



Robustness of a deep-learning dental radiograph classifier to compression and resolution loss for teledentistry

Robustez de un clasificador de aprendizaje profundo de radiografías dentales ante la compresión y la pérdida de resolución en teleodontología

Nairobi Hernández Bridon¹  

¹ Instituto de Investigaciones en Microbiología y Parasitología Médica. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Universidad de Buenos Aires. Ciudad Autónoma de Buenos Aires, Argentina

Submitted: 26-01-2025 | Revised: 03-04-2025 | Accepted: 15-06-2025 | Published: 16-06-2025

Corresponding Author: Nairobi Hernández Bridon 

ABSTRACT

Introduction: Teledentistry transmits radiographs through channels that compress and downsample images, but classifiers are almost always evaluated on uncompressed images. The gap between benchmark conditions and real transmission is rarely quantified. **Methods:** A tooth-level diagnostic classifier was built with transfer learning on 182 annotated teeth (four diagnosis classes) from 46 panoramic radiographs of the open DENTEX dataset, then re-evaluated after degrading the source images with JPEG compression (quality 75 to 10), downsampling (by 1/2 to 1/8) and Gaussian noise (sigma 10 to 50). **Results:** The classifier reached 67.30% accuracy on uncompressed images. Mild degradation was benign, and downsampling by half slightly improved accuracy to 72.70%. Severe degradation reduced accuracy, reaching 52.70% under JPEG quality 10, 54.50% under downsampling by 1/8, and 54.50% under the highest noise, corresponding to absolute losses of 14.50, 12.70 and 12.70 percentage points, respectively. **Conclusions:** A dental radiograph classifier tolerates the mild degradation typical of routine transmission but loses substantial accuracy under heavy compression, strong downsampling, and high noise. Transmission settings should therefore be treated as part of model specification rather than an afterthought.

Keywords: deep learning; radiograph; image compression; teledentistry; robustness.

RESUMEN

Introducción: La teleodontología transmite radiografías por canales que comprimen y reducen la resolución de las imágenes, pero los clasificadores casi siempre se evalúan sobre imágenes sin comprimir. La brecha entre las condiciones de referencia y la transmisión real rara vez se cuantifica. **Métodos:** Se construyó un clasificador diagnóstico a nivel de diente mediante aprendizaje por transferencia sobre 182 dientes anotados (cuatro clases diagnósticas) de 46 radiografías panorámicas del conjunto abierto DENTEX, y se reevaluó tras degradar las imágenes de origen con compresión JPEG (calidad 75 a 10), reducción de resolución (1/2 a 1/8) y ruido gaussiano (sigma 10 a 50). **Resultados:** El clasificador alcanzó 67,30% de exactitud en imágenes sin comprimir. La degradación leve fue benigna, y la reducción a la mitad elevó ligeramente la exactitud a 72,70%. La degradación severa redujo la exactitud a 52,70% con calidad JPEG 10, 54,50% con reducción 1/8 y 54,50% con el ruido máximo, pérdidas absolutas de 14,50, 12,70 y 12,70 puntos porcentuales, respectivamente. **Conclusiones:** Un clasificador de radiografías dentales tolera la degradación leve propia de la transmisión habitual, pero pierde exactitud sustancial bajo compresión intensa, reducción fuerte y ruido alto. Por tanto, los ajustes de transmisión deben tratarse como parte de la especificación del modelo y no como algo secundario.

Palabras clave: aprendizaje profundo; radiografía; compresión de imagen; teleodontología; robustez.

INTRODUCTION

Teledentistry has grown from a niche service into a routine channel for triage, consultation, and caries screening, and its workflow depends on sending images through networks that rarely preserve them unchanged. Smartphone photographs and radiographs are resized for upload, compressed to keep file sizes manageable, and transmitted over connections that add noise, so the image a remote clinician or algorithm receives is not the image that was captured.(1,2) The motivation for this growth is access: a store-and-forward model lets a mid-level provider or a patient capture images that a dentist reviews asynchronously, removing the need for a synchronous visit. (1)Diagnostic accuracy in this setting has been studied for human readers, who remain accurate across the mild degradation typical of store-and-forward systems, whether the images are taken by trained assistants or by patients themselves.(3,4) A systematic review of in-vivo teledentistry studies concluded that extraoral photography performs comparably to visual examination for caries detection, even when the images are compressed by consumer devices.(2)The same question has barely been asked of the deep-learning classifiers that are now proposed as automated readers for exactly these workflows, and that is the gap this study addresses.

Deep learning has become the dominant approach to automated analysis of dental radiographs, in line with its broader rise across medical imaging.(5,6) Convolutional networks now detect caries, number teeth, and classify dental pathology on panoramic images, and the field has converged on a shared recipe: a pretrained network is fine-tuned or used as a feature extractor on cropped, normalized, uncompressed images.(7-9) This recipe has produced strong benchmark numbers across cancer diagnosis and other imaging tasks, but it also bakes in an assumption that the evaluation image resembles the training image.(10) That assumption is precisely what a teledentistry deployment violates, because the deployment and the benchmark differ not in the task but in the input distribution, and a model is only as good as its weakest input.

The asymmetry is the point of departure for this study. Deep-learning models for dental radiographs are trained and evaluated on uncompressed, high-resolution images, and their reported accuracy reflects those ideal conditions.(7,8) A deployed model, by contrast, receives whatever survives transmission. Where the two diverge, reported performance overstates real performance, and the size of that overstatement is unknown for dental imaging. The question is not whether degradation matters but how much it matters, and which forms of degradation are most damaging, because that determines whether a teledentistry pipeline needs dedicated image handling. A model that is robust to realistic transmission can be deployed directly on the images a channel actually delivers, whereas a fragile one forces the pipeline to add preprocessing, file-size limits, or quality gates that add cost and points of failure. In a field that is moving quickly toward deployment, measuring this gap before deployment is cheaper than discovering it after a model has been trusted with real referrals.

Evidence from other imaging domains suggests the answer is not trivial. Studies of mammogram classification found that performance was largely preserved across JPEG quality levels until quality became very low,(11) and work on histopathology reported that JPEG 2000 compression could be applied aggressively before convolutional networks degraded.(12) Resolution, by contrast, is a stronger lever: endoscopy classifiers fell as image resolution was reduced, with the largest losses at the smallest sizes.(13) Deep networks are also known to be unstable under perturbations that are imperceptible or visually minor, which cautions against assuming that a small change in input leaves the output unchanged.(14) Together these studies establish that the answer to the robustness question depends on the modality, the task, and the specific degradation, which is why a dental-specific measurement is needed rather than an extrapolation from mammography or endoscopy. Dental panoramics are unusually large and high-resolution, and the clinically relevant findings are small and sparse within them, so the effect of losing either detail or pixels may differ from the other domains in either direction; a conclusion reached on breast or bowel images cannot simply be assumed to hold for teeth.

The contrast this study sets up is between a classifier's reference performance and its performance under the degradations that teledentistry actually imposes. We built a tooth-level diagnostic classifier on the open DENTEX dataset using transfer learning, then re-evaluated it after degrading the source images with three manipulations chosen to mirror transmission: JPEG compression, downsampling, and additive Gaussian noise, each across a ladder of severity spanning the range from what a well-configured system produces to what a degraded one can. The aim was to quantify the accuracy loss associated with each degradation and to identify the settings at which a teledentistry deployment would need to intervene. By measuring the same model under every condition, the design isolates the effect of the input itself, which is the quantity a deployment can

control and therefore the quantity worth measuring, and it avoids confounding the measurement with differences in model training or in the choice of test set.

METHODS

Dataset and task

The analysis used the DENTEX dataset, an open benchmark of panoramic dental radiographs released with a Creative Commons license and annotated with the FDI tooth numbering system and per-tooth diagnoses.⁽¹⁵⁾ The reuse of openly licensed dental imaging is itself a growing practice, and systematic reviews have catalogued the characteristics and licensing terms of the available oral and maxillofacial datasets, of which DENTEX is one.⁽¹⁶⁾ The validation split, which is the only split distributed with a diagnosis-level annotation file, contains 46 panoramic radiographs and 182 annotated abnormal teeth across four diagnosis classes: impacted tooth, caries, periapical lesion, and deep caries. Each annotation provides the bounding box of the affected tooth and its diagnosis category. The classification task was to assign each tooth region to one of the four diagnosis classes, a task that mirrors the abnormality classification performed in automated dental screening.^(7,17) The dataset was selected because its license permits reuse, because its annotations are publicly available in a machine-readable format without any login or request procedure, and because panoramic radiographs are the modality most commonly transmitted in teledentistry consultations, which makes this particular dataset a natural substrate for a transmission-robustness study.

Tooth crops and feature extraction

For every annotated tooth, the rectangular region was cropped from the source panoramic image and resized to 224 by 224 pixels with three color channels, with the grayscale radiograph replicated across the three channels to match the input format of the pretrained network. A ResNet-18 convolutional network pretrained on ImageNet was used as a fixed feature extractor: the final classification layer was removed, and the 512-dimensional representation after global average pooling was computed for each crop.^(18,19) This transfer-learning strategy is standard for small medical datasets and keeps the analysis feasible on central-processing-unit hardware, which is itself a constraint that many dental AI studies share. Feature extraction was performed in inference mode, so the representation of a given image is deterministic and identical across the reference and degraded conditions. The same crop-and-extract procedure was reused for every condition, so any difference in the resulting features is attributable to the degradation of the source image and not to variation in the processing, which is essential for a clean comparison.

Classifier

A multinomial logistic regression with balanced class weights was fitted to the extracted features. The 182 tooth crops were divided into a training set of 127 and a held-out test set of 55 with a fixed random seed and stratification by diagnosis class, so that each class was represented in both partitions in proportion to its prevalence. The classifier was trained once on the uncompressed training crops and held fixed for all subsequent evaluations, so that every degradation was assessed against the same model and the only thing that changed between conditions was the input image. Balanced class weights were used because the diagnosis classes are unevenly represented, and an unweighted model would favor the majority class at the expense of the clinically important minority classes.

Degradation protocols

Three degradation types were applied to the full-resolution source panoramics before cropping, so that each degraded tooth crop experienced the same pipeline a real transmitted image would, rather than degrading the crop after it was extracted. This ordering matters because a real transmission degrades the whole image, and the crop is then taken from the degraded result, which is the situation a deployed model faces; degrading the crop in isolation would not reproduce the interaction between compression and the surrounding image context, and it would ignore the fact that a downsampled image delivers fewer pixels for every tooth at once. JPEG compression saved each panoramic as a JPEG file at quality levels 75, 50, 25, and 10 and reloaded it, mimicking the lossy compression applied by messaging and storage services. Downsampling reduced the linear dimensions of each panoramic by factors of 2, 4, and 8 with bilinear interpolation, and the anno-

tated bounding boxes were scaled by the same factor before cropping, mimicking a pipeline that transmits a reduced-resolution image and therefore delivers fewer pixels for every tooth. Additive Gaussian noise with standard deviations of 10, 25, and 50 intensity levels was added independently to each pixel and the result clipped to the valid range, mimicking sensor and channel noise. These levels span the range from imperceptible to severe for grayscale radiographs, and they were chosen to bracket the conditions that a real deployment might encounter rather than to represent any single device; the coarsest settings are intentionally worse than a well-configured system would produce, so that the analysis maps the boundary at which degradation begins to matter.

Evaluation

Classification accuracy and macro-averaged F1 score were computed on the 55 held-out teeth under the reference condition and under each of the ten degraded conditions. The reference condition used the original uncompressed crops. No model was retrained on degraded images; the question was how the model trained on clean images behaves when its input degrades, because retraining on degraded images would be a different, more optimistic question about whether robustness can be learned rather than whether it holds out of the box. Because the classes were imbalanced, accuracy was reported together with macro F1, which weights every class equally and therefore reveals whether rare classes are being lost even when aggregate accuracy looks acceptable. The two metrics are complementary: accuracy answers whether the classifier is right on average, and macro F1 answers whether it remains right for the rare diagnoses that a screening tool exists to catch.

Reproducibility

All degradations used a fixed random seed, and the feature extractor was run in inference mode with deterministic operations, so re-running the analysis produces identical numbers from one execution to the next. The analysis was implemented in Python with PyTorch, torchvision, scikit-learn, NumPy, and Pillow, and every number reported here is generated by the analysis script rather than transcribed by hand, so the results can be recomputed exactly from the public dataset and its labels. The DENTEX images and annotation file are downloaded programmatically and require no login, which makes the entire pipeline, from data acquisition to final figure, reproducible by any reader with an internet connection.

RESULTS

Task and reference performance

The 182 annotated teeth were distributed across the four classes unevenly, with caries the most common and periapical lesion the least (Table 1). This imbalance reflects the epidemiology of the underlying conditions and is typical of dental radiograph datasets rather than an artifact of this sample, and it is the reason the evaluation relies on macro F1 alongside accuracy rather than on accuracy alone. The transfer-learning classifier achieved 67.30% accuracy and a macro F1 of 56% on the uncompressed reference images, a level consistent with a four-class task on a small, heterogeneous dataset. The gap between accuracy and macro F1 already indicates that the rare classes were classified less reliably than the majority class even before any degradation, and that gap becomes the baseline against which each degradation's additional harm is measured. All subsequent comparisons are therefore made against this same reference condition, so the reported losses are attributable to the degradations alone.

Effect of JPEG compression

JPEG compression degraded accuracy only weakly at moderate quality and more sharply at the lowest quality (Table 2, Figure 1). Accuracy was 65.50% at quality 75 and 52.70% at quality 10, a loss of 14.50 percentage points from the reference. The step from quality 25 to quality 10 was the steepest, whereas the differences among quality 75, 50, and 25 were within the noise of the held-out set, so for this task the practical tolerance of JPEG compression extends down to roughly quality 50 before meaningful losses appear. Macro F1 fell to its minimum at quality 10, indicating that the minority classes were disproportionately affected by heavy compression. An example crop shows that the diagnostic appearance of a lesion remains legible at quality 25 but becomes blocky at quality 10, which is consistent with the measured fall (Figure 2). The block artifacts introduced

at very low quality corrupt the same high-frequency edges that distinguish a small carious lesion from normal enamel, which is why the rare classes suffer first.

Table 1. Distribution of annotated tooth regions by diagnosis class.

Diagnosis	Tooth regions
Impacted tooth	40
Caries	101
Periapical lesion	9
Deep caries	32

Note: Each tooth region corresponds to one annotated abnormal tooth with a diagnosis label.

Table 2. Classifier accuracy under JPEG compression.

Condition	Accuracy (%)	Macro F1 (%)
Reference	67.3	56.0
JPEG quality 75	65.5	51.6
JPEG quality 50	61.8	51.3
JPEG quality 25	63.6	53.8
JPEG quality 10	52.7	31.5

Note: Reference accuracy on uncompressed images was 67.3%.

Effect of downsampling

Downsampling behaved differently from compression. Reducing resolution by half did not harm the classifier and slightly improved accuracy to 72.70%, consistent with mild smoothing removing high-frequency detail that the network did not use (Table 3, Figure 1). Resolution reduction by a quarter restored accuracy to the reference level, and reduction by an eighth produced the largest loss observed for downsampling, 54.50% accuracy, a drop of 12.70 percentage points. Macro F1 fell steeply at the coarsest resolution, indicating that rare classes became essentially undetectable. The contrast between the benign one-half reduction and the damaging one-eighth reduction is the clearest single result of the study: downsampling is not harmful until a resolution threshold is crossed, and that threshold lies somewhere between a quarter and an eighth of the original. For a panoramic image that starts at several thousand pixels across, this corresponds to a transmitted width of a few hundred pixels, which is well within the range of what aggressive smartphone recompression can produce.

Effect of additive noise

Table 3. Classifier accuracy under downsampling and additive noise.

Condition	Accuracy (%)	Macro F1 (%)
Downsample 1/2	72.7	60.0
Downsample 1/4	67.3	41.8
Downsample 1/8	54.5	17.6
Noise sigma 10	63.6	56.7
Noise sigma 25	61.8	38.9
Noise sigma 50	54.5	17.9

Note: Reference accuracy on uncompressed images was 67.3%.

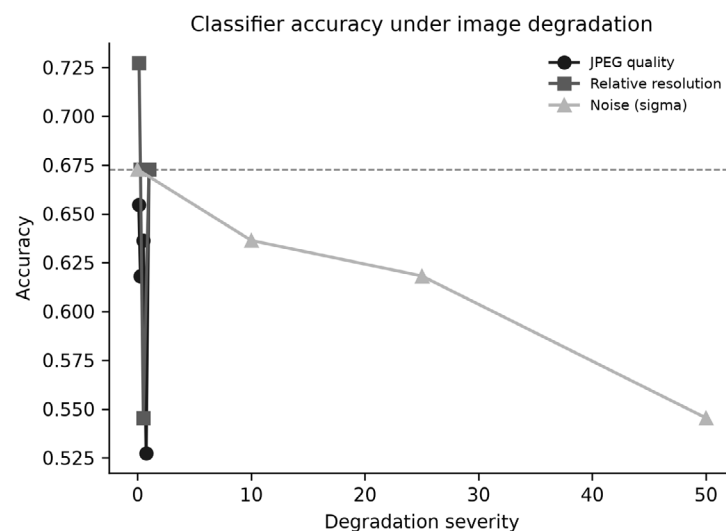
Additive Gaussian noise produced a monotonic decline in accuracy from 63.60% at the lowest noise level to 54.50% at the highest (Table 3, Figure 1). The loss at the highest noise, 12.70 percentage points, was comparable to the losses under the most severe compression and downsampling, so

the three degradations are roughly equally damaging at their extremes even though their thresholds differ. Macro F1 was more sensitive than accuracy, halving at the highest noise even while accuracy remained in the mid-fifties, which shows that noise erodes the rare classes before it erodes overall accuracy. Unlike compression and downsampling, noise has no benign range: even the lowest noise level tested cost a small amount of accuracy, which reflects that the network never trained on noisy inputs and therefore has no built-in tolerance for them. This asymmetry is worth emphasizing, because it means noise cannot be treated with the same relaxed attitude that the other two degradations permit.

Contrast across degradations

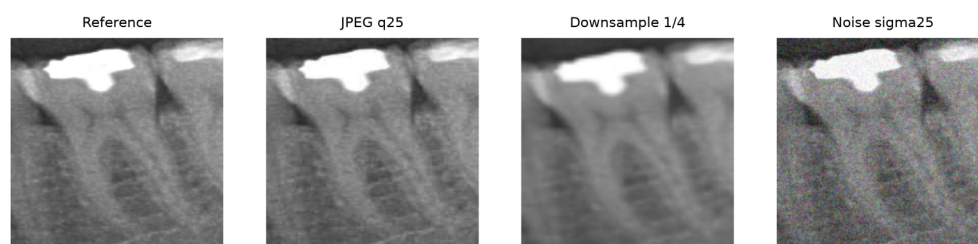
Comparing the three types of degradation directly, the classifier tolerated mild perturbation of every kind but failed in different ways as severity increased (Figure 1). Mild downsampling was the only manipulation that helped rather than hurt. At the severe end, all three degradations converged on roughly the same accuracy, but the macro F1 revealed that downsampling and noise collapsed the minority classes almost entirely, whereas heavy compression retained more of their signal. The three degradations therefore produce the same headline accuracy loss through different mechanisms: compression corrupts fine edges, downsampling removes pixels outright, and noise overwrites intensity values. The confusion matrices at reference and under the most severe JPEG compression show that misclassification concentrated at the boundaries between visually similar classes, and that the periapical-lesion class in particular lost most of its true positives (Figure 3). The per-class view confirms this pattern: the rare periapical-lesion class was the first to be lost under every degradation, which is the single most consequential observation of the study for clinical use, because periapical lesions and deep caries are the findings for which a screening referral matters most (Figure 4).

Figure 1. Classifier accuracy as a function of degradation severity.

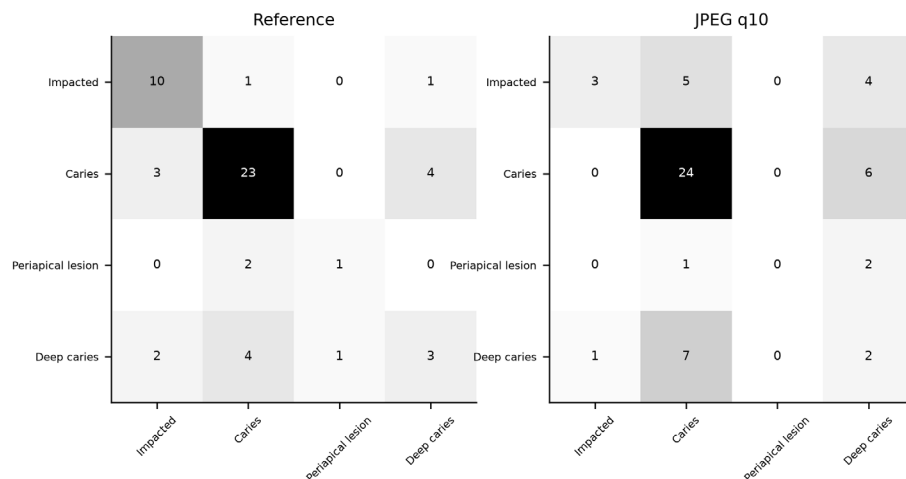


Note: The dashed line marks the reference accuracy on uncompressed images.

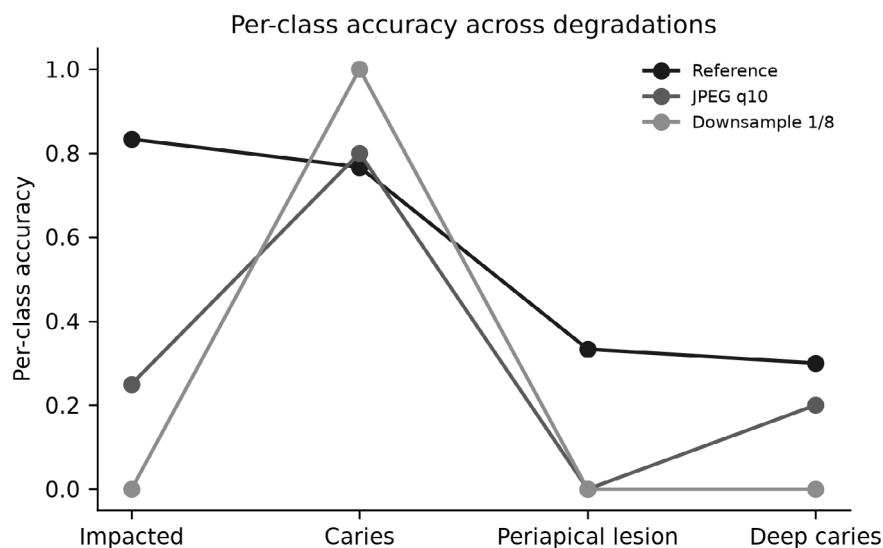
Figure 2. Example tooth crop under each degradation.



Source: DENTEX validation split.

Figure 3. Confusion matrices at reference and under the most severe JPEG compression.

Source: Held-out evaluation of 55 tooth regions.

Figure 4. Per-class accuracy at reference and under severe degradations.

Source: Held-out evaluation of 55 tooth regions.

DISCUSSION

The main finding is that a dental radiograph classifier degrades little under the mild image manipulation typical of routine transmission, but loses on the order of 14.50 to 12.70 percentage points of accuracy under the severe manipulation that a poorly configured or bandwidth-limited teledentistry pipeline can introduce. The contrast between the three degradation types is as informative as the overall loss: mild downsampling was benign or mildly helpful, whereas heavy compression, coarse downsampling, and strong noise each cut accuracy by a comparable amount, and the rare classes were the first to disappear. The practical reading is that transmission settings are not a neutral detail but a part of the model's specification. A benchmark that reports only clean-image accuracy is therefore answering a different question from the one a deployment actually asks, and the difference, measured here in percentage points, is large enough to matter for a screening decision, particularly for the minority diagnoses that a screening tool exists to catch.

The finding that mild downsampling did not hurt, and slightly helped, echoes the observation

in endoscopy that small reductions in resolution cost little and can even regularize the network by removing high-frequency noise.(13) It also aligns with the principle that dental findings of interest, such as a large carious lesion or an impacted tooth, occupy low-frequency structure that survives modest resizing, whereas the fine texture that downsampling removes is less diagnostic. Conversely, super-resolution methods exist precisely to recover detail lost to downsampling, which indicates how much information the coarsest levels discard, and such methods have become a standard tool in image restoration precisely because naive upsampling cannot reconstruct what was never transmitted.(20) The benign effect of mild downsampling also has a practical corollary: a pipeline can downsample to reduce file size and transmission cost at essentially no accuracy penalty, provided it does not cross the resolution threshold that the one-eighth reduction represents. This is the contrastive point of the study: robustness is not uniform across degradations, and the threshold at which performance falls depends on what kind of information the task requires.

The compression results are broadly consistent with the imaging literature. Mammogram classifiers held their accuracy across quality levels before failing at very low JPEG quality,(11) and histopathology networks tolerated aggressive JPEG 2000 compression.(12) Work that classifies directly in the JPEG-compressed domain has shown that the compression representation itself carries much of the signal the network needs, which helps explain why moderate quality loss is inexpensive,(21) and the same pattern has been reported for physiological signals.(22) Radiology has long treated moderate irreversible compression as diagnostically safe for human readers, with guidelines permitting substantial compression ratios, so the human and machine evidence have converged on the same practical tolerance.(23,24) Repeated JPEG compression introduces its own artifacts that can confound analysis, a caution relevant to any pipeline that re-encodes images at multiple stages, which is exactly what a store-and-forward system does when it resizes at capture and again at the server.(25) The dental evidence for human readers points the same way, with linear measurements on compressed panoramic radiographs remaining accurate(26) and early digital caries diagnosis surviving JPEG compression.(27) The present results extend that consensus to a deep-learning reader, with the caveat that the network, unlike a human, showed no benefit from any compression level and began to fail only at the extreme, so the safety margin enjoyed by human readers should not be assumed for automated ones.

The noise results are the most cautionary, because noise is the least controlled of the three in real transmission and because macro F1 collapsed while accuracy remained superficially acceptable. A radiograph transmitted through a noisy channel degrades the minority classes, periapical lesion and deep caries, before the majority class, caries, is affected, so a deployment that monitors only aggregate accuracy would miss the erosion of exactly the clinically important findings it is meant to flag. This is a known failure mode of deep classifiers under perturbation,(14) and it motivates the substantial literature on noise-robust and denoising networks that exists precisely because clean-training, noisy-testing degrades performance.(28,29) The practical lesson is that a teledentistry system should treat noise as a first-class concern, because it is the degradation most likely to be invisible at the point of review and the one that most silently undermines the rare and important diagnoses. A clinician receiving a noisy image can often still read it, and may not realize that the automated reader is failing on the same image, which is a subtle but real risk of pairing human and machine readers on a degraded channel. Reporting per-class performance, and not only aggregate accuracy, is the direct way to make this failure visible in a robustness evaluation, and it aligns with the broader emphasis on interpretability and trustworthy reporting in medical imaging.(30)

Several limitations bound these conclusions. The dataset is a single validation split of 46 images and 182 teeth, and the held-out set of 55 teeth is small, which makes the per-condition estimates noisy and explains the non-monotonic compression curve, where quality 25 appeared slightly better than quality 50. The class distribution is imbalanced, with only a handful of periapical-lesion teeth, so the per-class results for rare classes are indicative rather than precise, and the periapical-lesion class in particular should be read as suggestive. The classifier was a fixed-feature logistic regression rather than a fine-tuned network, which is a reasonable but conservative proxy for the stronger models used in production; a fine-tuned network might behave differently under degradation, and the direction of that difference is not obvious a priori, though a more powerful model could be either more or less sensitive to input degradation. The degradations were applied in isolation, whereas real transmission stacks compression, resizing, and noise together, and the combined effect may be larger than any single degradation suggests, so the single-degradation losses reported here are likely lower bounds on the true combined loss. Finally, the analysis used a single preprocessing pipeline, and results could shift with a different crop size, normalization, or feature extractor, even though the qualitative contrast between degradation types is unlikely to

reverse.

The implications for system design follow directly from the observed thresholds. A teledentistry pipeline that transmits at JPEG quality 50 or higher and at half resolution or better can expect its classifier to perform near its benchmark, which removes the need for special handling of those conditions. A pipeline that must operate at lower quality or resolution, whether because of bandwidth limits or because a smartphone downsamples before upload, should either revalidate its model at those settings or add a quality gate that rejects images below the validated threshold. The contrast between degradation types is operationally useful here: downsampling is the cheapest lever to relax, because mild downsampling is free, whereas noise is the most dangerous, because it has no safe range and because the channels that introduce it are the least controllable. System designers can therefore prioritize their safeguards where they matter most, and a transmission policy can be expressed as a simple pair of thresholds, a minimum quality and a maximum downsample factor, below which the classifier should not be trusted without revalidation.

There is also a methodological lesson in the design itself. The experiment separated the question of whether a model is robust out of the box from the separate question of whether robustness can be trained in, by holding the classifier fixed and varying only the input. This clean-training, degraded-testing setup is the harshest and most realistic framing for a deployed model, and it is the framing that a robustness evaluation should use when the goal is to predict real-world performance rather than to demonstrate that augmentation can improve it. The two questions are related but not interchangeable, and conflating them overstates how safe a model is in practice.

The degradations studied here are natural, in the sense that they arise from ordinary transmission rather than from a deliberate attack, and that distinguishes them from the adversarial perturbations that occupy much of the robustness literature.⁽¹⁴⁾ Natural degradations are the more relevant threat model for teledentistry, because no one is trying to fool a caries classifier, but everyone is subject to compression and bandwidth limits. The two threat models require different defenses, and a model hardened against adversarial attack is not automatically hardened against compression, so the present results are not a substitute for adversarial testing, only a complement to it.

These limitations define the next steps. A larger, multi-institution dataset would stabilize the per-condition estimates and permit per-class conclusions, and would test whether the thresholds observed here are specific to one collection or general. Evaluating a fine-tuned network would test whether the robustness pattern holds for the models actually deployed, and whether a stronger base model is more or less fragile under transmission. Applying degradations in combination would measure the realistic, worst-case loss that a pipeline can impose, and would provide the bound that a deployment actually needs to plan around. Robustness testing could also be folded into the reporting of new dental AI models, so that a paper states not only the accuracy on clean images but the accuracy at the compression and resolution settings its intended deployment will use, which is a modest and inexpensive addition that would materially improve the trustworthiness of the field. And the central recommendation follows directly from the contrast: teledentistry deployments should specify, and then validate against, the exact compression and resolution settings of their transmission channel, rather than assuming that a classifier benchmarked on clean images will perform identically on the images that actually arrive. This is a form of validation that costs little, requires no new data, and can be performed with the same public datasets used to train the model in the first place, which makes it a realistic expectation rather than an idealistic one. Given the demonstrated growth of dental AI and the licensing and quality issues that accompany the field's datasets, robustness testing against realistic transmission conditions is a natural addition to the reporting of any new model.⁽¹⁶⁾

CONCLUSIONS

A transfer-learning classifier for tooth-level diagnosis on dental panoramics tolerated mild JPEG compression, mild downsampling, and low noise with little loss, and mild downsampling slightly improved accuracy, but heavy compression, coarse downsampling, and very strong noise each cut accuracy by more than a tenth and largely collapsed the minority classes. Robustness was therefore not at all uniform across degradation types or severity, and the rare but clinically important diagnoses, periapical lesion and deep caries, were the first to be lost under every degradation. Transmission settings should be treated as part of the specification of any teledentistry classifier, robustness evaluations should report per-class performance rather than aggregate accuracy alone, and the thresholds measured here provide a concrete basis for setting those transmission limits.

REFERENCES

1. Estai M, Kanagasingam Y, Huang B, Shiikha J, Kruger E, Bunt S, et al. Comparison of a Smartphone-Based Photographic Method with Face-to-Face Caries Assessment: A Mobile Teledentistry Model. *Telemedicine and e-Health*. 2017;23(5):435-440. doi: 10.1089/tmj.2016.0122
2. Kargozar S, Jadidfard MP. Teledentistry accuracy for caries diagnosis: a systematic review of in-vivo studies using extra-oral photography methods. *BMC Oral Health*. 2024;24(1):828. doi: 10.1186/s12903-024-04564-4
3. Iranparvar P, Ghorbani Z, Zandieh S, Hartoonian S, Shirazi M. Evaluating Maternal Accuracy in Smartphone-Based Tele-Dentistry for Early Childhood Caries Detection. *Clinical & Exp Dental Res*. 2025;11(5):e70231. doi: 10.1002/cre2.70231
4. Lamas-Lara VF, Mattos-Vela MA, Evaristo-Chiyong TA, Guerrero ME, Jiménez-Yano JF, Gómez-Meza DN. Validity and reliability of a smartphone-based photographic method for detection of dental caries in adults for use in teledentistry. *Front Oral Health*. 2025;6:1470706. doi: 10.3389/froh.2025.1470706
5. Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. *Int J Multimed Info Retr*. 2022;11(1):19-38. doi: 10.1007/s13735-021-00218-1
6. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. *Evol Intel*. 2022;15(1):1-22. doi: 10.1007/s12065-020-00540-3
7. 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
8. Zancan BG, Carneiro JA, Martins CU, Tirapelli C, Capel CP, Dantas da Costa E, et al. AI-powered precision in dental radiographic analysis using tailored CNNs for tooth numbering and cavity detection. *PLOS Digit Health*. 2025;4(11):e0001074. doi: 10.1371/journal.pdig.0001074
9. Krichen M. Convolutional Neural Networks: A Survey. *Computers*. 2023;12(8):151. doi: 10.3390/computers12080151
10. Jiang X, Hu Z, Wang S, Zhang Y. Deep Learning for Medical Image-Based Cancer Diagnosis. *Cancers*. 2023;15(14):3608. doi: 10.3390/cancers15143608
11. Jo YY, Choi YS, Park HW, Lee JH, Jung H, Kim HE, et al. Impact of image compression on deep learning-based mammogram classification. *Sci Rep*. 2021;11(1):7924. doi: 10.1038/s41598-021-86726-w
12. Ghazvinian Zanjani F, Zinger S, Piepers B, Mahmoudpour S, Schelkens P, de With PHN. Impact of JPEG 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images. *J Med Imag*. 2019;6(02):1-9. doi: 10.1117/1.jmi.6.2.027501
13. Thambawita V, Strümke I, Hicks SA, Halvorsen P, Parasa S, Riegler MA. Impact of Image Resolution on Deep Learning Performance in Endoscopy Image Classification: An Experimental Study Using a Large Dataset of Endoscopic Images. *Diagnostics*. 2021;11(12):2183. doi: 10.3390/diagnostics11122183
14. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc Natl Acad Sci USA*. 2020;117(48):30088-30095. doi: 10.1073/pnas.1907377117
15. van Nistelrooij N, Ghoul KE, Xi T, Saha A, Kempers S, Cenci M, et al. Combining public datasets for automated tooth assessment in panoramic radiographs. *BMC Oral Health*. 2024;24(1):387. doi: 10.1186/s12903-024-04129-5
16. Hao J, Nalley A, Yeung AWK, Tanaka R, Ai QYH, Lam WYH, et al. Characteristics, licensing, and ethical considerations of openly accessible oral-maxillofacial imaging datasets: a systematic review. *npj Digit Med*. 2025;8(1):412. doi: 10.1038/s41746-025-01818-5
17. Adnan A, Rizwan MT, Attaullah HM, Basheer S, Quasim MT. Deep Architectural Classification of Dental Pathologies Using Orthopantomogram Imaging. *CMC*. 2025;85(3):5073-5091. doi: 10.32604/cmc.2025.068797
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84-90. doi: 10.1145/3065386
19. Shafiq M, Gu Z. Deep Residual Learning for Image Recognition: A Survey. *Applied Sciences*. 2022;12(18):8972. doi: 10.3390/app12188972
20. Dong C, Loy CC, He K, Tang X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans Pattern Anal Mach Intell*. 2016;38(2):295-307. doi: 10.1109/tpami.2015.2439281
21. Dong Y, Pan WD. Image Classification in JPEG Compression Domain for Malaria Infection Detection. *J Imaging*. 2022;8(5):129. doi: 10.3390/jimaging8050129
22. Dong T, Cook S, Oh J, Jang ES. Evaluation and analysis of image compression effect on neural network-based heart rate classification. *Sci Rep*. 2025;15(1):23173. doi: 10.1038/s41598-025-06031-8
23. Koff D, Bak P, Brownrigg P, Hosseinzadeh D, Khademi A, Kiss A, et al. Pan-Canadian Evaluation of Irreversible Compression Ratios ("Lossy" Compression) for Development of National Guidelines. *J Digit Imaging*. 2009;22(6):569-578. doi: 10.1007/s10278-008-9139-7
24. Koff D, Bak P, Matos A, Norman G. Evaluation of Irreversible Compression Ratios for Medical Images Thin Slice CT and Update of Canadian Association of Radiologists (CAR) Guidelines. *J Digit Imaging*. 2013;26(3):440-446. doi: 10.1007/s10278-012-9542-y
25. Wang Q, Zhang R. Double JPEG compression forensics based on a convolutional neural network. *EURASIP J on Info Security*. 2016;2016(1):23. doi: 10.1186/s13635-016-0047-y
26. Provedel LF, Pimentel BNADS, Jesus TMD, Lacerda APS, Rubira-Bullen IRF, Sarmiento VA. Effects of Image Compression on Linear Measurements of Digital Panoramic Radiographs. *Braz Dent J*. 2016;27(6):757-760. doi: 10.1590/0103-6440201601157
27. Wenzel A, Borg E, Hintze H, Gröndahl HG. Accuracy of caries diagnosis in digital images from charge-coupled device and storage phosphor systems: an in vitro study. *Dentomaxillofacial Radiology*. 1995;24(4):250-254. doi: 10.1259/dmfr.24.4.9161170

28. Momeny M, Latif AM, Agha Sarram M, Sheikhpour R, Zhang YD. A noise robust convolutional neural network for image classification. *Results in Engineering*. 2021;10:100225. doi: 10.1016/j.rineng.2021.100225
29. Ilesanmi AE, Ilesanmi TO. Methods for image denoising using convolutional neural network: a review. *Complex Intell Syst*. 2021;7(5):2179-2198. doi: 10.1007/s40747-021-00428-4
30. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys Med Biol*. 2021;66(4):04TR01. doi: 10.1088/1361-6560/abed17

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Nairobi Hernández Bridon.

Writing–review and editing: Nairobi Hernández Bridon.