scientific reports

OPEN

Critical evaluation of deep neural networks for wrist fracture detection

Abu Mohammed Raisuddin[1,5✉], Elias Vaattovaara[1,2,5], Mika Nevalainen[1,2], Marko Nikki[2], Elina Järvenpää[2], Kaisa Makkonen[2], Pekka Pinola[1,2], Tuula Palsio[1,4], Arttu Niemensivu[1], Osmo Tervonen[1,2] & Aleksei Tiulpin[1,2,3]

Wrist Fracture is the most common type of fracture with a high incidence rate. Conventional radiography (i.e. X-ray imaging) is used for wrist fracture detection routinely, but occasionally fracture delineation poses issues and an additional confirmation by computed tomography (CT) is needed for diagnosis. Recent advances in the field of Deep Learning (DL), a subfield of Artificial Intelligence (AI), have shown that wrist fracture detection can be automated using Convolutional Neural Networks. However, previous studies did not pay close attention to the difficult cases which can only be confirmed via CT imaging. In this study, we have developed and analyzed a state-of-the-art DL-based pipeline for wrist (distal radius) fracture detection—DeepWrist, and evaluated it against one general population test set, and one challenging test set comprising only cases requiring confirmation by CT. Our results reveal that a typical state-of-the-art approach, such as DeepWrist, while having a near-perfect performance on the general independent test set, has a substantially lower performance on the challenging test set—average precision of 0.99 (0.99–0.99) versus 0.64 (0.46–0.83), respectively. Similarly, the area under the ROC curve was of 0.99 (0.98–0.99) versus 0.84 (0.72–0.93), respectively. Our findings highlight the importance of a meticulous analysis of DL-based models before clinical use, and unearth the need for more challenging settings for testing medical AI systems.

Wrist fractures are the most common type of fractures[1] and typically indicate the fractures in the distal radius or ulna bones. The prevalence of wrist fractures is high, and according to the recent data, approximately 18 million hand and wrist fracture incidents occurred worldwide[2]. Population-wise, 162 cases of the distal radius or ulna fractures occur on average per 100, 000 inhabitants per year in the United States[3]. In the northern countries the incident rate is even higher. For example, in Finland the number of incidents is 258 per 100, 000 inhabitants annually[4].

Various types of treatments are available depending on the fracture's severity. Conservative casting and splinting are used for simple, acute, and nondisplaced fractures[5]. Besides, a large number of patients are treated with operative treatment (surgery)[6]. As an example from an economical point of view, Dutch Injury Surveillance System analysis shows that annual expenditure for wrist and hand injuries in the Netherlands is over 540, 000, 000 €[7]. In addition to the financial burden, wrist fractures significantly reduce the quality of life. A study on Australian older adults shows that the loss in Health Related Quality of Life due to wrist fracture takes around 18 months for recovery[8]. Due to the aforementioned facts, wrist fractures pose a significant healthcare burden worldwide.

Conventional radiography (X-ray imaging) is used routinely as the first-line tool for wrist fractures diagnosis[9]. All plain radiographs are taken in certain projection views: lateral (LAT), posteroanterior (PA), anteroposterior (AP), or oblique. For most of the cases, X-ray imaging is sufficient to keep the high quality of care, and it emits substantially less radiation to the patients than volumetric modalities do, such as computed tomography (CT)[10].

Wrist X-ray images are usually taken in an emergency room and visually inspected by the attending physician, or if available, by a radiologist. Diagnostic errors, especially misdiagnosis of fractures, are common issues in the haste of the emergency setting[11]. Generally, the diagnostic performance of a physician can be affected by multiple factors, such as work overload, fatigue, and lack of experience[12,13]. Many image interpretation errors could be avoided in the emergency room if the radiographs would be always instantly read by a radiologist or analyzed automatically providing support in the decision-making process.

[1]University of Oulu, Oulu, Finland. [2]Oulu University Hospital, Oulu, Finland. [3]Ailean Technologies Oy, Oulu, Finland. [4]City of Oulu, Oulu, Finland. [5]These authors contributed equally: Abu Mohammed Raisuddin and Elias Vaattovaara. ✉email: abu.raisuddin@oulu.fi
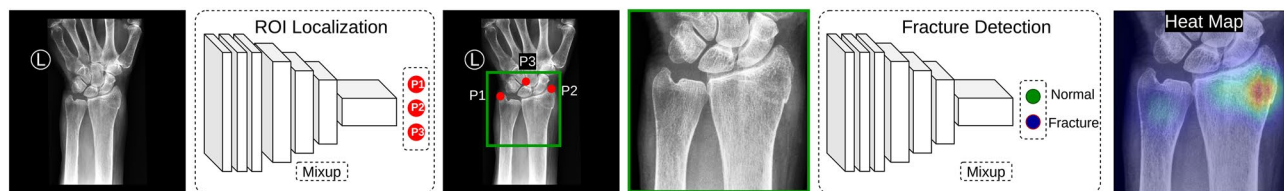
Figure 1. DeepWrist pipeline. From left to right: a wrist radiograph is passed to ROI (region of interest) localization block, which predicts three landmark points (P1, P2 and P3). Subsequently, these landmarks are used to crop an ROI image from the original radiograph. Finally, we utilize a fracture detection block which predicts whether the radiograph is normal or has a fracture. In addition to the prediction, we generate an explanation of the decision using a GradCAM technique.

During recent years, Deep Learning (DL) has been widely applied in the realm of musculoskeletal radiology. In the domain of automatic fracture detection, DL has been used in application to radiographs on various body parts: ankle[14], hip[15–17], humerus[18], and wrist[13,19–21]. The wrist fracture detection performances in these studies were reported to be relatively high—the Area Under the Receiver Operator Characteristics curve (AUROC) was of above or equal to 0.80 on a test set. However, all these studies lack the validation of the methods on difficult fractures, which are challenging to diagnose without CT, and can only be diagnosed by a very experienced professional. We note that in clinical practice, CT is applied rather seldom, mostly in the cases where a fracture is clinically obvious or heavily suspected, but the radiographs do not show any signs of it[22]. Therefore, having a reliable diagnostic process for these rare cases directly impacts patient care, and if one wants to establish a fully automatic assessment of wrist images in a clinical setting, a special attention needs to be paid to the challenging cases.

Generally, rare clinical cases are rather unaddressed as a separate stratum in the state-of-the-art medical imaging studies as a result their effects go unnoticed due to hidden stratification issue[23]. Recent studies on hidden stratification show that performance drop can be significantly high for an unaddressed stratum[23,24]. Uncertain wrist cases that needed CT imaging form such stratum, and are of primary interest in this work.

In this paper, we highlight the issue of hidden stratification in the realm of distal radius wrist fracture detection. In the sequel, we use the term wrist fractures for compactness, implying the fractures of the distal radius bone. The main contributions of our work can be summarized as follows:

- We develop an open-source wrist fracture diagnosis method—DeepWrist (see Fig. 1). This method is a two-stage pipeline, which utilizes anatomical landmark localization and image classification models, and reveals the local decision explanation using a GradCAM approach[25]. We show that on a general independent test set, this method yields high performance.
- For the first time in the realm of automatic wrist fracture detection, we show that a DL model trained on general population cases does not perform well on the difficult cases, which needed a CT imaging for diagnosis.
- We show that despite a prior belief on domain shift between the general population and difficult cases, state-of-the-art techniques for estimating uncertainty in DL, such as Deep Ensembles[26] are barely able to discriminate between these two sets of images.
- Finally, we compare the performance of our model and human physicians with various experience levels to investigate whether the aforementioned discrepancy is natural for them.

Materials and methods

Data. *Overview.* Our study leveraged three datasets, where one was used for training, and the other two for testing. These datasets consisted of referrals, PA and LAT images, and radiology reports. All the data were extracted from the Oulu University Hospital's (OUH) Picture Archiving and Communication System (PACS) and the Radiology Information System. We used pseudonymization to keep patients' identities protected. The project was approved by the Ethics Committee of Northern Ostrobothnia Hospital District (decision number: 126/2014), and the patients' informed consent requirement was waived due to the retrospective nature of this study. All methods of this research were performed in accordance with the Declaration of Helsinki.

Training dataset. To create the training set, we biased our data selection keeping the ratio of fractures 50%. Initially, our training dataset included 1000 cases with distal radius fractures. Subsequently, images, which had artifacts (reasons—non-diagnostic quality or implants) were removed leaving 953 distal radius fracture cases. In total, 1946 wrist studies (3873 PA and LAT images) were used in our training set. All the cases in this training set were the general fracture cases and it did not contain any challenging cases, for which an additional CT imaging was required.

We annotated the training images based on the radiology reports: every image was visually inspected, and an existing radiology report was then manually labeled as *normal* or *fracture* (only distal radius fractures are considered) by a medical student who received basic training in diagnostic radiology. Thereby, we assigned the same label to both PA and LAT images. Detailed label and projection view distributions of all datasets are shown in Table 1. The details on sex and age distribution can be found in Supplementary Section S1.

2

Dataset	# Cases	# Fracture cases	# Normal cases	View	# Radiographs	# Fracture radiographs	# Normal radiographs
Training set	1946	953	993	PA	1962	954	1008
				LAT	1911	946	965
Test set #1	207	129	78	PA	207	129	78
				LAT	207	129	78
Test set #2	105	20	85	PA	105	20	85
				LAT	105	20	85

Table 1. Datasets used in this study.

Landmark localization data. As our pipeline leveraged two parts—Region of Interest (ROI) localization block and fracture detection block, we had to perform the manual annotation for the ROI localization block. We annotated 3820 out of 3873 wrist radiographs from the training dataset with the anatomical landmarks (see Fig. 1) using the VGG Image Annotator (VIA)[27]. Here, 3056 radiographs were used for training, and 764 radiographs were used for measuring the accuracy of the ROI localization block. An analysis on intra-rater variability is discussed in Supplementary Section S2.

General population test set. The test set #1 or the general population test set initially consisted of 210 patient cases which were collected randomly from the Oulu University Hospital's PACS and did not require additional CT imaging for diagnosis. Three cases out of 210 had implants, thus were excluded from the final analysis leaving 207 cases with an equal number of PA and LAT radiographs where 129 of the cases were annotated as fracture and 78 as normal (see Table 1 for details). All images in this set were acquired from the emergency department. We utilized an annotation strategy similar to that of training dataset and used radiology reports to create the initial labels for these general population data. The reports in this dataset were created by the total of 16 radiology residents with work experience ranging from 16 to 53 months (median—35 months).

Besides the annotations produced from the radiology reports, all the radiographs in this dataset were re-read by two board-certified radiologists independently without the knowledge of initial radiology reports. Radiologists were specifically asked to give a yes-or-no answer whether there is a fracture in the distal radius or not to keep the labeling in line with the training data. In case of disagreement (3 cases), a consensus decision was made. Consensus-based labels were used as the ground truth for the test set #1. Beside the annotations from the board-certified radiologists, we included the annotations by other practitioners: the radiographs in this set were independently read by 2 primary care physicians with 3 and 4 years of clinical experience.

Challenging test set. The test set #2 or the test set of challenging cases had a total of 105 patient cases. These data were deemed hard for diagnosis from X-ray images, thus the presence of fracture was determined by CT imaging. Among the extracted 105 cases, 85 cases were found normal and 20 were found to have distal radius fracture from the radiology report (see Table 1 for details). The annotations derived from the CT report were used as the ground truth for this dataset. The two board-certified radiologists and two primary care physicians, who annotated the test set #1, also annotated the test set #2.

DeepWrist pipeline. *Overview and experimental setup.* Figure 1 shows a graphical illustration of our approach. The whole pipeline comprises two parts—ROI localization block by landmark localization and fracture detection block. The former part is based on the KNEEL method by Tiulpin et al.[28], and it was trained to localize three anatomical landmarks (see Fig. 1). Using these landmarks, we cropped the ROI to include the part of the image that contains the distal radius bone. The latter part of our pipeline is a CNN based classifier, pre-trained on ImageNet dataset, and subsequently trained on our training dataset.

All the experiments were conducted using PyTorch[29] with a PytorchLightning wrapper[30] for executing training and inference processes. SOLT[31] library (version 0.1.8) was used for data augmentation. We ran all our experiments using a single Nvidia Geforce RTX 2080 Ti GPU. For each view (PA and LAT), separate ROI localization and fracture detection blocks were trained.

Except for the final testing, all the experiments were conducted using cross-validation (CV) to determine the best hyperparameters. The classifiers' thresholds and the temperature hyperparameters of Deep Ensemble were maximized in an out-of-fold cross-validation setting. Supplementary Table S3 shows the settings used for hyperparameters selection. We used a fivefold CV to train the ROI localization block. To train the fracture detection block we also used the similar procedure. Here, we used the patient ID for group splitting to ensure that training and validation datasets did not intersect.

Pre-processing and augmentation. All the data were pre-processed before passing them through any of the blocks. After reading each radiograph, we used the global contrast normalization with initial clipping between the 5th and 99th intensity percentiles.

Due to the images being of large size, we used bi-linear interpolation, and re-scaled the images to a lower pixel-spacing. Specifically, we used the target pixel spacing of 0.27 mm for the PA view and 0.35 mm for the LAT view, to train the fracture detection block. For the ROI localization block, pixel spacing was not fixed, rather it was dependent on the expected size of input to the block in pixels which was 256×256.

For training, we used heavy data augmentations. We applied cutout[32], jittering, random color padding on a particular side, downscaling, flipping, rotation, shearing, padding, salt and pepper, blur, noise and gamma correction for the ROI localization block. For the fracture detection block we used similar augmentations. More details about the data augmentations are shown in the source code.

During inference, we did not use any augmentation for ROI localization block but we used Test-Time Augmentation (TTA) for fracture detection block to improve the performance. For the TTA, we used gray scale to color conversion, flipping and five-crop on both flipped and unflipped images.

ROI localization block. This module of the pipeline is a landmark localizer, which learns to identify three major key points in the wrist radiographs. After localization, we crop the ROI using the detected landmark points.

For PA view the landmarks were placed at the top of distal ulna, top of distal radius and the center of the wrist (see Fig. 1). For the LAT view, the landmarks were two distinguishable points on two sides of the top part of radio-ulna, and the center of wrist.

In short, our landmark localizer uses an hourglass network[33], with a soft-argmax layer to predict the landmark coordinates directly. We utilized the existing method and the open-source codebase from the KNEEL method[28]. To train this model, we used a Stochastic Gradient Descent(SGD) optimizer with a learning rate of $1e − 1$ with no momentum and a batch size of 24. The localization pipeline was trained for 300 epochs with a learning rate drop at 150th, 200th and 250th epochs by a factor of 10.

Since, ROI localization is a crucial part of our fracture detection pipeline, one has to ensure that the absence of failures on the datasets. Thus, to regularize the training, we used *mixup*[34]. Such strategy had shown to improve adversarial robustness, and improve generalization of Deep Neural Networks[34]. We also observed similar effects in our cross-validation experiments.

Briefly, *mixup* aims to convexify the training set by creating interpolated samples:

$$x_{mix} = \lambda x_1 + (1 − \lambda)x_2 \tag{1}$$

$$y_{mix} = \lambda y_1 + (1 − \lambda)y_2 \tag{2}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$. Empirically, we found that training with $\alpha = 0.4$ works the best with our data, and as recommended by the authors of KNEEL[28], we did not use the weight decay.

As mentioned earlier, we used the generated landmarks to create the ROI for fracture detection block. For that, we computed the center of mass of the landmark coordinates and added a top padding to the obtained coordinate point to calculate the center for cropping the ROI from the original DICOM image. In our experiments, the PA ROI had a size of 70 mm × 70 mm with a 15 mm top padding and LAT ROI has a size of 90 mm × 90 mm with a 20 mm top padding. These values for cropping the ROI were chosen empirically based on the visual inspection on CV. As mentioned earlier, we had 5 models from fivefold CV. During inference, we formed a 5-model ensemble, averaged the predicted landmarks coordinates from five models and used them as the predicted landmark coordinates of the block. After the ROI localization block was trained, we applied it to generate the ROIs for the whole training dataset to train the fracture detection block.

Fracture detection block. We used a SeresNet50[35] model pre-trained on ImageNet[36] dataset for fracture detection block. We added a *dropout* layer with 50% probability before the *fully connected* layer of the network (randomly reset to predict two classes, contrary to 1000 classes in ImageNet). The remaining part of the model architecture was taken from the work by Hu et al.[35]. Similar to the ROI localization block, we used an SGD optimizer with a learning rate of $1e − 1$, batch size of 32 and a weight decay of $1e − 4$. We did not use any momentum for the training. The model was trained for 300 epochs with a learning rate drop at 150th, 200th, 250th epochs by a factor of 10. For the first 10 epochs, we only trained the classifier part of the SeresNet50 and after that for the rest of the remaining epochs, we trained the full network.

Multi-view ensembling. To leverage the radiographs from both PA and LAT views, we created an Ensemble, which computed the average of the underlying blocks' predictions (5 models from each CV fold). We note that in the case of fracture detection block, we applied TTA to each individual item in the Ensemble before averaging. The whole prediction strategy is visualized in Supplementary Figure S1.

Evaluation of distribution shift. To rule out the possibility that the hard cases have a distribution shift from the distribution of general population cases which is negatively affecting the performance of fracture detection for hard cases, we conducted experiments using Deep Ensemble[26] approach to detect hard cases as out-of-distribution (OOD) data. We trained the above described fracture detection block, but without transfer learning, to ensure diversity in coverage of parameters' posterior distribution modes. For details about this experiment, see Supplementary Section S4.

Results interpretation. *Decision explanation via GradCAM.* To interpret the predictions of the fracture detection block, our pipeline produces a heat map focusing the part of radiograph, which positively affected the outcome of the model. For this, we used GradCAM[25] technique. In brief, GradCAM computes a weighted sum of the feature maps in the penultimate layer of the neural network. The weights for this summation are obtained by back-propagating the decision of choice (fracture in our case).

Metrics and statistical analyses. We used multiple metrics to interpret the results. In our notation, positive cases indicate fractures and negative indicates—their absence. We assessed the performance of the fracture detection block as the total performance of our pipeline. The main metrics were the AUROC and Area Under Precision-Recall Curve (AUPR). Using these two metrics in conjunction is important, as the label distribution of test set #2 is imbalanced (see Table 1). Apart from the metrics common in the machine learning literature, we also reported the metrics utilized by medical community—Sensitivity (also known as Recall or True Positive Rate), Specificity (also known as Selectivity or True Negative Rate), Precision (also known as Positive Predictive Value), F_1 Score and Balanced Accuracy. Beside these metrics, we also used the Cohen's quadratic kappa (κ) for the inter-rater analysis. Kappa measures the agreement between two raters for the same cases.

As the aforementioned metrics are not suitable to assess the anatomical landmarks prediction quality, we used the Euclidean distance between predicted landmark coordinates and ground truth. Here, we defined different precision thresholds and calculated the percentage of correctly classified key points within 1 mm, 1.5 mm etc.

To analyse the statistical significance, we used the stratified bootstrapping to compute the Confidence Interval (CI) of all the statistical metrics with 5000 iterations. We also used a logistic regression to assess the added value of our model to the confounding factors, such as age and sex on the test sets. We used *statsmodels*[37] for calculating the $p-value$.

Results

Localization of anatomical landmarks.
We analyzed the predictive performance of the landmark localizer as the predictive performance of ROI localization block. The landmarks are coarsely annotated for its training set as we do not need fine grained landmark coordinates for cutting a good ROI image. As a result the accuracy of landmark localizer (ROI localization block) is also evaluated with relaxation and tolerance. This block scores 0.70 (0.67−0.73) recall at 3 mm precision, 0.88 (0.86−0.80) recall at 4 mm precision and 0.96 (0.95−0.97) recall at 5 mm precision on the holdout test set. We found this accuracy sufficient for ROI localization due to the subsequent cropping strategy which was also confirmed by the visual inspection on the out-of-fold validation data. Therefore we did not aim to further improve this block of our method. A more detailed evaluation of the ROI localization is presented in Supplementary Section S2.

Fracture detection.
Cross-validation and threshold optimization. The out-of-fold validation accuracy was 0.95 (0.93−0.97) and 0.98 (0.97−0.99) for the PA and LAT views respectively. After training the models, we used validation predictions from all folds to identify the cut-off or threshold values. We found that F_1 Score was maximized when the probability threshold was of 0.41 for the PA view and of 0.58 for LAT view. For the final ensemble we used the average of these two thresholds (0.5).

To decide whether the mixup[34] technique would be used for this block, we trained this block with mixup ($\alpha = 0.7$) and without mixup and evaluated them on the out-of-fold validation data. We found out that mixup slightly improves the performance on out-of-fold validation data therefore we kept the mixup technique for this block.

Inter-rater agreement. Besides fracture detection performance we also analyzed inter-rater agreement among the human raters. We used Cohen's Quadratic Kappa for this purpose. The details of the inter-rater analyses can be found in Supplementary Section S3.

For test set #1, radiologist 2 had the most agreement with the consensus-based ground truth. Unlike radiologists, primary care physicians are not well trained on how to detect fracture accurately from plain radiographs which is reflected from the κ values of two primary care physicians (0.76 and 0.88) and two radiologists (0.98 and 0.99) with respect to consensus. The radiology resident's κ (0.93) lay between the primary care physicians (PCP1 and PCP2) and the radiologists (R1 and R2). It is notable that primary care physicians disagreed between themselves the most. In fact, the PCP1 and PCP2 had the worst agreement ($\kappa = 0.67$) among all the raters. For test set #2, all the raters have low agreement with the ground truth from the CT compared to the similar analyses in the test set #1.

Test set #1: general population test set.
For the test set #1 the AUROCs were 0.98 (0.97−0.99), 0.98 (0.97−0.99) and 0.99 (0.98−0.99) for PA view, LAT view and Ensemble respectively (see Table 2). In Fig. 2, we visualize the ROC curve for the test set #1 along with the performance of radiology resident, two radiologists and two primary care physicians. In terms of sensitivity and specificity, the radiologists and the resident performed better than our pipeline. But the primary care physicians had mixed scores: PCP1 scored a lower specificity but a higher sensitivity and PCP2 scored a higher specificity but a lower sensitivity than our pipeline's corresponding score (see Fig. 2 and Table 3 for details).

The AUPR on test set #1 is 0.99 for all views and the Ensemble. In Fig. 3, we visualize the Precision-Recall curve along with the performance of other raters where the radiologists and resident performed better than the pipeline in terms of precision and recall. But like before, the primary care physicians had mixed scores: PCP1 scored a higher recall but a lower precision, and PCP2 scored a lower recall but a higher precision than our pipeline's corresponding score (see Fig. 3 and Table 3 for details). Both AUROC and AUPR indicate that DeepWrist is a near-perfect classifier.

Test set #2: hard cases.
The AUROCs for the hard test set or test set #2 were of 0.81 (0.69−0.91), 0.83 (0.70−0.93) and 0.84 (0.72−0.93) for the PA view, LAT view and Ensemble respectively (see Table 2). In subplot (b) of Fig. 2, we show the performance of DeepWrist in terms of the sensitivity and specificity. Evidently

Dataset	Model	AUROC (95% CI)	AUPR (95% CI)	Sensitivity, recall, TPR (95% CI)	Specificity, selectivity, TNR (95% CI)	Precision PPV (95% CI)	F_1 score (95% CI)	BA (95% CI)
Test set #1	PA	0.98 (0.97–0.99)	0.99 (0.98–0.99)	0.97 (0.94–1.00)	0.88 (0.80–0.94)	0.93 (0.89–0.96)	0.95 (0.92–0.97)	0.93 (0.89–0.96)
	LAT	0.98 (0.97–0.99)	0.99 (0.98–0.99)	0.97 (0.94–1.00)	0.91 (0.84–0.96)	0.94 (0.91–0.97)	0.96 (0.93–0.98)	0.94 (0.90–0.97)
	Ensemble	0.99 (0.98–0.99)	0.99 (0.99–0.99)	0.97 (0.94–1.00)	0.87 (0.79–0.93)	0.92 (0.88–0.96)	0.95 (0.92–0.97)	0.92 (0.88–0.96)
Test set #2	PA	0.81 (0.69–0.91)	0.61 (0.44–0.80)	0.50 (0.30–0.70)	0.89 (0.82–0.95)	0.52 (0.33–0.73)	0.51 (0.31–0.68)	0.69 (0.58–0.80)
	LAT	0.83 (0.70–0.93)	0.57 (0.41–0.80)	0.50 (0.30–0.70)	0.94 (0.88–0.98)	0.66 (0.46–0.90)	0.57 (0.36–0.75)	0.72 (0.60–0.83)
	Ensemble	0.84 (0.72–0.93)	0.64 (0.46–0.83)	0.60 (0.40–0.80)	0.92 (0.87–0.97)	0.66 (0.48–0.87)	0.63 (0.44–0.80)	0.76 (0.65–0.87)

Table 2. DeepWrist's performance on trivial cases (test set #1) and hard cases (test set #2). Here, AUROC is area under the receiver operating characteristic, AUPR is the area under precision recall curve, CI is confidence interval, 95% CI is shown in parentheses, TPR is true positive rate, TNR is true negative rate and PPV is positive predictive value and BA stands for balanced accuracy.
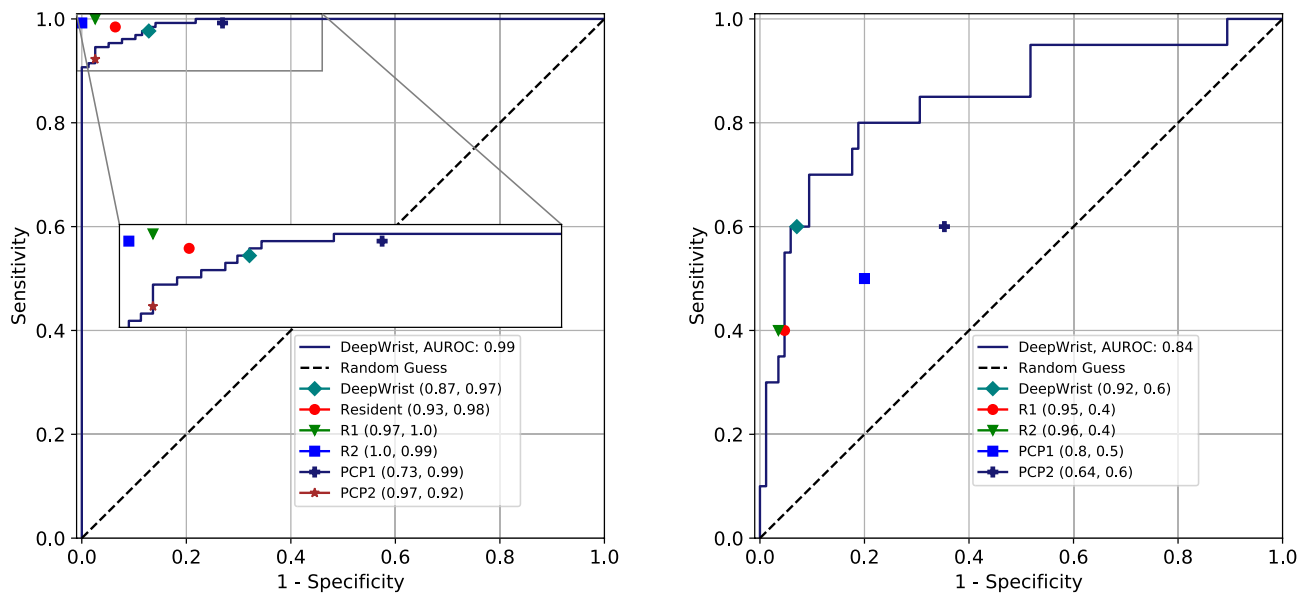
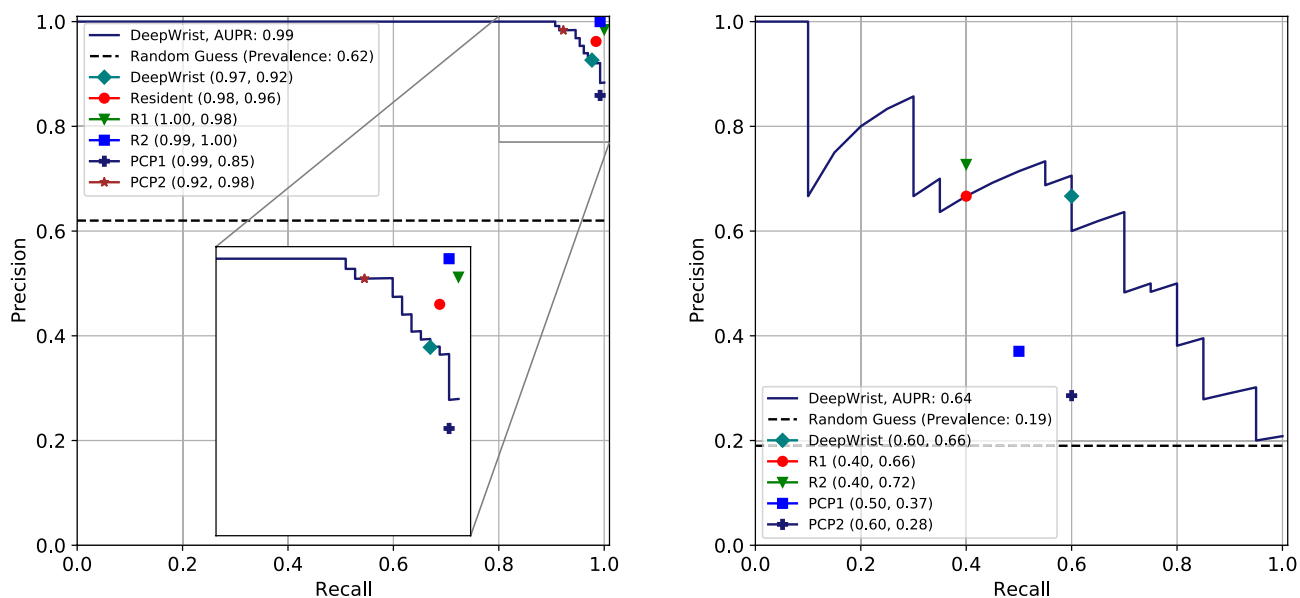(a) AUROC on test set #1 or trivial cases   (b) AUROC on test set #2 or hard cases

Figure 2. (**a**) AUROC performance of DeepWrist on test set #1 compared to a radiology resident, two radiologists (R1 & R2), and two primary care physicians (PCP1, PCP2), (**b**) AUROC performance of DeepWrist on test set #2.

the shown results are substantially lower compared to the results of test set #1. The PR curve (Fig. 3) also indicates the same findings. We note that human raters also showed the drop in performance (see Table 4).

Analysis of pitfalls. To analyse the pitfalls, we evaluated the impact of confounding factors (age and sex) using Logistic Regression to the predictions of our model. We found that for the test set #1, age and sex are significantly associated with the outcome ($p < 0.05$) but our model had also significant contributions ($p < 0.001$). However, for the test set #2 (hard cases), the p-value for DeepWrist was 0.43, indicating that our method did not contribute to the outcome more than the confounding factors did.

In addition to the statistical analyses, we visualized the GradCAM-based heatmaps (Fig. 4). For the True Positive cases in both datasets, on subplots (a)–(d), DeepWrist identified the correct zones, where distal radius fractures appear. The subplots (e) and (f) show that the model could not see these fractures, as they were not visually present in the image.

Is there a distribution shift between general and hard cases? Our results, show that for a 9 model Deep Ensemble, AUROC for OOD detection using predictive variance as uncertainty is 0.62 (0.56−0.68), which indicates that the hard cases are not well detected as OOD with reliable performance. For ensembles with a lower number of models, we observed a similar or worse performance. Further insights are shown in Supplementary Section S4.

(a) AUPR on test set #1 or trivial cases

(b) AUPR on test set #2 or hard cases

Figure 3. (**a**) AUPR performance on the test set #1 for DeepWrist, radiology resident, two radiologists (R1 & R2), and two primary care physicians (PCP1 & PCP2), (**b**) AUPR performance of DeepWrist and other graders on the test set #2. The plot highlights the drop in performance for both—human raters of different expertise, and our method.

	Radiology resident	Radiologist 1	Radiologist 2	Primary care Physician 1	Primary care Physician 2	DeepWrist
Sensitivity (95% CI)	0.98 (0.96–1.00)	1.00 (1.00–1.00)	0.99 (0.97–1.00)	0.99 (0.97–1.00)	0.92 (0.87–0.96)	0.97 (0.94–1.00)
Specificity (95% CI)	0.93 (0.87–0.98)	0.97 (0.93–1.00)	1.00 (1.00–1.00)	0.73 (0.62–0.82)	0.97 (0.93–1.00)	0.87 (0.79–0.93)
Precision (95% CI)	0.96 (0.92–0.99)	0.98 (0.96–1.00)	1.00 (1.00–1.00)	0.85 (0.81–0.90)	0.98 (0.95–1.00)	0.92 (0.88–0.96)
F_1 score (95% CI)	0.97 (0.95–0.99)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.92 (0.89–0.94)	0.95 (0.92–0.97)	0.95 (0.92–0.97)
BA (95% CI)	0.96 (0.92–0.98)	0.98 (0.96–1.00)	0.99 (0.98–1.00)	0.86 (0.81–0.91)	0.94 (0.91–0.97)	0.92 (0.88–0.96)

Table 3. Performance of 5 readers and DeepWrist on trivial cases (test set #1). 95% confidence intervals (CI) are shown in parentheses. BA stands for balanced accuracy.

	Radiologist 1	Radiologist 2	Primary care Physician 1	Primary care Physician 2	DeepWrist
Sensitivity (95% CI)	0.40 (0.20–0.60)	0.40 (0.20–0.60)	0.50 (0.30–0.70)	0.60 (0.40–0.80)	0.60 (0.40–0.80)
Specificity (95% CI)	0.95 (0.90–0.98)	0.96 (0.91–1.00)	0.80 (0.71–0.88)	0.64 (0.54–0.74)	0.92 (0.87–0.97)
Precision (95% CI)	0.66 (0.41–0.91)	0.72 (0.50–1.00)	0.37 (0.23–0.52)	0.28 (0.19–0.38)	0.66 (0.48–0.87)
F_1 score (95% CI)	0.50 (0.27–0.70)	0.51 (0.28–0.70)	0.42 (0.25–0.58)	0.38 (0.25–0.50)	0.63 (0.44–0.80)
BA (95% CI)	0.67 (0.57–0.78)	0.68 (0.57–0.79)	0.65 (0.53–0.76)	0.62 (0.50–0.73)	0.76 (0.65–0.87)

Table 4. Performance of 4 readers and DeepWrist on hard cases (test set #2). 95% confidence intervals (CI) are shown in parentheses. BA stands for balanced accuracy.

Discussion

In this study, we followed the recent works and trained a CNN-based pipeline for distal radius wrist fracture detection. Compared to recent studies on wrist fracture detection, on the general population dataset, our pipeline scored a better AUROC than others[13,19–21]. The important aim of our study, was to bring up the general issues of safety and robustness of AI in medical imaging to the attention of the reader. Earlier, this issue has been high-lighted by Oakden-Rayner et al.[23], and some results were shown on musculoskeletal image data from some of the public datasets. Our work is different from the prior art, as we investigated the problem on a real clinical dataset.
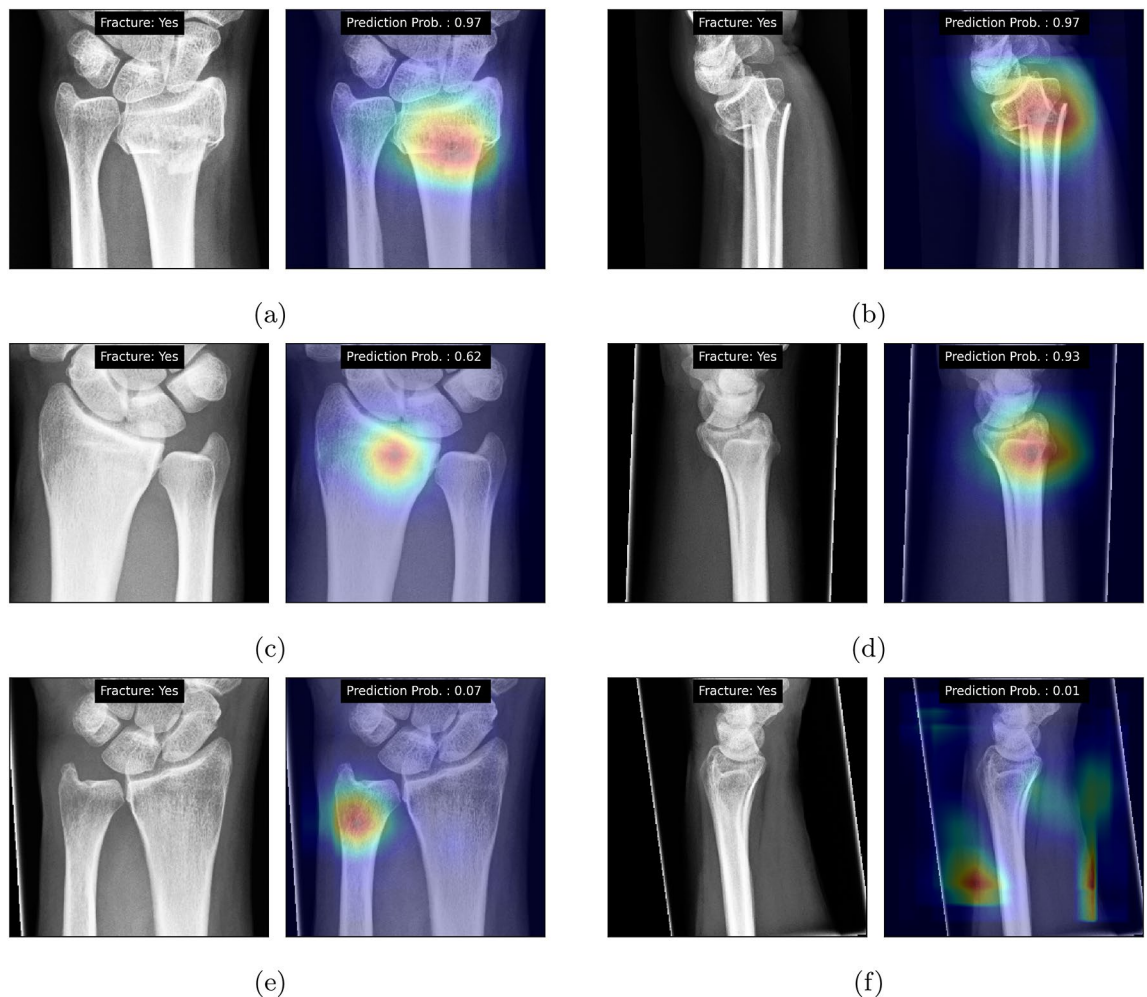
Figure 4. GradCAM-based heatmaps for the developed model. Each sub-figure shows input image and its label on the left side and GradCAM and prediction probability on the right side. (**a**) and (**b**) show PA and LAT views of a true positive case from test set #1 where the fracture is easily visible. (**c**) and (**d**) show both views of a true positive from test set #2 where the fracture is hardly visible and (**e**) and (**f**) show both views of a false negative case from test set #2 where the pipeline predicts them as normal.

A novelty of our work is that we used the validation on challenging cases to expose the safety and robustness issues. In the medical AI domain, most of the studies (for example the fracture detection studies[13,19–21]) do not investigate the challenging cases in the evaluation. However, in a real clinical scenario, all kinds of cases (trivial, hard or with incidental findings) can appear. We showed that even in a relatively well studied domain, there exist issues of AI robustness, which expose the requirement for an additional algorithm safety assessment in the medical AI realm.

On the general population test set (test set #1), we observed a near perfect classification performance (AUPR: 0.99, AUROC: 0.99), which, however, still could not surpass the best human rater in terms of Sensitivity, Specificity, Precision, F_1 Score or Balanced Accuracy. The second set of experiments has shown a sharp downfall of performance for test set #2. This dataset comprised the uncertain clinical cases, which could not be diagnosed by a radiologist from an X-ray image, and required an additional confirmation via CT imaging. We note that if we merge the uncertain cases with the general cases, the average performance remains still good, producing an AUROC of 0.97 (0.95−0.98) and an AUPR of 0.97 (0.96−0.98), matching the previous studies.

Along with the reported performance metrics, the inter-rater agreement analysis also shows similar results: all the raters have good agreement with the ground truth for the test set #1, while disagreeing with the ground truth for test set #2. In terms of fracture detection, Sensitivity, Precision, F_1 Score and Balanced Accuracy also decreased for all the raters on the test set #2, indicating that it is difficult for humans to make the decision of the challenging cases.

We investigated deeper whether our model learned any significant associations, which are predictive of fractures on the test set #2. We found that the predictions produced by our model are not more significant than the demographic variables on this dataset. This provides an opportunity for future studies to disentangle the prediction of fractures and the demographic variables.

Another aspect of our work is the assessment of the attention maps. We note that the GradCAM visualizations also confirmed that the DeepWrist did not find the signs of fractures in some of the images, and predicted

the cases as negative, while the CT imaging diagnosed fracture. However, it was interesting to observe that the attention maps did not point at the locations of possible fractures. We believe the assessment of such attention maps in the future can tell about the prediction uncertainty, and could, perhaps, allow to detect the cases, which are likely to be misdiagnosed. When making automatic decisions in clinical practice, such information could be useful, as it could allow for automatic referral of the image to a radiologist, when a machine is incapable of making a decision. We note that similar ideas have been investigated in other domains, such as fundus imaging[38], and we think that it is worth investigating them in the domain of musculoskeletal radiology. Our results show the attempt of using Deep Ensembles to quantify the total predictive uncertainty, however, we observed that the distinction between test set #1 and test set #2 is rather poor. We think different methods, which put a special focus on out-of-domain uncertainty may work better to analyze this problem.

Several limitations of this study should be mentioned. First, our training cases and the test set #1 were annotated from the radiology reports, which might contain misdiagnosis. However, we tried to combat this limitation, by manually verifying the quality of the report during the annotation. In relation to this limitation, we note that the ground truth for the test set #1 was derived from the consensus of R1 and R2, thereby yielding rather optimistic results in terms of the sensitivity and specificity. We think that future studies should also involve an independent set of readers, who will produce the ground truth. The second shortcoming of this work is that we had to exclude some of the cases from the statistical analysis due to their DICOM images having no age and sex metadata (see Supplementary Table S2). Therefore, we conducted the analysis of confounding factors using only the available data. The third limitation here is that the landmark annotations for training and the intra-rater variability analysis were done by a doctoral student (the first author). As a result, it is possible to have bias in the landmark annotation dataset. However, this limitation is rather minor, since after visual inspection of all our data processed by our landmark annotation method, we did not observe a single failure. The fourth limitation of the paper is that for the uncertainty estimation with Deep Ensembles, we were unable to use the power of transfer learning. Thereby, this could have affected the overall predictive performance of the ensemble. However we believe that despite this, the presented results are still indicative of how a state-of-the-art method for uncertainty estimation may perform in evaluating the domain shift. The fifth limitation of our work is limited data: the amount of challenging cases is much lower than the amount of general cases, and all data are taken from a single Hospital. We therefore think that the future studies need to conduct similar evaluations to ours across different hospitals and populations. The final, and major limitation of this work is that it rather poses a new challenge without proposing a solution for it. However, we considered the scope of this study to be in the realm of analysing the applicability of DL to the clinically challenging cases. As we already mentioned in the discussion of the attention maps, one could look at the uncertainty of predictions. The modern advances in Bayesian deep learning have potential to help with such matters[39,40].

To conclude, we believe that the integration of AI into the clinical practice should be taken with care, and new requirements for regulatory approval may need to be introduced. We believe that our work opens a new avenue for research in the realm of DL, and we consider that new methods, which are capable of robust out-of-domain predictive uncertainty estimation are needed to ensure the safety of using AI in healthcare.

Data availability
A Python implementation of DeepWrist is available at https://github.com/MIPT-Oulu/DeepWrist. The training and test data are not public. The repository contains Singularity and Docker containers for testing wrist radiographs in DICOM format.

References
1. Rundgren, J., Bojan, A., Navarro, C. M. & Enocson, A. Epidemiology, classification, treatment and mortality of distal radius fractures in adults: an observational study of 23,394 fractures from the national Swedish fracture register. *BMC Musculoskelet. Disord.* **21**, 88 (2020).
2. Crowe, C. S. *et al.* Global trends of hand and wrist trauma: a systematic analysis of fracture and digit amputation using the global burden of disease 2017 study. *Injury Prev.* **26**, i115–i124 (2020).
3. Karl, J. W., Olson, P. R. & Rosenwasser, M. P. The epidemiology of upper extremity fractures in the United States, 2009. *J. Orthop. Trauma* **29**, e242–e244 (2015).
4. Flinkkilä, T. *et al.* Epidemiology and seasonal variation of distal radius fractures in Oulu, Finland. *Osteoporos. Int.* **22**, 2307–2312 (2011).
5. Knott, P. T. *Casting and Splinting, Chapter 4* 4th edn, 31 (Elsevier Health Sciences, 2020).
6. Taljanovic, M. S. *et al.* Fracture fixation. *Radiographics* **23**, 1569–1590 (2003).
7. De Putter, C. *et al.* Economic impact of hand and wrist injuries: health-care costs and productivity costs in a population-based study. *JBJS* **94**, e56 (2012).
8. Abimanyi-Ochom, J. *et al.* Changes in quality of life associated with fragility fractures: Australian arm of the international cost and utility related to osteoporotic fractures study (ausICUROS). *Osteoporos. Int.* **26**, 1781–1790 (2015).
9. Basha, M. A. A., Ismail, A. A. A. & Imam, A. H. F. Does radiography still have a significant diagnostic role in evaluation of acute traumatic wrist injuries? A prospective comparative study. *Emerg. Radiol.* **25**, 129–138 (2018).
10. Smith-Bindman, R. *et al.* Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Arch. Intern. Med.* **169**, 2078–2086 (2009).
11. Guly, H. Diagnostic errors in an accident and emergency department. *Emerg. Med. J.* **18**, 263–269 (2001).
12. Hallas, P. & Ellingsen, T. Errors in fracture diagnoses in the emergency department-characteristics of patients and diurnal variation. *BMC Emerg. Med.* **6**, 4 (2006).
13. Lindsey, R. *et al.* Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci.* **115**, 11591–11596 (2018).
14. Kitamura, G., Chung, C. Y. & Moore, B. E. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. *J. Digit. Imaging* **32**, 672–677 (2019).

15. Adams, M. *et al.* Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. *J. Med. Imaging Radiat. Oncol.* **63**, 27–32 (2019).
16. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2**, 1–10 (2019).
17. Krogue, J. D. *et al.* Automatic hip fracture identification and functional subclassification with deep learning. *Radiol. Artif. Intell.* **2**, e190023 (2020).
18. Chung, S. W. *et al.* Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop.* **89**, 468–473 (2018).
19. Blüthgen, C. *et al.* Detection and localization of distal radius fractures: deep-learning system versus radiologists. *Eur. J. Radiol.* **126**, 108925 (2020).
20. Thian, Y. L. *et al.* Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiol. Artif. Intell.* **1**, e180001 (2019).
21. Kim, D. & MacKinnon, T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin. Radiol.* **73**, 439–445 (2018).
22. Welling, R. D. *et al.* MDCT and radiography of wrist fractures: radiographic sensitivity and fracture patterns. *Am. J. Roentgenol.* **190**, 10–16 (2008).
23. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning* 151–159 (2020).
24. Chedid, N. *et al.* Synthesis of fracture radiographs with deep neural networks. *Health Inf. Sci. Syst.* **8**, 21 (2020).
25. Selvaraju, R. R. *et al.* Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626 (2017).
26. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* 6402–6413 (2017).
27. Dutta, A. & Zisserman, A. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19* (ACM, 2019). https://doi.org/10.1145/3343031.3350535.
28. Tiulpin, A., Melekhov, I. & Saarakkala, S. Kneel: knee anatomical landmark localization using hourglass networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2019).
29. Paszke, A. *et al.* Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 8026–8037 (2019).
30. Falcon, W. Pytorch lightning. GitHub. Note:https://github.com/PyTorchLightning/pytorch-lightning (2019).
31. Tiulpin, A. Solt: streaming over lightweight transformations. https://doi.org/10.5281/zenodo.3702819 (2019).
32. DeVries, T. & Taylor, G. W. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017).
33. Newell, A., Yang, K. & Deng, J. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision* 483–499 (Springer, 2016).
34. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. Mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017).
35. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7132–7141 (2018).
36. Deng, J. *et al.* ImageNet: a large-scale hierarchical image database. ImageNethttp://www.image-net.org/ (2009).
37. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In *9th Python in Science Conference* (2010).
38. Leibig, C., Allken, V., Ayhan, M. S., Berens, P. & Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* **7**, 1–14 (2017).
39. Solovyev, R. *et al.* Bayesian feature pyramid networks for automatic multi-label segmentation of chest X-rays and assessment of cardio-thoratic ratio. In *International Conference on Advanced Concepts for Intelligent Vision Systems* 117–130 (Springer, 2020).
40. Farquhar, S., Osborne, M. A. & Gal, Y. Radial Bayesian neural networks: beyond discrete support in large-scale Bayesian deep learning. *STAT* **1050**, 7 (2020).

Acknowledgements

Author contributions

A.T., E.V., O.T. and M.N designed the experiments and organized the train data collection. E.V. collected the general population test datasets, organized all test sets' annotation and provided the clinical interpretation of the findings, and participate in the initial draft of the manuscript. A.N. collected the challenging test set. M.N(2)., E.J., K.M., P.P., and T.P. annotated the test data. A.M.R. annotated anatomical landmarks, conducted the experiments, gathered and formally analyzed the results, and wrote the first draft of the manuscript. A.T. supervised the project. All authors reviewed the manuscript and participated in its preparation.

Competing interests

Dr. Aleksei Tiulpin is a co-founder and a shareholder of Ailean Technologies Oy. Other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-021-85570-2.

Correspondence and requests for materials should be addressed to A.M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.