Skeletal Radiology (2022) 51:401–406
https://doi.org/10.1007/s00256-021-03880-y

SCIENTIFIC ARTICLE

# Can AI distinguish a bone radiograph from photos of flowers or cars? Evaluation of bone age deep learning model on inappropriate data inputs

**Paul H. Yi[1] · Anirudh Arun[2] · Nima Hafezi-Nejad[2] · Garry Choy[3] · Haris I. Sair[2] · Ferdinand K. Hui[2] · Jan Fritz[4]**

Received: 25 March 2021 / Revised: 15 July 2021 / Accepted: 25 July 2021 / Published online: 5 August 2021


## Abstract

**Objective** To evaluate the behavior of a publicly available deep convolutional neural network (DCNN) bone age algorithm when presented with inappropriate data inputs in both radiological and non-radiological domains.

**Methods** We evaluated a publicly available DCNN-based bone age application. The DCNN was trained on 12,612 pediatric hand radiographs and won the 2017 RSNA Pediatric Bone Age Challenge (concordance of 0.991 with radiologist ground-truth). We used the application to analyze 50 left-hand radiographs (appropriate data inputs) and seven classes of inappropriate data inputs in radiological (i.e., chest radiographs) and non-radiological (i.e., image of street numbers) domains. For each image, we noted if (1) the application distinguished between appropriate and inappropriate data inputs and (2) inference time per image. Mean inference times were compared using ANOVA.

**Results** The 16Bit Bone Age application calculated bone age for all pediatric hand radiographs with mean inference time of 1.1 s. The application did not distinguish between pediatric hand radiographs and inappropriate image types, including radiological and non-radiological domains. The application inappropriately calculated bone age for all inappropriate image types, with mean inference time of 1.1 s for all categories ($p = 1$).

**Conclusion** A publicly available DCNN-based bone age application failed to distinguish between appropriate and inappropriate data inputs and calculated bone age for inappropriate images. The awareness of inappropriate outputs based on inappropriate DCNN input is important if tasks such as bone age determination are automated, emphasizing the need for appropriate oversight at the data input and verification stage to avoid unrecognized erroneous results.

**Keywords** Artificial intelligence · Deep learning · Bone age · Quality · Safety

## Introduction

Deep learning has been met with enthusiasm and excitement by radiologists, as convolutional neural networks (DCNNs) have demonstrated the ability to perform radiologic tasks approaching or exceeding the levels of performance of expert radiologists in a variety of tasks [1–4]. Specifically, in musculoskeletal radiology, DCNNs have demonstrated wide-ranging utility for interpretation of radiographs for orthopedic trauma and implants [5–7], magnetic resonance imaging (MRI) for evaluation of internal derangement of the knee [8–10], and computed tomography (CT) for segmentation of pelvic muscles, fat, and bone [11].

Amidst the excitement over deep learning, concerns have been raised about the potential pitfalls and limitations of deep learning in radiology [12, 13]. For example, the variable generalizability of radiographic DCNN analysis has

✉ Jan Fritz
jan.fritz@nyulangone.org

1 University of Maryland Intelligent Imaging (UMII) Center, Department of Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

2 The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD, USA

3 Department of Radiology, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA

4 Department of Radiology, New York University Grossman School of Medicine, New York, NY, USA

been demonstrated for external datasets that have not been used in the initial model training [13], possibly due to differences in image acquisition or disease prevalence—a problem known as domain shift.

A similar problem with high clinical relevance is scenarios when DCNNs are provided data input of inappropriate format or type. For example, what would happen if a DCNN trained to predict bone age on left-hand radiographs was inadvertently provided a knee or wrist radiograph for analysis? The general assumption and expectation of radiologists may be that algorithms in place verify the appropriateness of image data input. Indeed, ideally, a DCNN would reject inappropriate data inputs and refuse to calculate a bone age; lest clinically inaccurate and nonsensical results are generated for a patient. However, the behavior of DCNNs when provided inappropriate data inputs has not been specifically evaluated for musculoskeletal applications.

Therefore, the purpose of our study was to evaluate the behavior of a publicly available deep convolutional neural network (DCNN) bone age algorithm when presented with inappropriate data inputs in both the radiological and non-radiological domains.

## Methods

### Description of 16Bit Bone Age Application

We evaluated a publicly available bone age application (https://www.16bit.ai/bone-age) [14], which analyzes pediatric left-hand posterior–anterior (PA) radiographs and automatically returns the predicted bone age. The application is described on the website as based on a DCNN trained on 12,612 pediatric left PA hand radiographs from two USA hospitals and was part of the Radiological Society of North America (RSNA) 2017 Pediatric Bone Age Machine Learning Challenge [15]. We note that the website does not explicitly instruct users to upload only left-hand radiographs or that the application has a check for appropriate vs. inappropriate data input. The DCNN uses both image and sex to predict bone age. It resizes input images to 500 × 500 pixels before analysis by an ensemble of DCNNs based on the Inception-V3 architecture [15]. This DCNN was the winner of the 2017 RSNA Pediatric Bone Age Challenge, achieving a concordance correlation coefficient (CCC) of 0.991 with radiologist-determined ground-truth and mean absolute difference of 4.265 months [15]. The web browser-based application is publicly available and accepts any standard image file via drag-and-drop upload or smartphone camera capture.

### Evaluation of DCNN behavior with appropriate and inappropriate data inputs

We collected 50 images from the public domain from each of the following categories: (1) pediatric left-hand PA radiographs [15], (2) pediatric lateral elbow radiographs [16], (3) adult chest radiographs [17], (4) street numbers (Google search), (5) human faces [18], (6) flowers [19], (7) cars (Google search), and (8) houses (Google search). The pediatric left-hand PA radiograph was defined as the "appropriate" data input, as this is what the DCNN was trained on for evaluation of bone age. All other image types were defined as "inappropriate" data inputs.

We evaluated each image using the bone age application and recorded if (1) the application distinguished between appropriate and data inputs (i.e., rejection or warning of inappropriate data inputs) and (2) the inference time for each image analysis (Fig. 1).

### Robustness study

Although the 16Bit Bone Age application was trained specifically to identify bone age on pediatric left-hand radiographs, it may be possible that the application generalizes to unseen data types of other types of bones. For the two non-hand radiograph categories evaluated in the inappropriate data input study (normal chest radiographs and elbow radiographs), we compared the predicted bone ages with the chronologic ages of the patients. For the elbow radiograph evaluation, we used 50 pediatric lateral elbow radiographs comprised of ages ranging from 1 year old to 18 years old (mean 10.3 years, standard deviation 5.2 years). For the chest radiograph evaluation, we used a random subset of 50 images from a publicly available pediatric chest radiograph dataset comprised of patients aged 1 to 5 years old [20].

### Statistical analysis

Statistical evaluations and computations were performed using VassarStats (http://vassarstats.net/). Descriptive statistics were used to summarize inference times using range, means, and standard deviation (SD). Mean inference times were compared between different image groups using analysis of variance (ANOVA). For the robustness study, we compared pediatric elbow radiograph predicted bone age with chronological age using paired *t*-tests. Because the pediatric chest radiographs did not provide image-level ages, we evaluated robustness by calculating the percentage of images with a predicted bone age of

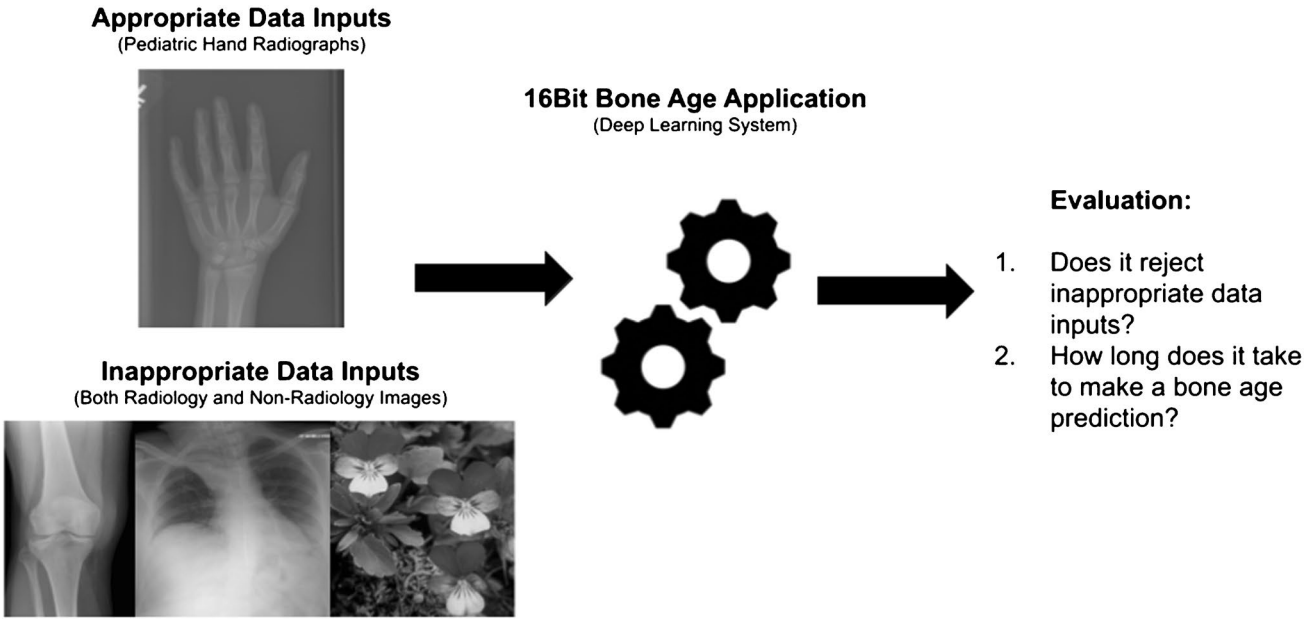


**Fig. 1** Study design for evaluation of bone age application

greater than 5 years (all of the pediatric chest radiographs were between 1 and 5 years old). $p$-values of $< 0.05$ were considered statistically significant.

## Results

### Appropriate vs. inappropriate data inputs

The bone age application appropriately calculated bone ages for all 50 pediatric hand radiographs with an average inference time of 1.1 s (range 1.0 to 1.2 s; SD 0.07 s).

The application did not distinguish between pediatric hand radiographs and inappropriate image types and calculated a bone age for all inappropriate images across all categories, including the radiographs and non-radiographs (Fig. 2). For the inappropriate radiographs, mean inference times were 1.1 s for pediatric elbow radiographs (range 1.0 to 1.6 s; SD 0.1 s) and 1.1 s for adult chest radiographs (range 1.0 to 1.3 s; SD 0.07 s). For the non-radiograph photographs, mean inference times were 1.1 s for street numbers (range 1.0 to 1.8 s; SD 0.1 s), 1.1 s for flowers (range 1.0 to 1.3 s; SD 0.09 s), 1.1 s for cars (range 1.0 to 1.1 s; SD 0.01 s), 1.0 s for faces (range 1.0 to 1.2 s; SD 0.04 s), and 1.0 s for houses (range 1.0 to 1.8 s; SD 0.1 s).

There was no significant difference in inference time between the groups ($p = 1$), which had an overall mean inference time of 1.1 s (range 1.0 to 1.8 s; SD 0.08 s).

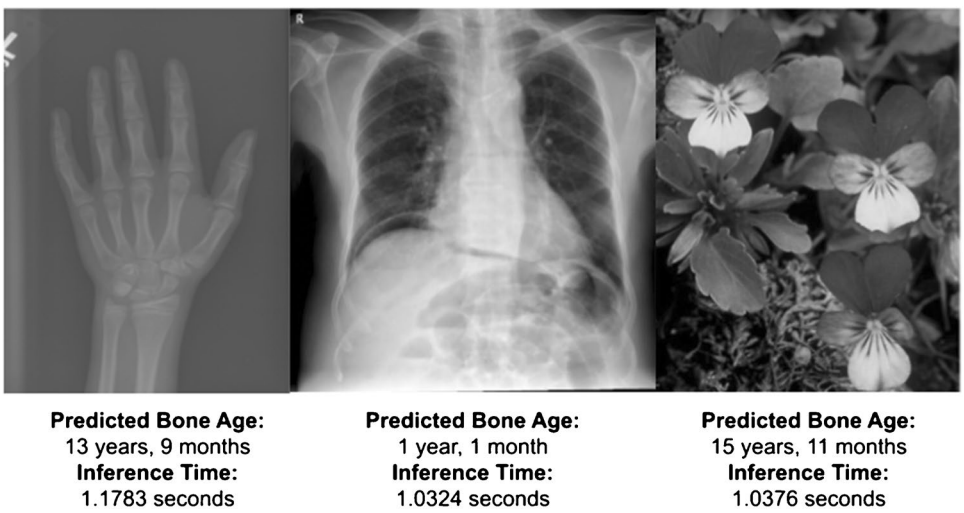


**Fig. 2** Bone age calculations of mean inference times of appropriate image data input (left-hand pediatric radiograph, left image) and inappropriate image data input (adult chest radiograph, center image; image with flower, right image)

### Robustness study results

For the pediatric elbow radiographs, the mean predicted bone age was 5.2 years (range 1.1 to 9.3 years, SD 2.0 years) compared to mean chronological age of 10.3 years (range 1 to 18 years, SD 5.2 years); this difference was statistically significant ($p < 0.0001$).

For the pediatric chest radiographs, all of which were between 1 and 5 years chronological age; the mean predicted bone age was 11.2 years (range 3.5 to 15.3 years, SD 2.8 years), with 49 of 50 (98%) being > 5 years.

## Discussion

Given the excitement created by deep learning in musculoskeletal radiology and potential implications for patient care, we wanted to evaluate potential limitations of DCNNs, particularly when presented with data of inappropriate input or type. A publicly available online bone age DCNN application appropriately and quickly calculated bone age for pediatric hand radiographs. However, the application failed to distinguish between appropriate and inappropriate data inputs, calculating bone age for all images across multiple categories. The inference time did not differ between hand radiographs and inappropriate images, suggesting that the algorithm does not hesitate or struggle with calculating a bone age on images that are overtly inappropriate inputs. Additionally, we found that there was poor generalizability of the bone age application towards prediction of age on pediatric chest and elbow radiographs. Our study is intended to raise awareness of erroneous DCNN results based on inappropriate data input and emphasize the need for appropriate oversight at the data input and verification stage to avoid unrecognized erroneous results.

DCNNs have been described as "savants," [21] in that they are exceptionally good at performing a specific task, such as calculation of bone age on left-hand PA radiographs, but are unable to perform other tasks. Indeed, we found that while the 16Bit Bone Age application calculated pediatric bone age in < 2 s, which is much faster than human radiologists, the application could not differentiate pediatric hand radiographs from radiographs of different anatomical areas (elbows and chest) and photos of flowers, cars, houses, and human faces, a task that a human radiologist would be able to perform easily.

A recent report titled "Can Your AI Differentiate Cats from Covid-19?" [22] similarly showed that DCNNs trained to detect COVID-19 pneumonia on chest radiographs cannot reject inappropriate data inputs, such as a photograph of a cat, and will unknowingly "diagnose" the photo as having COVID-19 with high confidence. Altogether, these findings highlight a major pitfall in the use of DCNNs for medical image analysis in their inability to reject grossly inappropriate data inputs, which raises concerns for ensuring the safe use of DCNNs in medical practice.

We also found that the 16Bit Bone Age application calculated bone age at a mean time of 1.1 s per image, regardless of the type of image, indicating that the DCNNs did not struggle or hesitate when evaluating the inappropriate images. Although this may not be intuitive, these findings make sense when considering how DCNNs "see" images. DCNNs view images as arrays of pixel intensities (i.e., rows and columns of numbers), from which they identify patterns of increasing complexity (i.e., features). Accordingly, when given a data input, DCNNs can easily analyze any image without hesitation or difficulty. This is in contrast to humans who would not be able to calculate or predict a bone age for these inappropriate image inputs, and again, demonstrates the limitation of DCNNs if they are provided inappropriate images.

Although the 16Bit bone age application was trained specifically to evaluate hand radiographs, one might wonder if the application could generalize to other bones. After all, recent work has shown that deep learning models can predict bone age accurately using only index finger radiographs for training [23], showcasing the flexible nature of deep learning models. To evaluate this, we performed two robustness studies on two types of pediatric radiographs, namely, chest radiographs and elbow radiographs. In both cases, we found large differences between the predicted bone ages and chronological ages, suggesting that the bone age application trained for left-hand radiographs only does not generalize to other types of bone radiographs. However, we acknowledge that this comparison is limited as we did not have access to corresponding bone ages on contemporaneously acquired left-hand radiographs for these patients and used chronological age as a proxy. Although this is a limitation, we do highlight that chronological age has been shown to correlate well with actual bone age in normal populations [24].

These findings may dampen the enthusiasm raised by deep learning for transforming radiology. Based on our findings, DCNNs are not ready to be deployed in a completely automated manner without safety checks for appropriate data input and output. For example, a labeling error of an image of a different anatomic area or laterality could result in clinically significant errors if DCNNs lack functions to identify such errors.

On the other hand, these limitations may also point to a path towards human–machine collaboration and synergy. While DCNN-based algorithms are unable to do anything other than their specific, trained task, they can execute their tasks at a superhuman pace. Therefore, if these models can be deployed to require human interaction for quality control for appropriate data input (e.g., *human-in-the-loop* approach), a synergistic outcome could be achieved for

improving efficiency and accuracy of radiologic tasks, such as bone age prediction.

Potential solutions to the problem of rejecting inappropriate data inputs that do not require human interaction or intervention have been explored by posing the question as a task of "out-of-distribution" detection. In this type of task, algorithms are designed to identify images that are dissimilar to the images that they were trained on ("out-of-distribution") using a variety of advanced computing methods, including autoencoders [25] and Bayesian neural networks [26]. Although out-of-distribution detection algorithms have been described in the medical literature [27], these techniques are relatively experimental and not widely implemented. One other solution may be to train DCNNs to identify the specific appropriate image input type before subsequent analysis, such as radiographic view [28–30] or anatomic area [31, 32], although this requires concerted efforts to curate labeled images for both appropriate and inappropriate image input types.

Our study has several limitations. First, we evaluated only a single bone age application for its behavior when presented with different data inputs, and our findings may not apply to other commercial applications or models. Second, we evaluated only eight image input types (one appropriate, seven inappropriate), which does not cover the entire spectrum of possible image input types. However, we intentionally chose a wide range of image types, from radiographs of inappropriate anatomic areas to photographs of humans and flowers, to cover a spectrum of different image types. Additionally, because DCNNs "see" images as arrays of numbers, it is unlikely that other image types would yield results different from the ones in our study. Third, there is another version of the 16Bit Bone Age application that is not free-for-use, but listed on a commercial radiology artificial intelligence (AI) marketplace [33] and marketed under the name "Physis," which may have an out-of-distribution detection mechanism. Nevertheless, the 16Bit Bone Age application that we evaluated is publicly available and listed on the 16Bit website as the RSNA Pediatric Bone Age Challenge winner, which is why we chose to evaluate it.

In conclusion, DCNN-based applications for bone age prediction may not have the ability to reject grossly invalid data and may deliver "clinical" results even when given inapplicable datasets. Such inappropriate analysis of images that a DCNN is not designed to analyze could pose a risk factor for clinical errors if tasks such as bone age determination are automated without appropriate oversight at the data input and verification stage. Clinically used algorithms' analyses of invalid data could lead to inappropriate conclusions and potentially deleterious decision-making. Accordingly, the ability of AI algorithms to identify inappropriate data inputs is important for patient safety, as well as adoption of AI, which is dependent on building trust in these algorithms. Thus, radiologists should be aware of these potential pitfalls and consider implementing safeguards against inappropriate and potentially misleading data analysis. Our findings suggest the need for the future development of "common sense" safeguards for AI algorithms and emphasizes the need for validation and regulation by entities like the Food and Drug Administration (FDA).

## Declarations

**Conflict of interest** The authors declare no competing interests.

## References


1. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15: e1002686. https://doi.org/10.1371/journal.pmed.1002686.
2. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–82. https://doi.org/10.1148/radiol.2017162326.
3. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology. 2017;287: 170236. https://doi.org/10.1148/radiol.2017170236.
4. Yala A, Schuster T, Miles R, Barzilay R, Lehman CA. Deep learning model to triage screening mammograms: a simulation study. Radiology. 2019;293:38–46. https://doi.org/10.1148/radiol.2019182908.
5. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol. 2018;48:239.
6. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci. 2018;115:11591–6.
7. Yi PH, Kim TK, Wei J, Li X, Hager GD, Sair HI, et al. Automated detection and classification of shoulder arthroplasty models using deep learning. Skeletal Radiol. 2020. https://doi.org/10.1007/s00256-020-03463-3.
8. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med. 2018;15:e1002699.
9. Fayad LM, Parekh VS, de Castro Luna R, Ko CC, Tank D, Fritz J, et al. A deep learning system for synthetic knee magnetic resonance imaging: is artificial intelligence-based fat-suppressed imaging feasible? Invest Radiol. 2020;56:357.
10. Germann C, Marbach G, Civardi F, Fucentese SF, Fritz J, Sutter R, et al. Deep convolutional neural network-based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee MRI cohorts with different pulse sequence protocols and 1.5-T and 3-T magnetic field strengths. Invest Radiol. 2021;55:499–506.
11. Hemke R, Buckless CG, Tsao A, Wang B, Torriani M. Deep learning for automated segmentation of pelvic muscles, fat, and

bone from CT studies for body composition assessment. Skeletal Radiol. 2019. https://doi.org/10.1007/s00256-019-03289-8.
12. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20:405.
13. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med. 2018;15:e1002683.
14. 16 Bit [Internet]. https://www.16bit.ai/bone-age. Accessed 8 Nov 2019
15. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA pediatric bone age machine learning challenge. Radiology. 2019;290:498–503.
16. UCSD Musculoskeletal Radiology [Internet]. http://bonepit.com/Normalforage/Normalforageindex.htm
17. Lakhani P, Gray DL, Pett CR, Nagy P, Shih G. Hello World deep learning in medical imaging. J Digit Imaging. 2018;31:283–9. https://doi.org/10.1007/s10278-018-0079-6.
18. Huang GB, Mattar M, Berg T, Learned-Miller E. Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. Tech Note Univ Massachusetts Amherst [Internet]. 2008; https://hal.inria.fr/inria-00321923/
19. Nilsback M-E, Zisserman A. 17 category flower dataset [Internet]. https://www.robots.ox.ac.uk/~vgg/data/flowers/17/index.html
20 Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122-1131.e9.
21. Deep learning: Einstein or Savant?—International Institute for Analytics [Internet]. https://www.iianalytics.com/blog/2017/6/15/deep-learning-einstein-or-savant
22. Mallick A, Dwivedi C, Kailkhura B, Joshi G, Han TY-J. Can your AI differentiate cats from Covid-19? Sample efficient uncertainty estimation for deep learning safety. ICML 2020 Work Uncertain Robustness Deep Learn [Internet]. 2020; https://api.deepai.org/publication-download-pdf/can-your-ai-differentiate-cats-from-covid-19-sample-efficient-uncertainty-estimation-for-deep-learning-safety
23. Reddy NE, Rayan JC, Annapragada AV, Mahmood NF, Scheslinger AE, et al. Bone age determination using only the index finger: a novel approach using a convolutional neural network compared with human radiologists. Pediatr Radiol. 2020;50:516–23.
24. Pan I, Baird GL, Mutasa S, Merck D, Ruzal-Shapiro C, Swenson DW, Ayyala RS, et al. Rethinking Greulich and Pyle: a deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs. Radiol Artif Intell. 2020;2:e190198.
25. Daxberger E, Hernández-Lobato JM. Bayesian variational autoencoders for unsupervised out-of-distribution detection. arXiv [Internet]. 2019; http://arxiv.org/abs/1912.05651
26. Mitros J, Mac Namee B. On the validity of Bayesian neural networks for uncertainty estimation. arXiv [Internet]. 2019; http://arxiv.org/abs/1912.01530
27. Gao L, Wu S. Response score of deep learning for out-of-distribution sample detection of medical images. J Biomed Inform. 2020;107:103442.
28. Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. High-throughput classification of radiographs using deep convolutional neural networks. J Digit Imaging. 2017;30:95–101. https://doi.org/10.1007/s10278-016-9914-9.
29. Kim TKTK, Yi PHPH, Wei J, Shin JWJW, Hager G, Hui FKFK, et al. Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs. J Digit Imaging. 2019;32:925–30.
30. Yi PH, Lin A, Wei J, Yu AC, Sair HI, Hui FK, et al. Deep-learning-based semantic labeling for 2D mammography and comparison of complexity for machine learning tasks. J Digit Imaging. 2019;32:565–70.
31. Cheng PM, Malhi HS. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J Digit Imaging. 2017;30:234–43.
32. Yi PH, Kim TK, Wei J, Shin J, Hui FK, Sair HI, et al. Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. Pediatr Radiol. 2019;49:1066–70.
33. physis by 16 Bit Inc. | Nuance Production [Internet]. https://subscriber.aimarketplace.nuance.com/apps/226933#!overview