Title

Assessing the Performance of Models from the 2022 RSNA Cervical Spine Fracture Detection Competition at a Level I Trauma Center.

Journal
Radiology: Artificial Intelligence, 6(6)

Authors
Hu, Zixuan
Patel, Markand
Ball, Robyn
et al.

Publication Date
2024-11-01

DOI
10.1148/ryai.230550

Peer reviewed

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Assessing the Performance of Models from the 2022 RSNA Cervical Spine Fracture Detection Competition at a Level I Trauma Center

Zixuan Hu, MSc • Markand Patel, BM, BSc, FRCR, PhD* • Robyn L. Ball, PhD • Hui Ming Lin, HBSc • Luciano M. Prevedello, MD • Mitra Naseri, MD • Shobhit Mathur, MD • Robert Moreland, MD, MSc • Jefferson Wilson, MD, PhD • Christopher Witiw, MD, MS • Kristen W. Yeom, MD • Qishen Ha, MEng • Darragh Hanley, BAI • Selim Seferbekov, MEng • Hao Chen, MSc • Philipp Singer, PhD • Christof Henkel, PhD • Pascal Pfeiffer, PhD • Ian Pan, MD • Harshit Sheoran, HBSc • Wuqi Li, MSc • Adam E. Flanders, MD • Felipe C. Kitamura, MD • Tyler Richards, MD • Jason Talbott, MD, PhD • Ervin Sejdić, PhD** • Errol Colak, MD, FRCPC***

From the Edward S. Rogers Department of Electrical and Computer Engineering (Z.H., W.L., E.S.), Department of Medical Imaging, Faculty of Medicine (M.P., S.M., R.M., E.C.), Faculty of Medicine (M.N., J.W., C.W.), and Division of Neurosurgery, Department of Surgery (J.W., C.W.), University of Toronto, 40 St George St, Toronto, ON, Canada M5S 3G4; Department of Medical Imaging (H.M.L., M.N., S.M., R.M., E.C.) and Li Ka Shing Knowledge Institute (S.M., J.W., C.W., E.C.), St Michael's Hospital, Unity Health Toronto, Toronto, Canada; The Jackson Laboratory, Bar Harbor, Maine (R.L.B.); Standard School of Medicine, Stanford University, Stanford, Calif (K.W.Y.); H2O.ai, Mountain View, Calif (Q.H., P.S., P.P.); School of Computer Science, University of Birmingham, Birmingham, UK (H.C.); DoubleYard, Edulab Group, Boston, Ireland (D.H.); Mapbox, London, UK (S.S.); NVIDIA, Santa Clara, Calif (C.H.); Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Mass (I.P.); University of London, Goldsmiths, London, UK (H.S.); Department of Radiology, The Ohio State University, Columbus, Ohio (L.M.P.); Department of Radiology, Division of Neuroradiology, Thomas Jefferson University, Philadelphia, Pa (A.E.F.); Universidade Federal de São Paulo (Unifesp), São Paulo, Brazil (F.C.K.); Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, Calif (J.T.); Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, Utah (T.R.); and North York General Hospital, Toronto, Canada (E.S.). Received December 1, 2023; revision requested February 26, 2024; revision received July 25; accepted September 6. **Address correspondence to** E.C. (email: *Errol.Colak@UnityHealth.to*).

* Z.H. and M.P. contributed equally to this work.

** E.S. and E.C. are co–senior authors.

E.C. supported by the Odette Professorship in Artificial Intelligence for Medical Imaging, St Michael's Hospital, Unity Health Toronto.

Conflicts of interest are listed at the end of this article.

See also commentary by Levi and Politi in this issue.

Radiology: Artificial Intelligence 2024; 6(6):e230550 • https://doi.org/10.1148/ryai.230550 • Content codes: AI NR

Purpose: To evaluate the performance of the top models from the RSNA 2022 Cervical Spine Fracture Detection challenge on a clinical test dataset of both noncontrast and contrast-enhanced CT scans acquired at a level I trauma center.

Materials and Methods: Seven top-performing models in the RSNA 2022 Cervical Spine Fracture Detection challenge were retrospectively evaluated on a clinical test set of 1828 CT scans (from 1829 series: 130 positive for fracture, 1699 negative for fracture; 1308 noncontrast, 521 contrast enhanced) from 1779 patients (mean age, 55.8 years ± 22.1 [SD]; 1154 [64.9%] male patients). Scans were acquired without exclusion criteria over 1 year (January–December 2022) from the emergency department of a neurosurgical and level I trauma center. Model performance was assessed using area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. False-positive and false-negative cases were further analyzed by a neuroradiologist.

Results: Although all seven models showed decreased performance on the clinical test set compared with the challenge dataset, the models maintained high performances. On noncontrast CT scans, the models achieved a mean AUC of 0.89 (range: 0.79–0.92), sensitivity of 67.0% (range: 30.9%–80.0%), and specificity of 92.9% (range: 82.1%–99.0%). On contrast-enhanced CT scans, the models had a mean AUC of 0.88 (range: 0.76–0.94), sensitivity of 81.9% (range: 42.7%–100.0%), and specificity of 72.1% (range: 16.4%–92.8%). The models identified 10 fractures missed by radiologists. False-positive cases were more common in contrast-enhanced scans and observed in patients with degenerative changes on noncontrast scans, while false-negative cases were often associated with degenerative changes and osteopenia.

Conclusion: The winning models from the 2022 RSNA AI Challenge demonstrated a high performance for cervical spine fracture detection on a clinical test dataset, warranting further evaluation for their use as clinical support tools.

Supplemental material is available for this article.

Traumatic cervical spinal injuries are common and associated with high morbidity and mortality rates (1). The incidence of cervical spine injuries is 16.5 per 100 000 individuals (2), and the prevalence is 1.7%–3.7% in patients with blunt trauma (3,4). CT is the reference standard modality for detection of cervical spine fractures (5), with trauma patients often undergoing CT examinations covering their entire neuroaxis, chest, abdomen, and pelvis. In busy clinical environments, where radiologists are facing increasing workloads, delays between imaging and interpretation can potentially lead to adverse outcomes (6). Up to a quarter of patients experience progression of their injuries due to delays in diagnosis or unwarranted manipulation (7). Early immobilization of unstable injuries can prevent neurologic deterioration (8), and prompt surgical intervention is associated with better outcomes (9).

The increasing volume of imaging studies and demand for rapid diagnosis have led to the exploration of machine learning (ML) to aid the imaging review process. For example, ML models can assist radiologists in the detection

Abbreviations

AUC = area under the receiver operating characteristic curve, Grad-CAM = gradient-weighted class activation mapping, ML = machine learning

Summary

Winning machine learning models from the RSNA 2022 Cervical Spine Fracture Detection competition demonstrated high performance on a large clinical test set of emergency department cervical spine CT scans from a level I trauma center.

Key Points

- Seven of the top-performing machine learning models from the RSNA 2022 Cervical Spine Fracture Detection artificial intelligence challenge generalized well to the clinical test dataset, with mean area under the receiver operating characteristic curve values of 0.89 (range: 0.79–0.92) for fracture detection on noncontrast CT scans and 0.88 (range: 0.76–0.94) on contrast-enhanced CT scans.
- The models achieved a mean sensitivity of 67.0% (range: 30.9%–80.0%) and mean specificity of 92.9% (range: 82.1%–99.0%) on noncontrast CT scans and a mean sensitivity of 81.9% (range: 42.7%–100.0%) and mean specificity of 72.1% (range: 16.4%–92.8%) on contrast-enhanced CT scans.
- The machine learning models identified 10 fractures missed by reporting radiologists out of 116 cases, and poorer model performance was most often attributed to contrast-enhanced scans and scans in patients with degenerative changes and osteopenia.

Keywords

Feature Detection, Supervised Learning, Convolutional Neural Network (CNN), Genetic Algorithms, CT, Spine, Technology Assessment, Head/Neck

and characterization of abnormalities, such as brain tumors (10), wrist fractures (11), and intracranial hemorrhage (12). Studies have also explored the use of ML for detection of spinal fractures with deep neural network models (13) showing high sensitivities (>95%) (14) and some matching the performance of radiologists (15). However, most of these studies focused on osteoporotic vertebral fractures, which are more likely to be stable and rarely found in the cervical spine. To date, there are relatively few studies exploring the application of ML models to aid cervical spine fracture detection in the acute trauma setting.

A major factor preventing widespread clinical implementation of ML models is the limited access to data. Clinical data are often fragmented, stored in disparate systems, and subject to privacy regulations (16), making it challenging to access a sufficiently large and diverse dataset. Even when accessible, data may have quality issues and biases (16,17). In addition, data annotation can be labor intensive and expensive and requires substantial medical expertise and time (16).

Initiatives such as the RSNA artificial intelligence challenges play a crucial role in mitigating some of these issues and have been ongoing for several years (18). Through these competitions, the RSNA is able to crowdsource multi-institutional and multinational datasets, expertise, and insights into relevant clinical issues. Top-performing models from prior RSNA competitions have been shown to generalize well to real-world external testing datasets (19,20). 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 (21) with 1108 global 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 (22).

This study examines the performance of the top ML models from the RSNA 2022 Cervical Spine Fracture Detection competition on a clinical validation dataset. Although the RSNA competition dataset is based on real-world multi-institutional data, it was curated with the intention of hosting a competition. Each contributing site was requested to provide an equivalent number of positive and negative cases, resulting in a substantially higher fracture prevalence than real-world rates. The identification and extraction of data were left to the discretion of each site (21), which introduces the potential for selection biases. The data also underwent filtration during curation, removing examinations with incomplete coverage of the cervical spine, prior surgery, intravenous contrast material, and motion artifacts. The clinical validation dataset in this study included all consecutive emergent CT scans that were acquired over the course of a calendar year at a busy urban neurosurgical and level I trauma center. In contrast to the RSNA competition dataset, this test set includes contrast-enhanced scans, as patients often receive contrast material as part of full-body trauma imaging at major trauma centers.

Materials and Methods

This retrospective study was approved by the institutional review board at Unity Health Toronto with a waiver of informed consent.

RSNA 2022 Competition Dataset

The RSNA 2022 Cervical Spine Fracture Detection competition took place from July 28 to October 27, 2022. The competition dataset, consisting of 3112 cervical spine noncontrast CT scans from 12 institutions, was used for model training and internal testing. The dataset was divided into training (2019 scans), public testing (304 scans), and private testing (789 scans) sets, with fracture prevalence rates of 47.6% (961 of 2019), 40.1% (122 of 304), and 45.9% (362 of 789), respectively, notably higher than typical real-world rates of 4%–7% (23,24). Detailed information about the dataset can be found in the work by Lin et al (21).

Models

We selected seven of the eight award-winning models based on their scores on the RSNA competition's private dataset (25) to rigorously evaluate their ability to generalize. 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 (Fig 1). A detailed description of the models is provided in Appendix S2.

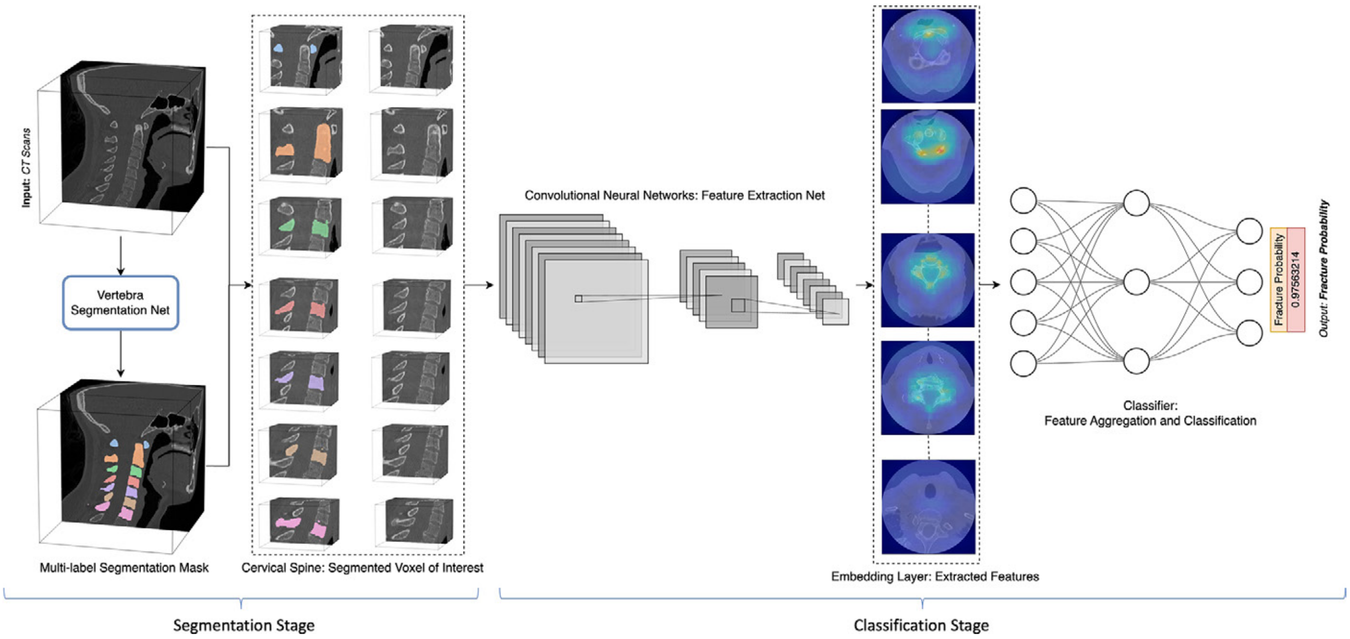
Input: CT Scans
Vertebra Segmentation Net
Multi-label Segmentation Mask
Cervical Spine: Segmented Voxel of Interest
Convolutional Neural Networks: Feature Extraction Net
Embedding Layer: Extracted Features
Classifier: Feature Aggregation and Classification
Fracture Probability
0.97563214
Output: Fracture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

Figure 1: Graphic displays an example of end-to-end architecture of a cervical spine CT fracture detection machine learning model, showcasing the segmentation stage to isolate the cervical spine's voxels of interest, followed by the classification stage for feature extraction, aggregation, and logits prediction.

Evaluation of Model Generalizability

Figure 2 displays the general workflow to evaluate model performance in the clinical setting. To better understand model generalizability, we analyzed model performance on a clinical test dataset composed of consecutive cervical spine CT scans obtained over the course of 1 year (January 1–December 31, 2022) for a traumatic indication at a busy urban neurosurgical and level I trauma center (St Michael's Hospital, Unity Health Toronto). Inclusion criteria were scans obtained in the emergency department for individuals older than age 18, with no exclusion criteria. The dataset included both noncontrast scans and contrast-enhanced scans. CT scans were downloaded via Philips Vue picture archiving and communication system (Philips Healthcare) and filtered for axial bone window images of the cervical spine measuring 1 mm or less in section thickness. Additional details about the dataset, including CT scan acquisition parameters, are provided in Appendix S1. The noncontrast scans were similar to the data used to train and evaluate models during the competition, while the contrast-enhanced scans allowed for the evaluation of model generalizability outside the distribution of the training data. Of note, this test dataset was not considered a purely "external" dataset, as our institution contributed data to the competition; however, none of those patients were represented in this dataset.

Reference Standard Labeling

To obtain reference standard labels, our radiology information system (Syngo; Siemens Medical Solutions) was searched for reports on emergency department cervical spine CT scans performed between January 1 and December 31, 2022, using mPower (Nuance Communications) in patients at least 18 years of age. Reports were classified as positive or negative for fracture at the patient and cervical spine segmental levels by a radiologist (M.N., 21 years of experience). The reference standard was established for equivocal reports by reviewing follow-up imaging examinations and clinical records. A random sample of 10% of the radiology reports were reviewed by a second radiologist (E.C., 15 years of experience), with 100% concordance at the segmental level. The presence or absence of intravenous contrast material was also established for each scan. Additional dataset curation details are provided in Appendix S1.

Review of False-Negative and False-Positive Cases

A neuroradiologist (S.M., 6.5 years of neuroradiology experience) reviewed every examination-level false-negative case to help determine the types of fractures missed by the ML models. CT scans that were misclassified as false positive at the examination 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 false-positive cases, whereas many of these were correctly classified by the other models. Heat maps from gradient-weighted class activation mapping (Grad-CAM) (26) were generated based on the averaged model outputs for false-positive classifications. Grad-CAM is a visualization technique that illuminates areas of an image influencing convolutional neural network prediction by highlighting these regions with heat maps. To interpret these maps, warmer colors (eg, red) indicate areas the model focused on more intensely, with brighter colors signifying higher influence on the model's decision. These heat maps allowed the neuroradiologist to concentrate on identifying commonly occurring features that may have misguided the model's judgment.

Statistical Analysis

The Youden *J* statistic (27) was used on the competition's public test dataset to determine optimal thresholds to binarize predicted probabilities that maximize the difference

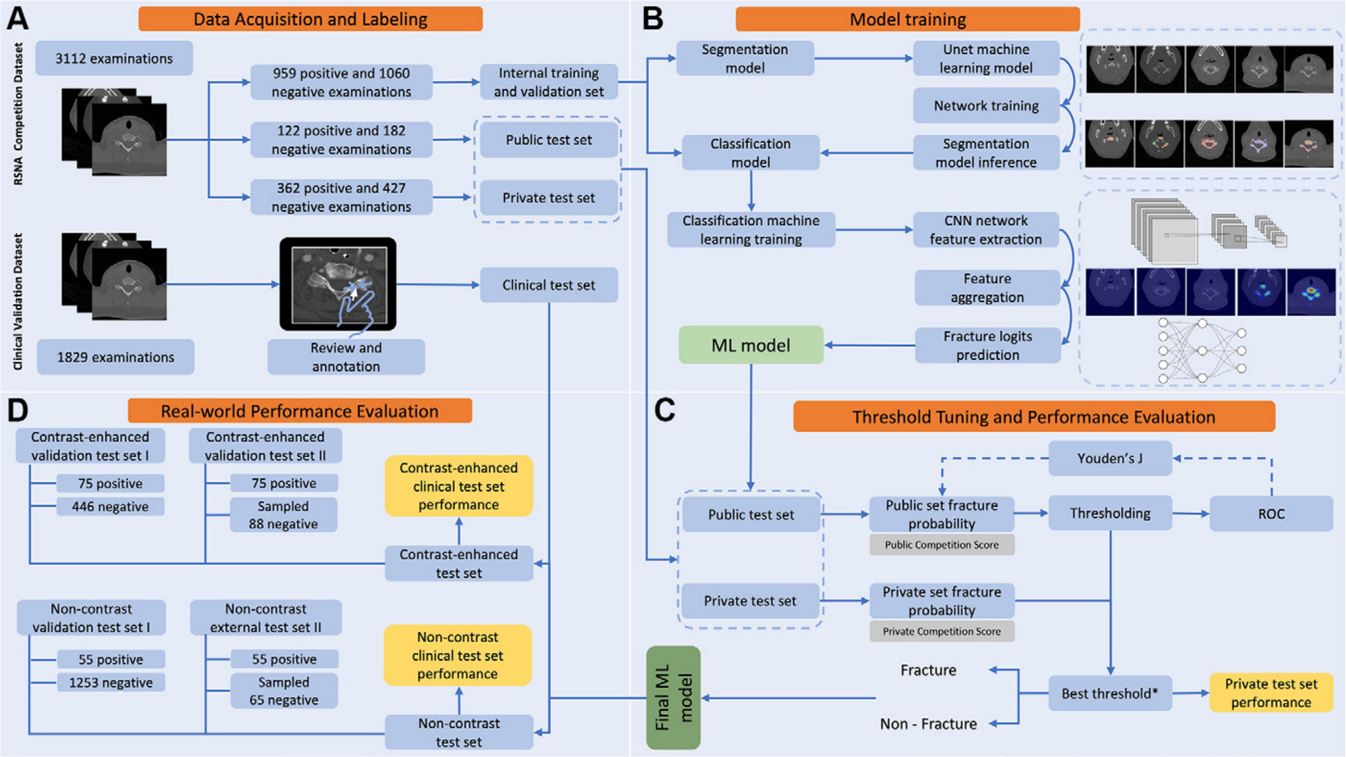
A
Data Acquisition and Labeling
RSNA Competition Dataset
3112 examinations
959 positive and 1060 negative examinations
Internal training and validation set
122 positive and 182 negative examinations
Public test set
362 positive and 427 negative examinations
Private test set
Clinical Validation Dataset
Review and annotation
Clinical test set
1829 examinations
B
Model training
Segmentation model
Unet machine learning model
Network training
Classification model
Segmentation model inference
Classification machine learning training
CNN network feature extraction
Feature aggregation
ML model
Fracture logits prediction
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 clinical 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 clinical test set performance
Non-contrast test set
Final 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
Best threshold*
Private test set performance
Non - Fracture

Figure 2: Flowchart of machine learning (ML) evaluation pipeline for cervical spine fracture detection. **(A)** The process starts with 3112 CT scans from the RSNA 2022 competition, divided into training, public test, and private test datasets. Additional CT scans from our institution are used as the clinical test dataset. **(B)** Each ML model has two main stages: segmentation, typically using two-dimensional or three-dimensional U-Net, and classification involving convolutional neural network (CNN) feature extraction, feature aggregation, and logits prediction. **(C)** Each model generates a fracture probability output binarized by applying an optimal threshold identified by the Youden *J* statistic on the public test dataset. Then the model's performance is assessed using the private test dataset. **(D)** The final evaluation comprises four subsets of the clinical test dataset: noncontrast scans, contrast-enhanced scans, bootstrap-sampled noncontrast scans, and bootstrap-sampled contrast-enhanced scans. ROC = receiver operating characteristic.

between the true-positive rate and the false-positive rate, effectively capturing the top-left-most point on the receiver operating characteristic curve. The thresholds for each of the seven models ($\text{Threshold}_{\text{Qishen}} = 0.72$, $\text{Threshold}_{\text{Darragh}} = 0.54$, $\text{Threshold}_{\text{Selim}} = 0.57$, $\text{Threshold}_{\text{Speedrun}} = 0.81$, $\text{Threshold}_{\text{Skecherz}} = 0.49$, $\text{Threshold}_{\text{QWER}} = 0.54$, $\text{Threshold}_{\text{Harshit}} = 0.72$) were then applied to the competition's private test set to establish baseline model performance on the competition dataset. Reference standard labels for each scan were compared with the ML model predictions. Sensitivity, specificity, positive predictive value, negative predictive value, accuracy, area under the receiver operating characteristic curve (AUC), and F1 score were the primary evaluation metrics.

Mean values and ranges (minimum, maximum) were calculated for each metric across all seven models. Additionally, the CIs were estimated separately for each individual model's performance. Specifically, the binomial method was used for accuracy, sensitivity, specificity, positive predictive value, and negative predictive value (28), while the Takahashi method was used for the F1 score (29) and the DeLong method for the AUC (30).

Model performances on the competition private test set and the clinical test set were compared to identify any differences or trends. Additional analyses, including those adjusting the clinical test dataset to match the competition dataset's prevalence, are detailed in Appendix S1.

Statistical analyses were performed using the Python libraries Scikit-learn (version 1.3.2), SciPy (version 1.11.4), and Confidenceinterval (version 1.0.4). Statistical significance of differences in model performances was not formally assessed in this study.

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 test dataset are not publicly available.

Our analysis and model implementation were conducted using Python (version 3.10.13) and torch (version 2.1.0). Additionally, we used a suite of Python packages to facilitate data analysis and results visualization, including SimpleITK (version 2.3.1), nibabel (version 5.2.0), torchvision (version 0.16.1), NumPy (version 1.26.2), scikit-image (version 0.22.0), opencv-python (version 4.8.1), pandas (version 2.1.4), matplotlib (version 3.8.0), and Grad-CAM (version 1.4.8). The detailed enumeration of the software and packages used aims to enhance the reproducibility and transparency of our study.

Results

Characteristics of the Clinical Test Set

The clinical test set used in this study was composed of 1829 series from 1828 cervical spine CT studies across 1779 adult patients (625 [35.1%] female patients, 1154 [64.9%] male pa-

Table 1: Patient Characteristics and Data Distribution of the Clinical Test Dataset

Attribute	Noncontrast CT Scan	Contrast-enhanced CT Scan	Overall
Total no. of patients	1268	520	1779
Mean age (y)	58.2 ± 22.3	50.0 ± 20.6	55.8 ± 22.1
Sex			
Female	481	145	625
Male	787	375	1154
Total no. of 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 cervical spine surgery	14	3	17
Positive	1	1	2
Negative	13	2	15
Postoperative material	14	2	16
Positive	1	1	2
Negative	13	1	14

Note.—Data are reported as numbers of patients or scans or mean ± SD. *Positive* and *negative* refer to positive or negative for a cervical spine fracture. A single CT scan may present with multiple fractures at different cervical spine level.

tients; age range, 18–101 years; mean age, 55.8 years ± 22.1 [SD]); a minority of patients had either pre- and postcontrast scans or repeat attendances to the emergency department. There were 130 scans positive for fracture. The dataset included 1308 noncontrast and 521 contrast-enhanced scans (Table 1).

Performance of Models on Each Dataset

Figure 3 shows the distribution of performance metrics for binary classification by the winning algorithms on different datasets. Detailed information regarding other analyses are provided in Appendix S1 and Tables S2–S5.

Competition Private Test Dataset

On the competition private test dataset, the seven top-scoring models had a mean AUC of 0.96 (range: 0.95–0.97), with a mean accuracy of 91.0% (range: 88.6%–92.6%). The mean sensitivity was 87.2% (range: 84.0%–89.8%), the mean specificity was 94.3% (range: 88.1%–96.7%), the mean positive predictive value was 93.0% (range: 86.4%–95.8%), and the mean negative predictive value was 89.7% (range: 87.5%–91.4%). Detailed individual model performances for the competition dataset are shown in Table 2.

Noncontrast and Contrast-enhanced CT Clinical Test Datasets

On the clinical test dataset, the models showed reduced AUC and accuracy on both subsets of the dataset. Specifically, accuracy was reduced in the contrast-enhanced dataset but remained high for the noncontrast subset. Due to the lower prevalence, model performances on both the noncontrast and contrast-enhanced datasets were characterized by a high negative predictive value (mean, 98.5% and 96.7%, respectively) and a notable decrease in positive predictive value (mean, 35.3% and 41.7%, respectively) compared with the competition dataset. In the noncontrast dataset with a real-world prevalence of 4.2%, the mean AUC across models was 0.89 (range: 0.79–0.92), mean accuracy was 91.8% (range: 81.9%–96.1%), mean sensitivity was 67.0% (range: 30.9%–80.0%), and mean specificity was 92.9% (range: 82.1%–99.0%). In the contrast-enhanced dataset with a real-world prevalence of 14.5%, the mean AUC across models was 0.88 (range: 0.76–0.94); mean accuracy was 73.5% (range: 28.4%–89.4%); mean sensitivity was 81.9% (range: 42.7%–100.0%); and mean specificity was 72.1% (range: 16.4%–92.8%). Individual model performance on the clinical test set is presented in Table 3 and Table 4 for noncontrast and contrast datasets, respectively.

Analysis of False-Positive and False-Negative Scans

There were 116 false-positive (47 noncontrast, 69 contrast-enhanced) and 78 false-negative (35 noncontrast, 43 contrast-enhanced) scans. On review, the ML models correctly identified 10 cases of true fractures that were initially missed by reporting radiologists (Fig 4). The most common influential regions identified on Grad-CAM heat maps for false-positive cases were vessels, present in 43 of 116 (37.1%) false-positive cases and in 39 of 69 (56.5%) false-positive contrast-enhanced studies. Other

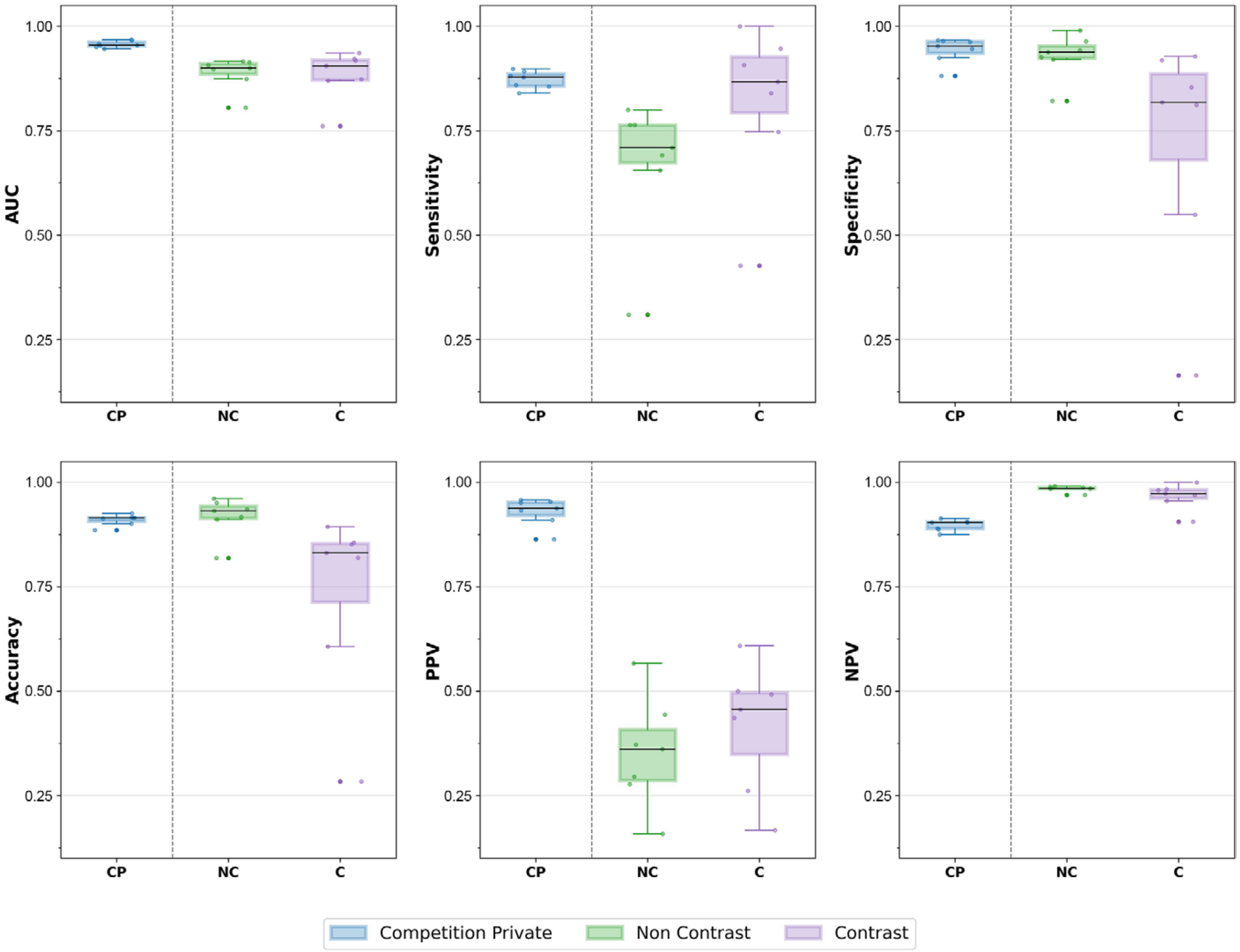

Figure 3: Box and whisker plots showcase the distribution of performance metrics for binary classification by the winning algorithms on different datasets. The metrics are the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Performance is displayed across the competition private dataset (CP), actual prevalence noncontrast (NC), and actual prevalence contrast (C) datasets. The box represents the IQR, the median is indicated by the black line within the box, and the whiskers show the full range excluding outliers, which are depicted as individual points. Data points are also shown as jitters for clarity.

less-common influential regions contributing to false-positive cases were related to chronic changes such as osteophytes, degenerative cortical irregularities, ligament and soft tissue calcification, vascular channels, and artifacts (Fig 5).

On review of the false-negative cases for the seven ML models, there were 135 fractures across 78 scans (42 contrast-enhanced and 36 noncontrast). 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 of 88), followed by osteopenia (23 of 88), artifact (eight of 88), healed chronic fracture (two of 88), and osseous lesions associated with pathologic fracture (two of 88). The most common sites of missed fractures were at the edge of the vertebral body end plate (36 of 135), transverse process (35 of 135), and spinous process (17 of 135). The most common levels of missed fractures were at C7 (19.3%; 26 of 135) followed by C6 (17.8%; 24 of 135).

Discussion

Award-winning models from the 2022 RSNA competition demonstrated strong performance with a mean AUC of 0.96 and accuracy of 91.0% on the competition test dataset and a mean AUC of 0.89 and accuracy of 91.8% for noncontrast scans and a mean AUC of 0.88 and accuracy of 73.5% for contrast-enhanced scans on the clinical test dataset. The major strength of this study is that every cervical spine CT scan performed in adults for a traumatic indication in the emergency department over a 1-year period was included in this study without exclusion criteria. Importantly, both contrast-enhanced and noncontrast CT scans were included in this dataset, as both are routinely encountered in clinical practice, despite models being trained solely only on noncontrast 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 a real-world clinical environment and was analyzed in balanced and unbalanced

Table 2: Individual Machine Learning Model Performances in Detecting Cervical Spine Fractures on the Competition Private Test Dataset (Fracture Prevalence = 45.8%)

Model	TP	FN	TN	FP	Sen (%)	Spec (%)	PPV (%)	NPV (%)	Acc (%)	F1 (%)	AUC
1. Qishen	318	44	413*	14*	87.8 (84.1, 90.8)	96.7 (94.6, 98.0)*	95.8 (93.0, 97.5)*	90.4 (87.3, 92.7)	92.6 (90.6, 94.3)*	91.6 (89.0, 94.3)*	0.97 (0.96, 0.98)*
2. RAWE	…	…	…	…	…	…	…	…	…	…	…
3. Darragh	310	52	412	15	85.6 (81.6, 88.9)	96.5 (94.3, 97.9)	95.4 (92.5, 97.2)	88.8 (85.6, 91.4)	91.5 (89.4, 93.3)	90.2 (87.4, 93.1)	0.97 (0.96, 0.98)
4. Selim	319	43	404	23	88.1 (84.4, 91.1)	94.6 (92.0, 96.4)	93.3 (90.1, 95.5)	90.4 (87.3, 92.8)	91.6 (89.5, 93.4)	90.6 (87.8, 93.4)	0.95 (0.94, 0.97)
5. Speedrun	304	58	407	20	84.0 (79.8, 87.4)	95.3 (92.9, 96.9)	93.8 (90.7, 96.0)	87.5 (84.2, 90.2)	90.1 (87.8, 92.0)	88.6 (85.6, 91.6)	0.95 (0.94, 0.97)
6. Skecherz	323	39	376	51	89.2 (85.6, 92.0)	88.1 (84.6, 90.8)	86.4 (82.5, 89.5)	90.6 (87.4, 93.0)	88.6 (86.2, 90.6)	87.8 (84.7, 90.8)	0.95 (0.94, 0.97)
7. QWER	325*	37*	395	32	89.8 (86.2, 92.5)*	92.5 (89.6, 94.6)	91.0 (87.6, 93.6)	91.4 (88.4, 93.7)*	91.3 (89.1, 93.0)	90.4 (87.6, 93.2)	0.96 (0.94, 0.97)
8. Harshit	311	51	411	16	85.9 (81.9, 89.1)	96.3 (94.0, 97.7)	95.1 (92.2, 97.0)	89.0 (85.8, 91.5)	91.5 (89.4, 93.3)	90.3 (87.4, 93.1)	0.95 (0.93, 0.96)

Note.—Values in parentheses are 95% CIs. A detailed description of the models is provided in Appendix S2. Acc = accuracy, AUC = area under the receiver operating characteristic curve, FN = false negative, FP = false positive, NPV = negative predictive value, PPV = positive predictive value, sen = sensitivity, spec = specificity, TN = true negative, TP = true positive.
* Best estimated value in the category.

Table 3: Individual Machine Learning Model Performances in Detecting Cervical Spine Fractures on the Noncontrast Subset of our Clinical Test Dataset with a Real-World Prevalence Rate of Fractures (Fracture Prevalence = 4.2%)

Model	TP	FN	TN	FP	Sen (%)	Spec (%)	PPV (%)	NPV (%)	Acc (%)	F1 (%)	AUC
1. Qishen	36	19	1208	45	65.5 (52.3, 76.6)	96.4 (95.2, 97.3)	44.4 (34.1, 55.3)	98.5 (97.6, 99.0)	95.1 (93.8, 96.1)	52.9 (42.8, 63.1)*	0.91 (0.87, 0.96)*
2. RAWE	…	…	…	…	…	…	…	…	…	…	…
3. Darragh	42	13	1182	71	76.4 (63.7, 85.6)	94.3 (92.9, 95.5)	37.2 (28.8, 46.4)	98.9 (98.1, 99.4)	93.6 (92.1, 94.8)	50.0 (40.7, 59.3)	0.90 (0.85, 0.95)
4. Selim	39	16	1160	93	70.9 (57.9, 81.2)	92.6 (91.0, 93.9)	29.5 (22.4, 37.8)	98.6 (97.8, 99.2)	91.7 (90.0, 93.0)	41.7 (32.8, 50.6)	0.91 (0.86, 0.95)
5. Speedrun	17	38	1240*	13*	30.9 (20.3, 44.0)	99.0 (98.2, 99.4)*	56.7 (39.2, 72.6)*	97.0 (95.9, 97.8)	96.1 (94.9, 97.0)*	40.0 (26.8, 53.2)	0.79 (0.72, 0.86)
6. Skecherz	42	13	1029	224	76.4 (63.7, 85.6)	82.1 (79.9, 84.1)	15.8 (11.9, 20.7)	98.8 (97.9, 99.3)	81.9 (79.7, 83.9)	26.2 (19.6, 32.7)	0.87 (0.81, 0.93)
7. QWER	44*	11*	1175	78	80.0 (67.6, 88.4)*	93.8 (92.3, 95.0)	36.1 (28.1, 44.9)	99.1 (98.3, 99.5)*	93.2 (91.7, 94.4)	49.7 (40.7, 58.7)	0.92 (0.87, 0.96)
8. Harshit	38	17	1154	99	69.1 (56.0, 79.7)	92.1 (90.5, 93.5)	27.7 (20.9, 35.8)	98.5 (97.7, 99.1)	91.1 (89.5, 92.6)	39.6 (30.8, 48.4)	0.90 (0.85, 0.95)

Note.—Values in parentheses are 95% CIs. A detailed description of the models is provided in Appendix S2. Acc = accuracy, AUC = area under the receiver operating characteristic curve, FN = false negative, FP = false positive, NPV = negative predictive value, PPV = positive predictive value, sen = sensitivity, spec = specificity, TN = true negative, TP = true positive.
* Best estimated value in the category.

groups, reflecting the matched higher prevalence of fractures in the competition test dataset and the lower prevalence of fractures encountered in clinical practice.

Previous studies exploring ML models for cervical spine fracture detection have shown varied performance. Zhang et al (31) reported an AUC up to 0.87, Salehinejad et al (32) achieved a classification accuracy of 71%–79%, and Golla et al (33) reached 87% sensitivity at the fracture level. BriefCase, a U.S. Food and Drug Administration–approved tool, demonstrated a sensitivity of 91.7% and specificity of 88.6% in its regulatory submission (34). However, external testing by Small et al (35) and Voter et al (36) reported lower sensitivities of 76% and 54.9%, respectively. Our results suggest that the top-performing ML models developed by participating teams for the RSNA competition have been able to achieve better model performance than the previously reported results from individual research groups. Although the models evaluated still do not match radiologist metrics,

Table 4: Individual Machine Learning Model Performances in Detecting Cervical Spine Fractures on the Contrast-enhanced Subset of our Clinical Test Dataset with a Real-World Prevalence Rate of Fractures (Fracture Prevalence = 14.5%)

Model	TP	FN	TN	FP	Sen (%)	Spec (%)	PPV (%)	NPV (%)	Acc (%)	F1 (%)	AUC
1. Qishen	68	7	365	81	90.7 (82.0, 95.4)	81.8 (78.0, 85.1)	45.6 (37.8, 53.6)	98.1 (96.2, 99.1)	83.1 (79.7, 86.1)	60.7 (53.2, 68.2)	0.94 (0.90, 0.97)*
2. RAWE	…	…	…	…	…	…	…	…	…	…	…
3. Darragh	63	12	381	65	84.0 (74.1, 90.6)	85.4 (81.8, 88.4)	49.2 (40.7, 57.8)	96.9 (94.7, 98.2)	85.2 (81.9, 88.0)	62.1 (54.2, 69.9)	0.92 (0.87, 0.96)
4. Selim	56	19	410	36	74.7 (63.8, 83.1)	91.9 (89.0, 94.1)	60.9 (50.7, 70.2)*	95.6 (93.2, 97.1)	89.4 (86.5, 91.8)*	67.1 (58.7, 75.4)*	0.92 (0.89, 0.96)
5. Speedrun	32	43	414*	32*	42.7 (32.1, 53.9)	92.8 (90.0, 94.9)*	50.0 (38.1, 61.9)	90.6 (87.6, 92.9)	85.6 (82.3, 88.4)	46.0 (35.8, 56.3)	0.76 (0.70, 0.83)
6. Skecherz	75*	0*	73	373	100.0 (95.1, 100.0)*	16.4 (13.2, 20.1)	16.7 (13.6, 20.5)	100.0 (95.0, 100.0)*	28.4 (24.7, 32.4)	28.7 (22.9, 34.4)	0.87 (0.83, 0.92)
7. QWER	65	10	362	84	86.7 (77.2, 92.6)	81.2 (77.3, 84.5)	43.6 (35.9, 51.6)	97.3 (95.1, 98.5)	82.0 (78.4, 85.0)	58.0 (50.4, 65.7)	0.91 (0.86, 0.95)
8. Harshit	71	4	245	201	94.7 (87.1, 97.9)	54.9 (50.3, 59.5)	26.1 (21.2, 31.6)	98.4 (95.9, 99.4)	60.7 (56.4, 64.8)	40.9 (34.2, 47.6)	0.87 (0.83, 0.91)

Note.—Values in parentheses are 95% CIs. A detailed description of the models is provided in Appendix S2. Acc = accuracy, AUC = area under the receiver operating characteristic curve, FN = false negative, FP = false positive, NPV = negative predictive value, PPV = positive predictive value, sen = sensitivity, spec = specificity, TN = true negative, TP = true positive.
* Best estimated value in the category.

with the sensitivity and specificity of radiologists to detect cervical spine fractures at CT shown to be 88.0%–93.0% and 96.0%–99.0%, respectively (35,37), these models hold promise as rapid auxiliary tools. In fact, our study showed that a small number of fractures missed by radiologists were retrospectively identified by the ML models. The models generated a cervical spine fracture prediction in just 10–30 seconds, while typically, it takes between 33 to 43 minutes from scan acquisition until a finalized report by radiologists (35). Therefore, ML models could be used as rapid triaging tools to flag the study to alert the radiologist of a possible fracture, some of which may be missed by radiologists.

In examining the performance metrics, it was noted that the models faced challenges when applied to the clinical test dataset, particularly with contrast-enhanced scans. Higher performance in the noncontrast subset is expected given that the training dataset consists of these exclusively. Average model sensitivity reductions were 20.2% for noncontrast scans and 5.2% for contrast-enhanced scans, while average specificity reductions were 1.4% for noncontrast scans and 22.2% for contrast-enhanced scans. Interestingly, accuracy for noncontrast scans slightly improved, with an average increase of 0.7%, whereas contrast-enhanced scans experienced an accuracy decline of 17.5%, both attributed to differences in fracture prevalence. These results underscore the broader challenge of transitioning from curated datasets to real-world clinical applications, as highlighted by external testing studies by Voter et al and Small et al, where sensitivity decreased from 91.7% to as low as 54.9% (35,36). Our study also revealed areas of strength and improvement for the models, through a comprehensive review of the false-negative and false-positive cases. Intravascular contrast material, 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 false-negative cases, certain types of fractures were missed most by the models, including fractures at the edge of the vertebral body end plate, transverse process, and spinous process locations, consistent with previous research (35,36). The most common cause for models to miss fractures were degenerative changes and osteopenia, also observed by Small et al (35), leading to underperformance in older patients (36). An understanding of these patterns can guide future model refinement by inclusion of greater numbers of imaging studies with underrepresented pathologies.

This study had limitations. The use of clinical data from a single center may limit the generalizability of our findings. Additionally, the training data were exclusively noncontrast CT scans focusing on acute fractures, which may not represent the complexity of cases in the clinical test dataset that included patients with previous surgical interventions and contrast-enhanced studies; therefore, models could underperform on our dataset. Last, although Grad-CAM was used for model transparency, its limitations in localizing multiple instances and capturing fine details, as noted by Mohamed et al (38), might have influenced the interpretability of our results, although major issues were not observed in this study. Future model enhancements will involve training on a more diverse array of scans and integrating feedback from real-world applications to boost accuracy and reliability in clinical settings.

In conclusion, evaluation of the top-performing ML models in the 2022 RSNA competition on a clinical test dataset demonstrated that the models fell short of their performance on the competition dataset. However, the models still performed favorably as compared with previously published cervical spine detection algorithms, including U.S. Food and

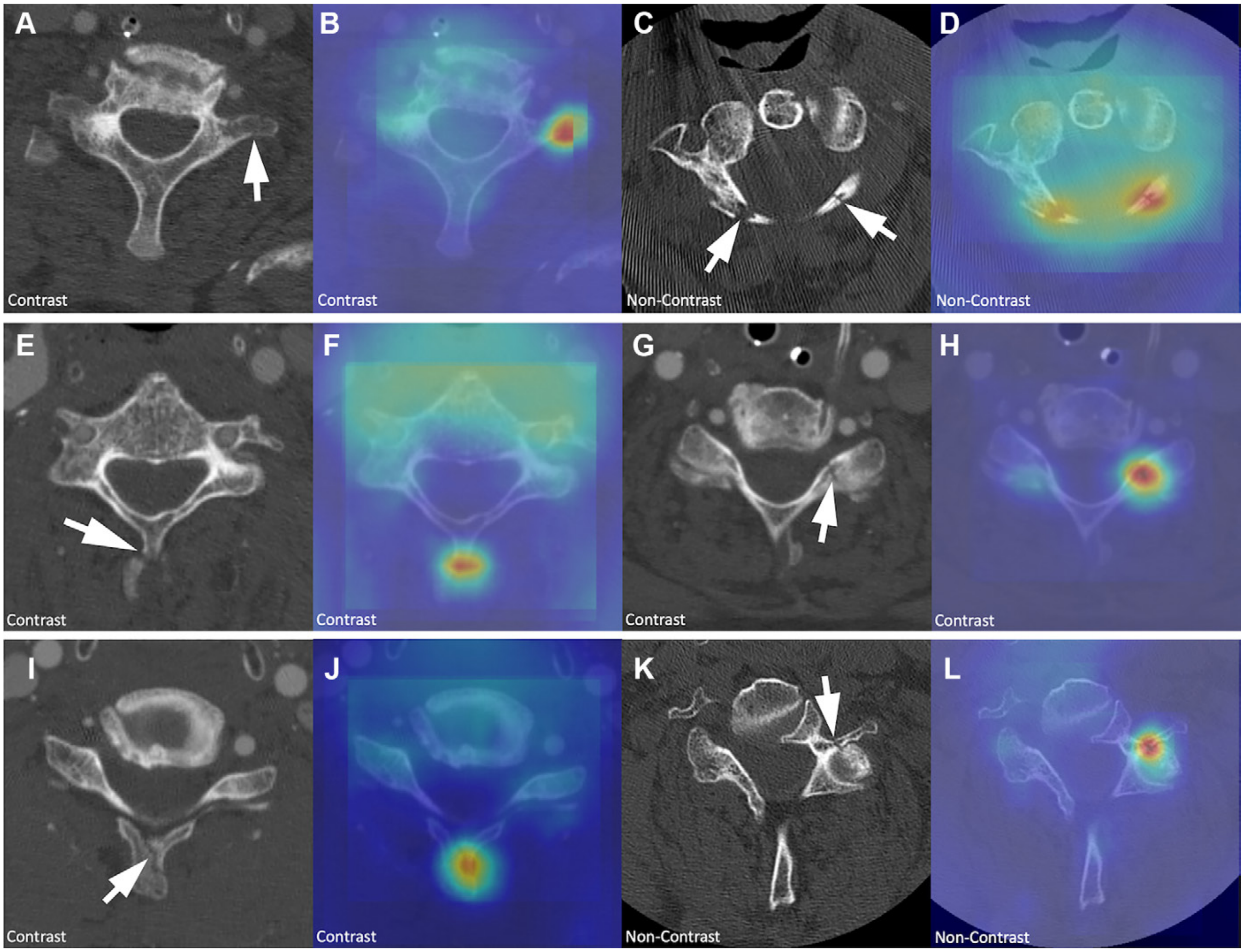

Figure 4: Example cases of fractures identified by the machine learning models but missed by reporting radiologists. Axial CT images of the cervical spine with associated gradient-weighted class activation heat maps show the most influential regions in the input image for the prediction. Warmer colors (eg, red) on the heat map indicate areas the model focuses on more intensely, with brighter colors signifying higher influence on the model's decision. The presence of contrast material at CT imaging is labeled in the bottom left corner of each image. **(A, B)** Minimally displaced left transverse process fracture (arrow in **A**). **(C, D)** Bilateral lamina fractures (arrows in **C**), moderately displaced on the right and undisplaced on the left. **(E, F)** Mildly displaced spinous process fracture (arrow in **E**). **(G, H)** Undisplaced left articular process and lamina fracture (arrow in **G**). **(I, J)** Minimally displaced spinous process fracture (arrow in **I**), and **(K, L)** minimally displaced left transverse process fracture (arrow in **K**).

Drug Administration–approved commercial models. The models showed potential to generalize to the analysis of contrast-enhanced scans and of patients with prior surgical intervention, despite being trained on a dataset that excluded these examinations. Addressing false-positive and false-negative cases through the inclusion of relevant imaging studies holds potential for future model refinement. These models may serve as valuable supplementary diagnostic tools for cervical spine fracture detection, emphasizing the necessity for ongoing improvement efforts and prospective evaluation of deployed models.

Author contributions: Guarantors of integrity of entire study, **Z.H., K.W.Y., Q.H., C.H., E.C.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **Z.H., M.P., L.M.P., S.M., R.M., K.W.Y., H.C., C.H., A.E.F., E.S., E.C.**; clinical studies, **M.P., M.N., S.M., K.M.Y., E.C.**; experimental studies, **Z.H., H.M.L., K.W.Y., Q.H., D.H., S.S., H.C., P.S., C.H., P.P., H.S., W.L., E.S.**; statistical analysis, **Z.H., M.P., R.L.B., C.H., E.S., E.C.**; and manuscript editing, **Z.H., M.P., R.L.B., H.M.L., L.M.P., S.M., R.M., J.W., C.W., K.W.Y., I.P., A.E.F., T.R., J.T., E.S., E.C.**

Disclosures of conflicts of interest: **Z.H.** No relevant relationships. **M.P.** No relevant relationships. **R.L.B.** Support, Radiological Society of North America; consulting fees, Radiological Society of North America, AI Committee, Radiological Society of North America; Statistical Consulting Section, American Statistical Association. **H.M.L.** No relevant relationships. **L.M.P.** Associate Editor of *Radiology: Artificial Intelligence*; member of the RSNA AI Committee. **M.N.** No relevant relationships. **S.M.** No relevant relationships. **R.M.** No relevant relationships. **J.W.** Grant or contract, CHIR; consulting fees, Stryker Canada. **C.W.** Institutional research funding, Cerapedics Inc. **K.W.Y.** No relevant relationships. **Q.H.** No relevant relationships. **D.H.** No relevant relationships. **S.S.** No relevant relationships. **H.C.** No relevant relationships. **P.S.** No relevant relationships. **C.H.** No relevant relationships. **P.P.** No relevant relationships. **I.P.** Consulting fee, Centaur Labs, MD.ai. **H.S.** No relevant relationships. **W.L.** No relevant relationships. **A.E.F.** Institutional grant, MIDRC NIBIB; RSNA Board of Directors. **F.C.K.** Consulting fees, MD.ai, GE HealthCare, Bunkerhill Health; honoraria, Sharing Progress in Cancer Care; Associate Editor at *Radiology: Artificial Intelligence*; Member of the RSNA AI Committee; member of the RSNA Radiology Informatics Council; early consultant to the Editor of *Radiology*; vice-chair of the Machine Learning Committee at SIIM. **T.R.** No relevant relationships. **J.T.** No relevant relationships. **E.S.** No relevant relationships. **E.C.** No relevant relationships.

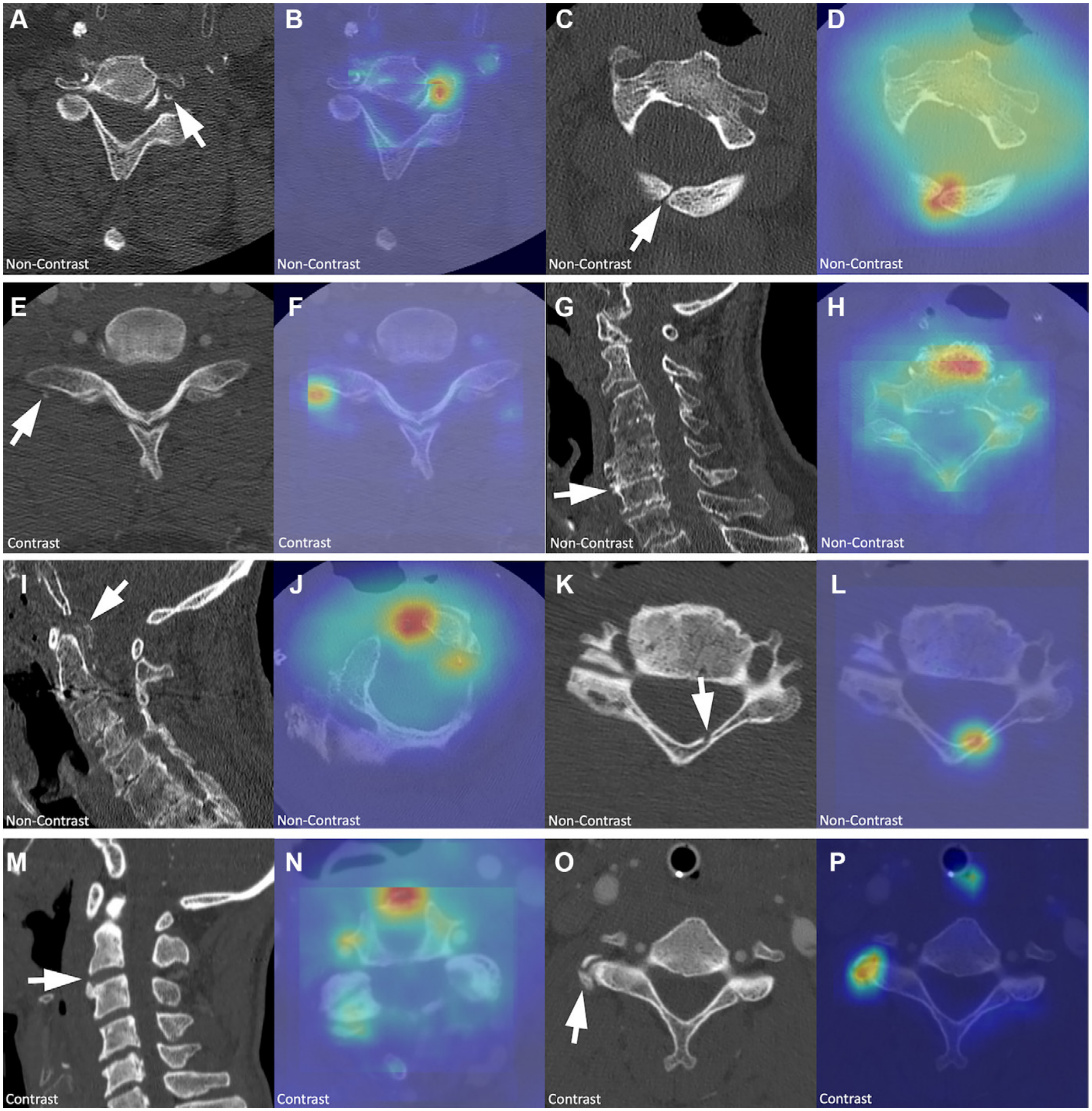

Figure 5: Example cases in the false-positive group incorrectly identified as fractures by the machine learning models. The CT images with associated gradient-weighted class activation heat maps show the most influential regions in the input image for the prediction. Warmer colors (eg, red) on the heat map indicate areas the model focuses on more intensely, with brighter colors signifying higher influence on the model's decision. The presence of contrast material at CT imaging is labeled in the bottom left corner of each image. CT images in **G**, **I**, and **M** are presented in the sagittal plane to better demonstrate pathology; all other images are in the axial plane. **(A, B)** Calcified atherosclerotic plaque in the left vertebral artery in the left transverse foramen (arrow in **A**). **(C, D)** Congenital lack of fusion of the posterior arch of C1 (arrow in **C**). **(E, F)** Contrast material within a small vessel in the right paraspinal region (arrow in **E**). **(G, H)** Chronic multilevel degenerative changes with reduced intervertebral disk spaces, osteophyte formation (arrow in **G**), and osteopenia. **(I, J)** Partially calcified pseudomass (arrow in **I**) posterior to the odontoid process of C2, secondary to calcium pyrophosphate dihydrate crystal deposition disease. **(K, L)** Nutrient vessel within the left lamina (arrow in **K**). **(M, N)** Chronic osteophyte arising from the superior-anterior vertebral body of C3 (arrow in **M**). **(O, P)** Chronic osteophytic changes associated with the right articular process (arrow in **O**).

References

1. Bank M, Gibbs K, Sison C, et al. Age and Other Risk Factors Influencing Long-Term Mortality in Patients With Traumatic Cervical Spine Fracture. Geriatr Orthop Surg Rehabil 2018;9:2151459318770882.
2. 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;22(1):78.
3. Milby AH, Halpern CH, Guo W, Stein SC. Prevalence of cervical spinal injury in trauma. Neurosurg Focus 2008;25(5):E10.
4. Inaba K, Byerly S, Bush LD, et al. Cervical spinal clearance: A prospective Western Trauma Association Multi-institutional Trial. J Trauma Acute Care Surg 2016;81(6):1122–1130.
5. 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;28(3):483–493.
6. 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;14(11):1498–1503.

7. 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;17(4):585–591.
8. 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;46(Suppl 1):S24–S27.
9. 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;36(Suppl 2):B13–B26.
10. Gull S, Akbar S. Artificial intelligence in brain tumor detection through MRI scans: advancements and challenges. In: Artificial Intelligence and Internet of Things. CRC, 2021; 241–276.
11. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73(5):439–445.
12. 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 USA 2019;116(45):22737–22745.
13. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016; 770–778.
14. Burns JE, Yao J, Summers RM. Vertebral Body Compression Fractures and Bone Density: Automated Detection and Classification on CT Images. Radiology 2017;284(3):788–797.
15. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 2018;98:8–15.
16. Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. Radiology 2020;295(1):4–15.
17. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol 2024;42(1):3–15.
18. Radiological Society of North America. AI challenges. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge. Accessed March 10, 2024
19. Salehinejad H, Kitamura J, Ditkofsky N, et al. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. Sci Rep 2021;11(1):17051.
20. 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;306(2):e220505.
21. Lin HM, Colak E, Richards T, et al. The RSNA Cervical Spine Fracture CT Dataset. Radiol Artif Intell 2023;5(5):e230034.
22. Kaggle. RSNA 2022 Cervical Spine Fracture Detection Leaderboard. https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/leaderboard. Accessed March 10, 2024.
23. Tang A, Pawar J, Bridge C, et al. Traumatic cervical spine fracture patterns on CT: a retrospective analysis at a level 1 trauma center. Emerg Radiol 2021;28(5):965–976.
24. 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;27(1):1–8.
25. Lee GR, Flanders AE, Richards T, et al. Performance of the Winning Algorithms of the RSNA 2022 Cervical Spine Fracture Detection Challenge. Radiol Artif Intell 2024;6(1):e230256.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017; 618–626.
27. 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;16(1):73–81.
28. Blyth CR, Still HA. Binomial confidence intervals. J Am Stat Assoc 1983;78(381):108–116.
29. Takahashi K, Yamamoto K, Kuchiba A, Koyama T. Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. Appl Intell 2022;52(5):4961–4972.
30. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett 2014;21(11):1389–1393.
31. Zhang M, Kim L, Cheong R, et al. Deep-learning artificial intelligence model for automated detection of cervical spine fracture on computed tomography (CT) imaging. In: 2019 AANS Annual Scientific Meeting. J Neurosurg 2019;131(1):2–116.
32. Salehinejad H, Ho E, Lin HM, et al. Deep Sequential Learning For Cervical Spine Fracture Detection On Computed Tomography Imaging. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021;1911–1914.
33. Golla AK, Lorenz C, Buerger C, et al. Cervical spine fracture detection in computed tomography using convolutional neural networks. Phys Med Biol 2023;68(11):115010.
34. U.S. Food and Drug Administration. K190896. https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190896.pdf. Accessed March 10, 2024.
35. Small JE, Osler P, Paul AB, Kunst M. CT Cervical Spine Fracture Detection Using a Convolutional Neural Network. AJNR Am J Neuroradiol 2021;42(7):1341–1347.
36. 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;42(8):1550–1556.
37. van der Kolk BBYM, van den Wittenboer GGJ, Warringa N, et al. 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;3(1):e12609.
38. Mohamed E, Sirlantzis K, Howells G. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. Displays 2022;73:102239.