Deep Learning for Fracture Detection in the Radius and Ulna on Conventional Radiographs

Jeroen Verboom
STUDENT NUMBER: 2030739

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Sharon Ong
Yash Satsangi

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 2021

Preface

I would like to express my gratitude to the people who helped me produce the master thesis. In particular I want to thank Nils Hendrix who was devoted to help me in terms of guidance, critical thinking and with coding examples. The last couple of moths were very insightful where I learned a lot from him. Without his support the product of this research would not be what it is today. Furthermore, I want to thank Sharon Ong who was a great supervisor keeping me sharp and helping me develop the correct structure this work and to cover the required sections. I am grateful to have had such support from both supervisors. Furthermore I want to thank Mattieu Rutten and Bram van Ginneken for setting up the AI for Health program and making this thesis opportunity possible.

Deep Learning for Fracture Detection in the Radius and Ulna on Conventional Radiographs

Jeroen Verboom

This work gives a compartmentalized overview of a fracture detection tool to detect and localize fractures in the radius and ulna on conventional radiographs using deep learning. This contrasts earlier studies in that it proposes a more efficient object detector, demonstrates the generalizability of fracture detection models to data from a different hospital, and employs more advanced class activation mapping methods for fracture localization. Both RadboudUMC and the Jeroen Bosch Ziekenhuis provided data to create a mutli-institutional dataset. The two data sources enabled me to demonstrate how fracture detection classifiers trained on data from only one institution significantly perform less when tested on data from another institution. Moreover, this study demonstrates a more efficient bone localization method that yields adequate performance to be used for cropping regions of interest, and a newer fracture localization method (ScoreCAM) that outperforms its predecessors in terms of highlighting less redundant information. I conclude that the algorithms presented in this work show the potential to be incorporated in a clinically usable fracture detection tool. However, more research needs to be conducted using multi-institutional for training fracture detection classifiers.

1 1 Introduction

1.1 Project description

Fracture epidemiology shows that the wrist is the most common body part to fracture (Brown & Ceasar, 2006). The societal cost of wrist injuries amounts up to 410 million a year, mainly caused by a loss of productivity in the working populace (Putter et al., 2016). The conventional way of diagnosing these fractures is by using radiography. The affordability and ease of use make it an attractive tool for diagnosis, however, the technique is limited in capturing details. The low resolution on radiographs poses a challenge for diagnosis because the images are inspected with the naked eye to determine the presence and severity of a fracture. Several studies have reported that the limited detail in radiographs forms a problem specifically in wrist and hand areas causing radiologists to overlook fractures. Welling, Jacobson, Jamadar, et al. (2008) showed that radiologists failed to detect 30% of wrist fractures on conventional radiographs that were visible in CT-scans. Kiuru, Haapamaki, Koivikko, et al. (2004) found that clinicians overlooked 14% of the fractures present while 37% of suspected fractures were proven not to be present by additional imaging. In a more recent study, Balci, Basara, Çekdemir, et al. (2015) reported similar findings claiming that one third of all wrist fractures are overlooked in conventional radiographs.
Wrongful interpretation of radiographs creates an issue particularly for the radius and associated ulna traumas because the distal parts of these bones are amongst the most

frequent bones to fracture in the body (Freed & Shields, 1984; Goldfarb, Yin, Gilul, 2001; Kiuru et al., 2004; Baig, 2017; Joeris, Lutz, Blumenthal, et al., 2017). Misdiagnosing fractures these fractures can lead to malunion, long-lasting pain, dysmorphia and nerve damage. The golden standard to detect or rule out fractures is by using a corresponding CT-scan, however CT-scans are notably more expensive, slower, and significantly less convenient to use (Brink, Steenbakkers, Holla et al., 2019) . The visual challenge for the human eye combined with inefficient alternative diagnostic procedures gives rise to the question how radiologists can be assisted in assessing hand and wrist radiographs.

1.2 Motivation

Implementing a fracture detection system for conventional wrist radiographs can have multiple societal implications. First, the application of automated fracture detection can contribute to the efficiency with which distal radius and ulna fractures are diagnosed. Currently, fractures in hand and wrist radiographs are visually inspected by radiologists which is expensive in both terms of time and money. Second, it can improve the quality of healthcare concerning hand and wrist fractures. The fracture detection tool can act as a "second opinion" pointing out fractures that radiologists would have overlooked. This can help train less experienced clinicians (Gan et al., 2019). Besides prevention of overlooking fractures, more accurate diagnosis will also lead to less over-treatment. Society loses 256 million euros a year by inability to work as a consequence of hand and wrist fractures (Putter, van Beeck, Polinder et al., 2016). Third, labelling medical data is a problem often encountered in medical image tasks (Seifert et al., 2010). A fracture detection tool could contribute to (pre-)annotating data for other clinical or scientific purposes.

Besides societal relevance, continued study on deep learning for fracture detection in wrist radiographs also has scientific significance. Current deep learning models for fracture detection in the radius and ulna fall short in two areas, creating a gap between academics and practice. The first drawback is generalizability of the deep learning model. Most algorithms for medical imaging are trained on data from one hospital. Using only once data source can result in overfitting on radiographs from one specific hospital. It is not yet clear how the methods presented in the literature perform on radiographs from different hospitals creating a caveat in terms of generalizability. Second, current fracture localization methods are aspecific compared to the diagnostic description of radiologists. Kim & MacKinnon (2017) merely focussed on classifying radiographs leaving out fracture localization tasks. Subsequently, Thian et al. (2019) added a bounding box to the output of the model to visually localize the fracture, and Raisuddin et al. (2020) highlighted the fractured area with a gradient class activation mapping (Grad-CAM). Unfortunately, both bounding boxes and Grad-CAM heatmaps are imprecise means of fracture localization as they highlight redundant pixels around the fracture. The absence of highlighting the exact fracture location leaves the radiologist unable to adequately assess the output of the model when put into practice. To bridge this gap, this thesis aims to develop a fracture detection tool that overcomes these limitations, setting the next step towards a clinically usable fracture detection tool.

1.3 Research questions

Training an accurate fracture detection model requires adequate resolution in the fractured area and minimal redundant information around the fracture to prevent overfitting. To achieve this, the radius and ulna will be cropped out from the radiograph before being fed to the fracture detection network. The ability to automate this cropping process with deep learning will be addressed through the first research question:

Research question1: *"How can deep neural networks localize distal radius and ulna on conventional hand and wrist radiographs?"*

After locating the bone of interest, the goal is to classify which radiographs contain a fracture and specify in the model's output where the fracture is located without highlighting much adjacent bone or tissue. Hence, the second research question is formulated:

Research question 2: *"How can deep convolutional neural networks accurately detect and localize fractures in radiographs from distal radius and the ulna?"*

To answer the second research question, two subquestions are defined. A radiograph diagnosis often includes projections from multiple angles, but the literature remains unclear about what input yields the best performance. Therefore, the first subquestion considers a comparison between classification performance given a frontal and lateral input for the model. This subquestion will help illustrate to what extent a convolutional neural network (CNN) can detect fractures in radiographs from distal radius and the ulna given different input angles.
Subquestion 1: *"Which input projection yield the best results from fracture detection in the radius and ulna on conventional radiographs?"*

After the model finds a fracture, it needs to highlight the abnormal region so radiologists can visually validate the extent in which the model can accurately localized the fracture. To assess the localization performance of the model, adequate techniques for concisely visualizing the model's rationale need to be investigated. Hence, the second subquestion is formulated as:
Subquestion 2: *"How can detected fractures be best visualized to explain the model output to the radiologist?"*

1.4 Summary of contributions

This research contributed to the body of literature concerning deep learning for bone localization, fracture detection and fracture localization. The findings show that a single shot detector such as YOLOv5 can be employed as regional proposal network for further fracture classification. Moreover, it presents the multi-institutional performance for several deep CNN architectures gaining insight in the generalizability of models trained on data from only one hospital. Finally, this research illustrates how detected fractures can be localized highlighting less excessive information than methods presented in the current literature. Besides contributions in the field of deep learning, this study created multiple datasets for both bone detection and fracture classification in radiographs capturing the radius and ulna. Future research at RadboudUMC and Jeroen Bosch Ziekenhuis can be accelerated using the datasets produced by this thesis.

2 Related work

2.1 Radiographs

Radiography is a procedure where a beam of electromagnetic radiation, called x-ray waves, is passed through the body. As the x-rays travel through the body, it encounters various tissues that influence how much of the x-ray wave reaches the digital receptors in the x-ray machine that produce the radiograph. Soft tissue barely intervenes with the transit of x-rays through the body casting a dark or black structure on the radiograph. Dense tissue like bones, however, block or scatter the electromagnetic waves casting a white 'shadow' on the receptors. This 'shadow' shows the bones on the radiograph in a light grey colour.

2.2 Deep learning for bone localization

The radiographs used for diagnosing fractures in the radius and ulna often include multiple other bones. Localizing and cropping the radius or ulna from the radiograph has multiple advantages for subsequent fracture detection. First, CNN architectures include down-sampling which will cause loss of resolution as it travels through the network. The loss of resolution leaves less information for the model to extract relevant features for frac ture detection. When the full image is cropped into only the bone of interest, the resolution is preserved, and the model has all the information (resolution) available to learn relevant features that imply a possible fracture. Second, using the whole radiograph means including redundant space around the fractured area. This can cause the model to overfit on uninformative noise. Third, the input size for the classification model needs to be uniform. Cropping out the radius or ulna can help create this uniform input size. Particularly since some images in the dataset contain both frontal and lateral projections in one landscape image this technique proves useful. Fourth, according to the study of Gan et al. (2019), model efficiency can be greatly improved by feeding the fracture detection CNN cropped images containing the region of interest. Since the input size is smaller, the CNN has less pixels to convolve over, shortening processing time. Object detection algorithms are therefore popular to include in the deep learning pipeline.
An approach commonly adopted for localizing (areas of the) bones is the Faster R-CNN architecture (Kotika, Demircioglu, Kim et al., 2018; Yahalomi et al., 2018; Gan et al., 2019; Thian et al., 2019; Qi, Zhao, Shi et al., 2020). This method includes an additional network that predicts potential regions of interest. This region proposal network works in tandem with the CNN classifier that assigns a class to the objects detected in the proposed region. Alternatively, bones can be detected using radom forrest regression voting or one-stage object detector (MobileNet) (Ebsim et al., 2019; Krogue, Cheng, Hwang et al., 2019). A comparative study by Westerberg (2020) tested two one-stage detectors (YOLOv3 & RetinaNet) and the most recent version of the region proposal network family, the Mask R-CNN. The results show that based on precision, recall and F1-score, the one-stage object detectors equally outperform the Mask R-CNN. Besides better accuracy, Benjdira, Khursheed, Koubaa, et al. (2019) also report that a one-stage object detector, like YOLO, is more efficient than a two staged regional proposal architecture scoring an average prediction time of 0.057 milliseconds versus 1.39 seconds respectively. These findings show that despite the frequent use of regional proposal networks, one-stage object detectors present an unexplored an potentially attractive alternative.

2.3 Deep learning for fracture detection in hand and wrist area

Deep learning models have been employed for fracture detection on radiograph of many different parts of the body. For example, Cheng, Ho, Lee et al. (2019) used a deep CNN to detect fractures on pelvic radiographs, Guan, Yao, Zhang et al. (2019) applied an altered CNN architecture on upper arm radiographs, and Yubin, Zhao, Shi et al. (2020) used a Faster R-CNN architecture to detect and classify fractures in radiographs containing the upper leg and knee.
Compared to other medical image analysis tasks, CNN-based fracture detecting in the radius and ulna is underrepresented in the literature. Olczak et al. (2017) were one of the first to develop a machine learning model to classify hand and wrist radiographs. They used a VGG-16 architecture to classify the radiographs achieving an accuracy of 83%. Multiple studies followed using alternative deep learning architectures. Inception networks have been one of the candidates thoroughly tested. Kim & MacKinnon (2017) retrained the top layer of an Inception-v3 network that was pre-trained on non-medical images, Gan et al. (2019) used the subsequent Inception-v4 model to classify fractures and Thian et al. (2019) used the Inception-ResNet-v2 architecture to detect and localize the fracture. CNN architectures like these are often combined with a regional proposal network (R-CNN) to mitigate redundant areas of the radiograph. Other deep learning architectures used for fracture detection on wrist radiographs include U-net and Squeeze-and-Excitation networks (Lindsey et al., 2018; Raisuddin et al., 2020). It is noteworthy that all the studies mentioned above used data from a single hospital. Little is known about the generalizability of these models, other than that deep learning models tend to decline in performance when tested on input from a different data source (Lian, Nguyen & Jiang, 2020). This highlights the need for further research which includes data from multiple sources putting these architectures to the test on less homogeneous input. Furthermore, many other architectures are yet to be tested on fracture detection in the radius and ulna. Popular networks like DenseNets and ResNet models can still be explored.

2.4 Frontal versus Lateral projections

Radiographs of the wrist can be made in various angles of which frontal and lateral angles are most common. Kiuru et al. (2004) show that frontal radiographs alone are often not sufficient in clearly capturing fractures. In 14% of their patients the frontal view did not show the fracture, while additional CT scans showed one was present. Besides, the frontal radiographs let radiologists believe that there was a fracture present when there was not in 37% of their patients. Medoff (2005) justifiably claims that the lateral angle is integral part of complete examination of wrist radiographs. Figure 1 shows the distinction between a healthy bone (top left and bottom left) and a fractured bone (top right and bottom right) from both frontal and lateral views. The red dotted line in illustrates a Barton fracture where the value of using the lateral projection becomes apparent. In a Barton fracture the top part of the distal radius stays intact, while the bottom part is separated from the rest of the bone. The lateral projection clearly shows dislocation of the radius relative to the carpal bones, whereas this is more difficult to spot on the frontal projection.

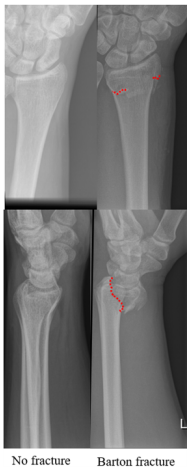


Figure 1
Barton fracture frontal & lateral projection.

Besides relevance for clinicians, studies on fracture detection in the radius and ulna also tend to use both projections (Yahalomi, Chernofsky & Werman, 2018; Blüthgen et al., 2020; Raisuddin et al., 2020). Notably is the study of Thian et al. (2019), which reports similar sensitivity scores for frontal and lateral inputs, but higher specificity scores for lateral inputs. Especially in paediatric radiographs and when casts are present, the lateral projection tends to outperform the frontal projection considerably. Based on visual relevance and superior specificity scores, one could argue that including both frontal and lateral projections yields the best results. However, the study of Yahalomi, Chernofsky & Werman (2018) contradicts this hypothesis claiming that "we found that the object detection neural network had better results when trained only on frontal images instead of frontal and lateral images together as the frontal and lateral images are substantially different. Another reason is in some of the lateral images the fracture does not appear because the other bones hide the fracture" (p.4). This highlights the need for additional research about which views yield the best result for the fracture detection and localization algorithms.

2.5 Fracture localization and decision-making transparency for deep CNNs

Although the performance of the deep neural networks is promising, the decision-making transparency in these models is unsatisfactory as these algorithms are not able to highlight the region in the radiograph that was most influential to the model's classification. Literature concerning fracture localization presents three methods for providing verifiable visual evidence for the fracture detection models. The first approach includes the use of bounding boxes (Figure 2, middle) to surround the fractured area in the radiograph (Yahalomi et al., 2018; Thian et al., 2019). The bounding boxes demonstrate the region on which the model based its prediction.
Figure 2 shows how the bounding box covers an area larger than the actual abnormality in the radiograph itself, highlighted by the red dotted line. Hence, validating the model's rationale is sub-optimal . This problem might worsen when fractures are aligned diagonally, enlarging the making the bounding even larger. The second approach therefore uses Gradient Class Activation Maps (Grad-CAM) to generate a heatmap (Figure 2, left). The heatmaps provide insight in which pixels in the input image contributed most to the decision made by the model (Krogue et al., 2019; Gupta, Demirer, Bigelow, et al., 2020; Raisuddin et al., 2020). However, upon visual inspection of the heatmaps, it was noted that the Grad-CAM heatmaps still include a significant region covering adjacent bone and tissue. The main reason for this aspecificity can be found in subsequent improvement studies claiming that the Grad-CAM algorithm is imprecise in covering a class region in an image as well as unable to capture objects in its entirety (Chattopadhay et al., 2018; Omeiza et al., 2019).
The third method to provide visual verifiable evidence for fracture locations is through dense conditional probability maps (DCPM). This approach also yields a heatmap, but in the DCPM heatmap "each pixel location represents the confidence that the corresponding pixel location in the input radiograph is part of a given fracture" (Lindsey et al., 2018, p.11593). Although his approach might yield accurate fracture localization, the method has two important drawbacks. First, annotation of pixels requires classifying every pixel in the image, this can be a challenging and time-consuming task. Second, the method is not suitable for displaced fractures as there is ambiguity in which class to give the pixels between the displaced bones. Being consistent in annotating different types of fractures can therefore be challenging. Reviewing these approaches, it becomes

clear that they are effective methods. However, they still require a trade-off between accurate localization and efficiency. More research is necessary to provide a method in which this trade-off is minimized.

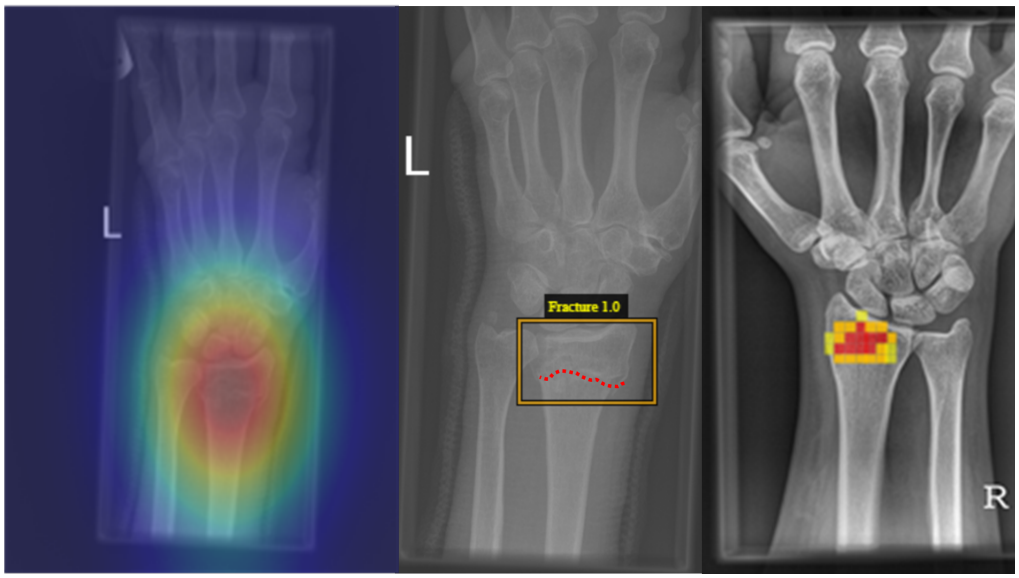


Figure 2
Fracture localization methods recreated. Left: Grad-CAM heatmaps adapted from Cheng et al. (2019). Middle: bounding boxes adapted from Yahalomi et al. (2018). Right: dense conditional probability maps adapted from Blütghen et al. (2020)).

3 Methods

The following sections elaborate and motivate the methods used to answer the research questions. The chapter will be split up into three parts addressing object detection, fracture detection and fracture localization.

3.1 Methods for Object detection

Research question one requires localizing the radius and ulna on the radiographs provided in the dataset. The related work on bone localization illustrated two suitable methods for this task, regional proposal networks and one stage object detectors. A recent study by Long, Deng, Wang, et al. (2020) provides an overview of the performance of the current state of the art object detectors. Even though their data did not consist of radiographic images, their findings support the results of Westerberg (2020) that one stage object detectors outperform other object detection architectures in both mean average precision and efficiency measured in frames per second (FPS indicates how many seconds it takes to analyse one image) on radiographic images. Taking these results into consideration, this study employs the most recent version of the one stage object detector, YOLOv5 version 5.0 (Jocher et al., 2020).

3.1.1 YOLOv5

The YOLO (You Only Look Once) family is popular amongst the object detectors because it can achieve high accuracy while keeping computation times low. YOLOv5 divides the image into an S x S grid and uses a single neural network to predict multiple bounding boxes and probabilities for each cell in the grid. After filtering out the bounding boxes with low probability, non max suppression is applied to select the bounding box with the highest combination of confidence and intersection over union (IoU). The name You Only Look Once was coined because the network looks at the entire image and only needs one forward pass to make predictions. This method contrasts that of regional proposal networks which perform detections on various region proposals and thus end up performing several predictions multiple times for various regions in an image.
In addition, the algorithm incorporates more contextual information, an important point for object detection. Where R-CNNs only work with the information in the proposed region, YOLO sees the entire images and therefore implicitly encodes information about classes and their appearances. As a result, less background patches are misclassified as objects. .

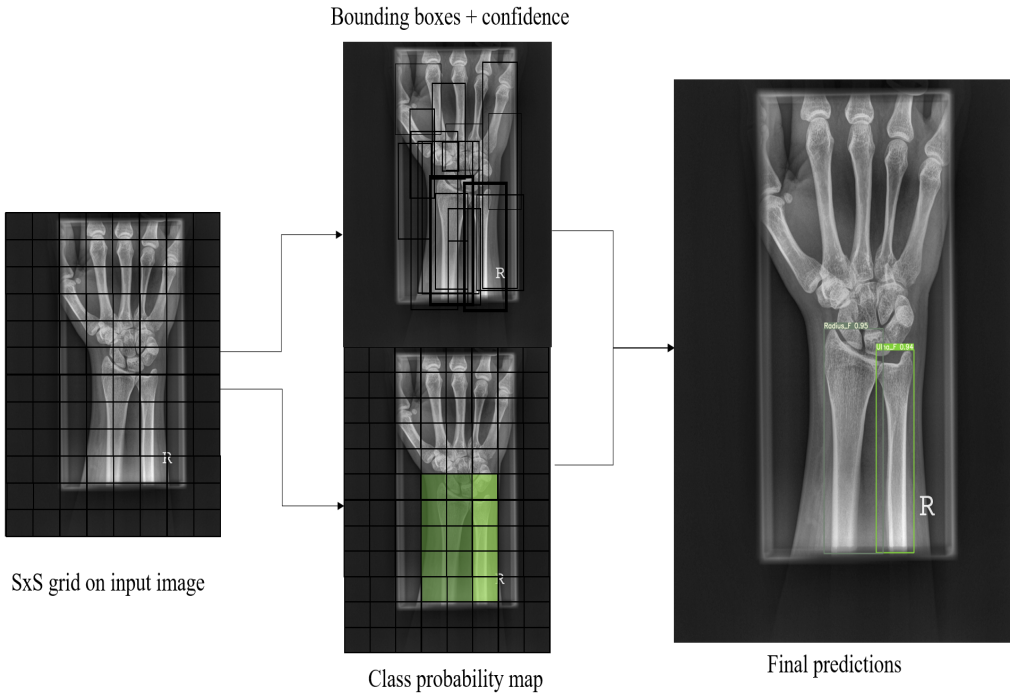


Figure 3
YOLO prediction steps: the image gets split up into a SxS grid (left) The multiple bounding boxes are predicted (middle top) and a class probability map is created (middle bottom). Non max suppression is used to extract the final bounding boxes (right).

3.2 Methods for fracture detection

This thesis compares three deep CNN architectures for the fracture detection task. The following sections will provide a brief summary on each of the architectures emphasizing how they address challenges encountered deep learning architectures.

3.2.1 Inception architecture

The Inception architecture was created by Szegedy et al. (2015) to address three problems of deep convolutional neural networks. First, information distribution in images needs to correspond to the kernel sizes in the network. When information is distributed globally, large filters are preferred, and vice versa if the information is distributed more locally. Furthermore, deep CNNs showed to be prone to overfitting and very demanding in terms of computational cost. The solution provided by this architecture is to have (1) multiple kernel sizes to address the information distribution challenge, and (2) introduce 1x1 convolutions to reduce the number of channels passed through the network, thus reducing dimensionality. The part of the network where multiple kernel sizes and 1x1 convolutions are applied is referenced to as the 'inception module'. This module has been refined in later versions to gain efficiency and prevent information loss throughout the network (Szegedy et al., 2016). Moreover, this architecture includes auxiliary classifiers in the middle of the network where an intermediate loss is calculated to prevent the vanishing gradients problem. This problem occurs when networks get deeper. During backpropagation, the gradients of the first couple of layers in a deep network get very small because the partial derivative is calculated repeatedly before reaching the early layers in the network. Since the gradients are responsible for updating the weights, tiny gradients hardly make changes in the weights of the network, and therefore the network barely learns. Inception resolves this issue by making total loss of the network is calculated by a weighted sum of these auxiliary classifiers combined with the model's final loss. The auxiliary classifiers help amplify the gradient signal that gets backpropagated through the network therefore preventing vanishing gradients .

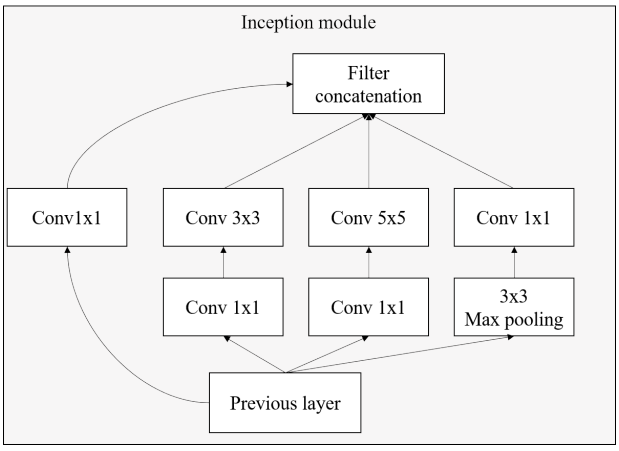


Figure 4
Inception module illustrating the use of multiple filter sizes. Adapted from Szegedy et al. (2015)

.
.

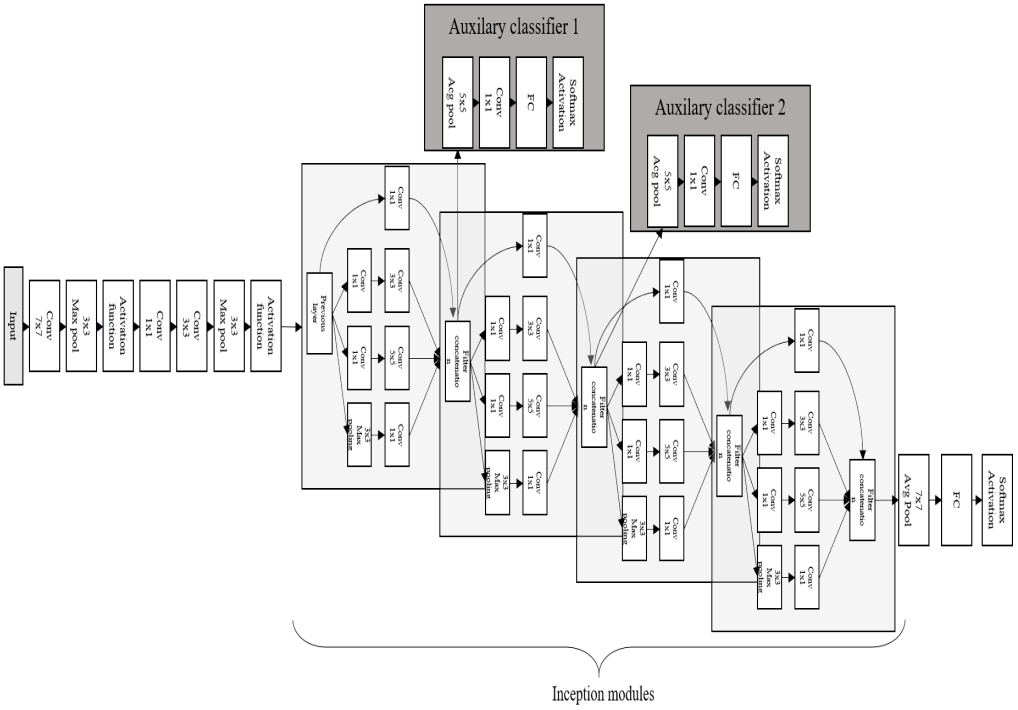


Figure 5
Inception architecture consisting of stacked inception modules and two auxillary classifiers to ampify the loss signal. Adapted from Szegedy et al. (2015)

3.2.2 ResNet architecture

As deep learning started to gain momentum in the field of image classification, deeper networks seemed to yield better results. However, researchers discovered that after a certain depth, the performance of deep CNNs started to degrade. In a response to this problem, He et al. (2015) developed ResNets. The ResNet architecture allows building significantly deeper networks by employing skip connections between the layers in the network. These connections directly combine the output from a previous layer with the input for a layer further on in the network. The network combines the feature maps through summation and handles difference in feature map size by applying 1x1 convolutions or zero padding. The skip connections have proven to help build deeper, better performing networks, without the drawbacks of the performance degradation problem. This study will employ an improved version of the ResNet architecture called ResNext that combines the concept of skip connections with the concept of cardinality and aggregated transformations. Like the inner works of a simple artificial neuron, the input is split up into C parts that each undergo a transformation and are then aggregated again. In this case C represents the cardinality specified for the model. After aggregation, the skip connection allows the original input to be combined with the output of the aggregated transformations. Xie et al. (2017) showed that the combination of skip connections, cardinality and aggregated transformation yielded better results than its counterpart architecture ResNet.

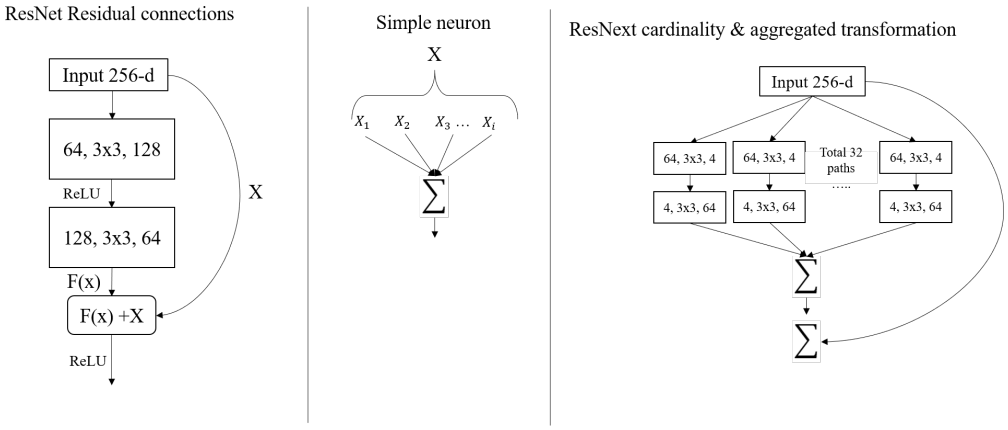


Figure 6
Residual connection (left) and aggregated transformation in simple neuron and in cardinality of the ResNext architecture (middle & right)

3.2.3 DenseNet architecture

The related work showed that CNN architectures have been used for fracture detection in the radius and ulna on conventional radiographs. Models like VGG-16 (Olczak, 2017; Yahalomi et al., 2019), inception architectures (Kim MacKinnon, 2017; Gan et al., 2019) and residual networks (Thian et al., 2019) have been explored. However, little is known about the suitability of densely connected CNNs, also known as DenseNets. This CNN-architecture provides an attractive alternative compared to the previously tested models because it has an improved information flow between the layers while using fewer parameters. This combination results in both high performance and efficiency (Huang, Liu, van der Maarten et al. (2017). The following section will discuss this architecture in more detail.

Huang, Liu, van der Maarten et al. (2017) developed DenseNets to maintain the information flow in deep convolutional neural networks. Good information flow between layers is necessary because it prevents the vanishing gradient problem described earlier. Like ResNets, DenseNets prevent vanishing gradients by creating multiple short paths that share information from earlier layers to layers further on in the network. The difference between the two architectures lies in how the information sharing between layers happens. Unlike a ResNet that the skip connection to pass information from one layer to another, a DenseNet connects information of all preceding layers to the input for the next layer. It does so by dividing the network into multiple dense blocks in which the size of the feature maps remains the same. Within this block, each layer receives the outputs from each earlier layer in the block. After passing through all layers in the block, its feature maps in it are concatenated and passed on to all the blocks that follow in the network. Through this way, a connection is established between all layers in the network. The transition layers that follow each dense block apply the convolutional and pooling operations through which the input is down sampled before it is passed on to the next dense block.

Tang, Tang, Peng, et al. (2020) studied a variety of CNN architectures, including DenseNets on a data set containing conventional chest radiographs. Their findings show that pre-trained DenseNets perform on par with pre-trained inception or residual

CNN architectures in classifying abnormalities in the chest. However, when learning from scratch, DenseNets outperformed all models in terms of accuracy, F1-score, sensitivity, and specificity. Considering the advantageous information flow, prediction performance and efficiency this study will compare the performance of several DenseNet models to the performance of the most recent ResNet and Inception models (ResNext50 and Inceptionv3).

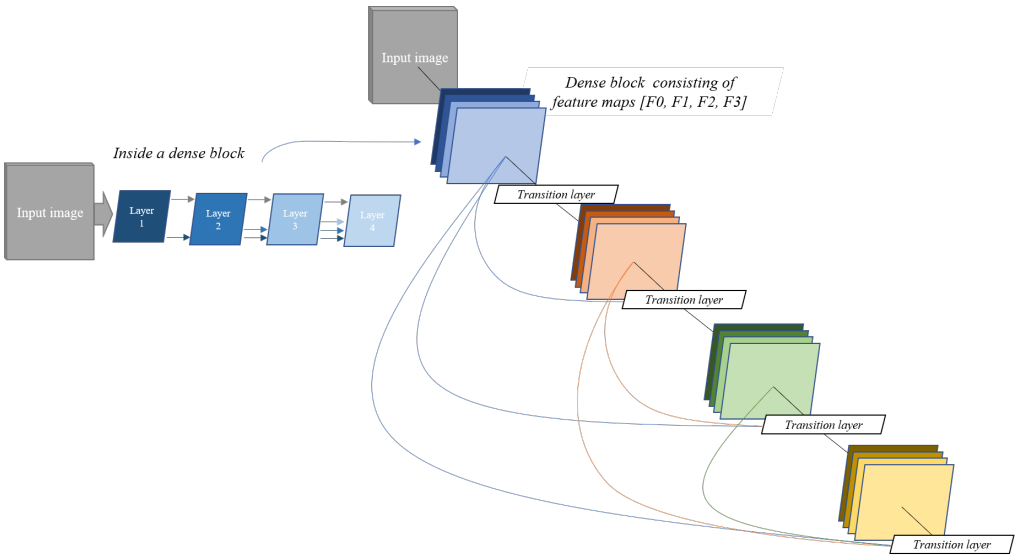


Figure 7
Information sharing within and between dense blocks in the DenseNet architecture. Adapted from Huang et al. (2018)

3.2.4 Anti-aliassing

Both the ResNext and DenseNet implementations in this research are anti-aliased versions of the architecture. Zhang (2019) showed that traditional CNNs are not shift-invariant, resulting in significantly different outputs when the same images was shifted a bit. Different results are a consequence of striding combined with a max pooling operation. Figure 8 illustrates how this process generates different outputs when the image is shifted. Adding a blurring step before and using average pooling instead, significantly reduces the performance decrease when the image is shifted. Zhang (2019) showed that anti-aliased versions of ResNet and DenseNet architectures not only improve in terms of shift invariance but also in terms of accuracy.

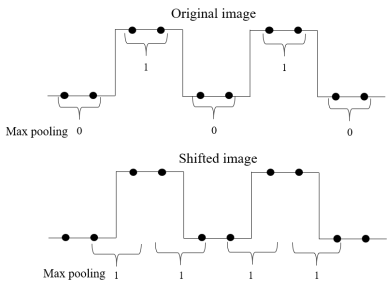


Figure 8
Max pooling operation combined with a stride of 2 on an an original image and a shifted image. The height of the line is a oversimplified representation of the pixel values. Adapted from Zhang (2019).

3.3 Methods for fracture localization

The second research question formulates accurate localization of fractures without highlighting too much redundant bone or tissue. Previous studies tried to achieve concise localization with methods that suffer from a trade-off between time consuming annotation (DCPM extracted from the U-Net) and aspecific visualization (bounding boxes and Grad-CAM). This study will try to find a balance between efficiency and accurate localization by using more advanced techniques in the field of class activation maps. Three class activation methods that are employed in this thesis will be summarized in the sections below.

3.3.1 Grad-CAM++

Gradient class activation maps (or Grad-CAM) help visualize the models rational after forward passing an image through the network. It combines gradient information of the predicted class with guided backpropagation to extract the influence of each feature map for the classified output. The algorithm then applies a ReLU function to only preserve only the features that had a positive influence on the classification. The output is subsequently transformed into a heatmap overlay on the image where the most important features light up more than other (Selvaraju et al., 2016). The more sophisticated successor of Grad-CAM, called GradCAM++, developed improved this method by making a more generalized formulation of the computation used in Grad-CAM. The improvements allowed the heatmaps to cover more of the classified object (Chattopadhyay et al., 2019).

3.3.2 SmoothGrad-CAM++

The SmoothGrad algorithm developed by Smilkov, Throat, Kim et al. (2017) similarly uses the gradients of the CNN's class activation scores to visualize which pixels contributed most to the output of the classification network. However, the fluctuation in these gradients creates noisy visualization. Smilkov et al. (2017) tackled this problem by repeatedly adding gaussian noise to the image and averaging the resulting sensitivity maps. This process effectively smooths out the fluctuation in the gradients that caused the noise. This resulted in a more accurate sensitivity map. Omeiza, Spekman, Cintas et al. (2019) combined the SmoothGrad technique with the Grad-CAM++ algorithm creating a more concise class activation map visualization. Compared to preceding CAM-techniques, SmoothGrad-CAM++ is reported to provide enhanced visual representation of relevant pixels for deep convolutional neural networks (Omeiza, Spekman, Cintas et al., 2019).

3.3.3 Score-CAM

The latest advancements in visualizing the logic behind the inferences of deep CNNs are published by Wang et al. (2020) that introduce Score-CAM. This method does not rely on the gradient information to determine the importance of the feature maps. Instead, it applies a two-stages process that first extracts and upsamples the activation maps with respect to the target label. In the second stage the algorithm uses the activation maps to mask the original image after which the masked image is propagated through the network. The forward-passing score on this masked image is used to generate a weight for the activation map which will be used to generate the heatmap. Hence the name 'ScoreCAM'. The second stage is repeated N times, where N corresponds to the

number of extracted activation maps. Wang et al. (2020) found that ScoreCAM outperformed Grad-CAM and Grad-CAM++ in quantifiable as well as qualitative measures. Subquestion two is devoted to investigating methods for localizing fractures without highlighting redundant bone. The methods described above provide solutions that allow to demonstrate the performance of the latest CAM methods for this task.

4 Experimental setup

4.1 Overall dataset

The data provided for this study contains 6947 images files from two institutions (6475 from Radboud UMC and 472 from Jeroen Bosch Ziekenhuis). The radiographs in the dataset were taken between 2000 and 2019. Figure 9 shows examples of radiographs containing frontal and lateral projections. Radiographs containing severe pathology (late staged tumours, intense arthritis etc.) were excluded from both datasets. The dysmorphological aspects in the radiograph made it hard to determine where the distal parts of the radius and ulna began or stopped. The pathology hindered accurate annotation of the bounding boxes that function as labels or crops for the classifiers

Figure 9
Example of a lateral radiograph (left) and a frontal radiograph (right)

4.2 Dataset curation

4.2.1 Dataset curation for object detector

The object detector was trained on a subset of the RadboudUMC archive consisting of 908 patients, 1949 studies making up 3350 manually annotated radiographs. The target labels were created by making tight crops around the radius and ulna on both frontal and lateral views. The following cases were excluded from the dataset; radiographs with a rotation of more than 45 degrees (as they might lead the algorithm to generating unnecessarily large bounding boxes on non-rotated images), radiographs containing osteosynthesis material (because the metal blocks the contours of the bones lying behind them making precise annotation difficult) and radiographs without the presence of a radius or ulna.

4.2.2 Dataset curation for fracture detection

The dataset for the fracture detection task contains radiology studies from 1448 patients resulting in 6475 radiographs. The target labels were established semi-automatically using a script that highlights regular expressions in radiology reports that indicate the presence or absence of a fracture. Additionally, the timeline of the diagnoses was checked to check for inconsistencies in the diagnoses. Furthermore, the researcher performed a visual inspection on the radiograph to exclude inconsistencies between the projections in the report and the actual radiographs. The crops for the classification dataset were manually annotated To illustrate the generalizability of the classifiers, the test set was created using radiographs from a different hospital (Jeroen Bosch Ziekenhuis). The same labelling procedure was applied for reports in the test set, however, the reports used in the test set concerned CT-scans. Brink et al. (2019) showed that sensitivity and specificity for fracture detection on conventional radiographs were 95% and 97% for the radius and 93% and 99% for the ulna, respectively. Both metrics rose to 100% when physicians used CT-scans to detect or rule out fractures. These findings underpin the value of using CT-scan reports as a reference to create trustworthy labels for the test set. After labelling the CT reports, all radiographs taken up to 4 weeks in advance were included to form the final test set.

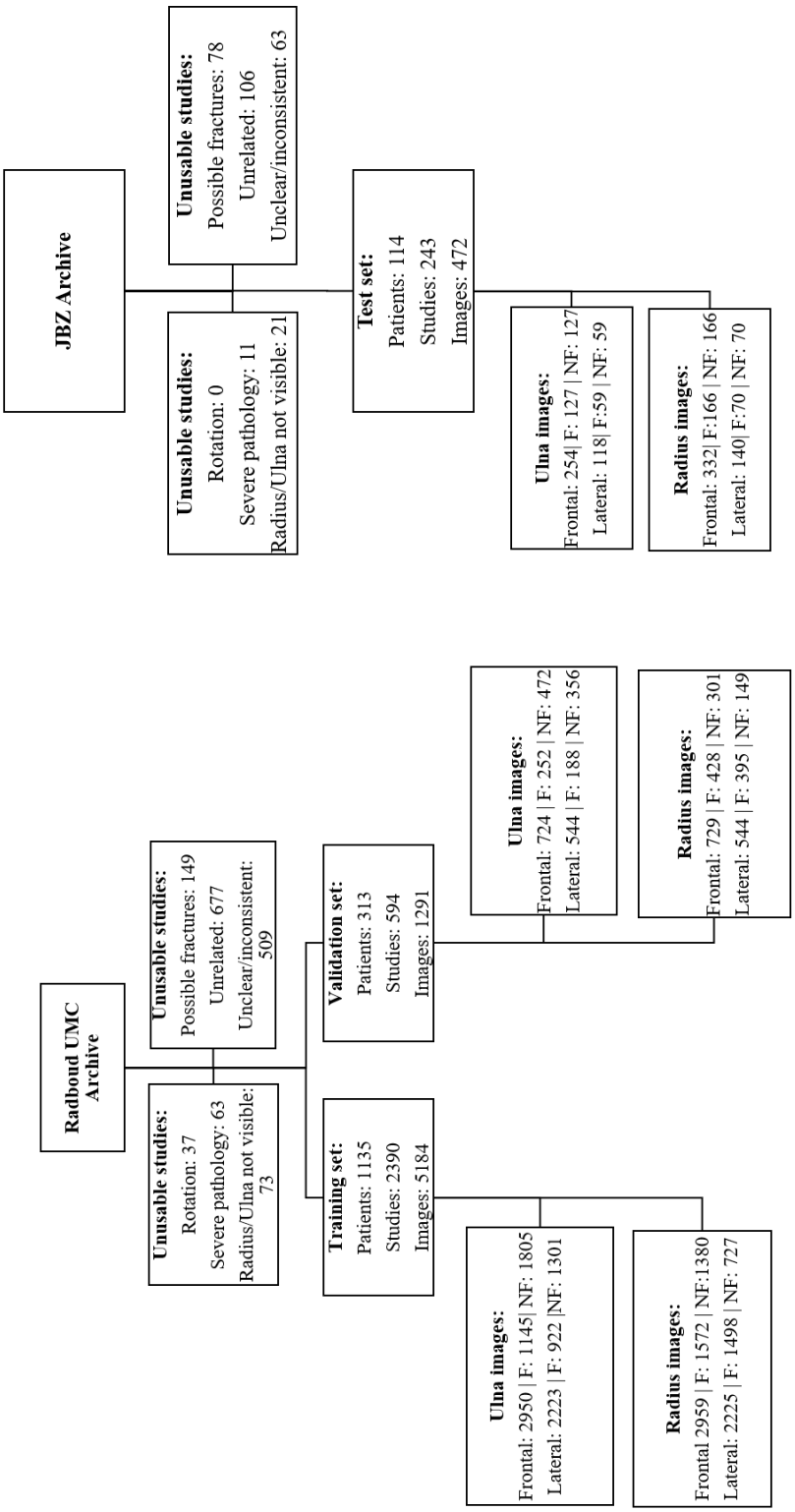


Figure 10
Dataset curation fracture detection dataset. F represents the number of images that include a fracture, and NF represents the number of images where no fracture is visible

4.3 Data pre-processing

4.3.1 Pre-processing for object detection

Two identical datasets were created, one containing the raw images and the other containing contrast enhanced images. No further deliberate pre-processing steps were undertaken augmentation steps included in the YOLOv5 architecture.

4.3.2 Pre-processing for fracture detection

After filtering out inconsistencies, additional pre-processing steps were applied before sending the images to the classifier. The following section will briefly elaborate on each step.

4.3.2.1 Rescale intensity and cropping the radiograph
First, the image intensity was rescaled to enhance the contrast in the image. Rescaling was applied using the range between the 3nd and 97th percentile to prevent outliers influencing the operation. Subsequently, the predicted bounding box coordinates were used to remove the redundant information in the radiograph whilst maintaining resolution.

4.3.2.2 Rotation, flipping and Gaussian noise
Next, the crops were augmented by applying both rotation and horizontal flipping with a 50% probability. These augmentation steps are meant to make the model more generalizable as radiographs are not taken with a consistent procedure. For the similar reasons gaussian noise is added to 50% of the images since some radiographs contain casts that might introduce noise over the fractures in the radius or ulna.

4.3.2.3 Normalization
Before the images are fed to the classifier, cropped images are reshaped to create a squared shape tensor which will be used as model input.

4.4 Evaluation metrics

4.4.1 Evaluation for object detectors

The performance concerning bone recognition will be evaluated using precision, recall and the mean average precision metric. The latter measures the average precision with which it classifies the detected bones averaged over all classes. Hence *mean* average precision. The mean average precision is formulated as:

$$mAP = \frac{\Sigma \, {}^{K}_{i=1} AP_i}{K}$$

The localization performance of the object detectors is evaluated using the intersection over union. This metric compares the bounding box of the ground truth with the predicted bounding box. Intersection over union is formulated as:

$$IoU = \frac{area\,(Bp \cap Bgt)}{area\,(Bp \cup Bgt)}$$

Through combining the mAP and the IoU, we can derive the metrics mAP0.5 and mAP0.5:0.95. The mAP0.5 metric refers to mAP using a IoU threshold of 0.5. The mAP 0.5 mainly indicates if the object detector can recognize the presence of an object in the given image. The mAP0.5:0.95 averages the mAP scores for an IoU thresholds between 0.5 and 0.95 with a step size of 0.05. This metric indicates to what extent the predicted bounding boxes correspond to the ground truth.

4.4.2 Evaluation for classification network

The classification performance will be measured using four metrics. The true positive rate (TPR, also known as recall or sensitivity) and the false positive rate (FPR or specificity) are measures widely used in medical studies to assess binary classification tasks. These measures are used to plot the receiver operating characteristics curve (ROC) and the corresponding area under the ROC curve (AUC). The AUC metric will provide a summary of the ROC curve giving insight in how good the model is able to distinguish between positive and negative samples. Furthermore, the model accuracy will be included creating an overall idea of how many cases it classified correctly in total.

$True\ positive\ rate = \frac{TP}{TP+FN}$ $False\ positive\ rate = \frac{FP}{TN+FP}$

4.4.3 Evaluation for fracture localization

The localization performance of the best model will be evaluated by plotting the heatmaps generated by experiment three into a matrix. This matrix will show an overview of how the Grad-CAM, SmoothGrad-CAM++ and ScoreCAM performed on the same images. These images will undergo a qualitative assessment looking at (1) the extent of highlighting excess information and (2) accurately localizing the fractured region.

4.5 Implementation details

This project will use Python 3.7 as the main programming language. PyTorch was used as the main deep learning framework for building both the object detector and the fracture detection model (Paszke, Gross, Chintala, et al. 2017). Additionally, the VGG image annotator software by Dutta, Gupta & Zisserman (2019) was used for creating annotations for the object detector. The contrast of the images was optimized using adaptive histogram equalization for accurate object annotation . The small YOLOv5 models were trained on 600 epochs using multiple samples fused into mosaic images.

The hue parameter was set to zero to ensure the model trained on grayscale images. Each epoch images were scaled with 10% to increase the robustness of the object detector. No further adjustments were made to the defaults settings of the YOLOv5 model version 5.0. The medium sized YOLOv5 models were trained using the same hyperparameter settings, however the number of epochs was decreased to 350. The fracture detection models were trained using the Adam optimizer, a batch size of 8 and a learning rate of 0.0001 decreasing after a minimum of 10 epochs. The models training sessions converged when the learning rate was lower than 0.000001 which occurred for most models between epoch 120 and epoch 200. Figure 11 and figure 12 illustrate the workflow for both object detection and classification tasks.

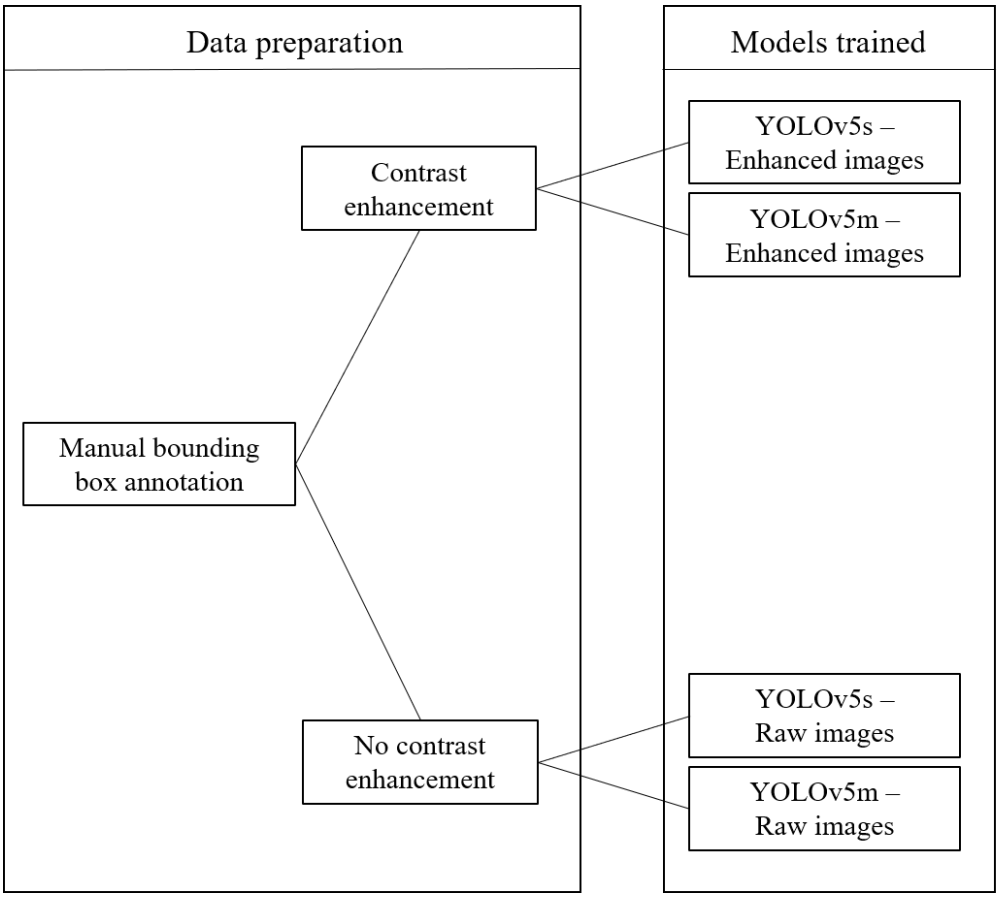


Figure 11
Workflow diagram for object detector

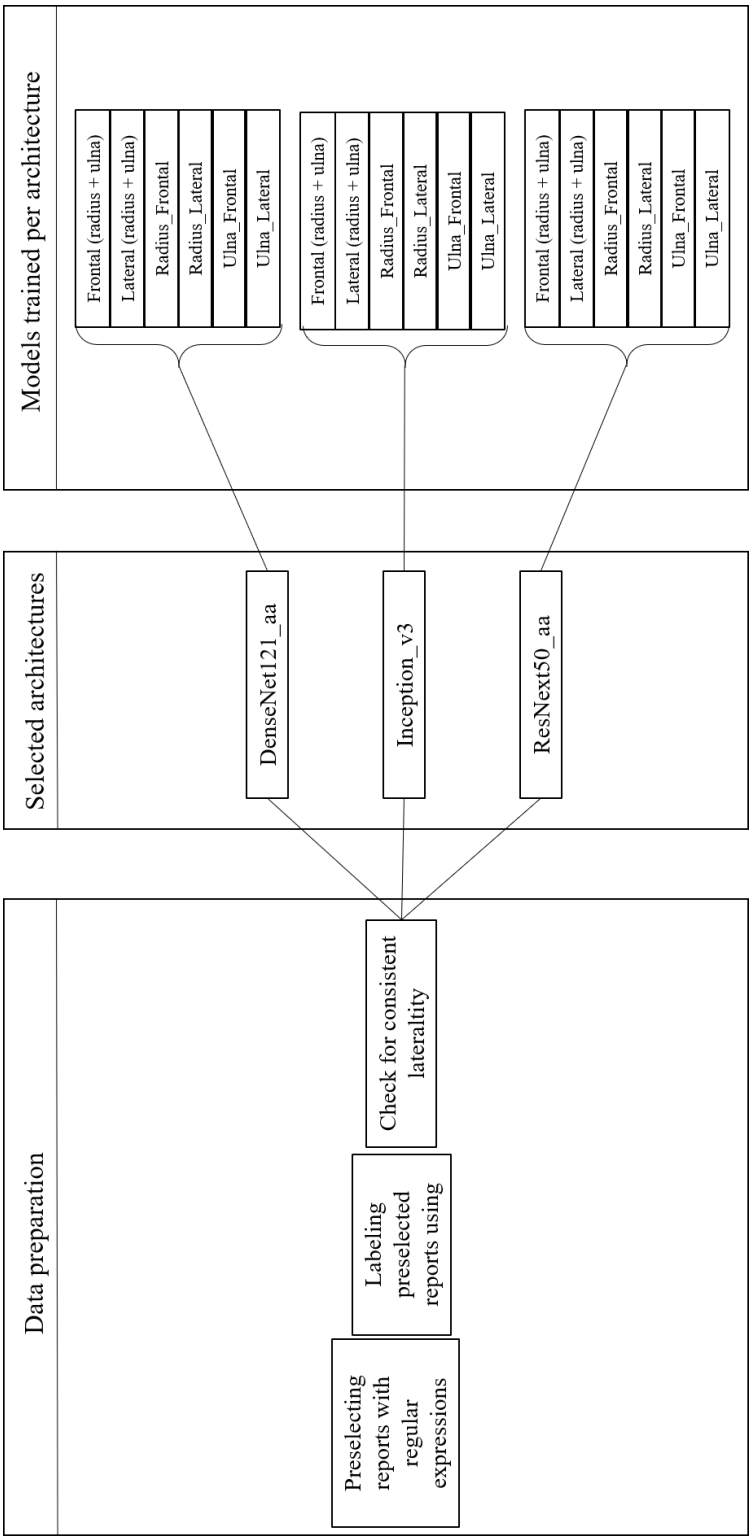


Figure 12
Workflow diagram classification models

5 Results

The following section will present the findings in three subsections. First, the object detector will be discussed,where the accuracy of the predicted bounding box for the distal part of the radius and ulna will be evaluated. Next, the performances of the fracture detection classifiers will be presented. Lastly, a qualitative comparison will be made between GradCAM, SmoothGradCAM++ and ScoreCAM to inspect how well the fracture is localized.

5.1 Object detection

The task for the object detector was to create bounding boxes for cropping for the radius and ulna on both AP and LAT projections. Two YOLOv5 versions (YOLOv5s & YOLOv5m) are tested on both raw and contrast enhanced images. The top section of Table 1 shows an overview of the performance on the test set on average and the bottom section shows the performance on the test set per bone per projection. The IoU threshold for correct the precision and recall was set on 0.5 and a the minimum confidence threshold for the predicted bounding boxes was set to 0.3.

Table 1
Average and bone-projection specific performances for YOLOv5s and YOLOv5m tested on raw images and contrast enhanced images.

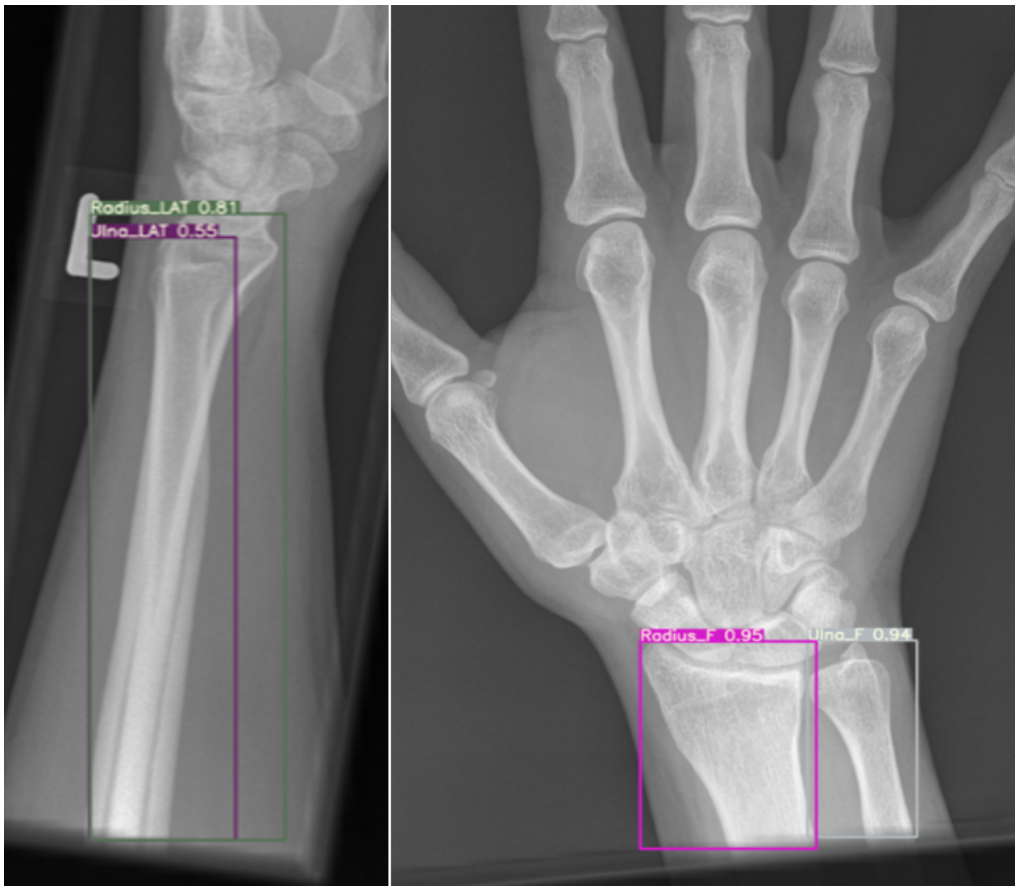
Average performance

Model	Data	Precision	Recall	mAP0.5	mAP0.5:0.95
YOLOvs	Enhanced	0.9782	0.972	0.9866	0.8148
YOLOvs	Raw	0.9798	0.9755	**0.9856**	0.8104
YOLOvm	Enhanced	0.9799	0.9741	0.9853	0.8141
YOLOvm	Raw	**0.9808**	**0.9798**	0.986	**0.834**

Performance per bone and projection

Model	Data	Metric	Radius AP	Ulna AP	Radius LAT	Ulna LAT
YOLOvs	Enhanced	Precision	0.991	0.986	0.975	0.956
		Recall	0.989	0.984	0.952	0.976
		mAP0.5	0.990	0.991	0.983	0.980
		mAP0.5:0.95	0.904	0.888	0.779	0.684
YOLOvs	Raw	Precision	0.989	0.985	0.976	0.97
		Recall	0.987	0.986	0.964	0.964
		mAP0.5	0.994	0.994	0.980	0.974
		mAP0.5:0.95	0.905	0.88	0.771	0.684
YOLOvm	Enhanced	Precision	0.983	0.984	0.976	0.951
		Recall	0.989	0.988	0.976	0.964
		mAP0.5	0.992	0.990	0.987	0.978
		mAP0.5:0.95	0.904	0.88	0.802	0.670
YOLOvm	Raw	Precision	0.991	0.991	0.976	0.958
		Recall	0.989	0.992	0.964	0.958
		mAP0.5	0.994	0.995	0.983	0.972
		mAP0.5:0.95	**0.910**	**0.893**	**0.811**	**0.731**

The predicted bounding boxes were visually inspected to assess the model's ability to create regions of interest for the fracture detection model. For this research it is relevant that the model captures the distal part of the radius and ulna, however, there is no convenient quantifiable method to consistently check if the model accurately this part of the bone. A visual inspection of the predicted results showed that the model missed the distal part of the bones in 1.7% of the images in the validation set and 1.47% in the test set.



5.2 Fracture detection

For the classification task six classifiers were trained per model architecture. The frontal model was trained on both the radius and ulna from the AP projection, the lateral model on both the radius and ulna on LAT projections, and the other four classifiers were trained on either the radius or the ulna given a specified projection (AP or LAT). The classifiers have been evaluated on four subsets of the test set, depending on what bone-projection combination they were trained on. The Frontal and Lateral classifiers were trained on both radius and ulna data, therefore they are each evaluated on two subsets. Table 2 shows the results of the best performing architectures on CT-referenced data from the Jeroen Bosch Ziekenhuis hospital. Compared to the performances on the validation set from the RadboudUMC, the results on the test set from the Jeroen Bosch Ziekenhuis are notably lower. These findings indicate that the generalizability of

classifiers trained on data from only one hospital is sub-optimal. Appendix A provides an overview of all the experiments that have been conducted, including the experiments on the validation set.
The rows highlighted with grey contain the best performing models on the respective subset of the test data. The Inception architecture yield the highest AUC in three of the four subsets. Only on the lateral projection of the radius DenseNet121_aa yielded higher AUC scores.

Table 2
Classifiers performances per subset of the CT-referenced testing data from Jeroen Bosch Ziekenhuis measured in sensitivity, specificity and AUC.

		Subset Radius_F (*N=166*)		
		Sensitivity	**Specificity**	**AUC**
DenseNet	Frontal model (radius + ulna)	0.7716	0.8588	0.9164
	Radius_Frontal	0.8881	0.8254	0.9180
Inception	Frontal model (radius + ulna)	0.8162	0.8707	0.9240
	Radius_Frontal	0.7962	0.9029	**0.9288**
ResNext50	Frontal model (radius + ulna)	0.7722	0.8289	0.8732
	Radius_Frontal	0.7647	0.8704	0.868

		Subset Ulna_F (*N=127*)		
		Sensitivity	**Specificity**	**AUC**
DenseNet	Frontal model (radius + ulna)	0.8268	0.874	0.8957
	Ulna_Frontal	0.9153	0.8897	0.9291
Inception	Frontal model (radius + ulna)	0.875	0.7761	0.9246
	Ulna_Frontal	0.8667	0.9076	**0.9433**
ResNext50	Frontal model (radius + ulna)	0.8116	0.7586	0.8607
	Ulna_Frontal	0.8692	0.8226	0.9211

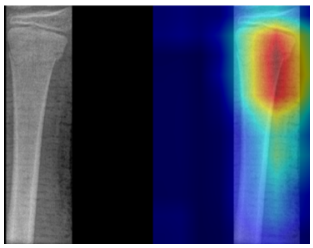
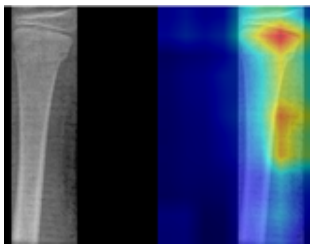
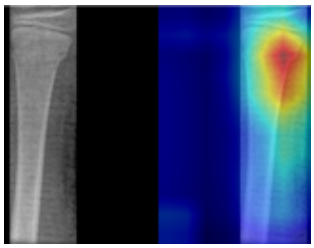
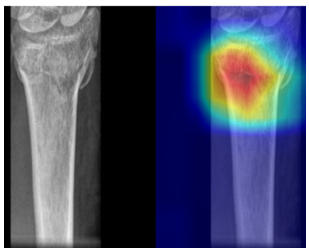
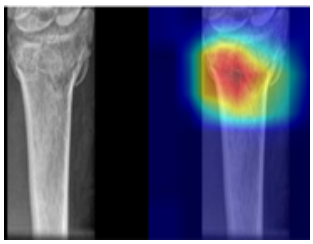
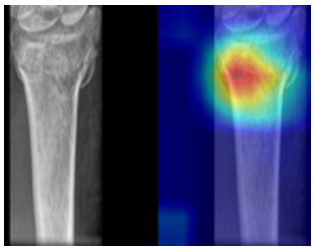
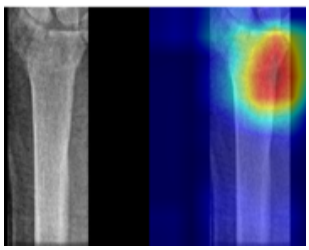
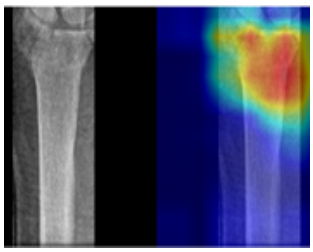
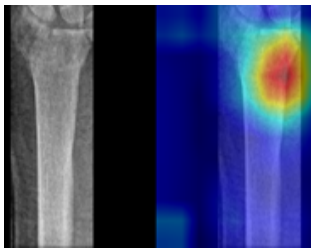
		Subset Radius_LAT *(N = 70)*		
		Sensitivity	**Specificity**	**AUC**
DenseNet	Lateral model (radius + ulna)	0.7465	0.6812	0.7677
	Radius_Lateral	0.7143	0.6164	**0.8868**
Inception	Lateral model (radius + ulna)	0.9155	0.5652	0.8157
	Radius_Lateral	0.7143	0.6883	0.7809
ResNext50	Lateral model (radius + ulna)	0.7188	0.6447	0.7593
	Radius_Lateral	0.6761	0.6667	0.7873

		Subset Ulna_LAT*(N = 59)*		
		Sensitivity	**Specificity**	**AUC**
DenseNet	Lateral model	0.7736	0.7231	0.7791
	Ulna_Lateral	0.8852	0.6842	0.8156
Inception	Lateral model	0.9796	0.6667	**0.9095**
	Ulna_Lateral	0.8871	0.7857	0.8967
ResNext50	Lateral model	0.7377	0.5263	0.7061
	Ulna_Lateral	0.9483	0.4667	0.7716

5.3 Fracture localization

For the fracture localization task, GradCAM, SmoothGradCAM++ and ScoreCAM were compared to highlight the most deterministic pixels from the activation layers. Table 3 shows a comparison illustrating examples of a fracture localization for all three methods. Compared to bounding box methods, the localization of ScoreCAM seems to be more specific, highlighting for example only the left part of the distal radius instead of the whole distal head (see Example 2 in Table 3). The ScoreCAM appears to yield the best results highlighting minimal redundant space around the fracture. In terms of localization, all methods are able to locate the abnormal region. Appendix B contains more samples comparing GradCAM++, SmoothGradCAM++ and ScoreCAM.

Table 3
Comparison between the three fracture localization methods GradCAM++, SmoothGradCAM++ and ScoreCAM.

Method	GradCAM++	SmoothGradCAM++	ScoreCAM
Example 1			
Example 2			
Example 3			

6 Discussion

This thesis used deep learning methods to address the tasks of bone localization, fracture detection and fracture localization. It employed a YOLOv5 algorithm to crop out the bones of interest. A consecutive CNN classifier comparison was conducted to investigate to what architecture performs best on fracture classification. Additionally, a qualitative evaluation between ScoreCAM, GradCAM++ and SmoothGradCAM++ demonstrated which method was most suitable for accurate fracture localization. The sections below will discuss the presented findings concerning these tasks.

6.1 interpretation of results

6.1.1 Object detection results

The findings show that the YOLOv5m model is able to localize and classify the radius and ulna correctly in almost all cases given the raw images. The precision and recall are both around 0.98 and its mAP 0.5 is at above 0.97 for all bone-projection combinations. These metrics indicate that the model is able to localize the bone, however, correctly capturing the entirety of the bone (with bounding boxes that have an IoU between 0.5 and 0.95) appears to be more challenging on lateral projections, especially for the ulna. A possible explanation for the lesser results on the lateral projection of the ulna might lie in the over-projection of the radial bone. Moreover, the lateral projection of the ulna was underrepresented in the object detection dataset which might have contributed to these findings. Despite not having identical metrics for object detection, an approximate comparison can be deducted from the paper of Gan et al. (2019). Their Faster R-CNN region proposal network achieved an average IoU of 87%, which is less than the 98% seen in the visual inspection. It appears that the YOLOv5m model can produce better results than the R-CNN alternatives, but more research needs to be conducted to establish a more detailed comparison.

6.1.2 Fracture detection results

Several conclusions can be deducted from the results regarding the fracture classification task. First, classifiers trained on frontal projections were found to achieve higher AUC scores than calssifiers trained on lateral images. Furthermore, classifiers specialized on a specific bone-projection combination (e.g. ulna frontal) tend to perform on par or outperform models trained on both bones in one projection. The Inception architecture performs best in most of the subsets from the testing data whereas the ResNext50 models do not meet the performances of the other two architectures. The significant difference in performance on the validation set from RadboudUMC and the test set from the CT-referenced radiographs from the Jeroen Bosch Ziekenhuis, demonstrate the relevance of using multiple data sources to train fracture detection models. Moreover, the findings are in line with Raisuddin et al. (2020) who showed that radiographs that correspond to CT-scans tend to be harder to classify for fracture detection algorithms.

6.2 Answers to the research questions

This thesis was split up into two main research questions. The first research question addressed the question: *"How can deep neural networks localize distal radius and ulna on conventional hand and wrist radiographs?"*. The object detection results show that the YOLOv5 algorithm was able to accurately locate the distal part of the radius and ulna in 98% of the cases. Although the model was not impeccable in capturing the entirety of the bone, its performance on detecting and classifying the distal part of the bones make it worthwhile to consider it as a regional proposal model for further fracture classification. The second research question was devoted to the detection and localization of the fractures. The question addressed: *"How can deep convolutional neural networks accurately detect and localize fractures in radiographs from distal radius and the ulna?"*. To answer the first part of the question, experiments were conducted using both projections (frontal and lateral) to see which view resulted in the best performing classifications. The findings show that the classifiers trained on frontal projections produce better results than their lateral counterparts. However, Table 2 also shows that specialization on either the radius or ulna given a projection results in equally good or better performance than training both bones on one projection. Subsequently, the high performing models were extended with three different fracture localization techniques to evaluate if a more precise fracture localization is achieved compared to their bounding box counterparts. The ScoreCAM method in particular achieves accurate localization, often highlighting only the fractured region which contrasts to surrounding the bounding box or GradCAM techniques that surround or highlight the whole breadth of the bone.

6.3 Limitations

An important limitation for the object detection model is that only one model was trained on both bones and both projections. It is possible that a lack of focus could have driven down the model's performance. Particularly localization of the ulna on lateral projections was sub-optimal, however, this could have also been cause by the over-projection caused by the radius.
There are three drawbacks with the classification task. First , the classifiers trained in this research use only *one* radiograph to determine the presence of a fracture, while radiologists tend to use multiple radiographs to form their diagnosis. Not combining the information of all radiographs in a series might have driven down the performance of the classifiers. Second, it is hard to compare the results of the fracture detection models with the performance of physicians or radiologists because the time constraints of this research did not leave room to collect data about human performance on the used data sets. Third, the integrity of both the training and test set labels have drawbacks. The labels of the training set were created by evaluating radiology studies, however, some of these studies included inconsistencies. For example, one radiologist mentioning a fracture in both radius and ulna while others mentioned only the radius fracture in the same week. Despite the extra precautions taken to combat inconsistent labels, there is no guarantee that all were filtered out. Furthermore, using CT-verified radiographs in the test set created a possible bias in the data because CT-scans are only made when there is a good reason to (oblique fractures, severe fractures, pathology etc.). Radiographs that required an additional CT-scan can therefore be more difficult to diagnose than the radiographs in the training and validation set which were explicit enough for direct diagnosis. Moreover, the size of the test set was constrained due to limited number of CT-scans available with a corresponding radiograph within the time frame of 28

days (the minimal time a fracture needs to completely heal). Combined with the low frequency of negative samples in the CT-verified radiographs, the size of the test set was significantly less than the validation set. Despite these drawbacks, the integrity of test set samples was outweighed the limited size and possible bias.

6.4 Societal impact

Implementing the deep learning methods presented by this thesis can significantly improve the efficiency with which distal radius and ulna fractures are diagnosed. The fracture detection pipeline can pre-classify incoming radiographs saving radiologist time in establishing their diagnosis. Moreover, the presented algorithms can serve as an extra check for physicians creating a second opinion that might contribute to the specificity and sensitivity with which fractures are detected by humans. In the process of conducting this research, several datasets were created concerning bone localization and fracture detection. Datasets like these can be used to as a foundation for more extensive research. I expect that, this enables future research will focus more on improving performance rather than creating annotations for their models producing a more accurate fracture detection mechanism.

6.5 Future work

Future research for object detectors might focus on creating a system that can handle osteosynthesis materials or severe pathology to increase the robustness of this component in the fracture detection pipeline. Furthermore, new research can study if object detectors specialized in only lateral or frontal projections yield better results than their generically trained competitors. The fracture detection classifiers in this study were specialized in the radius or ulna given a projection. Pan et al. (2019) showed that using an ensemble of models can yield better results. It would be interesting to see if these ensemble models also perform better in the domain of the radius and ulna given the fracture detection task. Also, the weights can of the current models can be used for transfer learning when training either these ensemble models or fracture detection models for other bones. This might be a better starting point than the default weights of an architecture. It would be interesting to see if radiologists misdiagnose less cases when assisted by a diagnostic algorithm like the one presented in this research. The results of this research showed that models trained on the data of one hospital do not generalize as well to data from other hospitals. Future studies concerning fracture detection models for the radius and ulna therefore should focus on establishing a diverse dataset sourced from different institutions. Finally, fractures might be localized with even more precision when a classifier is trained using a location scheme dividing the bone into multiple parts. Since radiologists specify the type and location of the fracture in their reports, a classifier using this scheme cannot 'cheat' by making its classification based on abnormalities that are in locations other than the fracture.

7 Conclusion

This thesis addressed the problem of misdiagnosis and over-treatment of distal radius and ulna fractures using deep learning. It provided a compartmentalized overview of a fracture detection tool demonstrating to what extent deep learning can be used to (1) localize the radius and ulna, (2) detect fractures in the radius and ulna, and (3) highlight the location of the fracture. Moreover, the findings reveal the generalizability of classifiers trained on data from a single institution. The datasets created in this thesis combined with the presented deep learning methods for bone localization, fracture detection and fracture localization contribute to the foundation of a clinically usable fracture detection mechanism.

References

Baig, M. (2017). A Review of Epidemiological Distribution of Different Types of Fractures in Paediatric Age. *Cureus*, *9*(8). https://doi.org/10.7759/cureus.1624

Balci, A., Basara, I., Çekdemir, E. Y., Tetik, F., Aktaş, G., Acarer, A., & Özaksoy, D. (2015). Wrist fractures: sensitivity of radiography, prevalence, and patterns in MDCT. *Emergency Radiology*, *22*(3), 251–256. https://doi.org/10.1007/s10140-014-1278-1

Benjdira, B., Khursheed, T., Koubaa, A., Ammar, A., & Ouni, K. (2019). Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3Benjdira, B., Khursheed, T., Koubaa, A., Ammar, A., &Ouni, K. (2019). Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3. 2019 1s. *2019 1st International Conference on Unmanned Vehicle Systems-Oman, UVS 2019*, 1–6.

Blüthgen, C., Becker, A. S., Vittoria de Martini, I., Meier, A., Martini, K., & Frauenfelder, T. (2020). Detection and localization of distal radius fractures: Deep learning system versus radiologists. *European Journal of Radiology*, *126*(February), 108925. https://doi.org/10.1016/j.ejrad.2020.108925

Brink, M., Steenbakkers, A., Holla, M., de Rooy, J., Cornelisse, S., Edwards, M. J., & Prokop, M. (2019). Single-shot CT after wrist trauma: impact on detection accuracy and treatment of fractures. *Skeletal Radiology*, *48*(6), 949–957. https://doi.org/10.1007/s00256-018-3097-z

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, *2018-Janua*, 839–847. https://doi.org/10.1109/WACV.2018.00097

Cheng, C. T., Ho, T. Y., Lee, T. Y., Chang, C. C., Chou, C. C., Chen, C. C., … Liao, C. H. (2019). Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *European Radiology*, *29*(10), 5469–5477. https://doi.org/10.1007/s00330-019-06167-y

Court-Brown, C. M., & Caesar, B. (2006). Epidemiology of adult fractures: A review. *Injury*, *37*(8), 691–697. https://doi.org/10.1016/j.injury.2006.04.130

De Putter, C. E., Van Beeck, E. F., Polinder, S., Panneman, M. J. M., Burdorf, A., Hovius, S. E. R., & Selles, R. W. (2016). Healthcare costs and productivity costs of hand and wrist injuries by external cause: A population-based study in working-age adults in the period 2008-2012. *Injury*, *47*(7), 1478–1482. https://doi.org/10.1016/j.injury.2016.04.041

Ebsim, R., Naqvi, J., & Cootes, T. F. (2019). Automatic detection of wrist fractures from posteroanterior and lateral radiographs: A deep learning-based approach.

In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11404 LNCS, pp. 114–125). Springer Verlag. https://doi.org/10.1007/978-3-030-11166-3_10

Freed, H. A., & Shields, N. N. (1984). Most frequently overlooked radiographically apparent fractures in a teaching hospital emergency department. *Annals of Emergency Medicine*, *13*(10), 900–904. https://doi.org/10.1016/S0196-0644(84)80666-6

Gan, K., Xu, D., Lin, Y., Shen, Y., Zhang, T., Hu, K., … Liu, Y. (2019). Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthopaedica*, *90*(4), 394–400. https://doi.org/10.1080/17453674.2019.1600125

Goldfarb, C. A., Yin, Y., Gilula, L. A., Fisher, A. J., & Boyer, M. I. (2001). Wrist fractures: What the clinician wants to know. *Radiology*, *219*(1), 11–28. https://doi.org/10.1148/radiology.219.1.r01ap1311

Guan, B., Yao, J., Zhang, G., & Wang, X. (2019). Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network. *Pattern Recognition Letters*, *125*, 521–526. https://doi.org/10.1016/j.patrec.2019.06.015

Gupta, V., Demirer, M., Bigelow, M., Yu, S. M., Yu, J. S., Prevedello, L. M., … Erdal, B. S. (2020). Using Transfer Learning and Class Activation Maps Supporting Detection and Localization of Femoral Fractures on Anteroposterior Radiographs. *Proceedings - International Symposium on Biomedical Imaging*, *2020-April*, 1526–1529. https://doi.org/10.1109/ISBI45749.2020.9098436

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition Kaiming. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.1002/chin.200650130

Huang, G., Liu, S., Maaten, L. Van Der, & Weinberger, K. Q. (2018). CondenseNet: An Efficient DenseNet Using Learned Group Convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2752–2761. https://doi.org/10.1109/CVPR.2018.00291

Joeris, A., Lutz, N., Blumenthal, A., Slongo, T., & Audigé, L. (2017). The AO Pediatric Comprehensive Classification of Long Bone Fractures (PCCF): Part I: Location and morphology of 2,292 upper extremity fractures in children and adolescents. *Acta Orthopaedica*, *88*(2), 123–128. https://doi.org/10.1080/17453674.2016.1258532

Kim, D. H., & MacKinnon, T. (2017). Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, *73*(5), 439–445. https://doi.org/10.1016/j.crad.2017.11.015

Kiuru, M. J., Haapamaki, V. V., Koivikko, M. P., & Koskinen, S. K. (2004). Wrist injuries; diagnosis with multidetector CT. *Emergency Radiology*, *10*(4), 182–185. https://doi.org/10.1007/s10140-003-0321-4

Koitka, S., Demircioglu, A., Kim, M. S., Friedrich, C. M., & Nensa, F. (2018). Ossification area localization in pediatric hand radiographs using deep neural networks for object detection. *PLoS ONE*, *13*(11), 1–12. https://doi.org/10.1371/journal.pone.0207496

Krogue, J. D., Cheng, K., Hwang, K. M., Toogood, P., Meinberg, E. G., Geiger, E. J., … Pedoia, V. (2019). Automatic hip fracture identification and functional subclassification with deep learning. *ArXiv*, (1), 1–14. https://doi.org/10.1148/ryai.2020190023

Liang, X., Nguyen, D., & Jiang, S. (2020). Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with Cone-Beam Computed Tomography (CBCT) to Computed Tomography (CT) image conversion. *ArXiv*. https://doi.org/10.1088/2632-2153/abb214

Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., … Potter, H. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(45), 11591–11596. https://doi.org/10.1073/pnas.1806905115

Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., … Wen, S. (2020). PP-YOLO: An effective and efficient implementation of object detector. *ArXiv*.

Medoff, R. J. (2005). Essential radiographic evaluation for distal radius fractures. *Hand Clinics*, *21*(3), 279–288. https://doi.org/10.1016/j.hcl.2005.02.008

Olczak, J., Fahlberg, N., Maki, A., Razavian, A. S., Jilert, A., Stark, A., … Gordon, M. (2017). Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? *Acta Orthopaedica*, *88*(6), 581–586. https://doi.org/10.1080/17453674.2017.1344459

Omeiza, D., Speakman, S., Cintas, C., & Weldemariam, K. (2019). Smooth Grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models. *ArXiv*, 1–10.

Pan, I., Thodberg, H. H., Halabi, S. S., Kalpathy-Cramer, J., & Larson, D. B. (2019). Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge. *Radiology: Artificial Intelligence*, *1*(6), e190053. https://doi.org/10.1148/ryai.2019190053

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., … Lerer, A. (2017). Automatic differentiation in PCF. *Proceedings of the ACM on Programming Languages*, *5*(POPL), 1–4. https://doi.org/10.1145/3434309

Qi, Y., Zhao, J., Shi, Y., Zuo, G., Zhang, H., Long, Y., … Wang, W. (2020). Ground Truth Annotated Femoral X-Ray Image Dataset and Object Detection Based Method for Fracture Types Classification. *IEEE Access*, *8*, 189436–189444. https://doi.org/10.1109/access.2020.3029039

Raisuddin, A. M., Vaattovaara, E., Nevalainen, M., Nikki, M., Järvenpää, E., Makkonen, K., … Tiulpin, A. (2020). Deep Learning for Wrist Fracture Detection: Are We There Yet? Retrieved from http://arxiv.org/abs/2012.02577

Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A., Huber, M., & Comaniciu, D. (2010). Semantic annotation of medical images. *Medical Imaging 2010: Advanced PACS-Based Imaging Informatics and Therapeutic Applications*, *7628*(May 2014), 762808. https://doi.org/10.1117/12.844207

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *ArXiv*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going Deeper with Convolutions Christian. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). https://doi.org/10.1002/jctb.4820

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Tang, Y. X., Tang, Y. B., Peng, Y., Yan, K., Bagheri, M., Redd, B. A., … Summers, R. M. (2020). Automated abnormality classification of chest radiographs using deep convolutional neural networks. *Npj Digital Medicine*, *3*(1). https://doi.org/10.1038/s41746-020-0273-z

Thian, Y. L., Li, Y., Jagmohan, P., Sia, D., Chan, V. E. Y., & Tan, R. T. (2019). Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. *Radiology: Artificial Intelligence*, *1*(1), e180001. https://doi.org/10.1148/ryai.2019180001

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., … Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*, 111–119. https://doi.org/10.1109/CVPRW50498.2020.00020

Xie, S., Girshick, R., & Doll, P. (2017). Aggregated Residual Transformations for Deep Neural Networks. In *IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Yahalomi, E., Chernofsky, M., & Werman, M. (2018). Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. Retrieved from http://arxiv.org/abs/1812.09025

Yang, A. Y., & Cheng, L. (2019). Long-Bone Fracture Detection using Artificial Neural Networks based on Contour Features of X-ray Images. Retrieved from http://arxiv.org/abs/1902.07897

Zhang, R. (2019). Making convolutional networks shift-invariant again. In *International Conference on Machine Learning* (pp. 7324–7334).

Appendix A: Validation scores on the data from RadboudUMC

	Classifier performance validation set			
DenseNet121_aa	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Radius_Frontal	0.859	0.8089	0.9286	0.9356
Ulna_Frontal	0.836	0.8212	0.8621	0.9134
Radius_Lateral	0.886	0.8956	0.8561	0.94
Ulna_Lateral	0.8676	0.8714	0.8561	0.9311
Frontal	0.8524	0.844	0.8682	0.9213
Lateral	0.8787	0.9005	0.8106	0.9305
Inception_v3	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Radius_Frontal	0.8537	0.8242	0.902	0.9252
Ulna_Frontal	0.8322	0.8322	0.8552	0.9209
Radius_Lateral	0.8824	0.8932	0.8485	0.9279
Ulna_Lateral	0.8805	0.9029	0.8106	0.931
Frontal	0.859	0.8681	0.902	0.9261
Lateral	0.8732	0.9029	0.7803	0.9301
ResNext50	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Radius_Frontal	0.8571	0.8531	0.8758	0.9284
Ulna_Frontal	0.8327	0.8241	0.8586	0.9146
Radius_Lateral	0.8401	0.8939	0.6959	0.9106
Ulna_Lateral	0.8	0.9011	0.697	0.911
Frontal	0.8271	0.8132	0.8514	0.8984
Lateral	0.8235	0.8592	0.7121	0.8699

Appendix B: Demographics dataset Radboud UMC

	Male	**Female**	**Average** age
Train	2122	3229	44.4
Validation	501	790	45.1

Appendix C: Additional examples of the CAM techniques

Table 1
Comparison between the three fracture localization methods GradCAM++, SmoothGradCAM++ and ScoreCAM.

GradCAM++	SmoothGradCAM++	ScoreCAM