Assessing the clinicians' pathway to embed artificial intelligence for assisted diagnostics of fracture detection

Carlos Francisco Moreno-García[1], Truong Dang[1], Kyle Martin[1], Manish Patel[2], Andrew Thompson[2], Lesley Leishman[1], Nirmalie Wiratunga[1]

Abstract. Fracture detection has been a long-standing paradigm on the medical imaging community. Many algorithms and systems have been presented to accurately detect and classify images in terms of the presence and absence of fractures in different parts of the body. While these solutions are capable of obtaining results which even surpass human scores, few efforts have been dedicated to evaluate how these systems can be embedded in the clinicians and radiologists working pipeline. Moreover, the reports that are included with the radiography could also provide key information regarding the nature and the severity of the fracture. In this paper, we present our first findings towards assessing how computer vision, natural language processing and other systems could be correctly embedded in the clinicians' pathway to better aid on the fracture detection task. We present some initial experimental results using publicly available fracture datasets along with a handful of data provided by the National Healthcare System from the United Kingdom in a research initiative call. Results show that there is a high likelihood of applying transfer learning from different existing and pre-trained models to the new records provided in the challenge, and that there are various ways in which these techniques can be embedded along the clinicians' pathway.

Keywords: Fracture detection, natural language processing, convolutional neural networks, clinicians' pathway

1 Introduction

In recent years, fracture detection has been one of the most cited challenges in medical imaging analysis, evidenced both by public competitions [18] and clinical trials [9] alike. The design of a system which aids clinicians in the automatic detection of fractures is of paramount to reduce the workload of the front line staff and allow them more time to focus on the most urgent cases. To address this issue, the Scottish Government, Opportunity North East (ONE) and the Small Business Research Initiative (SBRI) announced a challenge to carry out a project towards looking at this problem in the healthcare system in the northeast of Scotland[3]. A team comprised of members from the industry (Jiva.ai) and academia (Robert Gordon University) was formed to look at the problem and design a solution.

In addition, a key contribution of this work is the modelling of the clinician's pathway, exploring the current process of radiology imaging for fracture treatment. This was built through a series of co-creation sessions with a reporting radiographer, then verified by two clinicians and two other reporting radiographers. By designing this pathway, we identified three key stakeholders (clinician, radiologist, patient) and four sub-processes: (1) requesting radiology images; (2) acquiring radiology images; (3) reporting on radiology images; and (4) decision support. By understanding the current approach to radiology imaging for fracture treatment, we consider that we are capable of offering a valuable contribution that will allow research groups to pinpoint opportunities for smart automation of this process in future.

The contributions of this paper are as follows:

1. Identify the current processes involved with the various procedures affected by radiology imaging for fractures.
2. Explore where these processes can be improved through the implementation of AI.
3. Identify what form of AI would be most applicable in order to maximise the obtained benefits by the affected stakeholder.
4. Given the limited amount of samples provided by the challenge proposer, perform initial proof of concept tests using baseline methods to identify the potential of transfer learning in this domain.

2 Related Work

2.1 Image recognition

Only a handful of demonstrations of machine learning, computer vision and natural language processing for bone fracture detection appear in scientific literature. Lindsey et al. [9] demonstrated that a deep neural network trained on 256'000 x-rays could detect fractures with a similar diagnostic accuracy to a sub-specialised orthopaedic surgeon. Also, Olczak et al. [10] applied deep learning to analyse 256'458 x-rays, and concluded that artificial intelligence methods are applicable for radiology screening, especially during out-of-hours healthcare or at remote locations with poor access to radiologists

[1] Robert Gordon University, Garthdee Road, Aberdeen, Scotland, UK, Contact email: c.moreno-garcia@rgu.ac.uk
[2] Jiva.ai Ltd, 132 Street Lane, Leeds, UK, Contact email: info@jiva.ai
[3] http://cfmgcomputing.blogspot.com/2019/11/artificial-intelligence-for-fracture.html

or orthopedists. Smaller scale studies using tens to low thousands of images include Lawn et al. [8], Kim and MacKinnon [5], Tomita et al. [14], Dimililer [3] and Bandyopadhyay et al. [1], amongst others.

Three technical frames have been described as applicable for ML in radiology: image segmentation, registration and computer-aided detection and diagnosis [17]. Out of these, graph models, such as Bayesian networks and Markov random fields, have been identified as the two most widely used techniques for fracture modelling. However, recent advances in generative deep models (e.g. variational autoencoders) [6] have been applied to annotate both images and text; which are yet to be exploited in radiology and related applications. Similarly, multi-modal learners [16] have been used to learn from image and text to improve recognition of objects; unlike single modality learners these can combine mixed embedding spaces that unite different modalities. These have been applied to public text and images but digital health applications are yet to emerge. Given the relevance of both image and text to clinical radiology we expect to adapt these algorithms to create an innovative unified embedding suited to automated annotation by deep generative algorithms. Indeed, we can use state-of-the-art translation algorithms such as transformers [15] which exploit similarity and capitalise on adjacency information, to generate reports from both radiograms and clinical text.

3 Modelling a Clinician's Pathway

Any artificially intelligent solution which is suggested to resolve some problem within the field of radiology should be rooted in deep understanding of the domain and user requirements. With this in mind, we have received input from domain-experts to develop a clinician's pathway, detailing current process of radiology imaging for fracture treatment. In this paper, we aim to demonstrate the opportunities which offer potential for smart automation in this field. In particular, we aim to highlight the areas where the application of artificial intelligence would be impactful for increasing efficiency, improving patient experience and decreasing cost.

3.1 Co-Creation of Clinician's Pathway

We performed a multi-stage co-creation process to model the clinician's pathway which was indicative of real-world radiology practice. These stages were divided into a design phase, where we aimed to understand and model the current processes for the acquisition and reporting of radiology images, and a validation phase, in which we obtained unbiased feedback on our model from personnel external to our co-creation. The result is a clinician's workflow which has been developed alongside a reporting radiographer, and verified by two clinicians (one consultant radiologist and one senior accident & emergency doctor) and two reporting radiographers from within the National Health Service (NHS) Scotland. We are therefore confident in its accuracy and its suitability to describe real radiology processes. Although this pathway has been built with input from British radiologists, we suggest it can be generalised to wider radiology practices (within reason).

During the design phase, we organised two separate co-creation sessions. In the first session, we met with a reporting radiographer to discuss the complete journey of a patient who was given an appointment for radiology imaging for a suspected fracture. This session was useful to establish the process start-points and end-points. In the second session, we observed a reporting radiographer reporting on a series of x-ray images for suspected fractures. This session was intended to identify relevant technologies and the role they played in reporting on radiology images. The outcome of this two stage design phase was an initial draft of the clinician's pathway which could be validated by domain experts.

We then organised three sessions for the validation of the developed workflow. In the first session, we obtained feedback upon the pathway from the reporting radiographer who was directly involved in its formation. This allowed any errors which had arisen due to misunderstanding aspects of the design phase to be corrected. In the second session, a member of the research team met with a clinician and a reporting radiographer to explain and discuss the draft pathway and obtain feedback. This session was designed to ensure that the pathway could be generalised to more than just the single radiographer with whom it had been co-designed. In particular, the session resulted in a number of updates to the role of the clinician as an actor in the process. Finally, we used the third session as an opportunity to obtain blind feedback on the developed pathway as a form of litmus test regarding its accuracy to radiology practice. We presented the pathway to a new clinician and reporting radiographer, and requested feedback on any areas where they felt (a) that the pathway was not indicative of real-world practice and (b) that there were opportunities for artificial intelligence to make the process more efficient.

The results of the validation sessions were very valuable for the design process. As a methodology, by performing our co-creation of the clinician's pathway in this manner, we are confident it is accurate to real-world practice, and generalisable beyond simply an individual's viewpoint. In the final validation session, the clinician did not highlight any areas of the workflow which were not indicative of real-world practice, while the reporting radiologist suggested only a minor amendment to terminology. Furthermore, the areas which both of the participants suggested were suitable for artificial intelligence to make a process improvement very closely overlapped with our own findings as researchers. We discuss this in more detail in Section 3.3. In the following subsection, we will introduce and discuss the developed clinician's pathway in detail.

3.2 Resulting Clinician's Pathway

In modelling the process of radiology imaging for fracture treatment, we identified three key stakeholders (clinician, radiologist, patient) and four sub-processes: (1) requesting radiology images; (2) acquiring radiology images; (3) reporting on radiology images; and (4) decision support. The complete figure can be accessed via this link[4] and can be seen in Figure 1. We summarise our pathway using a workflow diagram which we will break down into respective processes in Figures 2, 3 and 4.

[4] https://www.dropbox.com/s/3gx7bicf43bn0lx/wrk_flw_comp_c.pdf?dl=0

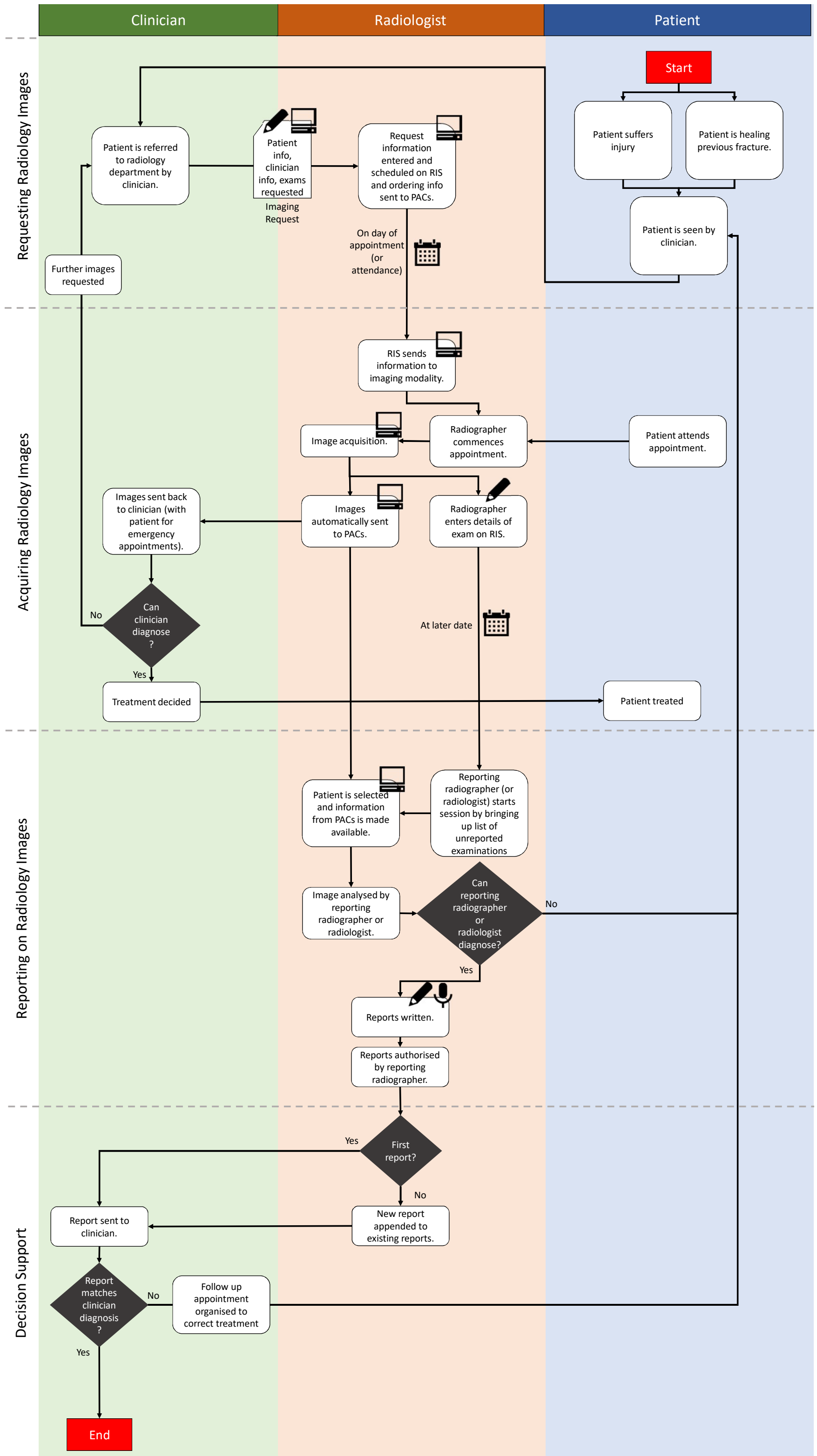

Figure 1. A complete depiction of the current workflow for requesting, acquiring and reporting upon radiology images

1. Requesting Radiology Images: When requesting radiography images for a patient, a clinician sends an imaging request to the radiology department. This request contains information on the patient's demographic (including a medical history), information about the clinician making the request and the examinations requested. It allows an appointment to be scheduled on the Radiology Information System (RIS), and the ordering information is dispatched to Picture Archiving and Communications System (PACS). PACS is an end-to-end system that supports the process of acquiring radiography images of a patient from referral of the patient until diagnosis and subsequent treatment are agreed. Contained within PACS is the RIS, where the textual components of patient information is stored.

2. Acquiring Radiology Images: Each day, the RIS automatically generates a work list for each imaging modality. This work list gives the radiologist (or the reporting radiographer) access to the request information created by the original referring clinician, helping them to understand the imaging requirements. This enables the radiologist to perform the requested imaging. After the imaging has been performed, the medical equipment will generate images in Digital Imaging and Communication in Medicine (DICOM) standard, and load them into a graphic user interface where they are made available for the radiologist to annotate and report. A graphical representation of this is displayed in Figure 2. DICOM defines an image standard and format for medical images. Images are high resolution and are linked directly with other patient data (such as name, gender and age). It was developed by the National Electrical Manufacturers Association (NEMA) as part of a set of standards that define best practice and inform the international standard for the capture, retrieval, storage and transmission of medical imaging data. DICOM is currently the most commonly used standard across the world for medical imaging, and is implemented in most radiology, cardiology and radiotherapy devices, as well as devices in other medical domains such as dentistry.

3. Reporting on Radiology Images: Radiologists can access these images by calling up a list of unreported examinations. For each patient, previous images and reports are also available to support diagnosis. PACS enables radiologists to annotate the images to highlight areas of interest or identify supporting evidence for their diagnosis. These annotations include the ability to perform simple measurements (length of objects, angles of intersections, etc) and to mark a Region of Interest (ROI) on the image. This allows the radiologist to use tools to capture metadata about the ROI, including its area, average pixel values, standard deviation, and range of pixel values.

The radiologist will then generate a textual report to summarise and describe their findings. These reports have no set template or length, but generally include a statement of whether a fracture has been detected, what type of fracture it is, where it is located, and the seriousness of the breakage. Furthermore, the reports may be appended to existing documentation on the patient (if previous radiology records exist) or may be used to begin a radiology record (if no previous visits have been recorded). Many countries then require the reports to be authorised by a radiologist before being released to a clinician. For example, within the United Kingdom the standards for fail-safe communication of radiology reports are governed by the Royal College of Radiologists (RCR)[5]. The result of these factors is that the reports are a complex textual data source with limited uniformity and describing a broad range of diagnosis and observations. We represent this information as part of the workflow displayed in Figure 3.

4. Decision Support In the existing pathway, decision support occurs after the radiology reports are generated. This is non-optimal; often for accident and emergency fracture cases (which make up the majority of fractures in a hospital) the clinician will attempt to read and comprehend the generated radiology images without any input from an expert radiologist. This can be seen in the current workflow in Figures 2 and 4. This occurs because experts can be unavailable - not all radiology staff are sufficiently trained to report on acquired images. As a result the clinician is forced to make a diagnosis and organise follow up treatment on the basis of their individual knowledge. This can lead to misdiagnosis, if the clinician's findings are not consistent with the radiologist's, which can have an impact on the patient's health as well as financial consequences for the hospital involved.

Having developed an understanding of the current procedure of radiology imaging for fracture treatment, we are motivated to make some recommendations where artificial intelligence could make improvements to this process.

3.3 Opportunities for Artificial Intelligence

The key outcome of this work is highlighting the applicability of artificial intelligence in two places: to reduce burden on radiologists by (1) autonomously classifying radiology images and (2) generating understandable and accurate medical reports to describe the intelligent system's findings. Applications of artificial intelligence to fill these gaps presents an opportunity to improve decision-support for clinicians by giving them access to the information immediately. This is a key factor that is missing from much of the research literature on this topic; although an artificial intelligence method for fracture recognition should enhance the efficiency of radiologists, it should also improve the decision-making of clinicians. Therefore, it should be of a suitable form to be absorbed by that user group.

This outcome is supported by the verification obtained from our test group of clinicians and radiologists. Based on their feedback, we have highlighted the most impacted area of the current clinical pathway in Figure 5.

As seen in our discussion of related work, there has been much exploration of autonomous classification of fracture images throughout the literature [1, 3, 5, 8, 9, 10, 14]. However, few works have considered how this could be integrated with existing medical processes. It is clear that from a clinician's perspective, it would be desirable to have the classification of the image and the report in order to support their decision-making. This suggests an ecosystem of artificial intelligence processes would be much more suitable than a standalone method.

4 Experiments on Image Classification

The main purpose of the experimental framework was to test the learning capabilities of different baseline algorithms and settings to classify the images provided in the challenge as fracture/no fracture. To do so, the first task consisted of having a specialist re-annotate the data provided by splitting it into fracture and no fracture labels, based both on the visual aspect of the radiography and on the information provided by the text reports. It was discovered that while most of the

[5] http://bit.ly/rcr-standards

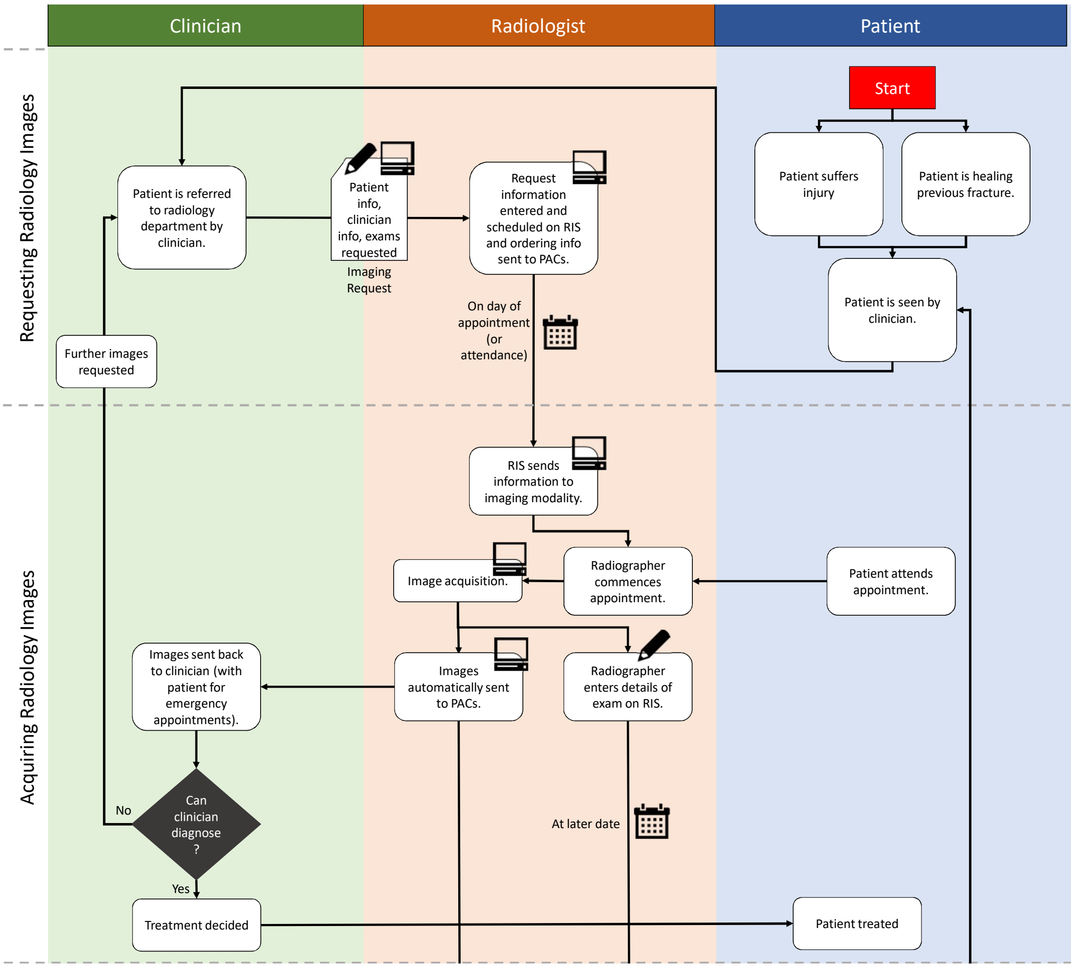

Figure 2. The co-created workflow for requesting and acquiring radiology images.

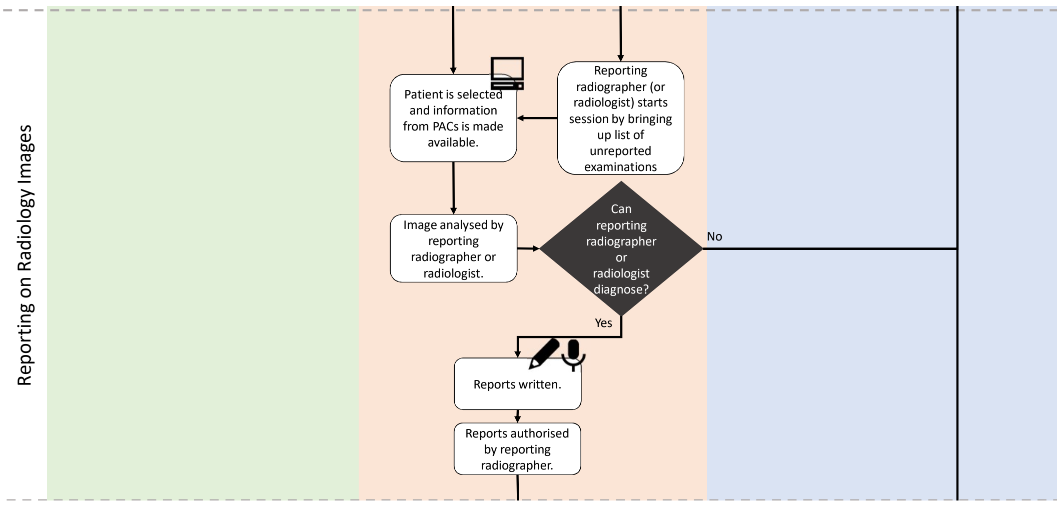

Figure 3. The co-created workflow for reporting on radiology images.

images corresponded to "regular" scenarios (where the purpose is to assess whether the patient has suffered a fracture or not), some other cases also contained follow-up reports (identified as POP) where the issue is not to identify the presence/absence of a fracture, but rather to give a follow up for a patient which already has had the fracture identified in a previous visit. We labelled 73 images as positive (i.e. with fracture) and 138 negative (i.e. no fracture). Moreover, six examples were POP fractures, and thus were not included in our experiments.

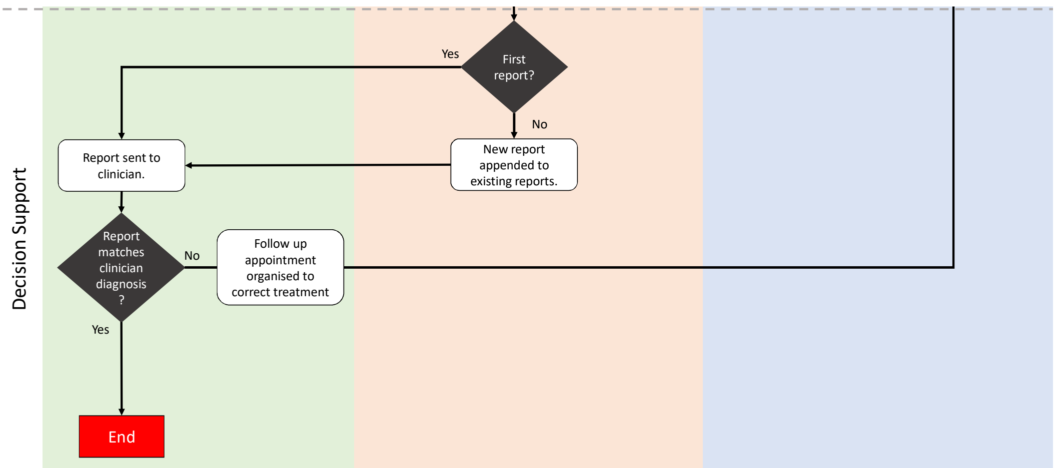

Figure 4. The co-created workflow for using radiology images for decision support.

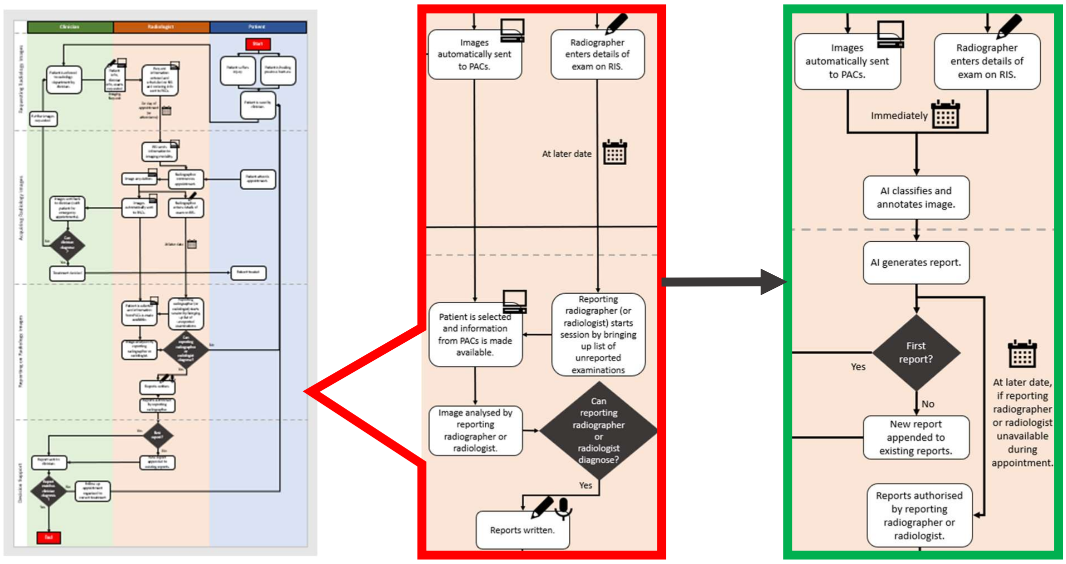

Figure 5. AI provides an opportunity to provide decision support much more quickly by classifying an acquired image and generating a suitable report to describe its findings.

4.1 Datasets

To demonstrate the potential of transfer learning capabilities of the selected algorithms towards the NHS provided records, we used the following publicly available dataset.

MURA: The MURA (MUsculoskeletal RAdiographs) dataset is a large dataset of bone X-rays. Algorithms are tasked with determining whether an X-ray study is normal or abnormal. It consists of X-ray scans on elbow, finger, forearm, hand, humerus, shoulder and wrist. The training set consists of 14'873 positive cases and 21'939 negative cases while the validation set has 1'530 positive examples and 1'667 negative examples. Among them, forearm and wrist are close to our problem, which consists of 7'443 negative and 5'094 positive examples. Images from this dataset can be accessed here[6].

[6] https://stanfordmlgroup.github.io/competitions/mura/

4.2 Image Preprocessing

To increase the likelihood of classification and the training sample size, We applied the following preprocessing techniques, which are the most commonly used in related literature [18]:

- horizontal flip, width shift by 0.1,
- height shift by 0.1,
- shearing with range 0.1,
- zoom with range from 0.9 to 1.25 and
- random rotation from 0 to 15 degrees.

4.3 Architecture Details

These followings baseline architectures were used in our experiments:

A **VGG 16** [11] is a Convolutional Neural Network (CNN) model proposed by Simonyan and Zisserman. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1'000 classes. It is an improvement over the classical AlexNet [7] by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. The original VGG16 was trained for weeks and was implemented using NVIDIA Titan Black GPU's.

B **Resnet 50** [4] is a CNN architecture of 50 layers deep, each of which is formulated as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. Because of these residual modules, the architecture can become very deep. This architecture won the 1^{st} place on the ILSVRC 2015 classification challenge.

C **Inception V3** [12, 13] is a CNN architecture which achieved improved utilisation of the computing resources inside the network by carefully crafted design that allows for increasing the depth and width of the network while keeping the computational budget constant. To optimize quality, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. The authors also proposed ways to scale up networks in ways that aim at using the added computation as efficiently as possible by suitably factorised convolutions and aggressive regularization. Tests were made on the ILSVRC 2012 dataset, in which with an ensemble of four models and multi-crop evaluation, authors reported 3.5% top-5 error on the validation set (3.6% error on the test set) and 17.3% top-1 error on the validation set.

4.4 Experiment Details

To have different points of comparison, we tested the results of using the three aforementioned classifiers to classify images from the MURA dataset. We distinguished between the following four configurations:

1. When the networks were pre-trained with ImageNet [2].
2. When they were initialised randomly.
3. When the networks were initialised randomly, trained on all the MURA dataset (except wrist and arm images) and then retrained on wrist and arm images.
4. When the networks were pre-trained from ImageNet randomly, trained on all of MURA (except wrist and arm) and then retrained on wrist and arm images.

The accuracy result can be seen in Table 4.4 and the run time in Table 4.4:

Table 1. Accuracy results for cases from (1) to (4)

	Case (1)	Case (2)	Case (3)	Case (4)
VGG16	0.82	0.535417	0.535417	0.798958
Resnet50	0.8083	0.535417	0.535417	0.783333
InceptionV3	0.677083	0.535417	0.535417	0.536458

The results showed that case 1 with VGG 16 and ResNet 50 delivered the best accuracy overall (82% and 80% respectively), implying that it is possible to obtain good accuracy provided that we can train the systems with sufficient data, regardless of its origin. Moreover, case 2 with these same architectures also showed good performance, (79% and 78% respectively), but even with the retraining on wrist and arm images, results were slightly worse than training only with ImageNet images. This may be due to the fact that some wrist/arm images had to be used for such retraining instead of testing. In terms of run time, we also found out that case 1 overall is faster to train and test.

Table 2. Running time (in seconds) for cases from (1) to (4)

	Case (1)	Case (2)	Case(3)	Case (4)
VGG16	1869.48	2024.83	5557.05	5374.36
Resnet50	1776.34	1308.95	11406.28	7378.9
InceptionV3	3128.09	3086.32	11429.92	8208.37

After this initial validation, we tested the transfer learning capability from MURA to the newly acquired images. We tested the following three cases:

5. When networks were pre-trained on ImageNet, trained on MURA and tested on the new dataset.
6. Mixing MURA and new images to generate both training and test sets (70% train, 30% test).
7. Same as the previous case, however the test set was composed of 70% of MURA images and 30% from the new dataset.

The accuracy results are shown in Table 4.4:

Table 3. Accuracy results for cases from (5) to (7)

	Case (5)	Case (6)	Case (7)
VGG16	0.668293	0.809524	0.704918
Resnet50	0.673171	0.76112	0.606557
InceptionV3	0.673171	0.727106	0.754098

In contrast to what was expected from the previous test, we observed that for case 5, all CNNs were unable to learn how to classify the new images. In contrast, it was more likely to obtain higher accuracy rates for case 6 and VGG 16 (81%), although this is a direct result of images from the MURA dataset being mixed within the test set. Meanwhile, case 7 and Inception V3 obtained 75% accuracy, but keeping in mind that the test set is only composed of new images, this was a clear indication that it is possible to transfer a model using a larger amount of images. In terms of run time, we discovered that it was faster to train networks through case 6, followed by case 6 and case 7 respectively. The complete run time results can be seen in Table 4.4.

Table 4. Running time (in seconds) for cases from (5) to (7)

	Case (5)	Case (6)	Case (7)
VGG16	2727.67	2700.84s	3718.74
Resnet50	1889.82	3495.72s	4280.42
InceptionV3	2763.39	10388.83s	9656.18

5 Conclusion

In this paper, we have presented a first step towards assessing the most proper way to embed machine learning, computer vision and natural language processing into the clinicians' pathway to improve assisted diagnostics of fracture detection. We have reviewed the most significant literature and designed a pipeline where we have annotated the most relevant action points where artificial intelligence can be used to improve the current practices. In addition, we have carried out some initial experiments to verify how current methods and transfer learning perform on identifying fractures in a reduced dataset provided by the British public health service. Results show that there is a great likelihood of being able to apply transfer learning for these purposes, and in the case that more images are provided by the challenge setter, then the accuracy can vastly improve.

We will continue this partnership to explore more ways in which we can further improve our findings and including other technologies to enhance the existing results. Finally, we will keep working with clinicians and radiographers to correctly assess their pathways and effectively applying these technologies in commercial settings.

Acknowledgment

We would like to acknowledge the Scottish Government's Small Business Research Initiative (SBRI) and Opportunity North East (ONE) for supporting this work, and to the Data Safe Haven (DaSH) from the National Health Service (NHS) for granting access to the test data.

REFERENCES

[1] Oishila Bandyopadhyay, Arindam Biswas, and Bhargab B. Bhattacharya. Long-bone fracture detection in digital X-ray images based on digital-geometric techniques, 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] Kamil Dimililer. IBFDS: Intelligent bone fracture detection system. In *Procedia Computer Science*, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv 1512.03385*, pages 770–778, 06 2016.

[5] D. H. Kim and T. MacKinnon. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, 73(5):439–445, 2018.

[6] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[8] Brian Lawn. *Fracture of brittle solids*. Cambridge university press, 1993.

[9] Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, and Hollis Potter. 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, nov 2018.

[10] Jakub Olczak, Niklas Fahlberg, Atsuto Maki, Ali Sharif Razavian, Anthony Jilert, André Stark, Olof Sköldenberg, and Max Gordon. 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, 2017.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *arXiv:1409.4842 [cs]*, September 2014. arXiv: 1409.4842.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[14] Naofumi Tomita, Yvonne Y. Cheung, and Saeed Hassanpour. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Computers in Biology and Medicine*, 98(February):8–15, 2018.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[16] O. Vinyals, D. Bohus, and R. Caruana. Learning Speaker, Addressee and Overlap Detection Models from Multimodal Streams. In *ICMI*, 2012.

[17] Xingwei Wang, Lihua Li, Wei Liu, Weidong Xu, Dror Lederman, and Bing Zheng. An interactive system for computer-aided diagnosis of breast masses. *Journal of Digital Imaging*, 25:570–579, 2012.

[18] Xulei Yang, Zeng Zeng, Sin G. Teo, Li Wang, Vijay Chandrasekhar, and Steven Hoi. Deep learning for practical image recognition: Case study on kaggle competitions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pages 923–931, New York, NY, USA, 2018. Association for Computing Machinery.