Lee *et al. Knee Surgery & Related Research* (2024) 36:24
https://doi.org/10.1186/s43019-024-00228-3

Knee Surgery
& Related Research

RESEARCH ARTICLE

Open Access

Accurate, automated classification of radiographic knee osteoarthritis severity using a novel method of deep learning: Plug-in modules

Do Weon Lee[1,5], Dae Seok Song[2], Hyuk-Soo Han[1,3] and Du Hyun Ro[1,2,3,4*]

Abstract

Background Fine-grained classification deals with data with a large degree of similarity, such as cat or bird species, and similarly, knee osteoarthritis severity classification [Kellgren–Lawrence (KL) grading] is one such fine-grained classification task. Recently, a plug-in module (PIM) that can be integrated into convolutional neural-network-based or transformer-based networks has been shown to provide strong discriminative regions for fine-grained classification, with results that outperformed the previous deep learning models. PIM utilizes each pixel of an image as an independent feature and can subsequently better classify images with minor differences. It was hypothesized that, as a fine-grained classification task, knee osteoarthritis severity may be classified well using PIMs. The aim of the study was to develop this automated knee osteoarthritis classification model.

Methods A deep learning model that classifies knee osteoarthritis severity of a radiograph was developed utilizing PIMs. A retrospective analysis on prospectively collected data was performed. The model was trained and developed using the Osteoarthritis Initiative dataset and was subsequently tested on an independent dataset, the Multicenter Osteoarthritis Study (test set size: 17,040). The final deep learning model was designed through an ensemble of four different PIMs.

Results The accuracy of the model was 84%, 43%, 70%, 81%, and 96% for KL grade 0, 1, 2, 3, and 4, respectively, with an overall accuracy of 75.7%.

Conclusions The ensemble of PIMs could classify knee osteoarthritis severity using simple radiographs with a fine accuracy. Although improvements will be needed in the future, the model has been proven to have the potential to be clinically useful.

Keywords Knee osteoarthritis, Deep learning, Classification

*Correspondence:
Du Hyun Ro
duhyunro@gmail.com
Full list of author information is available at the end of the article

Background

Kellgren–Lawrence grade (KLG) is one of the most commonly used criteria in classifying the severity of knee osteoarthritis (KOA) on a simple radiograph [1]. It categorizes KOA severity from grade 0–4 on the basis of joint space narrowing, osteophyte formation, subchondral sclerosis, and bony deformity observed in a simple radiograph of the knee. Despite its apparent simplicity, KLG varies even between expert surgeons or radiologists [2]. This is because the classification system is not a quantitative system and thus is often confusing, especially when diagnosed by clinicians with less experience in the field. For this reason, it would be useful to develop an accurate, automated prediction model of KLG.

During the past few years, there have been several efforts [3–10] to automatically classify radiographic severity of a knee with the aid of convolutional neural network (CNN), and the results have been promising. Deep learning (DL) methods are commonly used in this automatic KLG grading task since large scale data are utilized to improve model accuracy. However, these models are not flawless, including utilization of data with relatively low accuracy and low quality or volume. The accuracy levels were especially low when it came to discerning lower grades such as Kellgren–Lawrence grade [1] (KLG) 0 or 1. The accuracy of KLG 1 was only 11%, although the overall accuracy was 67% in a recent study [6]. This may be due to the fact that KLG 1 has less distinctive features than other higher grades because of the definition of "doubtful joint space narrowing and possible osteophytic lipping." Therefore, the authors felt the need to design a new model that better predicts the KLG of a knee image using a more robust and objective dataset.

Recently, a plug-in module (PIM) [11] that can be integrated to CNN-based or transformer-based networks has been proposed to provide strongly discriminative regions for fine-grained classification, and the results have outperformed those of previous DL methods. Fine-grained classification deals with data with a large degree of similarity, such as cat species or bird species, and similarly, KLG image classification is one such fine-grained classification task. PIM utilizes each pixel of an image as an independent feature and can subsequently better classify images with minor differences. Therefore, the authors hypothesized that applying PIM in this task (KLG classification) might outperform the previous CNN-based models, especially in discerning lower grades of KLG that have less distinctive features.

The authors hypothesized that, as a fine-grained classification tasks, knee osteoarthritis severity may be classified well through the application of PIMs. Therefore, the purpose of this study was to develop a prediction model that automatically assesses the KLG of a knee image applying PIM. The authors tried to develop a model with better accuracy and better generalization in classifying the KLG of a knee than the models provided in the previous literature [3–10].

Materials and methods

Data composition

A retrospective analysis on prospectively collected data was performed. The dataset used for the study was a combination of two different open source datasets, the Osteoarthritis Initiative (OAI) [12] and Multicenter Osteoarthritis Study (MOST) [13]. OAI, which is funded by the National Institutes of Health, National Institute on Aging, and National Institute of Arthritis and Musculoskeletal and Skin Diseases, holds clinical data and X-ray images from 4796 individuals (41.5% men and 58.5% women) aged between 45 and 79 years. MOST is a similar project on osteoarthritis and holds data from 3026 individuals (60% men and 40% women) aged between 50 and 79 years old, although it has been recently closed to the public due to financial reasons. The details about the acquisition and protocols in the OAI and MOST studies are available online at https://nda.nih.gov/oai and http://most.ucsf.edu, respectively.

The authors included all available knee radiographs from these two projects except for the images that did not have KLG labels. Subsequently, all images were cropped by range of interest (ROI) to include only one knee per image using a detection model (DAMO-YOLO [14]; Fig. 1). During the training process, 1439 images were used as training data, and the images were annotated either as a ROI of a knee image or of a metal implant in YOLO format. A single DAMO-YOLO model was developed to detect knee images while simultaneously ruling out the knees with metal implants. This detection model failed in only about 30 cases in which image contrast was significantly poor due to morbid obesity or inadequate radiograph filming. Finally, 63,688 images were selected, of which 46,648 images were used for training (37,462 images) and validation (9186 images) sets, and the remaining 17,040 images were used for the test set. The proportion of the training set:validation set was 8:2. The images from the MOST dataset were used as the test set, while training and validation sets were randomly assigned from the OAI dataset. Training and validation sets were divided so that the images of each set were from different patients, meaning that images of an individual patient are either in the training or validation set and not divided into both sets. This method was used to

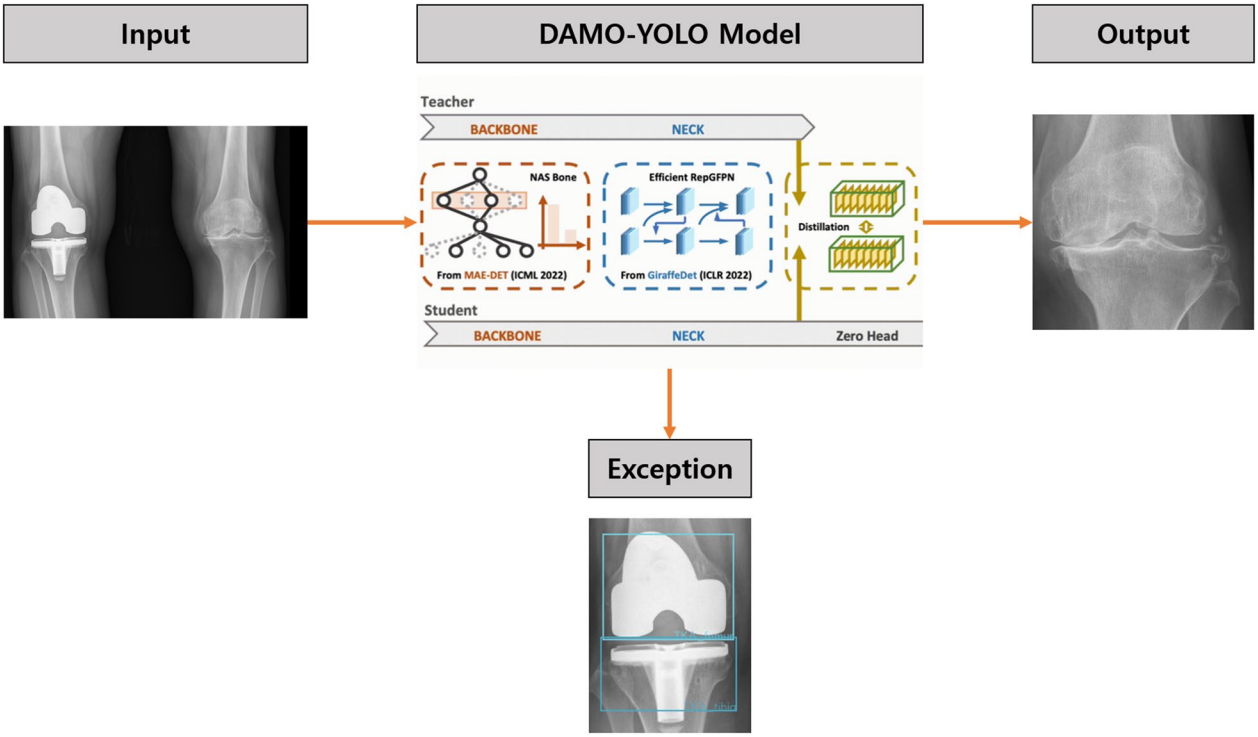

Fig. 1 Knee localization process of an image in this study

minimize the individual information unrelated to KLG that can be incorporated during model development.

Image preprocessing and augmentation

Image resizing was initially performed to 384 × 384 pixels and 448 × 448 pixels each for two different backbone models of PIM [11], Swin [15] and EfficientNet [16], respectively. Horizontal flipping was utilized, and two images were provided per test image: the original test image and horizontally flipped test image. Then, image quality augmentation [17] was performed using random brightness, sharpening, image compression, Gaussian noise, histogram equalization and contrast limited adaptive histogram equalization [18]. Lastly, grayscale values of each pixel of an image were normalized according to the average and standard deviation value of the whole pixels of all images used in the study (mean = 0.543, standard deviation = 0.203) [19].

Data labeling

Data labeling was performed using a vector for each KL (Kellgren–Lawrence) grade. The vector was designed also to either include one upper (or lower) grade to the initial KL grade to design a model that overestimates (or underestimates) the KL grade. This method, which has been first introduced as "label smoothing" in a previous literature [20], was attempted because prior KL grade models [3–6, 8, 9] and our trials showed that the models tend to confuse grade 1 and grade 2. Moreover, in a real clinical situation, even experts in this field sometimes overestimate or underestimate KLGs, and the authors tried to calibrate this. Several values were tried from 0.5 to 1 for the upper (or lower) grade labeling, and the vector that used 0.95 for the upper (or lower) grade labeling showed the most consistent accuracy for all grades. Therefore, each KL grade was labeled as shown below using this method.

Upper-grade labeling	Lower-grade labeling
Grade 0 → [1, 0.95, 0, 0, 0]	Grade 0 → [1, 0, 0, 0, 0]
Grade 1 → [0, 1, 0.95, 0, 0]	Grade 1 → [0.95, 1, 0, 0, 0]
Grade 2 → [0, 0, 1, 0.95, 0]	Grade 2 → [0, 0.95, 1, 0, 0]
Grade 3 → [0, 0, 0, 1, 0.95]	Grade 3 → [0, 0, 0.95, 1, 0]
Grade 4 → [0, 0, 0, 0, 1]	Grade 4 → [0, 0, 0, 0.95, 1]

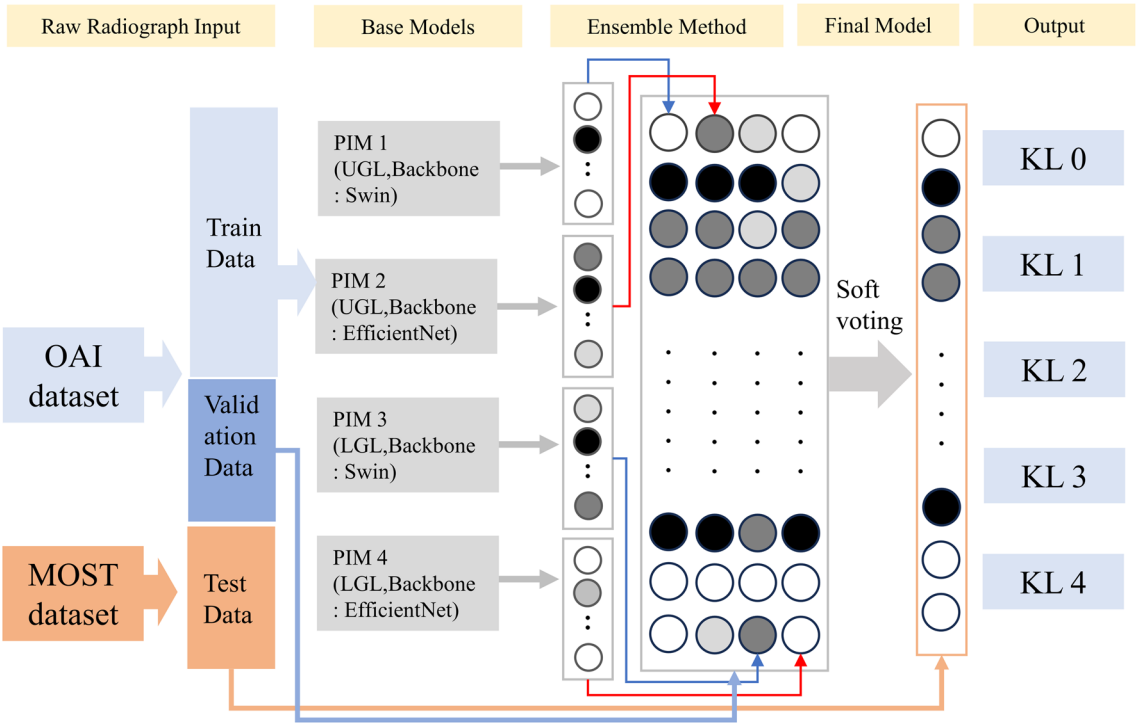

Fig. 2 The overall architecture of model development in this study; OAI, Osteoarthritis Initiative; MOST, Multicenter Osteoarthritis Study; PIM, plug-in module; UGL, upper-grade labeling; LGL, lower-grade labeling; KL, Kellgren–Lawrence

Final modeling

The final DL model was an ensemble of four different PIMs that used Swin [15] and EfficientNet [16] as the backbone models (two each). Soft voting [21], which uses the probabilities of each model to infer the value, was chosen as the ensemble method. Two of the models applied the upper-grade labeling method, one of which used Swin and the other EfficientNet as the backbone model. The other two models applied the lower-grade-labeling method instead, one of which used Swin and the other EfficientNet as the backbone model. PIM utilizes each pixel of an image as an independent feature, and these pixels are used as the inputs of the backbone blocks. The detailed structure of PIM is described in the original article [11]. The output of each model was a 1×5 vector that comprised weighted values per each class in which a higher value implies higher probability of a specific class. After training each of the four different PIMs, a Softmax function was additionally utilized to normalize the vector so that the sum of the values of the vector was 1. The Softmax function was not applied during each model (PIM) training but just afterwards. This additional process was performed before ensembling so that each PIM model could have equal contribution. Subsequently, weighted averages of the four models were used because the simple arithmetic sum of the four models seemed to overestimate the KLG in the validation set. After several trials, 2 was assigned for the two lower-grade-labeled models, while 1 was assigned for the two upper-grade-labeled models. Then, an arithmetic weighted sum average of four different vectors from the models was calculated as the final output. The class with the highest value in this average vector was selected as the class and compared with the ground-truth KLG, ranging from grade 0 to 4. The overall architecture of this model is depicted in Fig. 2, and the results of the model were augmented with Eigen-CAM [22] for visual explanations. Additionally, the computational complexity of the model was calculated using floating point operations (FLOPs) [23].

Results

The distribution of KL grades for training, validation, and test sets is presented in Table 1.

The confusion matrices of four different models that were used are shown in Fig. 3, and the accuracy was 62.0%, 71.7%, 63.6%, and 72.8%, respectively. Figures 4 and 5 present the confusion matrix and the receiver

Table 1 Data labeling method for each Kellgren-Lawrence grade in the study

Dataset (total number of images)	KL grade	Number of images (proportion, %)
Training set (37,462)	0	14,906 (39.8)
	1	7022 (18.7)
	2	9149 (24.4)
	3	4920 (13.1)
	4	1465 (3.9)
Validation set (9186)	0	3483 (37.9)
	1	1785 (19.4)
	2	2257 (24.6)
	3	1292 (14.1)
	4	396 (4.3)
Test set (17,040)	0	7148 (41.9)
	1	2689 (15.8)
	2	3024 (17.7)
	3	2922 (17.1)
	4	1257 (7.4)

KL, Kellgren–Lawrence

operating characteristic (ROC) curve of the ensemble DL model, respectively. The overall accuracy of the model was 75.7%, and the sensitivity and specificity for each KL grade are shown in Table 2. The accuracy was the lowest for KL grade 1 (46%) and the highest for KL grade 4 (93%). The FLOPs of the ensemble model were 565.34 G.

Prediction visual explanations

The visual explanations of the ensemble DL model for different grades of KL are shown in Fig. 6. We could identify different patterns across different class levels of KOA and the relevant features (represented as brighter uptake in the image) matched the expected KOA features (joint space narrowing and osteophytes) in the joint margins.

Discussion

The ensemble DL model of this study could predict KL grades with higher accuracy than most of the models that had been published previously (Table 3). Models with a minimum test set size of 500 were exclusively selected for appropriate comparison. Although the model still showed relatively lower accuracy on KLGs 1 and 2, this was superior to most of the previously proposed models. It is also noteworthy that the visual explanations of the ensemble model were in accordance with the relevant features of the model, further reinforcing the validity of the model.

There are several reasons why our model was successful in classifying the severity of KOA in a knee image. The most important basis of our model development was that classifying KOA severity using KLG is a fine-grained classification tasks. The previous 14 models [3–10, 24–29] were all CNN-based and have been successful in discerning KLGs 3 or 4 since these grades have distinct features: joint space narrowing. However, these previous models showed relatively low accuracy in lower grades (KLGs 0, 1, and 2). Therefore, in the current study, the ensemble of the PIM method was applied for this fine-grained classification task, and it improved accuracy in the lower grades, although discerning between KLGs 1 and 2 still has some room for improvement. The overall accuracy of KLG prediction in a recent model by Thomas et al. [4] was 71.0%. However, the accuracy for KLGs 1 and 2 was 27% and 66.8%, respectively, whereas in our study, the accuracy was significantly higher in comparison, with an accuracy of 43% and 70%, respectively. Applying vectors in KLG labeling is another unique feature that has not been proposed in the previous literature, in which scalar values were assigned for each KL grade. The authors believe that applying a vector instead of simple numbers may also have contributed to the improved accuracy since KLG is not a simple classification task but rather a classification system that becomes higher with increased KOA severity.

Swin and EfficientNet were each used as the backbones because the former transformer-based model learns using the general aspect of an image, while the latter CNN-based model learns using the local aspect of an image. The authors hypothesized that the combination or an ensemble of these two different characteristic models could lead to better predictive outcomes. In addition, by applying the upper- and lower-grade-labeling methods, the model was able to calibrate minor differences in KL grading because, in a real clinical situation, even experts in this field sometimes overestimate or underestimate KLGs. The original label smoothing method normalizes the probability of the labels to 1. However, instead of normalizing the probabilities so that the sum of the values of the vector becomes 1, the authors maintained the original KLG as 1 in the label. By utilizing this method, because a regression-type loss function was used for our model training, when the model correctly predicts the KLG, the value of the loss function becomes 0. On the contrary, if a normalized label is used (e.g., [0, 0.8, 0.2, 0, 0]), the value of the loss function is not 0, even when the model correctly predicts the KLG. Therefore, since every DL model tries to minimize the loss function, the model tends to deviate from the correctly predicted label when a normalized label is utilized. For this reason, the authors maintained the label of the

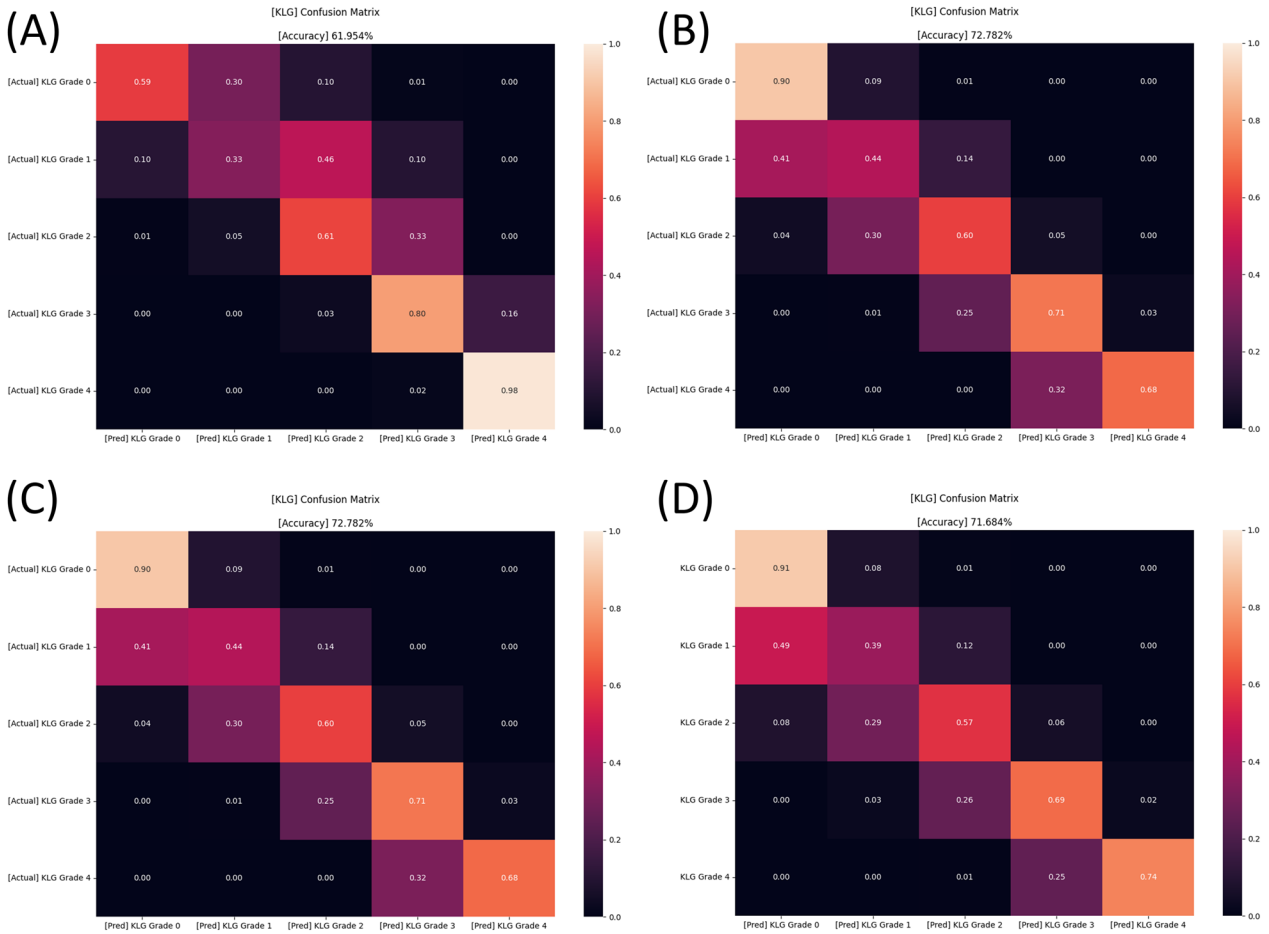

Fig. 3 Confusion matrices of four different models (before ensemble) in the study. A PIM that applied EfficientNet and upper-grade labeling (**A**), a PIM that applied EfficientNet and lower-grade labeling (**B**), a PIM that applied Swin and upper-grade labeling (**C**), and a PIM that applied Swin and lower-grade labeling (**D**). PIM, plug-in module

original KLG as 1. As a matter of fact, our experiments showed that the results of our labeling method were better than when the normalized labels were used. Possibly as a result of this "upper- and lower-grade labeling," our model was quite well balanced in predicting KLGs 2 and 3. For KLG 2, underestimation occurred in 15%, while overestimation occurred in 14% (Fig. 4). For KLG 3, underestimation occurred in 11%, while overestimation occurred in 8%. This was in contrast with a previous model by Thomas et al. [4] that showed similar overall accuracy in KLGs 2 and 3 to our model. In this previous model [4], underestimation occurred in 26%, while overestimation occurred in 7% in KLG 2, and underestimation occurred in 15%, while overestimation occurred in 4% in KLG 3.

Three previous models [3, 7, 10] showed superior accuracy: 87%, 98.9%, and 78.4%, respectively; however, our model was unique in that robust data from two different large cohorts, OAI and MOST (test set size of 17,040), were used. Although the average accuracy was lower than the model reported by Muhammad et al. [3], the training, validation, and test set size were significantly larger, and the test set was independent from the training set in our model. The evaluation of model accuracy on an independent dataset may have caused lower accuracy; however, because of this independent testing, our model is a more generalized model for future usage. Furthermore, using fewer ensemble base models (four in our model versus six in the previous model [3]) may lead to faster computing and subsequently better efficiency in clinical usage. A CNN-based model proposed by Abdullah et al. [7] reported the highest accuracy (98.9%) in classifying KLG, but the model was tested

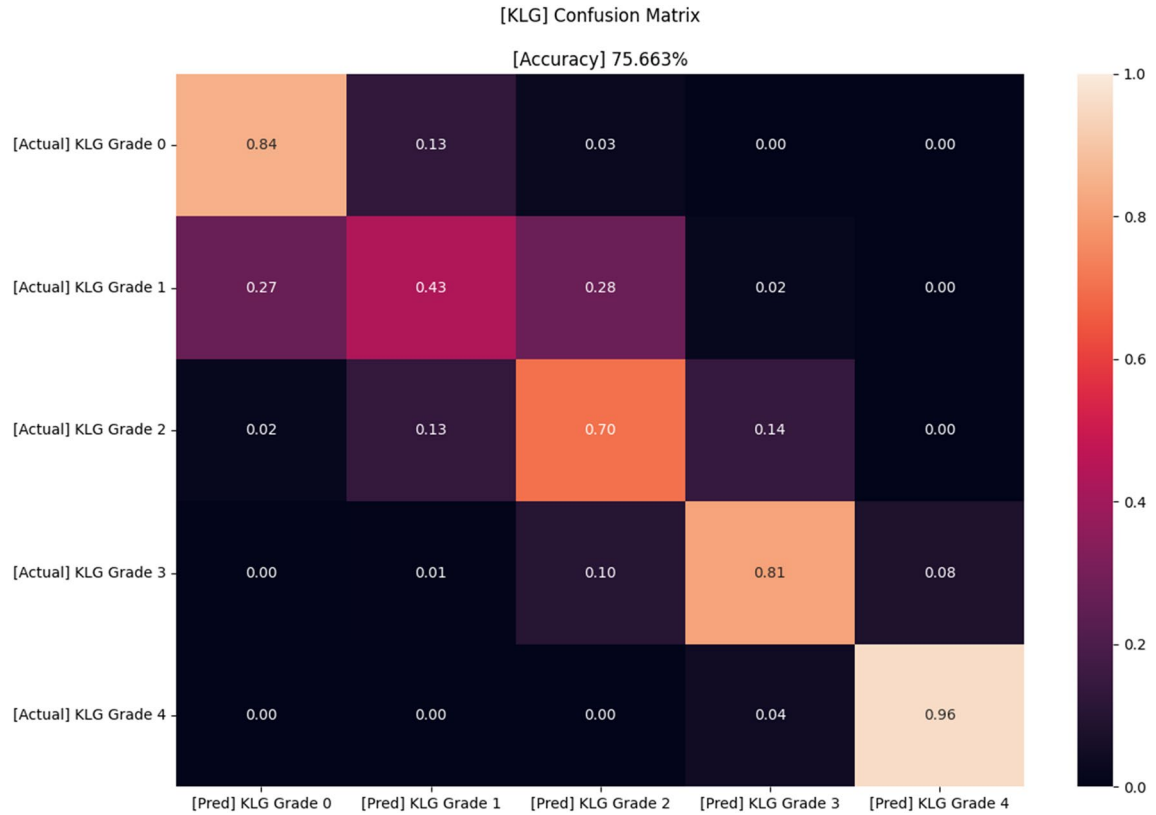

Fig. 4 Confusion matrix of the proposed model in this study

on a relatively small dataset (634 test images). Another CNN-based model proposed by Norman et al. [10] was tested on a relatively large dataset (5941 images) with a high average accuracy; however, the model did not discriminate between KLGs 0 and 1.

Comparing the accuracy of four individual models and the accuracy of the ensemble model, the ensemble method significantly increased the accuracy of the model by more than 3 percentage points when compared with each base model. Although this method improved the accuracy of KLG prediction, due to the usage of four different models in classifying an image, the computing time also doubled as a trade-off. Model training took about 4 full days with the state-of-the-art graphics processing unit (GPU; GeForce RTX 3090; NVIDIA, CA, USA) that was utilized in the study; however, classifying a single knee image takes a much shorter time in our model. Considering the fact that modern central processing unit (CPU) of a computer can perform around 100–200 GFLOPs (Giga FLOPs) per second, our model (565.34 GFLOPs) would take about 3–6 seconds to classify an image. Thus, our automated knee radiograph classification model can be useful in clinical practice by automatically providing the KLG of a knee image without any significant delay. Further investigations to lighten the model and reduce the computing time are needed in the future for better accessibility. High-temperature refinement and background suppression (HERBS) [30], which has been recently proposed for fine-grained classification tasks, outperformed PIM and thus could be an alternative to lighten and improve the accuracy.

There are several limitations in this study. First, our model provided a lower performance in classifying KLGs 1 and 2 compared with other grades, although it was higher than most of the previously reported literature. Second, although the KLGs of the image datasets (OAI and MOST) that were used in the study were labeled by thorough assessment and consensus by several radiologists, KLG itself can be variable even among experts due to its qualitative definition system. This may be the part of the reason why the automated classification models (including ours) show low accuracy on the

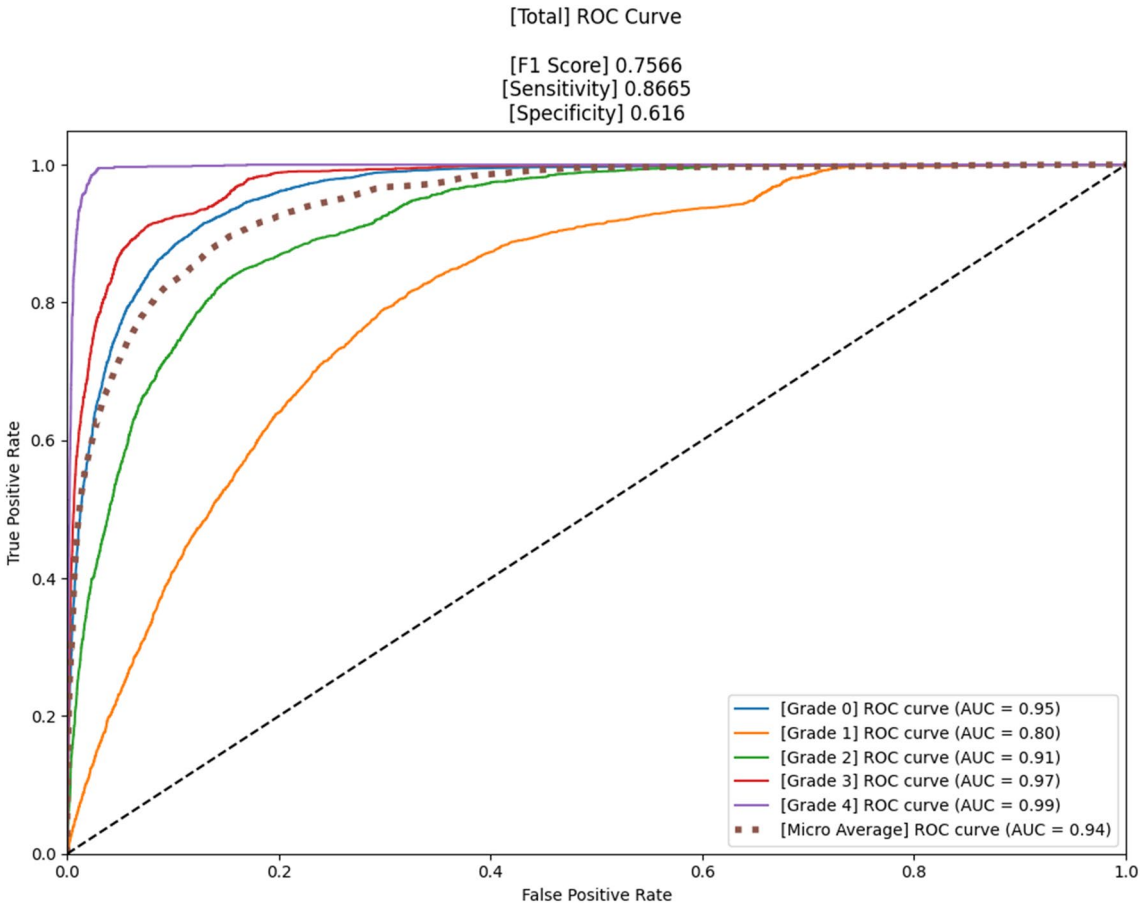

Fig. 5 The receiver operating characteristic (ROC) curve of the proposed model in this study

Table 2 Sensitivity and specificity of the proposed model for each Kellgren–Lawrence grade

KL grade	Sensitivity	Specificity
Grade 0	0.92	0.85
Grade 1	0.90	0.46
Grade 2	0.92	0.69
Grade 3	0.96	0.82
Grade 4	0.99	0.93

KL, Kellgren–Lawrence

lower grades. Third, overall accuracy of our model was 76.0%, which may raise concerns in clinical usage. However, the accuracy was generally high except for with grades 1 and 2; also, discerning between grade 1 and 2 actually does not critically alter decision-making in real practice. Therefore, the authors believe that the model can be readily used in the clinics, although there is much room for improvement. Last but most importantly, the model entirely relies on radiographs and does not synthesize other clinical records, such as pain or patient function. Future KOA severity grading should account for modalities other than simple radiographs to design an

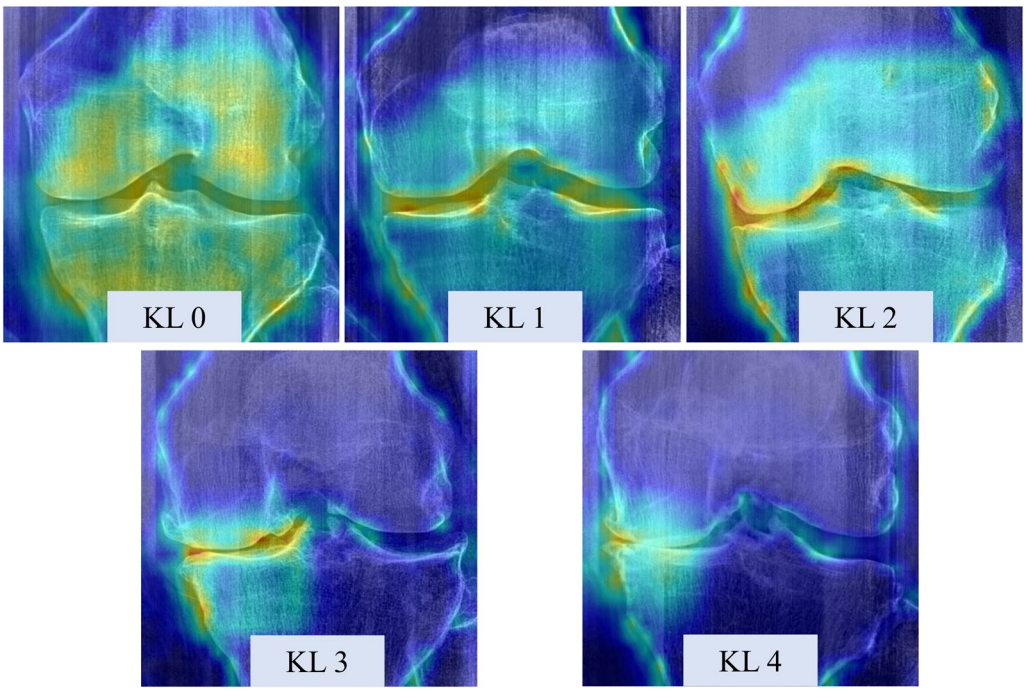

Fig. 6 Samples of the model visual explanation using Eigen-CAM for different knee osteoarthritis severity; KL, Kellgren–Lawrence

Table 3 Comparison with state-of-the-art DL-based methods (with a minimum test set size of 500) for a KOA severity assessment task

Year	Method	DL algorithm	Test set size	Accuracy for KLG 0 (%)	Accuracy for KLG 1 (%)	Accuracy for KLG 2 (%)	Accuracy for KLG 3 (%)	Accuracy for KLG 4 (%)	Average class accuracy (%)
2016	Reference [24]	CNN	2686	71	20	56	76	80	**59.60**
2017	Reference [25]	CNN	4400	86.9	6.0	60.2	73.0	78.1	**62.29**
2018	Reference [26]	Deep Siamese CNN	5960	78	45	52	70	88	**66.70**
2019	Reference [27]	CNN ensemble	1890	Not available					**69.50**
2019	Reference [28]	CNN	1495	Not available					**64.3**
2019	Reference [5]	Modified CNN	1385	89.8	55.6	82.6	36.0	100.0	**74.3**
2019	Reference [10]	CNN ensemble	5941	83.7	70.2		68.9	86.0	**78.4**
2020	Reference [29]	CNN	1175	79	52	58	59	85	**66.0**
2020	Reference [3]	CNN ensemble	7599	94	61	90	96	97	**87.0**
2020	Reference [6]	CNN ensemble	11,743	63.0	11.0	79.8	84.8	94.9	**68.0**
2020	Reference [4]	CNN	4090	86.5	27.0	66.8	80.9	85.8	**71.0**
2022	Reference [7]	ResNet-50 + AlexNet + TL	634	99.8	99.4	99.5	99.6	99.6	**98.9**
2024	Ours	PIM ensemble	17,040	85	46	69	82	93	**75.7**

The average class accuracy was highlighted in bold

DL, deep learning; KLG, Kellgren–Lawrence grade; CNN, convolutional neural network; TL, transfer learning; PIM, plug-in module

end-to-end model that could reflect the future prognosis of KOA. This newly proposed model would be more useful in the clinical settings than current radiograph-based models since the model directly reflects the prognosis of KOA, rather than inferring from radiographic severity.

Conclusions

The ensemble of PIMs classified KOA severity using simple radiographs with fine accuracy. Although improvements will be needed in the future, the model has been proven to have the potential to be clinically useful.

Abbreviations

KOA	Knee osteoarthritis
MRI	Magnetic resonance imaging
KLG	Kellgren—Lawrence grade
CNN	Convolutional neural network
PIM	Plug-in module
DL	Deep learning
OAI	Osteoarthritis Initiative
MOST	Multicenter Osteoarthritis Study
ROI	Range of interest
KL	Kellgren–Lawrence
FLOP	Floating point operations
ROC	Receiver operating characteristic

Acknowledgements

None.

Author contributions

D.W.L. contributed to acquisition of data, analysis and interpretation of data, and drafting of the article. D.S.S. and H.S.H. contributed to the acquisition and interpretation of data. D.H.R. contributed to the conception and design of the study and drafting of the article. All authors contributed to and approved the final manuscript.

Funding

None.

Availability of data and materials

The datasets used in the study are public-use datasets. The details about acquisition of these dataset are provided at "https://nda.nih.gov/oai" and "https://most.ucsf.edu/multicenter-osteoarthritis-study-most-public-data-sharing", respectively.

Declarations

Ethics approval and consent to participate

This article was prepared using the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST) public-use datasets. Ethical approval for collecting all subject information was provided by the OAI and MOST, and informed consent was obtained from all individual participants included in the studies. This article does not contain any studies with human participants performed by any of the authors. All participants provided informed consent before participating, in accordance with the Helsinki Declaration. Access, download, and analyses of the data were performed under the Data Use Agreement of OAI and MOST.

Consent for publication

Not applicable.

Competing interests

The corresponding author, D.H.R., is the CEO of CONNECTEVE Co., Ltd. However, this had no effect on the results of our research. No other author reports conflict of interest.

Author details

[1]Department of Orthopedic Surgery, Seoul National University College of Medicine, Seoul, South Korea. [2]CONNECTEVE Co., Ltd., Seoul, South Korea. [3]Department of Orthopedic Surgery, Seoul National University Hospital, 101 Daehak-Ro, Jongno-Gu, Seoul 110-744, South Korea. [4]Innovative Medical Technology Research Institute, Seoul National University Hospital, Seoul, South Korea. [5]Department of Orthopedic Surgery, Dongguk University Ilsan Hospital, Goyang, South Korea.

Received: 29 March 2024 Accepted: 30 July 2024
Published online: 13 August 2024

References

1. Kohn MD, Sassoon AA, Fernando ND (2016) Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res 474(8):1886–1893. https://doi.org/10.1007/s11999-016-4732-4
2. Wright RW, Group M (2014) Osteoarthritis Classification Scales: interobserver reliability and arthroscopic correlation. J Bone Joint Surg Am 96(14):1145–1151. https://doi.org/10.2106/JBJS.M.00929
3. Bany Muhammad M, Yeasin M (2021) Interpretable and parameter optimized ensemble model for knee osteoarthritis assessment using radiographs. Sci Rep 11(1):14348. https://doi.org/10.1038/s41598-021-93851-z
4. Thomas KA, Kidzinski L, Halilaj E, Fleming SL, Venkataraman GR, Oei EHG et al (2020) Automated classification of radiographic knee osteoarthritis severity using deep neural networks. Radiol Artif Intell 2(2):e190065. https://doi.org/10.1148/ryai.2020190065
5. Liu B, Luo J, Huang H (2020) Toward automatic quantification of knee osteoarthritis severity using improved faster R-CNN. Int J Comput Assist Radiol Surg 15(3):457–466. https://doi.org/10.1007/s11548-019-02096-9
6. Tiulpin A, Saarakkala S (2020) Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. Diagnostics (Basel). https://doi.org/10.3390/diagnostics10110932
7. Abdullah SS, Rajasekaran MP (2022) Automatic detection and classification of knee osteoarthritis using deep learning approach. Radiol Med 127(4):398–406. https://doi.org/10.1007/s11547-022-01476-7
8. Swiecicki A, Li N, O'Donnell J, Said N, Yang J, Mather RC et al (2021) Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. Comput Biol Med 133:104334. https://doi.org/10.1016/j.compbiomed.2021.104334
9. Kim DH, Lee KJ, Choi D, Lee JI, Choi HG, Lee YS (2020) Can additional patient information improve the diagnostic performance of deep learning for the interpretation of knee osteoarthritis severity. J Clin Med. https://doi.org/10.3390/jcm9103341
10. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S (2019) Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J Digit Imaging 32(3):471–477
11. Chou P-Y, Lin C-H, Kao W-C. A novel plug-in module for fine-grained visual classification. arXiv preprint. 2022. https://arxiv.org/abs/2202.03822.
12. Lester G (2006) The osteoarthritis initiative. Prot Cohort Study. 1:2
13. Segal NA, Nevitt MC, Gross KD, Hietpas J, Glass NA, Lewis CE et al (2013) The Multicenter Osteoarthritis study: opportunities for rehabilitation research. PM R 5(8):647–654. https://doi.org/10.1016/j.pmrj.2013.04.014
14. Xu X, Jiang Y, Chen W, Huang Y, Zhang Y, Sun X. Damo-yolo: A report on real-time object detection design. arXiv preprint. 2022. https://arxiv.org/abs/2211.15444
15. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021. pp. 10012–22.
16. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. International conference on machine learning: PMLR; 2019. pp. 6105–14.
17. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A (2021) A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol 65(5):545–563
18. Zuiderveld K (1994) Contrast limited adaptive histogram equalization. In: Zuiderveld K (ed) Graphics gems. Elsevier, Amsterdam, pp 474–485
19. Sane P, Agrawal R. Pixel normalization from numeric data as input to neural networks: For machine learning and image processing. In: 2017

International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET): IEEE; 2017. pp. 2221–5.
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 2818–26.
21. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng 2:40–46
22. Muhammad MB, Yeasin M. Eigen-cam: class activation map using principal components. 2020 international joint conference on neural networks (IJCNN): IEEE; 2020. pp. 1–7.
23. Katharopoulos A. Stop wasting my FLOPS: improving the efficiency of deep learning models. EPFL; 2022.
24. Antony J, McGuinness K, O'Connor NE, Moran K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: 2016 23rd international conference on pattern recognition (ICPR): IEEE; 2016. pp. 1195–200.
25. Antony J, McGuinness K, Moran K, O'Connor NE (2017) Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In: Perner P (ed) Machine learning and data mining in pattern recognition: 13th international conference, MLDM 2017. Springer, New York, pp 376–390
26. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S (2018) Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 8(1):1727
27. Muhammad MB, Moinuddin A, Lee MTM, Zhang Y, Abedi V, Zand R, et al. Deep ensemble network for quantification and severity assessment of knee osteoarthritis. In: 2019 18th IEEE International conference on machine learning and applications (ICMLA): IEEE; 2019. pp. 951–7.
28. Górriz M, Antony J, McGuinness K, Giró-i-Nieto X, O'Connor NE (2019) Assessing knee OA severity with CNN attention-based end-to-end architectures. International conference on medical imaging with deep learning. PMLR. 102:197–214
29. Kondal S, Kulkarni V, Gaikwad A, Kharat A, Pant A. Automatic grading of knee osteoarthritis on the Kellgren-Lawrence scale from radiographs using convolutional neural networks. arXiv preprint. 2020. https://arxiv.org/abs/2004.08572.
30. Chou PY, Kao YY, Lin CH. Fine-grained visual classification with high-temperature refinement and background suppression. arXiv preprint. 2023. https://arxiv.org/abs/2303.06442.

Publisher's Note