Hierarchical fracture classification of proximal femur X-Ray images using a multistage Deep Learning approach

Hierarchical fracture classification of proximal femur X-Ray images using a multistage Deep Learning approach

Leonardo Tanzi, Enrico Vezzetti, Rodrigo Moreno, Alessandro Aprato, Andrea Audisio, Alessandro Massè

ABSTRACT

Purpose - Suspected fractures are among the most common reasons for patients to visit emergency departments and often can be difficult to detect and analyze them on film scans. Therefore, we aimed to design a Deep Learning-based tool able to help doctors in diagnosis of bone fractures, following the hierarchical classification proposed by the Arbeitsgemeinschaft für Osteosynthesefragen (AO) Foundation and the Orthopaedic Trauma Association (OTA).
Methods - 2453 manually annotated images of proximal femur were used for the classification in different fracture types (1133 *Unbroken* femur, 570 type *A*, 750 type *B*). Secondly, the *A* type fractures were further classified into the types *A1*, *A2*, *A3*. Two approaches were implemented: the first is a fine-tuned InceptionV3 convolutional neural network (CNN), used as a baseline for our own proposed approach; the second is a multistage architecture composed by successive CNNs in cascade, perfectly suited to the hierarchical structure of the AO/OTA classification. Gradient Class Activation Maps (Grad-CAM) where used to visualize the most relevant areas of the images for classification. The averaged ability of the CNN was measured with accuracy, area under receiver operating characteristics curve (AUC), recall, precision and F1-score. The averaged ability of the orthopedists with and without the help of the CNN was measured with accuracy and Cohen's Kappa coefficient.
Results: We obtained an averaged accuracy of 0.86 (CI 0.84-0.88) for three classes classification and 0.81 (CI 0.79-0.82) for five classes classification. The average accuracy improvement of specialists was 14% with and without the CAD (Computer Assisted Diagnosis) system.
Conclusion: We showed the potential of using a CAD system based on CNN for improving diagnosis accuracy and for helping students with a lower level of expertise. We started our work with proximal femur fractures and we aim to extend it to all bone segments further in the future, in order to implement a tool that could be used in every-day hospital routine.

KEYWORDS
Deep Learning; X-Ray; Convolutional Neural Network; Bone Fracture; Orthopaedics

INTRODUCTION

Bone fractures are one of the most common injuries nowadays. Every year approximately 2.7 million fractures occur across the EU6 nations (France, Germany, Italy, Spain, Sweden, and the UK) [1]. Radiographs represent the first-line examination for suspected bone pathology, and classification systems for proximal femur fractures are primarily based on anteroposterior view of the hip [2]. Radiologists play a pivotal role in the diagnostic assessment of the trauma patients, as the correct and prompt identification of fractures strongly affects treatment outcomes. Nevertheless, the evaluation of x-ray images remains challenging: firstly, fractures could be missed because radiographically invisible or equivocal [3]; secondly, a long experience is needed to correctly identify different types of fractures; thirdly, doctors have often to act in emergency situations and

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

may be constrained by time and fatigue. Actually, it has been shown that performance of radiologist in the interpretation of musculoskeletal radiographs decrease in fracture detection at the end of the work day compared to beginning of work day [4]. In addition, radiographic interpretation often takes place in environments without the availability of qualified colleagues for second opinions [5]. A correct treatment and prognosis strongly depend on an accurate classification of the fractures type, such as those defined by the AO foundation. This is mainly accomplished by orthopedic surgeons, who participate in the diagnostic phase alongside radiologists, and successively classify fractures to guide treatment decision-making. In this work, proximal femur fractures were first taken into consideration as they represent the most common reason for admission to an acute orthopaedic ward especially for the elderly population [6]. The AO/OTA classification system was selected for this study because provides clinicians with a standardized methodology in describing fractures and dislocations. This classification is hierarchical, and is based on fracture localization and morphology [7]. In this context, a CAD system able to help doctors might have a direct impact in the outcome of the patients, as we aimed to demonstrate with this work using a Deep Learning [8] approach. Deep Learning is a subset of machine learning that is becoming more and more widely used in the world of computer vision technologies, giving astonishing results in different fields of application. When working with images, CNNs are the most used technology for their ability to capture the spatial and temporal dependencies in an image. Hence, we developed a CNN-based approach in order to implement what we consider the best suited approach for AO/OTA fractures classification, easily extendable to each bone in the human body.

PATIENTS AND METHODS

AO/OTA Proximal Femur Classification

In the AO/OTA classification, the proximal femur is coded as "31", and fractures are located as follows: type *A* fractures concerns the trochanteric region, type *B* fractures the femoral neck and type *C* fractures of the femoral head [9]. Each group is then subsequently divided in different levels of groups, in relation to the complexity of the fracture, considering the number of fracture lines as well as the displacement of fragments. In this study, just type *A* fractures were subclassified: *A1* represents simple pertrochanteric fractures, *A2* multifragmentary pertrochanteric, lateral wall incompetent fractures and *A3* intertrochanteric fractures. The classification process adopted in this study is showed in **Fig.1**.

Dataset

This retrospective study was conducted in a Level-I trauma center. All patients subjected to a pelvic radiograph in the Emergency Department for hip fractures between January 2013 and December 2019 were included in the study. **Table.1** describes baseline and clinical characteristics of the patients included in this study. Then, all anteroposterior pelvic radiographs and related radiological referral were collected anonymously using Synapse 3D (FUJIFILM Corporation). The mean age of patients was 83 (63–91) years. The initial dataset was labelled by a senior trauma surgeon with 17 years of experience and an orthopaedic specialist who has worked specifically on femur fractures in the past 5 years and was composed by 1.787 images of the entire or half hip bone. The first step was a cropping phase, where the areas related to the right and left femur were selected through a semi-automated cropping method and resized to 299 x 299. This technique concerned the use of the OpenCV [10] function *matchTemplate()*. Template matching is a technique for finding areas of an image that are similar to a template image, or to the same image but flipped horizontally (because we may have fractures in both legs). We improved the function using different scales and rotations

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

of the template image. Nevertheless, the percentage of success was around 75%. The remaining one has been cropped manually with an interactive GUI that allowed the user to select the boxes containing the proximal femurs. The second step was a cleaning phase, where images containing prosthesis (n=15), with a too low resolution (n=47), with the area around the femur partially hidden (n=23) or showing a lateral view (n=98) were excluded (n=183). Afterwards, the dataset was reviewed a second time by two radiologists from our medical team, to confirm the validity of the ground truth. The final number of images was: 1133 *Unbroken* femurs, 570 type *A*, 750 type *B* and 4 type *C*. Because of the low number of *C* fractures (n=4), we decided to exclude this class. Among the *A* type, 280 fractures were labelled as *A1*, 183 as *A2* and 107 as *A3*. This process is shown in **Fig.2** following the STARD 2015 flow diagram [11]. Type A fractures have been further classified A1, A2 and A3, as the AO classification in groups was found to be more reliable than other classification systems and showed good reproducibility between the observers [12]. Sub-classification of type *B* fractures, on the other hand, reported poor intra- and inter-observer reliability and limited predictive value for treatments and outcomes [13,14]. Some real X-Rays for each class taken from our dataset are shown in **Fig 3.**

Methods

Two different methods where developed, shown in the flow chart in **Fig.4** for the three classes classification case. After trying different architectures and methods, the model which gave best test accuracy was a simple InceptionV3 [15] network, with the last layer replaced with a Softmax layer for three and five classes classification and pre-trained on ImageNet [16]. We wrote a function to assign different weights to each class when computing the loss, in order to compensate for the unbalanced dataset, i.e. higher weights have been assigned to classes with fewer images. After trying different configurations, we obtained the best results using a batch size of 32 and Adam optimizer [17] with a learning rate of 0.0001 and beta values of 0.9 and 0.999 respectively. The function to calculate the loss was the *sparse categorical crossentropy*. We also implemented data augmentation in the training set using a rotation range of 10 degrees, horizontal flip and both height and width shift from 0.0 to 0.1 fraction of total height or width. We did not use any more complex transformations as shearing in order not to create artificial fractures. We carried on the same computation with VGG16 [18] and ResNet50 [19], before choosing InceptionV3 as the best architecture for the problem at hand, keeping every parameter the same (except the images input size that is 224 × 224 pixels for both of them). After selecting the best network and using its performance as a baseline, we switched to a second approach training our InceptionV3 network defined above to tackle a binary classification between *Broken* and *Unbroken* and *A* and *B* and a three classes classification between *A1, A2* and *A3*. We then tested the networks as a cascade of three stages: the first one recognizes between *Unbroken* and *Broken*, the second classifies between *A* and *B* the images labelled as *Broken* from the first network and the third classifies between *A1, A2,* and *A3* the ones resulting in *A* class. We then improved one more time the performances for *A1, A2* and *A3* class substituting the third stage with two sub-stages. This approach tried to emulate the method used by the specialists to classify fractures and its hierarchical structure is particularly appropriate for AO/OTA classification. Lastly, we used Grad-CAM [20] to visualize where the network was focusing for the different classifications.

Training, Framework and Evaluation

From the initial dataset, 20% of images for each class where kept apart for testing, resulting in a test set of 115 images of type *A*, 150 type *B* and 226 *Unbroken*. The networks were then trained and validated using 5-fold cross validation with the remaining images (455 type *A*, 600 type *B* and *907*

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

Unbroken), as shown in **Fig.5**. We ran the model for 150 epochs using *EarlyStopping* with a patience of 10 epochs. We used Keras [21], an open-source neural-network library written in Python, running on top of TensorFlow [22], on Ubuntu 16.04.5 LTS with NVIDIA GeForce GTX 1080 Ti. For each network, we computed the top-1 accuracy, the conventional accuracy for the deep CNN answer (top-1) being exactly the expected answer, among choices of *A*, *B* and *Unbroken* or *A1*, *A2*, *A3*, *B*, *Unbroken*. Then, the performances for single classes were measured using the area under the receiver operating curve (AUC), precision, recall and F1-score. Every value was averaged among the 5 folds. Performances of the specialists with and without the CAD system were measured using accuracy and Cohen's Kappa scores [23]. Firstly, one radiologist with 5 years of experience and one orthopaedic student with 1 year of experience evaluated the type of fracture of 150 hips without the help of the neural network. This set of images were taken from the test dataset and therefore not involved in the training process, in order to obtain comparable results. We decided to use two raters with different levels of experience to demonstrate that this tool could also be useful for educational purpose. It should be mentioned that the specialist participated in the creation of the ground truth classification together with a senior surgeon, although the evaluation for this study was conducted one year after the ground truth was created. Fourteen days later, in order to produce unbiased results, we gave them the prediction of the neural networks in cascade and the probability that the neural network assigned to each class.

STATISTICS

Statistics such as the top-1 accuracy, AUC, precision, recall and F1-score were computed with Python NumPy [24] and Scikit-learn [25] libraries, using a 95% confidence interval (CI). Comparisons between the CNN and humans were manually performed.

RESULTS

Baseline Method

Running the aforementioned InceptionV3 model, we obtained, using 5-fold cross validation, an average test accuracy of 0.87 (CI 0.85-0.88) for three classes classification and 0.78 (CI 0.76-0.79) for five classes classification. We chose InceptionV3 after running ResNet50 and VGG16 with the same parameters and obtaining lower test accuracies. ResNet50 gave an average test accuracy of 0.85 (CI 0.84-0.87) for three classes classification and 0.75 (CI 0.73-0.77) for five classes classification. VGG16 gave an average test accuracy of 0.82 (CI 0.79-0.85) for three classes classification and 0.77 (CI 0.76-0.78) for five classes classification. These values are shown in **Table.2**. We then computed precision, recall and F1-score for each class using InceptionV3 network (**Table.3**). The values are shown with related confidence interval. Lastly, we computed the confusion matrix and ROC curves for each fold. ROC and confusion matrix for the model that gave the best results both for three classes and five classes are shown in **Fig.6.** Average AUC among 5 folds for three classes classification was 0.95 (CI 0.93-0.97) for class *A*, 0.93 (CI 0.92-0.95) for class *B* and 0.96 (CI 0.95-0.97) for class *Unbroken*. Average AUC among 5 folds for five classes classification was 0.84 (CI 0.78-0.90) for class *A1*, 0.93 (CI 0.92-0.93) for class *A2*, 0.97 (CI 0.96-0.98) for class *A3*, 0.93 (CI 0.91-0.95) for class *B* and 0.97 (CI 0.96-0.98) for class *Unbroken*. Clearly, as noticed in **Table.3** and **Fig.6 (D)**, the network is not learning to recognize features of

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

A1, *A2* and *A3* fractures. This is the reason why we decided to implement the already mentioned multistage approach.

Multistage Method

Training three InceptionV3 networks for classification between *Broken-Unbroken*, *A-B* and *A1-A2-A3* with 5-folds cross validation, resulted in an average test accuracy of 0.91 (CI 0.89-0.93), 0.87 (CI 0.86-0.89) and 0.61 (CI 0.54-0.68) respectively. Values for each fold are shown in **Table.4**. Building the already described multi-stage architecture, we obtained an average accuracy of 0.86 (CI 0.84-0.88) for three classes classification and 0.80 (CI 0.77-0.82) for five classes classification. The other metrics are shown in **Table.5.** Notice that the performance for three classes resulted similar to the previous method. On the other hand, we had important improvements for five classes classification. However, these improvements were still not optimal, especially for *A1*, *A2*, and *A3* subclasses. For this approach we trained the last two networks with an increased dataset composed by new images of *A1*, *A2* and *A3* fractures, resulting in a total of 495 *A1*, 293 *A2* and 170 *A3*. With the help of the specialists, we noticed that the main problem was distinguishing between *A1* and *A2*. This was confirmed running a binary network to classify between *A1* and *A2* together against *A3* class, resulting in an average accuracy of 0.92 (CI 0.89-0.95). Thus, we added this step and a successive binary network to distinguish between *A1* and *A2* that reached an average accuracy of 0.68 (CI 0.65-0.70). The full pipeline is shown in **Fig.7.** Adding a new stage to the computation, we obtained an accuracy of 0.81 (CI 0.79-0.82) and values of precision, recall and F1-score increased, as shown in **Table.6.**

CAD-system

The evaluation of the type of fracture present in 150 bones images without the help of the neural network, performed by the radiologist and the student from our medical team, resulted in an accuracy of 0.99 and 0.95 respectively for *Broken-Unbroken* classification, 0.95 and 0.89 for *A-B* classification and 0.73 and 0.68 for *A1-A2-A3* classification. Fourteen days later, the same test was performed with the help of the neural networks in cascade, which obtained, with this particular set of images, an accuracy of 0.84. With the help of the CAD system, the accuracy of both the specialist and the student of *Broken-Unbroken* and *A-B* classification changed very little, while the accuracy for sub-fractures classification augmented to 0.86 and 0.83 respectively, that was the true aim of this work. The result is an average improvement of 14% in accuracy and a considerable reduction in time spent for evaluation. We then computed Cohen's Kappa scores, shown in **Table.7**, in order to observe the inter-agreement between the neural network and the raters with and without the CAD system's help.

DISCUSSION

In a previous paper [26] from our group, we reviewed some selected papers concerning this topic, starting from basic approaches to the main advanced solutions. Initial prior works for detection and classification of fractures [27,28] focused on conventional machine learning processes consisting of pre-processing, feature extraction and classification phases. Recently, impressive results have been obtained using Deep Learning methods. The majority of the existing works regarding fractures classification, focused mainly on the binary classification between broken and unbroken bones [29–31]. This aim unfortunately has a low impact on doctor diagnosis. To the best of our knowledge, Deep Learning has been applied to classify different types of fractures just in two previous papers by [32] and [33]. Nevertheless, results are still non optimal, especially for complex fractures, and a generalized approach still does not exist. The main objective of this work was to attempt to fill this

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

lack. We developed a system based on the AO/OTA classification for proximal femur fractures, because of its consistency in the classification process for all bone segments in case of further developments of this project to other districts. Using transfer learning with the InceptionV3 architecture, we reached an accuracy of 0.87 (CI 0.85-0.88) for fracture classification in types *A, B* and *Unbroken* and 0.78 (CI 0.76-0.79) for types *A1, A2, A3, B, Unbroken*. We used these results as baseline and we proposed a new method to increase the accuracy, especially for sub-classes, using different subsequent networks in cascade. This second approach is perfectly suitable for AO/OTA hierarchical classification and also allowed us to work on sub-fractures, such as *A1, A2* and *A3* and we obtained a 0.80 (CI 0.77-0.82) test accuracy for five classes classification, 2% more than the previous approach. We then added a new stage of classification to increase the performance to 0.81 (CI 0.79-0.82). In addition, to demonstrate that the network was actually learning, we implemented Grad-CAM, one variant of CAM [34] broadly applicable to any CNN-based architectures, which allowed to visualize where the network was focusing for the different classifications. Some examples of Grad-CAM heat maps are shown in **Fig.8** related to different types of fracture. This visualization confirmed that the network was focusing in the correct area of the bone: for *A* prediction the network focused in the lower part of the proximal femur, for *B* prediction around the neck. Unfortunately, for *A1, A2* and *A3* subgroups we were not able to find a recurrent pattern, a clear sign that this classification still needs to be improved. We finally showed the potential of this tool for helping in diagnosis with an average improvement of 14%. In **Table.7**, we could notice in the first row that the agreement between the neural network and the raters improved by 0.17 and 0.40 with and without the CAD system. This demonstrated that both of them changed their predictions accordingly with the neural network predictions, especially the student. Plus, we highlighted in red the agreement of the raters with and without the CAD system: from these values we could notice how the student changed his predictions accordingly to the neural network, resulting in a lower Cohen's Kappa (0.60) compared to the one of the specialist (0.73). In the end, from the values emphasized in green, we could see how the agreement between the specialist and the student increased with the use of the CAD system. For the specialist, the accuracy with the CAD system's help also surpassed the accuracy of the neural network alone, demonstrating the usefulness of a combined work between specialist and CAD system. For the student, we had an important improvement in accuracy and the help of the CAD system was also demonstrated by the growth of Cohen's Kappa coefficient.

LIMITS AND FUTURE PROSPECTS

These results underlined the utility of this tool both for practical and educational purpose. Nonetheless, many challenges need to be undertaken in order to overcome the limits of this approach. Firstly, our dataset is only based on the anteroposterior view of the hip. Our labelling phase was performed by both experienced radiologists and orthopedic surgeons and further checked using radiological referrals that were written using multiple projections of the affected hip. On the other hand, our tests were performed using just images from the dataset, therefore just the anteroposterior view of the fractured hip. For this reason, the physicians involved reported lower than expected accuracies for the classification of images, especially for subclassification of type *A* fractures. Secondly, we performed our tests on 150 images with and without the use of the CAD at 14-days distance. If the time between the two tests is reasonable, the tester's performances could be biased by a lack of further blinding to the test. Thirdly, we could not analyze 31*C* fractures due to the low number of images. These injuries are rare and occur in conjunction with hip dislocations in 5% to 15% of cases. To our knowledge, no inter and intra-observer accuracy of femoral head

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

fracture classifications like the AO/OTA or the Pipkin classification has been published. The low number of images is a real common issue in the Deep Learning world, especially in the medical field for the analysis of rare injuries. Techniques have been described to artificially increase the number of images, as Generative Adversarial Networks (GANs) [35]. In GANs, two networks compete using unsupervised machine learning. The first network, the generator, generates a data instance (for example, an image) which mimics real world data. This is fed to the second network, the discriminator, along with authentic examples from the real world. The generator's aim is to convince the discriminator that the generated data is authentic, while the discriminator has to discern between which data is real and which is replicated. In our case scenario, GANs could create new artificial images representing a certain type of fracture. On the other hand, reliability of the newly created images would be a major concern. We tried to artificially increase the number of fractures with GANs and the result is shown in **Fig.9**. We can notice that the network guessed the right shape of the bone but in some cases had still difficulty with the orientation of the femur. For this reason, these results are far from reliable. An alternative approach could be to feed a neural network with all types of fractures and generate fake samples. Then the doctors should select among these which ones can be considered realistic and adapted to build the dataset. Finally, our aim was to show the potential of using a CAD system based on CNNs to improve the diagnosis process, and to aid less experienced physicians in the identification of proximal femur fractures. In addition, our approach was designed to be generalizable to other bone segments. Our intent is to develop a comprehensive tool that can become part of our daily-practice for fractures' management.

FIGURES CAPTIONS

Fig.1 AO/OTA hierarchical classification determined by the localization and configurations of the fracture lines. Type *A* fractures concerns the trochanteric region, type *B* fractures the femoral neck and type *C* fractures of the femoral head. Each group is then subsequently divided in different levels of subgroups. In this figure, just type *A* fractures are showed: *A1* represents simple pertrochanteric fractures, *A2* multifragmentary pertrochanteric, lateral wall incompetent fractures and *A3* intertrochanteric fractures.
Fig.2 STARD 2015 Flow Diagram to define how the dataset is decomposed.
Fig.3 Some samples of real X-Rays images used for training the neural network after the cleaning and cropping phase.
Fig.4 Flow chart for the three classes classification case. After a semi-automated cropping phase, two approaches were presented: a classic CNN for classification used as baseline and a multistage one characterized by subsequent binary networks. Finally, Grad-CAM where used to visualize where the network was focusing.
Fig.5 Full composition of train, validation and test sets.
Fig.6 ROC curve with associated AUC (A, C) and confusion matrix (B, D) are shown for both three and five classes classification. The number of images used for testing was: 226 *Unbroken* images, 150 *B*, 114 *A*, 56 *A1*, 36 *A2* and 21 *A3*.
Fig.7 Full pipeline adding a new step of classification to distinguish the *A3* type fractures from *A1* and *A2*.
Fig.8 Grad-CAM output for *A* and *B* type fractures classification. The visualization confirms that the neural network is focusing in the correct area of the femur.
Fig.9 Artificial femur images produced with GANs technology.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

TABLES

		Total (n=1787))
Age (*yr*)	Median (IQR)	81 (73-86)
Sex	F	1206
	M	581
	%F	67.5%

Table.1 Baseline characteristics.

K-Fold	InceptionV3		VGG16		ResNet50	
	3 classes	5 classes	3 classes	5 classes	3 classes	5 classes
Fold1	0.87	0.76	0.84	0.78	0.88	0.75
Fold2	0.88	0.77	0.82	0.76	0.85	0.74
Fold3	0.85	0.77	0.80	0.76	0.82	0.76
Fold4	0.87	0.78	0.84	0.78	0.84	0.74
Fold5	0.86	0.79	0.79	0.77	0.86	0.77
Average	**0.87**	**0.78**	**0.82**	**0.77**	**0.85**	**0.75**
Standard Deviation	±0.01	±0.01	±0.02	±0.01	±0.02	±0.01

Table.2 Comparison between three and five classes classification for InceptionV3, VGG16 and ResNet50.

Class	Precision		Recall		F1-score		# of images for testing
	3 classes	5 classes	3 classes	5 classes	3 classes	5 classes	
Unbroken	**0.90** (0.86-0.93)	**0.91** (0.89-0.92)	**0.92** (0.89-0.95)	**0.92** (0.91-0.93)	**0.91** (0.89-0.92)	**0.91** (0.90-0.92)	226
B	**0.83** (0.80-0.87)	**0.78** (0.72-0.84)	**0.81** (0.76-0.86)	**0.83** (0.81-0.86)	**0.82** (0.80-0.84)	**0.81** (0.77-0.84)	150
A	**0.86** (0.80-0.91)		**0.84** (0.79-0.89)		**0.85** (0.84-0.86)		114
A1		**0.43** (0.36-0.50)		**0.29** (0.16-0.42)		**0.34** (0.25-0.43)	56
A2		**0.42** (0.38-0.46)		**0.50** (0.27-0.73)		**0.44** (0.35-0.53)	36
A3		**0.74** (0.56-0.93)		**0.57** (0.49-0.66)		**0.64** (0.55-0.72)	21

Table.3 Precision, Recall and F1-score for three and five classes classification using the InceptionV3 network. The number of images for testing is proportioned to the initial dataset unbalance.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

K-Fold	*Broken-Unbroken*	*A-B*	*A1-A2-A3*
Fold1	0.90	0.88	0.58
Fold2	0.92	0.87	0.54
Fold3	0.90	**0.89**	0.62
Fold4	**0.93**	0.86	0.62
Fold5	0.90	0.86	**0.68**
Average	0.91	0.87	0.61
Standard Deviation	±0.01	±0.01	±0.04

Table.4 Accuracy using 5-fold cross validation for the three networks used in the different stages.

Class	Precision		Recall		F1-score		# of images for testing
	3 classes	5 classes	3 classes	5 classes	3 classes	5 classes	
Unbroken	**0.91** (0.89-0.92)	**0.93(↑0.02)** (0.92-0.94)	**0.90** (0.86-0.93)	**0.90(↑0.02)** (0.86-0.93)	**0.90** (0.88-0.92)	**0.91(↑0.02)** (0.89-0.93)	226
B	**0.80** (0.76-0.84)	**0.85(↑0.07)** (0.82-0.87)	**0.83** (0.80-0.86)	**0.83 (=)** (0.80-0.86)	**0.82** (0.81-0.83)	**0.84(↑0.03)** (0.83-0.85)	150
A	**0.84** (0.79-0.89)		**0.82** (0.73-0.90)		**0.83** (0.78-0.87)		114
A1		**0.50(↑0.07)** (0.46-0.53)		**0.53(↑0.24)** (0.45-0.61)		**0.51(↑0.17)** (0.47-0.55)	56
A2		**0.45(↑0.03)** (0.35-0.55)		**0.56(↑0.06)** (0.40-0.72)		**0.49(↑0.05)** (0.40-0.58)	36
A3		**0.70(↑0.04)** (0.51-0.88)		**0.56(↓0.01)** (0.45-0.68)		**0.62(↓0.02)** (0.50-0.73)	21

Table.5 Precision, Recall and F1-score for three and five classes classification using the multistage approach. The number of images is proportioned to the initial dataset unbalance. Improvements for five classes classification are shown in parenthesis.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

Class	Precision	Recall	F1-score	# of images for testing
Unbroken	**0.93 (=)** (0.92-0.94)	**0.90 (=)** (0.86-0.93)	**0.91 (=)** (0.89-0.93)	226
B	**0.85 (=)** (0.82-0.87)	**0.83 (=)** (0.80-0.86)	**0.84 (=)** (0.83-0.85)	150
A1	**0.49 (↓0.01)** (0.45-0.54)	**0.54 (↑0.01)** (0.37-0.70)	**0.51 (=)** (0.42-0.60)	56
A2	**0.50 (↑0.05)** (0.41-0.58)	**0.55 (↓0.01)** (0.39-0.71)	**0.51 (↑0.02)** (0.43-0.60)	36
A3	**0.73 (↑0.03)** (0.54-0.93)	**0.73 (↑0.17)** (0.64-0.82)	**0.73 (↑0.11)** (0.62-0.84)	21

Table.6 Precision, Recall and F1-score for three and five classes classification using the multistage approach with 4 stages. Values for five classes classification increased, especially for A3, compared to the three stages approach, as shown in parenthesis.

Cohen's Kappa	Specialist No CAD	Specialist CAD	Student No CAD	Student CAD
Neural Network	**0.60**	**0.77 (↑0.17)**	**0.48**	**0.88 (↑0.40)**
Specialist No CAD		0.73	0.50	0.66
Specialist CAD			0.57	0.76
Student No CAD				0.60

Table.7: Cohen's Kappa scores to measure inter-agreement. Values between 0.41 and 0.60 indicate a moderate agreement, between 0.61 and 0.80 a substantial agreement and between 0.81 and 0.99 an almost perfect agreement.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

REFERENCES

[1] International Osteoporosis Foundation. Broken bones, broken lives: a roadmap to solve the fragility fracture crisis in Europe 2018. https://www.iofbonehealth.org/broken-bones-broken-lives (accessed September 20, 2019).

[2] Kani KK, Porrino JA, Mulcahy H, Chew FS. Fragility fractures of the proximal femur: review and update for radiologists. Skeletal Radiol 2019;48:29–45. https://doi.org/10.1007/s00256-018-3008-3.

[3] Lubovsky O, Liebergall M, Mattan Y, Weil Y, Mosheiff R. Early diagnosis of occult hip fractures MRI versus CT scan. Injury 2005;36:788–92. https://doi.org/10.1016/j.injury.2005.01.024.

[4] Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long Radiology Workdays Reduce Detection and Accommodation Accuracy. Journal of the American College of Radiology 2010;7:698–704. https://doi.org/10.1016/j.jacr.2010.03.004.

[5] Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department – characteristics of patients and diurnal variation. BMC Emerg Med 2006;6:4. https://doi.org/10.1186/1471-227X-6-4.

[6] Giannoulis D, Calori GM, Giannoudis PV. Thirty-day mortality after hip fractures: has anything changed? Eur J Orthop Surg Traumatol 2016;26:365–70. https://doi.org/10.1007/s00590-016-1744-4.

[7] Journal of Orthopaedic Trauma. Femur 2018;32:S33–44. https://doi.org/10.1097/BOT.0000000000001058.

[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[9] Meinberg E, Agel J, Roberts C, Karam M, Kellam J. Fracture and Dislocation Classification Compendium—2018: Journal of Orthopaedic Trauma 2018;32:S1–10. https://doi.org/10.1097/BOT.0000000000001063.

[10] Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools; 2000.

[11] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6:e012799. https://doi.org/10.1136/bmjopen-2016-012799.

[12] Jin W-J, Dai L-Y, Cui Y-M, Zhou Q, Jiang L-S, Lu H. Reliability of classification systems for intertrochanteric fractures of the proximal femur in experienced orthopaedic surgeons. Injury 2005;36:858–61. https://doi.org/10.1016/j.injury.2005.02.005.

[13] Blundell CM, Parker MJ, Pryor GA, Hopkinson-Woolley J, Bhonsle SS. Assessment of the AO classification of intracapsular fractures of the proximal femur. The Journal of Bone and Joint Surgery British Volume 1998;80-B:679–83. https://doi.org/10.1302/0301-620X.80B4.0800679.

[14] Masionis P, Uvarovas V, Mazarevičius G, Popov K, Venckus Š, Baužys K, et al. The reliability of a Garden, AO and simple II stage classifications for intracapsular hip fractures. Orthop Traumatol Surg Res 2019;105:29–33. https://doi.org/10.1016/j.otsr.2018.11.007.

[15] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE; 2016, p. 2818–26. https://doi.org/10.1109/CVPR.2016.308.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

[16] Fei-Fei L, Deng J, Li K. ImageNet: Constructing a large-scale image database. Journal of Vision 2010;9:1037–1037. https://doi.org/10.1167/9.8.1037.
[17] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. ArXiv:14126980 [Cs] 2017.
[18] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv:14091556 [Cs] 2015.
[19] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE; 2016, p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
[20] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE; 2017, p. 618–26. https://doi.org/10.1109/ICCV.2017.74.
[21] Chollet F, others. Keras. 2015.
[22] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015.
[23] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276–82.
[24] Oliphant T. NumPy: A guide to NumPy. 2006.
[25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–2830.
[26] Tanzi L, Vezzetti E, Moreno R, Moos S. X-Ray Bone Fracture Classification Using Deep Learning: A Baseline for Designing a Reliable Approach. Applied Sciences 2020;10:1507. https://doi.org/10.3390/app10041507.
[27] Cao Y, Wang H, Moradi M, Prasanna P, Syeda-Mahmood TF. Fracture detection in x-ray images through stacked random forests feature fusion. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA: IEEE; 2015, p. 801–5. https://doi.org/10.1109/ISBI.2015.7163993.
[28] Myint WW, Tun HM, Tun KS. Analysis on Detecting of Leg Bone Fracture from X-ray Images. IJSRP 2018;8. https://doi.org/10.29322/IJSRP.8.9.2018.p8150.
[29] 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 USA 2018;115:11591–6. https://doi.org/10.1073/pnas.1806905115.
[30] Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? Acta Orthopaedica 2017;88:581–6. https://doi.org/10.1080/17453674.2017.1344459.
[31] Pranata YD, Wang K-C, Wang J-C, Idram I, Lai J-Y, Liu J-W, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. Computer Methods and Programs in Biomedicine 2019;171:27–37. https://doi.org/10.1016/j.cmpb.2019.02.006.
[32] Chung SW, Han SS, Lee JW, Oh K-S, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthopaedica 2018;89:468–73. https://doi.org/10.1080/17453674.2018.1453714.
[33] Jiménez-Sánchez A, Kazi A, Albarqouni S, Kirchhoff C, Biberthaler P, Navab N, et al. Towards an Interactive and Interpretable CAD System to Support Proximal Femur Fracture Classification. ArXiv:190201338 [Cs] 2019.
[34] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. ArXiv:151204150 [Cs] 2015.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

[35] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. ArXiv:14062661 [Cs, Stat] 2014.

FIGURES

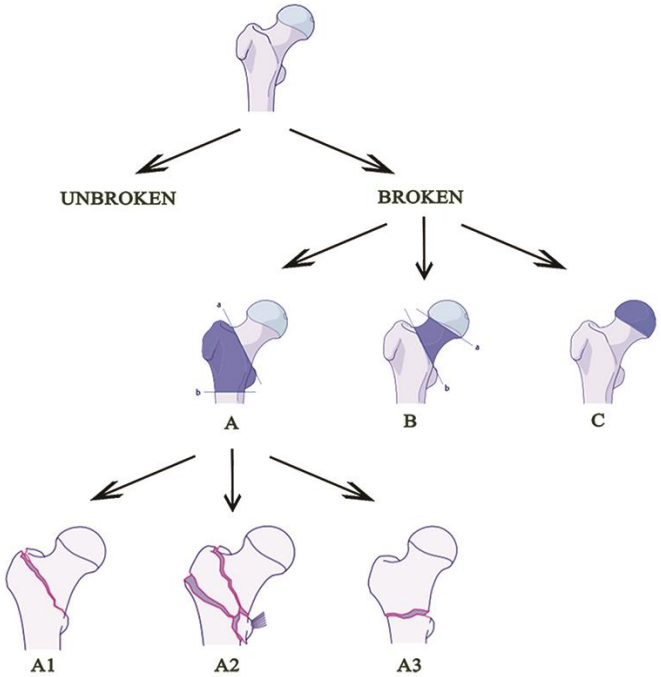

Fig.1

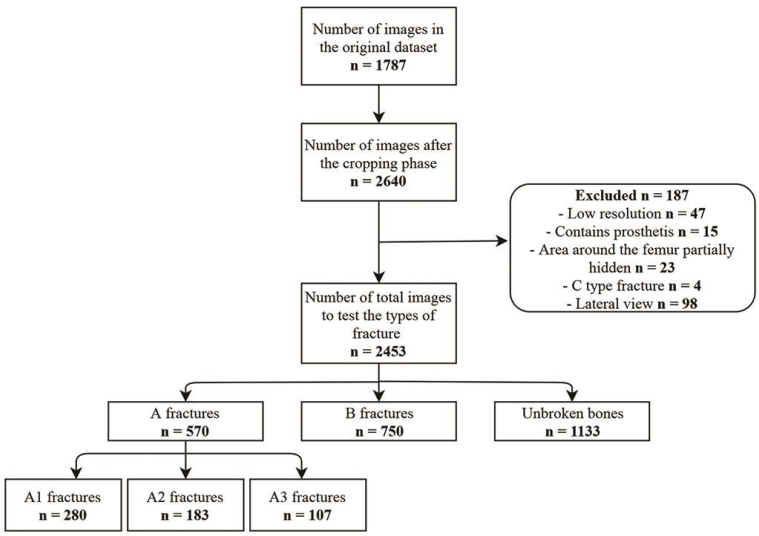

Fig.2

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

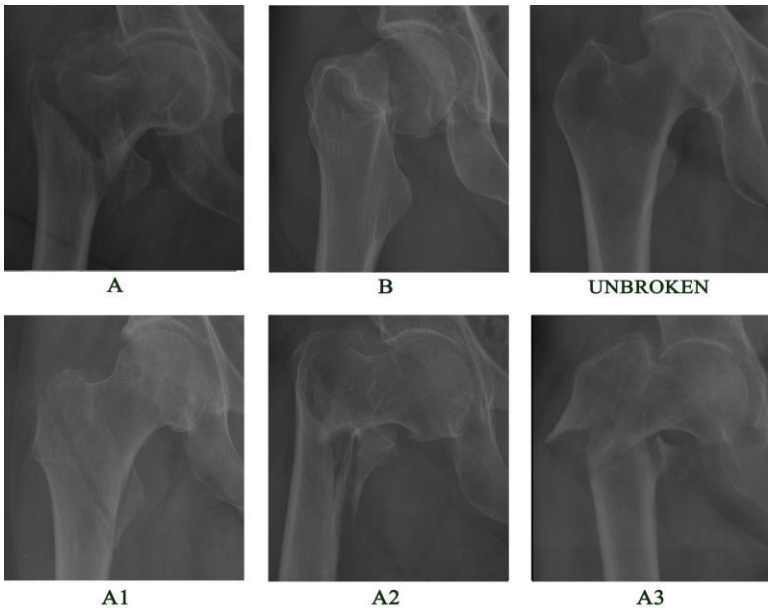

Fig.3

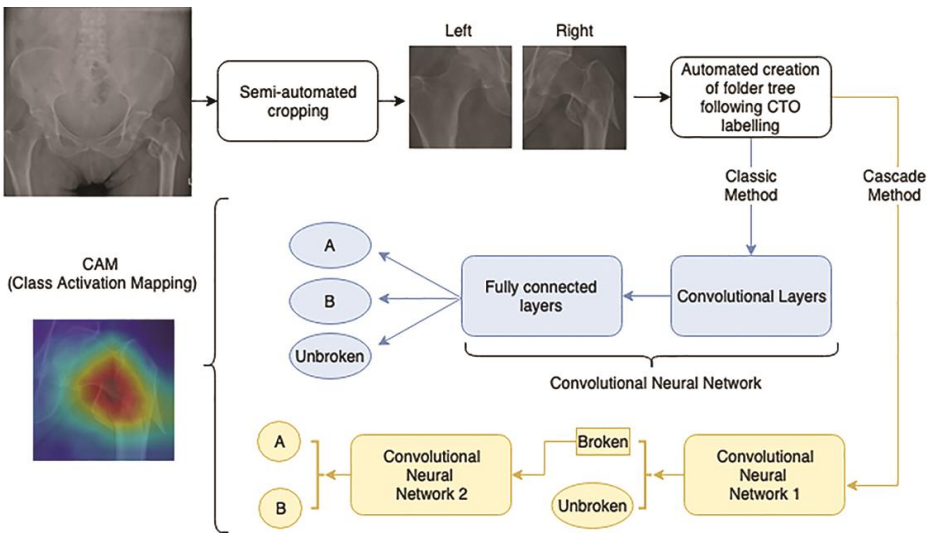

Fig.4.

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

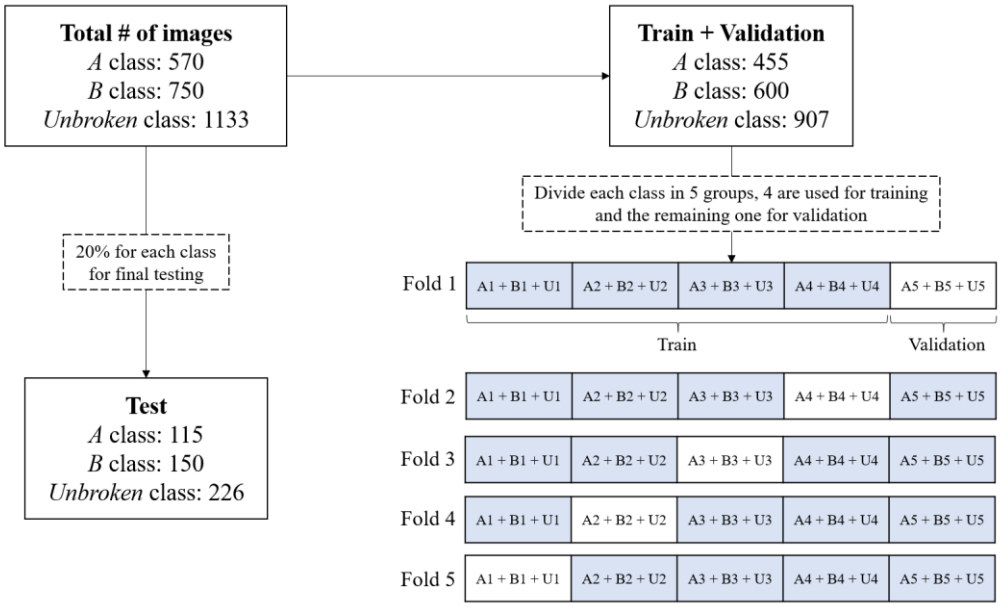

Fig.5

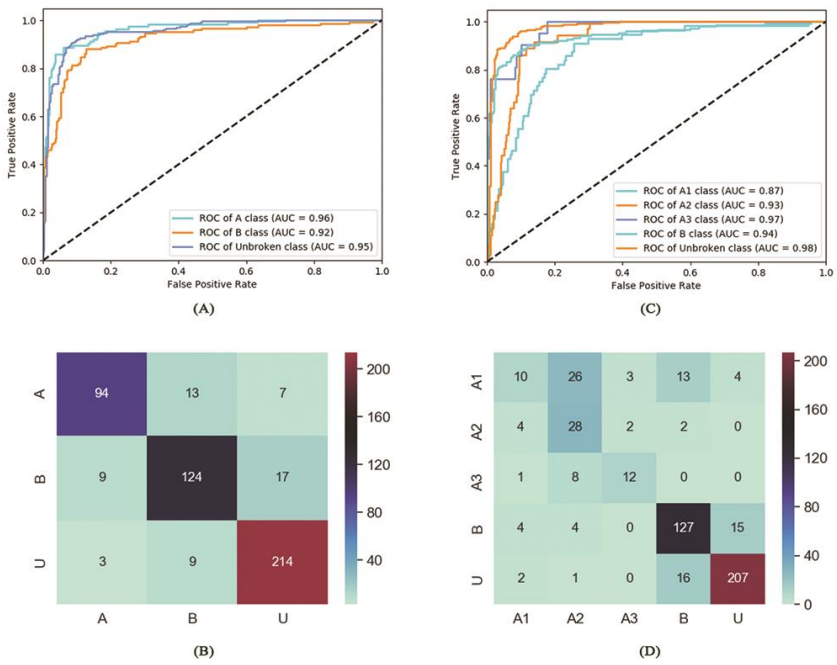

Fig.6

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

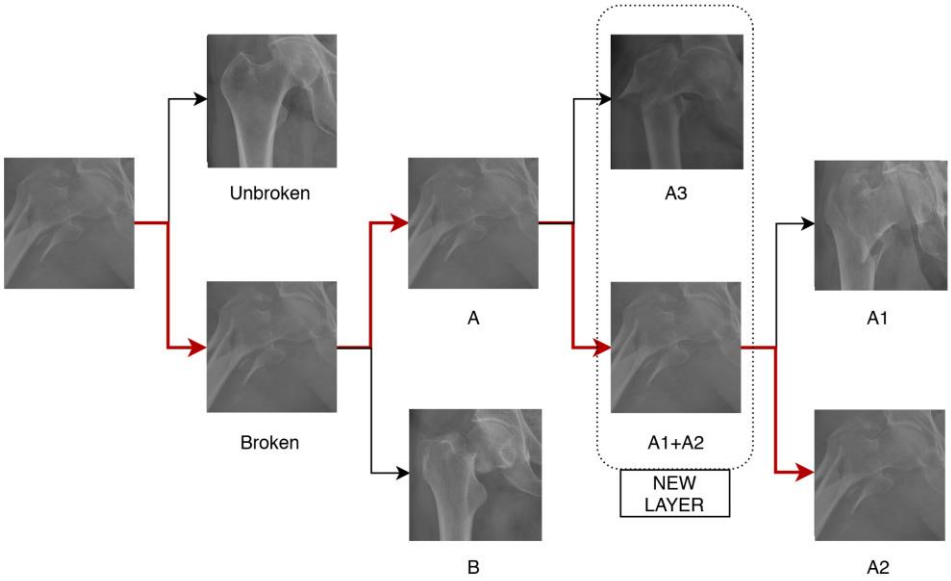

Fig.7

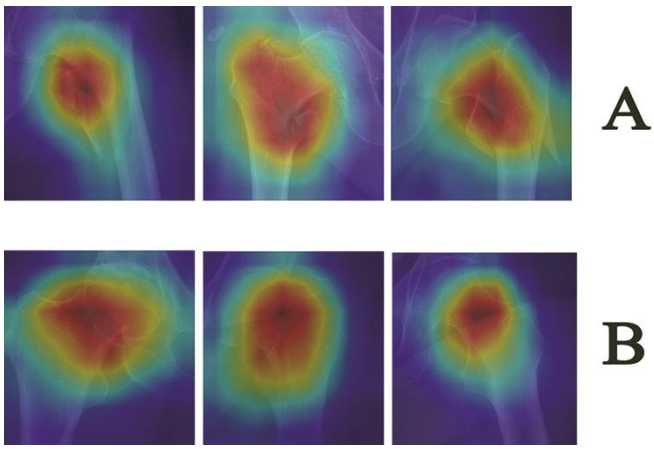

Fig.8

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.

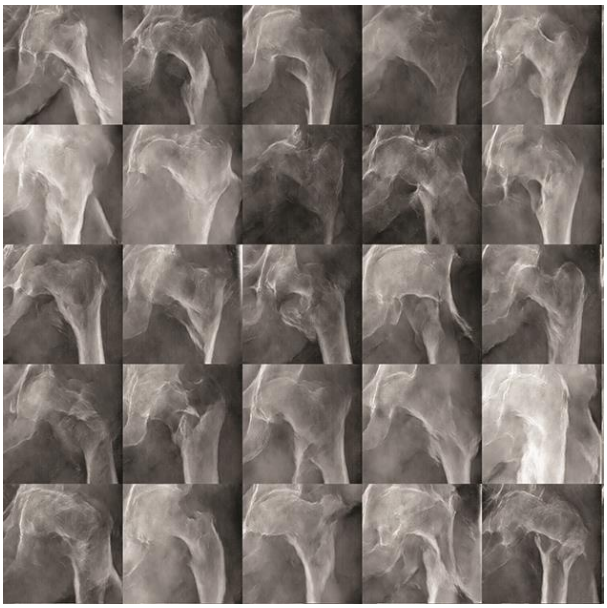

Fig.9

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association; CNN, convolutional neural network; Grad-CAM, Gradient Class Activation Maps; ROC, operating characteristics curve; AUC, area under receiver operating characteristics curve; CAD, Computer Assisted Diagnosis.