Radiology: Artificial Intelligence

Can AI Watch Your Back? Assessing the Performance of Models from the 2022 RSNA Cervical Spine Fracture Detection Competition at a Level I Trauma Center

Journal:	*Radiology: Artificial Intelligence*
Manuscript ID	RYAI-23-0550.R1
Manuscript Type:	Original Research
Manuscript Categorization Terms:	Feature detection < Vision < Application Domain < 8. MACHINE LEARNING, Supervised learning < Type of machine learning < 8. MACHINE LEARNING, Convolutional Neural Network (CNN) < Deep learning algorithms < Machine learning algorithms < 8. MACHINE LEARNING, Genetic algorithms < 8. MACHINE LEARNING, CT < 2. MODALITIES/TECHNIQUES, Spine < 5. STRUCTURES, Technology Assessment < 7. METHODOLOGY, Head/Neck < 4. AREAS/SYSTEMS, Adults < 1. SUBJECT MATTER

Manuscript Title: Can AI Watch Your Back? Assessing the Performance of Models from the 2022 RSNA Cervical Spine Fracture Detection Competition at a Level I Trauma Center

Article Type: Original Research

Summary

Winning machine learning models from the RSNA 2022 Cervical Spine Fracture Detection competition were evaluated on a clinical validation dataset of 1,829 emergency department cervical spine CT scans obtained over one year from a level I trauma center.

Key Points:

- Models generalized well to the clinical validation dataset with a mean AUC of 0.888 (95% CI: 0.851, 0.924) for fracture detection on non-contrast CT scans and 0.884 (95% CI: 0.829, 0.939) on contrast-enhanced CT scans.
- Compared to the RSNA competition dataset, all models exhibited reduced performance on the clinical validation dataset, especially on the contrast-enhanced scans which were not represented in the training dataset. Nevertheless, the models showed generalizability with a mean sensitivity of 0.819 (95% CI: 0.643, 0.996) and a mean specificity of 0.721 (95% CI: 0.465, 0.976).
- ML models demonstrated capability to identify fractures initially missed by reporting radiologists. Poorer performances were most often attributed to contrast-enhanced scans and scans on patients with degenerative changes and osteopenia.

Abstract

Purpose: Evaluate the top-performing models from the RSNA 2022 Cervical Spine Fracture Detection challenge on a clinical validation dataset. This study includes both non-contrast and contrast-enhanced CT scans, in comparison to the challenge which utilized a filtered curated dataset of only non-contrast scans.

Methods: Seven models achieving high performance in the RSNA Challenge were evaluated on a clinical validation dataset of 1,829 CT scans (130 positives, 1,699 negatives; 1,308 non-contrast, 521 contrast-enhanced), which were acquired without exclusion criteria over one-year from the emergency department of a neurosurgical and level I trauma center.

Results: Although all models exhibited decreased performance in the clinical validation dataset compared to the challenge dataset, the top-ranked models maintained a high performance. The models for non-contrast demonstrated a mean AUC of 0.888 (CI:0.851,0.924), sensitivity of 0.670 (CI:0.516,0.825), and specificity of 0.929 (CI:0.880,0.978). Contrast-enhanced scans achieved a mean AUC of 0.884 (CI:0.829,0.939), sensitivity of 0.819 (CI:0.643,0.996), and specificity of 0.721 (CI:0.465,0.976). ML models identified 10 fractures missed by radiologists. False-positives were more common in contrast-enhanced scans and seen in patients with degenerative changes on non-contrast scans, while false-negatives were often associated with degenerative changes and osteopenia.

Conclusion: The winning models from the 2022 RSNA Competition demonstrated a high performance for cervical spine fracture detection in the clinical validation dataset. To ensure optimal performance across diverse clinical environments, it may be necessary to integrate localized retraining or adjustments, even when the initial training set is broad. The study offers insights for future research and clinical support tool development.

Introduction

Traumatic cervical spinal injuries are frequently encountered and have a significant morbidity and mortality [1, 2]. The incidence of cervical spine injuries is 16.5 per 100,000 individuals [3] and the prevalence is 1.7-3.7% in all blunt trauma patients [4, 5, 6, 7]. Computed tomography (CT) is the gold standard modality for detection of cervical spine fractures due to its high spatial resolution and great delineation of bone [8]. Radiologists are facing increasing workloads, and trauma patients often have CT scans covering their entire neuroaxis, chest, abdomen, and pelvis. These consist of thousands of images, a significant volume for radiologists to interpret. In busy clinical environments, delays between patient imaging and interpretation can occur, potentially leading to adverse patient outcomes [9]. Up to a quarter of patients suffer progression of injuries due to delays in diagnosis or unwarranted manipulation in the emergency department [10]. Rapidly detecting fractures is essential to ensure early immobilization of unstable injuries to prevent neurologic deterioration [11], and to guide appropriate treatment [12]. Surgery is required in 18% of patients with cervical spine injuries [3], and early intervention is associated with better outcomes, particularly when there is spinal cord injury [13, 14].

The increasing volume of imaging studies in emergency situations and the demand for rapid diagnosis have led to the exploration of machine learning (ML) to aid the CT imaging review process [15]. Indeed, ML models have shown to be able to assist radiologists in medical image analysis in a variety of situations, for example detection and characterization of abnormalities such as brain tumors [16], wrist fractures [17], and intracranial hemorrhage [18]. Recent studies have explored ML models for fracture detection in the spine using a variety of architectures [19, 20]. Most work has focused on detection of osteoporotic vertebral fractures, which are more likely to be stable fractures and rarely found in the cervical spine. Some of these studies show high accuracies, with AUC and sensitivities greater than 95% [21, 22] and some matched the performance of radiologists [23], showcasing the potential of ML systems. There are however relatively few studies exploring the application of ML models to aid cervical spine fracture detection in a trauma setting.

ML models have also recently been proposed as a solution for automatically triaging and prioritizing medical imaging studies to reduce delays in diagnosis [24], however using these as clinical decision support tools can be complex, and there are several challenges to overcome to ensure they are generalizable to different populations. A major factor is the limited access to data, as clinical data is often fragmented, stored in disparate systems and subject to privacy regulations [25]. This makes it challenging to access a sufficiently large and diverse dataset, and even when data is accessible, it may suffer from quality issues and biases [25, 26]. In addition, data annotation can be labor-intensive, as well as requiring medical expertise, a significant amount of time, and can be expensive [25, 27].

Initiatives such as the RSNA competition play a crucial role in mitigating some of these issues and have been ongoing for several years [28]. Top-performing models from prior RSNA competition have shown high levels of generalizability on real-world external validation datasets [29, 30]. By organizing competitions that encourage the participation of hundreds of motivated competing teams from around the world, the RSNA crowdsource multi-institutional datasets, expertise, and insights into relevant clinical issues. They also recruit expert annotators who volunteer their time to provide annotations for the dataset, while also coordinating image donation across sites across the globe. The goal of the RSNA 2022 Cervical Spine Fracture Detection competition was to develop ML models that detect and localize fractures in the cervical spine [31] with 1,108 competitors participating. Eight participants were awarded the gold-prize for models that demonstrated exceptional performance on the private test set, with scores and rankings available on the competition's leaderboard [32]. Given the promising results, in this paper, we examined the performance of the top-performing ML models from this competition on a clinical validation dataset.

Although the RSNA competition dataset is based on real-world data collated from multiple institutions, it was curated with the specific intention of hosting a competition. Each contributing site was requested to provide an equivalent number of positive and negative cases which resulted in a substantially higher fracture prevalence than real world rates. The identification and extraction of data was left to the discretion of each site [31] which introduces the potential for selection biases. The data also underwent filtration during curation, removing exams with incomplete coverage of the cervical spine, prior surgery, intravenous contrast, and motion artifacts. In comparison to the RSNA competition dataset, the clinical validation dataset in this study included all consecutive emergent CT scans without exclusion criteria performed over the course of a calendar year, at a busy urban neurosurgical and level I trauma center. Notably it includes contrast-enhanced scans, which were not included in the RSNA competition dataset, as patients often receive contrast as part of full-body trauma imaging at major centers.

Methods

This retrospective study was approved by our Institutional Review Board with a waiver of informed consent.

Model Overview and Evaluation Framework

RSNA 2022 Competition Dataset

The RSNA 2022 Cervical Spine Fracture Detection competition dataset was used for model training and internal validation. This dataset consists of 3,112 cervical spine CT scans from 12 different institutions, with annotations provided by expert radiologists. Divided into training (prevalence of 47.5%), public testing set (prevalence of 40.1%), and private testing sets (prevalence of 45.9%), the dataset has a notably high prevalence of fractures, much higher than real-world rates which range between 4% and 7% [33, 34]. The private test set consisted of 789 non-contrast CT scans. Detailed information about the dataset can be found in work by Lin *et al.* [31]

Models

We selected the seven of the eight award-winning models based on their scores on the RSNA competition's private dataset to rigorously evaluate their generalization capabilities and robustness. The second-place model was excluded from our study as we were unable to reproduce its performance on the private competition test set using the provided source code and technical posts. The seven models we examined leveraged state-of-the-art techniques in computer vision and deep learning. The general strategy adopted by these models is a two-stage approach: segmentation and classification (Figure 1). A detailed description of the models is provided in Supplemental Materials S2.

Experiments and Evaluation

(Insert Figure 2 here)

Evaluation of Model Generalizability

The general workflow to evaluate model performance in the clinical setting is portrayed in Figure 2. To better understand model generalizability, we analyzed one year of consecutive CT scans from the emergency department setting of a busy urban neurosurgical and level I trauma center. The dataset included both non-contrast scans and contrast-enhanced scans, making it ideal for independent external model validation.

Clinical Validation Dataset

The clinical validation dataset used in this study is composed of 1,829 series from 1,828 cervical spine CT studies across 1,779 adult patients (625 females, 1,154 males; age range 18–101 years; mean age 55.8 ± 22.1 years); a minority of patients had either pre- and post-contrast scans or repeat attendances to the emergency department. There were 130 scans positive for fracture. The dataset was then divided into non-contrast and contrast-enhanced acquisitions. The non-contrast scans are similar to the data used to train and evaluate models during the competition while the contrast-enhanced scans allow for the evaluation of model generalizability outside the distribution of the training data. The dataset included 1,308 non-contrast and 521 contrast-enhanced scans (Table 1). To obtain ground truth labels, clinical reports for all the CT scans were manually reviewed by a radiologist and classified as positive or negative for fracture at the patient and cervical spine segmental levels. Ground truth was established for equivocal reports by reviewing follow-up imaging examinations and clinical records. A random sample of 10% of the radiology reports was reviewed by a second radiologist with 100% concordance at the segmental level. The presence or absence of intravenous contrast was also established for each scan. Additional dataset curation details are provided in the supplemental materials S1.

Evaluation

The Youden's *J* statistic [35] was employed on the competition's public test dataset to determine optimal thresholds to binarize predicted probabilities that maximize the difference between the true positive rate and the false positive rate, effectively capturing the top-left most point on the receiver operating characteristic (ROC) curve. The thresholds for each of the seven models (Threshold$_{Qishen}$=0.72, Threshold$_{Darragh}$=0.54, Threshold$_{Selim}$=0.57, Threshold$_{SpeedRun}$=0.81, Threshold$_{Skecherz}$=0.49, Threshold$_{QWER}$=0.54, Threshold$_{Harshit}$=0.72) were then applied to the competition's private test set to establish baseline model performance on the competition dataset. Ground truth labels for each scan were compared to the ML model predictions. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and AUC were the primary evaluative metrics; Balanced accuracy (BAcc, the average of sensitivity and specificity), and Matthew's correlation coefficient (MCC) were also calculated and are presented in the supplemental materials S1. To provide a comprehensive analysis, we estimated 95% confidence intervals around each metric on the competition private test dataset and clinical validation datasets separately. Specifically, we utilized the confidence interval Python library (v1.0.4) developed by Jacob [36] to find analytical confidence intervals (CI). Within this module, confidence intervals for accuracy, sensitivity, specificity, PPV, and NPV were estimated using the binomial method [37]. AUC confidence intervals were determined through the fast DeLong method [38].

Our evaluation included contrast-enhanced and non-contrast CT scans in the clinical validation dataset. Performance comparisons between the competition private test and clinical validation datasets were made to identify any differences or any trends. Additional analyses, including those adjusting the clinical validation dataset to match the competition dataset's prevalence, are detailed in the supplemental materials S1.

Review of False Negative and False Positive Cases

A neuroradiologist reviewed every exam level false negative (FN) case to help determine the types of fractures missed by the ML models. CT scans that were misclassified as false positive (FP) at the exam level by at least four of the seven models also underwent review. This approach was pursued as two models (Skecherz and Harshit) accounted for a substantial proportion of FP cases, whereas many of these were correctly classified by the other models. Grad-CAM [39] heatmaps were generated based on the averaged model outputs for FP classifications. Grad-CAM is a visualization technique that illuminates areas of an image influencing convolutional neural network (CNN) prediction, by highlighting these regions with heat maps. To interpret these maps, warmer colors (e.g., red) indicate areas the model focuses on more intensely, with brighter colors signifying higher influence on the model's decision. These heatmaps allowed the neuroradiologist to concentrate on identifying commonly occurring features that may have misguided the model's judgment.

Results

Figure 3 visualizes the distribution of performance metrics for binary classification by the winning algorithms on different datasets. Detailed information regarding individual model performance and other analyses are provided in the supplemental materials S1.

Competition Private Test Dataset

The models on the competition private test dataset had a mean AUC of 0.957 (95% CI: 0.950, 0.964), with a mean accuracy of 0.91 (95% CI: 0.898, 0.922). The mean sensitivity was recorded at 0.872 (95% CI: 0.853, 0.891), and mean specificity at 0.943 (95% CI: 0.914, 0.971).

Clinical Validation Non-Contrast and Contrast Datasets

The model results on the clinical validation dataset showed a reduction in AUC and accuracy on both subsets of the dataset. Accuracy was reduced in the contrast-enhanced dataset; however, it remained high for the non-contrast subset. Due to the lower prevalence, both non-contrast and contrast-enhanced datasets exhibited a high NPV and a notable decrease in PPV in comparison to the competition dataset. In the non-contrast dataset with a real-world prevalence of 4.2%, the mean AUC was 0.888 (95% CI: 0.851, 0.924), mean accuracy was 0.918 (95% CI: 0.875, 0.962), mean sensitivity was 0.67 (95% CI: 0.516, 0.825), and mean specificity was 0.929 (95% CI: 0.880, 0.978). In the contrast-enhanced dataset with a real-world prevalence of 14.5%, the mean AUC was 0.884 (95% CI: 0.829, 0.939), mean accuracy was 0.735 (95% CI: 0.532, 0.938), mean sensitivity was 0.819 (95% CI: 0.643, 0.996), and specificity was 0.721 (95% CI: 0.465, 0.976).

Analysis of False Positive and False Negative Scans

There were 116 false positives (FPs) (47 non-contrast, 69 contrast-enhanced) and 78 false negatives (FNs) (35 non-contrast, 43 contrast-enhanced) scans. On review, the ML models correctly identified 10 cases of true fractures, initially missed by reporting radiologists (Figure 4). The most common influential regions identified on Grad-CAM heatmaps for FP cases were vessels, present in 43/116 FP cases (37%), and in 39/69 FP contrast-enhanced studies (56%). Other, less common, influential regions contributing to FP cases were related to chronic changes such as osteophytes, degenerative cortical irregularities, ligament and soft tissue calcification, vascular channels, and artifacts (Figure 5).

On review of the FN cases for the 7 ML models, there were 134 fractures across 78 cases (42 contrast-enhanced and 36 non-contrast). There were two cases in which no definite fracture was identified, and these were reclassified as true negative cases. In cases of fracture, there were 88 underlying factors in the region of injury, possibly contributing to missed detection by the ML models. The most common were chronic and degenerative changes (53/88), followed by osteopenia (23/88), artifact (8/88), healed chronic fracture (2/88), and osseous lesions associated with pathological fracture (2/88). The most common sites of missed fractures were at the edge of the vertebral body endplate (36/135), transverse process (35/135), and spinous process (17/135). The most common levels of missed fractures were at C7 (19.2%) followed by C6 (17.7%).

Discussion

Award-winning models from the 2022 RSNA competition demonstrate reduced but still impressive performance on our clinical validation dataset from an urban neurosurgical and level I trauma center. The major strength of this study is that every cervical spine CT scan performed on adults in the emergency department over a one-year period was included in this study without exclusion criteria. Importantly, both contrast-enhanced and non-contrast CT scans were included in this dataset, as both are routinely encountered in clinical practice, particularly at major trauma centers, despite models being trained solely only on non-contrast scans. Although the competition dataset used real-world data collected from multiple institutions, the data underwent filtration during curation to help optimize it for competition purposes, which may not accurately reflect the clinical setting. Our dataset provides a more genuine representation of data encountered in real-world clinical environments and was analyzed in balanced and unbalanced groups, reflecting the matched higher prevalence of fractures in the competition test dataset and the lower prevalence of fractures encountered in clinical practice.

In examining the performance metrics, it was noted that the models faced challenges when applied to the clinical dataset, particularly with contrast-enhanced scans. Higher performance in the non-contrast subset is expected given that the training dataset consists of these exclusively. Average model sensitivity reductions were 0.202 for non-contrast scans and 0.052 for contrast-enhanced scans, while average specificity reductions were 0.014 for non-contrast scans and 0.222 for contrast-enhanced scans. Interestingly, accuracy for non-contrast scans slightly improved, with an average increase of 0.007, whereas contrast-enhanced scans experienced an accuracy decline of 0.175, both attributed to differences in fracture prevalence.

There are few studies exploring the application of ML models for cervical spine fracture detection. Zhang et al. pioneered an end-to-end deep-learning model, achieving an AUC of 0.87 at the image level and 0.85 at the scan level on a dataset of 1,347 CT scans [40]. Salehinejad et al. [41] proposed a deep convolutional neural network (DCNN) with a bidirectional long-short term memory (BLSTM) layer achieving a classification accuracy of 71-79% on a dataset of 3,666 studies and 71% for the balanced dataset. Their study highlights the importance of incorporating features from adjacent CT images for accurate fracture detection. Golla et al. [42] proposed a voxel classification-based approach, achieving 87% sensitivity at the fracture level. Currently, there is one FDA-approved software to detect fractures in the cervical spine (BriefCase, AIDOC Medical), which uses a two-stage ML model comprising a U-net masking stage followed by a false-positive reduction classification stage. It reported a sensitivity of 91.7% and specificity of 88.6% in the regulatory submission of 186 cases [43]. However, external validation studies by Small et al. showed a sensitivity of 76% and specificity of 97% [44], and Voter et al. showed a sensitivity of 54.9% and specificity of 94.1% [45].

Our study's observed drop in performance between curated competition datasets and actual clinical application aligns with the findings from Voter's and Small's external validation studies, which reported sensitivity reductions from 91.7% to 76% and 54.9% respectively [44, 45], emphasizing the challenges of transplanting models trained on datasets not native to the intended deployment environment. As part of a clinical decision support system, models triaging potential abnormal cases for more urgent radiologist review should have a high sensitivity, to help identify as many positive fracture cases as possible and reduce the risk of missing a case, however, models with high specificity may be more conservative and reduce false positive cases and unnecessary urgent review of imaging by the radiologist. Balancing these considerations is vital and optimization based on acceptable sensitivity and specificity should be used prior to incorporating these models into the clinical workflow. Additionally, providing the predicted probabilities may provide more nuance and assist radiologists in weighing the evidence of a fracture.

Our results suggest that ML models developed by participating teams for the RSNA competition have been able to achieve better model performance than the previous reported studies by individual research groups. Although the models evaluated still do not match radiologist metrics, with the sensitivity and specificity of radiologists to detect cervical spine fractures on CT shown to be 0.88-0.930 and 0.96-0.990 respectively [44, 46], these models hold promise as rapid, auxiliary tools. In our study, a small number of fractures missed by radiologists had been retrospectively identified by the ML models, and the models could generate a cervical spine fracture prediction in just 10 to 30 seconds. Currently it takes between 33 to 43 minutes from scan acquisition until a finalized report by radiologists [44], and therefore ML models could be used as rapid triaging tools to flag the study alerting the radiologist of a possible fracture, some of which may be missed by radiologists.

Our study has also revealed areas of strength and improvement for the models, through a comprehensive review of the FN and FP cases. Intravascular contrast, chronic changes, osseous channels, and artifacts can lead to falsely labeling studies as positive for fracture. For example, small opacified vessels closely related to the cervical spine can mimic the appearance of a fracture fragment. In the FN cases, there were certain types of fractures missed most by the models, including fractures at the edge of the vertebral body endplate, transverse process, and spinous process locations, consistent with previous research [44, 45]. The most common cause for models to miss fractures were degenerative changes and osteopenia, also observed by Small *et al.* [44], leading to underperformance in older patients [45]. An understanding of these patterns can guide future model refinement by inclusion of greater numbers of imaging studies with under-represented pathologies.

Our study has several limitations, including the use of clinical data from a single center, which can limit the generalizability of findings, and the retrospective nature of the study. While our institution contributed data to the competition, no patients were represented in both the competition and clinical validation datasets. The training data used to develop the models was narrowly focused on acute fractures identified through non-contrast CT scans and patients with surgical hardware were excluded due to the presence of streak artifact [34], and therefore models could underperform on our clinical validation dataset, as it included patients who had previous surgical intervention and contrast-enhanced studies. Additionally, our use of Grad-CAM for model transparency is limited by its difficulty in localizing multiple instances of the same class and in capturing fine-grained details, as highlighted by Mohamed [47] and Draelos [48]. Despite these challenges, the significant limitations of Grad-CAM were not observed in our study. For tasks requiring detailed localization, HiResCAM [48] could offer improved interpretability by addressing these specific limitations. ML models could likely be improved by additional training on patients with prior surgical intervention and contrast-enhanced studies which would potentially have resulted in increased performance on the clinical validation dataset. Furthermore, additional training on data from individual sites of deployment will also likely result in improved performance. Ongoing refinements will be essential for maximizing the models' effectiveness and improving patient outcomes in real-world healthcare settings, through large prospective multicenter studies.

Conclusion

In our clinical validation dataset, the top-performing ML models in the 2022 RSNA competition fell short of their performance on the competition dataset, however, they still performed favorably to previously published cervical spine detection algorithms, including FDA-approved commercial models. In this study, they showed potential to generalize to the analysis of contrast-enhanced scans and on patients with prior surgical intervention despite being trained on a dataset that excluded these examinations. Addressing false positive and negative cases through the inclusion of relevant imaging studies holds potential for future model refinement. These models serve as valuable supplementary diagnostic tools for cervical spine fracture detection, emphasizing the necessity for ongoing improvement efforts and prospective evaluation of deployed models.

Data and Model availability

The publicly available RSNA 2022 Cervical Spine Fracture Detection CT dataset and competition award winning models are available at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection. The competition private test and clinical validation dataset are not publicly available.

Our analysis and model implementation were conducted using Python (3.10.13), torch (2.1.0). Additionally, we utilized a suite of Python packages to facilitate data analysis and results visualization, including SimpleITK (2.3.1), nibabel (5.2.0), torchvision (0.16.1), numpy (1.26.2), scikit-image (0.22.0), opencv-python (4.8.1), scipy (1.11.4), pandas (2.1.4), matplotlib (3.8.0), grad-cam (1.4.8), and confidenceinterval (1.0.4). This detailed enumeration of the software and packages utilized aims to enhance the reproducibility and transparency of our study.

Acknowledgements

No acknowledgement.

REFERENCES

1. Bank M, Gibbs K, Sison C, Kutub N, Paptheodorou A, Lee S, Stein A, Bloom O. Age and Other Risk Factors Influencing Long-Term Mortality in Patients With Traumatic Cervical Spine Fracture. Geriatr Orthop Surg Rehabil. 2018 May 3;9:2151459318770882. doi: 10.1177/2151459318770882.
2. Spinal Cord Injury. World Health Organization. Available from: https://www.who.int/news-room/fact-sheets/detail/spinal-cord-injury. Accessed March 10, 2024
3. Fredø HL, Bakken IJ, Lied B, Rønning P, Helseth E. Incidence of traumatic cervical spine fractures in the Norwegian population: a national registry study. Scand J Trauma Resusc Emerg Med. 2014 Dec 18;22:78. doi: 10.1186/s13049-014-0078-7.
4. Milby AH, Halpern CH, Guo W, Stein SC. Prevalence of cervical spinal injury in trauma. Neurosurg Focus. 2008;25(5):E10. doi: 10.3171/FOC.2008.25.11.E10.
5. Inaba K, Byerly S, Bush LD, Martin MJ, Martin DT, Peck KA, Barmparas G, Bradley MJ, Hazelton JP, Coimbra R, Choudhry AJ, Brown CV, Ball CG, Cherry-Bukowiec JR, Burlew CC, Joseph B, Dunn J, Minshall CT, Carrick MM, Berg GM, Demetriades D; WTA C-Spine Study Group. Cervical spinal clearance: A prospective Western Trauma Association Multi-institutional Trial. J Trauma Acute Care Surg. 2016 Dec;81(6):1122-1130. doi: 10.1097/TA.0000000000001194.
6. Stiell IG, Wells GA, Vandemheen KL, Clement CM, Lesiuk H, De Maio VJ, Laupacis A, Schull M, McKnight RD, Verbeek R, Brison R, Cass D, Dreyer J, Eisenhauer MA, Greenberg GH, MacPhail I, Morrison L, Reardon M, Worthington J. The Canadian C-spine rule for radiography in alert and stable trauma patients. JAMA. 2001 Oct 17;286(15):1841-8. doi: 10.1001/jama.286.15.1841.
7. Hasler RM, Exadaktylos AK, Bouamra O, Benneker LM, Clancy M, Sieber R, Zimmermann H, Lecky F. Epidemiology and predictors of cervical spine injury in adult major trauma patients: a multicenter cohort study. J Trauma Acute Care Surg. 2012 Apr;72(4):975-81. doi: 10.1097/TA.0b013e31823f5e8e.
8. Minja FJ, Mehta KY, Mian AY. Current Challenges in the Use of Computed Tomography and MR Imaging in Suspected Cervical Spine Trauma. Neuroimaging Clin N Am. 2018 Aug;28(3):483-493. doi: 10.1016/j.nic.2018.03.009.
9. Glover M 4th, Almeida RR, Schaefer PW, Lev MH, Mehan WA Jr. Quantifying the Impact of Noninterpretive Tasks on Radiology Report Turn-Around Times. J Am Coll Radiol. 2017 Nov;14(11):1498-1503. doi: 10.1016/j.jacr.2017.07.023.
10. Malik SA, Murphy M, Connolly P, O'Byrne J. Evaluation of morbidity, mortality and outcome following cervical spine injuries in elderly patients. Eur Spine J. 2008 Apr;17(4):585-91. doi: 10.1007/s00586-008-0603-3.
11. Delcourt T, Bégué T, Saintyves G, Mebtouche N, Cottin P. Management of upper cervical spine fractures in elderly patients: current trends and outcomes. Injury. 2015 Jan;46 Suppl 1:S24-7. doi: 10.1016/S0020-1383(15)70007-0.
12. Fehlings MG, Perrin RG. The timing of surgical intervention in the treatment of spinal cord injury: a systematic review of recent clinical evidence. Spine (Phila Pa 1976). 2006 May 15;31(11 Suppl):S28-35; discussion S36. doi: 10.1097/01.brs.0000217973.11402.7f.
13. Fehlings MG, Perrin RG. The role and timing of early decompression for cervical spinal cord injury: update with a review of recent clinical evidence. Injury. 2005 Jul;36 Suppl 2:B13-26. doi: 10.1016/j.injury.2005.06.011.

14. Fehlings MG, Vaccaro A, Wilson JR, Singh A, W Cadotte D, Harrop JS, Aarabi B, Shaffrey C, Dvorak M, Fisher C, Arnold P, Massicotte EM, Lewis S, Rampersaud R. Early versus delayed decompression for traumatic cervical spinal cord injury: results of the Surgical Timing in Acute Spinal Cord Injury Study (STASCIS). PLoS One. 2012;7(2):e32037. doi: 10.1371/journal.pone.0032037.
15. Perotte R, Lewin GO, Tambe U, Galorenzo JB, Vawdrey DK, Akala OO, Makkar JS, Lin DJ, Mainieri L, Chang BC. Improving Emergency Department Flow: Reducing Turnaround Time for Emergent CT Scans. AMIA Annu Symp Proc. 2018 Dec 5;2018:897-906.
16. Gull S, Akbar S. Artificial intelligence in brain tumor detection through MRI scans: advancements and challenges. Artificial Intelligence and Internet of Things. 2021 Aug 25;241-76.
17. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018 May;73(5):439-445. doi: 10.1016/j.crad.2017.11.015.
18. Kuo W, Häne C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. Proc Natl Acad Sci U S A. 2019 Nov 5;116(45):22737-22745. doi: 10.1073/pnas.1908021116.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition. 2016;770-778.
20. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InMedical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 2015;234-241. Springer International Publishing.
21. Burns JE, Yao J, Summers RM. Vertebral Body Compression Fractures and Bone Density: Automated Detection and Classification on CT Images. Radiology. 2017 Sep;284(3):788-797. doi: 10.1148/radiol.2017162100.
22. Nicolaes J, Raeymaeckers S, Robben D, Wilms G, Vandermeulen D, Libanati C, Debois M. Detection of vertebral fractures in CT using 3D convolutional neural networks. InComputational Methods and Clinical Applications for Spine Imaging: 6th International Workshop and Challenge, CSI 2019, Shenzhen, China, October 17, 2019, Proceedings 6 2020;3-14. Springer International Publishing.
23. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med. 2018 Jul 1;98:8-15. doi: 10.1016/j.compbiomed.2018.05.011.
24. Jha S. Value of Triage by Artificial Intelligence. Acad Radiol. 2020 Jan;27(1):153-155. doi: 10.1016/j.acra.2019.11.002.
25. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP. Preparing Medical Imaging Data for Machine Learning. Radiology. 2020 Apr;295(1):4-15. doi: 10.1148/radiol.2020192224.
26. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, Matsui Y, Nozaki T, Nakaura T, Fujima N, Tatsugami F, Yanagawa M, Hirata K, Yamada A, Tsuboyama T, Kawamura M, Fujioka T, Naganawa S. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2024 Jan;42(1):3-15. doi: 10.1007/s11604-023-01474-3.
27. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018 Aug;18(8):500-510. doi: 10.1038/s41568-018-0016-5.

28. Radiological Society of North America. AI challenges. Available from: https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge. Accessed March 10, 2024
29. Salehinejad H, Kitamura J, Ditkofsky N, Lin A, Bharatha A, Suthiphosuwan S, Lin HM, Wilson JR, Mamdani M, Colak E. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. Sci Rep. 2021 Aug 23;11(1):17051. doi: 10.1038/s41598-021-95533-2.
30. Beheshtian E, Putman K, Santomartino SM, Parekh VS, Yi PH. Generalizability and Bias in a Deep Learning Pediatric Bone Age Prediction Model Using Hand Radiographs. Radiology. 2023 Feb;306(2):e220505. doi: 10.1148/radiol.220505.
31. Lin HM, Colak E, Richards T, Kitamura FC, Prevedello LM, Talbott J, Ball RL, Gumeler E, Yeom KW, Hamghalam M, Simpson AL, Strika J, Bulja D, Angkurawaranon S, Pérez-Lara A, Gómez-Alonso MI, Ortiz Jiménez J, Peoples JJ, Law M, Dogan H, Altinmakas E, Youssef A, Mahfouz Y, Kalpathy-Cramer J, Flanders AE; RSNA-ASSR-ASNR Annotators and the Dataset Curation Contributors. The RSNA Cervical Spine Fracture CT Dataset. Radiol Artif Intell. 2023 Aug 30;5(5):e230034. doi: 10.1148/ryai.230034.
32. Kaggle. RSNA 2022 Cervical Spine Fracture Detection Leaderboard. Available from: https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/leaderboard. Accessed March 10, 2024
33. Tang A, Pawar J, Bridge C, King R, Kakarmath S, Harris M, Khurana B. Traumatic cervical spine fracture patterns on CT: a retrospective analysis at a level 1 trauma center. Emerg Radiol. 2021 Oct;28(5):965-976. doi: 10.1007/s10140-021-01952-z.
34. Khanpara S, Ruiz-Pardo D, Spence SC, West OC, Riascos R. Incidence of cervical spine fractures on CT: a study in a large level I trauma center. Emerg Radiol. 2020 Feb;27(1):1-8. doi: 10.1007/s10140-019-01717-9.
35. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology. 2005 Jan;16(1):73-81. doi: 10.1097/01.ede.0000147512.81966.ba.
36. Jacob Gildenblat. A python library for confidence intervals. Available from: https://github.com/jacobgil/confidenceinterval. Accessed March 10, 2024
37. Blyth CR, Still HA. Binomial confidence intervals. Journal of the American Statistical Association. 1983 Mar 1;78(381):108-16.
38. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters. 2014 Jul 9;21(11):1389-93.
39. Takahashi K, Yamamoto K, Kuchiba A, Koyama T. Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. Appl Intell (Dordr). 2022 Mar;52(5):4961-4972. doi: 10.1007/s10489-021-02635-5.
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. InProceedings of the IEEE international conference on computer vision. 2017;618-626.
41. Zhang M, Kim L, Cheong R, Cohen-Wang B, Shpanskaya K, Wetstone J, Manoj N, Rajpurkar P, Yeom K. Deep-Learning Artificial Intelligence Model for Automated Detection of Cervical Spine Fracture on Computed Tomography (CT) Imaging. San Diego, CA. April 13–17, 2019.

2019 AANS Annual Scientific Meeting. J Neurosurg. 2019 Jul 1;131(1):2-116. doi: 10.3171/2019.7.JNS.AANS2019abstracts.

42. Salehinejad H, Ho E, Lin HM, Crivellaro P, Samorodova O, Arciniegas MT, Merali Z, Suthiphosuwan S, Bharatha A, Yeom K, Mamdani M. Deep sequential learning for cervical spine fracture detection on computed tomography imaging. In2021 IEEE 18th international symposium on biomedical imaging (ISBI) 2021 Apr 13;1911-1914.
43. Golla AK, Lorenz C, Buerger C, Lossau T, Klinder T, Mutze S, Arndt H, Spohn F, Mittmann M, Goelz L. Cervical spine fracture detection in computed tomography using convolutional neural networks. Phys Med Biol. 2023 May 29;68(11). doi: 10.1088/1361-6560/acd48b.
44. US Food and Drug Administration. K190896. Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190896.pdf. Accessed March 10, 2024.
45. Small JE, Osler P, Paul AB, Kunst M. CT Cervical Spine Fracture Detection Using a Convolutional Neural Network. AJNR Am J Neuroradiol. 2021 Jul;42(7):1341-1347. doi: 10.3174/ajnr.A7094.
46. Voter AF, Larson ME, Garrett JW, Yu JJ. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures. AJNR Am J Neuroradiol. 2021 Aug;42(8):1550-1556. doi: 10.3174/ajnr.A7179.
47. van der Kolk BBYM, van den Wittenboer GGJ, Warringa N, Nijholt IM, van Hasselt BAAM, Buijteweg LN, Schep NWL, Maas M, Boomsma MF. Assessment of cervical spine CT scans by emergency physicians: A comparative diagnostic accuracy study in a non-clinical setting. J Am Coll Emerg Physicians Open. 2022 Jan 20;3(1):e12609. doi: 10.1002/emp2.12609.
48. Mohamed E, Sirlantzis K, Howells G. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. Displays. 2022 Jul 1;73:102239.
49. Draelos RL, Carin L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. arXiv preprint arXiv:2011.08891. 2020 Nov 17.

Figure Legends

Figure 1. An example of end-to-end architecture of a cervical spine CT fracture detection ML model, showcasing the segmentation stage to isolate the cervical spine's voxels of interest (VOI), followed by the classification stage for feature extraction, aggregation, and logits prediction.

Figure 2. ML evaluation pipeline for cervical spine fracture detection. (a) The process starts with 3,112 CT scans from the RSNA 2022 competition, divided into training, public test, and private test datasets. An additional CT scans from our institution are used as external validation dataset. (b) Each ML model has two main stages: segmentation, typically using 2D or 3D UNet, and classification involving CNN feature extraction, feature aggregation and logits prediction. (c) Each model generates a fracture probability output binarized by applying an optimal threshold identified by *Youden's J* on the public test dataset. Then, the model's performance is assessed using the private test dataset. (d) The final evaluation comprises four external validation datasets: non-contrast scans, contrast-enhanced scans, bootstrap-sampled non-contrast scans, and bootstrap-sampled contrast-enhanced scans.

Figure 3. Box and whisker plots showcasing the distribution of performance metrics for binary classification by the winning algorithms on different datasets. Metrics include AUC, sensitivity, specificity, and accuracy. These are displayed across the competition private dataset (CP), simulated prevalence non-contrast (SPNC), simulated prevalence contrast (SPC), actual prevalence non-contrast (NC), and actual prevalence contrast (C) datasets. The box represents the interquartile range, the median is indicated by the line within the box, and the whiskers show the full range excluding outliers, which are depicted as individual points. Data points are also visualized as jitters for clarity.

Figure 4. Example cases of fractures identified by the ML model but missed by reporting radiologists. The CT images with associated Grad-CAM heatmaps show the most influential regions in the input image for the prediction. (a,b) Minimally-displaced left transverse process fracture. (c,d) Bilateral lamina fractures, moderately-displaced on the right and undisplaced on the left. (e,f) Mildly displaced spinous process fracture. (g,h) Undisplaced left articular process/lamina fracture. (i,j) Minimally displaced spinous process fracture, and (k,l) minimally displaced left transverse process fracture.

Figure 5. Example cases incorrectly identified as fractures by the ML model in the false positive group. The CT images with associated Grad-CAM heatmaps show the most influential regions in the input image for the prediction. (a,b) Calcified atherosclerotic plaque in the left vertebral artery in the left transverse foramen. (c,d) Congenital lack of fusion of the posterior arch of C1. (e,f) Contrast within a small vessel in the right paraspinal region on a contrast-enhanced CT examination. (g,h) Chronic multilevel degenerative changes with reduced intervertebral disc spaces, osteophyte formation and osteopenia. (i,j) Partially calcified pseudomass posterior to the odontoid process of C2, secondary to calcium pyrophosphate dihydrate crystal deposition disease. (k,l) Nutrient vessel within the left lamina. (m,n) Chronic osteophyte arising from the superior-anterior vertebral body of C3, and (o,p) chronic osteophytic changes associated with the right articular process.

Tables

Table 1: Clinical validation dataset data distribution details.

Attribute		Non-contrast	Contrast-enhanced	Overall
Total Series		1308	521	1829
Positive		55	75	130
	C1	5	19	24
	C2	14	26	40
	C3	8	10	18
	C4	7	14	21
	C5	11	16	27
	C6	18	22	40
	C7	17	21	38
Negative		1253	446	1699
Prior CSpine Surgery		14	3	17
Positive		1	1	2
Negative		13	2	15
Post-operative material		14	2	16
Positive		1	1	2
Negative		13	1	14
Total Patients		1268	520	1779
Female		481	145	625
Male		787	375	1154
Mean Age		58.2±22.3	50.0±20.6	55.8±22.1

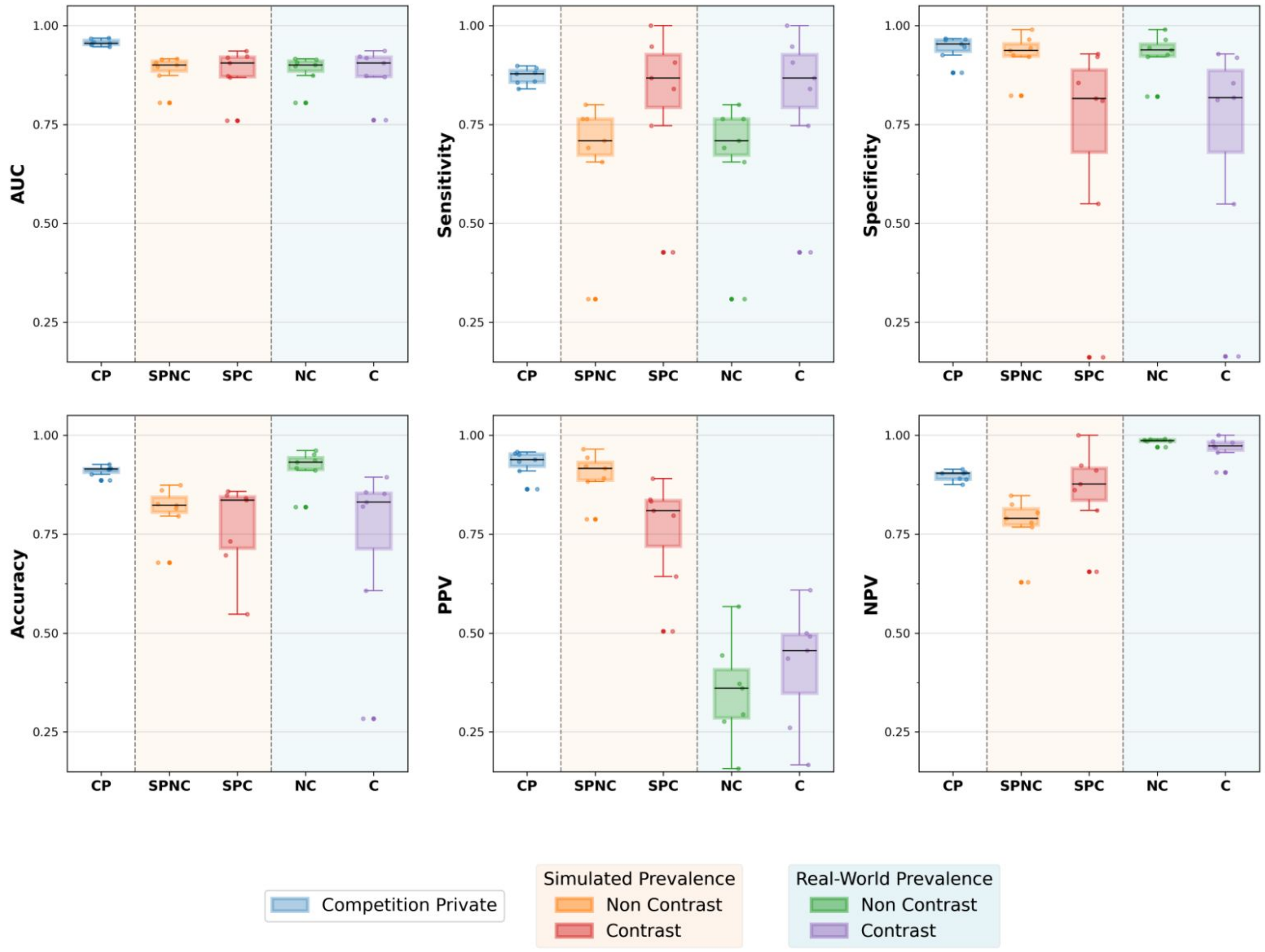
Appendix S1

Details of Clinical Validation Dataset

Our hospital's radiology information system (Syngo, Siemens Medical Solutions USA, Inc, Malvern, PA) was searched using Nuance mPower (Nuance Communications, Burlington, MA) for emergency patients at least 18 years of age that underwent a CT scan of the cervical spine between January 1 to December 31, 2022, for a traumatic indication. Every CT which included imaging of the cervical spine was included in this study, including both contrast-enhanced and non-contrast enhanced studies. All examinations were performed on a 64 row multi-detector CT scanner (Revolution, LightSpeed 64, or Optima 64, General Electric SMedical Systems, Milwaukee, WI, U.S.). A total of 1,971 reports were manually reviewed by a neuroradiologist and classified as positive or negative for fracture at the patient and cervical spine segmental levels. Imaging with equivocal reports categorized as indeterminate for fractures were considered positive cases in the assessment of the ML model's performance, as they would warrant prioritized review by a radiologist within a triage system. There is no overlap between these patients and CT scans with those we contributed to the RSNA 2022 Cervical Spine Fracture Detection competition dataset (314 CT scans in the training set). Eighty-nine scans were excluded from the study as they were performed for a non-traumatic indication. CT examinations were downloaded in Digital Imaging and Communications in Medicine (DICOM) format from the Picture Archiving and Communications System (Philips Vue PACS, Philips Healthcare, Netherlands) using RSNA Anonymizer (Oak Brook, IL). Fifty-three scans could not be downloaded from our local imaging archive. Axial bone window images of the cervical spine measuring 1 mm or less in image thickness were extracted using a custom Python script and de-identified by RSNA Anonymizer. A web-based annotation platform (MD.ai, New York, NY) was used by a radiologist to label each series as non-contrast or contrast-enhanced.

In this study, we aim to demonstrate each model's ability to distinguish between patients with and without c-spine fractures, rather than comparing models. Within this framework, our study extensively leveraged data from an entire calendar year, comprising 1,829 cervical spine CT scans obtained from a busy urban neurosurgical and trauma center. Specifically, we analyzed 1,308 non-contrast CT scans, of which 55 were positive for fractures, and 521 contrast-enhanced scans, with 75 positives. This gave us a lot more data than we initially thought we needed, making our study even stronger and more reliable. Having a full year's worth of data also means we covered a wide variety of real-life situations, which helps make our findings even more useful and relevant.

Details of the distribution of performance metrics



Supplemental Figure 1. Box and whisker plots showcasing the distribution of performance metrics for binary classification by the winning algorithms on different datasets. Metrics include AUC, sensitivity, specificity, and accuracy. These are displayed across the competition private dataset (CP), simulated prevalence non-contrast (SPNC), simulated prevalence contrast (SPC), real-world prevalence non-contrast (NC), and real-world prevalence contrast (C) datasets. The box represents the interquartile range, the median is indicated by the line within the box, and the whiskers show the full range excluding outliers, which are depicted as individual points. Data points are also visualized as jitters for clarity.

metric	**CP**	**SPC**	**SPNC**	**C**	**NC**
AUC	0.957 (0.950, 0.964)	0.883 (0.828, 0.938)	0.887 (0.851, 0.923)	0.884 (0.829, 0.939)	0.888 (0.851, 0.924)
Acc	0.910 (0.898, 0.922)	0.766 (0.659, 0.872)	0.810 (0.751, 0.870)	0.735 (0.532, 0.938)	0.918 (0.875, 0.962)
SEN	0.872 (0.853, 0.891)	0.819 (0.643, 0.996)	0.670 (0.516, 0.825)	0.819 (0.643, 0.996)	0.670 (0.516, 0.825)
SPEC	0.943 (0.914, 0.971)	0.720 (0.464, 0.976)	0.929 (0.881, 0.978)	0.721 (0.465, 0.976)	0.929 (0.880, 0.978)
NPV	0.897 (0.885, 0.910)	0.863 (0.762, 0.963)	0.777 (0.712, 0.843)	0.967 (0.939, 0.995)	0.985 (0.978, 0.991)
PPV	0.930 (0.899, 0.960)	0.759 (0.634, 0.885)	0.901 (0.848, 0.954)	0.417 (0.277, 0.558)	0.353 (0.233, 0.474)
F1	0.899 (0.887, 0.911)	0.762 (0.663, 0.861)	0.753 (0.633, 0.872)	0.519 (0.392, 0.647)	0.429 (0.345, 0.513)
MCC	0.821 (0.797, 0.846)	0.573 (0.417, 0.728)	0.635 (0.534, 0.735)	0.452 (0.303, 0.600)	0.434 (0.360, 0.509)

Supplemental Table 1. the distribution of performance metrics for binary classification by the winning algorithms on different datasets. Metrics include AUC, accuracy, sensitivity, specificity, NPV, PPV, F1 score, and MCC. These are displayed across the competition private dataset (CP), simulated prevalence non-contrast (SPNC), simulated prevalence contrast (SPC), real-world prevalence non-contrast (NC), and real-world prevalence contrast (C) datasets. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals.

Given the variation in fracture prevalence between datasets - 45.8% for competition's private test set, 4.2% for the non-contrast clinical validation dataset, and 14.5% for the contrast-enhanced clinical validation dataset; we used the bootstrap sampling method for clinical validation datasets to mirror the 45.8% prevalence rate. We performed 1000 bootstrap samples where for each bootstrap sample, we randomly sampled negative cases with replacement from each condition, sampling 166 cases from the contrast-enhanced subset and 120 negative cases from the non-contrast subset in each bootstrap sample. We retained all positive cases, 55 for the non-contrast subset and 75 for the contrast-enhanced subset. Models were then applied to each of these 1000 bootstrap samples. For the matched prevalence analysis, we calculated sensitivity, specificity, PPV, NPV, AUC, accuracy, balanced accuracy, F1-score, Kappa, and MCC.

With a simulated positive case prevalence of 45.8% to match the competition private dataset prevalence, models on the clinical validation dataset showed a reduction in AUC and accuracy on both subsets of the clinical validation dataset. For non-contrast scans, there was also a reduction in sensitivity and NPV, whereas for contrast-enhanced scans there was a reduction in specificity and PPV. On the non-contrast dataset there was a mean AUC of 0.887 (95% CI: 0.851, 0.923), and a mean accuracy of 0.81 (95% CI: 0.751, 0.870). Mean sensitivity was 0.67 (95% CI: 0.516, 0.825), and mean specificity was 0.929 (95% CI: 0.881, 0.978). For contrast-enhanced scans under a simulated prevalence, models achieved a mean AUC of 0.883 (95% CI: 0.828, 0.938), with a mean accuracy of 0.766 (95% CI: 0.659, 0.872). The mean sensitivity was high at 0.819 (95% CI: 0.643, 0.996), and the mean specificity was 0.72 (95% CI: 0.464, 0.976).

Individual Model Performances of Competition Private Test Dataset

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	318	44	**413**	**14**	0.878 (0.841, 0.908)	**0.967 (0.946, 0.980)**	**0.958 (0.930, 0.975)**	0.904 (0.873, 0.927)	**0.968 (0.955, 0.980)**	**0.926 (0.906, 0.943)**	**0.923 (0.797, 0.973)**	**0.916 (0.890, 0.943)**	**0.851 (0.814, 0.888)**	**0.854 (0.818, 0.888)**
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	310	52	412	15	0.856 (0.816, 0.889)	0.965 (0.943, 0.979)	0.954 (0.925, 0.972)	0.888 (0.856, 0.914)	0.967 (0.955, 0.979)	0.915 (0.894, 0.933)	0.911 (0.788, 0.965)	0.902 (0.874, 0.931)	0.828 (0.788, 0.867)	0.832 (0.795, 0.869)
4.Selim	319	43	404	23	0.881 (0.844, 0.911)	0.946 (0.920, 0.964)	0.933 (0.901, 0.955)	0.904 (0.873, 0.928)	0.954 (0.940, 0.969)	0.916 (0.895, 0.934)	0.914 (0.788, 0.968)	0.906 (0.878, 0.934)	0.831 (0.792, 0.870)	0.832 (0.792, 0.868)
5.Speedrun	304	58	407	20	0.840 (0.798, 0.874)	0.953 (0.929, 0.969)	0.938 (0.907, 0.960)	0.875 (0.842, 0.902)	0.955 (0.941, 0.969)	0.901 (0.878, 0.920)	0.896 (0.776, 0.956)	0.886 (0.856, 0.916)	0.799 (0.757, 0.841)	0.804 (0.763, 0.842)
6.Skecherz	323	39	376	51	0.892 (0.856, 0.920)	0.881 (0.846, 0.908)	0.864 (0.825, 0.895)	0.906 (0.874, 0.930)	0.951 (0.936, 0.967)	0.886 (0.862, 0.906)	0.886 (0.764, 0.949)	0.878 (0.847, 0.908)	0.771 (0.726, 0.815)	0.771 (0.725, 0.818)
7.QWER	**325**	**37**	395	32	**0.898 (0.862, 0.925)**	0.925 (0.896, 0.946)	0.910 (0.876, 0.936)	**0.914 (0.884, 0.937)**	0.958 (0.944, 0.973)	0.913 (0.891, 0.930)	0.911 (0.785, 0.967)	0.904 (0.876, 0.932)	0.824 (0.784, 0.863)	0.824 (0.785, 0.862)
8.Harshit	311	51	411	16	0.859 (0.819, 0.891)	0.963 (0.940, 0.977)	0.951 (0.922, 0.970)	0.890 (0.858, 0.915)	0.946 (0.928, 0.964)	0.915 (0.894, 0.933)	0.911 (0.788, 0.966)	0.903 (0.874, 0.931)	0.828 (0.788, 0.867)	0.831 (0.793, 0.869)

Supplemental Table 2. Individual ML model performances in detecting cervical spine fractures on the competition private test dataset (prevalence = 45.8%). TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values highlighted in bold represent the highest estimated mean values in each category.

All evaluated models demonstrated high performance with AUCs between 0.951-0.968 and accuracies of 0.886-0.926 (Supplemental Table 2) on the Kaggle competition private test dataset. The Qishen model secured first place in the competition based on a weighted log loss metric. It outperformed other models with estimates of an AUC of 0.968 (CI: 0.955, 0.980), accuracy of 0.926 (CI: 0.906, 0.943), sensitivity of 0.878 (CI: 0.841, 0.908), specificity of 0.967 (CI: 0.946, 0.980), PPV of 0.958 (CI: 0.930, 0.975), NPV of 0.904 (CI: 0.873, 0.927), Kappa agreement of 0.851 (CI: 0.814, 0.888), and MCC of 0.854 (CI: 0.817, 0.886). Meanwhile, the seventh-place QWER model posted the notable sensitivity and NPV, registering 0.898 (CI: 0.862, 0.925) and 0.914 (CI: 0.884, 0.937), respectively.

Individual Model Performances of Clinical Validation Dataset - Non-Contrast CT

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	36	19	62.8 (58.0, 65.0)	2.2 (0.0, 7.0)	0.655	0.965 (0.892, 1.000)	0.943 (0.837, 1.000)	0.768 (0.753, 0.774)	0.914 (0.887, 0.936)	0.823 (0.783, 0.842)	0.809 (0.773, 0.827)	0.772 (0.735, 0.791)	0.635 (0.556, 0.672)	0.664 (0.568, 0.712)
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	42	13	61.3 (58.0, 64.0)	3.7 (1.0, 7.0)	0.764	0.943 (0.892, 0.985)	0.921 (0.857, 0.977)	0.825 (0.817, 0.831)	0.9 (0.875, 0.919)	0.861 (0.833, 0.883)	0.854 (0.818, 0.882)	0.835 (0.808, 0.857)	0.716 (0.661, 0.761)	0.726 (0.665, 0.778)
4.Selim	39	16	60.1 (56.0, 64.0)	4.9 (1.0, 9.0)	0.709	0.925 (0.862, 0.985)	0.891 (0.812, 0.975)	0.79 (0.778, 0.800)	0.906 (0.876, 0.930)	0.826 (0.792, 0.858)	0.818 (0.785, 0.847)	0.789 (0.757, 0.821)	0.644 (0.576, 0.709)	0.657 (0.580, 0.733)
5.Speedrun	17	38	**64.4 (62.0, 65.0)**	**0.6 (0.0, 3.0)**	0.309	**0.99 (0.954, 1.000)**	**0.965 (0.850, 1.000)**	0.629 (0.620, 0.631)	0.794 (0.749, 0.836)	0.678 (0.658, 0.683)	0.649 (0.633, 0.655)	0.468 (0.453, 0.472)	0.316 (0.276, 0.326)	0.421 (0.352, 0.442)
6.Skecherz	42	13	53.5 (47.0, 59.0)	11.5 (6.0, 18.0)	0.764	0.823 (0.723, 0.908)	0.788 (0.700, 0.875)	0.804 (0.783, 0.819)	0.874 (0.844, 0.900)	0.796 (0.742, 0.842)	0.792 (0.736, 0.836)	0.775 (0.730, 0.816)	0.588 (0.483, 0.678)	0.589 (0.485, 0.683)
7.QWER	**44**	**11**	60.9 (56.0, 63.0)	4.1 (2.0, 9.0)	**0.8**	0.937 (0.862, 0.969)	0.916 (0.830, 0.957)	**0.847 (0.836, 0.851)**	**0.916 (0.887, 0.935)**	**0.874 (0.833, 0.892)**	**0.868 (0.831, 0.885)**	**0.854 (0.815, 0.871)**	**0.744 (0.663, 0.779)**	**0.75 (0.664, 0.788)**
8.Harshit	38	17	59.8 (55.0, 63.0)	5.2 (2.0, 10.0)	0.691	0.921 (0.846, 0.969)	0.883 (0.792, 0.950)	0.779 (0.764, 0.788)	0.896 (0.868, 0.922)	0.815 (0.775, 0.842)	0.806 (0.761, 0.830)	0.775 (0.738, 0.800)	0.622 (0.542, 0.674)	0.636 (0.546, 0.698)

Supplemental Table 3. Performance metrics of individual ML models for detecting cervical spine fractures on the non-contrast subset of the clinical validation dataset with a simulated equal prevalence as the competition private test dataset (prevalence = 45.8%). TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values in bold represent the highest estimated mean values in each category.

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	36	19	1208	45	0.655 (0.523, 0.766)	0.964 (0.952, 0.973)	0.444 (0.341, 0.553)	0.985 (0.976, 0.990)	0.914 (0.868, 0.960)	0.951 (0.938, 0.961)	0.809 (0.646, 0.908)	**0.529** **(0.428, 0.631)**	**0.505** **(0.399, 0.610)**	**0.515** **(0.400 , 0.614)**
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	42	13	1182	71	0.764 (0.637, 0.856)	0.943 (0.929, 0.955)	0.372 (0.288, 0.464)	0.989 (0.981, 0.994)	0.900 (0.846, 0.955)	0.936 (0.921, 0.948)	0.853 (0.682, 0.941)	0.500 (0.407, 0.593)	0.470 (0.373, 0.567)	0.504 (0.396, 0.582)
4.Selim	39	16	1160	93	0.709 (0.579, 0.812)	0.926 (0.910, 0.939)	0.295 (0.224, 0.378)	0.986 (0.978, 0.992)	0.907 (0.860, 0.954)	0.917 (0.900, 0.930)	0.817 (0.653, 0.914)	0.417 (0.328, 0.506)	0.380 (0.286, 0.475)	0.420 (0.330, 0.498)
5.Speedrun	17	38	**1240**	**13**	0.309 (0.203, 0.440)	**0.990** **(0.982, 0.994)**	**0.567** **(0.392, 0.726)**	0.970 (0.959, 0.978)	0.794 (0.725, 0.863)	**0.961** **(0.949, 0.970)**	0.649 (0.515, 0.763)	0.400 (0.268, 0.532)	0.382 (0.247, 0.517)	0.403 (0.270, 0.526)
6.Skecherz	42	13	1029	224	0.764 (0.637, 0.856)	0.821 (0.799, 0.841)	0.158 (0.119, 0.207)	0.988 (0.979, 0.993)	0.874 (0.812, 0.935)	0.819 (0.797, 0.839)	0.792 (0.635, 0.893)	0.262 (0.196, 0.327)	0.206 (0.122, 0.291)	0.292 (0.226, 0.351)
7.QWER	**44**	**11**	1175	78	**0.800** **(0.676, 0.884)**	0.938 (0.923, 0.950)	0.361 (0.281, 0.449)	**0.991** **(0.983, 0.995)**	**0.916** **(0.867, 0.964)**	0.932 (0.917, 0.944)	**0.869** **(0.694, 0.951)**	0.497 (0.407, 0.587)	0.466 (0.371, 0.561)	0.507 (0.417, 0.586)
8.Harshit	38	17	1154	99	0.691 (0.560, 0.797)	0.921 (0.905, 0.935)	0.277 (0.209, 0.358)	0.985 (0.977, 0.991)	0.897 (0.847, 0.946)	0.911 (0.895, 0.926)	0.806 (0.643, 0.906)	0.396 (0.308, 0.484)	0.357 (0.263, 0.451)	0.400 (0.305, 0.479)

Supplemental Table 4. Individual ML model performances in detecting cervical spine fractures on the non-contrast subset of our clinical validation dataset with a real-world prevalence rate of fractures (prevalence = 4.2%). TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values highlighted in bold represent the highest estimated mean values in each category.

Models demonstrated an AUC of 0.794-0.916 and accuracies of 0.678-0.874 (Supplemental Table 3) on the bootstrapped sampled clinical validation dataset. The seventh-place model (QWER) had the highest estimated scores with an AUC of 0.916 (CI: 0.887, 0.935), accuracy of 0.874 (CI: 0.833, 0.892), balanced accuracy of 0.868 (CI: 0.831, 0.885), sensitivity of 0.800, specificity of 0.937 (CI: 0.862, 0.969), PPV of 0.916 (CI: 0.830, 0.957), NPV of 0.847 (CI: 0.836, 0.851), F1 score of 0.854 (CI: 0.815, 0.871), Kappa agreement of 0.744 (CI: 0.663, 0.779), and MCC of 0.750 (CI: 0.664, 0.788). The fifth-place model (Speedrun) had the highest observed values of specificity and PPV across all models, of 0.990 (CI: 0.954, 1.000) and 0.965 (CI: 0.850, 1.000) respectively, but had the lowest observed values of sensitivity and NPV of 0.309 and 0.629 (CI: 0.620, 0.631) respectively.

With a real-world fracture prevalence, models demonstrated an AUC of 0.794-0.916 and accuracies of 0.819-0.961 (Supplemental Table 4). All models showed a higher accuracy, lower PPV, higher NPV, lower F1 score, lower kappa and lower MCC in the clinical validation dataset compared to the matched prevalence dataset. The first-place model (Qishen) demonstrated notable metrics with an AUC of 0.914 (CI: 0.868, 0.960), accuracy of 0.951 (CI: 0.938, 0.961), sensitivity of 0.655 (CI: 0.523, 0.766), specificity of 0.964 (CI: 0.952, 0.973), PPV of 0.444 (CI: 0.341, 0.553), NPV of 0.985 (CI: 0.976, 0.990), F1 score of 0.529 (CI: 0.428, 0.631), Kappa agreement of 0.505 (CI: 0.399, 0.610), and MCC of 0.514 (CI: 0.400, 0.614). The fifth-place model (Speedrun) achieved among the higher values in specificity at 0.990 (CI: 0.982, 0.994), PPV of 0.567 (CI: 0.392, 0.726), and accuracy of 0.961 (CI: 0.949, 0.970), but had a sensitivity of 0.309 (CI: 0.203, 0.440) which was among the lower values observed across all models.

Individual Model Performances of Clinical Validation Dataset - Contrast-Enhanced CT

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	68	7	71.8 (64.0, 77.0)	16.2 (11.0, 24.0)	0.907	0.816 (0.727, 0.875)	0.809 (0.739, 0.861)	0.911 (0.901, 0.917)	**0.935 (0.915, 0.951)**	**0.858 (0.810, 0.890)**	**0.861 (0.817, 0.891)**	**0.855 (0.814, 0.883)**	**0.717 (0.624, 0.779)**	**0.722 (0.637, 0.780)**
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	63	12	75.2 (69.0, 82.0)	12.8 (6.0, 19.0)	0.840	0.855 (0.784, 0.932)	0.833 (0.768, 0.913)	0.862 (0.852, 0.872)	0.919 (0.901, 0.935)	0.848 (0.810, 0.890)	0.848 (0.812, 0.886)	0.836 (0.803, 0.875)	0.695 (0.620, 0.776)	0.695 (0.622, 0.779)
4.Selim	56	19	81.0 (77.0, 86.0)	7.0 (2.0, 11.0)	0.747	0.921 (0.875, 0.977)	**0.890 (0.836, 0.966)**	0.810 (0.796, 0.816)	0.922 (0.900, 0.941)	0.841 (0.816, 0.871)	0.834 (0.811, 0.862)	0.812 (0.789, 0.842)	0.675 (0.627, 0.736)	0.684 (0.630, 0.754)
5.Speedrun	32	43	**81.6 (76.0, 85.1)**	**6.4 (2.9, 12.0)**	0.427	**0.928 (0.864, 0.967)**	0.837 (0.727, 0.914)	0.655 (0.639, 0.665)	0.760 (0.718, 0.803)	0.697 (0.663, 0.718)	0.677 (0.645, 0.697)	0.565 (0.538, 0.582)	0.368 (0.299, 0.410)	0.417 (0.326, 0.476)
6.Skecherz	**75**	**0**	14.3 (8.0, 22.0)	73.7 (66.0, 80.0)	**1.000**	0.162 (0.091, 0.250)	0.505 (0.484, 0.532)	**1.000 (1.000, 1.000)**	0.872 (0.835, 0.900)	0.548 (0.509, 0.595)	0.581 (0.545, 0.625)	0.671 (0.652, 0.694)	0.152 (0.084, 0.235)	0.285 (0.210, 0.365)
7.QWER	65	10	71.2 (63.0, 78.0)	16.8 (10.0, 25.0)	0.867	0.810 (0.716, 0.886)	0.797 (0.722, 0.867)	0.877 (0.863, 0.886)	0.905 (0.879, 0.924)	0.836 (0.785, 0.877)	0.838 (0.791, 0.877)	0.830 (0.788, 0.867)	0.672 (0.574, 0.753)	0.675 (0.584, 0.753)
8.Harshit	71	4	48.4 (39.0, 57.0)	39.6 (31.0, 49.0)	0.947	0.550 (0.443, 0.648)	0.643 (0.592, 0.696)	0.923 (0.902, 0.932)	0.869 (0.824, 0.905)	0.732 (0.675, 0.785)	0.748 (0.695, 0.797)	0.766 (0.728, 0.802)	0.480 (0.373, 0.579)	0.530 (0.441, 0.612)

Supplemental Table 5. Performance metrics of individual ML models for detecting cervical spine fractures on the contrast-enhanced subset of the clinical validation dataset with a simulated equal prevalence as the competition private test dataset (prevalence = 45.8%). TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values in bold represent the highest estimated mean values in each category.

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	68	7	365	81	0.907 (0.820, 0.954)	0.818 (0.780, 0.851)	0.456 (0.378, 0.536)	0.981 (0.962, 0.991)	**0.936** **(0.901,** **0.970)**	0.831 (0.797, 0.861)	**0.863** **(0.708,** **0.942)**	0.607 (0.532, 0.682)	0.514 (0.424, 0.604)	0.563 (0.486, 0.640)
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	63	12	381	65	0.840 (0.741, 0.906)	0.854 (0.818, 0.884)	0.492 (0.407, 0.578)	0.969 (0.947, 0.982)	0.918 (0.874, 0.962)	0.852 (0.819, 0.880)	0.847 (0.686, 0.933)	0.621 (0.542, 0.699)	0.537 (0.445, 0.628)	0.566 (0.477, 0.644)
4.Selim	56	19	410	36	0.747 (0.638, 0.831)	0.919 (0.890, 0.941)	**0.609** **(0.507,** **0.702)**	0.956 (0.932, 0.971)	0.922 (0.887, 0.957)	**0.894** **(0.865,** **0.918)**	0.833 (0.671, 0.924)	**0.671** **(0.587,** **0.754)**	**0.609** **(0.515,** **0.702)**	**0.612** **(0.514,** **0.703)**
5.Speedrun	32	43	**414**	**32**	0.427 (0.321, 0.539)	**0.928** **(0.900,** **0.949)**	0.500 (0.381, 0.619)	0.906 (0.876, 0.929)	0.761 (0.697, 0.825)	0.856 (0.823, 0.884)	0.677 (0.536, 0.792)	0.460 (0.358, 0.563)	0.378 (0.264, 0.492)	0.379 (0.265, 0.488)
6.Skecherz	**75**	**0**	73	373	**1.000** **(0.951,** **1.000)**	0.164 (0.132, 0.201)	0.167 (0.136, 0.205)	**1.000** **(0.950,** **1.000)**	0.873 (0.829, 0.917)	0.284 (0.247, 0.324)	0.582 (0.515, 0.646)	0.287 (0.229, 0.344)	0.053 (-0.101, 0.208)	0.165 (0.139, 0.193)
7.QWER	65	10	362	84	0.867 (0.772, 0.926)	0.812 (0.773, 0.845)	0.436 (0.359, 0.516)	0.973 (0.951, 0.985)	0.905 (0.860, 0.950)	0.820 (0.784, 0.850)	0.839 (0.684, 0.926)	0.580 (0.504, 0.657)	0.481 (0.389, 0.573)	0.529 (0.442, 0.599)
8.Harshit	71	4	245	201	0.947 (0.871, 0.979)	0.549 (0.503, 0.595)	0.261 (0.212, 0.316)	0.984 (0.959, 0.994)	0.870 (0.827, 0.913)	0.607 (0.564, 0.648)	0.748 (0.623, 0.842)	0.409 (0.342, 0.476)	0.237 (0.126, 0.348)	0.347 (0.292, 0.404)

Supplemental Table 6. Individual ML model performances in detecting cervical spine fractures on the contrast-enhanced subset of our clinical validation dataset with a real-world prevalence rate of fractures (prevalence = 14.5%). TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values highlighted in bold represent the highest estimated mean values in each category.

Models demonstrated an AUC of 0.760-0.935 and accuracies of 0.548-0.858 (Supplemental Table 5) on the bootstrapped sampled clinical validation dataset. The first-place model (Qishen) had the highest observed scores with an AUC of 0.935 (CI: 0.915, 0.951), accuracy of 0.858 (CI: 0.810, 0.890), balanced accuracy of 0.861 (CI: 0.817, 0.891), F1 score of 0.855 (CI: 0.814, 0.883), Kappa agreement of 0.717 (CI: 0.624, 0.779), and MCC of 0.722 (CI: 0.637, 0.780). The sixth-place model (Skecherz) achieved the notable sensitivity and NPV, both at 1.000 (CI: 1.000, 1.000), but its specificity and PPV were comparatively low at 0.162 (CI: 0.091, 0.250) and 0.505 (CI: 0.484, 0.532) respectively.

With a real-world fracture prevalence, models demonstrated an AUC of 0.761-0.936 and accuracies of 0.284-0.894 (Supplemental Table 6). All models showed a lower PPV, higher or unchanged NPV, lower F1 score, lower kappa and lower MCC in the real-world prevalence dataset compared to the matched prevalence dataset. There was variation in change of model accuracies between the two datasets. The fourth-place model (Selim) showed the highest estimated scores with an AUC of 0.922 (CI: 0.887, 0.957), accuracy of 0.894 (CI: 0.865, 0.918), sensitivity of 0.747 (CI: 0.638, 0.831), specificity of 0.919 (CI: 0.891, 0.941), PPV of 0.609 (CI: 0.507, 0.702), NPV of 0.956 (CI: 0.932, 0.972), F1 score of 0.671 (CI: 0.587, 0.754), Kappa agreement of 0.609 (CI: 0.515, 0.702), and MCC of 0.612 (CI: 0.514, 0.703).

Individual Model Peformances of Competition Private test set - Sensitivity test

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	15.8 (12.0, 17.0)	2.2 (1.0, 6.0)	**413**	**14**	0.878 (0.667, 0.944)	**0.967**	**0.529 (0.462, 0.548)**	0.995 (0.986, 0.998)	**0.968 (0.892, 0.997)**	**0.964 (0.955, 0.966)**	**0.922 (0.817, 0.984)**	**0.660 (0.545, 0.694)**	**0.642 (0.523, 0.677)**	**0.665 (0.532, 0.706)**
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	15.4 (12.0, 18.0)	2.6 (0.0, 6.0)	412	15	0.858 (0.667, 1.000)	0.965	0.506 (0.444, 0.545)	0.994 (0.986, 1.000)	0.967 (0.889, 0.994)	0.961 (0.953, 0.966)	0.909 (0.816, 0.982)	0.636 (0.533, 0.706)	0.617 (0.510, 0.690)	0.641 (0.521, 0.725)
4.Selim	15.9 (11.0, 17.0)	2.1 (1.0, 7.0)	404	23	0.881 (0.611, 0.944)	0.946	0.407 (0.324, 0.425)	0.995 (0.983, 0.998)	0.954 (0.890, 0.991)	0.944 (0.933, 0.946)	0.912 (0.779, 0.945)	0.557 (0.423, 0.586)	0.531 (0.391, 0.562)	0.577 (0.413, 0.613)
5.Speedrun	15.1 (12.0, 18.0)	2.9 (0.0, 6.0)	407	20	0.837 (0.667, 1.000)	0.953	0.428 (0.355, 0.459)	0.993 (0.985, 1.000)	0.954 (0.876, 0.991)	0.948 (0.942, 0.955)	0.896 (0.838, 0.977)	0.566 (0.480, 0.643)	0.542 (0.452, 0.622)	0.577 (0.473, 0.672)
6.Skecherz	16.0 (14.0, 18.0)	2.0 (0.0, 4.0)	376	51	0.890 (0.778, 1.000)	0.881	0.239 (0.215, 0.261)	0.995 (0.989, 1.000)	0.950 (0.858, 0.994)	0.881 (0.876, 0.885)	0.885 (0.829, 0.940)	0.377 (0.337, 0.414)	0.334 (0.293, 0.374)	0.424 (0.367, 0.479)
7.QWER	**16.2 (14.0, 18.0)**	**1.8 (0.0, 4.0)**	395	32	**0.900 (0.778, 1.000)**	0.925	0.336 (0.304, 0.360)	**0.995 (0.990, 1.000)**	0.960 (0.871, 0.993)	0.924 (0.919, 0.928)	0.911 (0.851, 0.963)	0.489 (0.438, 0.529)	0.457 (0.403, 0.500)	0.523 (0.455, 0.577)
8.Harshit	15.5 (12.0, 18.0)	2.5 (0.0, 6.0)	411	16	0.860 (0.667, 1.000)	0.963	0.491 (0.429, 0.529)	0.994 (0.986, 1.000)	0.946 (0.840, 0.994)	0.958 (0.951, 0.964)	0.912 (0.815, 0.981)	0.625 (0.522, 0.692)	0.604 (0.497, 0.675)	0.631 (0.510, 0.714)

Supplemental Table 7. Individual ML model performances in detecting cervical spine fractures on the competition private test dataset with a simulated equal prevalence (4.2%) as the non-contrast clinical validation dataset. TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values highlighted in bold represent the highest estimated mean values in each category.

model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	63.4 (57.0, 68.0)	8.6 (4.0, 15.0)	**413**	**14**	0.881 (0.792, 0.944)	**0.967**	**0.819** **(0.803, 0.829)**	0.980 (0.965, 0.990)	**0.968** **(0.932, 0.988)**	**0.955** **(0.942, 0.964)**	**0.923** **(0.886, 0.963)**	**0.849** **(0.797, 0.883)**	**0.822** **(0.763, 0.862)**	**0.823** **(0.763, 0.864)**
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	61.7 (54.0, 67.0)	10.3 (5.0, 18.0)	412	15	0.857 (0.750, 0.931)	0.965	0.804 (0.783, 0.817)	0.976 (0.958, 0.988)	0.967 (0.938, 0.987)	0.949 (0.934, 0.960)	0.910 (0.864, 0.948)	0.829 (0.766, 0.870)	0.800 (0.727, 0.847)	0.801 (0.728, 0.849)
4.Selim	63.5 (57.0, 69.0)	8.5 (3.0, 15.0)	404	23	0.882 (0.792, 0.958)	0.946	0.734 (0.712, 0.750)	0.979 (0.964, 0.993)	0.955 (0.921, 0.978)	0.937 (0.924, 0.948)	0.914 (0.876, 0.952)	0.801 (0.750, 0.841)	0.764 (0.705, 0.811)	0.768 (0.707, 0.820)
5.Speedrun	60.4 (54.0, 67.0)	11.6 (5.0, 18.0)	407	20	0.838 (0.750, 0.931)	0.953	0.751 (0.730, 0.770)	0.972 (0.958, 0.988)	0.954 (0.924, 0.978)	0.937 (0.924, 0.950)	0.897 (0.852, 0.935)	0.792 (0.740, 0.843)	0.755 (0.695, 0.813)	0.756 (0.695, 0.818)
6.Skecherz	64.2 (59.0, 69.0)	7.8 (3.0, 13.0)	376	51	0.892 (0.819, 0.958)	0.881	0.557 (0.536, 0.575)	0.980 (0.967, 0.992)	0.951 (0.912, 0.979)	0.882 (0.872, 0.892)	0.885 (0.850, 0.919)	0.686 (0.648, 0.719)	0.618 (0.574, 0.657)	0.644 (0.593, 0.690)
7.QWER	**64.5** **(59.0, 69.0)**	**7.5** **(3.0, 13.0)**	395	32	**0.895** **(0.819, 0.958)**	0.925	0.668 (0.648, 0.683)	**0.981** **(0.968, 0.992)**	0.958 (0.918, 0.981)	0.921 (0.910, 0.930)	0.911 (0.872, 0.942)	0.765 (0.724, 0.798)	0.718 (0.671, 0.757)	0.730 (0.677, 0.773)
8.Harshit	61.8 (55.0, 67.0)	10.2 (5.0, 17.0)	411	16	0.859 (0.764, 0.931)	0.963	0.794 (0.775, 0.807)	0.976 (0.960, 0.988)	0.946 (0.901, 0.978)	0.948 (0.934, 0.958)	0.911 (0.856, 0.947)	0.825 (0.769, 0.865)	0.794 (0.731, 0.840)	0.795 (0.731, 0.843)

Supplemental Table 8. Individual ML model performances in detecting cervical spine fractures on the competition private test dataset with a simulated equal prevalence (14.5%) as the contrast-enhanced clinical validation dataset. TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values highlighted in bold represent the highest estimated mean values in each category.

In our sensitivity analysis, we employed bootstrap sampling techniques to assess the robustness of the ML models' performance mirroring the same prevalence seen in both the non-contrast dataset (4.2%, supplemental Table 7) and the contrast-enhanced dataset (14.5%, supplemental Table 8). In this approach, all 427 negative cases were retained, and only the positive cases were sampled, resulting in 17 and 71 cases respectively. As with the previous experiments, metrics, means, standard deviations, and 95% CIs were systematically calculated. The metrics like SEN, SPEC, and AUC stayed consistent no matter the prevalence, with spans of 0.880-0.898, 0.880-0.967 and 0.947-0.967 respectively. However, with prevalence reducing from 0.458 to 0.042, we noted that PPV decreased from the ranges of 0.863-0.958% to 0.239-0.529, while NPV increased from 0.815-0.914 to 0.993-0.995, which is a typical response. Additionally, ACC saw a rise, moving from 0.886-0.926 to 0.881-0.964. Despite the models' consistent discriminatory ability, key metrics like F1, Kappa, and MCC experienced declined. They fell from their initial ranges of 0.877-0.916, 0.771-0.850, and 0.771-0.853 to 0.377-0.660, 0.334-0.642, and 0.424-0.665 respectively in the face of diminished prevalence. Although the models could effectively distinguish between positive and negative cases, their overall performance concerning accurate classifications and harmony with actual labels diminished in the context of reduced prevalence.

Prior Cervical Spine Surgery Analysis

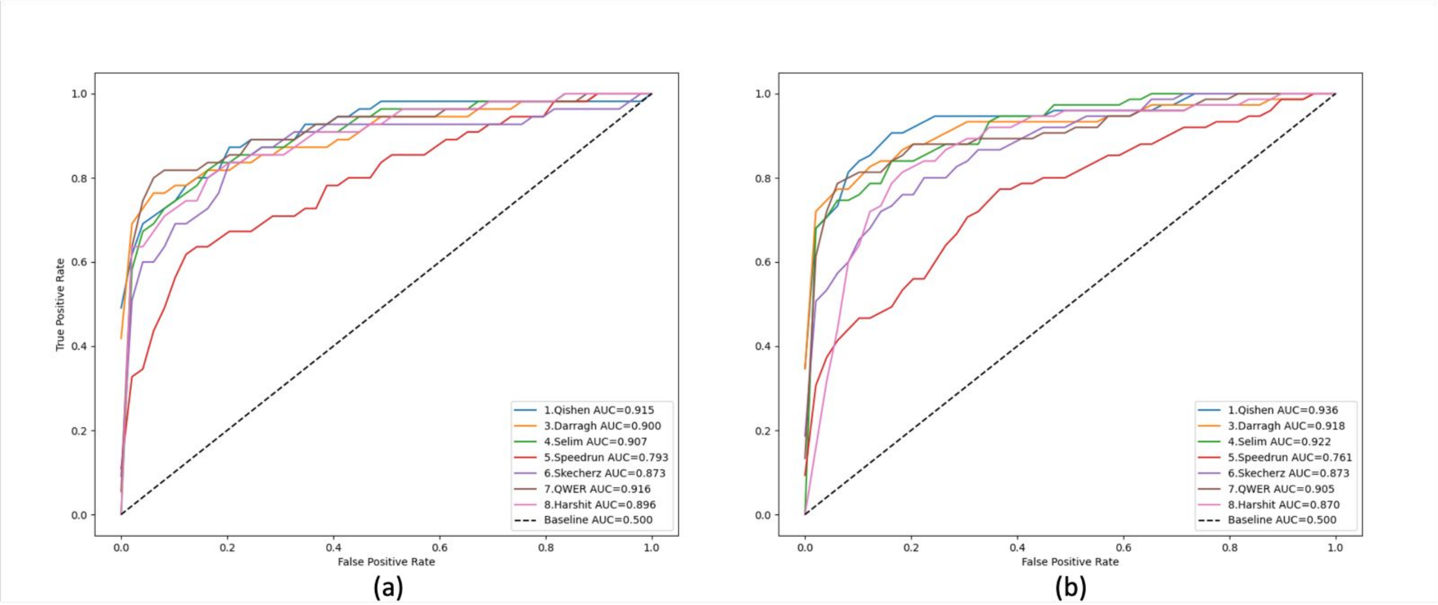
model	TP	FN	TN	FP	SEN	SPEC	PPV	NPV	AUC	Acc	BAcc	F1	Kappa	MCC
1.Qishen	1.0 (0.0, 1.0)	1.0 (1.0, 2.0)	13.1 (11.0, 15.0)	1.9 (0.0, 4.0)	0.493 (0.000, 0.500)	0.871 (0.733, 1.000)	0.352 (0.000, 1.000)	0.930 (0.857, 1.000)	0.799 (0.333, 1.000)	0.826 (0.706, 1.000)	0.682 (0.333, 0.967)	0.387 (0.000, 0.800)	0.298 (-0.202, 0.767)	0.316 (-0.169, 1.000)
2.RAWE	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3.Darragh	1.0 (0.0, 2.0)	1.0 (0.0, 2.0)	10.0 (7.0, 14.0)	5.0 (1.0, 8.0)	0.504 (0.000, 1.000)	0.669 (0.467, 0.933)	0.172 (0.000, 0.500)	0.912 (0.800, 1.000)	0.837 (0.533, 1.000)	0.650 (0.471, 0.882)	0.587 (0.267, 0.933)	0.250 (0.000, 0.571)	0.094 (-0.224, 0.605)	0.120 (-0.306, 0.658)
4.Selim	1.0 (0.0, 1.0)	1.0 (1.0, 2.0)	10.0 (7.0, 14.0)	5.0 (1.0, 8.0)	0.486 (0.000, 0.500)	0.669 (0.467, 0.933)	0.170 (0.000, 0.500)	0.908 (0.778, 1.000)	0.695 (0.267, 1.000)	0.647 (0.353, 0.824)	0.577 (0.233, 0.900)	0.244 (0.000, 0.571)	0.087 (-0.232, 0.485)	0.109 (-0.344, 0.566)
5.Speedrun	1.0 (0.0, 1.0)	1.0 (1.0, 2.0)	15.0 (15.0, 15.0)	0.0 (0.0, 0.0)	0.484 (0.000, 0.500)	1.000 (1.000, 1.000)	0.737 (0.000, 1.000)	0.937 (0.882, 0.938)	0.968 (0.867, 1.000)	0.939 (0.882, 0.941)	0.742 (0.500, 0.750)	0.568 (0.000, 0.667)	0.554 (0.000, 0.638)	0.577 (0.000, 0.685)
6.Skecherz	2.0 (2.0, 2.0)	0.0 (0.0, 0.0)	2.0 (0.0, 4.0)	13.0 (11.0, 15.0)	1.000 (1.000, 1.000)	0.133 (0.000, 0.267)	0.134 (0.133, 0.133)	0.883 (0.000, 1.000)	0.725 (0.267, 1.000)	0.235 (0.118, 0.353)	0.566 (0.500, 0.633)	0.237 (0.235, 0.235)	0.037 (0.035, 0.035)	0.124 (0.000, 0.203)
7.QWER	1.0 (1.0, 2.0)	1.0 (0.0, 1.0)	11.0 (6.0, 13.0)	4.0 (2.0, 9.0)	0.507 (0.500, 1.000)	0.731 (0.400, 0.867)	0.206 (0.000, 0.667)	0.919 (0.833, 1.000)	0.706 (0.267, 1.000)	0.705 (0.412, 0.824)	0.619 (0.333, 0.967)	0.285 (0.000, 0.571)	0.146 (-0.202, 0.767)	0.172 (-0.236, 0.789)
8.Harshit	1.0 (1.0, 2.0)	1.0 (0.0, 1.0)	8.0 (5.0, 13.0)	7.0 (2.0, 10.0)	0.507 (0.500, 1.000)	0.536 (0.333, 0.867)	0.128 (0.000, 0.333)	0.891 (0.750, 1.000)	0.740 (0.333, 1.000)	0.533 (0.353, 0.824)	0.522 (0.200, 0.867)	0.202 (0.000, 0.500)	0.020 (-0.238, 0.393)	0.029 (-0.387, 0.494)

Supplemental Table 9. Individual ML model performances in detecting cervical spine fractures on the CT scans with prior cervical spine surgery in the clvalidation dataset. TP true positive, FN false negative, TN true negative, FP false positive, SEN sensitivity, SPEC specificity, PPV positive predictive value, NPV negative predictive value, AUC area under the receiver operating curve, Acc accuracy, BAcc balanced accuracy, F1 F1 score, Kappa Kappa coefficient, and MCC Matthews correlation coefficient. Metrics are displayed as the means, accompanied by their corresponding 95% confidence intervals, calculated using Efron's accelerated bootstrap method. Values highlighted in bold represent the highest estimated mean values in each category.

We examined the influence of surgical intervention and the presence of post-operative hyperdense materials such as screws, rods, and surgical clips on the performance of ML models. Out of 17 cases (comprising 14 non-contrast scans and 3 contrast scans) where patients had undergone cervical spine surgery, 16 had postoperative material present. Among these, 2 cases (1 non-contrast scan and 1 contrast-enhanced scan) were identified as positive. The detailed metrics in this condition can be found in supplemental Table 9.

Clinical Validation Dataset – ROC Curve

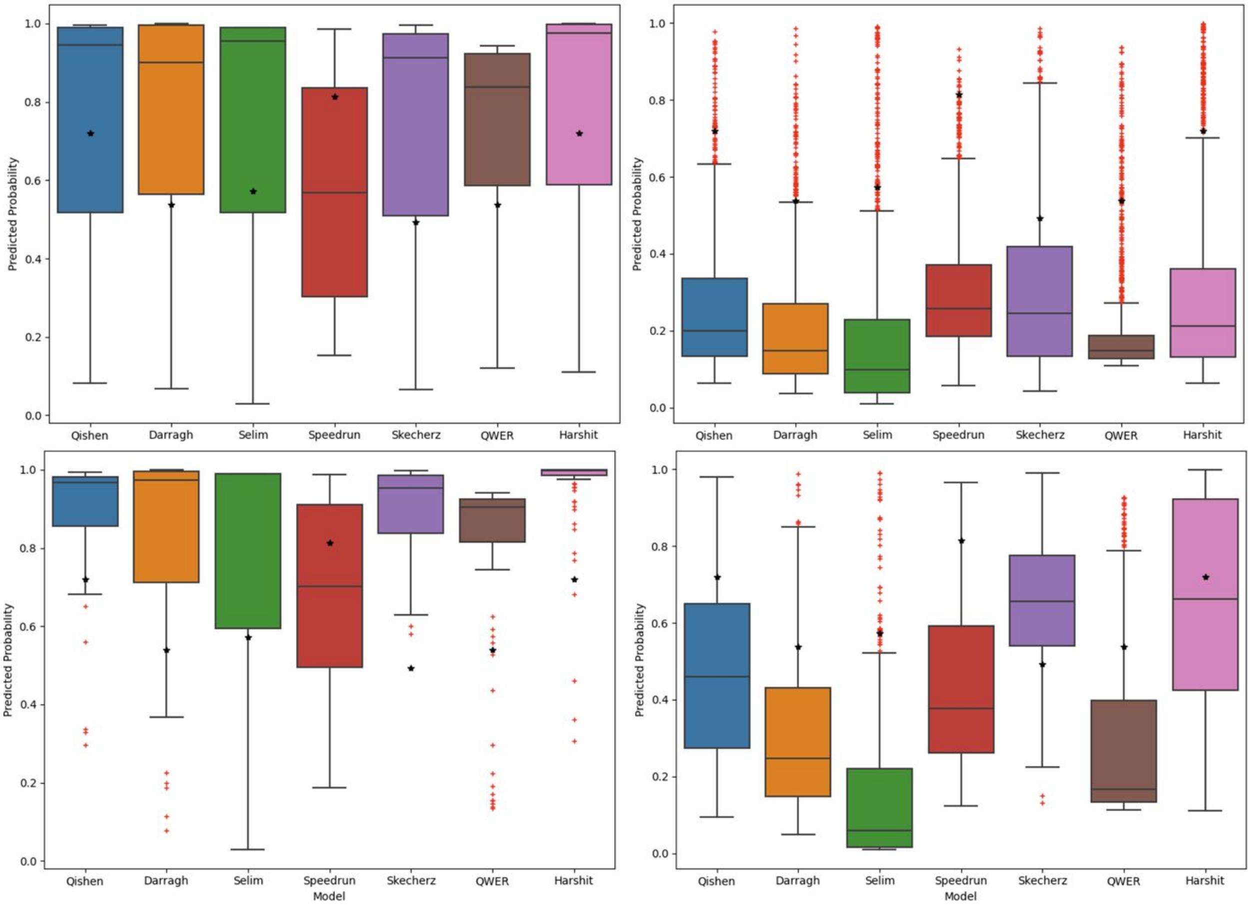
ROC curves were generated for the clinical validation dataset based on model predictions for both contrast and non-contrast conditions with real-world prevalence.



Supplemental figure 2. Receiver operating characteristic (ROC) curves for the clinical validation dataset. (a) Probabilities generated per model for non-contrast studies, (b) probabilities for contrast-enhanced studies

Clinical Validation Dataset - Distribution of cervical spine fracture probabilities

Supplemental figure 3 illustrates the distribution of cervical spine fracture probabilities as predicted by ML models on the clinical validation dataset, evaluated at the patient level.

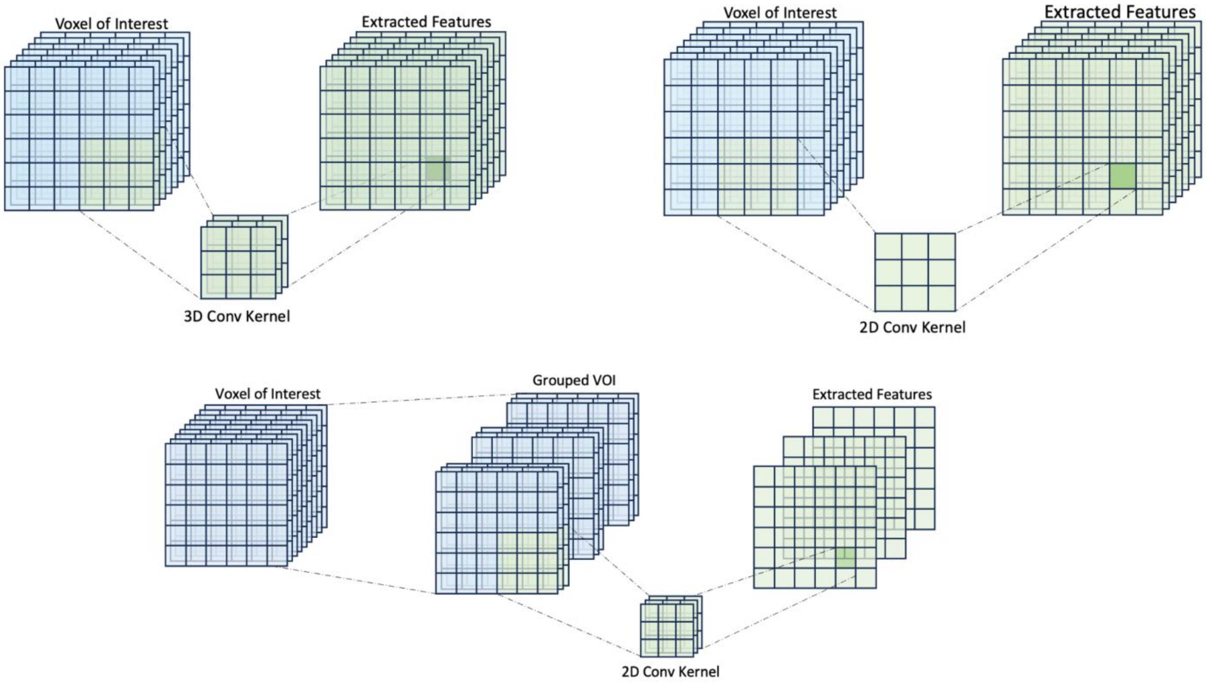


Supplemental figure 3. Probability distribution of predicted labels for ground truth cases. (a) top-left, Positive ground truth in non-contrast studies, (b) top-right, negative ground truth in non-contrast studies, (c) bottom-left, positive ground truth in contrast-enhanced studies, and (d) bottom-right, negative ground truth in contrast-enhanced studies. The central red line signifies the median, while the box's lower and upper edges represent the 25th and 75th percentiles, respectively. Whiskers extend to the furthest data points not classified as outliers, which are individually marked with red "+" symbols. The threshold determined by the Bayesian optimizer is indicated by a black "*" symbol. Cases with probabilities exceeding this threshold are categorized as either positive or negative.

Appendix S2

ML Model Overview

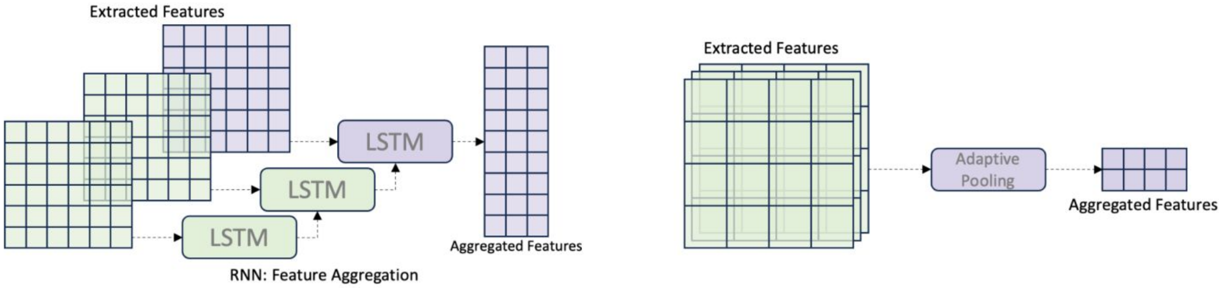
In the segmentation stage, the original CT images are processed to isolate the voxels of interest (VOI), which corresponds to the location of the cervical spine. Typically, a U-Net architecture [1] is employed for this task. Following this, the classification stage involves feature extraction from the isolated VOI, feature aggregations, and logits prediction.



Supplemental Figure 4. Illustration of feature extraction strategies: a) 3D CNN applied to 3-dimensional voxels, b) 2.5D CNN with multiple channels capturing information along the z-axis within a 2D CNN framework, and c) 2D CNN applied to image-level images.

In terms of feature extraction, three main strategies are widely adopted: 3D CNN applied to 3-dimensional voxels, 2-dimensional CNN used on 2-dimensional axial images, and a variant known as multi-channel 2D CNN or 2.5D CNN. In this 2.5D CNN approach, multiple channels encapsulate information along the *z*-axis, allowing for a user-defined convolutional depth while still utilizing a 2D CNN framework.

When features are extracted from 2D or 2.5D images, further aggregation is often necessary as CT scans and fractures are inherently 3D. Recurrent Neural Networks (RNNs), particularly long short-term memory (LSTM) or gated recurrent units (GRUs) structures, are popular choices for this aggregation. These RNN architectures are adept at handling sequences, thus providing a natural way to aggregate 2D features over a sequence of images to form a 3D representation. Conversely, if feature extraction is performed in 3D space, adaptive pooling is often used. Adaptive pooling allows for the compression and aggregation of features across different dimensions, effectively condensing the 3D feature space into a form that is more manageable for classification. This enables the model to retain critical spatial hierarchies while reducing computational complexity.



Supplemental Figure 5. Illustration of feature aggregation methods: a) RNN structures like LSTM or GRUs, ideal for aggregating 2D features over a sequence to form a 3D representation, and b) adaptive pooling, used for compressing and aggregating features in 3D space.

In the final layer, fully connected (FC) layers with activation functions are employed to form a dense network that takes the aggregated features to make logit predictions. The logits generated go through a sigmoid function to yield the final output.

Visualization of predicted areas of fractures was performed using feature maps generated from the last CNN block before the feature aggregation step in models using the GradCAM method [2]. Such visualizations are useful in serving as a validation tool, ensuring that the ML model is effectively focusing on the correct regions.

Deep Dive: Insights into Individual ML Model Solutions

Team Qishen Ha

The first model, developed by Qishen Ha, winner of the RSNA 2022 Cervical Spine Fracture Detection competition, employs a two-stage pipeline. The first stage uses 3D semantic segmentation with ResNet-18 [3] or EfficientNetV2-s [1] and U-Net model, trained on 87 studies of segmentation samples with a size of 128x128x128 for segmenting cervical vertebras. The subsequent stage applies a 2.5D classification with LSTM, offering two model variants that consider either a single vertebra or an entire patient as a training sample. Preprocessing of the second stage involves using predicted 3D masks to crop out vertebras from an original 3D image and extracting adjacent images to form images with 5 channels. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/362607

Team RAWE

The second-place model, developed by Ryan Rong, employs a two-stage pipeline. The segmentation stage employs a 2.5D CNN with a U-Net for segmentation, based on EfficientNet-b0. The classification stage is a binary classification phase that incorporates a combination of a 2D CNN (based on ResNet-50 [3] or EfficientNetV2-s [4]), Bidirectional Gated Recurrent Unit (BiGRU), and an Attention mechanism. The input data for this stage consists of 24 uniformly distributed slices of a single vertebra, cropped by the vertebra VOI predicted from the first stage along the cranio-caudal axis. Ryan and his team assumed that fractures in the seven segments occur independently and aggregated these predicted fracture probabilities into a case-level prediction. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/365115

Team Darragh

The third-place model, developed by Darragh, also employs a two-stage pipeline. The segmentation stage involves a 2.5D CNN with EfficientNet-v2 [4] for bounding box prediction, which is used to crop the vertebrae in the preprocessing step. The classification stage involves a 2.5D CNN with a 1D RNN for fracture prediction. For the model of classification stage, Darragh first trained an intermediate model to predict the visibility ratio of the vertebra and the fracture probability for each image. The final model of the second stage was then derived by using the feature extractor head of the intermediate model, trained on the vertebra fracture ground truth. The model is noted for its unique approach of using vertebrae ratio as an additional target for training, which provides a valuable insight for future work in this field. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/362643

Team Selim Seferbekov

The fourth-place model, developed by Selim Seferbekov, utilizes a two-stage pipeline, drawing inspiration from Tran et al.'s work on Channel-Separated Convolutional Networks (CSN), including interactions preserved (IP) and interactions reduced (IR) CSN [5]. The first stage involves a 3D segmentation using a U-Net-like multiclass segmentation with an IR-CSN-50 encoder and a standard U-Net decoder with (2+1) D convolutions. The second stage involves a 3D classification for each vertebra crop using IR(IP)-CSN-152 with global max pooling. The input data for this stage are 40 images around each vertebra. The model is noted for its approach of using a pretrained CSN network from the ig65m+kinetics. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/364837

Team Speedrun

The model that secured the 5th place in the competition shares a similar idea with the 3rd place model, both aiming to predict the fracture ratio of each vertebra at the image-level and then aggregating these image-level predictions into a case-level prediction. The 5th place model employs a three-stage pipeline. In the first stage, it uses a 2D EfficientNet-B5 [4] based CNN for vertebrae classification and an EfficientNet-B3-UNet for vertebrae segmentation. The second stage involves a hybrid 2.5D and 3D CNN for predicting fractures at the image-level. images that includes a stack of interest along with two neighboring stacks, each stack containing three consecutive images. In the final stage, a simple feedforward neural network is used to aggregate the image-level predictions into a case-level prediction. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/363232

Team Skecherz

Ian Pan developed a three-stage model for cervical spine fracture detection using 3D CNN and Temporal-Depth CNNs (TD-CNNs). The first stage of the model focuses on cervical spine segmentation, effectively isolating the region of interest. In the second stage, the model employs two distinct strategies for feature extraction. One strategy utilizes a 3D CNN network to extract spatial features, while the other employs a Temporal-Depth CNN, a 2D CNN feature extractor combined with a sequence model head, specially, to capture cranio-caudal direction patterns. The final stage of the mode used transformers to combine features extracted from the 3D CNN and TD-CNN to predict fracture probabilities. This innovative approach breaks down the complex task into manageable and readable subcomponents, making the process more comprehensible. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/362651

Team qwer

The 7th place solution employs a three-stage 3D segmentation method for fracture detection. In the first stage, the model segments the cervical vertebrae using a 3D-nnUNet. The second stage involves generating pseudo fracture segmentation masks using data augmentations and bounding box labels, and then using another 3D-nnUNet to segment the bone fracture region. In the final stage, the outputs from the previous stages are grouped and a small 3D-CNN is trained to get the final fracture prediction at case-level. This model's approach, particularly its innovative use of data augmentation for pseudo-labeling and the application of a 3D nnU-Net for fracture region prediction, could potentially be applied to other similar tasks in the future. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/364848

Team Harshit Sheoran

Harshit Sheoran, who secured the 8th position, implemented a three-stage model consisting of two segmentation stages and one classification stage. For the segmentation stage, it involved training a U-Net binary segmentation model on sagittal view images to classify which image corresponds to a specific vertebra. A second U-Net model was then trained on axial view images for bone segmentation and ROI bounding box prediction. In the classification stage, an efficientnet-b5 model was employed for the extraction of image features. Then GRU bidirectional layers, an Attention mechanism, and a 1D CNN were employed for the feature aggregations. More details can be found at https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/discussion/362669

Kaggle Evaluation

The model performance during the competition was evaluated using a weighted multi-label logarithmic loss function. The function took into account the ground truth labels and applied specific weights to each label. The binary weighted log loss function for label *j* on exam *i* is formulated as:

$$L_{ij} = -w_j * [y_{ij} * lo(p_{ij}) + (1 - y_{ij}) * log(1 - p_{ij})]$$

Different weights were assigned for negative and positive segments and patients. The competition's metric was an average of this weighted log loss across all predictions, with a lower score indicating better model performance.

References

1. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.
2. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
4. Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In International conference on machine learning (pp. 10096-10106). PMLR.
5. Tran, D., Wang, H., Torresani, L., & Feiszli, M. (2019). Video classification with channel-separated convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5552-5561).

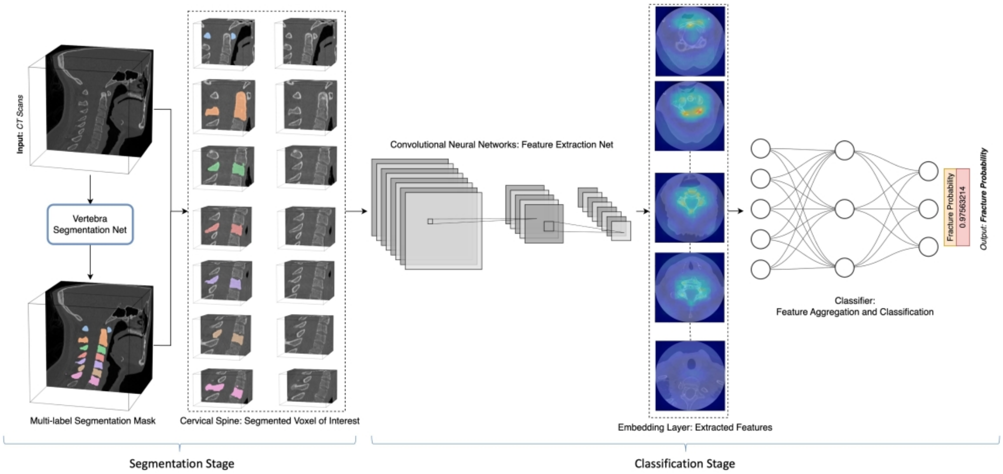

Input: CT Scans
Vertebra Segmentation Net
Multi-label Segmentation Mask
Cervical Spine: Segmented Voxel of Interest
Convolutional Neural Networks: Feature Extraction Net
Embedding Layer: Extracted Features
Classifier:
Feature Aggregation and Classification
Fracture Probability
0.97563214
Output: Fracture Probability
Segmentation Stage
Classification Stage

Figure 1. An example of end-to-end architecture of a cervical spine CT fracture detection ML model, showcasing the segmentation stage to isolate the cervical spine's voxels of interest (VOI), followed by the classification stage for feature extraction, aggregation, and logits prediction.

324x156mm (330 x 330 DPI)

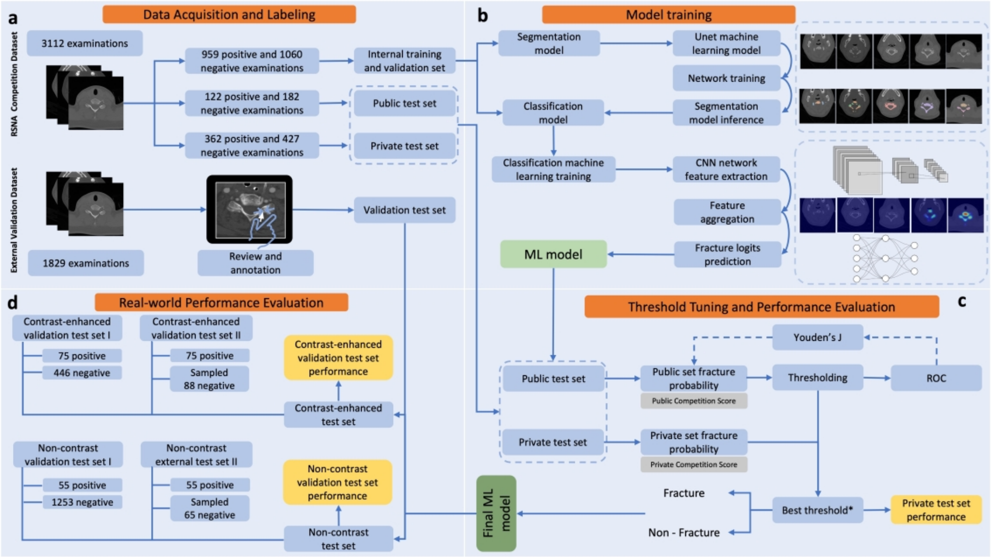

a
Data Acquisition and Labeling
RSNA Competition Dataset
3112 examinations
959 positive and 1060 negative examinations
Internal training and validation set
122 positive and 182 negative examinations
Public test set
362 positive and 427 negative examinations
Private test set
External Validation Dataset
1829 examinations
Review and annotation
Validation test set
b
Model training
Segmentation model
Unet machine learning model
Network training
Segmentation model inference
Classification model
Classification machine learning training
CNN network feature extraction
Feature aggregation
Fracture logits prediction
ML model
c
Threshold Tuning and Performance Evaluation
Youden's J
Public test set
Public set fracture probability
Public Competition Score
Thresholding
ROC
Private test set
Private set fracture probability
Private Competition Score
Fracture
Non - Fracture
Best threshold*
Private test set performance
Final ML model
d
Real-world Performance Evaluation
Contrast-enhanced validation test set I
75 positive
446 negative
Contrast-enhanced validation test set II
75 positive
Sampled 88 negative
Contrast-enhanced validation test set performance
Contrast-enhanced test set
Non-contrast validation test set I
55 positive
1253 negative
Non-contrast external test set II
55 positive
Sampled 65 negative
Non-contrast validation test set performance
Non-contrast test set

Figure 2. ML evaluation pipeline for cervical spine fracture detection. (a) The process starts with 3,112 CT scans from the RSNA 2022 competition, divided into training, public test, and private test datasets. An additional 1,829 CT scans from our institution are used as a clinical validation dataset. (b) Each ML model has two main stages: segmentation and classification involving CNN feature extraction, feature aggregation and logits prediction. (c) Each model generates a fracture probability output binarized by applying an optimal threshold identified by Youden's J on the public test dataset. Then, the model's performance is assessed using the private test dataset. (d) The final evaluation comprises four clinical validation datasets: non-contrast scans, contrast-enhanced scans, bootstrap-sampled non-contrast scans, and bootstrap-sampled contrast-enhanced scans.

344x193mm (330 x 330 DPI)

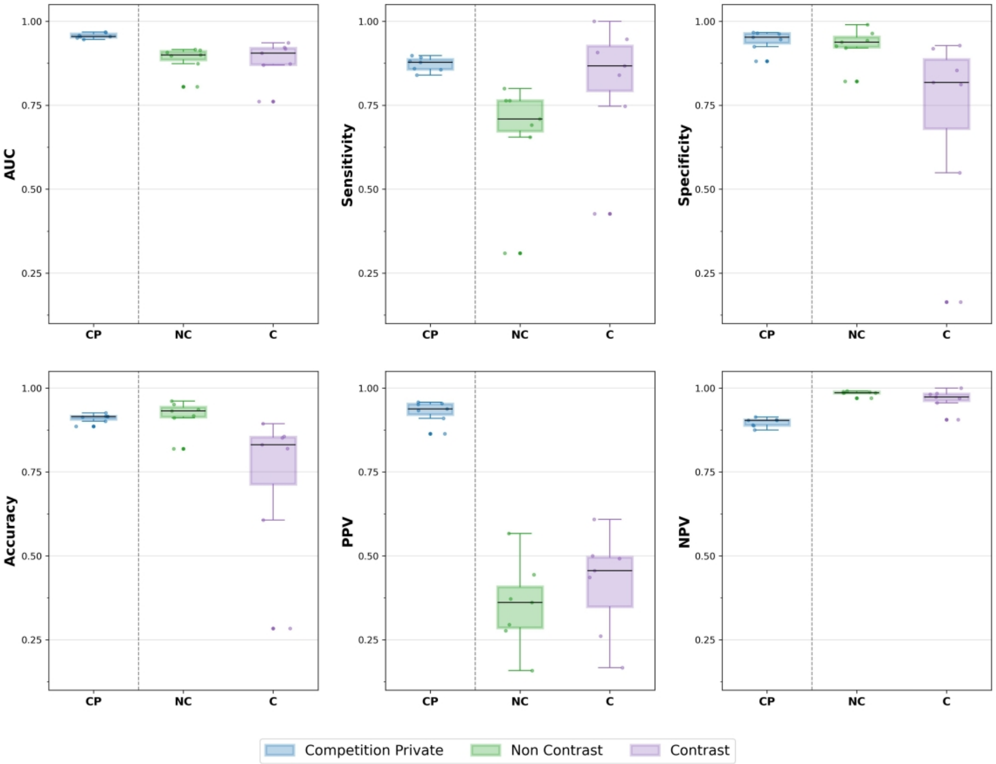


Figure 3. Box and whisker plots showcasing the distribution of performance metrics for binary classification by the winning algorithms on different datasets. Metrics include AUC, sensitivity, specificity, and accuracy. These are displayed across the competition private dataset (CP), clinical validation non-contrast (NC), and clinical validation contrast (C) datasets. The box represents the interquartile range, the median is indicated by the line within the box, and the whiskers show the full range excluding outliers, which are depicted as individual points. Data points are also visualized as jitters for clarity.

378x291mm (300 x 300 DPI)

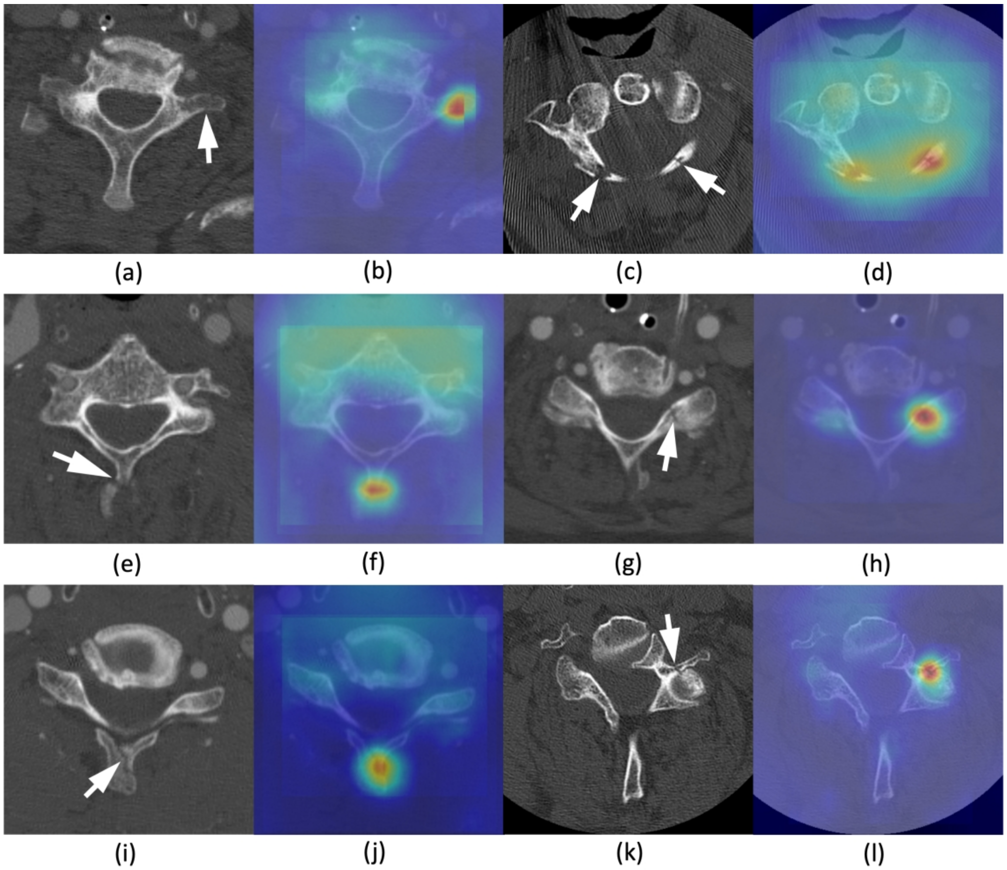


Figure 4. Example cases of fractures identified by the ML model but missed by reporting radiologists. The CT images with associated Grad-CAM heatmaps show the most influential regions in the input image for the prediction. (a,b) Minimally-displaced left transverse process fracture. (c,d) Bilateral lamina fractures, moderately-displaced on the right and undisplaced on the left. (e,f) Mildly displaced spinous process fracture. (g,h) Undisplaced left articular process/lamina fracture. (i,j) Minimally displaced spinous process fracture, and (k,l) minimally displaced left transverse process fracture.

800x700mm (96 x 96 DPI)

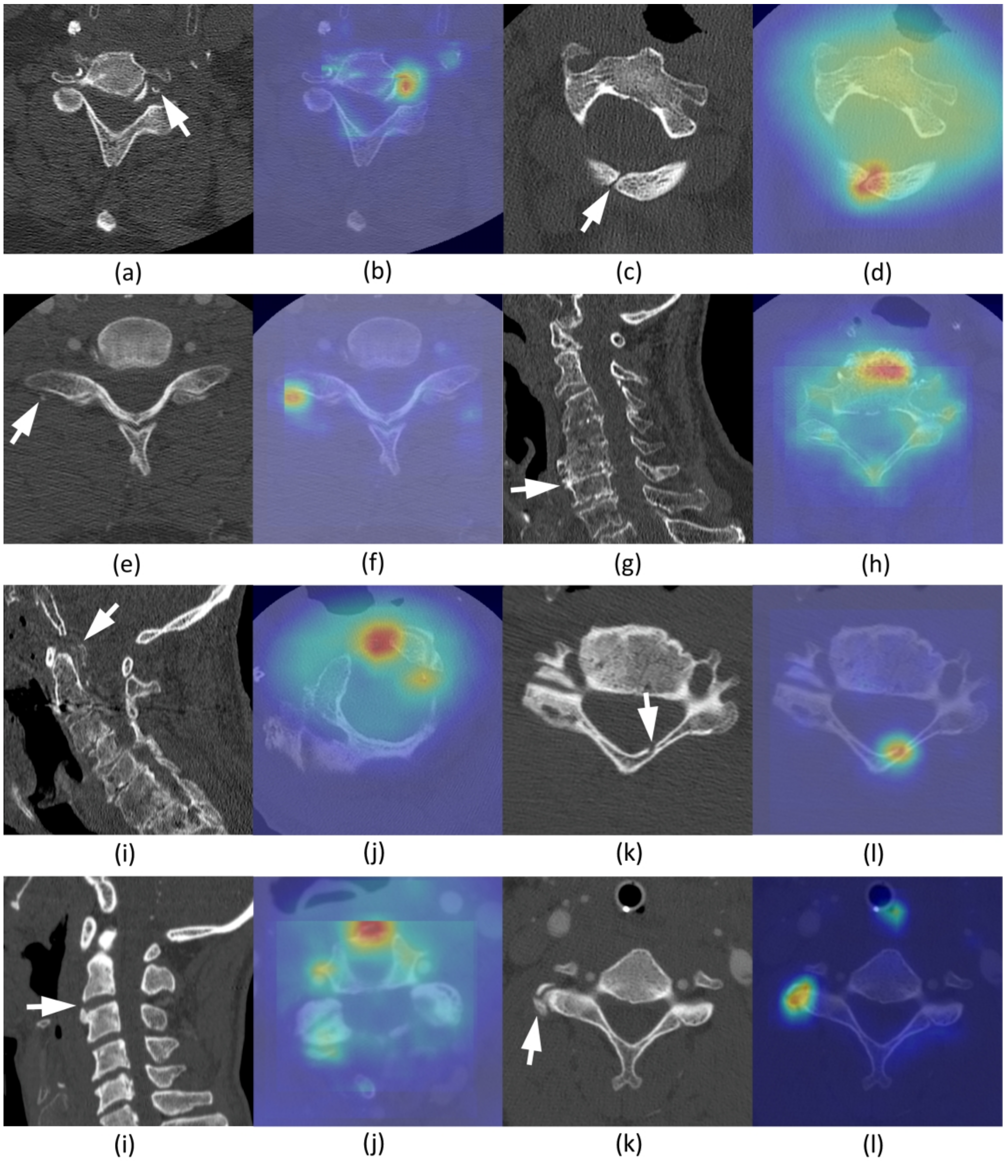


Figure 5. Example cases incorrectly identified as fractures by the ML model in the false positive group. The CT images with associated Grad-CAM heatmaps show the most influential regions in the input image for the prediction. (a,b) Calcified atherosclerotic plaque in the left vertebral artery in the left transverse foramen. (c,d) Congenital lack of fusion of the posterior arch of C1. (e,f) Contrast within a small vessel in the right paraspinal region on a contrast-enhanced CT examination. (g,h) Chronic multilevel degenerative changes with reduced intervertebral disc spaces, osteophyte formation and osteopenia. (i,j) Partially calcified pseudomass posterior to the odontoid process of C2, secondary to calcium pyrophosphate dihydrate crystal deposition disease. (k,l) Nutrient vessel within the left lamina. (m,n) Chronic osteophytes arise from the superior-anterior vertebral body of C3, and (o,p) chronic osteophytic changes associated with the right articular process.

800x935mm (96 x 96 DPI)

CLAIM: Checklist for Artificial Intelligence in Medical Imaging

Section / Topic	No.	Item	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	x
	2	Structured summary of study design, methods, results, and conclusions	x
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	X
	4	Study objectives and hypotheses	x
METHODS			
Study Design	5	Prospective or retrospective study	X
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	x
Data	7	Data sources	X
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	X
	9	Data pre-processing steps	X
	10	Selection of data subsets, if applicable	N/A
	11	Definitions of data elements, with references to Common Data Elements	
	12	De-identification methods	X
	13	How missing data were handled	X
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	X
	15	Rationale for choosing the reference standard (if alternatives exist)	X
	16	Source of ground-truth annotations; qualifications and preparation of annotators	X
	17	Annotation tools	X
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	N/A
Data Partitions	19	Intended sample size and how it was determined	N/A
	20	How data were assigned to partitions; specify proportions	N/A
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	N/A

Model	**22**	Detailed description of model, including inputs, outputs, all intermediate layers and connections	**X**
	23	Software libraries, frameworks, and packages	**X**
	24	Initialization of model parameters (e.g., randomization, transfer learning)	**X**
Training	**25**	Details of training approach, including data augmentation, hyperparameters, number of models trained	**X**
	26	Method of selecting the final model	**X**
	27	Ensembling techniques, if applicable	**X**
Evaluation	**28**	Metrics of model performance	**X**
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	**X**
	30	Robustness or sensitivity analysis	**X**
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	**X**
	32	Validation or testing on external data	**X**
RESULTS			
Data	**33**	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	**X**
	34	Demographic and clinical characteristics of cases in each partition	**X**
Model performance	**35**	Performance metrics for optimal model(s) on all data partitions	**X**
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	**X**
	37	Failure analysis of incorrectly classified cases	**X**
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	**X**
	39	Implications for practice, including the intended use and/or clinical role	**X**
OTHER INFORMATION			
	40	Registration number and name of registry	**N/A**
	41	Where the full study protocol can be accessed	**X**
	42	Sources of funding and other support; role of funders	**X**

Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2(2):e200029. https://doi.org/10.1148/ryai.2020200029