University of Twente

Detection of Hip Fractures from Radiographs using Deep Neural Networks

Madis Lemsalu (m.s.lemsalu@student.utwente.nl)
University College Twente

June 24, 2019

Supervisors:
Dr. C. Christin Seifert
Dr. Ir. Jasper Homminga
Dr. Ir. Maurice Van Keulen
Bsc. Jeroen Geerdink

Abstract

Hip fractures are a leading cause for hospitalization worldwide and consequently incur a serious financial burden on societies. In this paper, different deep learning architectures are compared in the task of predicting hip fractures in radiographs of both anteroposterior and lateral views. The training data consisted of 6952 radiographs where 43 percent of images had a fracture. The best performing model architecture was Densenet169 with an accuracy of 96.5 percent on the validation set, while the worst performing architecture of Alexnet had an accuracy of 88 percent. In this paper, important pixel regions that had the biggest impact on the model's classification decision were also visualized by a human interpretable heatmap. Moreover, in order to test the robustness of the trained classifiers, the input images were manipulated with an adversarial attack to induce a false classification. The attack successfully changed the classification outcome and this phenomenon highlights the fragile nature of deep neural networks. Finally, the paper outlines an ethical outlook regarding the future of the field of radiology.

Keywords— computer vision, medical imaging, adversarial vulnerability, machine ethics

1 Introduction

Hip fractures represent a significant clinical and public health problem worldwide. Its consequences cause excessive morbidity that leads to increased mortality [1]. Less than 50% of all patients who have suffered a hip fracture regain the ability to live independently [2] and hip fractures are one of the leading causes for hospitalization in the elderly worldwide [3]. Subsequently, hip fractures incur a serious financial burden on societies [1].

In case there is a fracture suspicion of an elderly or a middle-aged patient, clinical guidelines suggest ordering a hip radiograph, also known as a x-ray scan, first [4]. However, fractures are commonly missed due to a radiologist not noticing an abnormality [5]. Concurrently, not all fractures are detectable on radiographs [6, 7] and further imaging in case of indeterminate radiograph usually warrants a pelvic magnetic resonance imaging (MRI) [4]. Additional imaging is possible by a combination of nuclear medicine bone scans, MRI, and computed tomography (CT) based on the availability of hospital resources. These advanced imaging modalities increase hospital's resource utilization and diagnostic costs, which is particularly burdening in remote and under-serviced regions. More importantly, delayed or missing diagnosis extends the length of hospitalization that entails an increase in cost [8].

Machine learning models can both detect fractures and aid radiologists in detecting fractures. Medical analysis with machine learning has gained unprecedented popularity due to significant increase in the availability of computing power and advances in deep learning [9]. The number of imaging studies has doubled every ten years over the past two decades [10]. Not surprisingly, in the scientific literature, classification of medical images is the most published research area among the uses of deep learning in healthcare [11].

In this research paper, a full pipeline of building a hip fracture classifier is explored. Firstly, different Convolutional Neural Networks (CNNs) are trained on a dataset of 6952 hip radiographs in an attempt to find the most accurate algorithm for classifying hip fractures. Secondly, visualization of the fracture location is implemented to shed the 'black-box' image of the classifier as well as provide human interpretable results for a radiologist.

Furthermore, there are domain-specific vulnerabilities of medical learning systems that have to be taken into account in the context of the project. Namely, deep learning algorithms are highly susceptible to small and carefully engineered changes in input to alter the output. Changes in the output consequently mislead the system and cause misclassifications that potentially render the usage of the medical classifier dangerous. Adversarial examples also expose fundamental blind spots in our training algorithms. Consequently, in the context of the paper the author will explore the technological vulnerabilities of the system and exemplify a concrete way to manipulate the classifier. Lastly, one of the key ethical concerns for Artificially Intelligence (AI) imaging revolves around the accountability and liability of an algorithm's decision. Machine ethics in the realm of AI imaging is explored in the light of the results from this research.

2 Previous work

Despite the prevalence of hip fractures in the clinical medicine, research of hip fracture detection using deep neural networks is severely under-researched with few prior research publications [12, 13, 4, 14]. The key obstacle is lack of large and well-annotated datasets of hip fractures [15, 16]. Annotation of medical images is a time-consuming and very expensive procedure to undertake [17].

Gale et al. implemented a type of CNN, called Densenet, that has successfully detected hip fractures with the accuracy of 97 percent [12]. The model was trained on approximately 53,000 hip x-ray images with the fracture rate of 12 percent. They used an approach, coined as cascaded-CNN, where they trained series of CNNs each for a specific task. First, a CNN was trained to categorize different radiographs types such as frontal or lateral. Then a second CNN was applied to crop the radiograph in the region of the neck of the femur where fractures are located. Another CNN was trained to exclude images that had metal rods in them. Finally, the image of the cropped hip joint was feed into a final CNN classified that used a customized DenseNet169 as its architecture. The input image size was 1024 x 1024 pixels. In this research, cascaded-CNN are not implemented; conversely, one convolutional neural network is used for the task of image classification, despite the angle of imaging and the presence of metal rods. In this research, the author will compare the accuracy of using one CNN with a smaller training set and image size in juxtaposition to the results of Gale et al. where a cascaded-CNN approach was used.

When developing deep neural networks for tasks that are high impact and high risk, such as medical diagnosis, it is of utmost importance for researchers to be able to explain the algorithm's process for arriving at the prediction [18]. In order to increase transparency and intuitiveness of a deep neural network, Selvaraju et al. developed a method to visualize input regions that are 'important' for predictions [19]. The method is called Gradient-weighted Class Activation Mapping (Grad-CAM) and it utilizes the flow of gradient information within the last convolutional layer to visualize the 'most predictive' input regions. GRAD-CAM is used in this research to produce heatmaps that visualize the regions of high importance for the classification of a hip fracture.

Furthermore, pervasive vulnerability of deep neural networks has attracted significant attention in recent years [20]. Research has shown that by altering relatively small perturbations in the input vector, one can alter the algorithm's classification decision [21]. This phenomenon is called 'adversarial attacks', where imperceptible alteration of natural pixel values in images induce erroneous predictions for state-of-the-art deep neural network classifiers. This rises ethical and moral concerns of implementing vulnerable algorithms for critical tasks in the healthcare sector.

3 Methods

In this paper, seven pre-trained ImageNet models were compared in the task of hip fracture classification. When training deep learning models, selecting the correct hyperparameters are of crucial importance, as the seven models' architectures differ from each other. Subsequently, specific hyperparameters were found for each individual model. The hyperparameters that were manually determined were weight decay, learning rate, batch size, and momentum.

The hyperparameters of weight decay and learning rate were determined by a

learning rate range test, which is outlined by Smith et al [22]. The test consists of starting from a very low learning rate and increasing it incrementally after each mini-batch. The loss for each learning rate value is recorded and the process is continued until the loss increases exponentially. The learning rate range test is ran twice. Once in the beginning when only the last neural network layer is trained, and once when all of the network layers are unfrozen. The batch size was kept constant at 32. The image size was downscaled to 224x224 pixels due to RAM restrictions. The models used were pre-trained on ImageNet dataset and consequently our data was normalized by the mean and standard deviation of the training data from ImageNet.

The models were trained according to the practices of transfer learning where pre-trained models are used to initialize the starting model's weights. Firstly, only the last layers of the network were trained for eight epochs. Afterwards, all of the layers were unfrozen and trained with a cyclical learning policy [22]. The cyclical learning policy was used in order to prevent the model from getting stuck at saddle point. The training continued until there was no further improvement in accuracy of the validation set for seven epochs. All of the models were trained using stochastic gradient descent with Adam optimizer and Rectified Linear Unit (ReLU) as the activation function. The evaluation criteria was accuracy of the validation set and the loss function was cross-entropy loss. The input was flattened, before it was fed to the loss function.

To regularize the model, dropout was applied with the value of 0.5. Moreover, in order to further regularize and invoke model generalization, data augmentation was done for the training data. Data augmentation consisting of rotation and flipping has yielded high accuracy on radiology images in prior research [23, 24]. In this research, extensive data augmentation consisting of rotation, symmetric warping, zooming, and transformations of lighting/contrast were applied.

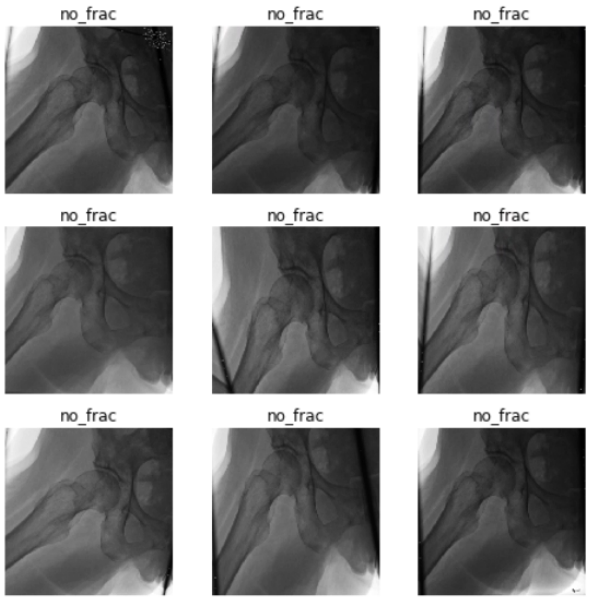

Figure 1: Resulting data augmentation for one image.

3.1 Model Interpretability

This paper utilizes the GRAD-CAM technique to visualize the output of any CNN model by using a coarse localization map in order to highlight areas that the model deems important for classification decision. The author decided to implement this due to the importance of explaining the decision to human radiologists in order to evaluate the decision areas of the model and aid in the decision making process. Moreover, the visual explanation yields transparency to evaluate the pixel regions that the model deems important.

3.2 Adversarial Examples

A phenomenon called adversarial examples allows for imperceptible alteration of natural pixel values in images to induce erroneous predictions for state-of-the-art deep neural network classifiers.

Researchers have not reached an agreement for the root cause of this phenomena. Previous work claims that aberrations may arise due to statistical fluctuations [25] in the training data or input spaces' high dimensional nature [26].

Recent research has found that attacks consist of only rotations and translations are sufficient in deceiving a deep neural network [20]. Even altering one pixel has successfully changed the predictive outcome of CNNs [21]. Potential solutions to increase model robustness against adversarial attacks is to use test-time augmentation (TTA) [27, 20]. TTA consists of application of different augmentations to a test image and making a separate prediction for each augmented image. Then, the final prediction is the average prediction from all of the individual predictions. Furthermore, image augmentation during training has shown to be an effective technique to increase prediction accuracy and robustness of an image against adversarial attacks [28]. Another possibility to combat adversarial attacks is to increase the amount of training data [29]. Finally, injecting adversarial examples into the training set has also lead to increased model robustness [30].

On the other hand, one research paper has pointed out that:

> *"Adversarial vulnerability is a direct result of our models' sensitivity to well-generalizing features in the data."* [31]

This means that by training a classifier that maximizes the accuracy of the validation set, the classifier will use any available signal to derive predictive features from the data - even if the signals look incomprehensible from a human perspective. This would mean that adversarial vulnerability would be a purely human-centric phenomenon, as from the perspective of supervised learning, adversarial features can be as important as non-adversarial features.

Concurrently, approaches aiming at enhancing human interpretability, may hide features that are meaningful for the models themselves. Conclusively, training models to produce human-meaningful explanations cannot be achieved independently from training of the models themselves.

In this research, a simple adversarial attack was carried out called the Fast Gradient Sign Method" (FGSM) [32]. FGSM adds noise in the same direction as the gradient of the cost function. The amount of noise is controlled by the magnitude of ϵ.

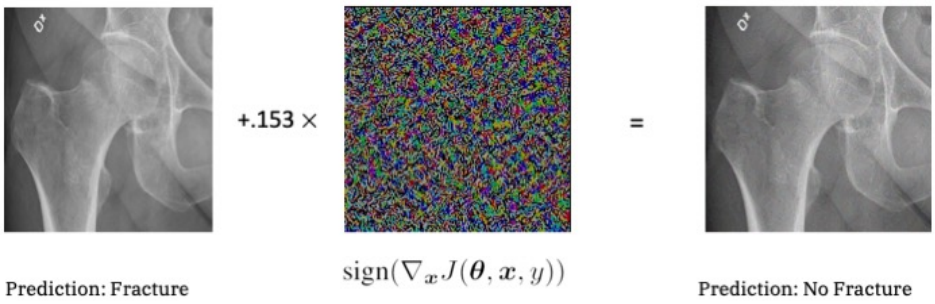

Figure 2: A demonstration of a simple fast adversarial example. A small altercation, which is nearly invisible to the human eye, can change the classification of the image. Epsilon stands for a vector whose elements are equal to the elements of cost function's gradient in respect to the input. In this case, epsilon takes the value of 0.153.

The attack was carried out successfully and the outcome of the classification changed. The altered image looks undifferentiable to the human eye.

4 Experiment

4.1 Dataset

The dataset consists of 8643 images in DICOM format and are provided by the radiology department of the Hospital Group Twente (ZGT). The data was converted to a PNG image format from the original DICOM file, while discarding rest of the meta-data. The images were annotated manually by two radiologists. The data set was split into a training and test set with a ratio of 1:5, which resulted in a training set (6952 images) and a validation set (1676 images). There was no overlap of patients between the training and validation set. The prevalence of fractures was 43,1% in the training set and 42,5% in the validation set.

4.2 Model architectures

The model architectures used in this experiment are from the PyTorch library. Different model architectures were implemented to evaluate whether a certain type of architecture would have a predominant accuracy on the validation set. The following is a a short description comparing different architectures used in this experiment.

Alexnet significantly outperformed prior competition on Imagenet in 2012. The layers consist of dropout, max pooling, and it uses stochastic gradient descent with momentum. All of the following architectures build upon the fundamentals of Alexnet. VGG_19 is a architecture built in 2014 that is similar to Alexnet, but uses only 3x3 convolutions and far more filters. Consequently, VGG_19 is the architecture with the largest amount of parameters (138 million). Resnets were developed in 2015 and use "skip connections", where some of the layers in the neural network are skipped. SquuezeNet was developed in 2016 with the aim to use as few parameters as possible while achieving similar accuracy on Imagenet than Alexnet. Densenets on the other hand use the

input from all preceding layers. This allows for the network to be compact by having fewer number of channels.

4.3 Results

The best performing model was Densenet169, which was the same architecture that was used by Gale et al.'s research. Densenet169 had an accuracy of 0.965, recall of 0.968, precision of 0.970, and a f1-score of 0.967.

Resnet50 model architecture was the second best performing algorithm that reached an accuracy of 0.963, recall of 0.950, precision of 0.98, and a f1-score of 0.958.

The following heatmaps are implemented with GRAD-CAM and they showcase classified radiographs by the best performing Densenet169.

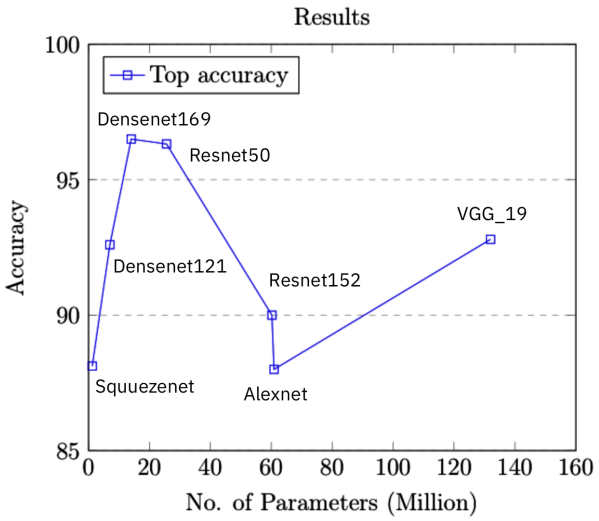

Figure 3: Classifcation Accuracy

The final results:

Model	Accuracy
Densenet169	96.50
Resnet50	96.32
VGG_19	92.80
Densenet121	92.60
Resnet152	90.00
Squuezenet	88.12
Alexnet	88.00

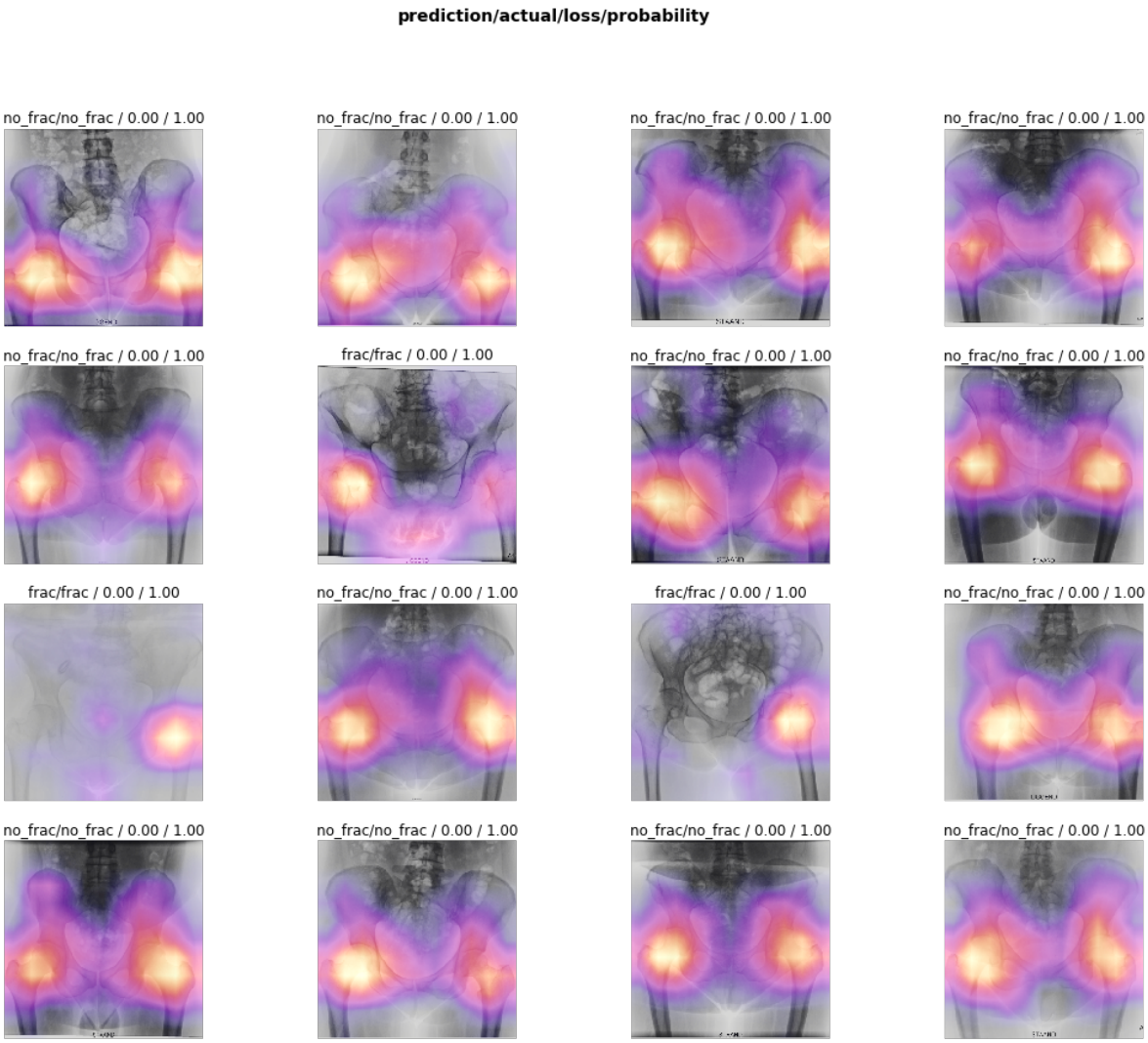

Figure 4: Radiographs that the best performing Densenet169 model was 100% confident about and classified correctly. The model has generalized best to pelvic radiographs.

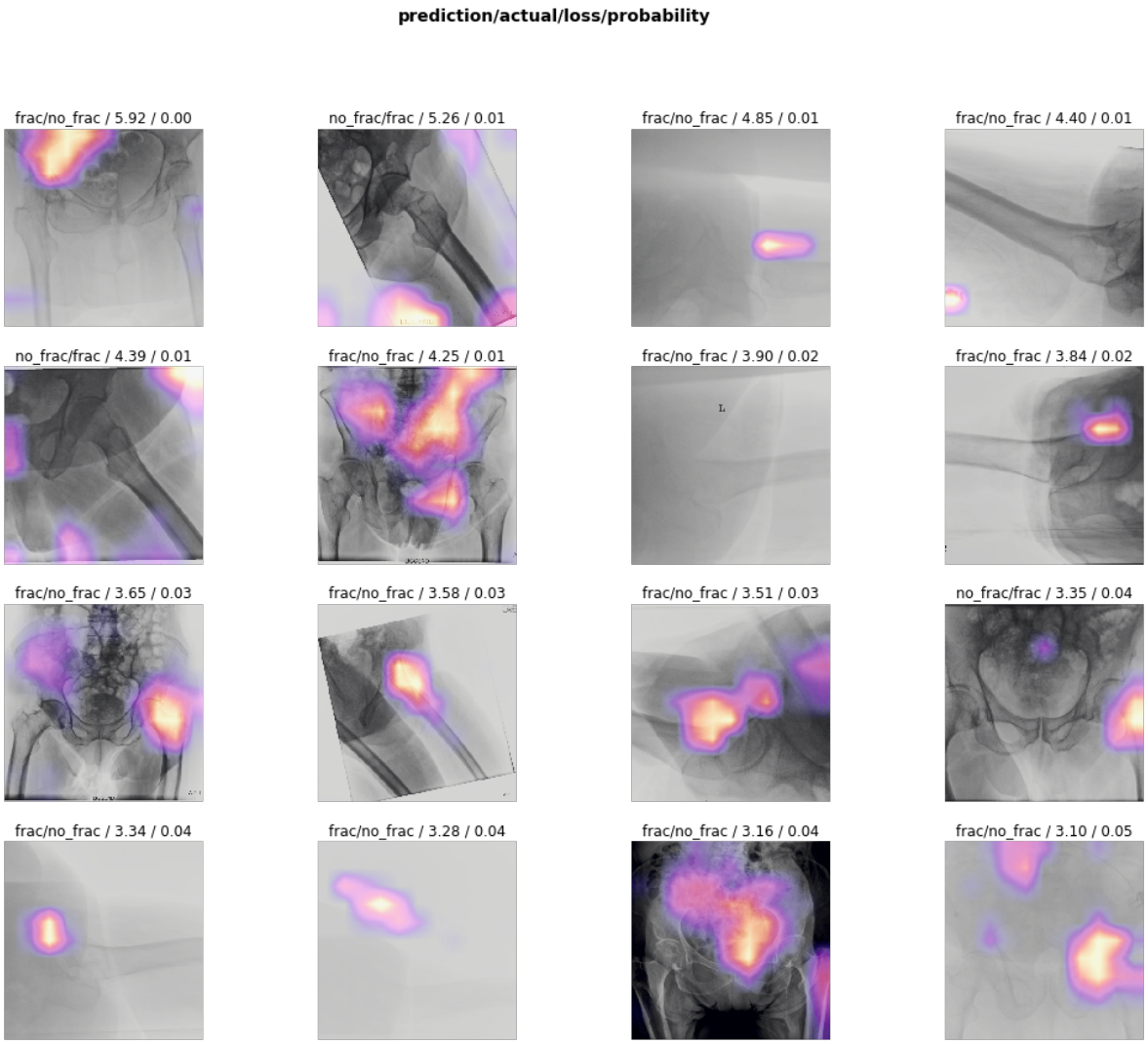

Figure 5: Radiographs that the best performing Densenet169 model had the lowest confidence in and classified incorrectly.

5 Discussion

The author believes there are multitude of factors contributing to the inability to reach the level of accuracy reported in the work of Gale et al. Consequently, by taking into account these factors for further research, one would be able to improve the accuracy of the models further. Several of the factors will be explained in the following paragraphs.

Firstly, a significant difference lies in the substantial discrepancy between the size of the training sets. Gale et al.'s research had around 50,000 images, while this study had 6952 images. Generally speaking, an increase in the size of the training data improves the accuracy of a machine learning model and its ability to generalize [33, 34].

Moreover, in the original research paper, they separated different radiograph types, such as frontal or lateral, by using a cascaded-CNN approach. In this research, the cascaded-CNN approach was not used. The trained model generalized best to

frontal pelvic radiographs, while it had low confidence in its classification of lateral x-rays. A cascaded-CNN approach could potentially increase the accuracy of the model by separating different radiograph types. Lateral images tend to be difficult to accurately classify due to the positioning of the patient and by having a separate CNN trained specifically for lateral images, the chances for better predictions increase as the model wouldn't have to generalize to different types of images.

Furthermore, some of the lateral x-rays had a bad image quality where the pixel values had a uniform distribution. This was due to the fact that the patient was experiencing pain during the radiograph and moved while the image was taken. Additional problems were caused by incorrectly labeled data as well as images where the fracture was located outside of the image and consequently the algorithm couldn't detect the fracture.

When contrasting the performance between radiologists and classification algorithms, Kasai et al. has performed a clinical trial to evaluate whether algorithms can effectively augment the performance of radiologists. The results indicated that radiologists working together with algorithms were significantly better at detecting vertebral fractures [35]. The same effect has been corroborated by other studies [36].

In the light of this research, the algorithm for detecting hip fractures could be used to augment the performance of radiologists. One option to do so, would involve the algorithm predicting the presence of a fracture with a heatmap showcasing the regions of high importance for the algorithm's decision.

At some point, the algorithm may be allowed to autonomously make the decision whether a fracture has occurred. In that case, prerequisites need to be in place for evaluating the performance of the algorithm. For example, a radiologist could verify a small sub-sample of the classifications conducted by the algorithm.

A way to increase prediction accuracy and safeguard patient safety, would be to implement a cut off point for classification confidence. For example, when the algorithm's classification confidence falls below certain threshold, say 99,5 percent, then the radiograph would be reviewed by a radiologist with the help of the generated heatmap. That is to say, when the algorithm's prediction confidence is not 100 percent, a manual review by a radiologist would be conducted.

Overall, the research conducted in this paper is a promising step towards potential augmentation of radiologists in classification of hip radiographs at the ZGT.

5.1 Future of Radiology

Radiologists have been on the forefront of digital era and development ever since the beginning of the profession. Already in 1980s radiologists were pioneering work in medical imaging perception [37]. They have guided the process by being the pioneers in adopting computer science and can now be considered as the most digitally aware healthcare professionals [38].

As a consequence of the implementation of computer science in the routines at their hospitals radiologists were able reduce the time required for image interpretation. The so saved time could, thus, be spent on focusing on improving patient care. By reducing the time required for image interpretation, radiologists could focus on improving patient care.

Moreover, radiologists continue to bear duties such as consideration of patients' values, medical judgment, quality assurance, and communication of diagnosis that

cannot be performed by computers alone [39]. For example, radiologists could lead multidisciplinary meetings for drafting policy, in addition to verifying reports and judgment calls. Consequently, radiologists remain essential for medical practice due to ingenuity that requires unique human characteristics.

However, with the advancement of AI, the role of radiologist is bound to shape from someone who is dealing mainly with image data, to applications that cover a broader spectrum of patient care. History of automation has shown that jobs are often not lost, but often reshaped [40].

5.2 Ethical Concerns

Historically, data collection of patient's medical data served the purpose in care of that patient and it was often deleted over time due to cost of storage. With the advent of personalized medicine that focuses on providing treatment suitable for each individual, a new purpose for medical data has emerged: to inform the care of other patients. This secondary use of data presents a multitude of ethical and legal challenges for AI in radiology. It raises, for example, questions about the level of autonomy and liability of algorithms. Moreover, concerns arise regarding confidentiality, ownership and sharing of the patient's data as well as over data protection [36].

In respect to data protection, one can classify radiology images data as 'data concerning health'. These data include data on the physiological state of a patient and is, thus, considered as a 'special category of personal data' according to Art. 4 para. 15 in the GDPR [41]. This means that its usage is - in general - prohibited according to Art. 9 in the GDPR. This is due to the fact, that this kind of data is considered extremely sensitive and a misuse of this data could have severe consequences for the individual. However, there are several exceptions to that general prohibition. The two most applicable ones for this paper are the ones in Art. 9 para. 2 lit. h and j of the GDPR. According to the sub-paragraphs, special category of personal data can be processed for medical treatments of a patient or for research purposes [42]. However, special protection of radiology images data under the GDPR leave still multiple ethical and legal questions unanswered.

5.3 Future Improvement

There are various possibilities to further improve the results by the classifier. Firstly, increased image resolution and training time has shown to improve the accuracy of convolutional neural networks [33]. An increased resolution would also increase the precision of the GRAD-CAM visualization heatmap, due to increased pixel density. Secondly, a cascaded-CNN can be used to detect different types of images (frontal, lateral) and to build separate classifiers for each type of image. Alternatively, separation of image types could already be accomplished during the annotation process and different classifiers could be trained for each image type. Furthermore, there are multiple images where the patient has experienced pain and has consequently moved during imaging, which has obscured the quality of the image that has resulted in a blurred and low contrast image. Such images could be detected filtered out of the dataset and assigned to a radiologist. Finally, newer architectures, such as EfficientNet, that have outperformed model architectures used in this study at the Imagenet dataset, could potentially further increase the accuracy of the classifier.

6 Conclusion

In conclusion, different machine learning architectures were trained on a dataset of 6952 radiographs of hip fractures. The best performing algorithm was Densenet169 that reached an accuracy of 96.5 percent. To make the output of the algorithm more transparent, GRAD-CAM was implemented in order to visualise important pixel regions for the classification decision. The trained models were susceptible to adversarial attacks that highlight the fragile nature of the classifiers. Finally, an ethical review of implementation of such classifiers was carried out.

References

[1] Brent C Taylor, Pamela J Schreiner, Katie L Stone, Howard A Fink, Steven R Cummings, Michael C Nevitt, Paula J Bowman, and Kristine E Ensrud. Long-term prediction of incident hip fracture risk in elderly white women: study of osteoporotic fractures. *Journal of the American Geriatrics Society*, 52(9):1479–1486, 2004.

[2] R Sean Morrison, Mark R Chassin, and Albert L Siu. The medical consultant's role in caring for patients with hip fracture. *Annals of internal medicine*, 128(12_Part_1):1010–1020, 1998.

[3] Carmen A Brauer, Marcelo Coca-Perraillon, David M Cutler, and Allison B Rosen. Incidence and mortality of hip fractures in the united states. *Jama*, 302(14):1573–1579, 2009.

[4] Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(1):31, 2019.

[5] Jennifer J Donald and Stuart A Barnard. Common patterns in 558 diagnostic radiology errors. *Journal of medical imaging and radiation oncology*, 56(2):173–178, 2012.

[6] Jesse Cannon, Salvatore Silvestri, and Mark Munro. Imaging choices in occult hip fracture. *The Journal of emergency medicine*, 37(2):144–152, 2009.

[7] Matthew W Kirby and Charles Spritzer. Radiographic detection of hip and pelvic fractures in the emergency department. *American Journal of Roentgenology*, 194(4):1054–1060, 2010.

[8] Nicole Simunovic, PJ Devereaux, and Mohit Bhandari. Surgery for hip fractures: Does surgical delay affect outcomes? *Indian journal of orthopaedics*, 45(1):27, 2011.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[10] Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. Rising use of diagnostic medical imaging in a large integrated health system. *Health affairs*, 27(6):1491–1502, 2008.

[11] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016.

[12] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv preprint arXiv:1711.06504*, 2017.

[13] Anees Kazi, Shadi Albarqouni, Amelia Jimenez Sanchez, Sonja Kirchhoff, Peter Biberthaler, Nassir Navab, and Diana Mateus. Automatic classification of proximal femur fractures based on attention models. In *International Workshop on Machine Learning in Medical Imaging*, pages 70–78. Springer, 2017.

[14] Thao P Ho-Le, Jacqueline R Center, John A Eisman, Tuan V Nguyen, and Hung T Nguyen. Prediction of

hip fracture in post-menopausal women using artificial neural network approach. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4207–4210. IEEE, 2017.

[15] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.

[16] C Krittanawong. The rise of artificial intelligence and the uncertain future for physicians. *European journal of internal medicine*, 48:e13–e14, 2018.

[17] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741M. International Society for Optics and Photonics, 2018.

[18] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[19] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[20] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

[21] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019.

[22] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[23] R Ogawa, T Kido, and T Mochizuki. Effect of augmented datasets on deep convolutional neural networks applied to chest radiographs. *Clinical Radiology*, 2019.

[24] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*, 2018.

[25] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. iclr'15. *arXiv preprint arXiv:1412.6572*, 2015.

[26] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *arXiv preprint arXiv:1809.03063*, 2018.

[27] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126, 2015.

[28] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten.

Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[29] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

[30] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[31] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[33] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[34] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[35] Satoshi Kasai, Feng Li, Junji Shiraishi, and Kunio Doi. Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs. *American Journal of Roentgenology*, 191(1):260–265, 2008.

[36] Charlene Liew. The future of radiology augmented with artificial intelligence: a strategy for success. *European journal of radiology*, 102:152–156, 2018.

[37] Elizabeth A Krupinski. The future of image perception in radiology: synergy between humans and computers. *Academic radiology*, 10(1):1–3, 2003.

[38] Francesco Sardanelli. Trends in radiology and experimental research, 2017.

[39] S Russell and J Bohannon. Artificial intelligence. fears of an ai pioneer. *Science (New York, NY)*, 349(6245):252–252, 2015.

[40] Saurabh Jha and Eric J Topol. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Jama*, 316(22):2353–2354, 2016.

[41] Marta Otto. Regulation (eu) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation–gdpr). In *International and European Labour Law*, pages 958–981. Nomos Verlagsgesellschaft mbH & Co. KG, 2018.

[42] Xiang Su, Jarkko Hyysalo, Mika Rautiainen, Jukka Riekki, Jaakko Sauvola, Altti Ilari Maarala, Harri Hirvonsalo, Pingjiang Li, and Harri Honko. Privacy as a service: Protecting the individual in healthcare data

processing. *Computer*, 49(11):49–59, 2016.