often outperform features trained from scratch when data are scarce, and that the advantage depends on the similarity between the source and target domains, with natural-image features transferring well to radiographs despite the visual gap. What is less clear is how fast the transfer-learning advantage saturates, and whether a dental classifier reaches its ceiling after a handful of examples or only after hundreds, which is the distinction that determines whether annotation effort should be concentrated or spread thin.(4,10)

The question is sharpened by the class structure of dental diagnoses, because a curve that plateaus on overall accuracy may still be climbing on the rare classes that matter clinically, and the sample size at which the overall curve flattens is not necessarily the sample size at which every clinically important class has been adequately learned.

The present study contrasted two feature representations, namely pre-trained ResNet-18 features and randomly initialized ResNet-18 features, across a range of training sizes, so that the contribution of pre-training could be separated from the contribution of the architecture itself.

This contrast is informative because the two representations share exactly the same dimensionality and the same downstream classifier, and differ only in whether the backbone weights carry the statistics of natural images, which cleanly isolates the effect that transfer learning is claimed to provide.(9,10)

A random-feature baseline is a stronger control than no baseline at all, because it matches the architecture and therefore distinguishes the benefit of pre-training from the benefit of simply having a fixed nonlinear feature map, a distinction that is lost when transfer learning is compared only against a hypothetical from-scratch network that is never actually trained. The study also traced the curve separately for each diagnosis class, because a single overall curve can conceal the divergent behaviour of a majority class that saturates early and a rare class that remains unlearned, and the per-class view was expected to be at least as informative as the overall one. The objective was to trace the learning curve of a dental radiograph classifier as a function of the number of annotated training teeth and to identify the sample size at which performance plateaus, comparing pre-trained and random features.

METHODS

Dataset and feature extraction

This was an experimental learning-curve analysis on a public, openly licensed dataset, with no human participants and no ethics approval required beyond the terms of the licence. The DENTEX validation subset(2)provided 50 panoramic radiographs with 182 annotated teeth.

Each annotated tooth was cropped to its bounding box, resized to 224 by 224 pixels and normalized with the ImageNet statistics, and the diagnosis label of each tooth came from the four-class

taxonomy of the dataset, namely impacted tooth, caries, periapical lesion and deep caries. Two feature representations were extracted for every tooth with the same ResNet-18 architecture.⁽⁵⁾

The pre-trained representation used weights pre-trained on ImageNet, and the random representation used the identical architecture with its weights drawn from the default random initialization, so that the two feature sets differed only in the origin of the backbone weights and not in their dimensionality, their receptive field or their downstream task. In both cases the final fully connected layer was removed and the 512-dimensional penultimate representation was retained, giving two matrices of 182 rows and 512 columns that fed the same downstream classifier, and the extraction was performed in inference mode so that no gradient information was used.

Learning-curve construction

The learning curve was constructed by training a logistic regression classifier on stratified subsamples of increasing size and measuring its accuracy on a fixed held-out test set. The held-out test set was a single stratified 30% split of the 182 teeth, so that the same 55 test teeth were used at every training size and the accuracy values across the curve were directly comparable, which is essential because a comparison of accuracies measured on different test sets would conflate the effect of training size with the effect of test-set composition.

The training sizes were 5, 10, 20, 40, 80 and 127 teeth, where 127 was the full training split, and the sizes were chosen to span the range in which a dental classifier might plausibly plateau, from a handful of examples to the entire available collection. At each size the subsample was drawn to preserve the proportion of each diagnosis class, with at least one tooth per class where the class was present, so that the curve would not be distorted by a subsample that omitted an entire class by chance.

The logistic regression was fit after standardizing the features to zero mean and unit variance with a scaler computed on the training subsample alone, so that no information from the test set leaked into the preprocessing, and the regularization was fixed rather than tuned per subsample, so that the only variable changing across the curve was the number of training teeth. Five independent subsamples were drawn at each size, and the test accuracy was averaged across them, so that the reported curve reflects subsample-to-subsample variability rather than a single lucky or unlucky draw.

The same procedure was applied identically to the pre-trained and the random feature matrices, so that the comparison between the two curves was not confounded by any difference in the subsampling, the classifier or the evaluation, and any observed difference could be attributed to the feature representation alone.

Plateau definition

The plateau was defined as the smallest training size at which the mean test accuracy reached 95% of the accuracy obtained with the full training split, a relative convention that captures the point of diminishing returns without requiring the curve to be literally flat or presupposing a particular absolute accuracy as the target.⁽⁴⁾

The plateau was computed separately for the pre-trained and the random feature curves, because the two representations reached different full-data accuracies, and comparing them against a common absolute threshold would have conflated a difference in ceiling with a difference in data efficiency.

The 95% threshold was chosen in preference to a stricter 99% or a looser 90% because it marks the point at which the remaining five percent of accuracy would cost the largest share of additional annotation effort, which is precisely the information that an annotation budget most needs in order to decide where to stop.

The per-class accuracy at the 20-tooth training size was also recorded for the pre-trained features, to describe whether the small-sample performance was concentrated in the majority class or spread across the diagnosis classes, and that size was chosen because it fell within the early plateau where the overall curve first appeared to flatten.

Statistical analysis

The primary comparison was descriptive rather than confirmatory, because the interest lay in the shape and the plateau of the learning curve rather than in a single hypothesis test, and the accuracy at each size was reported as a mean with its standard deviation across the five subsamples.

The comparability of the two curves was ensured by construction, since the two feature matrices were extracted from the same teeth, split into the same train and test sets, and passed through

the same classifier and scaler at every size, so that no difference in preprocessing, regularization or evaluation could masquerade as a difference between the representations.

Five subsamples per size were chosen as a balance between the stability of the mean and the computational cost of retraining, and the standard deviations reported alongside the means allow the reader to judge how much of the curve reflects a consistent trend rather than subsample noise. Analyses were implemented in Python 3.13 with PyTorch 2.14 and scikit-learn 1.9, and the complete script was executed twice to confirm that every reported figure was reproducible and independent of run-to-run randomness.

RESULTS

Dataset characteristics

The dataset contributed 182 annotated teeth obtained from 50 panoramic radiographs (Table 1). Caries was the most frequent diagnosis with 101 teeth, followed by impacted teeth with 40, deep caries with 32 and periapical lesions with 9, an imbalance that is central to the interpretation of the learning curve, because the number of examples available for each class at a given training size is proportional to that class's prevalence rather than equal across classes.

Table 1. Characteristics of the DENTEX validation subset.

Diagnosis class	Annotated teeth	Share (%)
Impacted	40	22.0
Caries	101	55.5
Periapical lesion	9	4.9
Deep caries	32	17.6
Total	182	100.0

Note: The 182 annotated teeth come from 50 panoramic radiographs. Source: Elaborated from the DENTEX dataset (Hugging Face, cc-by-nc-sa-4.0).

Pre-trained versus random features across training sizes

The pre-trained representation outperformed the random representation at almost every one of the six training sizes tested, and the gap between the two curves behaved differently at small and large sample sizes, widening as the sample grew rather than shrinking as the data accumulated (Table 2).

With 5 training teeth the pre-trained features reached 50.2% accuracy compared with 43.3% for the random features, a difference of 6.9 percentage points. By 127 training teeth the pre-trained features reached 67.3% compared with 58.2% for the random features, so the gap widened to 9.1 percentage points at the largest size. The pre-trained advantage was therefore present from the very smallest sample and did not narrow as data accumulated over the six training sizes, which indicates that the benefit of pre-training is not a small-sample crutch that disappears once the target data become adequate but a persistent property of the representation.

The two curves also differed in their trajectories across the six training sizes (Figure 1). The random-feature curve rose to a peak of 59.3% at 40 training teeth and then declined slightly to 58.2% at the full 127 teeth, whereas the pre-trained curve was flat between 20 and 40 teeth at 56.4% and then rose again to 64.7% at 80 teeth and 67.3% at the full sample.

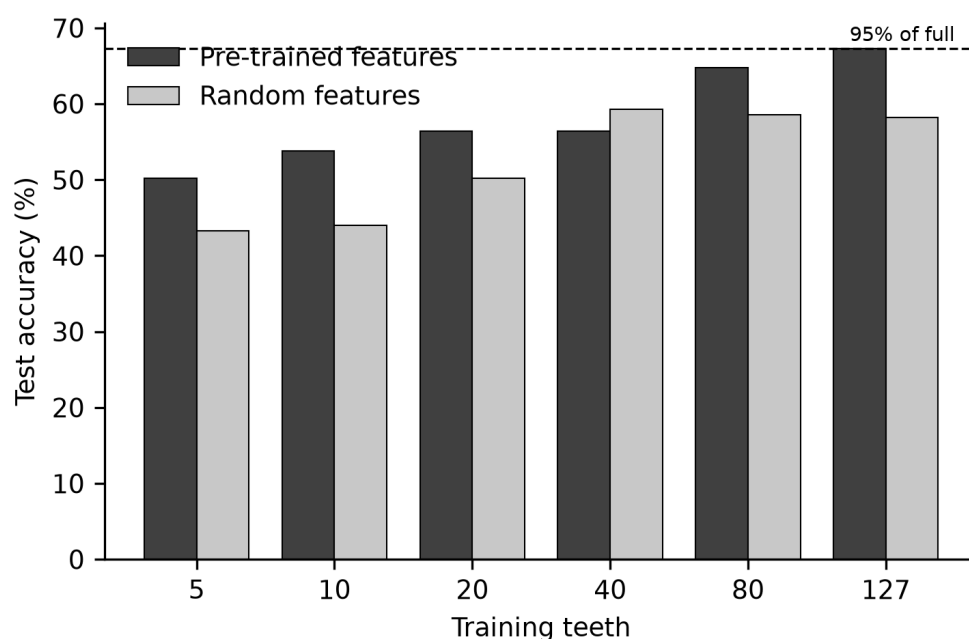
The pre-trained curve therefore showed a two-phase pattern, with an early plateau followed by a later gain, while the random-feature curve rose to its peak and then reversed direction after 40 training teeth, declining slightly at the two largest sizes. The two curves crossed only in the middle of the range: at 40 teeth the random representation briefly matched the pre-trained representation, at 59.3% against 56.4%, before the pre-trained curve pulled ahead again at the larger sizes, so the ranking between the two representations was not monotonic across the full range of training sizes.

The subsample-to-subsample variability was largest at the very smallest training sizes, reaching a standard deviation of 12.9 percentage points for the random representation at 5 teeth, and it narrowed as the sample grew, which is the expected consequence of a small training set whose composition varies substantially from draw to draw.

Table 2. Test accuracy by training size for pre-trained and random features.

Training teeth	Pre-trained features (%)	Random features (%)
5	50.2	43.3
10	53.8	44.0
20	56.4	50.2
40	56.4	59.3
80	64.7	58.5
127	67.3	58.2

Note: Accuracy is mean across five stratified subsamples. The full training split contains 127 teeth. Source: Elaborated from the learning-curve analysis (DENTEX).

Figure 1. Test accuracy as a function of training size for pre-trained and random features.

Note: Bars show the mean accuracy across five stratified subsamples. The dashed line marks 95% of the full-data accuracy of the pre-trained curve. Source: Elaborated from the learning-curve analysis (DENTEX).

Plateau size

The pre-trained curve reached 95% of its full-data accuracy of 67.3%, a threshold of 63.9%, only at 80 training teeth, whereas the random-feature curve reached 95% of its own lower full-data accuracy of just 58.2%, a threshold of only 55.3%, at 40 training teeth. In the pre-trained representation the plateau therefore occurred at 80 teeth, twice the plateau size of the random representation, but the earlier plateau of the random representation was reached at a lower absolute accuracy, so the two plateau sizes are not directly comparable as a measure of data efficiency.

The contrast is more informative when read against the absolute accuracy at each plateau: the random representation saturated early at a modest 58.2% ceiling, whereas the pre-trained representation continued to climb to a 67.3% ceiling, so pre-training traded a later plateau for a higher one.

Per-class behaviour at small sample sizes

At the 20-tooth training size the accuracy of the pre-trained classifier was concentrated almost entirely in the majority class (Table 3). The classifier reached 80.7% accuracy on caries but only 51.7% on impacted teeth, 6.0% on deep caries and 0.0% on periapical lesions, so the apparently reasonable overall accuracy at 20 teeth concealed a near-total failure on the three minority classes.

This per-class pattern reversed the overall impression: whereas the overall curve suggested that 20 teeth captured much of the achievable performance, the per-class breakdown showed

that a 20-tooth sample effectively bought accuracy on caries alone, and that the rarer the class the more completely it remained unlearned at that size.

Table 3. Per-class accuracy at 20 training teeth for pre-trained features.

Diagnosis class	Accuracy at 20 training teeth (%)
Impacted	51.7
Caries	80.7
Periapical lesion	0.0
Deep caries	6.0

Note: Accuracy is mean across five seeds; the periapical class had no correctly classified test tooth in this configuration. Source: Elaborated from the learning-curve analysis (DENTEX).

DISCUSSION

One would expect a transfer-learning classifier to saturate within a handful of examples, on the reasoning that its backbone is already trained and only a small head needs to adapt to the dental domain; the learning curve instead showed that the pre-trained representation did not reach 95% of its full-data accuracy until 80 annotated teeth, and that below 20 teeth the classifier was effectively a caries detector that failed on the other three classes.

The random-feature baseline, by contrast, reached its own plateau earlier and at a lower ceiling, which clarifies what pre-training actually contributes to the learning curve: not an earlier plateau, but a higher and later one that continues to improve as more dental examples arrive.

This reversal of the intuitive expectation is the central contribution of the study, because it means that the convenience of transfer learning should not be mistaken for an exemption from the data-collection problem, and that a research group relying on a pre-trained backbone still faces the unavoidable task of accumulating enough annotated examples to cover the rare classes it cares about before the model can be considered clinically complete.

The contrast between the pre-trained and random representations is the clearest summary of the results. Compared with the random features, the pre-trained features began with a 6.9 percentage-point advantage at 5 teeth and widened that advantage to 9.1 points at 127 teeth, which is consistent with the established finding that ImageNet pre-training provides a useful starting representation for medical tasks but does not by itself teach the domain-specific features that a dental classifier needs.^(6,7)

The widening gap also matches the observation that the benefit of a better or better-matched source model grows with the target dataset size, so that pre-training and annotation effort are complements rather than substitutes.^(9,10) In the random-feature group the curve peaked at 40 teeth and then reversed, a sign that the random representation began to overfit the noise in the larger training sample rather than extracting additional signal, whereas the pre-trained curve, in the same training sizes, continued to rise.

The reversal of the random-feature curve is itself informative: it shows that a non-pre-trained representation can be made worse by more data when the representation is too noisy to benefit from it, a failure mode that the pre-trained curve did not exhibit. This non-monotonic behaviour is a reminder that a learning curve need not be increasing, and that an observed decline with more data is a diagnostic signal that the representation, rather than the amount of data, is the limiting factor, which is exactly the situation that transfer learning is designed to remedy.

The two-phase shape of the pre-trained curve, flat between 20 and 40 teeth and then rising again after 40, is not an artifact but a consequence of the class imbalance documented in the dataset.

The first phase, in which accuracy reached about 56%, is the point at which the classifier has learned the majority class, caries, which alone accounts for more than half the teeth; the second phase, beginning after 40 teeth, is where enough minority-class examples have accumulated for the classifier to begin distinguishing impacted teeth and deep caries, which is why the overall accuracy did not reach 95% of its ceiling until 80 teeth.

This reading is supported directly by the timing: 20 teeth contain roughly 11 caries examples but only about 1 or 2 periapical examples, so a classifier trained on 20 teeth has effectively seen the majority class and little else, and a single periapical example is not enough for a linear classifier to learn a reliable boundary for that class. The per-class results at 20 teeth make this reading concrete,

because the 80.7% accuracy on caries coexisted with 6.0% on deep caries and 0.0% on periapical lesions, so the early plateau was a majority-class plateau rather than a genuine saturation of the task.(3,4) The ordering of the per-class accuracies followed the class prevalence almost exactly, with the most prevalent class the most accurate and the rarest class the least accurate, which is the signature of a classifier whose performance is governed by the number of examples per class rather than by any intrinsic diagnostic difficulty of the classes themselves.

The comparison with the wider medical imaging literature places these results firmly in context. Transfer learning has been reported to require fewer examples than training from scratch for tasks as varied as pneumonia, osteoporosis and kidney classification, and the compendious reviews of transfer learning for cancer diagnosis consistently find a pre-training advantage, but those studies rarely report the full learning curve or a plateau size, so the present result that pre-training delays rather than advances the plateau is a refinement rather than a contradiction.(11-14)

The same pattern recurs in studies that compare transfer-learning pipelines for osteoporosis prediction and in ensemble transfer approaches for Alzheimer diagnosis, which report the accuracy of the final model without tracing how that accuracy accumulates with data.(15,16) Work on effective data selection and on multi-domain meta-datasets has begun to ask the complementary question of which examples matter most, and it suggests that the value of an added label is not uniform, which is consistent with the majority-class concentration observed here and with the idea that active or targeted annotation of the rare classes would bend the curve upward faster than random annotation.(17,18)

The present results therefore connect the sample-size question to the data-selection question: if each added label for a rare class is worth more than a label for the majority class, then the optimal annotation strategy is to oversample the rare classes, and the learning curve provides the evidence for that strategy by showing how little the majority-class labels contribute once the majority class has already saturated.

The present curve therefore sits within an emerging literature that treats both the number and the choice of training examples as first-class design variables rather than as an afterthought of model selection or an unexamined default.

The contrast with the classical sample-size literature is instructive: where a conventional power calculation asks how many subjects are needed to detect an effect at a given significance, a learning-curve analysis asks how many examples are needed to reach a given accuracy, and the two questions give different kinds of answers, the first probabilistic and the second empirical.(3) The learning curve is in this sense the more direct instrument for a builder of classifiers, because it reports the observed relationship between data and performance rather than an assumption about that relationship, and it can be re-estimated whenever the dataset, the set of classes or the training pipeline changes.(4)

The dental-specific literature likewise reports accuracy without reporting how that accuracy depends on the number of annotated teeth, so the present learning curve provides the sample-size evidence that the field has been missing.(19-21)

Dental studies of ankylosis, supernumerary teeth and other uncommon findings are particularly exposed to the sample-size problem, because their target classes are rare and their reported accuracy may rest on a small number of positive examples.(22,23) Comparative evaluations of vision transformers and convolutional networks for dental radiographs similarly focus on architecture rather than on the amount of data, so the present curve complements them by isolating the data dimension.(24)

The finding that a small annotated sample buys majority-class performance while leaving the rare classes essentially unlearned is directly relevant to the many dental studies that train on a few hundred images and report a single overall accuracy, because that single number can conceal the same majority-class concentration observed here.(25,26)

The practical consequence for annotation planning is that the number of images a dental task requires cannot be read from a single target accuracy without knowing the class balance. A group that needs only to flag caries, the majority class, may find that a few dozen labelled teeth suffice, because the majority class saturates early, whereas a group that needs to detect periapical lesions or deep caries must annotate enough images to accumulate rare-class examples, which in this dataset meant approaching the full sample before the overall accuracy approached its ceiling. This distinction between majority-class and rare-class sample requirements is the main actionable message of the curve, and it is the reason the plateau size alone, without the per-class breakdown, would be misleading.(3,4)

The message is amplified by the shape of the curve rather than by any single point on it: a group that stops annotating when the overall accuracy first flattens, at around 20 to 40 teeth, will

believe it has reached the ceiling when it has in fact only finished learning the majority class, and will leave the clinically important rare conditions undetected. A more reliable rule of thumb is to monitor the accuracy on the rarest class that the model must detect and to continue annotating until that class-specific curve, not the overall curve, reaches its own plateau, even though that threshold will typically demand many more labelled examples than the overall accuracy would suggest.

The strengths of this study are the paired design, in which the pre-trained and random features were extracted from the same teeth and passed through the same split, scaler and classifier, and the use of an openly licensed dataset that makes the curve reproducible from the public source alone. The principal limitation is the size of the dataset, because 182 teeth and a 55-tooth test set make the per-class accuracies at small training sizes noisy, and a larger multi-institution collection would trace the minority-class curves with useful precision rather than leaving the periapical class with a zero at the small sizes.

A second limitation is that the analysis used a single backbone and a single downstream classifier, so the exact plateau size of 80 teeth should not be transplanted to a different architecture or a different dental task without re-measuring it; the qualitative two-phase pattern, however, is likely to generalize because it follows from class imbalance rather than from the specifics of ResNet-18.^(1,5) A third limitation is that the random-feature baseline used the same architecture as the pre-trained backbone, and a comparison against a genuinely task-trained-from-scratch network or a foundation model would place the transfer-learning advantage on a fuller scale.^(27,28)

A fourth limitation is that the learning curve was measured with a single fixed test split, so the absolute accuracy values carry the variance of that particular split, even though the comparison between the two feature representations is protected by the paired design. These limitations do not change the central finding, but they bound how far the specific numbers should be generalized, and they point toward a larger, multi-class, multi-institution learning-curve study as the natural extension.

The implication for dental artificial intelligence is that annotation effort should be guided by the class that is hardest to learn rather than by the class that is easiest, and that a learning curve reported with a per-class breakdown is a more honest guide to data requirements than a single accuracy at a single size. Where rare conditions are the clinical target, the curve suggests that annotation should continue well past the point at which overall accuracy appears to plateau, because the overall plateau masks the continued learning of the minority classes, and stopping at that point would leave those classes effectively undetected.

This recommendation holds broadly across health systems, because the class imbalance that drives it is a property of disease prevalence rather than of any particular institution, and a small dental service in a low-resource setting faces the same majority-class concentration when it trains on a modest labelled set.^(29,30)

The portability of the recommendation follows from the fact that the two-phase curve arises from prevalence, which is broadly similar across populations for the common dental conditions studied here, even though the absolute class sizes and the achievable accuracies will differ between settings. The learning curve itself is a portable tool rather than a local one: any dental group can trace its own curve with the same subsampling procedure on its own labelled collection, and the resulting plateau and per-class breakdown will be specific to that group's prevalence and imaging protocol, which is precisely the information that a generic sample-size recommendation cannot provide.

CONCLUSIONS

A transfer-learning dental radiograph classifier with pre-trained ResNet-18 features reached 95% of its full-data accuracy only at 80 annotated teeth, whereas a random-feature baseline plateaued earlier, at 40 teeth, but at a lower accuracy, so pre-training delayed the plateau while raising its ceiling.

The pre-trained representation outperformed the random representation at every training size, with an advantage that widened from 6.9 to 9.1 percentage points as the sample grew, which shows that pre-training and annotation effort are complements rather than substitutes.

At small training sizes the accuracy was concentrated in the majority class, caries, while the minority classes remained largely unlearned, so the number of annotations a dental task requires is governed by the class balance and by the rarest condition that must be detected rather than by the overall accuracy alone, and a learning curve with a per-class breakdown is the appropriate instrument for planning that annotation effort.

REFERENCES

1. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017;42:60-88. doi: 10.1016/j.media.2017.07.005
2. Ibrahim Hamamci et al. DENTEX: an abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays [Internet]. 2023. Available from: <https://huggingface.co/datasets/ibrahimhamamci/DENTEX>
3. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12(1):8. doi: 10.1186/1472-6947-12-8
4. Viering T, Loog M. The shape of learning curves: a review [Internet]. 2023. Available from: <https://arxiv.org/abs/2103.10948>
5. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-778. doi: 10.1109/cvpr.2016.90
6. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? [Internet]. 2014. Available from: <https://arxiv.org/abs/1411.1792>
7. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?. *IEEE Trans Med Imaging*. 2016;35(5):1299-1312. doi: 10.1109/tmi.2016.2535302
8. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging*. 2016;35(5):1285-1298. doi: 10.1109/tmi.2016.2528162
9. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging [Internet]. 2019. Available from: <https://arxiv.org/abs/1902.07208>
10. Kornblith S, Shlens J, Le QV. Do better ImageNet models transfer better? [Internet]. 2019. Available from: <https://arxiv.org/abs/1805.08974>
11. Katreddi S, Midatani A, Roy AP, Velpuri U, Kasani S. Pediatric pneumonia X-ray image classification: predictive model development with DenseNet-169 transfer learning. *J Med Artif Intell*. 2025;8:37-37. doi: 10.21037/jmai-24-356
12. Qureshi MB, Sani M, Raza A, Qureshi MS, Alghamdi W, Akbar S, et al. Deep-learning based osteoporosis classification in knee X-rays using transfer-learning approach. *Sci Rep*. 2025;15(1):38448. doi: 10.1038/s41598-025-24338-4
13. Pimpalkar A, Saini DKJB, Shelke N, Balodi A, Rapate G, Tolani M. Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Sci Rep*. 2025;15(1):10741. doi: 10.1038/s41598-025-94905-2
14. Kaur N, Hans R. Transfer Learning for Cancer Diagnosis in Medical Images: A Compendious Study. *Int J Comput Intell Syst*. 2025;18(1):62. doi: 10.1007/s44196-025-00772-0
15. Kaur N, Khan S, Alazman I, Bin-Asfour M, Alam MN, Balasubramaniam V, et al. Transfer learning based osteoporosis prediction using enhanced medical imaging and fuzzy fusion. *Sci Rep*. 2025;15(1):41960. doi: 10.1038/s41598-025-25946-w
16. Varzaneh ZA, Mousavi SM, Khoshkangini R, Moosavi Khaliji SM. An ensemble model based on transfer learning for the early detection of Alzheimer's disease. *Sci Rep*. 2025;15(1):34634. doi: 10.1038/s41598-025-22025-y
17. Lee H, Kwak JY, Lee E. Effective data selection via deep learning processes and corresponding learning strategies in ultrasound image classification. *Sci Rep*. 2025;15(1):16138. doi: 10.1038/s41598-025-00416-5
18. Woerner S, Jaques A, Baumgartner CF. A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset. *Sci Data*. 2025;12(1):666. doi: 10.1038/s41597-025-04866-4
19. Raeisi Z, Rokhva S, Rahmani F, Goodarzi A, Najafzadeh H. Multi-label diagnosis of dental conditions from panoramic x-rays using attention-enhanced deep learning. *Oral Maxillofac Surg*. 2025;29(1):166. doi: 10.1007/s10006-025-01463-y
20. Pornprasertsuk-Damrongsri S, Vachmanus S, Papasratorn D, Kitisubkanchana J, Chaikantha S, Arayasantiparb R, et al. Clinical application of deep learning for enhanced multistage caries detection in panoramic radiographs. *Sci Rep*. 2025;15(1):33491. doi: 10.1038/s41598-025-16591-4
21. Eftimie S, Ileni T, Iacob L, Hedeşiu M, Dioşan L. Tooth-level detection and mapping of dental pathologies on panoramic radiographs using YOLOv11 and RT-DETR. *MethodsX*. 2025;15:103696. doi: 10.1016/j.mex.2025.103696
22. N. Y. Artificial intelligence for detecting dental ankylosis in primary molars using panoramic radiographs—a retrospective study. *JOC PD*. 2025;49(6):120. doi: 10.22514/jocpd.2025.133
23. Uzel İ, Ghabchi B, Çoğulu D. Deep learning-based automated detection of supernumerary teeth in pediatric panoramic radiographs. *PLoS One*. 2025;20(11):e0335845. doi: 10.1371/journal.pone.0335845
24. Felek T, Tercanlı H, Gök RŞ. Evaluating vision transformers and convolutional neural networks in the context of dental image processing: a systematic review. *BMC Oral Health*. 2025;25(1):1626. doi: 10.1186/s12903-025-07036-5
25. Sukegawa S, Yoshii K, Hara T, Tanaka F, Taki Y, Inoue Y, et al. Optimizing dental implant identification using deep learning leveraging artificial data. *Sci Rep*. 2025;15(1):3724. doi: 10.1038/s41598-025-87579-3
26. Pishghadam N, Esmaeilifard R, Paknahad M. Explainable deep learning for age and gender estimation in dental CBCT scans using attention mechanisms and multi task learning. *Sci Rep*. 2025;15(1):18070. doi: 10.1038/s41598-025-03305-z

27. Tavakoli N, Shakeri Z, Gowda V, Samsel K, Bedayat A, Ghasemiesfe A, et al. Generative AI and Foundation Models in Radiology: Applications, Opportunities, and Potential Challenges. *Radiology*. 2025;317(2):e242961. doi: 10.1148/radiol.242961
28. Wang Z, Sun Q, Zhang B, Wang P, Zhang J, Zhang Q. PM2: A new prompting multi-modal model paradigm for few-shot medical image classification. *Computer Methods and Programs in Biomedicine*. 2025;272:109042. doi: 10.1016/j.cmpb.2025.109042
29. Benakatti V, Nayakar RP, Anandhalli M, Sukhasare RC. Efficacy of deep learning models and dental professionals in identifying dental implants. *Imaging Sci Dent*. 2025;55(4):351. doi: 10.5624/isd.20250048
30. Achararit P, Manaspon C, Jongwannasiri C, Kulthanaamondhita P, Itthichaisri C, Chantarangsu S, et al. Impacted lower third molar classification and difficulty index assessment: comparisons among dental students, general practitioners and deep learning model assistance. *BMC Oral Health*. 2025;25(1):152. doi: 10.1186/s12903-025-05425-4

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Verónica Karina Arias Murillo, Melissa Castillo Jimenez.

Writing–review and editing: Verónica Karina Arias Murillo, Melissa Castillo Jimenez.