Original Article | Pediatric Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2020.0051
Korean J Radiol 2021;22(4):612-623

Korean Journal of Radiology
KJR

Diagnostic Performance of a New Convolutional Neural Network Algorithm for Detecting Developmental Dysplasia of the Hip on Anteroposterior Radiographs

Hyoung Suk Park, PhD[1]*, Kiwan Jeon, PhD[1]*, Yeon Jin Cho, MD[2], Se Woo Kim, MD[2, 3], Seul Bi Lee, MD[2], Gayoung Choi, MD[2], Seunghyun Lee, MD[2], Young Hun Choi, MD[2, 3], Jung-Eun Cheon, MD[2, 3, 4], Woo Sun Kim, MD[2, 3, 4], Young Jin Ryu, MD[5], Jae-Yeon Hwang, MD[6]

[1]Division of Medical Mathematics, National Institute for Mathematical Sciences, Daejeon, Korea; [2]Department of Radiology, Seoul National University Hospital, Seoul, Korea; [3]Department of Radiology, Seoul National University College of Medicine, Seoul, Korea; [4]Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Korea; [5]Department of Radiology, Seoul National University Bundang Hospital, Seongnam, Korea; [6]Department of Radiology, Pusan National University Yangsan Hospital, Yangsan, Korea

Objective: To evaluate the diagnostic performance of a deep learning algorithm for the automated detection of developmental dysplasia of the hip (DDH) on anteroposterior (AP) radiographs.

Materials and Methods: Of 2601 hip AP radiographs, 5076 cropped unilateral hip joint images were used to construct a dataset that was further divided into training (80%), validation (10%), or test sets (10%). Three radiologists were asked to label the hip images as normal or DDH. To investigate the diagnostic performance of the deep learning algorithm, we calculated the receiver operating characteristics (ROC), precision-recall curve (PRC) plots, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) and compared them with the performance of radiologists with different levels of experience.

Results: The area under the ROC plot generated by the deep learning algorithm and radiologists was 0.988 and 0.988–0.919, respectively. The area under the PRC plot generated by the deep learning algorithm and radiologists was 0.973 and 0.618–0.958, respectively. The sensitivity, specificity, PPV, and NPV of the proposed deep learning algorithm were 98.0, 98.1, 84.5, and 99.8%, respectively. There was no significant difference in the diagnosis of DDH by the algorithm and the radiologist with experience in pediatric radiology ($p = 0.180$). However, the proposed model showed higher sensitivity, specificity, and PPV, compared to the radiologist without experience in pediatric radiology ($p < 0.001$).

Conclusion: The proposed deep learning algorithm provided an accurate diagnosis of DDH on hip radiographs, which was comparable to the diagnosis by an experienced radiologist.

Keywords: *Deep learning; Artificial intelligence; Hip dysplasia; Child; Radiography*

INTRODUCTION

Developmental dysplasia of the hip (DDH) is a spectrum of structural abnormalities ranging from mild dysplasia and subluxation to dislocation of the femoral head (1). The incidence of DDH is approximately 1.5 to 35 in 1000

Received: January 21, 2020 **Revised:** June 26, 2020 **Accepted:** July 22, 2020
This research was supported by Research Program 2019 funded by Seoul National University College of Medicine Research Foundation. H.S.P and K.J. were supported by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korean government (No. NIMS-B20900000). This work utilized a software for screening of DDH (C-2019-015787) developed by National Institute for Mathematical Sciences.

*These authors contributed equally to this work.

Corresponding author: Yeon Jin Cho, MD, Department of Radiology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
• E-mail: blue1010c@gmail.com

Korean Journal of Radiology
KJR

persons, and the incidence of sonographic screening is 25 to 50 in 1000 persons (1, 2). For the normal development of the hip joint, articulation of the femoral head and acetabulum is essential; thus, early diagnosis and treatment of DDH are important (1). The recommended diagnostic methods differ according to the patient's age (3). Clinical screening is one of the most widely used screening methods, but the results have low reproducibility (4). Sonography and conventional radiography are the most widely used screening tools. Conventional radiography is a useful diagnostic tool as it is readily available at a low cost. The anteroposterior (AP) view, frog-leg lateral view, or von Rosen view can be used for the diagnosis of DDH; however, the hip AP view is usually sufficient for screening DDH (5). The diagnostic performance of conventional radiography for the detection of DDH in infants can vary according to several factors such as the patient's age, technical adequacy, and reader's experience (6). In particular, the interpreter's experience can significantly affect the diagnostic accuracy of conventional radiography.

Recently, machine learning has made tremendous progress and is considered to be an emerging technique for classification of images (7). Recent studies have shown the potential of deep learning in lesion detection and classification on radiologic images (8-13). Using conventional radiography, several studies have shown that machine learning can be helpful for the detection of fractures in skeletal bones and lesions on chest radiography (9, 14).

To our knowledge, there have been few attempts to apply machine learning to the detection of DDH using conventional radiography. Some studies have adopted the classical machine learning technique (i.e., logistic regression) to detect DDH from the clinical features extracted from two- and three-dimensional ultrasounds (15, 16). In another study, a convolutional neural network (CNN)-based deep learning algorithm was applied to automatically measure the Sharp's angle from the hip radiography (17).

In this study, we used a CNN to directly diagnose the DDH from conventional hip AP radiographs. The purpose of this study was to develop a CNN-based deep learning algorithm to diagnose DDH from hip AP radiographs and to validate the diagnostic performance of the algorithm.

MATERIALS AND METHODS

This was a retrospective study. The study was approved by the Institutional Review Boards of three different tertiary medical centers Seoul National University Hospital (SNUH), Seoul National University Bundang Hospital (SNUBH), and Pusan National University Yangsan Hospital (PNUYH) who waived the need for informed consent (IRB No. H-1808-037-964).

Dataset and Labeling

At SNUH, patients younger than 12 months of age who were suspected of DDH and who had undergone hip AP radiography between January 2011 and June 2018 were enrolled in our study. At SNUBH and PNUYH, patients younger than 12 months of age who had undergone hip AP radiography between January 2016 and June 2018 were enrolled in this study. All patient data were anonymized, and reviewers were blinded to the diagnosis or medical history.

A total of 2601 hip AP radiographs from three different hospitals were collected. As two hip joint images were included in a single hip AP radiograph, a total of 5202 hip images were obtained. Patients who underwent corrective surgeries (87 images), had severe congenital skeletal dysplasia (five images) or had radiographs of suboptimal quality, including images taken in an inappropriate position (34 images), were excluded from the dataset. After excluding 126 inappropriate images, a total of 5076 hip images were included in this study. The data set consisted of 3433 hip images from SNUH, 1036 hip images from SNUBH, and 607 hip images from PNUYH (Table 1).

To generate a reference standard, and with the use of

Table 1. Training, Validation, and Test Datasets for DDH Detection

Hospitals	Total	Training Set		Validation Set		Test Set	
		Normal	DDH	Normal	DDH	Normal	DDH
SNUH	3433	2406	341	300	43	300	43
SNUBH	1036	800	32	97	5	97	5
PNUYH	607	452	19	65	3	66	2
Total	5076	3658	392	462	51	463	50

DDH = developmental dysplasia of the hip, PNUYH = Busan National University Yangsan Hospital, SNUBH = Seoul National University Bundang Hospital, SNUH = Seoul National University Hospital

 https://doi.org/10.3348/kjr.2020.0051

clinical information, the hip radiographs were independently reviewed and labeled by two pediatric radiologists (with 7 and 13 years of experience, respectively). In cases of disagreement between the two pediatric specialists, the case in question was discussed by the two and a diagnosis was made in consensus. If consensus could not be achieved after discussion, the corresponding ultrasound exams were reviewed and discussed again. Each hip joint was evaluated in a single hip radiograph and classified as normal or DDH. A total of 16 hip radiographs showed disagreement in the diagnosis of DDH, and so sonographic findings were used to achieve a diagnosis. A diagnosis of DDH was made when the following criteria were met: 1) high acetabular index (> 30°), 2) abnormal acetabular morphology (shallow acetabulum) and delayed femoral head ossification (evidently smaller size compared to the normal side or no ossification center at 8 months of age), 3) abnormal femoral head location out of the inferior medial quadrant of the acetabulum, or 4) disruption of Shenton's line (3). The positive and negative cases collected from each hospital were randomly split into three sets: 80% for training, 10% for validation, and 10% for testing. Among the cases included in the test set (513 images), 41 hip joint images (8.0%) were obtained in infants under 4 months.

Data Preparation and Preprocessing

From each hip radiograph, both left and right hip joint images were extracted using the template matching, a technique that allows identification of an area similar to that of a target image. A diagram of the image extraction process is shown in Figure 1. Automatically cropped images were checked and all of the images were satisfactory for diagnosis and included a femoral head, acetabulum, and pubic bones. We collected 5076 labeled images 414 x 414 in size, which included 494 DDH images and 4582 normal images. To avoid overfitting, the training datasets were augmented using operators such as rotation (randomly within ± 15 degrees), translation along the vertical and horizontal axes (randomly within ± 45 pixels), flipping, and scaling within a range (0.9–1.1). To deal with the imbalanced data in our class distribution, training datasets for DDH images were augmented by a factor of 10 times, whereas training datasets for normal images were augmented by a factor of 4 times. As a result, 3920 DDH and 14632 normal images were used for training. Training was performed after resizing the images from 414 x 414 to 128 x 128.

Deep Learning Algorithm

The proposed CNN classifier evaluates the abnormality of hip joint images extracted from hip AP radiographs (18). The architecture of the proposed network is shown in Figure 1. Each of the four brown boxes and one green box consists of multiple layers such as 3 x 3 convolution with a stride of 1, rectified linear unit (19) activation function, and batch normalization (20). Each box is followed by max-pooling with a stride of 2 and a dropout layer with a dropout rate of 25%. The last three bars (yellow and purple) denote the fully connected layers, each followed by a dropout layer. The last fully connected layer in the purple bar was followed by the Softmax layer (21) instead of a dropout layer. The number below each box denotes the number of feature maps (or output units). Cross-entropy (22) was adopted as a loss function for classification. The proposed network was minimized using the Adam

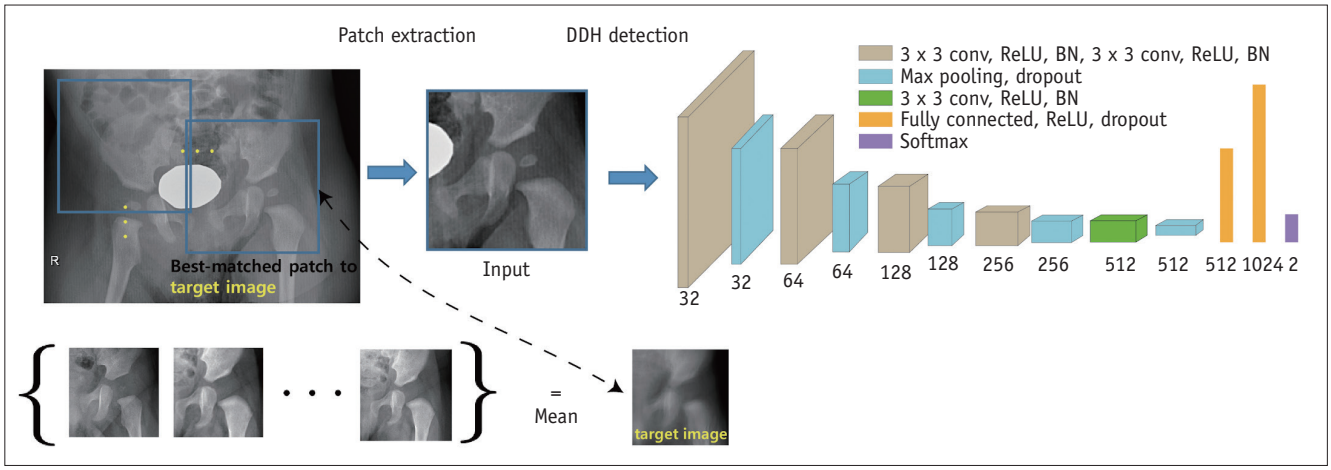


Fig. 1. Diagram of the proposed method for detection of DDH. BN = batch normalization, conv = convolution, DDH = developmental dysplasia of the hip, ReLU = rectified linear unit

https://doi.org/10.3348/kjr.2020.0051

Korean Journal of Radiology
KJR

optimizer (23), and a learning rate of 0.0001, mini-batch size of 16, and 100 epochs were used for training. The training was implemented using Tensorflow (24) on a GPU (NVIDIA, Titan Xp. 12GB) system. We generated heatmaps to determine which portion of the image the deep learning algorithm recognized to differentiate DDHs from normal hips. Heatmaps generated by a gradient-weighted class activation mapping were combined with the corresponding hip joint images (25).

Diagnostic Performance of the Deep Learning Algorithm

The test set was evaluated using the trained CNN algorithm. With an optimal cut-off probability value of 0.001, we constructed 2 x 2 tables and calculated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. The optimal cut-off was determined such that the sum of sensitivity and specificity was maximized.

We also evaluated the diagnostic performance of multiclass classification. Based on the probability values, cases with a probability value of less than 0.001 were classified as normal, and cases with a probability value of more than 0.999 were classed as DDH. The remainders, with a probability value between 0.001 and 0.999, were considered indeterminate. PPV and NPV were calculated with this multiclass classification.

We also constructed a receiver operating characteristics curve (ROC) plot and a precision-recall curve (PRC) plot. The areas under the ROC (AUROC) and PRC (AUPRC) plots were calculated. We generated heatmaps to determine which portion of the image the deep learning algorithm recognized to differentiate DDHs from normal hips.

Human Readout by a Radiologist

Three invited radiologists performed image reviews of the test set. Reviewer 1 had nine years of experience in radiology, including pediatric radiology. Reviewer 2 had five years of experience in radiology without any experience in pediatric radiology. Reviewer 3. had three years of experience in radiology without any experience in pediatric radiology. All three reviewers were asked to independently label patched hip images on a 5-point scale for DDH (1, definitely normal; 2, probably normal; 3, indeterminate; 4, probable DDH; and 5, definite DDH). The reviewers were blinded to the clinical information of each patient, and the reviewers labeled the images using patched unilateral hip images without a contralateral hip image.

Comparison with Human Readers

The sensitivity, specificity, PPV, NPV, and accuracy of the diagnostic performance of the three human readers were also calculated. To calculate the sensitivity and specificity of the diagnosis of DDH, the labels were dichotomized into normal (label 1 and 2) and DDH (label 3, 4, and 5). McNemar's test was conducted to compare the diagnostic performance of the deep learning algorithm with that of each of the three radiologists.

We constructed ROC and PRC plots for human readout. The AUROC and AUPRC were compared between the deep learning algorithm and each of the three different human reviewers. All data were analyzed using MedCalc version 12.7 (MedCalc Software).

Subgroup Analysis by Age Group

To analyze the diagnostic performance of the proposed algorithm and radiologists by patient age group (patients under 4 months versus patients over 4 months of age) a subgroup analysis was performed. We constructed 2 x 2 tables and calculated the sensitivity, specificity, PPV, NPV, and accuracy in two different age groups. We also calculated the AUROC and AUPRC for the two different age groups. An independent ROC comparison analysis was performed to compare the diagnostic performance between the two age groups.

Table 2. Diagnostic Performance of Deep Learning in Diagnosing DDH

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC of ROC Plot	AUC of PRC Plot
Deep learning algorithm	94.0 (83.5–98.7)	98.9 (97.5–99.6)	90.4 (79.7–95.8)	99.4 (98.1–99.8)	0.988 (0.974–0.995)	0.973 (0.937–0.995)
Radiologist 1	96.0 (86.3–99.5)	99.1 (97.8–99.8)	92.3 (81.9–97.0)	99.6 (98.3–99.9)	0.988 (0.974–0.995)	0.958 (0.897–0.919)
Radiologist 2	96.0 (86.3–99.5)	89.0 (85.8–91.7)	48.5 (41.9–55.1)	99.5 (98.1–99.9)	0.959 (0.939–0.975)	0.835 (0.728–0.919)
Radiologist 3	84.0 (70.9–92.8)	85.8 (82.2–88.8)	38.9 (33.0–45.1)	98.0 (96.3–98.9)	0.919 (0.892–0.941)	0.618 (0.468–0.761)

Data in the parentheses are 95% confidence intervals. AUC = area under the curve, NPV = negative predictive value, PPV = positive predictive value, PRC = precision-recall, ROC = receiver operating characteristics

https://doi.org/10.3348/kjr.2020.0051

RESULTS

Table 2 shows the diagnostic performance of the deep learning algorithm and three human reviewers. Figure 2 shows the confusion matrices of the human readers and the developed model at the optimal cut-off probability value.

Evaluation of the Diagnostic Performance of the Deep Learning Algorithm

The sensitivity and specificity of the deep learning algorithm in diagnosing DDH were 98.0% and 98.1%, respectively, while the PPV and NPV were 84.5% and 99.8%, respectively. With a binary classification, among 513 cases in the test set, there was one false negative case (0.2%) and nine false positive cases (1.8%). When the images were labeled 0 for normal cases and 1 for DDHs, most cases (497 of 513; 96.9%) had probability values less than 0.001 or more than 0.999. Sixteen out of 513 cases (3.1%) had a probability value between 0.001 and 0.999, and all false positive cases in the binary classification were included

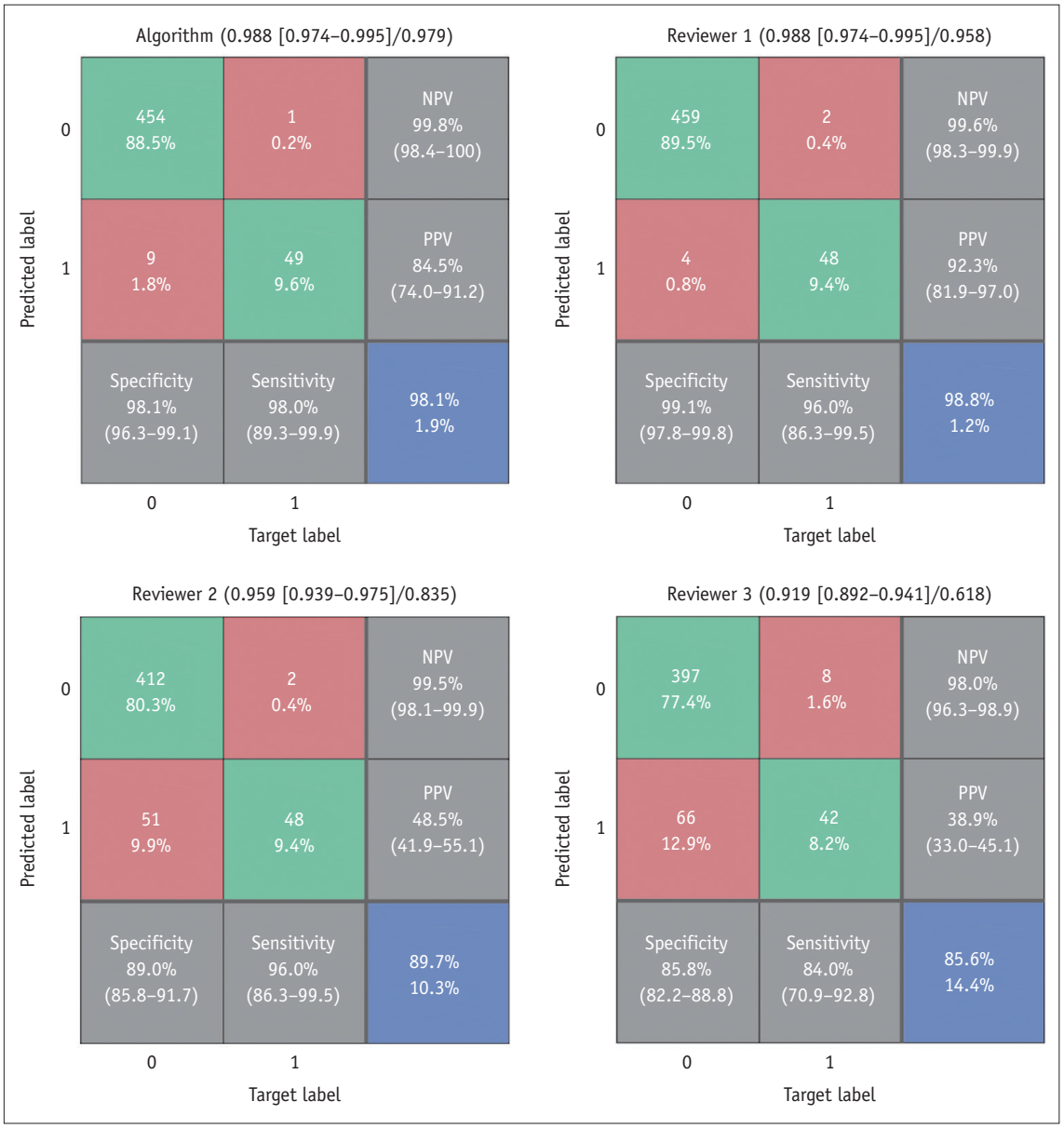


Fig. 2. Confusion matrices of the model at the optimal operating point and for each human reader in the test set. In the confusion matrices, predicted label and target label refer to labeling by algorithm or radiologists and the reference standard, respectively. Label 0 and 1 denote "no DDH" and "DDH," respectively. The figures in the blue box indicate the diagnostic accuracy. NPV = negative predictive value, PPV = positive predictive value

https://doi.org/10.3348/kjr.2020.0051

Korean Journal of Radiology
KJR

in this range. On multiclass classification, PPV and NPV increased to 100% and 99.8%, respectively. The AUROC of the deep learning algorithm was 0.988 in the test set (Fig. 3A), and the AUPRC was 0.973 (Fig. 3B).

In the generated heatmaps, the intensities of the heatmaps were concentrated around the hip joints in most cases, regardless of the presence or absence of ossification centers at the femoral head.

Representative cases from the test set are shown in Figures 4, 5, and 6. Figure 4 shows cases correctly diagnosed as DDH by the proposed deep learning algorithm. Figures 5 and 6 show representative false negative and false positive cases.

Comparison with Human Readers

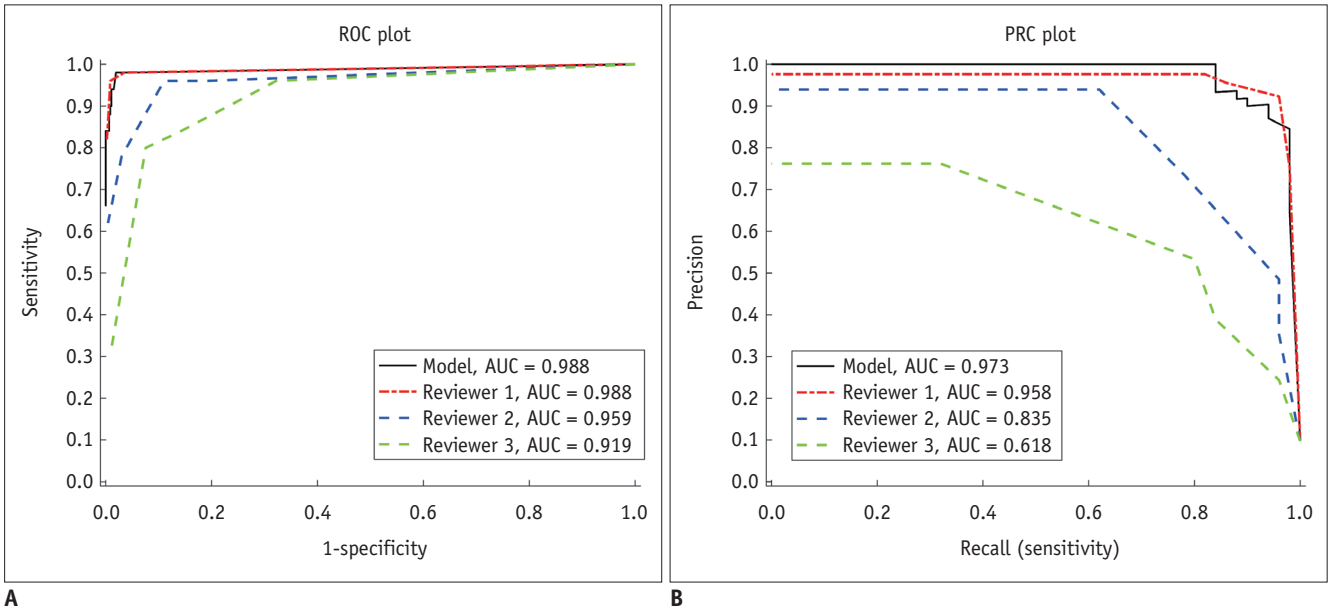
The sensitivity and specificity of the diagnoses made by the radiologists ranged from 84.0 to 96.0% and 85.8 to 99.1%, respectively, while the PPV and NPV ranged from 38.9 to 92.3% and 98.0 to 99.6%, respectively. The accuracy of the three human reviewers ranged from 85.6 to 98.8%.

Based on the McNemar's test, there was no significant difference in the diagnosis of DDH ($p = 0.180$) made by the algorithm and the experienced pediatric radiologist (reviewer 1). However, there were significant differences in the diagnosis of DDH between the inexperienced radiologists (reviewer 2 and 3) compared to the experienced pediatric radiologist ($p < 0.001$) and the algorithm ($p < 0.001$).

Figure 3 shows the AUROC and AUPRC of the proposed algorithm and the three different human reviewers. The AUROCs for the three reviewers were 0.988, 0.959, and 0.919. In the ROC comparison, the AUROC of the proposed algorithm was not significantly different from that of reviewer 1 and reviewer 2 ($p = 0.988$ and $p = 0.147$, respectively). The AUROC curve of the proposed algorithm was significantly higher than that of reviewer 3 ($p < 0.001$). There was also a significant difference in AUROC between reviewer 1 and reviewer 3 ($p = 0.003$). For the three human reviewers, the AUPRC plots were 0.958, 0.835, and 0.618 for radiologists 1, 2, and 3, respectively. In the PRC comparison, the AUPRC of the proposed algorithm was not significantly different from that of reviewer 1 ($p = 0.630$). The AUPRC of the proposed algorithm was significantly higher than those of reviewers 2 and 3 ($p = 0.003$ and $p < 0.001$, respectively).

Subgroup Analysis

Tables 3 and 4 show the diagnostic performance of the proposed algorithm and the radiologists in the two different



A B

Fig. 3. ROC plots and PRC plots with the corresponding areas under the curves for the deep learning algorithm and human reviewers.

A. ROC plots with AUCs for the deep learning algorithm and human reviewers. The area under the ROC of the proposed algorithm was not significantly different from those of reviewer 1 and reviewer 2 ($p = 0.988$ and $p = 0.147$, respectively) but, it was significantly higher than that of reviewer 3 ($p < 0.001$). **B.** PRC plots with AUCs for the deep learning algorithm and human reviewers. The area under the PRC of the proposed algorithm was not significantly different from that of reviewer 1 ($p = 0.630$) but, it was significantly higher than those of reviewers 2 and 3 ($p = 0.003$ and $p < 0.001$, respectively). AUC = area under the curve, PRC = precision-recall curve, ROC = receiver operating characteristics

https://doi.org/10.3348/kjr.2020.0051

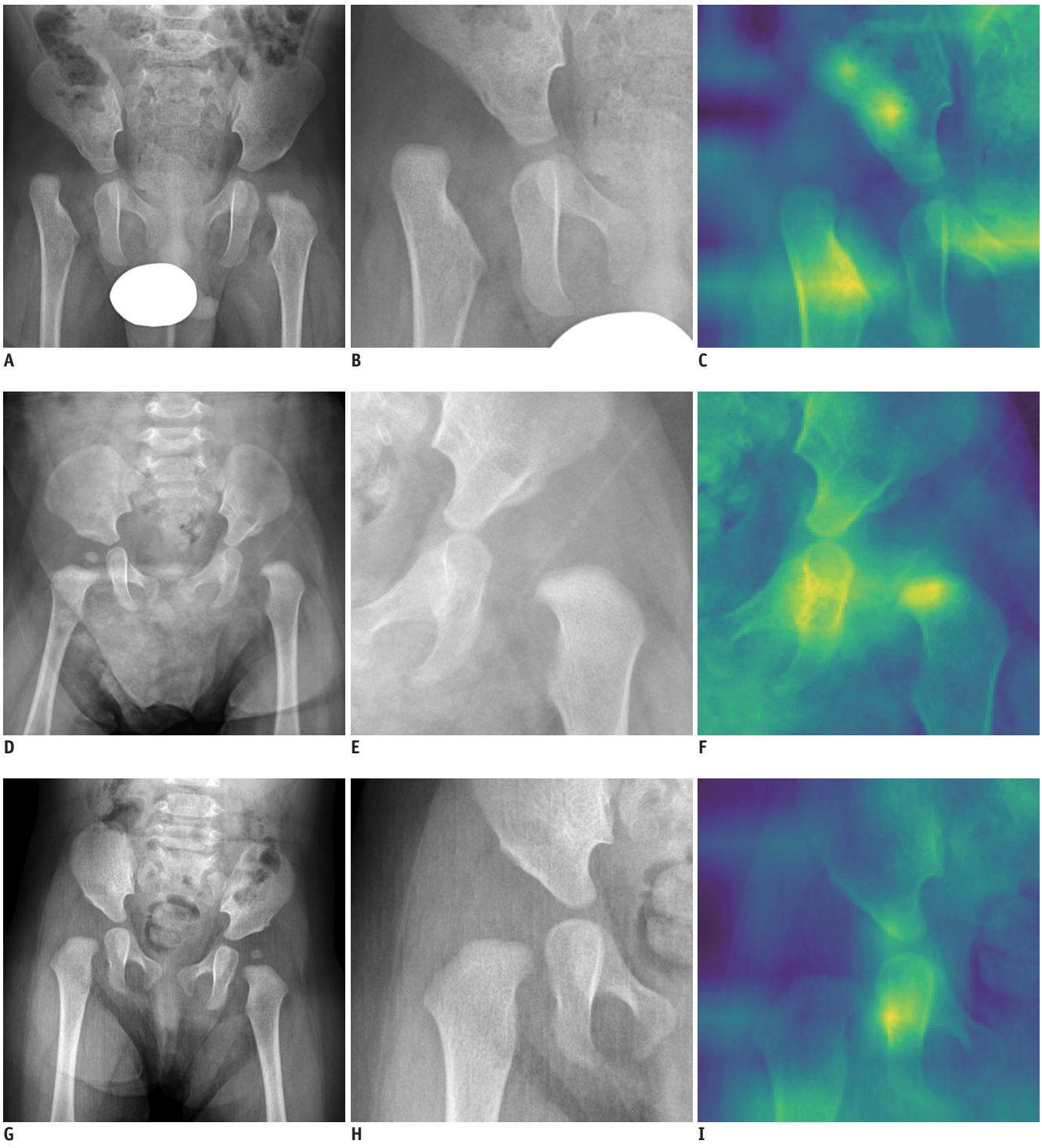


Fig. 4. Representative cases with hip anteroposterior radiographs, cropped image, and heatmap using class activation mapping.
A. A case of DDH on the right side. **B, C.** A cropped image and heatmap showing dislocation of the right hip and increased acetabular angle with disruption of Shenton's line. The probability value derived by the deep learning algorithm was 1.000. Three invited radiologists labeled this case as DDH (label 5). **D.** A case of DDH on the left side. **E, F.** A cropped image and heatmap showing mild lateral subluxation of the left femoral head and increased acetabular angle with disruption of Shenton's line. The probability value derived by the deep learning algorithm was 1.000. Three invited radiologists labeled this case as DDH (label 5). **G.** A case of DDH on the right side. **H, I.** A cropped image and heatmap showing an increased acetabular angle with disruption of Shenton's line. The probability value derived by the deep learning algorithm was 1.000. Two invited radiologists labeled this case as DDH (label 4 from radiologist 1 and label 3 from radiologist 2) and one radiologist (radiologist 3) misclassified this case as normal (label 2).

https://doi.org/10.3348/kjr.2020.0051

Korean Journal of Radiology
KJR

age groups. Independent ROC comparison analysis (p > 0.05) revealed no significant differences in the diagnostic performance of the proposed algorithm nor the three radiologists between the two different age groups (< 4 months of age vs. ≥ 4 months of age).

DISCUSSION

We proposed a deep learning-based method to evaluate DDH from hip joint images automatically extracted from hip AP radiographs. Since the proposed algorithm was trained to evaluate the abnormality of cropped hip joint images instead of entire hip AP radiographs, it can effectively learn the features of DDH from a limited number of hip radiographs. The algorithm was able to provide a fast and accurate diagnosis of DDH. The implementation duration to evaluate DDH from the hip AP view is under a few milliseconds.

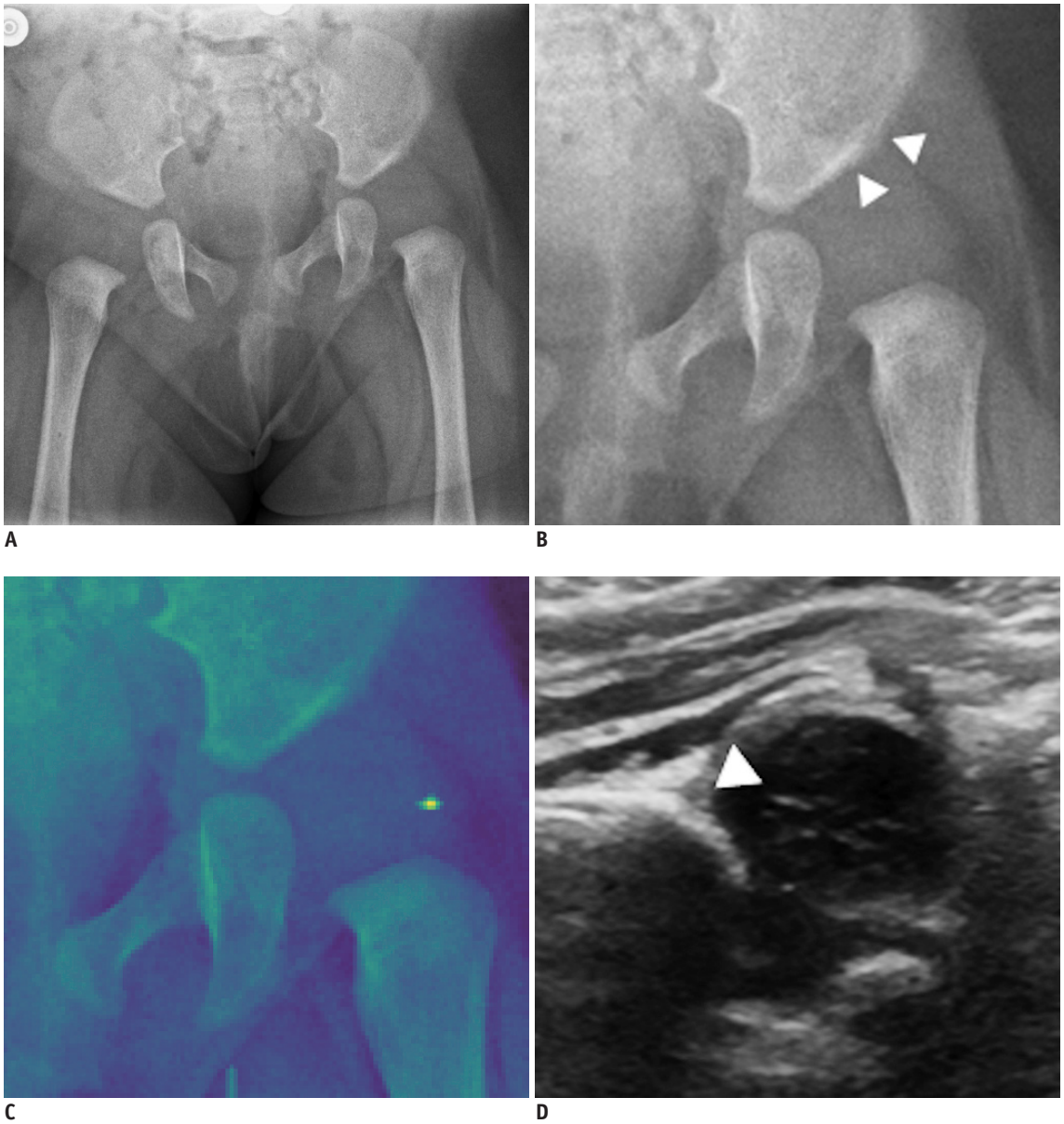


Fig. 5. A false negative case using the deep learning algorithm.
A. A 3-month-old girl with DDH on the left side. There were no specific findings in the physical examination. **B, C.** A cropped image and heatmap of the left hip on conventional hip anteroposterior radiographs showing increased acetabular angle with mild dysplasia of acetabulum (arrow heads). **D.** The left hip demonstrated a reduced alpha angle (53–55°) with a round acetabular edge (arrowhead) in the ultrasound exam. The probability value derived by the deep learning algorithm was 0. Two invited radiologists labeled this case as DDH (label 3 from radiologist 1 and label 5 from radiologist 2) and one radiologist (radiologist 3) misclassified this case as normal (label 1).

https://doi.org/10.3348/kjr.2020.0051

A B C

Fig. 6. A false positive case using the deep learning algorithm.
A. A case with normal hip configuration was misclassified as DDH on the left side. **B, C.** The cropped image showed normal acetabular angle and intact Shenton's line without subluxation of the femoral head. The probability value derived by the deep learning algorithm was 0.890. All three invited radiologists correctly diagnosed this case as normal.

Table 3. Diagnostic Performance of Deep Learning in Diagnosing DDH in Infants under 4 Months of Age

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC of ROC Plot	AUC of PRC Plot
Deep learning algorithm	80.0 (28.4–99.5)	94.4 (81.3–99.3)	66.7 (32.7–89.2)	97.1 (85.5–99.5)	0.886 (0.748–0.964)	0.863 (0.536–1.000)
Radiologist 1	100 (47.8–100.0)	100 (90.3–100.0)	100	100	1.000 (0.914–1.000)	0.994 (1.000–1.000)
Radiologist 2	100 (47.8–100.0)	66.7 (49.0–81.4)	29.4 (20.8–39.8)	100	0.942 (0.821–0.991)	0.720 (0.264–1.000)
Radiologist 3	80.0 (28.4–99.5)	88.9 (73.9–96.9)	50.0 (26.5–73.5)	97.0 (84.7–99.5)	0.836 (0.687–0.933)	0.590 (0.125–1.000)

Data in the parentheses are 95% confidence intervals.

Table 4. Diagnostic Performance of Deep Learning in Diagnosing DDH in Infants over 4 Months of Age

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC of ROC Plot	AUC of PRC Plot
Deep learning algorithm	100 (92.1–100.0)	98.4 (96.7–99.3)	86.5 (75.5–93.1)	100	0.998 (0.989–1.000)	0.986 (0.965–0.998)
Radiologist 1	95.6 (84.9–99.5)	99.1 (97.6–99.7)	91.5 (80.2–96.6)	99.5 (98.2–99.9)	0.986 (0.971–0.995)	0.954 (0.887–0.997)
Radiologist 2	95.6 (84.9–99.5)	90.9 (87.7–93.4)	52.4 (44.8–60.0)	99.5 (98.0–99.9)	0.961 (0.939–0.977)	0.863 (0.761–0.937)
Radiologist 3	84.4 (70.5–93.5)	85.5 (81.8–88.7)	38.0 (32.0–44.3)	98.1 (96.3–99.0)	0.927 (0.900–0.949)	0.628 (0.470–0.767)

Data in the parentheses are 95% confidence intervals.

As summarized in Table 2, in this study, the diagnostic performance of the deep learning algorithm was better than that of the radiologists without experience in pediatric radiology and was comparable to that of an experienced pediatric radiologist. The AUROC curve of the developed deep learning algorithm for the identification of DDH in conventional hip radiography in pediatric patients under 12 months of age was 0.988. As they are more informative than ROC plots, we constructed PRC plots for evaluating binary classifiers on imbalanced datasets (26). The usage and interpretation of an imbalanced dataset are some of the challenges in the field of machine learning, and the dataset used in this study was imbalanced with more negative cases than positive cases. The area under the PRC plot of the deep learning algorithm showed a higher value than those of human reviewers.

In the test set, 10 (1.9%) out of 513 cases were misdiagnosed by the proposed deep learning algorithm. Even though the number of misdiagnosed cases was low in the binary classification, the diagnostic performance can be improved with multiclass classification. Most (9 of 10 cases) misdiagnosed cases had probability values between 0.001 and 0.999. If the probability value obtained by the proposed algorithm is in the range of 0.001 and 0.999, it can be classified as an indeterminate case. The AP radiographs, classified into the indeterminate group,

https://doi.org/10.3348/kjr.2020.0051

Korean Journal of Radiology
KJR

can be manually diagnosed by pediatric experts, or further evaluated using ultrasound exams, if necessary. Such a semi-automatic approach could reduce the misdiagnosis of DDH patients in clinical practice.

In heatmaps, the intensities were mostly concentrated at the hip joint and not at other bone structures or the genital shield. This indicates that the deep learning algorithm we developed recognized the hip joint well in these images and that the configuration of hip joints was used for the determination of hip dysplasia. The intensities of the heatmaps were concentrated at the femoral head, acetabulum, ischium, and pubic bones, regardless of the presence or absence of ossification centers on the femoral head. However, in the case of a false prediction, the intensities were not concentrated in the femoral head or acetabulum.

Whether universal screening for DDH is beneficial remains controversial (27). In most studies, there is no evidence that universal screening decreases late diagnosis or improves clinical outcomes (28-30). However, selective screening may be worthwhile, especially for groups at high-risk of DDH (27, 31). There are three important methods used to diagnose or screen DDH: physical examination, ultrasound, and conventional radiography. Physical examination is an important component and a cornerstone for referral to radiologic examination. However, the diagnostic performance of the physical exam alone is variable and has low accuracy; the sensitivity of the physical examination has been reported to be 13–60% (4). Ultrasound is the study of choice for evaluating the hip in neonates and infants under 6 months of age when the femoral head ossifies (1, 3). One of the advantages of ultrasound over conventional radiography is that it shows the cartilage and soft tissue of the hip (1). However, ultrasound requires a learning curve to achieve the appropriate level of performance, and the results have inter-observer variability (32-34). Conventional hip radiography is more readily available at a lower cost than ultrasound (27). Moreover, conventional radiography has a lower false positive rate than ultrasound between 4 and 6 months of age, which is a watershed period (27). Ultrasonography is particularly susceptible to mild dysplasia, and the likelihood of false positives or false negatives is relatively high (31). After 6 months of age, at the onset of the ossification of the femoral head, radiography is the standard imaging method to evaluate DDH (3). Conventional radiography in 4–6 months old infants is usually appropriate for detecting DDH (35, 36). The proposed algorithm showed excellent diagnostic performance, which was comparable to that of experienced radiologists in patients over 4 months of age. Using the proposed algorithm might allow efficient and accurate screening of DDH without an experienced clinician or radiologist. Patients could subsequently be referred to a specialist. Although there was no significant difference in the diagnostic performance of the proposed algorithm between age groups younger than 4 months and age groups older than 4 months, the assessment of the clinical feasibility of the proposed algorithm in infants younger than 4 months of age was limited by the small number of participants in this age group. In addition, the use of hip radiography is usually not appropriate for the diagnosis or screening of DDH in infants younger than 4 months of age (37). Thus, the use of the proposed algorithm in infants under 4 months old is limited, and ultrasound may be more appropriate for the diagnosis or screening of DDH in this age group.

This study has several limitations. First, human readout did not perfectly reflect the diagnostic performance of DDH in real clinical situations. The developed deep learning algorithm used in this study was designed to analyze a patched unilateral hip without any information about contralateral hip configuration. To compare the diagnostic performance between the developed algorithm and the human reviewers, three reviewers were asked to evaluate the images with unilateral hip information only which is different from an actual clinical setting. Since information from the contralateral hip helps diagnose DDH, the diagnostic performance of human reviewers was relatively under-estimated. Second, the developed algorithm and the radiologists were blinded to the patient's age, which can help the diagnosis of DDH through the assessment of delays in femoral head ossification. The lack of external validation is another limitation of this study, and the internal validation may exaggerate the performance of the developed deep learning algorithm.

In conclusion, the proposed deep learning algorithm provided an accurate diagnosis of DDH on conventional hip radiographs, which was comparable to that of an experienced radiologist.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

https://doi.org/10.3348/kjr.2020.0051

ORCID iDs
Hyoung Suk Park
https://orcid.org/0000-0003-0032-4630
Kiwan Jeon
https://orcid.org/0000-0002-2460-7478
Yeon Jin Cho
https://orcid.org/0000-0001-9820-3030
Se Woo Kim
https://orcid.org/0000-0001-8350-7584
Seul Bi Lee
https://orcid.org/0000-0002-5163-3911
Gayoung Choi
https://orcid.org/0000-0002-2004-5228
Seunghyun Lee
https://orcid.org/0000-0003-1858-0640
Young Hun Choi
https://orcid.org/0000-0002-1842-9062
Jung-Eun Cheon
https://orcid.org/0000-0003-1479-2064
Woo Sun Kim
https://orcid.org/0000-0003-2184-1311
Young Jin Ryu
https://orcid.org/0000-0001-5222-3749
Jae-Yeon Hwang
https://orcid.org/0000-0003-2777-3444

REFERENCES

1. Siegel MJ. *Pediatric sonography*, 4th ed. Philadelphia: Lippincott Williams & Wilkins, 2011:607
2. Kotlarsky P, Haber R, Bialik V, Eidelman M. Developmental dysplasia of the hip: what has changed in the last 20 years? *World J Orthop* 2015;6:886-901
3. Starr V, Ha BY. Imaging update on developmental dysplasia of the hip with the role of MRI. *AJR Am J Roentgenol* 2014;203:1324-1335
4. Price KR, Dove R, Hunter JB. The use of X-ray at 5 months in a selective screening programme for developmental dysplasia of the hip. *J Child Orthop* 2011;5:195-200
5. Coley, BD. *Caffey's pediatric diagnostic imaging*, 13th ed. Philadelphia: Elsevier, 2019:1316-1322
6. Sahin S, Akata E, Sahin O, Tuncay C, Özkan H. A novel computer-based method for measuring the acetabular angle on hip radiographs. *Acta Orthop Traumatol Turc* 2017;51:155-159
7. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* 2015;115:211-252
8. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *AJR Am J Roentgenol* 2017;209:1374-1380
9. Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol* 2017;52:281-287
10. Mannil M, von Spiczak J, Manka R, Alkadhi H. Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: unveiling the invisible. *Invest Radiol* 2018;53:338-343
11. Ciritsis A, Rossi C, Wurnig MC, Phi Van V, Boss A. Intravoxel incoherent motion: model-free determination of tissue type in abdominal organs using machine learning. *Invest Radiol* 2017;52:747-757
12. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol* 2017;52:434-440
13. 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
14. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. *Radiology: Artificial Intelligence* 2019;1:e180015
15. Zonoobi D, Hareendranathan A, Mostofi E, Mabee M, Pasha S, Cobzas D, et al. Developmental hip dysplasia diagnosis at three-dimensional US: a multicenter study. *Radiology* 2018;287:1003-1015
16. Hareendranathan AR, Zonoobi D, Mabee M, Cobzas D, Punithakumar K, Noga M, et al. Toward automatic diagnosis of hip dysplasia from 2D ultrasound. IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017); 2017 Apr 18-21; Melbourne, Australia: IEEE; p. 982-985
17. Li Q, Zhong L, Huang H, Liu H, Qin Y, Wang Y, et al. Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of Sharp's angle on standardized anteroposterior pelvic radiographs. *Medicine (Baltimore)* 2019;98:e18500
18. Krizhevsky A, Sutskever I, Hinton GE. Advances in neural information processing systems. In: Krizhevsky A, Sutskever I, Hinton GE, eds. *Imagenet classification with deep convolutional neural networks*. San Diego: Neural Information Processing Systems Foundation, 2012:1097-1105
19. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML-10); 2010 Jun 21-24; Haifa, Israel; ICML; p. 807-814
20. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint* 2015;arXiv:1502.03167
21. Bishop CM. *Pattern recognition and machine learning*. New York: Springer, 2006:203-204

22. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep learning*. Cambridge: MIT Press, 2016:131-132
23. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint* 2014;arXiv:1412.6980
24. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint* 2016;arXiv:1603.04467
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22-29; Venice, Italy: ICCV; p. 618-626
26. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432
27. Shaw BA, Segal LS; Section on Orthopaedics. Evaluation and referral for developmental dysplasia of the hip in infants. *Pediatrics* 2016;138:e20163107
28. Shipman SA, Helfand M, Moyer VA, Yawn BP. Screening for developmental dysplasia of the hip: a systematic literature review for the US preventive services task force. *Pediatrics* 2006;117:e557-e576
29. Holen KJ, Tegnander A, Bredland T, Johansen OJ, Saether OD, Eik-Nes SH, et al. Universal or selective screening of the neonatal hip using ultrasound? A prospective, randomised trial of 15,529 newborn infants. *J Bone Joint Surg Br* 2002;84:886-890
30. Woolacott NF, Puhan MA, Steurer J, Kleijnen J. Ultrasonography in screening for developmental dysplasia of the hip in newborns: systematic review. *BMJ* 2005;330:1413
31. Karmazyn BK, Gunderman RB, Coley BD, Blatt ER, Bulas D, Fordham L, et al. ACR appropriateness criteria on developmental dysplasia of the hip--child. *J Am Coll Radiol* 2009;6:551-557
32. Peterlein CD, Schüttler KF, Lakemeier S, Timmesfeld N, Görg C, Fuchs-Winkelmann S, et al. Reproducibility of different screening classifications in ultrasonography of the newborn hip. *BMC Pediatr* 2010;10:98
33. Hell AK, Becker JC, Rühmann O, Lewinski Gv, Lazovic D. Inter- und intraindividuelle Messabweichungen in der Säuglingshüftsonografie nach Graf. *Z Orthop Unfall* 2008;146:624-629
34. Quader N, Schaeffer EK, Hodgson AJ, Abugharbieh R, Mulpuri K. A systematic review and meta-analysis on the reproducibility of ultrasound-based metrics for assessing developmental dysplasia of the hip. *J Pediatr Orthop* 2018;38:e305-e311
35. Tudor A, Sestan B, Rakovac I, Schnurrer Luke-Vrbanić T, Prpić T, Rubinić D, et al. The rational strategies for detecting developmental dysplasia of the hip at the age of 4-6 months old infants: a prospective study. *Coll Antropol* 2007;31:475-481
36. Garvey M, Donoghue VB, Gorman WA, O'Brien N, Murphy JF. Radiographic screening at four months of infants at risk for congenital hip dislocation. *J Bone Joint Surg Br* 1992;74:704-707
37. Nguyen JC, Dorfman SR, Rigsby CK, Iyer RS, Alazraki AL, Anupindi SA, et al. ACR appropriateness criteria® developmental dysplasia of the hip-child. *J Am Coll Radiol* 2019;16:S94-S103

 https://doi.org/10.3348/kjr.2020.0051