Diagnosing the Severity of Knee Osteoarthritis Using Regression Scores From Artificial Intelligence Convolution Neural Networks

Michael Fei, BA; Sarah Lu, BA; Jun Ho Chung, MD; Sherif Hassan, MD, PhD; Joseph Elsissy, MD; Brian A. Schneiderman, MD

abstract

Background: This study focused on using deep learning neural networks to classify the severity of osteoarthritis in the knee. A continuous regression score of osteoarthritis severity has yet to be explored using artificial intelligence machine learning, which could offer a more nuanced assessment of osteoarthritis. **Materials and Methods:** This study used 8260 radiographic images from The Osteoarthritis Initiative to develop and assess four neural network models (VGG16, EfficientNetV2 small, ResNet34, and DenseNet196). Each model generated a regressor score of the osteoarthritis severity based on Kellgren-Lawrence grading scale criteria. Primary performance outcomes assessed were area under the curve (AUC), accuracy, and mean absolute error (MAE) for each model. Secondary outcomes evaluated were precision, recall, and F-1 score. **Results:** The EfficientNet model architecture yielded the strongest AUC (0.83), accuracy (71%), and MAE (0.42) compared with VGG16 (AUC: 0.74; accuracy: 57%; MAE: 0.54), ResNet34 (AUC: 0.76; accuracy: 60%; MAE: 0.53), and DenseNet196 (AUC: 0.78; accuracy: 62%; MAE: 0.49). **Conclusion:** Convolutional neural networks offer an automated and accurate way to quickly assess and diagnose knee radiographs for osteoarthritis. The regression score models evaluated in this study demonstrated superior AUC, accuracy, and MAE compared with standard convolutional neural network models. The EfficientNet model exhibited the best overall performance, including the highest AUC (0.83) noted in the literature. The artificial intelligence-generated regressor exhibits a finer progression of knee osteoarthritis by quantifying severity of various hallmark features. Potential applications for this technology include its use as a screening tool in determining patient suitability for orthopedic referral. [*Orthopedics*. 2024;47(5):e247-e254.]

Osteoarthritis (OA) of the knee is a common degenerative bone pathology characterized by cartilage degeneration, synovial distention, synovial inflammation, and reactive bone hyperplasia.[1] Early detection and accurate classification can provide the patient with a wider range of treatment options, dictated by the patient's symptoms and radiographic evidence of disease.[2] Plain radiography remains a critical tool in evaluating OA severity and directing appropriate management. The radiographic hallmarks of OA include joint space narrowing, osteophyte formation, subchon-

From Creighton University School of Medicine, Phoenix, Arizona (MF); and California University of Science and Medicine, Colton (SL), the Department of Orthopedic Surgery, Loma Linda University, Loma Linda (JHC, JE, BAS), and the University of California, Riverside, Riverside (SH), California.

Disclosure: JE is a consultant for DePuy Synthes and Arbutus Medical. The remaining authors have disclosed no potential conflicts of interest, financial or otherwise.

Address correspondence to Michael Fei, BA, 31 E Thomas Rd, Apt 246, Phoenix, AZ 85012; email: Mfei1225@gmail.com.

Submitted: March 9, 2024. Accepted: June 20, 2024. Published online: July 31, 2024.

doi: 10.3928/01477447-20240718-02

dral sclerosis and cysts, and osseous deformity, which can all be visualized on plain radiography.[3] Various other imaging modalities can be used to assess OA, such as computed tomography and magnetic resonance imaging. However, the limited clinical benefits offered by these studies over plain radiography do not routinely justify the associated time and cost.

The Kellgren-Lawrence (KL) grading scale has commonly been used to grade OA severity radiographically. With this method, plain radiographs are used to classify knee OA on a discrete scale ranging from 0 to 4, with 0 representing no evidence of OA and 4 signifying advanced disease. While considered the gold standard for assessing OA on radiographic imaging, the KL scale is often criticized for being subjective and having suboptimal interrater reliability.[4] Other solutions have been proposed to address this limitation but ultimately provide less granularity in describing OA severity.[5,6]

Deep learning neural networks, a subsection of machine learning (ML), may be able to address the reproducibility issues inherent to the KL grading among other benefits. A ML solution potentially provides a standardized and rapid radiographic analysis to clinicians, which can be used to decrease subjectivity. In a study examining the efficacy of artificial intelligence compared with physicians when diagnosing fractures radiographically, a ML algorithm improved sensitivity, maintained specificity, and shortened reading time.[7] There may be benefits in OA diagnosis as well. Another benefit of ML is that it may allow for the use of the KL scale as a continuous score, rather than as discrete values, which could provide further granularity when determining OA severity. Currently, there are no studies investigating the use of ML to grade knee OA on a continuous spectrum. Unlike traditional ML models, neural network models iteratively learn and improve from previous mistakes, making them one of the most powerful ML models.[8] Convolutional neural networks (CNNs) are a subset of neural networks designed to analyze and learn from images. CNNs process groups of pixels together rather than an individual pixel, allowing them to identify and find patterns within a given image.[9]

The specific aims of this study were to train deep learning neural networks to provide a KL grade on a continuous regression scale of 0 to 4 (ie, decimal numbers ranging from 0 to 4) and to assess their performance in characterizing OA based on plain radiographs.

Materials and Methods

Dataset

The Osteoarthritis Initiative is a public database that contains normal and pathologic knee radiographs with varying severity of OA. The images are labeled by their KL classification by two orthopedic surgeons. The KL grading scale criteria are as follows: grade 0, no pathological features; grade 1, doubtful narrowing of joint space and possible osteophytic lipping; grade 2, definite osteophytes and possible narrowing of joint space; grade 3, moderate multiple osteophytes, definite narrowing of joint space, some sclerosis, and possible deformity of bony ends; and grade 4, large osteophytes, marked narrowing of joint space, severe sclerosis, and definite deformity of bone ends.[10]

The current study analyzed all 8260 available radiographic images from The Osteoarthritis Initiative, an open-source dataset of 4796 patients 45 to 79 years old. Throughout various disciplines that use neural networks, a standard practice is to partition datasets into 80% for training, 10% for model evaluation, and 10% for testing.[6,11,12] The more data used to train the model, the stronger the performance. However, a sufficiently large test dataset needs to be preserved to benchmark and test the model's abilities to generalize to unseen data. Because there were no special constraints within our dataset, the standard 80/10/10 split was used. Therefore, 6608 images were used to train the model, 826 images were used to evaluate the model's performance after each training cycle, and 826 images were used for testing.

Data Preprocessing

The radiographic images were first transformed using contrast-limited adaptive histogram equalization to improve the contrast with a clip size of 5 and tiles of (8, 8) as proposed by Ahmed and Mstafa.[13] Random augmentations were then performed, a process by which training images were randomly rotated between 0° and 15°, randomly zoomed in or out 0% to 10%, and/or randomly flipped over the y-axis.[6,14] This technique artificially creates new training images and improves algorithm performance.[15] It also prevents overfitting, a phenomenon in which the model memorizes training images rather than learns to recognize features of interest. Overfitting can lead to poor external validity.[16]

Model Architecture

The neural network was developed using the PyTorch[17] and MONAI[18] frameworks. The models were run on a local workstation using a graphics processing unit (RTX 3080; Nvidia Corporation). The base CNN architectures from PyTorch Torchvision were loaded with their pre-trained values.[19] For our experiments, the CNN architectures used were VGG16, EfficientNetV2 small, ResNet34, and DenseNet196. These pre-built model architectures are widely researched, are publicly available, and have been used in many medical applications.[20-23]

By this point, the models had identified key radiographic features associated with OA, such as osteophytes or decreased joint space, storing them as smaller matrices that represent part of the original image. The models were then wrapped with a rectified linear unit layer and global average pooling layer that assigned each feature a value based on the likelihood a feature is present in the image. Finally, a fully connected layer linked to a singular

output node provided the final severity score. A visual representation of this pipeline is provided in **Figure 1**.

Training Parameters

Neural networks need to be trained by looping through the data multiple times. The models used in this study were trained for 50 epochs (an entire pass of training data through the algorithm). After each epoch, the performance is evaluated using a loss function. A mean square loss function was used in this regression model to account for the ordinal nature of KL classification of OA (KL3 is more similar to KL2 than KL1). The loss function assigns a loss value that is inversely proportional to the model's performance, thus gaining a penalty when the predicted value is incorrect (**Figure 2**).

Using the loss function, a gradient vector, a vector that points in the direction that minimizes the loss function, is calculated with back-propagation. The weights and bias of the neural network are stepped in the direction of the gradient, decreasing the loss and allowing the model to learn. RMSprop optimizer was used to step the model in the direction of the gradient. The code used to preprocess and run the model can be found at https://github.com/mfei1225/OA_Knee_Regression.

Evaluation

The models evaluated the test dataset and produced a decimal predicted score for each image. Because KL grades are reported in discrete values, predicted scores were rounded to the nearest whole number according to **Table 1**. Scores greater than 4 were rounded down to 4, and scores less than 0 were rounded to 0. From these rounded values, precision, recall, and F-1 were calculated for each KL grade. Standard evaluation metrics used to assess performance in classification tasks were calculated, including accuracy, the area under the receiver operating characteristic curve (AUC), and the mean absolute error (MAE).[24] An AUC score of 0.5 means no discrimination, 0.7 to 0.8 indicate acceptable performance, and 0.8 to 0.9 indicate excellent performance.[25] MAE is another popular metric. A high MAE is suggestive of large discrepancy between true labels and predictions, while a MAE of 0 means perfect algorithm performance. Finally, a *t* test of the absolute error was performed between each model to determine significant statistical differences.

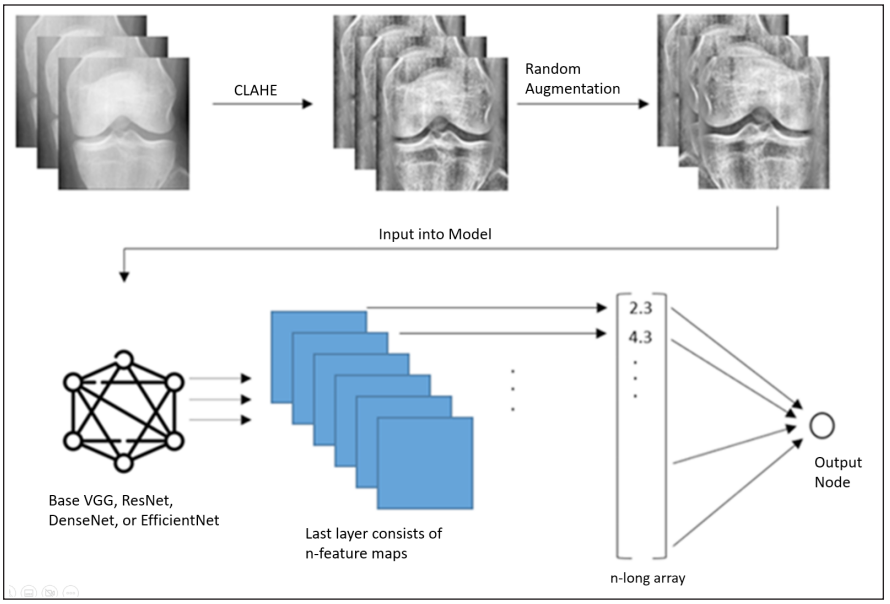

Figure 1: Pipeline of image transformation and model architecture. Abbreviation: CLAHE, contrast-limited adaptive histogram equalization.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Label\_value - Predicted\_value)^2$$

Figure 2: Equation for mean square error (MSE) where N is number of samples.

Table 1

Conversion of Model Score to Kellgren-Lawrence (KL) Score

Predicted model score	Rounded KL score
Model score<0.5	0
0.5≤model score<1.5	1
1.5≤model score<2.5	2
2.5≤model score<3.5	3
3.5≤model score	4

Results

The accuracy of VGG, ResNet, DenseNet, and EfficientNet was 57%, 60%, 62%, and 71%, respectively. AUC score was 0.74, 0.76, 0.78, and 0.83 for the respective models. MAE score was 0.54, 0.53, 0.49, and 0.42 (**Table 2**). When examining the precision for each KL class, EfficientNet had the best performance of all the models with 0.81, 0.42, 0.79, 0.75, and 0.63 for KL0, KL1, KL2, KL3, and KL4, respectively. When examining the recall for each KL class, EfficientNet had the best performance for KL0, KL2, KL3, and KL4 with 0.84, 0.59, 0.79, and 1.00, respectively. Only the VGG model had better recall for KL1 with 0.59. For the F-1 scores, which combine the precision and recall scores, EfficientNet outperformed the other models with scores of 0.82, 0.45, 0.68, 0.77, and 0.78 for KL0, KL1, KL2, KL3, and KL4, respectively. The *t* test suggested that the absolute error of EfficientNet was significantly less than the absolute error of the other three models with *P* values less than .05 (**Figure 3**).

Table 2

Performance Metrics of the Models

Model/metric	KL0	KL1	KL2	KL3	KL4	Overall
VGG16						
Precision	0.77	0.30	0.66	0.73	0.58	
Recall	0.68	0.59	0.35	0.60	0.74	
F-1	0.72	0.40	0.46	0.66	0.65	
MAE						0.54
Accuracy						57%
AUC						0.74
ResNet34						
Precision	0.73	0.31	0.66	0.60	0.61	
Recall	0.74	0.38	0.46	0.70	0.89	
F-1	0.73	0.34	0.54	0.65	0.72	
MAE						0.53
Accuracy						60%
AUC						0.76
DenseNet196						
Precision	0.73	0.31	0.66	0.60	0.61	
Recall	0.74	0.38	0.46	0.70	0.89	
F-1	0.73	0.34	0.54	0.65	0.72	
MAE						0.49
Accuracy						62%
AUC						0.78
EfficientNetV2 small						
Precision	0.81	0.42	0.79	0.75	0.63	
Recall	0.84	0.48	0.59	0.79	1.00	
F-1	0.82	0.45	0.68	0.77	0.78	
MAE						0.42
Accuracy						71%
AUC						0.83

Abbreviations: AUC, area under the receiver operating characteristic curve; KL, Kellgren-Lawrence; MAE, mean absolute error.

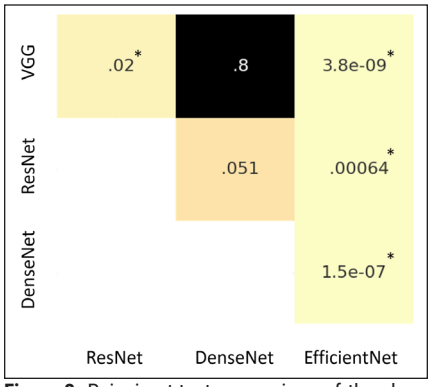

Figure 3: Pairwise *t* test comparison of the absolute error from each model with *P* values. *$P<.05$.

Discussion

This study investigated whether CNNs could sufficiently diagnose OA severity from radiographs using the KL grading scale. These results may have clinical applications, such as being used to make the KL system more consistent with less interobserver error or to create predictive disease progression models to inform physicians and patients about prognosis.

In this study, four different models were employed. The EfficientNet model performed significantly better the other models when comparing the absolute error. This is also confirmed with the highest accuracy of 71%, highest AUC of 0.83, and lowest MAE of 0.42. The low MAE suggests that even when the model was wrong, the prediction was very close and on average 0.42 away from the labeled grade. Precision, recall, and F-1 scores for individual KL grades show that all four models tend to have strong performance in correctly predicting arthritis at stages KL2, KL3, and KL4. In contrast, KL1 was often misclassified by all models as KL0, which is consistent with other models in the literature.[13,21] Generally, a higher precision means more false-negative results and thus a lower recall, and a higher recall means more false-positive results and thus a lower precision. However, for EfficientNet, the precision, recall, and F-1 suggest that there was little tradeoff between precision and recall, with this model having the highest metric in almost all KL classes. In the one case where the VGG model has the higher recall for the KL1 score, there is a relatively large tradeoff resulting in more false-positive results (**Figure 4**), lower precision, and overall lower F-1 score.

Comparison With Previous Literature

CNN models evaluated in this study demonstrated similar or improved performance compared with other previously reported KL classification models using the same dataset.[21,26-28] Tiulpin et al[26] used a Siamese deep neural network to achieve a 66.71% accuracy, albeit they employed a neural network with a more complex architecture. The complexity in the model was shown to add no additional predictive power at the expense of computing resource, which is limited and valuable in a hospital setting.

Furthermore, the current study used a mean square loss function. This differs from the traditional cross-entropy loss function, which is used when classification problems have no relationship between outputs. If a cross-entropy loss function were used in the context of the KL scale, the model would only learn from correct predictions. With a mean square loss function, the model learns from incorrect predictions as well. Chen et al[27] attempted to address the issues associated with a cross-entropy loss function by proposing an ordinal loss function that achieved 69.7% accuracy. However, the values in the ordinal loss function were manually chosen, making it difficult to reuse the same methods on a similar task with an ordinal scale.

Another ML model to evaluate knee OA was investigated by Wang et al.[21] It achieved a comparable accuracy of 71.7%. However, the authors rebalanced the dataset before training, possibly creating duplicate images in the dataset. This biases the model performance by creating more samples in KL4, KL3, and KL2, which have been shown to be significantly easier to identify than mild cases such as KL0 and KL1, influencing the overall accuracy. Thomas et al[28] combined grade 0 and grade 1 as one class, and grades 2, 3, and 4 as another class. They achieved an accuracy of 0.84. This method simplifies the problem but gives less useful information clinically. Regardless, if the same method is applied to the results of the EfficientNet model, an accuracy of 0.89 is achieved, indicating stronger performance. Methods of the current study built upon these previous protocols to decrease bias of the studied models.

Finally, when analyzing model performance using the AUC, a more encompassing metric in ML classification problems,[29] the EfficientNet AUC was 0.83. Janvier et al[30] achieved the highest AUC of 0.73 while using a standard logistic regression approach. Tiulpin et al[31] achieved an AUC of 0.79 using a multimodal ML approach. Given that the same dataset was used, the high AUC of EfficientNet demonstrates the strongest performance with less manual labor and less information fed into the model.

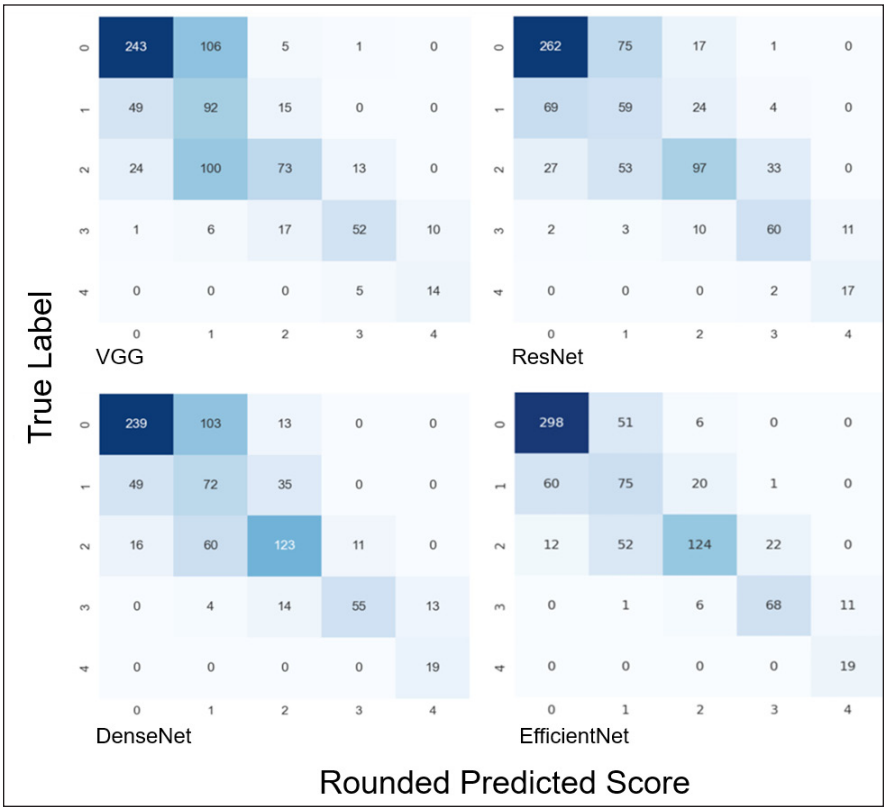

Figure 4: Confusion matrix for each model. The y-axis represents the actual Kellgren-Lawrence rating of the image, while the x-axis represents the rounded predicted output value.

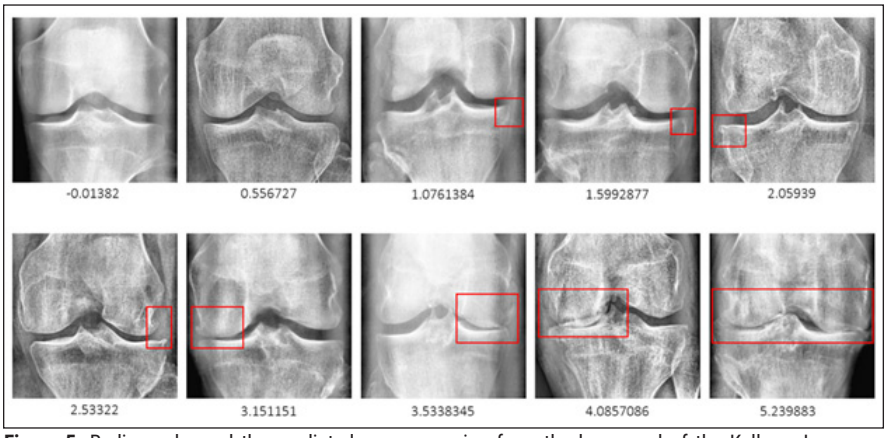

Figure 5: Radiographs and the predicted scores ranging from the lower end of the Kellgren-Lawrence spectrum to the upper end of the Kellgren-Lawrence spectrum. Red boxes outline features of osteoarthritis—osteophytic lipping, osteophytes, subchondral sclerosis, and joint space narrowing.

Qualitative Analysis of Model and Clinical Applications

This study was successful in creating a CNN model to identify and classify knee OA using images from a publicly available and validated dataset. The EfficientNet model was similarly powerful, if not more accurate, compared with other models used in the literature. To the best of our knowledge, this is the first study to propose a CNN model to evaluate and re-

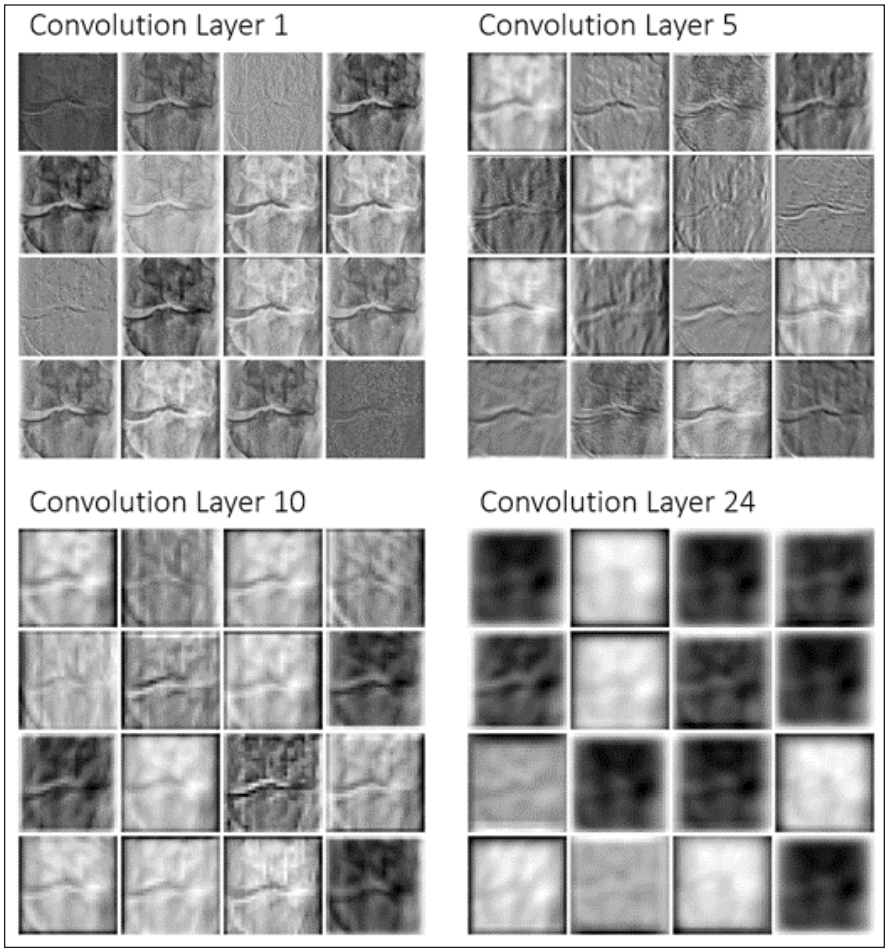

Figure 6: Sample filters from different layers of the EfficientNet model. Different filters at different convolution layers highlight different areas of the radiograph that the model was focused on, such as the femoral condyle, the tibia plateau, joint space, and cartilage.

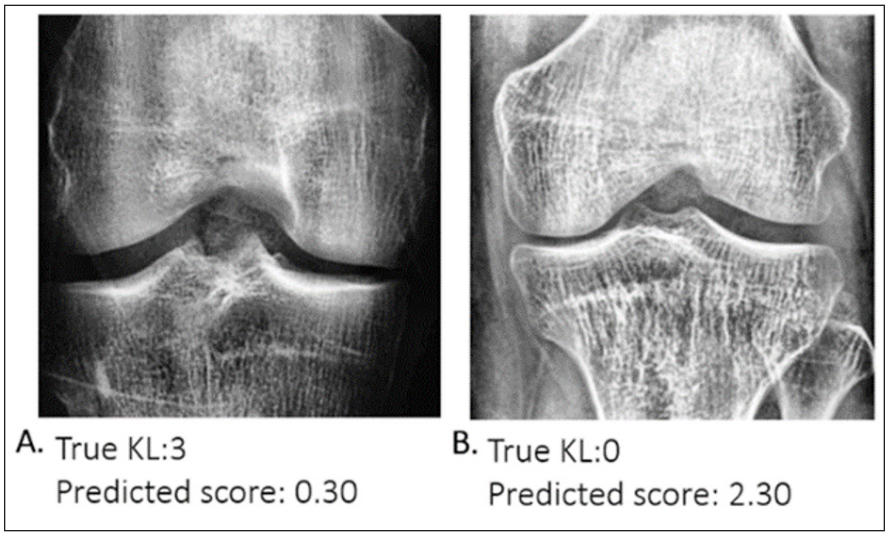

Figure 7: Anteroposterior radiographs that the EfficientNet model underestimated (A) and overestimated (B). Abbreviation: KL, Kellgren-Lawrence.

port KL knee OA grade on a continuous scale. That is, the model can subclassify radiographs by providing decimal values instead of whole number scores. By using a continuous output, OA grade can be evaluated on a more granular level than the traditional KL classification, which employs discrete values (**Figure 5**). A continuous score could aid in tracking the progression of OA severity over the time course of multiple visits; thus, future models could be created to estimate disease progression and provide a prognostic timeline. In addition, the filters seen in **Figure 6** provide some insight into the process by which models quantitatively interpret images. Different filters focus on and evaluate the severity of different individual OA features, namely, the degree of joint space narrowing, osteophyte formation, and subchondral sclerosis. As such, these features may have utility for providers when considering a patient's radiographic findings.

Potential applications of this technology are numerous. Particularly, it may have value as a clinical aid in the patient referral process as a diagnostic aid. While OA treatment is dictated by patient symptoms in addition to radiographic findings of disease, the ML algorithm described in this study may be a useful adjunct when screening patients and directing care appropriately. For an aging population in which joint replacement surgery is expected to grow significantly during the coming decades, automation of radiographic diagnostics associated with OA may allow for a more efficient and cost-effective process.[32] This model could offer primary care physicians, insurance representatives, and other health care professionals valuable guidance when considering a patient for referral to an orthopedic surgeon. In addition, other deep learning models have already used the KL score to accurately predict clinical presentation from radiological findings.[33] Thus, the regressor score can be incorporated to aid in estimating disease prognosis and when

intervention is needed to prevent impairment of activities of daily living.

Limitations

This study used images from The Osteoarthritis Initiative and is thus limited to the confinements of the database. Among the 8260 images analyzed, some outlying radiographs were identified that did not seem to have features consistent with the labeled KL grade. For instance, **Figure 7A** shows non-obvious osteophytes near the epiphysis but the true label KL grade is 3. Conversely, **Figure 7B** shows a knee radiograph with uneven joint space narrowing but the true label KL grade is 0. The poorly predicted images provide valuable insight in that the model is able to make consistent/precise decisions based on the features present in the knee radiograph.

A limitation with all ML models is the quality of the data. The model is only as good as the provided grading assessment labels. As with all ML studies, the provided labels, which were from two physicians, were assumed to be ground truth. However, the KL grading scale has been criticized for lacking objectivity and interrater reliability. With only two graders provided by The Osteoarthritis Initiative, there is potential for bias, error, and subjectivity that would influence the predictive nature of the CNN. To further address this limitation, a follow-up study is planned to compare the performance of the CNN against a larger group of practicing orthopedic surgeons.

Conclusion

This study presented a novel approach to consistently grading knee OA using a regression score. The model exhibited similar or improved performance compared with those currently reported in the literature. These features could be a diagnostic aid to health care professionals when they are deciding whether to refer to orthopedic surgeons or physical medicine and rehabilitation specialists for treatment of knee OA. Future studies can be performed to expand this algorithm to other imaging modalities.

References

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *Lancet*. 2019;393(10182):1745-1759. https://doi.org/10.1016/S0140-6736(19)30417-9 PMID:31034380
2. Chu CR, Williams AA, Coyle CH, Bowers ME. Early diagnosis to enable early treatment of pre-osteoarthritis. *Arthritis Res Ther*. 2012;14(3):212. https://doi.org/10.1186/ar3845 PMID:22682469
3. Glyn-Jones S, Palmer AJR, Agricola R, et al. Osteoarthritis. *Lancet*. 2015;386(9991):376-387. https://doi.org/10.1016/S0140-6736(14)60802-3 PMID:25748615
4. Hayes B, Kittelson A, Loyd B, Wellsandt E, Flug J, Stevens-Lapsley J. Assessing radiographic knee osteoarthritis: an online training tutorial for the Kellgren-Lawrence grading scale. *MedEdPORTAL*. 2016;12:10503. https://doi.org/10.15766/mep_2374-8265.10503 PMID:30984845
5. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J Digit Imaging*. 2019;32(3):471-477. https://doi.org/10.1007/s10278-018-0098-3 PMID:30306418
6. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol*. 2018;73(5):439-445. https://doi.org/10.1016/j.crad.2017.11.015 PMID:29269036
7. Guermazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*. 2022;302(3):627-636. https://doi.org/10.1148/radiol.210937 PMID:34931859
8. Krogh A. What are artificial neural networks? *Nat Biotechnol*. 2008;26(2):195-197. https://doi.org/10.1038/nbt1386 PMID:18259176
9. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: *2017 International Conference on Engineering and Technology (ICET)*. IEEE; 2017:1-6. https://doi.org/10.1109/ICEngTechnol.2017.8308186
10. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. *Ann Rheum Dis*. 1957;16(4):494-502. https://doi.org/10.1136/ard.16.4.494 PMID:13498604
11. Song J, Wang H, Liu Y, et al. End-to-end automatic differentiation of the coronavirus disease 2019 (COVID-19) from viral pneumonia based on chest CT. *Eur J Nucl Med Mol Imaging*. 2020;47(11):2516-2524. https://doi.org/10.1007/s00259-020-04929-1 PMID:32567006
12. Wang YC, Chan TCH, Sahakian AV. Real-time estimation of lesion depth and control of radiofrequency ablation within ex vivo animal tissues using a neural network. *Int J Hyperthermia*. 2018;34(7):1104-1113. https://doi.org/10.1080/02656736.2017.1416495 PMID:29301446
13. Ahmed SM, Mstafa RJ. Identifying severity grading of knee osteoarthritis from x-ray images using an efficient mixture of deep learning and machine learning models. *Diagnostics (Basel)*. 2022;12(12):2939. https://doi.org/10.3390/diagnostics12122939 PMID:36552945
14. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. 2018;321:321-331. https://doi.org/10.1016/j.neucom.2018.09.013
15. Howard AG. Some improvements on deep convolutional neural network based image classification. *arXiv [cs.CV]*. Published online December 19, 2013.
16. Salman S, Liu X. Overfitting mechanism and avoidance in deep neural networks. *arXiv [cs.LG]*. Published online January 19, 2019.
17. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. *arXiv [cs.LG]*. Published online December 3, 2019.
18. Cardoso MJ, Li W, Brown R, et al. MONAI: an open-source framework for deep learning in healthcare. *arXiv [cs.LG]*. Published online November 4, 2022.
19. Albardi F, Kabir HMD, Bhuiyan MMI, Kebria PM, Khosravi A, Nahavandi S. A comprehensive study on Torchvision pre-trained models for fine-grained inter-species classification. In: *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE; 2021:2767-2774. https://doi.org/10.1109/SMC52423.2021.9659161
20. Marques G, Agarwal D, de la Torre Díez I. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Appl Soft Comput*. 2020;96:106691. https://doi.org/10.1016/j.asoc.2020.106691 PMID:33519327
21. 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):512-521. https://doi.org/10.1049/cit2.12079
22. Yan K, Lu L, Summers RM. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE; 2018:1022-1025. https://doi.org/10.1109/ISBI.2018.8363745
23. Zhou T, Ye X, Lu H, Zheng X, Qiu S, Liu Y. Dense convolutional network and its application in medical image analysis. *BioMed Res Int*. 2022;2022:2384830. https://doi.org/10.1155/2022/2384830 PMID:35509707

24. Vujovic ŽĐ. Classification model evaluation metrics. *Int J Adv Comput Sci Appl*. 2021;12(6). https://doi.org/10.14569/IJACSA.2021.0120670

25. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d PMID:20736804

26. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci Rep*. 2018;8(1):1727. https://doi.org/10.1038/s41598-018-20132-7 PMID:29379060

27. Chen P, Gao L, Shi X, Allen K, Yang L. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Comput Med Imaging Graph*. 2019;75:84-92. https://doi.org/10.1016/j.compmedimag.2019.06.002 PMID:31238184

28. Thomas KA, Kidzinski L, Halilaj E, et al. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiol Artif Intell*. 2020;2(2):e190065. https://doi.org/10.1148/ryai.2020190065 PMID:32280948

29. Ling CX, Huang J, Zhang H. AUC: a better measure than accuracy in comparing learning algorithms. In: Xiang Y, Chaib-draa B (Eds), *Advances in Artificial Intelligence.* Canadian AI 2003. Lecture Notes in Computer Science, vol 2671. Springer. https://doi.org/10.1007/3-540-44886-1_25

30. Janvier T, Jennane R, Toumi H, Lespessailles E. Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the Osteoarthritis Initiative. *Osteoarthritis Cartilage*. 2017;25(12):2047-2054. https://doi.org/10.1016/j.joca.2017.09.004 PMID:28935435

31. Tiulpin A, Klein S, Bierma-Zeinstra SMA, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Sci Rep*. 2019;9(1):20038. https://doi.org/10.1038/s41598-019-56527-3 PMID:31882803

32. Quintana JM, Arostegui I, Escobar A, Azkarate J, Goenaga JI, Lafuente I. Prevalence of knee and hip osteoarthritis and the appropriateness of joint replacement in an older population. *Arch Intern Med*. 2008;168(14):1576-1584. https://doi.org/10.1001/archinte.168.14.1576 PMID:18663171

33. Guan B, Liu F, Mizaian AH, et al. Deep learning approach to predict pain progression in knee osteoarthritis. *Skeletal Radiol*. 2022;51(2):363-373. https://doi.org/10.1007/s00256-021-03773-0 PMID:33835240