

Enhancing Clinical Decision Support Through Explainable AI: A Comparative Study of CNN and Vision Transformer Models for Skull Fracture Detection in CT Scans

R. A. S. A. Ranmuthu

Department of Statistics and Computer Science, University of Peradeniya

Saluka R. Kodituwakku

Department of Statistics and Computer Science, University of Peradeniya

Lakshika S. Nawarathna

lakshikas@sci.pdn.ac.lk

Department of Statistics and Computer Science, University of Peradeniya

Research Article

Keywords: Attention Rollout, convolutional neural networks, DeiT-Tiny, explainable artificial intelligence, Grad-CAM, ResNet-50, skull fracture detection, transfer learning, traumatic brain injury, vision transformers

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10478728/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Skull fractures are critical indicators of traumatic brain injury (TBI), yet their detection on computed tomography (CT) scans remains challenging due to overlapping anatomical structures and high radiologist workload. Though Deep learning models offer automated detection potential their clinical adoption is limited by opaque decision-making.

Methods

A convolutional neural network (ResNet-50) and a data-efficient vision transformer (DeiT-Tiny) for slice-level skull fracture classification in CT scans were compared. Axial CT slices from 15 patients (4,914 CT slices) were collected under ethical approval at the National Hospital, Kandy, Sri Lanka. Preprocessing included Hounsfield unit windowing and augmentation of fractured slices. Both models were trained with ImageNet-pretrained weights and evaluated using a patient-level train-validation-test split. Explainability was assessed using Grad-CAM (ResNet-50) and Attention Rollout (DeiT-Tiny), with quantitative alignment measured against radiologist-annotated fracture regions.

Results

ResNet-50 achieved superior classification performance (accuracy 0.726, AUC-ROC 0.800, F1-score 0.428) compared to DeiT-Tiny (accuracy 0.673, AUC-ROC 0.746, F1-score 0.384). The AUC difference was statistically significant (mean difference = 0.054; 95% CI: 0.035–0.074; $p < 0.001$). DeiT-Tiny demonstrated marginally better spatial alignment in explainability maps (mean IoU 0.021 vs 0.006; mean Dice 0.041 vs 0.012), though absolute values remained low.

Conclusions

ResNet-50 provides statistically robust classification performance under limited-data conditions, while DeiT-Tiny offers modest advantages in the spatial coherence of visual explanations. These findings highlight a trade-off between predictive accuracy and interpretability in deep learning-based clinical decision support. Future work should focus on multi-centre validation, patient-level classification, and improved explainability frameworks for small medical imaging datasets.

Background

Traumatic brain injury (TBI) is a leading cause of morbidity, disability, and death worldwide, with particularly high incidence in low- and middle-income countries where road traffic accidents and falls predominate [1]. Skull fractures are clinically important markers of TBI severity, frequently associated

with intracranial haemorrhage and cerebrospinal fluid leak, and their early detection is critical for guiding triage and treatment decisions [2, 3]. Computed tomography (CT) remains the gold standard for acute head trauma assessment due to its speed, wide availability, and sensitivity to acute intracranial pathology [4].

Despite its established role, CT-based fracture detection carries a significant risk of missed or delayed diagnosis. Reported missed-diagnosis rates for mild TBI vary widely, from approximately 7% for isolated facial fractures to over 77% in general emergency department populations, depending on injury type, clinical setting, and radiologist workload [5–8]. Subtle linear fractures, whose appearance can closely mimic normal anatomical structures such as sutures and vascular grooves, are particularly prone to oversight [9].

Deep learning, and convolutional neural networks (CNNs) in particular, has demonstrated strong performance in medical image analysis across a range of tasks including fracture detection [10]. ResNet-50, with its residual skip-connection architecture, has emerged as a widely adopted baseline for medical imaging classification [11]. However, CNNs primarily process images in small, localized regions, making it difficult for them to capture long-distance spatial connections; a capability that may be relevant for detecting fractures extending across large portions of the skull surface.

Vision Transformers (ViTs), which process images as sequences of patches and model global context through self-attention mechanisms, represent an alternative paradigm [12]. Data-efficient variants such as DeiT-Tiny, which incorporates knowledge distillation to reduce data requirements, have extended the practical applicability of transformers to smaller datasets [13]. Nonetheless, their relative merits for skull fracture detection under limited-data clinical conditions remain unexplored.

A key barrier to clinical deployment of deep learning models is their lack of interpretability. Explainable AI (XAI) methods that produce visual saliency maps, such as Gradient-weighted Class Activation Mapping (Grad-CAM) for CNNs and Attention Rollout for ViTs, offer a means of interrogating model decision-making and verifying alignment with clinical anatomy [14, 15]. Prior work has highlighted both the promise and the limitations of these approaches in medical imaging contexts [16].

This study directly addresses the lack of quantitative explainability benchmarks comparing CNNs and data-efficient Vision Transformers (ViTs) on small clinical datasets by providing a rigorous head-to-head comparison of ResNet-50 and DeiT-Tiny for slice-level skull fracture classification in CT images, using a small ethically approved clinical dataset from Sri Lanka. Performance is evaluated across standard classification metrics with statistical significance testing, and explainability quality is assessed quantitatively using radiologist-annotated ground truth. To our knowledge, few studies have directly compared CNN and ViT architectures for skull fracture detection with quantitative explainability evaluation against radiologist annotations. This study makes four main contributions: (1) a direct comparison between a CNN (ResNet-50) and a Vision Transformer (DeiT-Tiny), (2) statistical comparison of model performance, (3) quantitative evaluation of explainability using radiologist annotations and (4) evaluation on a real-world clinical dataset from Sri Lanka.

Methods

Ethics and Dataset

Axial CT scan slices were collected from 15 patients presenting with head trauma at the National Hospital, Kandy, Sri Lanka, following ethical approval from the hospital ethics review committee (Ethical Clearance No: NHK/ERC/30/2025). All procedures were conducted in compliance with the National Guidelines for the Establishment and Functioning of Ethical Review Committees in Health Care Institutions in Sri Lanka (2013). Individual patient consent was waived by the ethics committee as this was a retrospective analysis of anonymised imaging data.

Image Preprocessing

All CT images were processed from Digital Imaging and Communications in Medicine (DICOM) format using the PyDICOM library (v3.0.1). Pixel intensities were converted to Hounsfield Units (HU) using rescale slope and intercept parameters extracted from DICOM headers. A bone window transformation (window centre: 300 HU; window width: 1,500 HU) was applied to enhance visualisation of cortical bone structures [17]. Bone window settings were selected because cortical bone structures are most clearly visualised within this window, facilitating fracture identification.

Images were subsequently normalised to the range $[0, 1]$, resized to 224×224 pixels using area interpolation, and converted to three-channel format by channel replication to ensure compatibility with ImageNet-pretrained models. For compatibility with ImageNet-pretrained models, input images were standardised using channel-wise mean and standard deviation values derived from the ImageNet dataset.

Data Augmentation

To mitigate class imbalance (approximately 2:1 normal-to-fractured ratio) and improve model generalisation, data augmentation was applied exclusively to fractured slices in the training set. However, this may introduce distributional differences. Augmentation operations included random rotation ($\pm 10^\circ$), random brightness and contrast adjustment (factors 0.9–1.1), additive Gaussian noise ($\sigma = 5$), probabilistic Gaussian blur (3×3 kernel, $p = 0.5$), and horizontal flipping ($p = 0.5$). These transformations were selected to reflect realistic variation in scanner protocols, patient positioning, and image reconstruction parameters while preserving clinically relevant fracture features. Following augmentation, the training set contained 5,186 slices with balanced class distribution (2,593 per class). Augmentation was intentionally restricted to the minority class to alleviate class imbalance while avoiding unnecessary duplication of majority-class samples.

This strategy was designed to simulate realistic variability in imaging conditions while preserving clinically relevant fracture features. No augmentation was applied to validation or test data. Partitioning

was performed at the patient level to prevent data leakage between splits [19]. The 15 patients were stratified by fracture status and randomly divided into training, validation, and test sets, yielding 3,351 training, 648 validation, and 915 test slices. The test set was held out completely until final evaluation.

Model Architectures

ResNet-50

ResNet-50 [11] was selected as the CNN baseline. Pre-trained ImageNet weights were used for initialisation. The original 1,000-class output layer was replaced with a classification head comprising global average pooling, a dropout layer ($p = 0.5$), and a two-unit fully connected layer with SoftMax activation. Training proceeded in two phases: the classification head alone was trained for five epochs (convolutional layers frozen), followed by full fine-tuning with discriminative learning rates (convolutional layers trained at one-tenth the learning rate of the classification head). ResNet-50 consists of approximately 25.6 million parameters.

DeiT-Tiny

DeiT-Tiny [13] was selected as the vision transformer model. It comprises 12 transformer encoder layers with an embedding dimension of 192 and three self-attention heads, totalling approximately 5.7 million parameters. The model processes input images (224×224) as sequences of 196 non-overlapping 16×16 patches with learnable positional embeddings and a prepended CLS token for classification. The pre-trained classification layer was replaced with a two-unit output layer. The CLS token is heavily regularised throughout the network's depth using mechanisms built into the transformer block, such as Stochastic Depth or DropPath. An additional dropout layer at the final head is often unnecessary and sometimes detrimental to ViT s. All parameters were fully fine-tuned, as gradual unfreezing offers less benefit for transformer architectures [12].

Training Procedure

Both models were optimised using AdamW [22] with binary cross-entropy loss. Architecture-specific hyperparameters were tuned independently via grid search on the validation set. For ResNet-50, the optimal configuration was: learning rate = 1×10^{-4} , weight decay = 0.01, batch size = 16, dropout = 0.5. For DeiT-Tiny: learning rate = 5×10^{-4} , weight decay = 0.05, batch size = 16, attention dropout = 0.1, projection dropout = 0.1. Cosine annealing learning rate scheduling was applied [22]. Training ran for a maximum of 30 epochs with early stopping triggered by absence of improvement in validation loss for five consecutive epochs. All experiments were conducted on an NVIDIA Tesla V100 GPU (Google Colab) with a fixed random seed (42).

Model Evaluation

Both models were evaluated on the held-out test set (915 slices; 723 normal, 192 fractured). Classification performance was quantified using accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Bootstrap resampling with 1,000 iterations was used to compute 95% confidence intervals for AUC-ROC. Statistical comparison between models was performed using McNemar's test for paired binary predictions and a bootstrap-based permutation test for AUC difference, with a significance threshold of $p < 0.05$.

Explainability Analysis

Grad-CAM [15] was used to generate saliency maps for ResNet-50 by computing gradients of the predicted class score with respect to the final convolutional layer's feature maps, then forming a class activation map by globally averaging those gradients and upsampling to image resolution. For DeiT-Tiny, Attention Rollout [14] was applied by recursively multiplying attention matrices from all transformer layers and extracting attention weights at the CLS token, which were then reshaped to a spatial heatmap.

To enable quantitative evaluation of explainability, a consultant radiologist annotated fracture locations on 30 fractured test slices using rectangular bounding boxes. These annotations were binarized to produce ground-truth fracture masks. For each model, the corresponding saliency map was thresholded at the median intensity value as a simple and model-agnostic binarisation strategy; however, alternative thresholding methods (e.g., Otsu, percentile-based) may yield different alignment results.. Spatial alignment with the ground-truth mask was computed using Intersection over Union (IoU) and Dice Similarity Coefficient (DSC).

Reporting Standards and Reproducibility

This study was conducted in accordance with emerging reporting recommendations for artificial intelligence in medical research, including principles from TRIPOD-AI. All methodological steps, including preprocessing, model configuration, and evaluation procedures, are described to facilitate reproducibility. Code and implementation details are available from the corresponding author upon reasonable request.

Results

Classification Performance

Performance metrics on the held-out test set are summarised in Table 1. ResNet-50 outperformed DeiT-Tiny across all evaluated metrics. ResNet-50 achieved an AUC-ROC of 0.800 compared to 0.746 for DeiT-Tiny, with non-overlapping 95% bootstrap confidence intervals (Table 1), providing an initial indication of statistically significant separation.

Table 1

Classification performance of ResNet-50 and DeiT-Tiny on the held-out test set. (Values shown are mean accuracy, precision, recall, F1-score, and AUC-ROC with 95% confidence intervals.)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC (95% Confidence Interval)	
ResNet-50	0.726	0.381	0.490	0.428	0.800	0.765–0.833
DeiT-Tiny	0.673	0.317	0.484	0.384	0.746	0.712–0.780

Despite moderate AUC, both models exhibited moderate recall (~ 0.49 for both), with a tendency toward false positives, a pattern consistent with the inherent difficulty of distinguishing subtle linear skull fractures from normal anatomical variants such as sutures and vascular channels. DeiT-Tiny produced a higher false-positive rate than ResNet-50, resulting in lower specificity. Confusion matrix data for ResNet-50 (570 true negatives, 94 true positives, 153 false positives, 98 false negatives) and DeiT-Tiny (523 true negatives, 93 true positives, 200 false positives, 99 false negatives) reflect this pattern (Fig. 1).

Training Dynamics

ResNet-50 converged rapidly, with training loss stabilising below 0.01 and peak validation AUC of approximately 0.944 achieved at epochs 10–11, followed by moderate overfitting. DeiT-Tiny demonstrated a slightly higher peak validation AUC (≈ 0.967) at epochs 9–11 but exhibited more pronounced post-peak instability and a sharper decline in validation performance during extended training, consistent with the greater data requirements of transformer architectures [12, Fig. 2].

Statistical Significance Testing

McNemar's test confirmed a highly significant difference in classification behaviour between the two models ($\chi^2 = 14.16$; $p = 0.00017$). Bootstrap-based AUC difference testing yielded a mean AUC difference of 0.054 (95% CI: 0.035–0.074; $p < 0.001$), confirming that the performance advantage of ResNet-50 is statistically robust and unlikely to reflect sampling variation.

Explainability and Spatial Alignment

Quantitative alignment metrics between saliency maps and radiologist-annotated fracture regions are presented in Table 2 and Fig. 4.

Table 2

Partial alignment of explainability maps with radiologist-annotated fracture regions. (Mean Intersection over Union [IoU] and Dice Similarity Coefficient [DSC] are reported for ResNet-50 (Grad-CAM) and DeiT-Tiny (Attention Rollout).)

Model (Method)	Mean IoU	Mean Dice Coefficient
ResNet-50 (Grad-CAM)	0.006	0.012
DeiT-Tiny (Attention Rollout)	0.021	0.041

DeiT-Tiny demonstrated approximately 3.5-fold higher mean IoU and Dice compared to ResNet-50. However, absolute alignment values were low for both models, which is consistent with the intrinsic properties of post-hoc explainability methods: Grad-CAM heatmaps and Attention Rollout maps are smooth, low-resolution visualisations that reflect regions of influence rather than precise fracture boundaries, making pixel-level overlap with bounding-box annotations an inherently conservative measure.

Visual inspection of saliency maps (Fig. 3) revealed that both methods frequently directed attention toward anatomically plausible regions in the vicinity of fracture sites and toward structural boundaries of the skull vault, suggesting that each model has learned clinically meaningful spatial patterns and partially align with clinically relevant regions.

Discussion

Key Findings

This study provides the first direct comparison of ResNet-50 and DeiT-Tiny for slice-level skull fracture detection in CT images under limited-data conditions. Three principal findings emerged: (1) ResNet-50 significantly outperformed DeiT-Tiny across all classification metrics, with a statistically robust advantage in AUC-ROC; (2) DeiT-Tiny produced explainability maps with modestly better spatial alignment to radiologist-annotated fracture regions; and (3) absolute alignment scores were low for both models, underscoring the current limitations of post-hoc explainability methods in medical imaging.

Clinical Implications

The superior classification performance of ResNet-50 reflects the inductive bias of convolutional architectures toward local spatial features, which is advantageous in small medical datasets [12, 13, 18]. However, the moderate sensitivity (~ 0.49) observed for both models limits their suitability as standalone diagnostic tools. Instead, these models may serve as assistive systems to flag suspicious slices or support radiologists in high-volume emergency settings [2, 5, 6]. DeiT-Tiny's relatively more coherent attention maps, although not precise enough for localisation, may enhance clinician trust by providing interpretable visual cues in decision support workflows [16, 23].

Limitations

This study has several limitations that should be acknowledged. First, the dataset comprised only 15 patients from a single clinical site, which restricts the generalisability of the findings across different scanners, imaging protocols, and patient populations [19]. Second, classification was performed at the slice level rather than the patient level, which does not fully reflect clinical diagnostic workflows where radiologists integrate findings across all slices of a scan [24]. Third, fracture regions were annotated using rectangular bounding boxes rather than pixel-precise masks, which constrained the accuracy of explainability evaluation and likely underestimated the true spatial alignment of both models. In addition, computational constraints limited the depth of hyperparameter optimisation, particularly for transformer

architectures, which may have affected the comparative performance of DeiT-Tiny. Finally, the models were evaluated only on internal data, and external validation on multi-centre datasets will be necessary before these approaches can be considered for clinical translation [19].

Future Directions

Future research should prioritise multi-centre data collection to improve generalisability [19], patient-level classification strategies to better reflect clinical workflows, and pixel-level fracture annotations for more precise evaluation of explainability [16]. Hybrid architectures that integrate convolutional and transformer components may help reconcile the observed trade-off between predictive accuracy and interpretability [25]. In addition, the development of dedicated explainability frameworks tailored to small medical imaging datasets will be critical for clinical translation [16, 23].

Conclusions

Summary of Findings

This study demonstrates that ResNet-50 provides statistically superior and more reliable skull fracture classification compared to DeiT-Tiny under limited-data conditions, consistent with its inductive bias toward local spatial features. In contrast, DeiT-Tiny offered modest advantages in the spatial coherence of its attention-based explainability maps, though absolute alignment values remained low.

Clinical Relevance

From a clinical perspective, the moderate sensitivity of both models (~ 0.49) constrains their use as standalone diagnostic tools. Nevertheless, they hold promise as assistive systems to support radiologists in high-volume trauma settings, where automated flagging of suspicious slices could reduce missed diagnoses and improve workflow efficiency. The observed performance–explainability trade-off underscores the importance of balancing predictive accuracy with interpretability to foster clinician trust in AI-driven decision support.

Future Outlook

Further research should focus on multi-centre validation, patient-level classification strategies, and pixel-level fracture annotations to enhance generalisability and localisation accuracy. Hybrid architectures that integrate convolutional and transformer components may help reconcile the trade-off between accuracy and interpretability. Dedicated explainability frameworks tailored to small medical imaging datasets will be essential for safe and effective clinical translation.

Abbreviations

AUC

Area under the receiver operating characteristic curve

CNN
Convolutional Neural Network
CT
Computed Tomography
DeiT
Data-efficient Image Transformer
DICOM
Digital Imaging and Communications in Medicine
DSC
Dice Similarity Coefficient
Grad-CAM
Gradient-weighted Class Activation Mapping
HU
Hounsfield Unit
IoU
Intersection over Union
TBI
Traumatic Brain Injury
ViT
Vision Transformer
XAI
Explainable Artificial Intelligence

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Review Committee of the National Hospital, Kandy, Sri Lanka (Ethical Clearance No: NHK/ERC/30/2025). All procedures were conducted in accordance with the National Guidelines for the Establishment and Functioning of Ethical Review Committees in Health Care Institutions in Sri Lanka (2013). Requirement for individual patient consent was waived by the ethics committee as all data were fully anonymised prior to analysis.

Consent for publication:

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors and was self-funded.

Author Contribution

RASAR conducted the literature review, collected and analyzed the data, and prepared the initial draft of the manuscript. LSN and SRK supervised the research, provided guidance on the study design and interpretation of the findings, critically reviewed and revised the manuscript, and contributed to improving the overall quality of the paper. All authors read and approved the final manuscript.

Acknowledgement

The authors sincerely thank Dr. Ganganath Rodrigo, Department of Radiology, National Hospital, Kandy, for his valuable contribution to the CT scan data collection and radiological annotation. The authors also express their gratitude to Dr. Palitha Ratnayake for his valuable support throughout this study.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

References

1. Maas AI, Menon DK, Manley GT, Abrams M, Åkerlund C, Andelic N, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. *Lancet Neurol*. 2022;21(11):1004–60. 10.1016/S1474-4422(22)00260-7.
2. Iacobellis F, Abu-Omar A, Crivelli P, Galluzzo M, Danzi R, Trinci M, et al. Current standards for and clinical impact of emergency radiology in major trauma. *Int J Environ Res Public Health*. 2022;19(1):539. 10.3390/ijerph19010539.
3. Hofman PA, Nelemans P, Kemerink GJ, Wilmlink JT. Value of radiological diagnosis of skull fracture in the management of mild head injury: meta-analysis. *J Neurol Neurosurg Psychiatry*. 2000;68(4):416–22. 10.1136/jnnp.68.4.416.
4. Brenner DJ, Hall EJ. Computed tomography; an increasing source of radiation exposure. *N Engl J Med*. 2007;357(22):2277–84. 10.1056/NEJMr072149.
5. Pozzato I, Meares S, Kifley A, Craig A, Gillett M, Vu KV, et al. Challenges in the acute identification of mild traumatic brain injuries: results from an emergency department surveillance study. *BMJ Open*. 2020;10(2):e034494. 10.1136/bmjopen-2019-034494.

6. Peixoto C, Buchanan D, Nahas R. Missed emergency department diagnosis of mild traumatic brain injury in patients with chronic pain after motor vehicle collision. *Pain Physician*. 2023;26(1):101–10. 10.36076/ppj.2023.26.101.
7. Sharma B, Bradbury CL, Mikulis D, Green R. Missed diagnosis of traumatic brain injury in patients with traumatic spinal cord injury. *J Rehabil Med*. 2014;46(4):370–3. 10.2340/16501977-1261.
8. Kokko L, Puolakkainen T, Thorén H, Piippo-Karjalainen A, Suominen A, Snäll J. Traumatic brain injury in patients with facial fracture – a challenge for the clinician? *Journal of Stomatology. Oral Maxillofacial Surg*. 2025;126(4):102302. 10.1016/j.jormas.2025.102302.
9. Jin KS, Lee H, Sohn JB, Han YS, Jung DU, Sim HY, et al. Fracture patterns and causes in the craniofacial region: an 8-year review of 2076 patients. *Maxillofacial Plast Reconstr Surg*. 2018;40(1):29. 10.1186/s40902-018-0168-y.
10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. 10.1016/j.media.2017.07.005.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016. pp. 770–778. 10.1109/CVPR.2016.90
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al. An image is worth 16×16 words: transformers for image recognition at scale. In: *Proceedings of the International Conference on Learning Representations (ICLR)*; 2021. 10.48550/arXiv.2010.11929
13. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*; 2021. Vol. 139, pp. 10347–10357. 10.48550/arXiv.2012.12877
14. Abnar S, Zuidema W. Quantifying attention flow in transformers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2020. pp. 4190–4197. 10.18653/v1/2020.acl-main.385
15. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; 2017. pp. 618–626. 10.1109/ICCV.2017.74
16. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: an overview for clinical practitioners – saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110787. 10.1016/j.ejrad.2023.110787.
17. Murphy A, Feger J, Ismail MA. Windowing (CT). *Radiopaedia.org*. 2017. <https://DOI.org/10.53347/RID-52108>
18. Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of neural network representations revisited. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*; 2019. Vol. 97, pp. 3519–3529.
19. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional

- study. PLoS Med. 2018;15(11):e1002683. 10.1371/journal.pmed.1002683.
20. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In: Proceedings of IJCAI*; 1995. Vol. 14, pp. 1137–1145.
 21. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. 2017;145(Pt B):166–79. 10.1016/j.neuroimage.2016.10.038.
 22. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. *In: Proceedings of the International Conference on Learning Representations (ICLR)*; 2019. 10.48550/arXiv.1711.05101
 23. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf Decis Mak*. 2020;20:310. 10.1186/s12911-020-01332-6.
 24. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Panykh OS, et al. Current applications and future impact of machine learning in radiology. *Radiology*. 2018;288(2):318–28. 10.1148/radiol.2018171820.
 25. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med Image Anal*. 2024;97:103280. 10.1016/j.media.2024.103280.

Figures

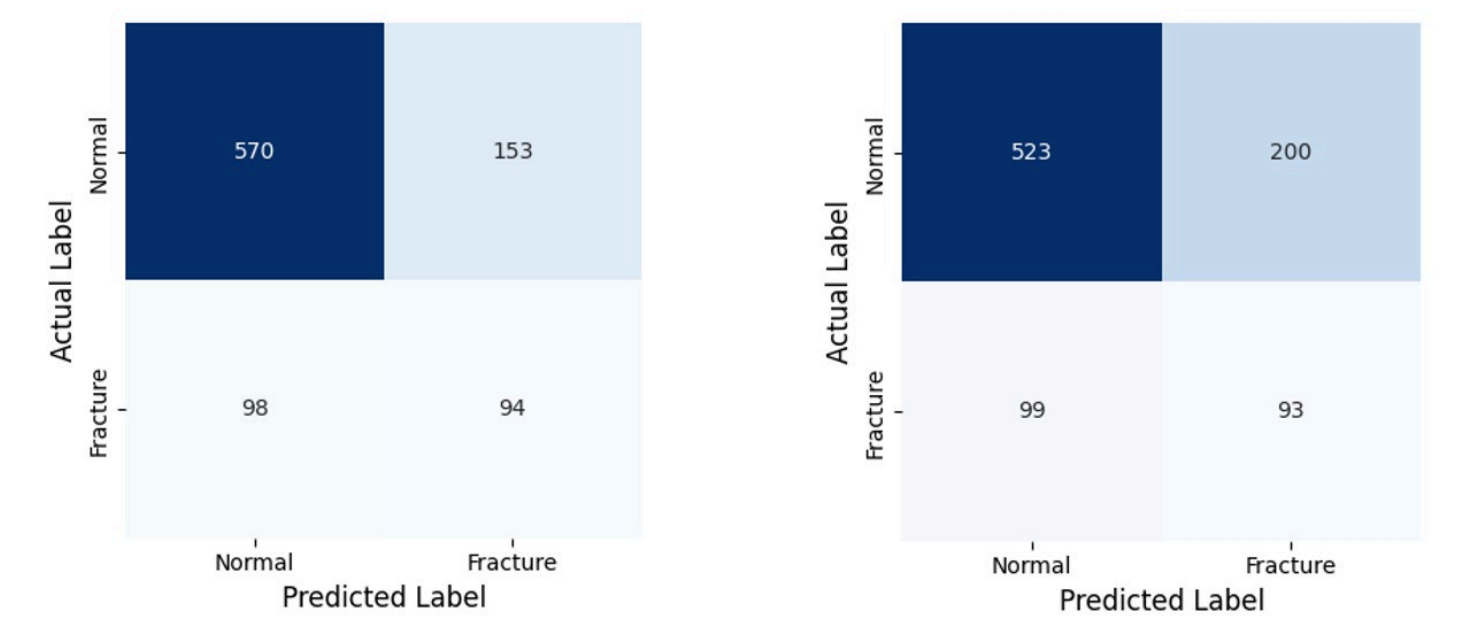
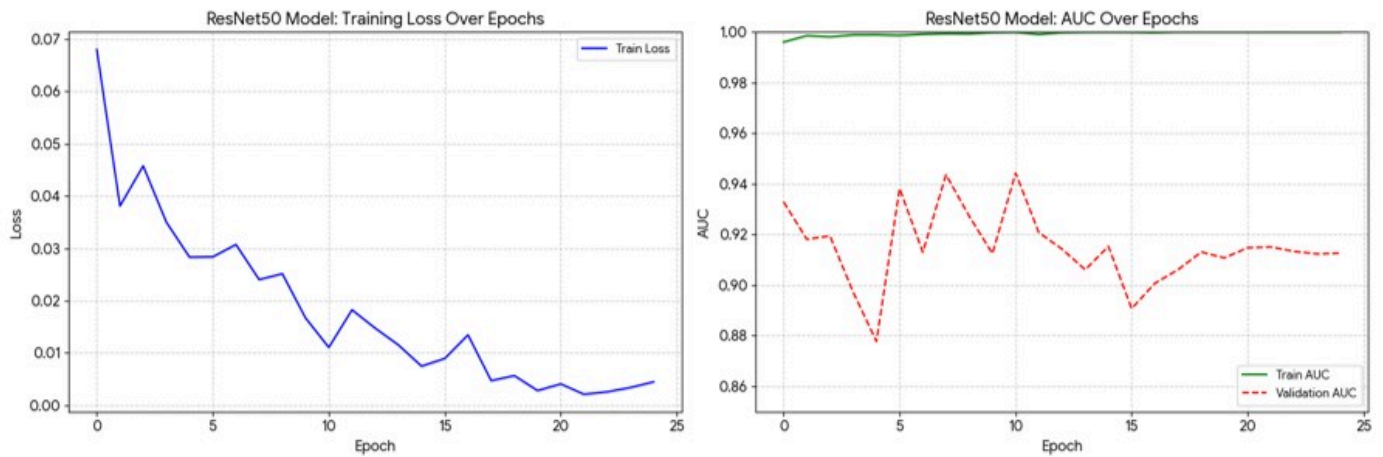
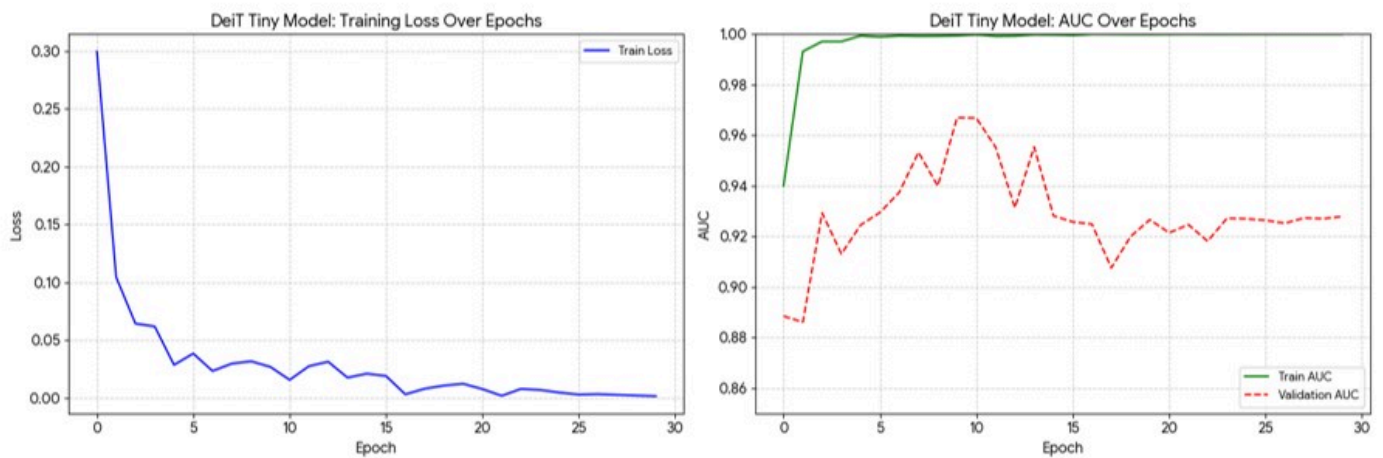


Figure 1

Confusion matrix results for ResNet-50 and DeiT-Tiny on the held-out test set.
(Left: ResNet-50; Right: DeiT-Tiny.)



(a) Model Training History of ResNet-50 Model



(b) Model Training History of DeiT-Tiny Model

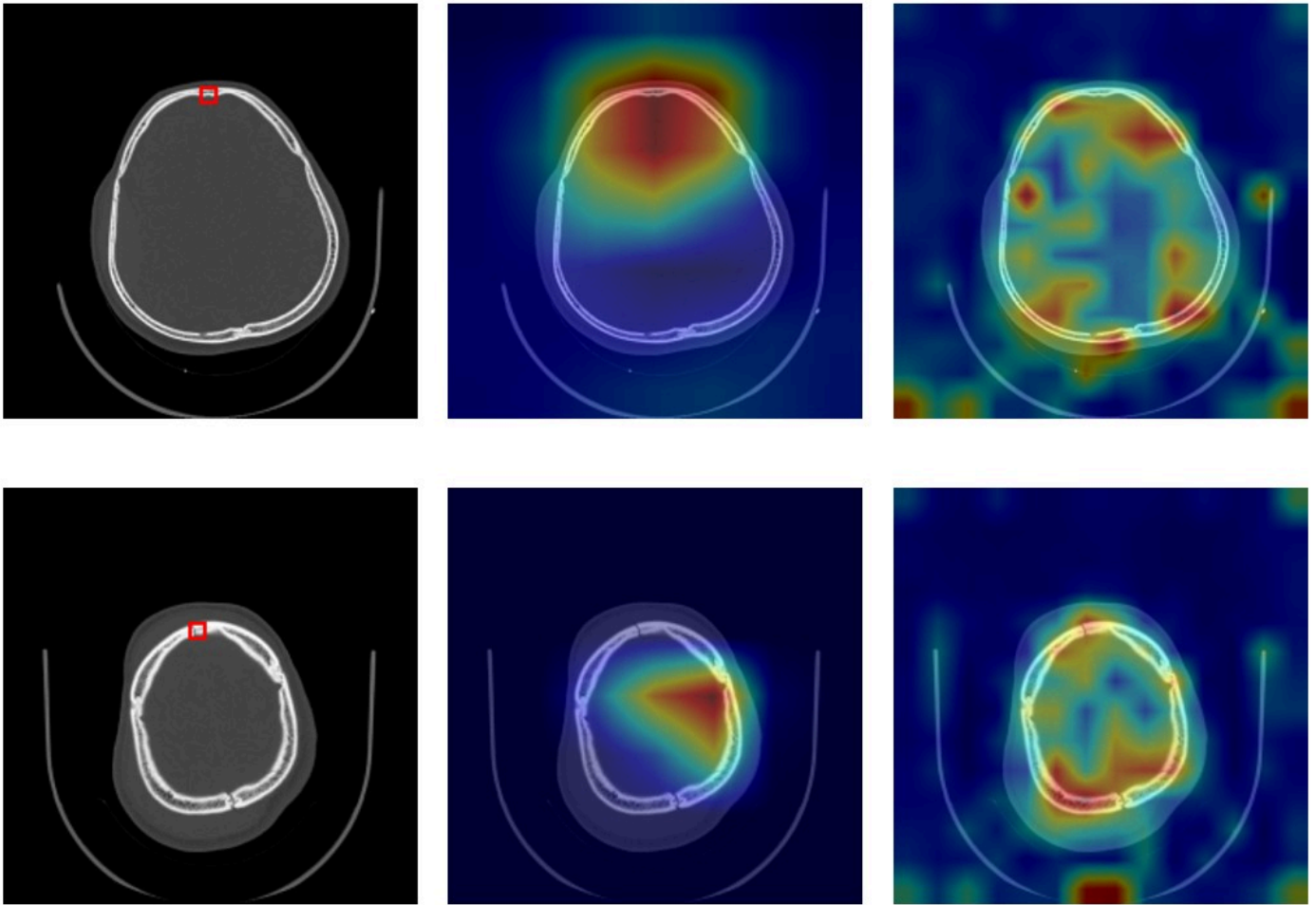
Figure 2

Training history for ResNet-50 (left) and DeiT-Tiny (right): training loss (top) and validation AUC (bottom) over epochs.

Original

ResNet-50

DeiT-Tiny

**Figure 3**

Representative saliency map comparisons. (Original CT slice [left], ResNet-50 Grad-CAM overlay [centre], DeiT-Tiny Attention Rollout [right]. Red boxes indicate radiologist-annotated fracture regions.)

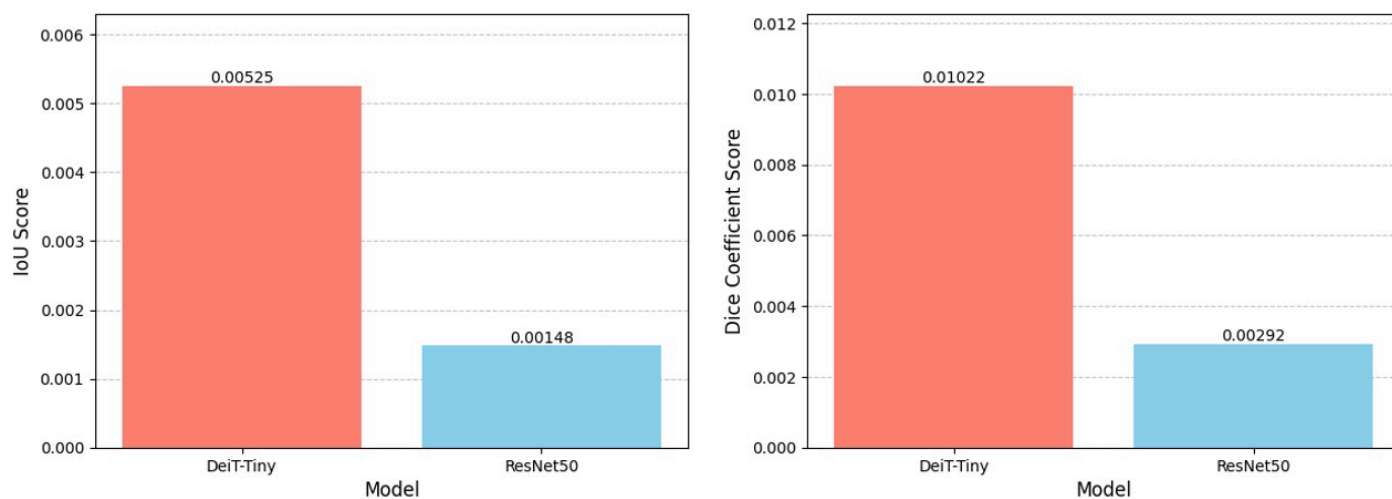


Figure 4

Quantitative explainability evaluation. (Mean Intersection over Union [IoU, left] and Dice Similarity Coefficient [DSC, right] for ResNet-50 and DeiT-Tiny saliency maps compared against radiologist annotations.)