Original Article | Pediatric Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2021.0449
Korean J Radiol 2022;23(3):343-354

Korean Journal of Radiology
KJR

Deep Learning-Assisted Diagnosis of Pediatric Skull Fractures on Plain Radiographs

Jae Won Choi[1, 2], Yeon Jin Cho[1, 3], Ji Young Ha[4], Yun Young Lee[5], Seok Young Koh[3], June Young Seo[3], Young Hun Choi[1, 3], Jung-Eun Cheon[1, 3, 6], Ji Hoon Phi[7], Injoon Kim[8], Jaekwang Yang[9], Woo Sun Kim[1, 3, 6]

[1]Department of Radiology, Seoul National University College of Medicine, Seoul, Korea; Departments of [2]Radiology and [8]Emergency Medicine, Armed Forces Yangju Hospital, Yangju, Korea; [3]Department of Radiology, Seoul National University Hospital, Seoul, Korea; [4]Department of Radiology, Gyeongsang National University Changwon Hospital, Changwon, Korea; [5]Department of Radiology, Chonnam National University Hospital, Gwangju, Korea; [6]Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Korea; [7]Division of Pediatric Neurosurgery, Seoul National University Children's Hospital, Seoul, Korea; [9]Army Aviation Operations Command, Icheon, Korea

Objective: To develop and evaluate a deep learning-based artificial intelligence (AI) model for detecting skull fractures on plain radiographs in children.

Materials and Methods: This retrospective multi-center study consisted of a development dataset acquired from two hospitals (n = 149 and 264) and an external test set (n = 95) from a third hospital. Datasets included children with head trauma who underwent both skull radiography and cranial computed tomography (CT). The development dataset was split into training, tuning, and internal test sets in a ratio of 7:1:2. The reference standard for skull fracture was cranial CT. Two radiology residents, a pediatric radiologist, and two emergency physicians participated in a two-session observer study on an external test set with and without AI assistance. We obtained the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity along with their 95% confidence intervals (CIs).

Results: The AI model showed an AUROC of 0.922 (95% CI, 0.842–0.969) in the internal test set and 0.870 (95% CI, 0.785–0.930) in the external test set. The model had a sensitivity of 81.1% (95% CI, 64.8%–92.0%) and specificity of 91.3% (95% CI, 79.2%–97.6%) for the internal test set and 78.9% (95% CI, 54.4%–93.9%) and 88.2% (95% CI, 78.7%–94.4%), respectively, for the external test set. With the model's assistance, significant AUROC improvement was observed in radiology residents (pooled results) and emergency physicians (pooled results) with the difference from reading without AI assistance of 0.094 (95% CI, 0.020–0.168; p = 0.012) and 0.069 (95% CI, 0.002–0.136; p = 0.043), respectively, but not in the pediatric radiologist with the difference of 0.008 (95% CI, -0.074–0.090; p = 0.850).

Conclusion: A deep learning-based AI model improved the performance of inexperienced radiologists and emergency physicians in diagnosing pediatric skull fractures on plain radiographs.

Keywords: *Deep learning; Artificial intelligence; Skull fracture; Pediatric; Plain radiograph*

INTRODUCTION

Pediatric head trauma is a significant cause of morbidity and mortality worldwide. The increasing number of emergency department visits for head trauma is a public health concern [1]. Compared with adults, pediatric clinical assessment is often more problematic, and asymptomatic intracranial injury is more common in pediatric head trauma patients [2,3]. As most patients have minor head trauma, it is important to identify the exact patients at risk of long-term neurological devastation or requiring immediate intervention [4,5].

In diagnosing traumatic head injuries in children, computed tomography (CT) is considered the most accurate, and evidence-based guidelines, such as the American College of Radiology Appropriateness Criteria, consider that skull radiography is inadequate [4]. However, many

Received: June 2, 2021 **Revised:** October 27, 2021
Accepted: November 7, 2021
Corresponding author: Yeon Jin Cho, MD, PhD, Department of Radiology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
• E-mail: blue1010c@gmail.com

pediatric head trauma patients undergo skull radiography [6], and its use varies among healthcare providers, institutions, and nations [6-8]. Skull radiography may be used as a screening tool in cases where CT is not clinically indicated, and in children with skull fractures scheduled to undergo CT examination. It also plays a crucial role as part of a skeletal survey for suspected physical abuse [9,10]. Non-accidental head injuries are the most common cause of death due to child abuse. Although it is relatively common, it is a serious cause of morbidity and mortality in children [11,12]. Furthermore, the interpretation of pediatric skull radiographs is challenging. Variable appearances of primary and accessory sutures may complicate the detection of skull fractures [13]. Vascular channels may also mimic skull fracture [14]. Skull radiography can be more challenging if a radiologist with pediatric expertise is unavailable because physicians may have limited ability in identifying skull fractures [15]. Recently, deep learning using convolutional neural networks, a rapidly advancing subfield of artificial intelligence (AI), has shown promising performance in medical image analysis [16]. Many studies have demonstrated the application of deep learning in musculoskeletal radiology [17]. Here, we aimed to develop and evaluate a deep learning model that detects pediatric skull fractures on plain radiographs.

MATERIALS AND METHODS

This retrospective study was approved by the Institutional Review Boards of three participating hospitals: Hospital #1 (Seoul National University Hospital, IRB No. 1910-144-1072), Hospital #2 (Chonnam National University Hospital, IRB No. 2021-069), and Hospital #3 (Gyeongsang National University Changwon Hospital, IRB No. 2021-05-025), with a waiver for informed consent.

Data Curation

An overview of the datasets is shown in Figure 1. The development dataset comprised 413 consecutive patients from Hospitals #1 and #2. The inclusion criteria were 1) patients with head trauma who presented to the pediatric emergency department (age < 20 years), 2) underwent both anteroposterior (AP) and lateral skull radiography, and 3) concurrent cranial CT. Patients with previous head surgeries were excluded from the study. We included 87 fracture-positive and 62 fracture-negative patients from eligible patients who visited Hospital #1 between January 2013 and December 2019. Similarly, in Hospital #2, we included 99 fracture-positive and 165 fracture-negative patients who presented between January 2016 and August 2019. The development dataset was randomly split into training, tuning, and internal test sets in an approximate ratio of 7:1:2 at the patient level in a stratified manner

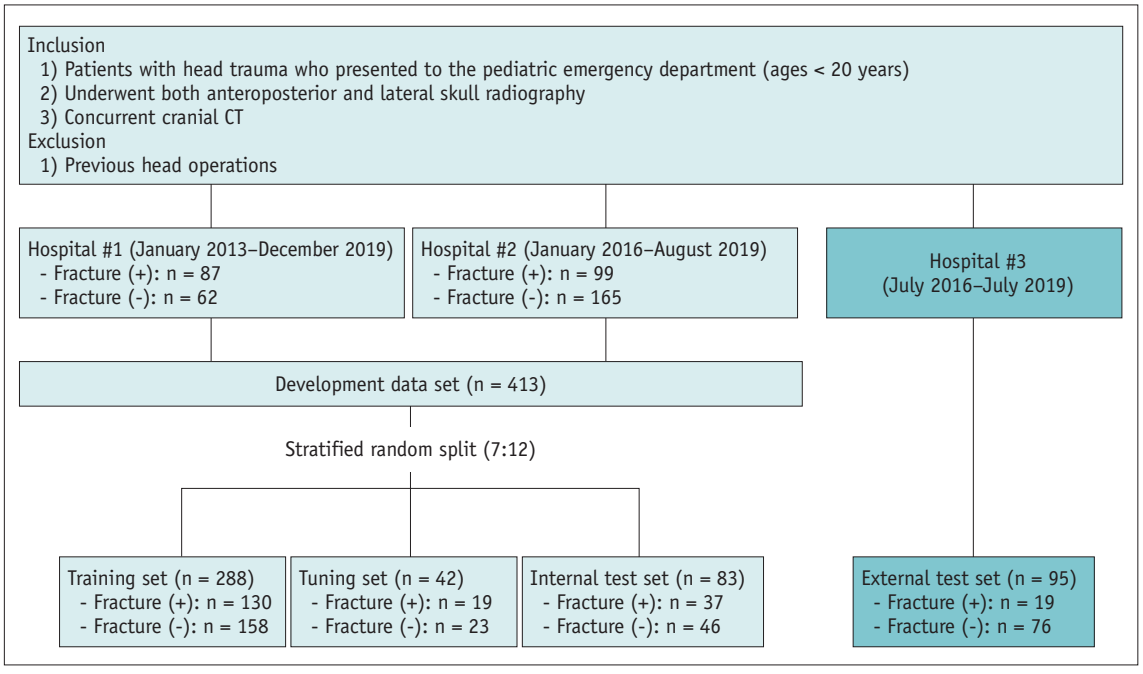

Fig. 1. Overview of datasets.

https://doi.org/10.3348/kjr.2021.0449

Korean Journal of Radiology
KJR

based on the labels. For the external test set, we collected consecutive patients using the following criteria: 1) patients who met the same inclusion and exclusion criteria as the development dataset, and 2) visited Hospital #3 between July 2016 and July 2019. As a result, the external test set consisted of 19 fracture-positive and 76 fracture-negative patients.

For both the development and external test set patients, we acquired all available skull radiographs, including AP and lateral views, as well as the Towne view. There were 413, 558, and 18 images of AP, lateral, and Towne view radiographs, respectively, in the development set. In the external test set, the numbers were 95, 178, and 95, respectively.

Reference Standard and Annotation

The reference standard for skull fracture diagnosis was cranial CT. In the development dataset, two pediatric radiologists (16 and 8 years of experience, respectively) retrospectively reviewed radiographs along with cranial CT and annotated fractures on the radiographs in consensus using the polyline tool of an image annotation freeware (VGG Image Annotator [18]). However, in the external test set, we performed only per-patient labeling based on cranial CT.

Deep Learning Model Development

We used the YOLOv3 architecture [19], which is one of the best-known object detection deep learning frameworks, to perform a per-image detection of skull fractures. The outputs of the model were the coordinates of the predicted bounding boxes and scores in the range of 0 to 1. Technical details regarding data preprocessing and model development are provided in Supplementary Material. After training the model for interpretation as per patient, we defined the prediction score of a patient as the maximum score of all candidate bounding boxes predicted from the patient's images.

Observer Study

An observer study was conducted using an external test set. Two radiology residents (with 2 years of experience), a pediatric radiologist (with 8 years of experience), and two emergency physicians (both with 5 years of experience) participated in a two-session review of the skull radiographs in the external test set. We provided them anonymized original Digital Imaging and Communications in Medicine files, except for age and sex, and the readers were aware that the study consisted of pediatric head trauma patients. Only radiographs were obtained during the first session. The second session, which was held two weeks after the first session and altered the review order of the patients, included model assistance. Plain radiographs were presented along with images annotated with bounding boxes and scores predicted using the deep learning model (Supplementary Fig. 1). In both sessions, the readers recorded the final likelihood of skull fracture that they decided (either with or without AI results) in each patient on a 5-point scale (1, definitely normal; 2, probably normal; 3, indeterminate; 4, probable fracture; 5, definite fracture).

Statistical Analysis

The area under the receiver operating characteristic curve (AUROC) was calculated. For binary classification with the model, we chose three cutoff points according to the results from the internal test set: an optimal cutoff point yielding the maximum value of the Youden index [20], a high-sensitivity cutoff point yielding 90% sensitivity, and a high-specificity cutoff point yielding 95% specificity. For human readers, AUROC was obtained using 5-point diagnostic

Table 1. Summary of Patient Characteristics

Parameter	Training Set (n = 288)	Tuning Set (n = 42)	Internal Test Set (n = 83)	External Test Set (n = 95)
Age, years				
< 2	119 (90, 29)	21 (15, 6)	38 (26, 12)	25 (12, 13)
≥ 2	169 (40, 129)	21 (4, 17)	45 (11, 34)	70 (7, 63)
Sex				
Male	173 (82, 91)	25 (13, 12)	49 (24, 25)	69 (8, 61)
Female	115 (48, 67)	17 (6, 11)	34 (13, 21)	26 (11, 15)
Label				
Fracture-positive	130	19	37	19
Fracture-negative	158	23	46	76

Data represent the total number of patients. The data within parentheses represent patients with and without fractures.

 https://doi.org/10.3348/kjr.2021.0449

confidence levels, and they were dichotomized into normal (score 1 to 3) and fracture (scores 4 and 5) for binary diagnosis. We obtained sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) from the confusion matrices. We used the DeLong et al. [21] method to compare individual AUROC values and the McNemar test to compare the sensitivity and specificity values. For the comparison of AUROC values pooled across readers, we performed a multi-reader multi-case (MRMC) ROC analysis using the Obuchowski-Rockette method for fixed-reader random case [22,23]. Preverbal (< 2 years of age) children are considered separate from older children in clinical decision rules for pediatric head trauma due to their higher risk of injuries [24]. Therefore, we performed subgroup analyses based on patient age (< 2 years vs. ≥ 2 years). For subgroup analyses and comparisons with human readers, the model's performance at the optimal cutoff was used. Statistical significance was set at $p < 0.05$. We used

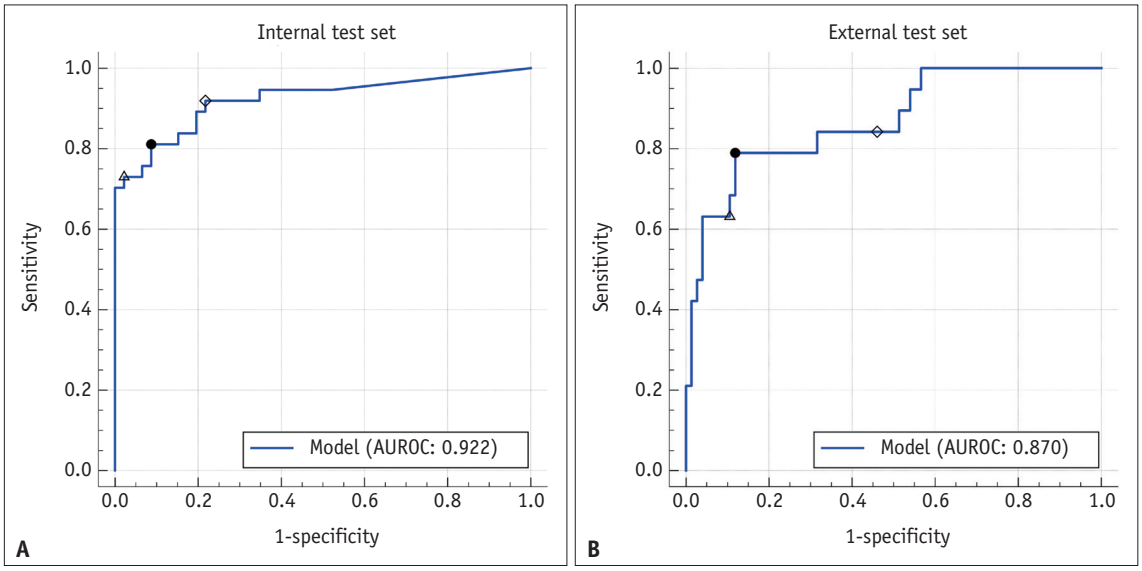

Fig. 2. AUROC curves of the model for per-patient diagnosis of skull fracture in the internal test set (A) and external test set (B) for all patients (•: optimal cutoff [0.43], ◊: high-sensitivity cutoff [0.05], △: high-specificity cutoff [0.63]). AUROC = area under the receiver operating characteristic curve

Table 2. Standalone Performance of Deep Learning Model

	AUROC	Binary Classification				
		Cutoff*	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Internal test set						
All patients (n = 83)	0.922 (0.842–0.969)	Not applicable				
All patients (n = 83)		Optimal: 0.43	81.1 (64.8–92.0)	91.3 (79.2–97.6)	88.2	85.7
All patients (n = 83)		High sensitivity: 0.05	91.9 (78.1–98.3)	78.3 (63.6–89.1)	77.3	92.3
All patients (n = 83)		High specificity: 0.63	73.0 (55.9–86.2)	97.8 (88.5–99.9)	96.4	81.8
External test set						
All patients (n = 95)	0.870 (0.785–0.930)	Not applicable				
Ages < 2 years (n = 25)	0.885 (0.694–0.977)	Not applicable				
Ages ≥ 2 years (n = 70)	0.778 (0.663–0.868)	Not applicable				
All patients (n = 95)		Optimal: 0.43	78.9 (54.4–93.9)	88.2 (78.7–94.4)	62.5	94.4
Ages < 2 years (n = 25)		Optimal: 0.43	91.7 (61.5–99.8)	84.6 (54.6–98.1)	84.6	91.7
Ages ≥ 2 years (n = 70)		Optimal: 0.43	57.1 (18.4–90.1)	88.9 (78.4–95.4)	36.7	94.9
All patients (n = 95)		High sensitivity: 0.05	84.2 (60.4–96.6)	53.9 (42.1–65.5)	31.4	93.2
All patients (n = 95)		High specificity: 0.63	63.2 (38.4–83.7)	89.5 (80.3–95.3)	60.0	90.7

Data are presented in percentage with 95% confidence intervals in parentheses, if available. *Operating points were derived from the internal test set. Optimal indicates the cutoff value yielding the maximum value of the Youden index. AUROC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value

https://doi.org/10.3348/kjr.2021.0449

Korean Journal of Radiology
KJR

the RJafroc package [25] in R version 4.1.1 (R Project for Statistical Computing, https://www.r-project.org) for the MRMC ROC analysis. All other data were analyzed using MedCalc version 12.7 (MedCalc Software).

RESULTS

Patient Characteristics

The development dataset included a total of 413 patients (median age and interquartile range, 3 years, 0–7 years; 247 male, 166 female; 186 with fracture, 227 without fracture) and the external test set included a total of 95 patients (median age and interquartile range, 7.5 years, 3–13 years; 69 males, 26 females; 19 with fracture, 76 without fracture). Patient characteristics of the datasets are summarized in Table 1.

Standalone Performance of Deep Learning Model

The developed deep learning model showed an AUROC of 0.922 (95% confidence interval [CI], 0.842–0.969) in the internal test set and 0.870 (95% CI, 0.785–0.930) in the external test set (Fig. 2). Table 2 shows the sensitivity, specificity, PPV, and NPV of the proposed model. When the cutoff by the maximum Youden index value was applied, the

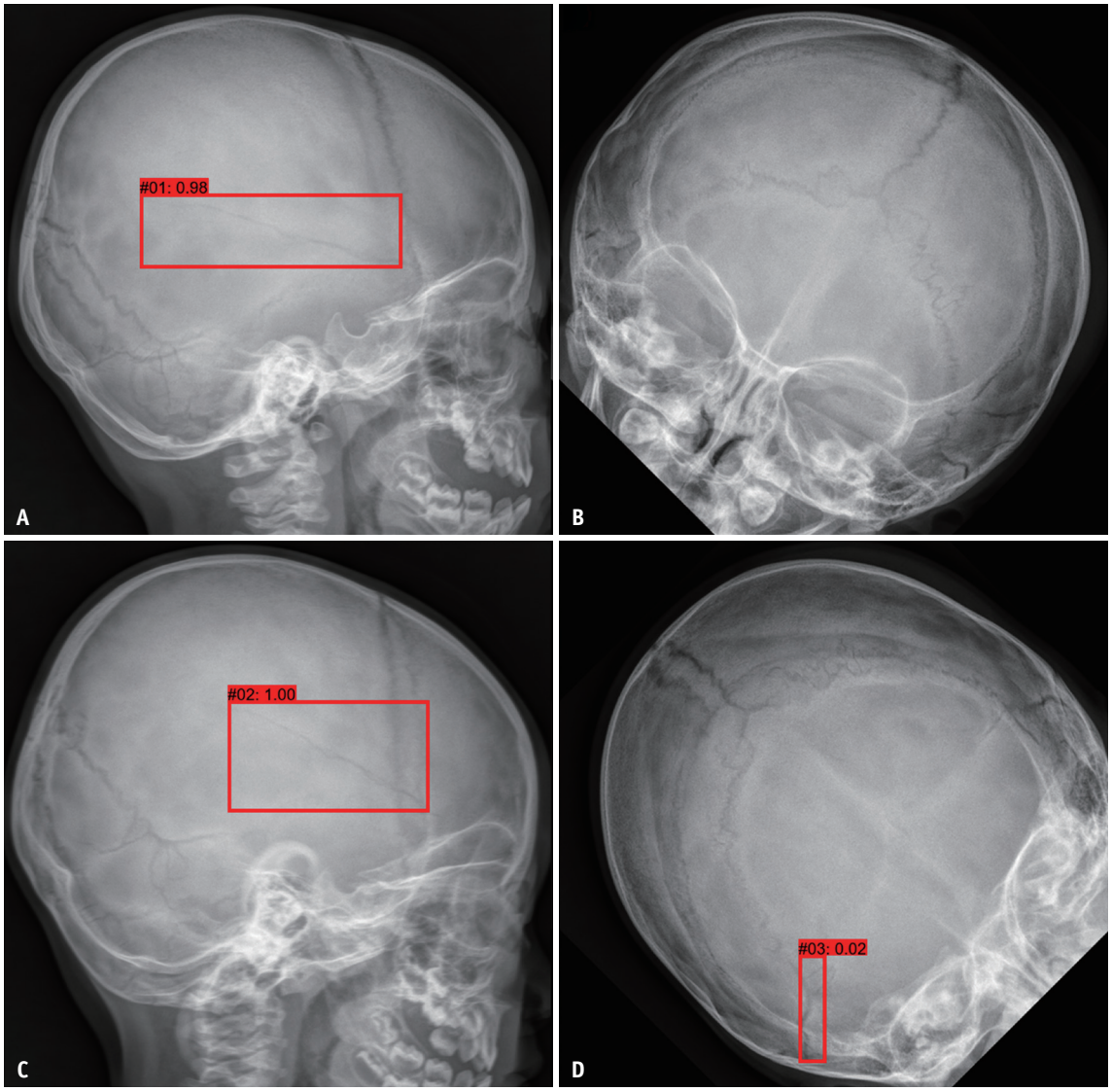

Fig. 3. Representative true-positive case: a 1-year-old boy with a left parietal bone fracture.
A-D. Skull left lateral **(A)**, anteroposterior **(B)**, right lateral **(C)**, and Town view **(D)** radiographs show a left parietal bone fracture. The model correctly detected the fracture on the skull left **(A)** and right lateral radiograph **(C)** with a prediction score of 0.98 and 1.00, respectively. All radiology residents and emergency physicians rated this case as from "probably normal" to "probable fracture" in the first session. However, all changed to "definite fracture" in the second session with the model's assistance.

https://doi.org/10.3348/kjr.2021.0449

model had a sensitivity of 81.1% (95% CI, 64.8%–92.0%) and a specificity of 91.3% (95% CI, 79.2%–97.6%) in the internal test set and a sensitivity of 78.9% (95% CI, 54.4%–93.9%) and specificity of 88.2% (95% CI, 78.7%–94.4%) in the external test set. The results for subgroups of age < 2 years and ≥ 2 years are provided in Table 2.

The median (range) number of positive and false-positive calls by the model (score above the optimal cutoff [0.43]) per patient in the external test set was 0 (0–6) and 0 (0–4), respectively. The number of total and false-positive bounding boxes (score above 0.001) was 2 (0–6) and 1 (0–6), respectively, per patient. Figures 3-6 illustrate representative true-positive, false-positive, false-negative, and true-negative cases, respectively, from the external test set.

Observer Performance with and without Deep Learning Model Assistance

Table 3 and Figure 7 show the diagnostic performance of human readers in the external test set with and without the model's assistance. In the first session, the AUROCs of the observers ranged from 0.684 to 0.949 and showed no significant differences compared with the model in all patients and the age subgroups ($p > 0.05$; see Supplementary Table 1 for details). The sensitivity and specificity of the observers ranged from 0.0% to 91.7% and from 46.2% to 96.8%, respectively.

In the second session with the model's assistance, improvement was noted for some of the performance parameters in some of the readers (Table 3). Significant AUROC improvements were observed by pooling the results of radiology residents (0.094 [95% CI, 0.020–0.168], $p = 0.012$) or the results of emergency physicians (0.069 [95% CI, 0.002–0.136], $p = 0.043$), but not in the pediatric radiologist (0.008 [95% CI, -0.074–0.090], $p = 0.850$). Compared with the first session, all readers showed comparable or higher sensitivities (improvements of 0.0%–10.5%) and higher specificities (improvements of 2.6%–17.1%), but statistical significance was achieved only in the specificity of one radiology resident ($p = 0.002$). For patients younger than 2 years, the pooled AUROC improvements with the model's assistance were not significant in radiology residents (0.146 [95% CI, -0.027–0.318], $p = 0.097$), pediatric radiologist (-0.067 [95% CI, -0.153–0.018], $p = 0.124$), or emergency physicians (0.032 [95% CI, -0.108–0.172], $p = 0.654$). A significant AUROC improvement was observed in one radiology resident (0.231 [95% CI, 0.027–0.434], $p = 0.026$), while other individual readers showed no significant differences in AUROC, sensitivity, and specificity ($p > 0.05$). For patients aged 2 years and older, no significant pooled AUROC improvements with the model's assistance were demonstrated in radiology residents (0.093 [95% CI, -0.074–0.260], $p = 0.276$), pediatric radiologists (0.108 [95% CI, -0.072–0.287], $p = 0.240$), or emergency physicians (0.117 [95% CI, -0.021–0.256], $p = 0.097$). An emergency physician

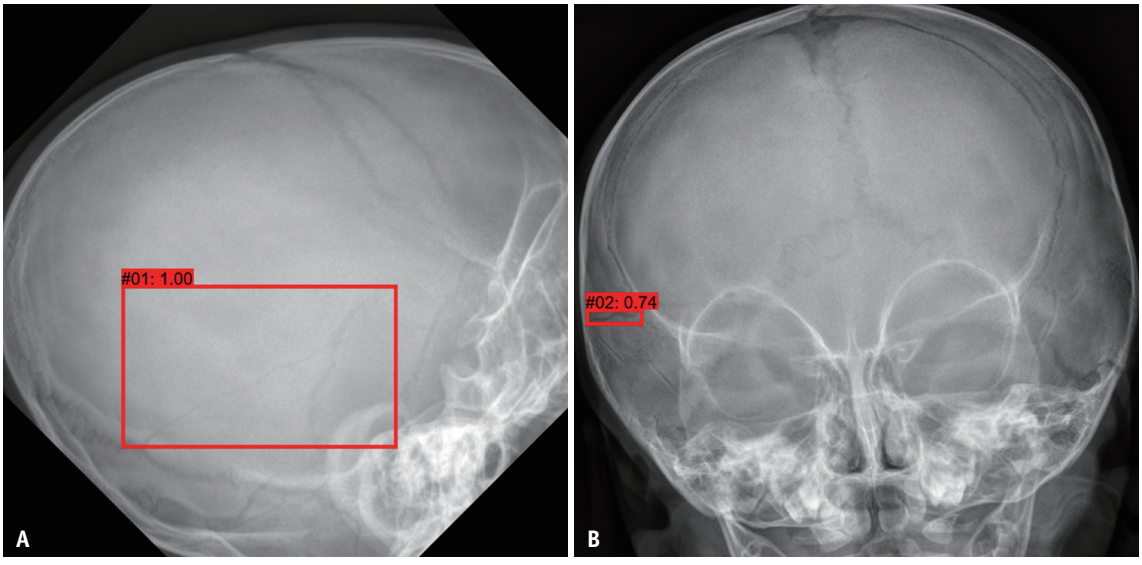

Fig. 4. Representative false-positive case: a 9-month-old girl with no skull fracture.
A, B. The model marked bounding boxes with prediction scores of 1.00 on the skull right lateral radiograph **(A)** and 0.74 on the skull anteroposterior radiograph **(B)**. Additionally, the pediatric radiologist and one radiology resident rated this case as "probable fracture." However, the concurrent cranial CT confirmed that there was no skull fracture.

https://doi.org/10.3348/kjr.2021.0449

Korean Journal of Radiology
KJR

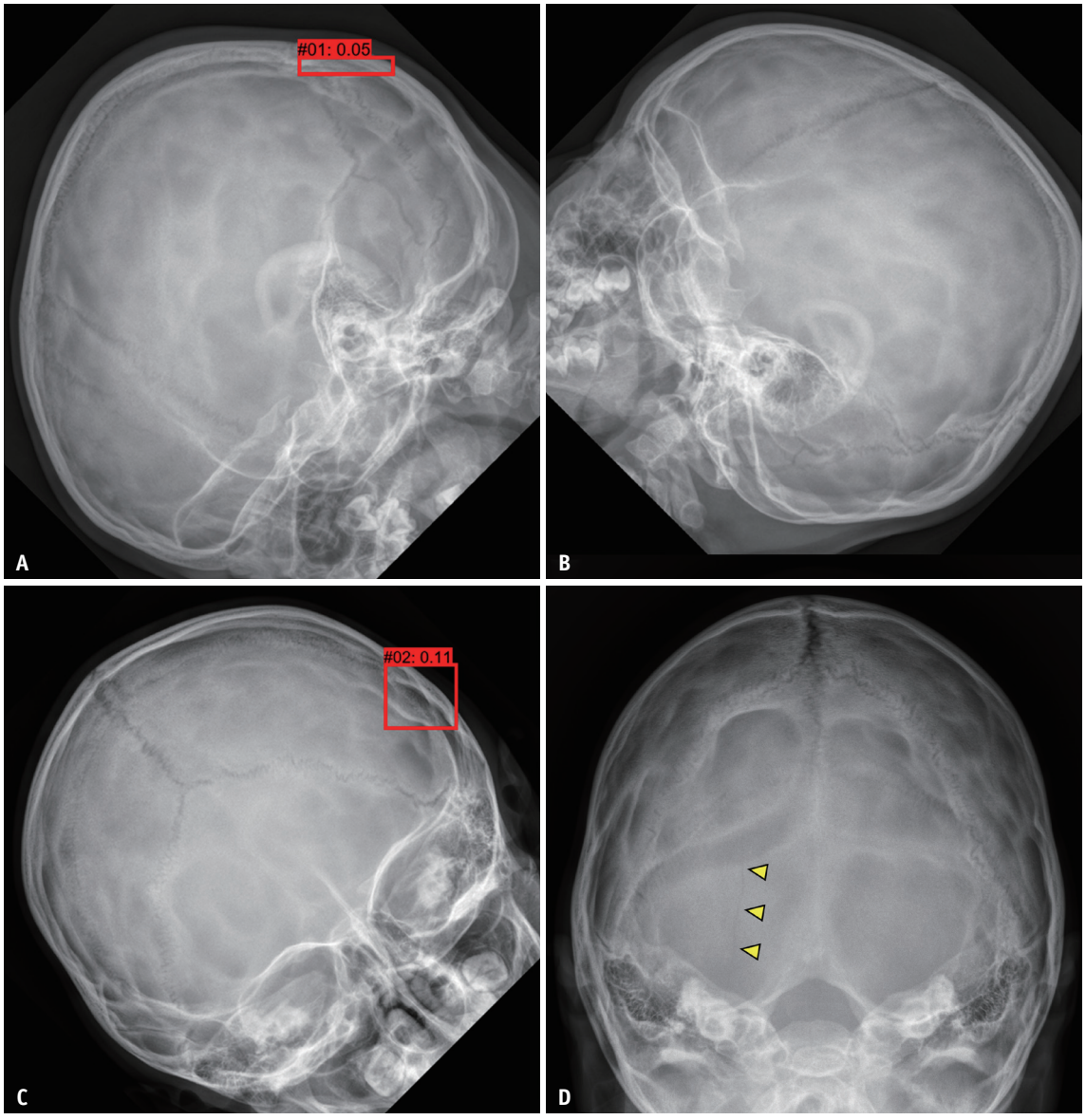

Fig. 5. Representative false-negative case: a 3-year-old boy with a right occipital bone fracture.
A-D. The model showed the highest prediction score of 0.11 on the skull anteroposterior radiograph **(C)**. However, the Towne view **(D)** showed a radiolucent line (arrowheads) in the right occipital bone, which was not evident in the right lateral **(A)**, left lateral **(B)**, and anteroposterior **(C)** views. The concurrent cranial CT confirmed the right occipital bone fracture. All radiologists and one emergency physician correctly rated this case as "definite fracture" or "probable fracture."

showed a significant AUROC improvement (0.173 [95% CI, 0.015–0.332], $p = 0.032$) and a radiology resident showed a significant improvement in specificity (76.2% to 88.9%, $p = 0.039$), while other readers showed no significant differences in diagnostic performance ($p > 0.05$).

DISCUSSION

We implemented and validated a deep learning model for the automated detection of pediatric skull fractures on plain radiographs. To the best of our knowledge, this is the first study to demonstrate the feasibility and clinical validity of a deep learning algorithm for the diagnosis of skull fractures on plain radiography. Although many recent studies have utilized deep learning to detect fractures on radiographs [26], few have involved the pediatric population [27]. Furthermore, we not only compared the stand-alone performance of our developed model with radiologists and emergency physicians, but also demonstrated the effect of the assistance of the model on the readers' performance.

https://doi.org/10.3348/kjr.2021.0449

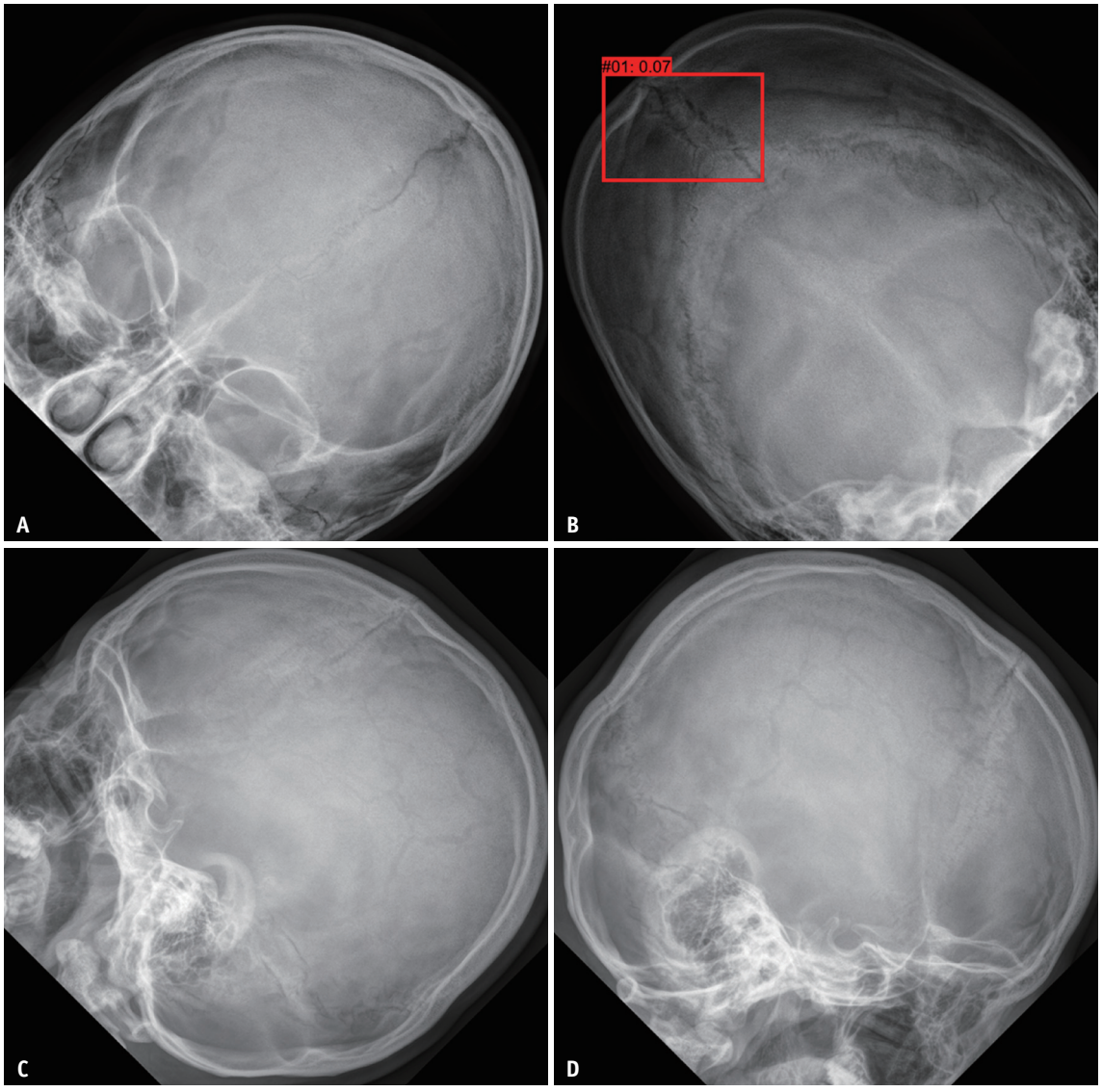

Fig. 6. Representative true-negative case: an 8-year-old boy with no skull fracture.
A-D. Skull anteroposterior **(A)**, Town view **(B)**, right lateral **(C)**, and left lateral **(D)** radiographs show no definite skull fracture. The model showed the highest prediction score of 0.07 on the Towne view radiograph **(B)**. All emergency physicians and one radiology resident rated this case as "probable fracture" or "definite fracture" in the first session. They all changed to "definitely normal" in the second session with the model's assistance.

Many patients undergo skull radiography, even though CT is the modality of choice for pediatric head trauma [6,7]. A significant drawback of CT is radiation exposure [28], which is resource-intensive and poses additional risks for patients who require sedation [29]. Thus, there exist several clinical decision rules, such as the Pediatric Emergency Care Applied Research Network (PECARN) [24] rules, that were developed to reduce unnecessary CT examinations. The PECARN rules demonstrate accurate recommendations for performing and avoiding CT in high-risk and low-risk head injuries, respectively [30,31]. However, the management of intermediate-risk injuries involves clinical settings with other factors, including physician experience and parental preference [24]. In such cases, for fractures screened on skull radiography, clinical signs of head injury may be decisive in determining further CT evaluation.

Despite the high number of skull radiographs performed in pediatric head trauma, interpreting them can be a diagnostic challenge [32,33]. The sensitivity of radiography for pediatric skull fractures is 74%–81% [32,34], which is similar to or slightly higher than our external test results from the radiologists without the model's assistance

https://doi.org/10.3348/kjr.2021.0449

Korean Journal of Radiology
KJR

Table 3. Observer Performance with and without Deep Learning Model Assistance

	Model-Unassisted R1	Model-Assisted R1	Model-Unassisted R2	Model-Assisted R2	Model-Unassisted PR	Model-Assisted PR	Model-Unassisted EM1	Model-Assisted EM1	Model-Unassisted EM2	Model-Assisted EM2	Model-Unassisted vs. Model-Assisted (*P* value)				
											R1	R2	PR	EM1	EM2
All (n = 95)															
AUROC	0.765 (0.667–0.846)	0.882 (0.800–0.939)	0.846 (0.757–0.912)	0.917 (0.843–0.964)	0.908 (0.831–0.958)	0.916 (0.841–0.963)	0.848 (0.760–0.914)	0.914 (0.839–0.962)	0.812 (0.719–0.885)	0.885 (0.803–0.941)	0.030	0.080	0.850	0.088	0.165
Sensitivity, %	68.4 (43.5–87.4)	78.9 (54.4–93.9)	73.7 (48.8–90.9)	78.9 (54.4–93.9)	73.7 (48.8–90.9)	84.2 (60.4–96.6)	52.6 (28.9–75.6)	63.2 (38.4–83.7)	73.7 (48.8–90.9)	73.7 (48.8–90.9)	0.688	1.000	0.625	0.500	0.688
Specificity, %	71.1 (59.5–80.9)	88.2 (78.7–94.4)	84.2 (74.0–91.6)	89.5 (80.3–95.3)	94.7 (87.1–98.5)	98.7 (93.0–100.0)	92.1 (83.6–97.1)	94.7 (87.1–98.6)	78.9 (68.1–87.5)	84.2 (74.0–91.6)	0.002	0.388	1.000	1.000	0.481
Ages < 2 years (n = 25)															
AUROC	0.686 (0.471–0.855)	0.917 (0.735–0.989)	0.849 (0.650–0.960)	0.910 (0.727–0.987)	0.949 (0.780–0.997)	0.881 (0.780–0.997)	0.929 (0.753–0.993)	0.824 (0.620–0.946)	0.875 (0.682–0.972)	0.942 (0.771–0.996)	0.026	0.495	0.124	0.427	0.254
Sensitivity, %	75.0 (42.8–94.5)	91.7 (61.5–99.8)	91.7 (61.5–99.8)	91.7 (61.5–99.8)	91.7 (61.5–99.8)	91.7 (61.5–99.8)	83.3 (51.6–97.9)	91.7 (61.5–99.8)	83.3 (51.6–97.9)	91.7 (61.5–99.8)	0.625	NA	NA	1.000	1.000
Specificity, %	46.2 (19.2–74.9)	84.6 (54.6–98.1)	76.9 (46.2–95.0)	76.9 (46.2–95.0)	84.6 (54.6–98.1)	84.6 (54.6–98.1)	84.6 (54.6–98.1)	84.6 (54.5–98.1)	61.5 (31.6–86.1)	76.9 (46.2–95.0)	0.063	1.000	NA	0.500	0.688
Ages ≥ 2 years (n = 70)															
AUROC	0.684 (0.562–0.790)	0.785 (0.670–0.874)	0.757 (0.640–0.852)	0.842 (0.736–0.919)	0.790 (0.676–0.878)	0.898 (0.802–0.958)	0.693 (0.571–0.798)	0.866 (0.764–0.936)	0.684 (0.562–0.790)	0.745 (0.627–0.842)	0.394	0.354	0.240	0.032	0.553
Sensitivity, %	57.1 (18.4–90.1)	57.1 (18.4–90.1)	42.9 (9.9–81.6)	57.1 (18.4–90.1)	42.9 (9.9–81.6)	71.4 (29.0–96.3)	0.0 (0.0–41.0)	14.3 (0.4–57.9)	57.1 (18.4–90.1)	42.9 (9.9–81.6)	1.000	1.000	0.625	1.000	1.000
Specificity, %	76.2 (63.8–86.0)	88.9 (78.4–95.4)	85.7 (74.6–93.3)	92.1 (82.4–97.4)	96.8 (89.0–99.6)	98.4 (91.5–100.0)	93.7 (84.5–98.2)	96.8 (89.0–99.6)	82.5 (70.9–90.9)	85.7 (74.6–93.2)	0.039	0.289	1.000	0.688	0.774

Data except for AUROC are presented in percentage with 95% confidence intervals in parentheses. AUROC = area under the receiver operating characteristic curve, EM = emergency physician, NA = not applicable, PR = pediatric radiologist, R = radiology resident

Fig. 7. AUROC curves of the model and the human reviewers for diagnosis of skull fracture on the external test set for all patients. Dashed and solid lines indicate the first (without model's assistance) and second (with model's assistance) sessions, respectively. AUROC = area under the receiver operating characteristic curve, EM = emergency physician

(68.4%–73.7%). Our developed model showed high sensitivities of 78.9% (95% CI, 54.4%–93.9%) in the external test set and 81.1% (95% CI, 64.8%–92.0%) in the internal test set. The difference in the diagnostic performance of the model between the internal and external test sets may be attributed to several factors. The model underperformed in patients aged ≥ 2 years, particularly in terms of sensitivity, and the external test set had more patients aged ≥ 2 years with fractures (37%) than the internal test set (29.7%). Moreover, only a few Towne view studies were in the development dataset compared with the external test set. Thus, we expected the model to miss fractures identified only in the Towne view (Fig. 5). In addition, three of the four fractures missed by the model were occipital fractures, which are usually better depicted on the Towne view. Two of them were correctly diagnosed by all the radiologists. Nevertheless, the relatively low sensitivity in the external test set is reasonable, as overestimating the model's performance during internal validation due to overfitting is a well-known problem in deep learning [35].

In the observer study, significant pooled AUROC improvements were observed in radiology residents (0.094 [95% CI, 0.020–0.168], $p = 0.012$) and emergency

https://doi.org/10.3348/kjr.2021.0449

physicians (0.069 [95% CI, 0.002–0.136], $p = 0.043$). The improvements in diagnostic performance tended to be higher in specificity than sensitivity, which can be attributed to the high stand-alone specificity of the model. A greater improvement in specificity may be important in certain clinical scenarios. Clinical decision rules for neuroimaging in pediatric head trauma are very sensitive [36,37], and skull radiography is not recommended if CT is indicated [4]. Thus, skull radiography is often performed in patients with a low risk of brain injury. False-positive detections of fractures on skull radiographs would lead to unnecessary CT examinations and prolonged hospital stays. Several previous studies have suggested the potential role of a deep learning model as a second reader for inexperienced radiologists or physicians [38,39]. We also believe that our developed model may be beneficial in reducing the number of false-positive interpretations of skull radiographs.

We performed subgroup analysis based on the age of 2 years, as included in the PECARN rules [24]. Preverbal (< 2 years of age) children have traditionally been considered separate from older children because they are more difficult to assess, have a higher risk of injuries, and have a higher incidence of asymptomatic intracranial injuries and skull fractures due to minor trauma [40]. Our model tended to perform better in patients younger than 2 years than in older children, particularly with a higher sensitivity (91.7% vs. 57.1%) but a comparable specificity (84.6% vs. 88.9%). The human readers also showed such a tendency, as previous studies reported similar patterns of models and radiologists not only in diagnosis but also in misdiagnosis [38,41]. This tendency implies that it is more likely to be due to a demographic factor rather than being specific to the model. In older children, sutures and vascular grooves become more prominent on plain radiographs and thus may mimic fracture lines or be even more vivid than actual fracture lines.

Recent studies have used deep learning for fracture detection using classification models [26], but we implemented an object detection model because it has an advantage over a classification model in providing more explainable output. The transparency of a model can be crucial in computer-aided diagnosis, where experts must understand and validate the model's prediction [42]. There are several methods, including class-activation maps, to visualize the localization information of a classification model [42]. However, they are unable to produce localization information at a high resolution or for multiple objects. Conversely, bounding boxes with probabilities from a detection model are direct indicators of how the model predicts and enables precise localization, even for multiple objects.

Our study has several limitations. First, because not all pediatric head trauma patients undergo both skull radiography and CT, our datasets may not represent the general pediatric head trauma population. Second, this was a retrospective study with a limited amount of data. A further prospective study with a larger cohort is warranted to improve the diagnostic performance and generalizability of the deep learning model. Lastly, the learning effect from the model-unassisted reading might have affected the results of the model-assisted session.

In conclusion, a deep learning model could improve the performance of inexperienced radiologists and emergency physicians in the diagnosis of pediatric skull fractures on plain radiographs.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2021.0449.

Availability of Data and Material

The datasets generated or analyzed during the study are not publicly available due to patient-related data but are available from the corresponding author on reasonable request.

Conflicts of Interest

Young Hun Choi and Jung-Eun Cheon who is on the editorial board of the *Korean Journal of Radiology* was not involved in the editorial evaluation or decision to publish this article. All remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Jae Won Choi, Yeon Jin Cho, Young Hun Choi, Jung-Eun Cheon, Woo Sun Kim. Data curation: Jae Won Choi, Yeon Jin Cho, Ji Young Ha, Yun Young Lee, Young Hun Choi. Formal analysis: Jae Won Choi, Yeon Jin Cho. Investigation: Jae Won Choi, Yeon Jin Cho, Seok Young Koh, June Young Seo, Injoon Kim, Jaekwang Yang. Methodology: Jae Won Choi, Yeon Jin Cho, Young Hun Choi, Woo Sun Kim. Project administration: Yeon Jin Cho, Woo Sun Kim. Resources: Yeon Jin Cho, Woo Sun Kim, Ji Young

Korean Journal of Radiology
KJR

Ha, Yun Young Lee. Software: Jae Won Choi. Supervision: Young Hun Choi, Jung-Eun Cheon, Ji Hoon Phi, Woo Sun Kim. Visualization: Jae Won Choi, Yeon Jin Cho. Writing—original draft: Jae Won Choi, Yeon Jin Cho. Writing—review & editing: Jae Won Choi, Yeon Jin Cho, Young Hun Choi, Ji Hoon Phi, Woo Sun Kim.

ORCID iDs
Jae Won Choi
https://orcid.org/0000-0002-5937-7238
Yeon Jin Cho
https://orcid.org/0000-0001-9820-3030
Ji Young Ha
https://orcid.org/0000-0001-5769-3045
Yun Young Lee
https://orcid.org/0000-0002-1200-9462
Seok Young Koh
https://orcid.org/0000-0003-1853-3166
June Young Seo
https://orcid.org/0000-0002-9572-7573
Young Hun Choi
https://orcid.org/0000-0002-1842-9062
Jung-Eun Cheon
https://orcid.org/0000-0003-1479-2064
Ji Hoon Phi
https://orcid.org/0000-0002-9603-5843
Injoon Kim
https://orcid.org/0000-0001-6892-489X
Jaekwang Yang
https://orcid.org/0000-0002-1336-2137
Woo Sun Kim
https://orcid.org/0000-0003-2184-1311

Funding Statement
None

REFERENCES

1. Marin JR, Weaver MD, Yealy DM, Mannix RC. Trends in visits for traumatic brain injury to emergency departments in the United States. *JAMA* 2014;311:1917-1919
2. Greenes DS, Schutzman SA. Clinical indicators of intracranial injury in head-injured infants. *Pediatrics* 1999;104:861-867
3. Schutzman SA, Greenes DS. Pediatric minor head trauma. *Ann Emerg Med* 2001;37:65-74
4. Expert Panel on Pediatric Imaging, Ryan ME, Pruthi S, Desai NK, Falcone RA Jr, Glenn OA, et al. ACR appropriateness criteria® head trauma-child. *J Am Coll Radiol* 2020;17:S125-S137
5. Burstein B, Upton JEM, Terra HF, Neuman MI. Use of CT for head trauma: 2007-2015. *Pediatrics* 2018;142:e20180814
6. Kim HB, Kim DK, Kwak YH, Shin SD, Song KJ, Lee SC, et al. Epidemiology of traumatic head injury in Korean children. *J Korean Med Sci* 2012;27:437-442
7. Furtado LMF, da Costa Val Filho JA, Dos Santos AR, E Sá RF, Sandes BL, Hon Y, et al. Pediatric minor head trauma in Brazil and external validation of PECARN rules with a cost-effectiveness analysis. *Brain Inj* 2020;34:1467-1471
8. Carrière B, Clément K, Gravel J. Variation in the use of skull radiographs by emergency physicians in young children with minor head trauma. *CJEM* 2014;16:281-287
9. Expert Panel on Pediatric Imaging, Wootton-Gorges SL, Soares BP, Alazraki AL, Anupindi SA, Blount JP, et al. ACR appropriateness criteria® suspected physical abuse-child. *J Am Coll Radiol* 2017;14:S338-S349
10. Tang PH, Lim CC. Imaging of accidental paediatric head trauma. *Pediatr Radiol* 2009;39:438-446
11. Paul AR, Adamo MA. Non-accidental trauma in pediatric patients: a review of epidemiology, pathophysiology, diagnosis and treatment. *Transl Pediatr* 2014;3:195-207
12. Rajaram S, Batty R, Rittey CD, Griffiths PD, Connolly DJ. Neuroimaging in non-accidental head injury in children: an important element of assessment. *Postgrad Med J* 2011;87:355-361
13. Idriz S, Patel JH, Ameli Renani S, Allan R, Vlahos I. CT of normal developmental and variant anatomy of the pediatric skull: distinguishing trauma from normality. *Radiographics* 2015;35:1585-1601
14. George CLS, Harper NS, Guillaume D, Cayci Z, Nascene D. Vascular channel mimicking a skull fracture. *J Pediatr* 2017;181:326
15. Chung S, Schamban N, Wypij D, Cleveland R, Schutzman SA. Skull radiograph interpretation of children younger than two years: how good are pediatric emergency physicians? *Ann Emerg Med* 2004;43:718-722
16. Do S, Song KD, Chung JW. Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning. *Korean J Radiol* 2020;21:33-41
17. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. *Skeletal Radiol* 2020;49:183-197
18. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. Proceedings of the 27th ACM International Conference on Multimedia; 2019 Oct 21-25; New York, NY, USA: Association for Computing Machinery; 2019; p. 2276-2279
19. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv [Preprint]. 2018 [cited 2020 December 14]. Available at: https://arxiv.org/abs/1804.02767
20. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32-35
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing

the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837-845
22. Obuchowski Jr NA, Rockette Jr HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Commun Stat Simul Comput* 1995;24:285-308
23. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat Med* 2007;26:596-619
24. Kuppermann N, Holmes JF, Dayan PS, Hoyle JD Jr, Atabaki SM, Holubkov R, et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *Lancet* 2009;374:1160-1170
25. Chakraborty DP. *Observer performance methods for diagnostic imaging*: *foundations, modeling, and applications with r-based examples*. Boca Raton: CRC Press, 2017
26. Yang S, Yin B, Cao W, Feng C, Fan G, He S. Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. *Clin Radiol* 2020;75:713.e17-713.e28
27. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, et al. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. *Invest Radiol* 2020;55:101-110
28. Miglioretti DL, Johnson E, Williams A, Greenlee RT, Weinmann S, Solberg LI, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA Pediatr* 2013;167:700-707
29. Goldwasser T, Bressan S, Oakley E, Arpone M, Babl FE. Use of sedation in children receiving computed tomography after head injuries. *Eur J Emerg Med* 2015;22:413-418
30. Babl FE, Lyttle MD, Bressan S, Borland M, Phillips N, Kochar A, et al. A prospective observational study to assess the diagnostic accuracy of clinical decision rules for children presenting to emergency departments after head injuries (protocol): the Australasian Paediatric Head Injury Rules Study (APHIRST). *BMC Pediatr* 2014;14:148
31. Easter JS, Bakes K, Dhaliwal J, Miller M, Caruso E, Haukoos JS. Comparison of PECARN, CATCH, and CHALICE rules for children with minor head injury: a prospective cohort study. *Ann Emerg Med* 2014;64:145-152
32. Kim YI, Cheong JW, Yoon SH. Clinical comparison of the predictive value of the simple skull X-ray and 3 dimensional computed tomography for skull fractures of children. *J Korean Neurosurg Soc* 2012;52:528-533
33. Oh CK, Yoon SH. The significance of incomplete skull fracture in the birth injury. *Med Hypotheses* 2010;74:898-900
34. Martin A, Paddock M, Johns CS, Smith J, Raghavan A, Connolly DJA, et al. Avoiding skull radiographs in infants with suspected inflicted injury who also undergo head CT: "a no-brainer?" *Eur Radiol* 2020;30:1480-1487
35. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean J Radiol* 2021;22:442-453
36. Lorton F, Poullaouec C, Legallais E, Simon-Pimmel J, Chêne MA, Leroy H, et al. Validation of the PECARN clinical decision rule for children with minor head trauma: a French multicenter prospective study. *Scand J Trauma Resusc Emerg Med* 2016;24:98
37. Ide K, Uematsu S, Tetsuhara K, Yoshimura S, Kato T, Kobayashi T. External validation of the PECARN head trauma prediction rules in Japan. *Acad Emerg Med* 2017;24:308-314
38. Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 2019;293:573-580
39. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019;2:e191095
40. Schutzman SA, Barnes P, Duhaime AC, Greenes D, Homer C, Jaffe D, et al. Evaluation and management of children younger than two years old with apparently minor head trauma: proposed guidelines. *Pediatrics* 2001;107:983-993
41. Kim Y, Lee KJ, Sunwoo L, Choi D, Nam CM, Cho J, et al. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. *Invest Radiol* 2019;54:7-15
42. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, von Tengg-Kobligk H, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell* 2020;2:e190043

https://doi.org/10.3348/kjr.2021.0449