The Knee 52 (2025) 212–219

Contents lists available at ScienceDirect

The Knee

journal homepage: www.elsevier.com/locate/thekne

Machine learning is better than surgeons at assessing unicompartmental knee replacement radiographs

S Jack Tu [a,*], Sara Kendrick [b], Karthik Saravanan [a], Christopher Dodd [c], David W Murray [a,c], Stephen J Mellon [a]

[a] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Windmill Road, Oxford OX3 7LD, United Kingdom
[b] Indiana University School of Medicine, 340 W. 10th Street Fairbanks Hall., Indianapolis, IN 46202, United States
[c] Oxford University Hospitals NHS Foundation Trust, Nuffield Orthopaedic Centre, Old Road, Oxford OX3 7HE, United Kingdom

ARTICLE INFO

Article history:
Received 6 August 2024
Revised 4 November 2024
Accepted 8 November 2024

Keywords:
Convolutional Neural Network (CNN)
Transfer learning
Radiograph
Clinical outcomes

ABSTRACT

Background:: Poor results occasionally occur after unicompartmental knee replacement (UKR). It is often difficult, even for experienced surgeons, to determine why patients have poor outcomes from radiographs. The aim was to compare the ability of experienced surgeons and machine learning to predict whether patients had poor or excellent outcomes from radiographs.

Methods:: 924 one-year anterior-posterior radiographs post-UKR were used to train a machine learning model (ResNet50v2) with a transfer learning approach based on their one-year Oxford Knee Score categories. Two experienced surgeons and the model assessed and categorised 70 radiographs (14 Poor scores; 56 Excellent scores) not used for training according to their expected outcome.

Results:: The ResNet50v2 model correctly identified 71% (n = 10) of the patients with a poor score and 46 (82%) of those with an excellent score. In contrast, one surgeon could not identify patients with Poor scores (0%) and the other identified one (7%). Both misidentified 3 of those with Excellent scores. The model visualisation method suggested that estimated classifications were made from image features around the implants.

Conclusion:: The results suggest that there are radiographical features that relate to poor outcomes, which the surgeons are unaware of. Those the model did not identify may have an extra-articular cause for their poor outcome. Further analysis to identify the features associated with poor outcomes could potentially suggest ways that indications or techniques could be improved so as to decrease the incidence of poor results.

* Corresponding author.
E-mail addresses: jack.tu@ndorms.ox.ac.uk (S Jack Tu), stephen.mellon@ndorms.ox.ac.uk (S.J Mellon).

https://doi.org/10.1016/j.knee.2024.11.007
0968-0160/

So what does this mean for the knee surgeon?

It is difficult for surgeons to predict whether a patient will have a poor outcome from radiographs taken after knee replacements or if they have a poor outcome to determine why it occurred. We found that a trained machine learning (ML) model was much better than surgeons at predicting which patients will have a poor outcome following Unicompartmental knee replacement and that it based its decision on image features near the implants and medial tibial spine. Further analysis should allow us to identify these features and modify the operation to decrease the risk of a poor outcome. This approach is different from how ML is usually used. Instead of replicating tasks we can already do, it is identifying unknown image features. If rolled out on a national scale, based on routine radiographs and Registry data, it has the potential to improve the results of knee replacement.

1. Introduction

Poor results occasionally occur after all types of knee replacement. There are many causes for these poor results. For example, there may be problems with the implant and how it was implanted, or with retained parts of the knee or referred pain from elsewhere. Poor outcomes may lead to revisions, which may or may not be successful. This is a particular problem with Unicompartmental Knee replacement (UKR): Although UKR have fewer poor outcomes than Total Knee Replacements (TKR), they are about five times more likely to be revised if they do have a poor outcome, and many of these revisions do not help [1]. This is in part because surgeons reviewing radiographs often cannot determine why patients have poor outcomes and, therefore, have difficulty deciding how best to treat them. If it were possible to develop, using machine learning (ML), a better method to interpret post-operative radiographs to determine why patients have poor outcomes, this would aid in the management of these patients. More importantly, it could suggest ways that implants, indications or techniques could be improved so as to decrease the incidence of poor results.

Deep Convolutional Neural Networks (DCNN) can potentially improve the interpretation of post-operative radiographs and determine their correlation with poor outcomes. It is widely agreed that DCNN models can be trained to replicate how clinicians diagnose conditions from radiographs. For instance, a DCNN was trained and used for automatic Kellgren-Lawrence classification for knee osteoarthritis, achieving an accuracy of 81% [2]. DCNNs can also be trained to detect specific patterns and relationships within the images, which can be used to diagnose known medical conditions [3]. For example, it can detect signs of osteoporosis in bone scans [4], recognise fractures [5,6], and identify tumours [7,8]. While the current trend is to train an AI model to reproduce human classification, we wanted to explore whether a DCNN can be trained to predict poor patient-reported outcomes from anterior-posterior radiographs.

Training DCNN for medical applications can be challenging due to the need for significant amounts of labelled data, which is often impractical. Transfer learning offers a practical solution using a pre-trained model that can be retrained with a smaller dataset for a related task. This approach fine-tunes the existing algorithm and statistical weights for specific new uses. In this study, ResNet50v2, a deep model trained on over a million ImageNet dataset images and capable of classifying up to 1000 image categories [9,10], is a suitable candidate to be applied to classify patient-reported outcomes based on radiographs.

The primary aim of the current study was to train a Convolutional Neural Network (CNN) model to look at one-year post-operative anterior-posterior (AP) X-rays and automatically categorise them as patients with either poor or excellent clinical outcomes. In our UKR practice, about 5% of patients have poor clinical outcomes at one year based on the Oxford Knee Score (OKS). Secondly, with a separate set of radiographs that the model had not seen, to determine how successful it was at identifying patients with poor outcomes compared to two experienced orthopaedic surgeons. The third objective was to attempt to identify image features associated with poor outcomes.

2. Material and methods

2.1. Dataset

Cementless Oxford UKR were implanted by two experienced surgeons for the recommended indications of anteromedial osteoarthritis or avascular necrosis and the recommended surgical technique [11]. One year post-operatively, patients completed an OKS, which was categorised into Excellent, Good, Fair and Poor [12]. Radiographs, including an AP aligned on the tibial component, were obtained.

Five hundred and twenty-seven well-aligned anterior-posterior radiographs were used for the study. Given that the images contained both left and right knees, we opted to augment the dataset by duplicating the images and flipping them. Subsequently, 937 images out of 1054 were used for model training. We extracted 70 images to be used as the testing set, of

which 14 were from patients with a poor score. The images carried some text in the corners, such as codes, scales, and left/right labels. We cropped these out by setting a region of interest around the components and also to standardise the images in terms of content. We employed a square shape, with the top boundary placed just above the patella and the lower boundary just below the fibular head. Fig. 1 demonstrates the manual crop step before model training.

Training and validation datasets were prepared using the Keras Python package [13,14]. Cropped images sized 480-by-480 pixels were fed into the model in batch sizes of 8. Bilinear interpolation was used to ensure consistent sizing of the images, and a 90–10 training-validation split was implemented using a random function with a seed value of 42.

Pre-processing steps were performed on the images before they were fed into the neural network for training; these were resizing and rescaling layers. A resizing function was used to compress the images into 224-by-224 pixels images. We also performed a step that rescaled each pixel values from their original RGB values in the [0, 255] range to a [0, 1] binary range. These preprocessing steps were done for compatibility with the CNN model we chose (ResNe50v2).

2.2. Transfer learning method with pre-trained CNN model

We employed "transfer learning" by using the pre-trained ResNet50v2 model. It was developed to perform image classification by training on datasets that typically contain millions of images which can be assigned to a thousand categories. To adapt the model to our specific task, we fine-tuned and re-trained the model by retraining its layers to classify knee X-ray images into three categories based on Oxford Knee Score: 'Fair and Poor' (below 34 points), 'Good' (35–40 points), and 'Excellent' (above 41 points). Fig. 2 illustrates the complete architecture of our CNN model, encompassing both the retained pre-trained layers and the newly trained elements.

2.3. The training process

The CNN was constructed and trained on a regular PC using open-source software: TensorFlow-Keras. The computer had a 2.66 GHz 6-core Intel Core i7 processor, an AMD Radeon Pro 5300 M graphics card, and 16 GB of memory. During the 20 rounds (or epochs) of training, we monitored important metrics such as accuracy and loss to ensure the model was effectively learning. The training process and an example operation are shown in Fig. 3. When a new image is shown to the model, after training is complete, a score indicating the probability that the image belongs to each category is generated. The highest score is taken as the model's estimation.

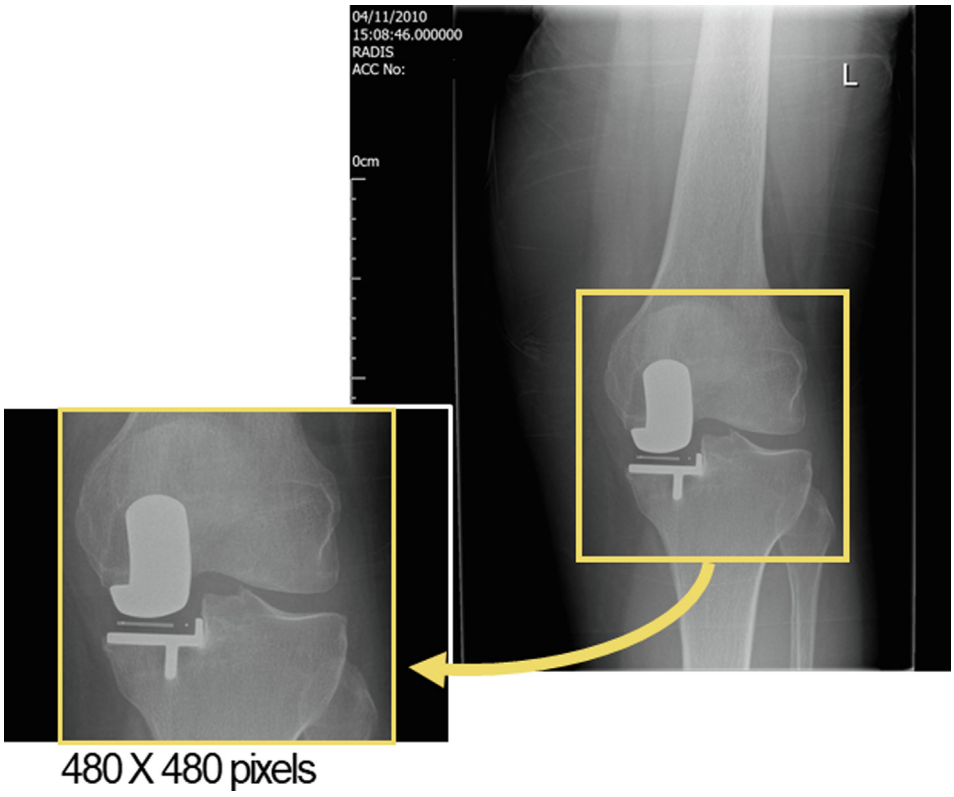

Fig. 1. Manual pre-processing before CNN training: A square box was used to crop aligned AP images, with the top edge along the edge of the patella and the entire tibiofemoral joint line included.

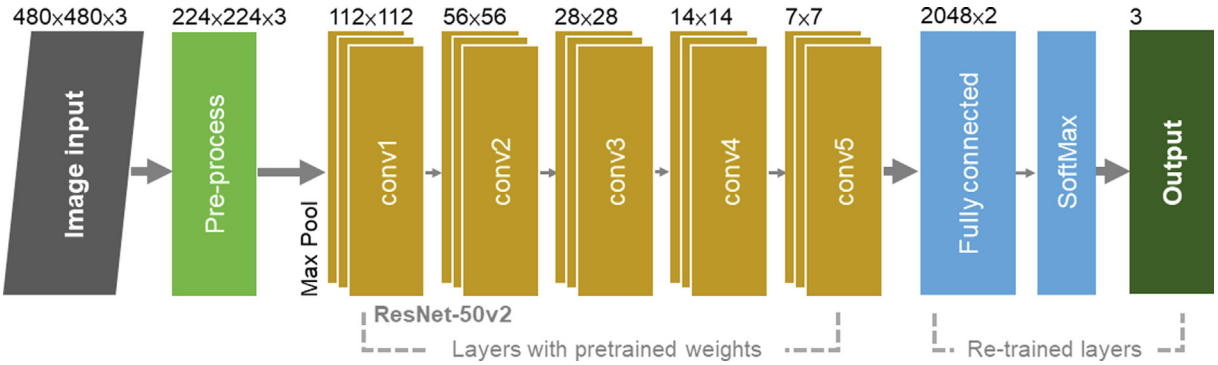

Fig. 2. The customised ResNet50 architecture deployed in the proposed patient outcome classification task.

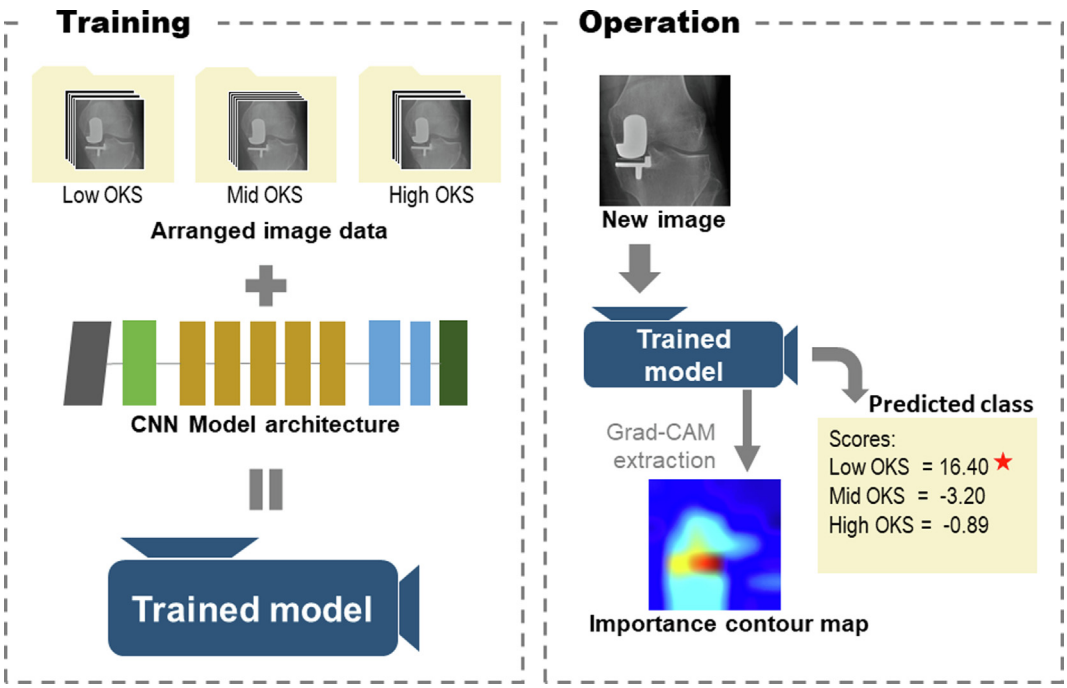

Fig. 3. Overview of how the CNN model was trained and implemented for new images.

Since our training data was small and the images in each category were similar, randomly splitting the data for validation could have a significant influence on the results. To make sure the results were reliable, we trained the model 10 times separately. The best-performing model was the one that correctly identified the most number of images with a poor OKS.

2.4. Comparison of model with orthopaedic surgeons

We performed a test to compare the performance of the best-performing ResNet50v2 model with that of experienced surgeons. The test dataset comprised 70 images with only Excellent and Poor OKS. Two experienced orthopaedic surgeons independently reviewed the images and produced categorised lists under a double-blinded process. We evaluated the performance of an AI model and surgeons using accuracy and the F1 score as critical metrics. Accuracy measures the percentage of correct predictions. However, accuracy alone can be misleading when the distribution of cases is imbalanced, such as in this study where the number of poor and excellent cases is not equal. For instance, even if a rater classifies all cases as excellent, the accuracy would still be 0.80. On the other hand, the F1 score considers both false positive and false negative errors and provides a single metric to evaluate the overall performance. A higher F1 score indicates a better ability to identify both positive and negative cases, which can help to gain a better understanding of the model's performance.

2.5. Evaluation of model classification

After training the model, we employed a Gradient-weighted Class Activation Mapping (Grad-CAM) method to try to visually verify how the model predicted OKS outcomes from radiographs. Grad-CAM uses gradients from the final convolutional processing step (layer) of the trained model to generate a coarse localisation map, which helps identify important regions within the images for prediction and categorisation purposes. By using this approach, we were able to extract a contour map that highlights the image areas most crucial for the model's classification decisions.

3. Results

Table 1 compares the classification results of 10 independent instances of trained ResNet50v2 to check if the CNN model accurately identified Poor cases. Model number 04 identified ten poor images (71%) that were used in the evaluation test against humans. The worst-performing trained model was number 5, which identified only 4 poor images (28.5%). On average, across all ten instances, the correct classification was 6.4 images (45.7%). Four images (numbers 03, 04, 06, and 12) were picked up by more than 70% of the trained models, and less than 30% of the models identified four images (numbers 05, 10, 11 and 14) as Poor images. The other six images were identified half the time by various trained models.

The confusion matrices presented in Table 2 illustrate the results of the classification performed by a computer model and two experienced orthopaedic surgeons. The rows indicate the actual class based on OKS, with a total of 14 poor outcome images. The columns represent the estimated class, showing that out of those 14, the computer identified 10 as having poor outcomes (71%). The computer identified 46 out of the 56 Good/Excellent results correctly. Overall, the model accuracy was 0.80 and F1 score 0.59. One of the experienced orthopaedic surgeons identified a single knee with a poor outcome, while the other was unable to identify any. The surgeons both identified 53 out of 56 patients with a good outcome. The accuracy of the surgeons was 0.77 (F1 score 0.11) and 0.76 (unable to calculate F1 score).

The importance of image features in each area towards the final classification estimation is demonstrated by the confusion matrix depicted in Fig. 4a. This matrix includes four featured contour maps generated by the Gradient-weighted Class Activation Mapping method based on the final convolutional layer from the CNN model. Red is indicative of the most critical features, while blue indicates the least important ones. Additionally, Fig. 4b illustrates the average grad-cam matrices across all images in each quadrant of the confusion matrix.

Table 1
Poor list comparison between ten independent trained ResNet50v2 models.

	Poor AP image numbers:													
	01	02	03	04	05	06	07	08	09	10	11	12	13	14
Model #01			○	○		○			○		○	○	○	
Model #02		○	○	○		○						○		
Model #03			○	○	○	○		○			○		○	
Model #04	○	○	○	○	○		○	○	○			○		○
Model #05	○		○			○						○		
Model #06		○		○		○						○		
Model #07		○	○				○	○				○		
Model #08			○	○		○			○			○	○	
Model #09	○		○			○	○				○	○	○	
Model #10	○	○	○		○		○	○	○			○		

*○: correctly classified by the ResNet50v2 training instance.

Table 2
Confusion matrices and performance metrics of the classification performed by a computer model and two surgeons.

		ResNet-50V2		Orthopaedic I		Orthopaedic II	
		Estimation					
		Poor	Not Poor	Poor	Not Poor	Poor	Not Poor
Data	Poor	10	4	1	13	0	14
	Not Poor	10	46	3	53	3	53
	Precision	0.50		0.25		0.00	
	Recall	0.71		0.07		0.00	
	F1-score	0.59		0.11		n/a	
	Accuracy	0.80		0.77		0.76	

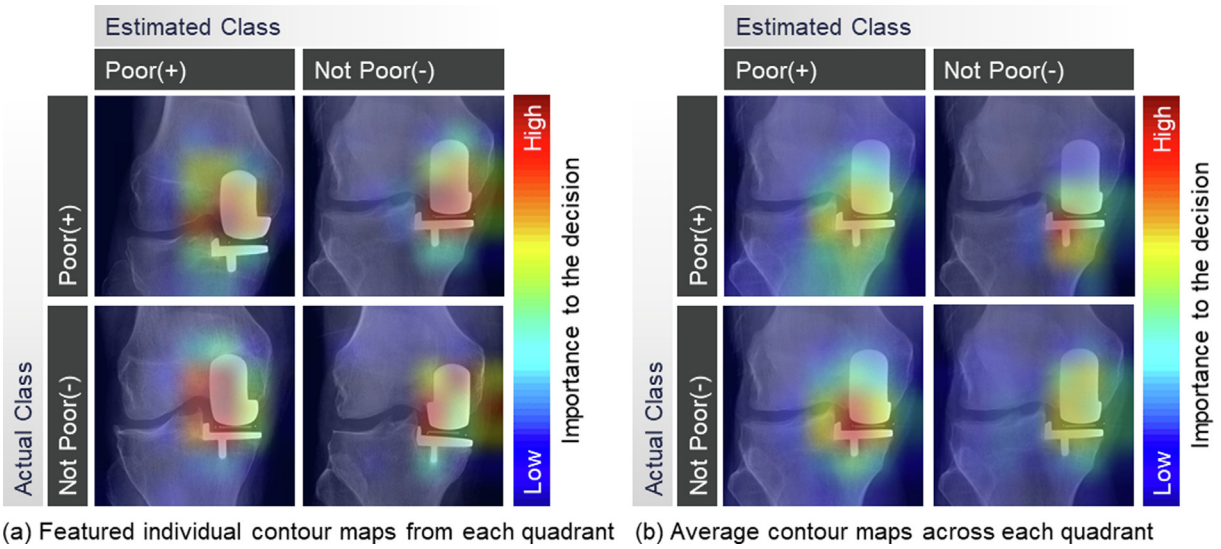

Fig. 4. Grad-CAM visualisation: contour maps demonstrate the importance of each area of the image in the classification. The areas marked in red contain features that contributed the most to the output.

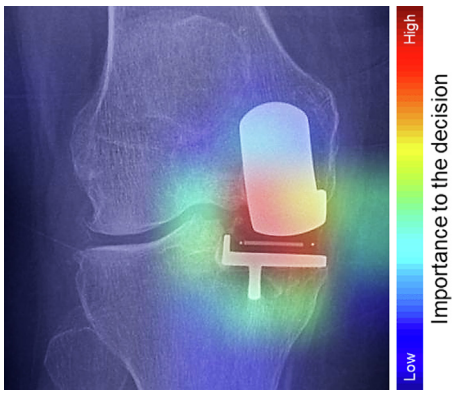

Fig. 5. Example image related to a poor OKS. The surgeon suspects the patient has a poor outcome as developing lateral osteoarthritis is seen, but the computer model estimates the poor outcome based on the image features around the implant.

Fig. 5 is the only image associated with a poor outcome that was correctly identified by a surgeon. The image is overlapped with Gradient-weighted Class Activation Mapping to demonstrate some key areas of the image that contributed to the classification of the ResNet50v2.

4. Discussions

The results of this study suggest that there are hitherto unknown features on AP knee radiographs of patients with OUKR at one year post-op that are predictive of the clinical outcome at that time point. There were fourteen radiographs from patients with poor outcomes: One surgeon could not identify any of these radiographs, and the other surgeon identified one. In contrast, the ResNet50v2 model identified ten patients who had reported a poor outcome in their one-year OKS.

CNNs are commonly described as operating in a ”black box” fashion. However, we utilised Grad-CAM to extract importance scores from subregions of the input images, allowing us to visualise which areas were most important to the classification. Using this approach, it was found that whether the classification was right or wrong, the most important features contributing to the final classification were always around the components. Interestingly, for those cases considered to be poor by the ResNet50v2 model, the Grad-CAM tended to identify more features in the region of the intercondylar notch and tibial spine, whereas, for those considered to be good, it tended to identify more features in the region of the femoral component and bearing. Features relating to the lateral side or patella did not seem to be important in making the decision. This is perhaps not surprising as it is known that the state of the Patello-femoral joint does not influence the outcome of the OUKR, and at one year the lateral side is unlikely to cause a problem.

By typical machine learning standards, the model in the present work has low overall accuracy (0.80) and F1 score (0.59), as achieving an accuracy of above 0.99 is a common expectation. However, in certain use cases, such as the present cases of the UKR patient cohort, even if the model fails to meet this criterion, it can still be valuable and provide meaningful clinical information. For instance, the images with poor outcomes that were not identified by the CNN model may have an extra-articular problem with referred pain causing their poor outcome. It is important to note that experienced orthopaedic surgeons were unable to reliably identify any of the images with a poor outcome. Therefore, applying the CNN model can potentially provide invaluable insights into possible pathophysiology of the joint, and it is worth studying further what subtle changes in the images are related to poor clinical outcomes and learning how to avoid the cause.

One surgeon correctly identified one of the radiographs associated with a poor outcome. However, this may have been identified for the wrong reason (Fig. 5). The surgeon believed this patient had a poor outcome because of lateral OA associated with a narrow joint space. The other surgeon, although identifying possible narrowing, did not feel that it was severe enough, at one year, to result in a poor outcome. The ResNet50v2 algorithm correctly identified this case based on information from around the implants rather than the appearance of the lateral side. This suggests that the poor outcome was not related to the lateral side.

We conducted several training sessions to ensure the strength of our CNN model. Since having a small dataset poses a challenge, we compared the lists generated by each independent training session to find any overlaps or inconsistencies to ensure the reliability of our model.

Upon analysis, we observed that our model had difficulty in consistently identifying certain images, specifically numbers 5, 10, 11, and 14. On the other hand, our model was able to correctly classify other images, such as numbers 3, 4, 6, and 12, in the majority of instances. These findings suggest that images 3, 4, 6 and 12 are highly likely to have abnormalities on the X-ray associated with the poor outcome, whereas 5, 10, 11, and 14 may have an extra-articular problem. The fact that neither of the surgeons identified images 3, 4, 6 and 12 suggests that the features on these images associated with poor outcomes are unknown to the surgeons. We expect that further research based on the current dataset would allow us to identify these features, particularly if it were enlarged by including more of our patients. We can achieve this through experimentation with various CNN models or by leveraging explainable AI. Once identified, we should be able to determine if the features are related to a cause of the poor results and, if so, make modifications to the procedure so as to decrease the incidence of poor results.

This approach is novel as it deviates from the conventional use of CNN to replicate feasible visual tasks with an automatic algorithm, and we believe it has the potential to dramatically improve the results of knee replacement if it were introduced on a national scale. This would require collecting pre and post-operative radiographs of patients who have primary and revision knee replacements and combining this with data from a National Registry that collects pre-operative and post-operative (6 months or one year) patient-reported outcomes. With this data, it should be possible to identify commonly used implants and the optimal way to implant them: For individual surgeons, it should be possible to make recommendations about their technique, and for patients, it should be possible to identify the best treatment, whether they are considering either primary or revision surgery.

5. Conclusions

AI identified over half of the patients with poor outcomes from AP radiographs alone. The experienced orthopaedic surgeons could not reliably identify any of the patients with poor outcomes. The results suggest that some information in the X-rays relates to a poor outcome, which was hitherto unknown to us, and the Grad-CAM method suggested that the key features are around the implants. Further work is needed to identify these features. This would not only aid surgeons in the management of patients with poor outcomes but would also lead to improvements in indications, surgical techniques, and possibly implants.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:[One or more of the authors have received or will receive benefits for professional use from a commercial party, Zimmer Biomet, related indirectly to the subject of this article. In addition, benefits have been directed to a research fund, foundation, educational institution, or other non-profit organisation with which more of the authors are associated. (Versus Arthritis & Orthopaedic Research UK)].

Acknowledgements

SJT is an Orthopaedic Research UK Ealy Career Research Fellow (563). This study has been delivered through the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of Orthopaedic Research UK, the NIHR or the Department of Health and Social Care.

References

[1] Goodfellow JW, O'Connor JJ, Murray DW. A critique of revision rate as an outcome measure: re-interpretation of knee joint registry data. J Bone Joint Surg Br 2010;92(12):1628–31. doi: https://doi.org/10.1302/0301-620X.92B12.25193.

[2] Wang Y, Li S, Zhao B, Zhang J, Yang Y, Li B. A resnet-based approach for accurate radiographic diagnosis of knee osteoarthritis. CAAI Trans Intell Technol 2022;7(3):512–21. doi: https://doi.org/10.1049/cit2.12079.

[3] Lisacek-Kiosoglous AB, Powling AS, Fontalis A, Gabr A, Mazomenos E, Haddad FS. Artificial intelligence in orthopaedic surgery. Bone Jt Res 2023;12(7):447–54. doi: https://doi.org/10.1302/2046-3758.127.BJR-2023-0111.R1.

[4] Wani IM, Arora S. Osteoporosis diagnosis in knee x-rays by transfer learning based on convolution neural network. Multimed Tools Appl 2023;82(9):14193–217. doi: https://doi.org/10.1007/s11042-022-13911-y.

[5] Beyaz S, Acici K, Sumer E. Femoral neck fracture detection in x-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg 2020;31(2):175–83. doi: https://doi.org/10.5606/ehc.2020.72163.

[6] Pranata YD, Wang K-C, Wang J-C, Idram I, Lai J-Y, Liu J-W, Hsieh I-H. Deep learning and surf for automated classification and detection of calcaneus fractures in ct images. Comput Methods Prog Biomed 2019;171:27–37. doi: https://doi.org/10.1016/j.cmpb.2019.02.006.

[7] Park C-W, Oh S-J, Kim K-S, Jang M-C, Kim IS, Lee Y-K, Chung MJ, Cho BH, Seo S-W. Artificial intelligence-based classification of bone tumors in the proximal femur on plain radiographs: System development and validation. PLOS ONE 2022;17:1–14. doi: https://doi.org/10.1371/journal.pone.0264140.

[8] Zhou X, Wang H, Feng C, Xu R, He Y, Li L, Tu C. Emerging applications of deep learning in bone tumors: current advances and challenges. Front Oncol 2022;12:908873. doi: https://doi.org/10.3389/fonc.2022.908873.

[9] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conf Comput Vision Pattern Recogn (CVPR) 2016;2016:770–8. doi: https://doi.org/10.1109/CVPR.2016.90.

[10] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Computer Vision – ECCV 2016, Springer International Publishing; 2016. p. 630–45.

[11] Goodfellow J, O'Connor J, Pandit H, Dodd C, Murray D. Medial Indications other than AMOA. In: Unicompartmental Arthroplasty with the Oxford Knee. 2nd ed. Oxford: Goodfellow Publishers; 2015. p. 163–70. doi: https://doi.org/10.23912/978-1-910158-45-6-4359.

[12] Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, Dawson J. The use of the Oxford hip and knee scores. J Bone Jt Surg Br 2007;89-B(8):1010–4. doi: https://doi.org/10.1302/0301-620X.89B8.19424.

[13] Chollet F. Keras (2015) [cited 01/05/2024]. https://keras.io.

[14] Chollet F. Deep learning with Python, 2nd ed. New York: Manning Publications; 2021.