



Membership inference vulnerability and the mitigating influence of regularization and differential privacy in dental radiograph classifiers

Vulnerabilidad a la inferencia de pertenencia en clasificadores de radiografías dentales y el efecto de regularización y privacidad diferencial

Teng He¹

Wenya Long¹

Yingjuan Xie¹

¹ Zhengzhou University, The First Affiliated Hospital, Department of Hematology. Zhengzhou, China.

Submitted: 17-05-2026 | Revised: 10-07-2026 | Accepted: 23-09-2026 | Published: 25-09-2026

Corresponding Author: Wenya Long 

ABSTRACT

Introduction: Deep learning classifiers for dental panoramic radiographs are trained on patient images, and the privacy of that training data is rarely audited. **Methods:** ResNet-18 features were extracted from 182 annotated teeth in the openly licensed DENTEX dataset and used to train a four-class diagnostic classifier under four configurations (a baseline, L2 weight decay, dropout, and differentially private stochastic gradient descent). Membership inference was mounted with loss- and confidence-threshold attacks and scored with the area under the receiver operating characteristic curve across ten 70/30 train/test splits. **Results:** The baseline classifier was vulnerable, with a loss-based attack AUC of 0.648 (95% CI 0.62-0.68). L2 weight decay and dropout reduced the AUC only marginally, to 0.647 and 0.633, whereas differentially private training reduced it to 0.537 (95% CI 0.53-0.55) and brought the confidence-based attack to chance (0.499), while lowering test accuracy from 70.4% to 54.5%. **Conclusions:** Dental radiograph classifiers leak membership information; conventional regularization offers little protection, whereas differential privacy substantially reduces the risk at the cost of accuracy.

Keywords: membership inference; differential privacy; dental radiograph; deep learning; privacy.

RESUMEN

Introducción: Los clasificadores de aprendizaje profundo para radiografías panorámicas dentales se entrenan con imágenes de pacientes, y la privacidad de esos datos de entrenamiento rara vez se audita. **Métodos:** Se extrajeron características de ResNet-18 de 182 dientes anotados del conjunto abierto DENTEX y se entrenó un clasificador diagnóstico de cuatro clases en cuatro configuraciones (línea de base, decaimiento de pesos L2, abandono y descenso de gradiente estocástico diferencialmente privado). La inferencia de pertenencia se montó con ataques por umbral de pérdida y de confianza y se midió con el área bajo la curva característica operativa del receptor en diez particiones 70/30. **Resultados:** El clasificador de línea de base fue vulnerable, con un AUC de ataque por pérdida de 0,648 (IC 95 % 0,62-0,68). El decaimiento L2 y el abandono lo redujeron solo marginalmente, a 0,647 y 0,633, mientras que el entrenamiento diferencialmente privado lo redujo a 0,537 (IC 95 % 0,53-0,55) y llevó el ataque por confianza al azar (0,499), a costa de bajar la exactitud de prueba de 70,4 % a 54,5 %. **Conclusiones:** Los clasificadores de radiografías dentales filtran información de pertenencia; la regularización convencional protege poco, mientras que la privacidad diferencial reduce sustancialmente el riesgo a costa de la exactitud.

Palabras clave: inferencia de pertenencia; privacidad diferencial; radiografía dental; aprendizaje profundo; privacidad.

INTRODUCTION

Panoramic radiography is a routine examination in dental practice, and deep learning classifiers trained on these images now detect caries, restorations, impacted teeth, periapical lesions, root anomalies and other conditions with accuracy that approaches or matches experienced clinicians. (1-3) Systematic reviews of this literature report sensitivities and specificities that would have been difficult to imagine a decade ago, and the field has moved from proof-of-concept toward products that are integrated into imaging workstations and clinical decision support. (4) The clinical promise of these systems, however, rests on large collections of patient radiographs, which are concentrated in the training sets of hospital groups, software vendors and research consortia. Because a single panoramic radiograph can reveal the anatomy, age, prior treatment and oral health of an identifiable person, the training set itself has become a valuable and sensitive asset, and its confidentiality has not received the same scrutiny as the diagnostic performance of the models it produces.

Membership inference is the adversary's attempt to learn whether a specific record was part of a model's training data, and it is now a mature and reproducible privacy threat for machine learning systems. (5-7) An attacker who observes only the model's confidence scores, or in some settings only its hard labels, can often distinguish members from non-members with accuracy well above chance, and the leakage is stronger for models that overfit their training distribution. (8,9) The problem matters in health care in particular because the training records are clinical images or records of real patients, and because an inference that a person's radiograph was used to train a model can reveal a diagnosis, a treatment history or a data-sharing agreement that the patient never authorized. (10,11) Related attack families, including model inversion and attribute inference, extend the same principle and demonstrate that the release of a medical model can be tantamount to the release of information about the individuals behind it. (10)

The same techniques have been demonstrated in medical imaging, where classifiers for retinal, chest and other clinical images have been shown to reveal the presence of individual patients in their training sets. (12,13) Studies that audit collaboratively trained models and those that mount query-based attacks on medical data converge on the conclusion that medical deep learning is not immune to membership leakage. (12,13) Defenses have been proposed that add generative noise or synthesize protective images, but they typically trade diagnostic accuracy for privacy. (14) What remains underexplored is whether the transfer-learning classifiers that dominate dental radiograph analysis exhibit the same vulnerability, and whether the defenses proposed in the privacy literature actually reduce the risk in this specific setting. Dental imaging has distinct properties that could matter: panoramic radiographs are highly standardized, the diagnostic tasks are often unbalanced, and the prevailing pipeline reuses a frozen general-purpose backbone, all of which shape how much a model memorizes and how much an attacker can extract.

The regulatory environment raises the stakes of this question. Data protection frameworks that govern health data in Europe, North America and a growing number of jurisdictions in Africa, Asia and Latin America treat a patient radiograph as a biometric or health datum that cannot be repurposed for model training without a lawful basis, and they impose accountability obligations on whoever processes it. If a trained classifier leaks the identity of its training subjects, that leakage becomes a potential breach of those obligations, with consequences that differ across health systems. Privacy audits such as membership inference therefore serve a compliance function as well as a research one, and their results should inform the governance of dental AI wherever patient images are pooled. (11,15)

Two families of mitigation are available. The first is conventional regularization, such as L2 weight decay and dropout, which reduces overfitting and is therefore expected, under the overfitting hypothesis, to limit membership leakage. (5,16) Its effect has been inconsistent across studies, with some reporting only modest gains against a determined attacker and others finding that the reduction in leakage is too small to be practically meaningful. (17) The second is differential privacy, formalized as a bound on the influence of any single training example, and made practical for deep learning by the differentially private stochastic gradient descent (DP-SGD) algorithm. (18-20) Differential privacy carries a well-documented utility cost, and in medical imaging that cost has been large enough to discourage deployment in several benchmark studies. (21-23) Whether that cost is acceptable for a dental diagnostic classifier, and whether it buys a meaningful reduction in membership inference, has not been established.

This study audits the membership inference risk of a transfer-learning classifier for four common dental diagnoses in panoramic radiographs and compares the mitigating influence of regularization and differential privacy. The objective was to quantify membership inference risk in a dental radiograph classifier and to measure how much L2 weight decay, dropout and DP-SGD

reduce that risk. The following hypotheses were tested: H1, a baseline classifier without privacy protection is vulnerable to membership inference, with attacker AUC above chance; H2, L2 weight decay and dropout reduce membership inference AUC relative to the baseline; and H3, DP-SGD reduces membership inference AUC to approximately chance.

METHODS

Study design and dataset

This was an experimental privacy audit conducted on a public, openly licensed image dataset, with no human participants and no ethics approval required beyond the terms of the licence. The DENTEX validation subset(24)provided 50 panoramic radiographs with 182 annotated teeth.

Each annotation contained a bounding box and a diagnosis label drawn from four classes: impacted tooth, caries, periapical lesion and deep caries. The dataset also carried quadrant and enumeration labels, which were not used in this analysis.

The dataset was used exactly as released; no images were excluded and no labels were modified. The choice of a validation subset reflected the fact that it is the smallest openly licensed partition with per-tooth diagnosis labels, and its use kept the experiment reproducible and feasible on a central processing unit while preserving the full transfer-learning pipeline that would be applied to a larger training set.

The DENTEX dataset was selected over alternative dental image collections because it is hosted on a public repository with an explicit licence, requires no registration to download, and has been used as a benchmark in prior dental imaging studies, which makes the present audit directly comparable to the performance literature on the same images.(3,24)

Feature extraction and classification model

Every annotated tooth was cropped from its panoramic image using its bounding box, resized to 224 by 224 pixels and standardized with the ImageNet mean and standard deviation. A ResNet-18 convolutional network pre-trained on ImageNet (25) served as a frozen feature extractor: its final fully connected layer was removed, and the 512-dimensional penultimate representation of each crop was computed in forward mode without gradient.

The crops were processed in batches, and the resulting feature matrix of 182 rows and 512 columns was the sole input to the classifier. The classifier itself was a single linear layer mapping these 512 features to the four diagnosis logits.

This transfer-learning design reflects the dominant practice in dental imaging, in which a pre-trained backbone is reused and only a lightweight head is trained, and it isolates the privacy behaviour of the trained head, which is the component whose parameters would be released or shared in a typical deployment. A single linear head also has a direct bearing on the analysis: its parameters are few, its training is fast and deterministic enough to be repeated many times on a central processing unit, and its per-sample gradients can be computed exactly, which is a prerequisite for the differentially private update described below.

The linear head is therefore not a simplification of convenience but a deliberate choice that makes the privacy mechanisms measurable in closed form.

Training configurations

Four training configurations operationalized the hypotheses. The baseline trained the linear head with stochastic gradient descent for 200 epochs, a learning rate of 0.01, no weight decay and no dropout. The L2 configuration added a weight decay coefficient of 0.01 to the same optimizer, penalizing the squared norm of the head weights.

The dropout configuration inserted a dropout layer with probability 0.5 immediately before the linear head and otherwise matched the baseline, so that each feature was masked with probability one half during training. These two configurations are the most common regularizers in dental image classification and represent the mitigation that an engineer would plausibly apply without specialized privacy tooling.

The DP-SGD configuration replaced the optimizer with a manual differentially private update, following the algorithm of Abadi and colleagues.(18)In each epoch the per-sample gradient of the cross-entropy loss with respect to the head parameters was computed for every training example, clipped so that its L2 norm did not exceed a bound of 1.0, summed across the batch,

and perturbed with Gaussian noise whose standard deviation was set by multiplying the clipping bound by a noise multiplier of 1.1.

The noisy sum was divided by the batch size and applied as a gradient step at a learning rate of 0.05 for 60 epochs. The clipping bound and noise multiplier were chosen to produce a non-trivial privacy guarantee while remaining within the utility range of a small dataset. All training configurations used a fixed random seed for reproducibility, and the model was reset to the same initialization before each configuration.

Membership inference attack

After each model was trained on its training split, a threshold membership inference attack was mounted against it. For every sample in the pooled training and test sets, the attacker computed two candidate scores: the per-sample cross-entropy loss, on which members tend to score lower because the model was optimized to minimize their loss, and the maximum softmax confidence, on which members tend to score higher for the same reason.

Each score was evaluated separately as a binary discriminator between members and non-members by sweeping a decision threshold over the range of observed scores, and its effectiveness was summarized by the area under the receiver operating characteristic curve, with 0.5 corresponding to chance. This is the standard loss- and confidence-threshold attack described in the membership inference literature and is available to any adversary who can query the model and obtain its output probabilities.(5,6,26)

The attack used the same feature representation as the model itself, which corresponds to a strong but realistic adversary who knows the feature extractor, a common assumption in the white-box privacy literature.(16)The area under the curve was chosen because it summarizes the attack's discrimination across all possible thresholds without requiring the defender to commit to a particular operating point, and because it is the standard summary used across the membership inference literature.(7,27)

Statistical analysis

The experiment was repeated across ten random stratified 70/30 train/test splits, so that each configuration produced ten values of test accuracy and ten attack AUCs for each score. The stratification preserved the proportion of each diagnosis class in both partitions, which was necessary because the periapical lesion class contained only nine examples. Means and standard deviations were computed across the ten seeds, and 95% confidence intervals for the headline AUCs were derived from the t distribution. Each hypothesis was operationalized and resolved by a pre-specified test and decision criterion.

For H1, the outcome was the baseline loss-based attack AUC, the measure was its mean over seeds, and the test was a one-sided one-sample t test of that mean against 0.5; the hypothesis was accepted if the mean exceeded 0.5 with p below 0.05. For H2, the outcome was the difference between the baseline loss-based AUC and the loss-based AUC of each regularized configuration, and the test was a one-sided paired t test across seeds; the hypothesis was accepted for each regularizer if its AUC was significantly lower than the baseline. For H3, two criteria were applied: a two-sided one-sample t test of the DP-SGD loss-based AUC against 0.5, to establish whether it was distinguishable from chance, and a one-sided paired t test of the DP-SGD loss-based AUC against the baseline, to establish whether the reduction was significant. Test accuracy was reported as a percentage and served as the utility reference for the privacy-utility trade-off. 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.

RESULTS

The study included 182 annotated teeth obtained from 50 panoramic radiographs (Table 1). The diagnosis classes were uneven: caries predominated with 101 teeth, followed by impacted teeth with 40, deep caries with 32 and periapical lesions with 9.

H1: Vulnerability of the baseline classifier to membership inference

The baseline classifier was vulnerable to membership inference (Table 2; Figure 1). The loss-based attack achieved an AUC of 0.648, with a 95% confidence interval of 0.62 to 0.68, which exceeded the chance value of 0.5 ($p < 0.001$). The confidence-based attack was weaker, reaching an AUC of

0.582 with a standard deviation of 0.057 across seeds. The baseline test accuracy across the ten splits was 70.4% with a standard deviation of 3.9%. The loss-based and confidence-based attacks therefore diverged in strength by a margin of roughly six percentage points, a gap that recurred in every training configuration and that reflects the greater information carried by the loss signal.

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.

Table 2. Test accuracy and membership inference AUC across training configurations.

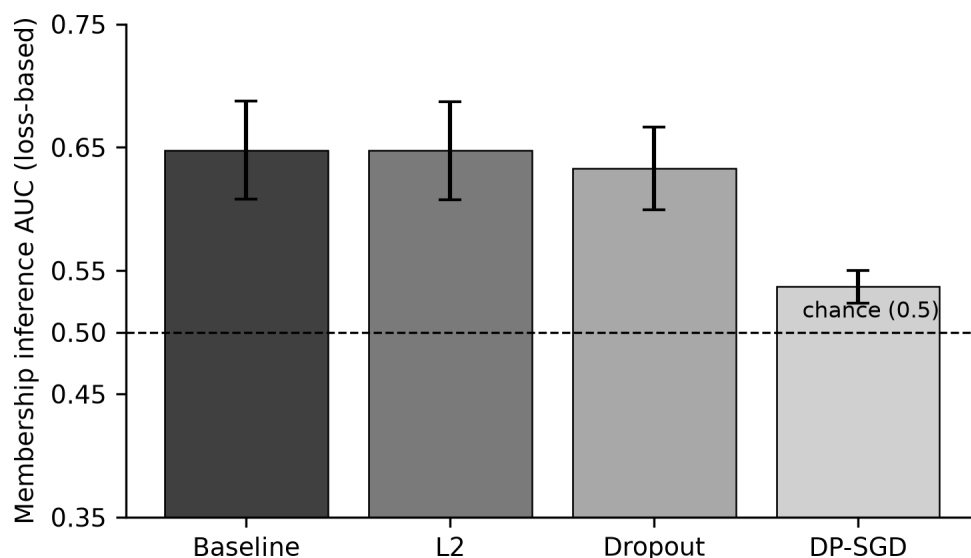
Configuration	Test accuracy (%)	MIA AUC (loss)	MIA AUC (confidence)
baseline	70.4 ± 3.9	0.648 ± 0.040	0.582 ± 0.057
l2	70.4 ± 3.9	0.647 ± 0.040	0.581 ± 0.057
dropout	70.4 ± 3.4	0.633 ± 0.034	0.563 ± 0.049
dpsgd	54.5 ± 0.0	0.537 ± 0.013	0.499 ± 0.031

Note: Values are mean ± standard deviation across 10 seeds. MIA = membership inference; AUC = area under the receiver operating characteristic curve; loss- and confidence-based attacks are threshold scores.

H2: Effect of regularization on membership inference risk

Regularization reduced membership inference risk only marginally (Table 2). The loss-based AUC fell from 0.648 in the baseline to 0.647 with L2 weight decay ($p = 0.005$) and to 0.633 with dropout ($p < 0.001$). The confidence-based AUC moved in the same direction, from 0.582 to 0.581 with L2 and to 0.563 with dropout. Neither form of regularization altered test accuracy, which remained at 70.4% in both configurations, with a slightly smaller standard deviation of 3.4% under dropout.

Figure 1. Membership inference attack AUC across training configurations.



Note: Bars show the mean loss-based membership inference AUC with one standard deviation across 10 seeds. The dashed line marks the chance value of 0.5.

H3: Effect of differential privacy on membership inference risk

Differentially private training produced the largest reduction in membership inference risk, together with a clear loss of utility (Table 2). The loss-based AUC decreased from 0.648 to 0.537, with a 95% confidence interval of 0.53 to 0.55, and the paired comparison against the baseline was significant ($p < 0.001$). The confidence-based AUC fell to 0.499, which was indistinguishable from chance. The residual loss-based AUC of 0.537 remained slightly above 0.5 ($p < 0.001$), with a small standard deviation of 0.013 across seeds. Test accuracy dropped from 70.4% to 54.5%, a value close to the prevalence of the majority class, and the standard deviation collapsed to zero, indicating that the differentially private model predicted a constant class in every split.

DISCUSSION

The main finding of this audit was that a transfer-learning dental panoramic radiograph classifier trained on an openly licensed dataset leaked membership information, and that the choice of mitigation determined whether that leakage could be brought to chance. The baseline loss-based attack AUC of 0.648 places dental radiograph classifiers among the machine learning models for which membership inference is a realistic threat.(5,27)

The magnitude is consistent with the modest but systematic leakage reported for overfitting deep classifiers in other domains, and with audits of clinical image models that found membership detectable from model outputs alone.(12,13) Dental images are no exception to a pattern that has been documented for retinal and chest imaging, and the result argues against the assumption that a medical classifier is private merely because its parameters are not released.

H1 was supported, and the finding is consistent with the theoretical account that ties membership leakage to overfitting. Models that memorize the idiosyncrasies of their training examples expose those examples to inference, and the gap between the loss distribution of members and non-members is the observable trace of that memorization.(8)

The confidence-based attack was weaker than the loss-based attack in this study, a pattern that matches prior work showing that loss is a more informative signal than confidence for threshold attacks, and that label-only attacks are weaker still.(6,9) The direction of these differences matters because it identifies which output a defender must protect: a model that publishes calibrated probabilities still leaks through its losses, and the loss signal is available to any party that can compute the training objective.

The finding also aligns with the observation that membership inference attacks are easier on difficult, poorly separated tasks, which describes a four-class diagnosis over an unbalanced and small training set.(17) In that respect the dental classifier is not an outlier but a representative case of the conditions under which leakage is most pronounced.

H2 was supported statistically, but the effect size was negligible for practice. L2 weight decay reduced the loss-based AUC by only 0.001 and dropout by 0.015, and neither changed test accuracy.

The result is in line with studies that found regularization to be a weak membership inference defense: because the attacker's decision boundary is fitted to the loss surface rather than to the model's generalization gap, a small reduction in overfitting does not erase the separation between members and non-members.(16,17)

The near-null effect of L2 is unsurprising under this account, since weight decay on a single linear head hardly constrains the memorization that drives leakage, and the modest effect of dropout is consistent with its role as a stochastic regularizer that reduces, but does not eliminate, the confidence gap between members and non-members.(5)

Regularization should therefore not be relied upon as a privacy control, even though it is already present in many dental AI pipelines for other reasons. Its real value is that it costs nothing in accuracy, which makes it a sensible default but not a safeguard. These results contrast with the more optimistic view that regularization is a sufficient membership inference defense, and they reinforce the conclusion of comprehensive analyses that no amount of weight decay or dropout, applied to a well-fitting model, closes the gap to chance.(16,26)

The reason is structural rather than parametric: a linear head trained to convergence on a small feature set reproduces the same separation between members and non-members regardless of how its norm is penalized.

H3 was largely supported, with an important qualification. DP-SGD reduced the loss-based AUC from 0.648 to 0.537 and brought the confidence-based attack to chance at 0.499, which is the expected outcome for a mechanism that bounds the influence of each individual training example.(18,20) The residual loss-based leakage above 0.5 is consistent with the small-sample setting, in

which the noise required for a meaningful privacy guarantee also limits how far the model can be pushed toward a stable optimum, and with the known difficulty of reaching exactly the chance level on every attack score in practice.(23,28)

The qualification is the utility collapse: test accuracy fell from 70.4% to 54.5%, near the majority-class rate, and the zero variance across splits indicates that the differentially private model predicted a constant class in every run, effectively abandoning the minority classes. This trade-off mirrors the pattern reported in medical imaging, where differentially private training has often degraded performance enough to raise doubts about clinical viability, and where privacy-utility reconciliation remains an active research problem.(21,22,29,30)

For a dental classifier, the price of near-chance membership inference was a model that no longer distinguished periapical lesions from the majority class, which is not acceptable for a diagnostic tool.

The mechanism behind the collapse of the differentially private model is instructive. In a small, unbalanced dataset the gradient signal for the minority classes is weak, and per-sample clipping together with the additive noise masks that signal, so the optimizer settles on the majority class that minimizes the expected loss.

This is not a failure of the algorithm but a direct consequence of the privacy budget: at the chosen noise level the guarantee consumes the statistical power that a linear head needs to separate nine periapical lesions from the rest. The same mechanism explains why the loss-based attack retained a small advantage over chance even after the confidence-based attack was neutralized; the loss encodes a residual difference in the per-example objective that the clipped and noisy updates never fully erase, whereas the confidence collapses toward the constant output of a degenerate model.

This distinction between the two attack scores is not a curiosity but a practical warning, because it shows that a defense that appears to reach chance on one metric can still leave a measurable signal on another.(28,31)

These findings generalize across health systems in a specific sense. Membership inference is a property of the training algorithm and the geometry of the dataset rather than of the patient population, so a classifier trained in a high-income setting is, all else being equal, no more or less private than one trained elsewhere.

What varies across jurisdictions is the size and class balance of the available training data and the regulatory weight attached to the outcome. A dental center in a low- or middle-income country that can assemble only a few hundred labelled images faces the same privacy-utility frontier documented here, and the data protection law under which it operates, whether modeled on European rules or a sectoral health regulation, determines how much residual leakage is tolerable.

The utility collapse observed under DP-SGD is therefore a caution for any deployment that relies on small, class-imbalanced dental datasets, which are common across Africa, Latin America and parts of Asia where large annotated collections are scarce. The same principle applies in reverse: a well-resourced center with tens of thousands of labelled images can afford the accuracy loss that a smaller center cannot, so the optimal privacy strategy is a function of data abundance rather than of any universal rule.

The comparison between the three hypotheses resolves the question posed by the study in a way that should guide practitioners. Conventional regularization is insufficient, differential privacy is effective but costly at this dataset scale, and the residual risk under DP-SGD, though small, is not zero. The finding that the confidence-based attack reached chance while the loss-based attack did not further implies that even a privacy-conscious deployer who hides raw probabilities cannot assume safety, because the loss signal can often be reconstructed from repeated queries and is frequently exposed through auxiliary outputs such as uncertainty estimates.(9,16)

The trade-off observed here also bears on the broader governance of dental AI, in which federated and privacy-preserving training is increasingly proposed as a route to multi-institutional datasets; those approaches inherit the same tension between confidentiality and utility documented in the medical imaging literature, and they add the complication of heterogeneous participating sites.(15,32,33)

A dental institution that releases or shares a classifier should treat the training set as exposed by default, measure the actual membership inference risk with an audit of the kind performed here, and select a privacy budget explicitly rather than assuming that any single technique is sufficient.(31)

Concretely, a vendor that publishes a model card should report the attack AUC alongside accuracy, a research group that deposits a dental classifier should disclose whether differentially private training was used and at what budget, and a regulator assessing a dental AI submission

should ask whether membership inference was tested before granting authorization. These are low-cost steps that convert a technical finding into an operational safeguard, and they are portable across the jurisdictions in which dental AI is deployed.

Several strengths support the validity of these results. The audit used a public dataset with an explicit licence, a fixed random seed and a script that was executed twice with identical output, so the findings are reproducible. The attack was evaluated over ten independent splits rather than a single partition, which stabilizes the AUC estimates and allowed the hypothesis tests to reflect run-to-run variability rather than a single lucky or unlucky partition. The use of a frozen pre-trained backbone matches the transfer-learning practice that dominates dental imaging, so the vulnerability measured here is the one that would confront a typical deployed system rather than an idealized end-to-end network.

The explicit mapping of each hypothesis to a test and a decision criterion, decided before the analysis, reduces the risk of post hoc interpretation. The methodology also transferred unchanged from the general privacy literature to a dental task, which means the audit can be repeated by any dental group with a labelled radiograph set and a pre-trained backbone, without specialized privacy infrastructure. Finally, reporting both loss- and confidence-based attacks avoided the single-metric optimism that can occur when a defense is evaluated against only one attack score.

Limitations

- Size of the dataset: The 182 annotated teeth in the DENTEX validation subset are fewer than a production training set, and the small number of periapical lesions (nine) limited the model's ability to learn that class under all configurations. A larger, multi-institution dataset would yield more stable utility estimates and a more precise picture of the privacy-utility frontier; future work should repeat this audit on a full training split.
- The attack assumed access to model probabilities, which is not always granted in a deployed system; label-only attacks are weaker, so the reported AUCs represent an upper bound on risk for models that expose only hard predictions. An audit that restricts the adversary to labels would clarify how much of the leakage survives that constraint.
- The single noise configuration for DP-SGD: The clipping bound and noise multiplier were fixed, and other choices would produce a different point on the privacy-utility curve; a sweep over the privacy budget would map that curve and identify whether a configuration exists that preserves clinically useful accuracy while keeping membership inference at chance.
- The use of a linear head on frozen features, which may underestimate the leakage of an end-to-end fine-tuned network that memorizes more input detail; fine-tuning the backbone under each privacy configuration would test whether the same ordering of mitigations holds when the full network is trained.
- The audit considered only passive, threshold attacks on a released model and did not evaluate active or adaptive attackers who can influence training or mount shadow-model and white-box attacks with full parameter access. Adaptive attacks that train shadow models on the same distribution can recover more of the membership signal than the fixed thresholds used here; future audits should include shadow-model attacks to bound the worst case.

CONCLUSIONS

A baseline dental panoramic radiograph classifier was vulnerable to membership inference, with a loss-based attack AUC substantially above chance.

L2 weight decay and dropout reduced membership inference risk only negligibly and did not alter accuracy.

Differentially private training reduced the risk to near chance but collapsed diagnostic accuracy, so the choice of privacy mechanism is a measured trade-off rather than a cost-free safeguard.

The findings support the routine inclusion of membership inference audits in the development and governance of dental artificial intelligence.

REFERENCES

1. Ayhan B, Ayan E, Atsü S. Detection of dental caries under fixed dental prostheses by analyzing digital panoramic radiographs with artificial intelligence algorithms based on deep learning methods. *BMC Oral Health*. 2025;25(1):216. doi: 10.1186/s12903-025-05577-3
2. Pornprasertsuk-Damrongsri S, Vachmanus S, Papasratorn D, Kitisubkanchana J, Chaikantha S, Arayasan-tiparb 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
3. Bonfanti-Gris M, Herrera A, Salido Rodríguez-Manzaneque MP, Martínez-Rus F, Pradíes G. Deep learning for tooth detection and segmentation in panoramic radiographs: a systematic review and meta-analysis. *BMC Oral Health*. 2025;25(1):1280. doi: 10.1186/s12903-025-06349-9
4. Silva B, Fontinele J, Vieira CLZ, Tavares JMRS, Cury PR, Oliveira L. A holistic approach for classifying dental conditions from textual reports and panoramic radiographs. *Medical Image Analysis*. 2025;105:103709. doi: 10.1016/j.media.2025.103709
5. Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks Against Machine Learning Models. 2017 IEEE Symposium on Security and Privacy (SP). 2017:3-18. doi: 10.1109/sp.2017.41
6. Salem A, Zhang Y, Humbert M, Berrang P, Fritz M, Backes M. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. *Proceedings 2019 Network and Distributed System Security Symposium*. 2019. doi: 10.14722/ndss.2019.23119
7. Hu H, Salicic Z, Sun L, Dobbie G, Yu PS, Zhang X. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput Surv*. 2022;54(11s):1-37. doi: 10.1145/3523273
8. Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. 2018 IEEE 31st Computer Security Foundations Symposium (CSF). 2018:268-282. doi: 10.1109/csf.2018.00027
9. Li Z, Zhang Y. Membership Leakage in Label-Only Exposures. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021:880-895. doi: 10.1145/3460120.3484575
10. Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015:1322-1333. doi: 10.1145/2810103.2813677
11. Krüger FP, Östman J, Mervin L, Tetko IV, Engkvist O. Publishing neural networks in drug discovery might compromise training data privacy. *J Cheminform*. 2025;17(1):38. doi: 10.1186/s13321-025-00982-w
12. Huang Y, Huang CY, Li X, Li K. A Dataset Auditing Method for Collaboratively Trained Machine Learning Models. *IEEE Trans Med Imaging*. 2023;42(7):2081-2090. doi: 10.1109/tmi.2022.3220706
13. Pedarla LP, Zhang X, Zhao L, Khan H. Evaluation of Query-Based Membership Inference Attack on the Medical Data. *Proceedings of the 2023 ACM Southeast Conference*. 2023:191-195. doi: 10.1145/3564746.3587027
14. Paul W, Cao Y, Zhang M, Burlina P. Defending Medical Image Diagnostics Against Privacy Attacks Using Generative Methods: Application to Retinal Diagnostics. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2021. p. 174-187. doi: 10.1007/978-3-030-90874-4_17
15. Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G. A survey on security and privacy of federated learning. *Future Generation Computer Systems*. 2021;115:619-640. doi: 10.1016/j.future.2020.10.007
16. Nasr M, Shokri R, Houmansadr A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. 2019 IEEE Symposium on Security and Privacy (SP). 2019:739-753. doi: 10.1109/sp.2019.00065
17. Shafran A, Peleg S, Hoshen Y. Membership Inference Attacks are Easier on Difficult Problems. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:14800-14809. doi: 10.1109/iccv48922.2021.01455
18. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016:308-318. doi: 10.1145/2976749.2978318
19. Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*. 2014;9(3-4):211-487. doi: 10.1561/04000000042
20. Dwork C. Differential Privacy: A Survey of Results. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg; 2008. p. 1-19. doi: 10.1007/978-3-540-79228-4_1
21. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2(6):305-311. doi: 10.1038/s42256-020-0186-1
22. Ziller A, Usynin D, Braren R, Makowski M, Rueckert D, Kaissis G. Medical imaging deep learning with differential privacy. *Sci Rep*. 2021;11(1):13524. doi: 10.1038/s41598-021-93030-0
23. Ziller A, Mueller TT, Stieger S, Feiner LF, Brandt J, Braren R, et al. Reconciling privacy and accuracy in AI for medical imaging. *Nat Mach Intell*. 2024;6(7):764-774. doi: 10.1038/s42256-024-00858-y
24. 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>
25. 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
26. Truex S, Liu L, Gursoy ME, Yu L, Wei W. Demystifying Membership Inference Attacks in Machine Learning as a Service. *IEEE Trans Serv Comput*. 2021;14(6):2073-2089. doi: 10.1109/tsc.2019.2897554
27. Carlini N, Chien S, Nasr M, Song S, Terzis A, Tramèr F. Membership Inference Attacks From First Principles. 2022 IEEE Symposium on Security and Privacy (SP). 2022:1897-1914. doi: 10.1109/sp46214.2022.9833649

28. Benladgham R, Hadjila F, Belloum A. Efficient Privacy-Utility Optimization for Differentially Private Deep Learning. *Int j electr comput eng syst (Online)*. 2025;16(5):377-395. doi: 10.32985/ijeces.16.5.3
29. Zheng L, Cao Y, Yoshikawa M, Shen Y, Rashed EA, Taura K, et al. Sensitivity-Aware Differential Privacy for Federated Medical Imaging. *Sensors*. 2025;25(9):2847. doi: 10.3390/s25092847
30. Adams T, Birkenbihl C, Otte K, Ng HG, Rieling JA, Näher AF, et al. On the fidelity versus privacy and utility trade-off of synthetic patient data. *iScience*. 2025;28(5):112382. doi: 10.1016/j.isci.2025.112382
31. Wang Z, Zhang C, Chen Y, Baracaldo N, Kadhe SR, Yu L. Membership Inference Attacks as Privacy Tools: Reliability, Disparity and Ensemble. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*. 2025:1724-1738. doi: 10.1145/3719027.3744818
32. Raza A, Guzzo A, Ianni M, Lappano R, Zanolini A, Maggiolini M, et al. Federated Learning in radiomics: A comprehensive meta-survey on medical image analysis. *Computer Methods and Programs in Biomedicine*. 2025;267:108768. doi: 10.1016/j.cmpb.2025.108768
33. Elabd E. Dynamic differential privacy technique for deep learning models. *Sci Rep*. 2025;15(1):39353. doi: 10.1038/s41598-025-27708-0

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Teng He, Wenya Long, Yingjuan Xie.

Writing–review and editing: Teng He, Wenya Long, Yingjuan Xie.