



Energy and carbon cost of training dental image classifiers: how model size affects efficiency

Costo energético del entrenamiento de clasificadores de imágenes dentales según el tamaño del modelo

Leodanis Hernández Cabrera¹  

Maikro Osvaldo Chávez Moya¹  

Neisy Pérez Ramos²  

¹ Universidad de Ciencias Médicas de Villa Clara, Facultad de Medicina. Santa Clara, Cuba

² Universidad de Ciencias Médicas de Villa Clara, Facultad de Medicina, Departamento de Ciencias Biomédicas. Santa Clara, Cuba

Submitted: 31-03-2025 | Revised: 02-06-2025 | Accepted: 12-09-2025 | Published: 27-09-2025

Corresponding Author: Leodanis Hernández Cabrera 

ABSTRACT

Introduction: Artificial intelligence in dentistry is growing, but the energy and carbon cost of training dental image classifiers is rarely reported, so researchers and clinicians have no basis for choosing a model size that balances accuracy against environmental cost. **Methods:** Four pretrained models of increasing size (MobileNetV2, ResNet18, ResNet34, ResNet50) were fine-tuned on the same 182-tooth, four-class diagnosis task from the DENTEX dataset under an identical protocol. Training energy and carbon emissions were measured with the CodeCarbon library, and accuracy was evaluated on a held-out set. **Results:** Energy increased from 0.26 mWh for MobileNetV2 to 0.86 mWh for ResNet50, a 3.30-fold difference across a tenfold range in parameters. Accuracy did not rise with size: ResNet18 reached 62.20% accuracy, matching ResNet50 at 62.20% while using roughly one third of its energy. ResNet18 achieved the highest accuracy per unit energy. **Conclusions:** For dental image classification at this scale, larger models add energy and carbon without adding accuracy, and the smallest adequate model is the efficient choice. Reporting energy alongside accuracy should become routine when dental AI models are developed.

Keywords: deep learning; energy efficiency; carbon footprint; dental radiography; green AI.

RESUMEN

Introducción: La inteligencia artificial en odontología crece, pero el costo energético y de carbono de entrenar clasificadores de imágenes dentales rara vez se reporta, de modo que investigadores y clínicos carecen de base para elegir un tamaño de modelo que equilibre exactitud y costo ambiental. **Métodos:** Cuatro modelos preentrenados de tamaño creciente (MobileNetV2, ResNet18, ResNet34, ResNet50) se ajustaron sobre la misma tarea de diagnóstico de cuatro clases de 182 dientes del conjunto DENTEX con un protocolo idéntico. La energía y las emisiones de carbono del entrenamiento se midieron con la biblioteca CodeCarbon, y la exactitud se evaluó en un conjunto de prueba. **Resultados:** La energía aumentó de 0,26 mWh para MobileNetV2 a 0,86 mWh para ResNet50, una diferencia de 3,30 veces en un rango de parámetros diez veces mayor. La exactitud no creció con el tamaño: ResNet18 alcanzó 62,20% de exactitud, igual que ResNet50 (62,20%) con aproximadamente un tercio de su energía. ResNet18 logró la mayor exactitud por unidad de energía. **Conclusiones:** Para la clasificación de imágenes dentales a esta escala, los modelos mayores añaden energía y carbono sin añadir exactitud, y el modelo más pequeño adecuado es la elección eficiente. Reportar la energía junto a la exactitud debería ser rutinario al desarrollar modelos de IA dental.

Palabras clave: aprendizaje profundo; eficiencia energética; huella de carbono; radiografía dental; IA verde.

INTRODUCTION

Dental artificial intelligence has reached the point where a clinician can reasonably expect an automated reader for caries, periodontal bone loss, tooth numbering, or implant identification on a panoramic radiograph, and a researcher can expect to reproduce or extend such a system in a few days of work.(1-3)The applications now span detection, segmentation, and diagnosis across multiple imaging modalities, and reviews of the field catalogue a rapidly expanding set of models.(4,5) The growth has been driven by the ready availability of pretrained networks that can be fine-tuned on modest labeled datasets, which has lowered the barrier to building a dental model to the point that architecture choice is often made on convenience or habit rather than on any measured criterion, and that a developer who owns a powerful machine may default to the largest model it can run without a specific reason to choose otherwise. What is almost never reported alongside these systems is what it cost, in energy and carbon, to train them. The omission is not a curiosity; it is a decision made without information, because a developer who is choosing among model architectures has no way of knowing whether the largest model they can run is worth the environmental price of running it, and a reviewer has no standard against which to judge whether a given training run was efficient or wasteful.

The concern is no longer hypothetical for machine learning generally. The field has documented that training large models carries a substantial carbon footprint, and that this footprint scales steeply with model size and computational budget, in some cases producing emissions comparable to those of an entire transcontinental flight or the lifetime of several cars.(6,7) The scaling of models toward ever-larger architectures has made the energy question a mainstream research concern rather than a niche one.(8)The relevant comparison is not whether any single training run is large, but whether the energy is spent for a return, and the literature has repeatedly shown that it is often not: gains in accuracy that follow from doubling a model's size are frequently marginal, while the energy cost doubles or more.(9)This is the distinction between accuracy and efficiency that the Green AI movement has worked to make explicit, and it applies with particular force to transfer learning, where the pretrained weights already carry most of the representational capacity and the fine-tuning step is where a smaller model can often match a much larger one at a small fraction of the cost. Open-source libraries can now measure the energy and emissions of a training run directly, and comparative studies have shown that architecture and hardware choices can change the energy cost of the same task by an order of magnitude or more.(10-12)The principle that emerged, summarized under the heading of Green AI, is that accuracy should not be pursued without regard to cost, and that efficiency itself is a measurable research contribution worth reporting.(9)

This principle has begun to reach medicine. Sustainability frameworks now ask AI developers to account for the environmental footprint of their models alongside their performance, and reviews of medical AI have started to catalog the carbon implications of increasingly large models in clinical workflows.(13-15)The medical imaging community, in particular, has begun to discuss the environmental cost of the large models and large datasets that its benchmarks depend on, and to argue that efficiency should be a criterion in model selection rather than an afterthought.(15)Dentistry, as a discipline that is adopting the same models and the same transfer-learning workflow, is subject to the same considerations, even though its datasets are smaller and its models are typically fine-tuned rather than trained from scratch. The smaller scale of dental AI does not exempt it from the efficiency question; it merely changes the unit, and the same principle of spending energy only for a measurable return applies whether a training run costs milliwatt-hours or megawatt-hours. Within dentistry specifically, the literature on deep learning has grown rapidly, with studies spanning detection, segmentation, and diagnosis across panoramic, periapical, and cone-beam imaging, but it remains focused on accuracy and clinical performance, and the energy and carbon cost of dental AI has not, to our knowledge, been measured and reported.(16-20)A dental researcher or clinic deploying models therefore faces an efficiency question with no measured answer in their own field.

The practical stakes are concrete rather than abstract. A dental clinic that deploys models at the edge, on a workstation or a dedicated inference device, pays for its electricity and dissipates its heat, and a researcher who runs many candidate architectures during model development multiplies that cost many times over before a single model is chosen. The choice of model size is one of the few energy-relevant decisions that a developer controls directly, and it is also one of the easiest to make well, because it requires only a measurement of energy and accuracy rather than any change to the underlying task or data. A measured efficiency result therefore has immediate, actionable value: it tells a developer which of the available pretrained models is the responsible

default, and it gives a reviewer or a reader a concrete number against which to judge a reported model.

This study provides that measurement. We fine-tuned four pretrained models spanning an order of magnitude in parameter count, from MobileNetV2 through ResNet50, on the same dental diagnosis task under an identical protocol, and measured the energy and carbon of each training run with CodeCarbon. The aim was to answer a question that a practitioner can act on directly: how does the energy and carbon cost of training a dental image classifier scale with model size, and how much accuracy is gained per unit of energy? The answer matters because the choice of model size is one of the few energy-relevant decisions a developer controls, and because efficiency, once measured, becomes a criterion that can guide that choice just as accuracy does.(21)

METHODS

Dataset and task

The task was tooth-level diagnosis on panoramic dental radiographs. We used the publicly available DENTEX dataset, an open benchmark released under a Creative Commons license with per-tooth bounding boxes and diagnosis labels.(3)The validation split provides 182 annotated abnormal teeth across four diagnosis classes: impacted tooth, caries, periapical lesion, and deep caries. Each annotated tooth was cropped from its panoramic image and resized to 224 by 224 pixels with three color channels, giving the classification task used throughout the study. The four classes were unevenly represented, with caries the most common and periapical lesion the least, which mirrors the epidemiology of the underlying conditions and is typical of dental imaging benchmarks, and this imbalance, while a source of noise in the accuracy estimates, does not affect the energy measurement because every model trained on the identical set of examples. The task was chosen because it is representative of the abnormality classification that dental AI systems perform in practice, and because the dataset's license permits reuse, which is a prerequisite for a reproducible energy study. A licensed, freely downloadable dataset also matters for the practical reproducibility of the result, since a reader who wishes to re-measure the energy on their own hardware can obtain the identical images and the identical annotations without any administrative barriers. Using the same task and data for every model ensures that the energy comparison isolates the effect of the architecture and nothing else. The tooth crops were generated once and reused for every model, and the same preprocessing, normalization, and input size were applied throughout, so that no difference in the input pipeline could confound the energy or accuracy measurements.

Models

Four convolutional networks of increasing size were compared, spanning roughly a tenfold range in parameter count: MobileNetV2 (approximately 2.2 million parameters), ResNet18 (11.2 million), ResNet34 (21.3 million), and ResNet50 (23.5 million). MobileNetV2 was included as the small, depthwise-separable architecture designed for efficiency on limited hardware, and the ResNet family was included as the widely used residual baseline with increasing depth and capacity. All four were initialized from ImageNet-pretrained weights, and the final classification layer of each was replaced with a four-way output. This design holds the task, data, and initialization source constant and varies only the architecture, so any measured difference in energy is attributable to the model rather than to the data or the training procedure, and the comparison spans both the architecture families and the size range that a dental developer would realistically consider. The parameter counts, from 2.2 million to 23.5 million, cover the span between a deliberately frugal model and a standard deep model, and the energy measured across that span therefore captures the cost of choosing capacity over frugality.

Training protocol

Each model was fine-tuned under an identical protocol so that the energy comparison was fair. The 182 tooth crops were divided into a training set of 145 and a held-out test set of 37 by a fixed, stratified random split shared across all models, so that every model was trained and evaluated on exactly the same examples. All parameters of each network were trained for four epochs with the Adam optimizer at a learning rate of $1e-4$, a batch size of 16, and cross-entropy loss, on the same central-processing-unit hardware and with the same software versions. The same random seed was used throughout, and no data augmentation or early stopping was applied, so that the only

systematic difference between runs was the architecture. Four epochs and a small batch size were chosen to keep the workload modest and reproducible rather than to reach the best achievable accuracy, which is the appropriate design for an energy-scaling comparison, because a longer training protocol would conflate the energy question with the separate question of how long each model needs to converge. Full fine-tuning of all parameters, rather than freezing the backbone, was used so that the measured energy reflected the entire network and the comparison was not distorted by treating the largest model's backbone as a fixed feature extractor, which would have understated the energy cost of the larger models relative to the smaller ones. The protocol therefore represents a middle ground between the two common transfer-learning variants, feature extraction and full fine-tuning, and it is the variant in which the energy difference between architectures is most directly comparable.

Energy and carbon measurement

Energy and carbon emissions were measured with the open-source CodeCarbon library, which samples hardware power draw during the training loop and converts it to energy in kilowatt-hours and to carbon dioxide equivalents using a regional electricity grid intensity.^(11,22) A fresh emissions tracker was started immediately before each training loop and stopped immediately after, so the measured energy corresponds to the fine-tuning phase alone and excludes data loading and model setup, which keeps the comparison clean but means the reported energy is a lower bound on the full workflow that includes data preparation and model serialization. Accuracy was computed on the held-out test set after each run. We report energy directly, rather than only the derived carbon figure, so that the result can be re-expressed for any electricity grid, and because energy is the quantity most directly controlled by a developer's hardware and architecture choices, whereas the carbon conversion depends on a grid mix that is outside the developer's control.⁽²³⁾ The measurement boundary and the energy-only reporting were chosen deliberately to make the result portable and reproducible across settings, so that a reader in a different country or on different hardware can compare their own measurements against ours on the energy axis without the grid mix obscuring the comparison.

Efficiency metric

To compare models on a like-for-like basis, we computed accuracy per unit energy as the held-out accuracy divided by the training energy, which expresses how much classification performance a unit of energy buys. Because the absolute energy of a single small training run is tiny, the primary outputs are the relative scaling of energy with model size and the efficiency ranking across models rather than the absolute carbon footprint. This metric follows the Green AI convention of reporting an efficiency ratio, such as accuracy per unit of computation or energy, alongside accuracy itself, on the grounds that a model should be judged not only by what it achieves but by what it costs to achieve it.⁽⁹⁾ The metric is intentionally simple and interpretable: it is the slope of the accuracy-versus-energy curve, and a higher value means more diagnostic performance for the same energy budget. Because the four models are trained on the same data under the same protocol, differences in this metric are attributable to the architectures alone, which makes the efficiency ranking a property of model choice rather than of the experimental setup.

RESULTS

Energy and carbon scale with model size

Measured energy increased monotonically with model size, from 0.26 milliwatt-hours for MobileNetV2 to 0.86 milliwatt-hours for ResNet50, a 3.30-fold increase across a tenfold range in parameters (Table 1, Figure 1). Estimated carbon emissions followed the same pattern and remained below one milligram of carbon dioxide equivalent in every case, because the absolute workload was small. The scaling was not linear with parameter count: ResNet50 used more than three times the energy of ResNet18 for only about twice the parameters, a consequence of the deeper bottleneck architecture's higher per-parameter computational cost, and this nonlinearity is itself a reason to prefer measured energy over parameter count as the quantity to track. The energy differences were therefore driven more by the computational structure of each architecture than by the raw parameter count alone, which is why parameter count, while a convenient proxy for model size, is not by itself a complete or reliable measure of the energy cost. The monotonic increase in energy

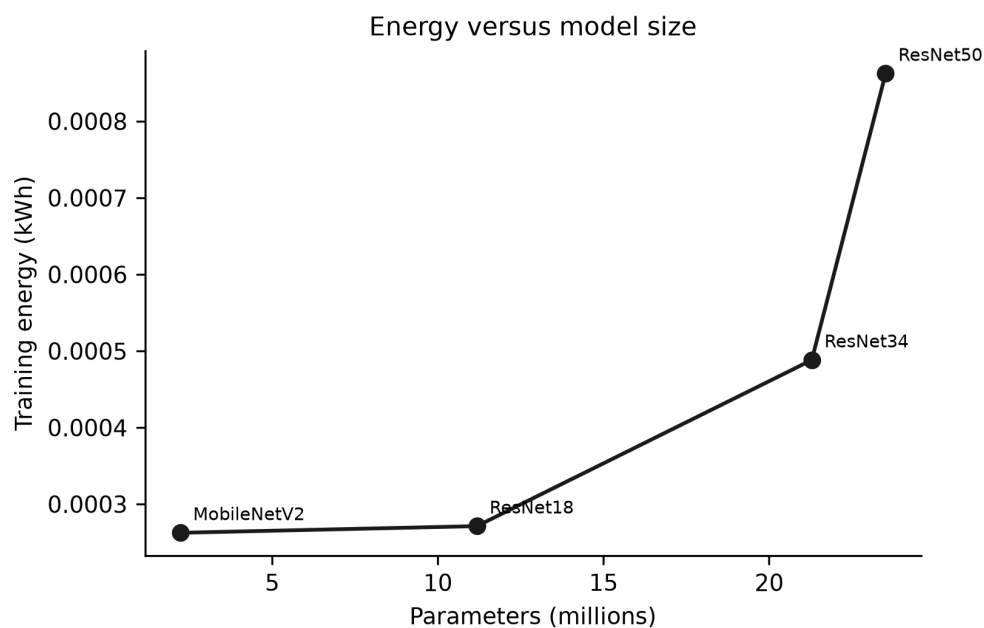
with size, despite the non-monotonic pattern in accuracy, is the core contrast of the study, and it is this contrast, rather than either quantity alone, that makes the efficiency comparison informative.

Table 1. Parameters, training energy, emissions and accuracy by model.

Model	Parameters (millions)	Energy (mWh)	CO2e (mg)	Accuracy (%)
MobileNetV2	2.2	0.26	0.09	56.80
ResNet18	11.2	0.27	0.10	62.20
ResNet34	21.3	0.49	0.17	56.80
ResNet50	23.5	0.86	0.31	62.20

Note: mWh = milliwatt-hours; CO2e = carbon dioxide equivalents. Accuracy is on a held-out test set. Source: Fine-tuning on 182 DENTEX tooth crops with CodeCarbon measurement.

Figure 1. Training energy versus model size.



Source: CodeCarbon measurement of four fine-tuning runs.

Accuracy does not rise with size

Accuracy did not increase with model size in this particular task. ResNet18 reached 62.20% accuracy, and ResNet50 reached the same 62.20%, while MobileNetV2 and ResNet34 both reached 56.80% (Table 1). The two largest models therefore delivered no accuracy advantage over the second-smallest, and ResNet34 was no better than the smallest model despite using nearly twice its energy. The accuracy ceiling of roughly 62.20% was set by the small, class-imbalanced dataset rather than by model capacity, which is why additional parameters did not help. This saturation is the key empirical fact behind the efficiency result, because it means the energy spent on larger models was spent for nothing, and it is a pattern that a developer should expect whenever a dataset is the limiting factor rather than the model. In such a setting, the rational strategy is not to buy more capacity but to spend the same energy on better data, on augmentation, or on a better validation protocol, all of which address the actual bottleneck rather than the apparent one.

Accuracy per unit energy favors the smaller adequate model

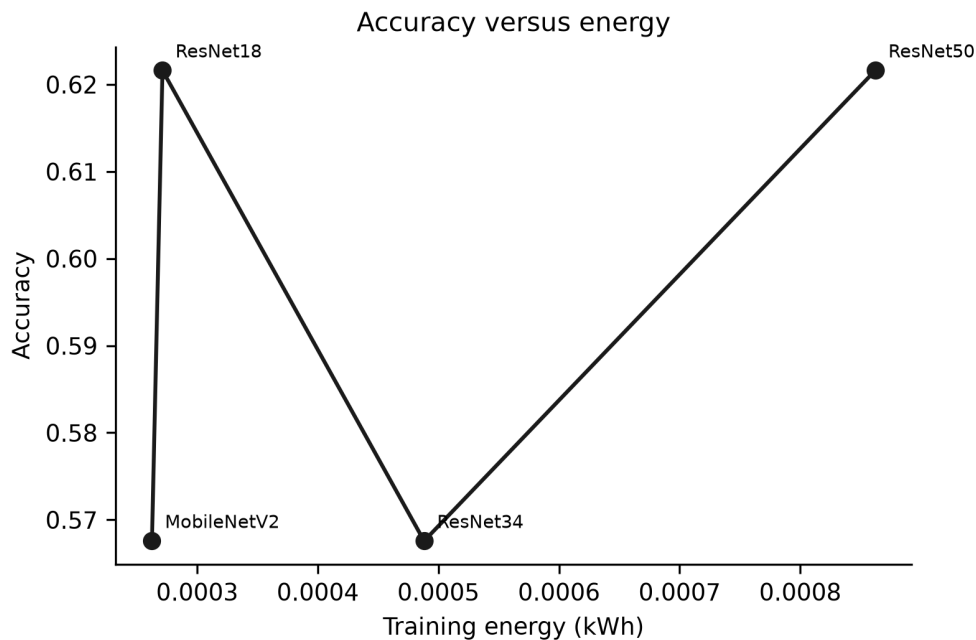
ResNet18 achieved the highest accuracy per unit energy, followed closely by MobileNetV2, while the larger models fell sharply (Table 2, Figure 2). The efficiency ranking was the practical output of the study: the smallest model that reached the accuracy ceiling, ResNet18, was also the most efficient, whereas ResNet50 spent roughly three times the energy of ResNet18 to reach the same accuracy and therefore offered about one third of its accuracy per unit energy. Expressed the other way around, a developer who chose ResNet50 over ResNet18 paid a threefold energy premium for zero accuracy gain, which is the concrete waste that the efficiency metric is designed to expose. ResNet34 was the least attractive option overall, combining the low accuracy of MobileNetV2 with a substantially higher energy cost, and it illustrates the risk of selecting an architecture by habit rather than by measurement. The clear separation between the efficient and inefficient models, a threefold range in accuracy per unit energy, is large enough to be operationally meaningful even though the absolute energy values are small, because the same ratio would hold and would simply scale upward if the training were repeated on more data, for more epochs, or across more candidate architectures.

Table 2. Accuracy per unit energy by model.

Model	Accuracy per kWh
MobileNetV2	2162
ResNet18	2292
ResNet34	1163
ResNet50	721

Note: Higher values indicate more classification performance per unit of energy. Source: Held-out accuracy divided by CodeCarbon-measured training energy.

Figure 2. Accuracy versus training energy.



Source: CodeCarbon measurement of four fine-tuning runs.

DISCUSSION

The practical finding is that, for dental image classification at this scale, larger models add energy and carbon without adding accuracy, and the smallest adequate model is the efficient choice. ResNet50 consumed more than three times the energy of ResNet18 to reach the same 62.20% accuracy, and ResNet34 consumed nearly twice the energy of MobileNetV2 for no accuracy gain. These are the two concrete comparisons that make the waste visible, and they are the ones a decision-maker should retain. The conclusion a developer can act on is direct: when fine-tuning a dental radiograph classifier, begin with the smallest model that reaches the task's accuracy ceiling, and move to a larger architecture only if a measured, demonstrable accuracy shortfall justifies the additional energy.

This result is consistent with the broader energy-aware machine-learning literature, which has established that model size and training cost scale together and that efficiency should be reported alongside accuracy.(6,9,10) Comparative studies have repeatedly found that the largest model is not automatically the best, and that architecture choice can change energy cost by orders of magnitude for the same performance.(12,24) Software-level measurement studies have likewise shown that the choice of model and hardware, rather than the choice of algorithm alone, dominates the energy budget of a training run, and that accurate energy accounting requires care in where the measurement boundary is drawn.(22,25) Our contribution is to bring this measurement to dental imaging specifically, where it had not been made, and to express the result in the terms a dental developer actually faces: which pretrained model should they fine-tune.(26,27) The threefold range in accuracy per unit energy that we observed is of the same character as the larger efficiency ranges reported in the general literature, which supports confidence that the measurement, despite its small absolute scale, is capturing a real and general phenomenon rather than noise. The comparison to other medical imaging fields is also instructive: just as radiologists and pathologists have begun to weigh the energy cost of their increasingly large models, dental researchers can apply the same reasoning to the transfer-learning workflows that dominate their field, where the cost is smaller in absolute terms but the same principle of spending energy only for a return applies.(15,26)

The observed saturation of accuracy is informative in its own right. With only 145 training teeth across four imbalanced classes, the dataset, not the model, was the binding constraint, and adding parameters beyond a modest threshold simply spent energy without returning performance. This is a common situation in dental AI, where annotated clinical images are scarce relative to the large natural-image datasets on which the pretrained models were trained, and it implies that the efficiency advantage of small models is likely to generalize to other dental tasks with comparably small labeled datasets. For tasks with much larger datasets, where a bigger model might genuinely use the extra capacity, the efficiency ranking could shift, and our result should not be read as a claim that larger models never help, only that their help must be demonstrated rather than assumed. The non-monotonic accuracy across sizes, with ResNet18 and ResNet50 outperforming MobileNetV2 and ResNet34, is also consistent with the known fact that depth and width interact in complex ways with the difficulty of the task and the amount of data, and it reinforces the recommendation that a developer should measure accuracy and energy for their specific task rather than rely on a general rule about which architecture is best. The fact that ResNet34 underperformed ResNet18 despite being strictly deeper is a reminder that architectural changes do not always translate into better results on a small dataset, and it is exactly the kind of outcome that only measurement can reveal.

The absolute carbon footprint measured here was tiny, fractions of a milligram of carbon dioxide equivalent per training run, because a four-epoch fine-tune on fewer than two hundred images is a small workload. We emphasize this so the result is not over-read: the study is about relative efficiency, not about claiming that dental AI is a major carbon emitter. The relevance lies in the scaling relationship and the efficiency ranking, which hold regardless of the absolute scale, and which would magnify in proportion if a developer trained for longer, on more data, or across many candidate architectures, as is typical during model development and hyperparameter search.(7,28) A realistic development workflow that sweeps over many architectures and hyperparameter settings, each run repeated several times for stability, would multiply the single-run energy measured here by a large factor, and it is precisely in that aggregate that the choice of an efficient architecture compounds into a meaningful saving. The single-run measurement is therefore best read as a unit cost that scales linearly with the amount of development work undertaken.

Several limitations should be weighed, each with its remedy. First, the dataset is a single validation split with 182 teeth, and the held-out set of 37 teeth is small, so the accuracy estimates are noisy; a larger, multi-institution dataset would stabilize the efficiency ranking and would test whether the accuracy ceiling observed here is a property of this dataset or of the task. Second,

energy was measured on a single CPU, and absolute values would differ on a GPU or on different hardware, though the relative scaling among models is a property of the architectures and should persist; future work should repeat the measurement on the hardware where dental models are actually trained, and report both the CPU and GPU numbers. Third, only four epochs were run, so the comparison reflects a fixed training budget rather than convergence to each model's best accuracy; a longer protocol would test whether larger models eventually earn their cost, though it would also raise the absolute energy and could be expected to widen, not narrow, the efficiency gap. Fourth, the carbon estimates depend on the assumed grid intensity, which varies by region, and we report energy alongside carbon so that the result can be re-expressed for any grid. Fifth, only four architectures were compared, and the efficiency frontier could be refined with more models and with distillation or pruning, which are themselves energy-relevant techniques that can compress a large model into a smaller one after training.(21)

For edge deployment, where a model runs on a clinic's own hardware rather than in a data center, the finding has a second, related implication: the same architecture that is efficient to train is also likely to be efficient to run, and a model that is needlessly large carries a permanent inference cost that accrues with every patient image processed, long after training has finished. A classifier that examines hundreds of patient radiographs a day will, over its lifetime, spend far more energy on inference than it ever did on training, so an architecture chosen for training efficiency tends to remain the efficient choice in production, and the training-time measurement reported here is therefore a reasonable proxy for the lifetime ranking of the models. This is especially relevant for the sustainability framing that now accompanies clinical AI, in which the environmental footprint of a deployed system is part of its overall assessment rather than an afterthought.(13) The training energy measured here is a lower bound on the model's lifetime energy, because a deployed classifier is used many more times than it is trained, and choosing an efficient architecture therefore compounds its savings over the life of the system. This connection between training efficiency and deployment efficiency is worth stating explicitly, because a developer who optimizes only for training cost may still leave an avoidable inference cost on the table.

The recommendation that follows from these measurements is that dental AI development should treat energy as a first-class, reported quantity alongside accuracy, as the Green AI literature has urged for machine learning generally.(9,13) A developer who fine-tunes a ResNet50 for a task that a ResNet18 performs equally well is spending energy and carbon for no clinical return, and the only reason such a choice goes unnoticed is that the cost is not measured. Reporting two numbers, accuracy and energy, would make efficient model selection a matter of routine rather than of guesswork, and would let the dental community demonstrate that its adoption of AI is sustainable rather than merely fast.(29,30) The measurement itself is inexpensive to add, requiring only an open-source library such as CodeCarbon and a few lines of instrumentation around the training loop, so the barrier to making energy reporting routine is low relative to the information that the measurement provides. A practical decision rule follows from the data: for a small, class-imbalanced dental dataset, begin with ResNet18 or a comparable small network, measure its accuracy, and only escalate to a larger model if the measured accuracy is inadequate and the expected gain justifies the additional energy. On the evidence of this study, a ResNet50 is not the right default for such a task, and a developer should be able to say what it bought them before choosing it. The rule is deliberately conservative, in the sense that it treats the smaller model as the default and requires a justification for anything larger, which inverts the common practice of defaulting to the largest available model. This inversion is the practical substance of the Green AI principle applied to a specific clinical domain, and it costs nothing at all to adopt once energy is measured. Future work could extend this measurement to a wider range of architectures, to the inference phase where the recurring energy cost lives, and to the effects of model compression techniques such as pruning and quantization, all of which would sharpen the efficiency frontier that a developer can use to make an informed choice, and would build toward a standing benchmark of energy efficiency for dental AI that future models can be judged against.

CONCLUSIONS

The energy cost of training dental image classifiers increased substantially and monotonically from the smallest to the largest model tested, while accuracy did not rise with size beyond a ceiling reached already by ResNet18, which was also the most efficient model in accuracy per unit energy. For dental radiograph classification at this scale, larger models add energy and carbon without adding accuracy, and the smallest adequate model is the efficient choice. Energy should

be reported alongside accuracy in dental AI development so that model selection can weigh diagnostic performance against environmental cost, and a developer should begin with a small model and justify any move to a larger one by a measured accuracy gain rather than by habit. The measurement itself is cheap to perform, and the information it yields is what turns an environmental consideration into an ordinary engineering criterion.

REFERENCES

1. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiology*. 2020;49(1):20190107. doi: 10.1259/dmfr.20190107
2. Moharrami M, Farmer J, Singhal S, Watson E, Glogauer M, Johnson AEW, et al. Detecting dental caries on oral photographs using artificial intelligence: A systematic review. *Oral Diseases*. 2024;30(4):1765-1783. doi: 10.1111/odi.14659
3. Ba-Hattab R, Barhom N, Osman S, Naceur I, Odeh A, Asad A, et al. Detection of Periapical Lesions on Panoramic Radiographs Using Deep Learning. *Applied Sciences*. 2023;13(3):1516. doi: 10.3390/app13031516
4. Department of Orthodontics, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India, Batra P, Tagra H, HCL Technologies Ltd, New Delhi, India, Katyal S, Department of Orthodontics, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India. Artificial Intelligence in Teledentistry. *Discoveries*. 2022;10(3):e153. doi: 10.15190/d.2022.12
5. Moeini A, Torabi S. The Role of Artificial Intelligence in Dental Diagnosis and Treatment Planning. *JODHN*. 2025;2(1):14-26. doi: 10.61838/kman.jodhn.2.1.2
6. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019:3645-3650. doi: 10.18653/v1/p19-1355
7. Patterson D, Gonzalez J, Le Q, Liang C. Carbon emissions and large neural network training [Internet]. 2021. Available from: <https://arxiv.org/abs/2104.10350>
8. Sheng C, Wang L, Huang Z, Wang T, Guo Y, Hou W, et al. Transformer-Based Deep Learning Network for Tooth Segmentation on Panoramic Radiographs. *J Syst Sci Complex*. 2023;36(1):257-272. doi: 10.1007/s11424-022-2057-9
9. Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. *Commun ACM*. 2020;63(12):54-63. doi: 10.1145/3381831
10. Lacoste A, Luccioni A, Schmidt V, Dandres T. Quantifying the carbon emissions of machine learning [Internet]. 2019. Available from: <https://arxiv.org/abs/1910.09700>
11. Budenny SA, Lazarev VD, Zakharenko NN, Korovin AN, Plosskaya OA, Dimitrov DV, et al. eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Dokl Math*. 2022;106(S1):S118-S128. doi: 10.1134/s1064562422060230
12. Argerich MF, Patiño-Martínez M. Measuring and Improving the Energy Efficiency of Large Language Models Inference. *IEEE Access*. 2024;12:80194-80207. doi: 10.1109/access.2024.3409745
13. Lokmic-Tomkins Z, Davies S, Block LJ, Cochrane L, Dorin A, von Gerich H, et al. Assessing the carbon footprint of digital health interventions: a scoping review. *J Am Med Inform Assoc*. 2022;29(12):2128-2139. doi: 10.1093/jamia/ocac196
14. Bignami E, Darhour LJ, Bellini V. Sustainable AI in medicine: navigating innovation, challenges, and environmental impact. *Health Econ Rev*. 2025;15(1):110. doi: 10.1186/s13561-025-00704-w
15. Kocak B, Ponsiglione A, Romeo V, Ugga L, Huisman M, Cuocolo R. Radiology AI and sustainability paradox: environmental, economic, and social dimensions. *Insights Imaging*. 2025;16(1):88. doi: 10.1186/s13244-025-01962-2
16. Chen IDS, Yang CM, Chen MJ, Chen MC, Weng RM, Yeh CH. Deep Learning-Based Recognition of Periodontitis and Dental Caries in Dental X-ray Images. *Bioengineering*. 2023;10(8):911. doi: 10.3390/bioengineering10080911
17. Liu RW, Ong W, Makmur A, Kumar N, Low XZ, Shuliang G, et al. Application of Artificial Intelligence Methods on Osteoporosis Classification with Radiographs—A Systematic Review. *Bioengineering*. 2024;11(5):484. doi: 10.3390/bioengineering11050484
18. Mao YC, Huang YC, Chen TY, Li KC, Lin YJ, Liu YL, et al. Deep Learning for Dental Diagnosis: A Novel Approach to Furcation Involvement Detection on Periapical Radiographs. *Bioengineering*. 2023;10(7):802. doi: 10.3390/bioengineering10070802
19. Alharbi SS, Alhasson HF. Exploring the Applications of Artificial Intelligence in Dental Image Detection: A Systematic Review. *Diagnostics*. 2024;14(21):2442. doi: 10.3390/diagnostics14212442
20. Çelik B, Mikaeili M, Genç MZ, Çelik ME. Can super resolution via deep learning improve classification accuracy in dental radiography?. *Dentomaxillofacial Radiology*. 2025;54(6):473-487. doi: 10.1093/dmfr/twaf029
21. Rawat R. Optimizing ML models for cybercrime detection: balancing performance, energy consumption, and carbon footprint through multi-objective optimization. *Discov Artif Intell*. 2025;5(1):38. doi: 10.1007/s44163-025-00262-4
22. Raffin G, Trystram D. Dissecting the Software-Based Measurement of CPU Energy Consumption: A Comparative Analysis. *IEEE Trans Parallel Distrib Syst*. 2025;36(1):96-107. doi: 10.1109/tpds.2024.3492336
23. Guldner A, Bender R, Calero C, Fernando GS, Funke M, Gröger J, et al. Development and evaluation of a reference measurement model for assessing the resource and energy efficiency of software products and components—Green Software Measurement Model (GSMM). *Future Generation Computer Systems*. 2024;155:402-418. doi: 10.1016/j.future.2024.01.033
24. Rodriguez C, Chevalier M, Tran H, Truong K. Evaluating the energy consumption of machine learning: systematic literature review and experiments [Internet]. 2024. Available from: <https://inria.hal.science/hal-04762843>
25. Rajput S, Widmayer T, Shang Z, Kechagia M, Sarro F, Sharma T. Enhancing Energy-Awareness in Deep Learning through Fine-Grained Energy Measurement. *ACM Trans Softw Eng Methodol*. 2024;33(8):1-34. doi: 10.1145/3680470
26. Ezhov M, Gusarev M, Golitsyna M, Yates JM, Kushnerev E, Tamimi D, et al. Clinically applicable artificial intelligence system for dental diagnosis with CBCT. *Sci Rep*. 2021;11(1):15006. doi: 10.1038/s41598-021-94093-9

27. Sadr H, Nazari M, Khodaverdian Z, Farzan R, Yousefzadeh-Chabok S, Ashoobi MT, et al. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *Eur J Med Res*. 2025;30(1):418. doi: 10.1186/s40001-025-02680-7
28. Hasan SM, Shahid AR, Imteaj A. Evaluating Sustainability and Social Costs of Adversarial Training in Machine Learning. *IEEE Consumer Electron Mag*. 2025;14(3):27-34. doi: 10.1109/mce.2024.3458350
29. Wong YL, Aslam TM, Chang DF, Erny B. Sustaining Ophthalmic Practices for the Future: A High-Value Care Approach to Environmental Responsibility. *Ophthalmol Ther*. 2025;14(6):1199-1218. doi: 10.1007/s40123-025-01146-7
30. Bachina L, Kanagala A, Korapu S, Ratnaraju P. Sustainable materials for artificial intelligence (AI) technology adoption for energy-efficient patient-centric healthcare solutions. *J Educ Health Promot*. 2025;14(1):4. doi: 10.4103/jehp.jehp_527_24

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Leodanis Hernández Cabrera, Maikro Osvaldo Chávez Moya, Neisy Pérez Ramos.

Writing–review and editing: Leodanis Hernández Cabrera, Maikro Osvaldo Chávez Moya, Neisy Pérez Ramos.