Automated Hand Osteoarthritis Classification Using Convolutional Neural Networks

Carmine Guida
Department of Computer Science
Pace University
New York, NY, USA
cguida@pace.edu

Ming Zhang
School of Computing and Data Science
Wentworth Institute of Technology
Boston, MA, USA
zhangm1@wit.edu

Jordan Blackadar
School of Computing and Data Science
Wentworth Institute of Technology
Boston, MA, USA
blackadarj@wit.edu

Zilong Yang
School of Computing and Data Science
Wentworth Institute of Technology
Boston, MA, USA
yangz1@wit.edu

Jeffrey B. Driban
Division of Rheumatology, Allergy, & Immunology
Tufts Medical Center
Boston, MA, USA
jeffrey.driban@tufts.edu

Jeffrey Duryea
Department of Radiology, Brigham and Women's Hospital and Harvard Medical School
Boston, MA, USA
jduryea@bwh.harvard.edu

Lena Schaefer
Department of Radiology, Brigham and Women's Hospital and Harvard Medical School
Boston, MA, USA
lenafranziskaschaefer@yahoo.com

Charles B. Eaton
Center for Primary Care & Prevention, Alpert Medical School of Brown University
Pawtucket, RI, USA
cbeaton51@gmail.com

Timothy McAlindon
Division of Rheumatology
Tufts Medical Center
Boston, MA, USA 02111
tmcalindon@tuftsmedicalcenter.org

Juan Shan
Department of Computer Science
Pace University
New York, NY, USA
jshan@pace.edu

***Abstract*—Osteoarthritis (OA) is the most common form of arthritis and often occurs in joints such as the knees, hips, and hands. Given there is no cure for OA, early detection and prevention are required to avoid further damage to the joint. Typically, joints are given a Kellgren and Lawrence (KL) grade of 0 to 4 with KL ≤ 1 meaning non-OA and KL ≥ 2 being positive for OA. Overall hand OA is determined by a positive OA rating of a joint on more than one finger. Therefore, to detect hand OA, one needs to detect worrisome hand joints first. This study uses a convolutional neural network (CNN) and proposes a custom architecture to automatically classify joints from hand X-rays into 5 KL categories as well as 2 categories of non-OA/OA. Post-processing is used to determine overall hand OA. Using a dataset of 3,556 hand X-rays, our custom CNN architecture was able to achieve a 5-category finger joint classification accuracy of 82.7% with a Matthews correlation coefficient (MCC) of 0.61. For 2-category classification, our model achieved an accuracy of 92.9% with an MCC of 0.74 and an area under the curve (AUC) score of 0.965. Based on the joint-level classification results of each hand, our model achieved an accuracy of 88.6% to classify the hand-level OA, i.e., to distinguish hand X-rays with and without OA. To our knowledge, this is the first work that uses CNN to classify hand joints into KL grades and detect overall hand OA based on individual hand joints.**

Keywords—Hand Osteoarthritis, X-ray, Machine Learning, Convolutional Neural Networks

I. Introduction

The most common form of arthritis is osteoarthritis (OA) [1]. OA occurs in the joints and most commonly the knees, hips, and hands [2]. Characteristics of OA include pain during activity, reduced function, stiffness, and joint instability [3]. OA is the leading cause of disability in older adults and given an aging population and longer lifespans will become more common [4,5]. Unfortunately, there is no drug treatment method that can cure OA [6] therefore, early detection and prevention are needed.

Knee OA can limit walking, stair climbing, and other daily activities and affect the overall quality of life [7]. Like knee OA, hip OA can also cause a lack of mobility as well as a lack of independence and increased use of health care services [8]. While many studies have been conducted on knee and hip OA, hand OA is the next most common with patients reporting pain, stiffness and disability, which is not well-studied [9]. The cartilage loss and resulting disintegration of the joint can progress to a point where they become harmful and interfere with hand functions [10]. Early detection of hand OA is needed as having baseline OA in a joint showed an increased chance of developing OA in another joint within the same row or ray [11].

Assessment for hand OA can be made by examining radiographic (X-ray) images which are inexpensive and widely available [10]. OA can be diagnosed by observing the degrading of cartilage through joint erosion (JE) and joint space narrowing (JSN) [13]. The Kellgren–Lawrence (KL) scoring system for OA has 5 grades (0 – 4) with KL=0 meaning no OA, KL=1 meaning doubtful, KL=2 being minimal, KL=3 being moderate, and KL=4 indicating severe OA [14]. For hand OA, a KL grade can be assigned to various joints including the metacarpophalangeal (MCP), proximal interphalangeal (PIP), and distal interphalangeal (DIP) joints in the hand [15].

Machine learning methods including support vector machines (SVM) [16] and artificial neural networks (ANN)

This research was funded by the National Science Foundation, with grant numbers NSF-173420 and NSF-173429.

have been used for the classification of OA [16,17,18]. More recent machine learning techniques such as convolutional neural networks (CNN) are able to use multidimensional data such as images [19]. CNNs are often referred to as "deep learning" due to the several layers of processing [20]. CNNs have been used for medical image classification including pneumonia in chest X-rays [21] and COVID-19 [22]. CNNs have also been used for detection tasks such as locating discs in the spine [23]. Medical images differ from natural images in that they are standardized and regulated for quality which makes them well suited for machine learning purposes [24].

There are recent related works involving automating KL and OA classification using X-rays and CNNs. These studies include joints such as the knee [25-27] as well as the hip [28]. While there are studies using machine learning techniques to classify Rheumatoid Arthritis (RA) in the hand [29,30], work involving automated hand joint classification of OA is limited. A related work [31] [SJ1]used classic CNN architectures to classify the entire hand as OA/non-OA. For this study, we created a custom CNN architecture to classify the 5 KL grade categories on individual hand joints and evaluated its performance with other models.

II. Materials and Methods

A. Dataset

The Osteoarthritis Initiative (OAI) is a multi-center longitudinal study of 4796 men and women ages 45–79 [32]. The study includes a publicly available dataset of X-ray images as well as magnetic resonance images (MRI). While the OAI was focused on knee OA, hand X-ray images were also collected for the baseline and 48-month visit at each center. A trained radiologist labeled the KL grades for 12 joints on the dominant hand for 3,519 participants [33]. Intra-reader agreement based on weighted kappas was good (weighted kappa > 0.84) [33,34].

For our study, we used data from the 48-month visit as it had more KL=4 grades present. We manually located the MCP, PIP, and DIP joints on the pinky, ring, middle, and index fingers. We additionally stored the orientation of each joint. Individual hands were assigned to training (70%), validation (15%), and testing (15%) sets. For each hand, the individual joints are cropped from the image. Table I shows the distribution of KL grades. The available dataset is unbalanced in which KL=0 is about 73% of the dataset with KL=4 (severe OA) making up under 1% of the dataset.

TABLE I. Distribution of KL grades into Training, Validation, and Testing sets.

Set	KL=0	KL=1	KL=2	KL=3	KL=4	Total
Training	21807	3231	3942	687	213	29880
Validation	4559	721	887	168	61	6396
Testing	4582	706	895	155	58	6396
Total	30948	4658	5724	1010	332	42672

B. Preprocessing

The majority of the X-ray images (72%) were of only the right hand. For bilateral images (which contain both hands) the image was split. The KL graded hands in the dataset were mostly right hands with under 1% being left hands and as a result, we flipped all left hands to more resemble right hands.

The original X-ray images contain noise and other information not needed for KL classification. We developed a separate model using the U-net [35] architecture to generate masks of the hand. We manually masked the hand for a small subset as training data with 36 images and applied the trained U-net to the whole dataset. Fig. 1 shows the original image, mask generated by the U-net model, and the result after applying the mask to the original image. This leaves the hand isolated with a clean background. The mask also helps to remove any markers in the X-ray.

As a result of the different manufacturers of X-ray equipment used, images in the dataset had varying pixel dimensions. These images are in the DICOM format and have a pixel spacing attribute which indicates the amount of physical space between each pixel. We rescaled each image to a pixel spacing of 0.15mm. This size was the most common in the dataset resulting in about 28% not needing to be scaled.

Using manually labeled locations and angles, each joint was cropped with a size of 180×180 and then rotated in the reverse direction to give all joints the same orientation. Fig. 2 shows the location and angles of an example pinky finger with the final orientations of each joint.

Fig. 1. Original image, automatically generated mask, and hand X-ray after mask is applied.

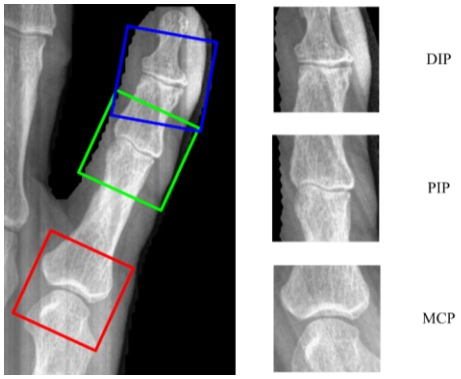

Fig. 2. Individual joints are cropped at an angle and rotated to have the same orientation.

Additionally, given the various sizes of fingers across the X-rays of different patients, scaling was needed to normalize the size of the joints in the samples. Using the masked image, the width of the finger at the PIP and DIP joints were calculated by searching for pixels with intenisty=0 to the left and right. For the MCP we used the PIP width and an additional 25%. These joint images were then scaled to match the target width of 180. Fig. 3 shows the original cropped image and the result after scaling.

C. CNN Model

The model used in this study originated as a basic model with a few layers. This model underwent several iterations and tuning. The final CNN architecture used in this study can be seen in Fig. 4. Each convolutional layer in the model is followed by an activation layer using the ReLU function. The first convolutional layer in the model uses 32 filters with a size of 7×7 and a stride of 2×2. The second convolutional layer uses 64 filters with a size of 5×5 and stride of 1×1. The third layer uses 128 filters with a size of 3×3 and stride of 1×1. Following this first block of layers is an average pooling layer. The next 3 convolution layers use 256, 512, and 1024 filters, respectively, all with a size of 3×3.

Global average pooling is used before the fully connected layers. Each fully connected layer has 1024 nodes and uses L2 regularization. Dropout layers are used with a rate of 50% in order to reduce overfitting. Finally, a SoftMax layer is used to output the 5 KL categories. The convolutional layers used He initialization [SJ2][36]. During training, the model uses the Adam optimizer with a learning rate of 0.0001, a batch size of 64, and early stopping based on validation loss.

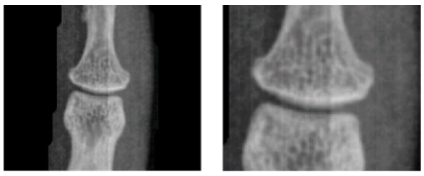

Fig. 3. A small finger joint is scaled in place to fit the target width size.

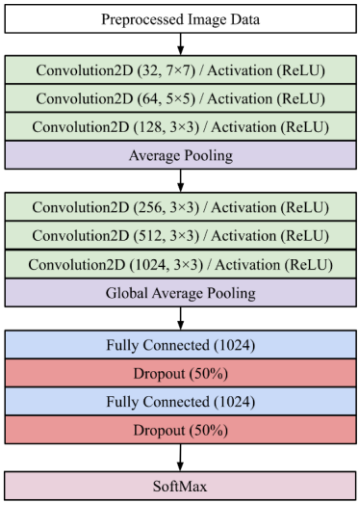

Fig. 4. The CNN architecture used in this study.

The model was built in Python using the Keras library with TensorFlow as the backend. The study was performed using a high-performance computer with a NVIDIA Tesla V100 32G GPU. Training time averaged 81 minutes.

III. Experiment and Results

Given the imbalance in our dataset, besides the commonly used overall accuracy, we employed the Mathews correlation coefficient (MCC) as evaluation metrics in the discussion of our results. For binary classification, the formula for MCC [37] can be seen in equation (1). Here TP is the true positive count, TN is the true negative, FP being the false positive, and FN being the false negative.

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}. \quad (1)$$

While originally proposed for binary classification, MCC was extended to multiple categories [38]. For multiple categories, MCC is defined as a confusion matrix C for K categories and can be described using the following intermediate variables [39]:

- $t_k = \sum_i^K C_{ik}$ The number of times k truly occurred.
- $p_k = \sum_i^K C_{ki}$ The number of times k was predicted.
- $c = \sum_k^K C_{kk}$ The total correctly predicted samples.
- $s = \sum_i^K \sum_j^K C_{ij}$ The total number of samples.

The MCC formula for multiple categories is:

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}} \quad (2)$$

Having trained models for both binary OA/non-OA classification and 5-category KL classification, both formulas of MCC will be used for evaluation. An MCC measure of 1.0 represents a perfect prediction with 0 being random and -1.0 meaning entirely incorrect. When there are more than two categories, the minimum value will be between -1 and 0 depending on the true distribution.

A. Results for KL Classification

A validation set was used during the training of our model. The results of the validation set can be seen in Table II. Reviewing this result, we can see that category KL=0 was the most accurate. This may be due to the majority of samples having this label. Categories KL=1, KL=2, KL=3 and KL=4 were often misclassified into the next lowest category. Finally, KL=4 with only 213 joints (0.7% of the training samples) was difficult for our model to classify. The overall performance of the validation set was an accuracy of 82.7% with an MCC of 0.62. Fig. 5 shows examples of different joint types with various KL grades from the validation set. As the KL grade gets higher (more severe OA) the space between the joints can be seen as getting more narrow.

B. Dataset Balancing

As seen in Table 1, the training set KL=0 category has 21807 samples but KL=4 category only has 213 samples. To better train a CNN model, we want to balance the sample across all categroies. We experimented with downsampling the KL=0 category during training combined with oversampling of the other KL grades. For instance, when KL=0 is limited to 5000 samples, image augmentation is used on the other KL grades to bring each up to 5000 samples and balance the dataset. Image augmentation methods included horizontal flipping, rotation between -20 and 20 degrees, and shifting the image horizontally and vertically. The validation set is not changed, and the results are in Table III. Balancing the dataset during training resulted in the model predicting more of KL=4 and increased accuracy in this category. However, it also caused decreased accuracy of the other categories and the overall accuracy dropped as well. The overall performance when limited to 10,000 is an accuracy of 82.1% with an MCC of 0.59. Performance when limited to 5,000 is an accuracy of 81.2% with an MCC of 0.57.

TABLE II. CONFUSION MATRIX FOR KL GRADES USING THE VALIDATION SET.

Act / Pred	P0	P1	P2	P3	P4	Total	Acc
A0	4263	181	115	0	0	4559	93.5%
A1	321	191	209	0	0	721	26.5%
A2	57	97	707	26	0	887	79.7%
A3	3	0	46	115	4	168	68.5%
A4	1	0	4	46	10	61	16.4%
Total	4645	469	1081	187	14	6396	82.7%

A0 denotes category is actually KL=0 while P0 means predicted as KL=0 by the model.

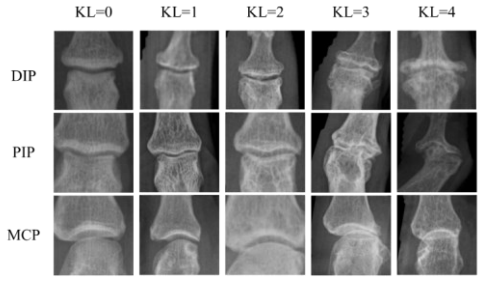

Fig. 5. Examples of MCP, PIP, and DIP joints with various KL grades from the validation set.

C. Individual Models for Different Joint Types

Our original model is trained by using all three joint types: MCP, PIP, and DIP (see Fig. 2). We additionally trained three separate models where each model processed a single joint type. The results of each individually trained model as well as the combined model can be seen in Table IV. In the dataset, there are few cases where the MCP joints have a grade of KL=4 with most being KL=0. PIP joints also have few KL=4 cases. The DIP joints have the most diverse types of KL grades. A reflection of the KL grade distribution can be seen in the MCC scores for each joint type. The overall performance of the MCP trained model was an accuracy of 93.3% with an MCC of 0.56. The PIP trained model had an accuracy of 76.7% with an MCC of 0.54. The DIP trained model had an accuracy of 78.6% with an MCC of 0.64. Using the best MCP, PIP and DIP trained models there was a slight increase in performance versus a single model trained on all three (82.9% vs. 82.7%). In summary, combining the three best individual models had a slightly higher accuracy of 82.9% than the original one-model-for-all while the MCC dropped slightly from 0.62 to 0.61.

TABLE III. ACCURACY FOR KL GRADES WHEN DOWNSAMPLING KL=0 SAMPLES DURING TRAINING (MCC IN PARENTHESES).

KL Grade	Original Dataset	Limited to: 10000 + AUG	Limited to: 5000 + AUG
KL=0	93.5%	94.4%	93.8%
KL=1	26.5%	20.3%	27.2%
KL=2	79.7%	73.1%	65.2%
KL=3	68.5%	73.2%	65.5%
KL=4	16.4%	45.9%	54.1%
Overall	82.7% (0.62)	82.1% (0.60)	81.2% (0.57)

TABLE IV. ACCURACY FOR KL GRADES WHEN INDIVIDUALLY TRAINING EACH JOINT TYPE (MCC IN PARENTHESES).

KL Grade	Original Method	MCP	PIP	DIP	Combined
KL=0	93.5%	98.2%	91.2%	93.3%	94.7%
KL=1	26.5%	25.0%	34.8%	30.9%	32.2%
KL=2	79.7%	67.1%	69.6%	73.0%	71.3%
KL=3	68.5%	42.1%	38.7%	78.0%	66.7%
KL=4	16.4%	0.0%	0.0%	15.0%	9.8%

KL Grade	Original Method	MCP	PIP	DIP	Combined
Overall	82.7% (0.62)	93.3% (0.56)	76.7% (0.54)	78.6% (0.64)	82.9% (0.61)

D. Transfer Learning

We compared our custom model built for this study with classic CNN architectures. Table V shows the performance of our model against VGG16 [40], ResNet50 [41], and DenseNet121 [42]. We used the pre-trained ImageNet weights for transfer learning with the final pooling and SoftMax layer set as trainable. Using transfer learning required the images to be scaled to 224×224 as well as copying the grayscale images into additional channels to form a RGB image. VGG16 classified most joints into KL=0 with some as KL=2. ResNet50 classified almost all samples into KL=0. DenseNet121 had the highest accuracy of these classic architectures as well as the best distribution of classification of KL grades, but the accuracy is still lower than that of the proposed custom model.

E. Non-OA / OA

Using the KL grade, a 2-category classification of non-OA (KL ≤ 1) and OA (KL ≥ 2) can be determined. We additionally trained a model with 2-category input by pre-processing the 5-category input into 2-categories of non-OA and OA. The results of the model trained on 2-categories can be seen in Table VI. For comparison, we also trained the previously mentioned classic architectures described in section III.D on 2-categories using transfer learning. The receiver operating characteristic (ROC) curves of our model and the classic CNN architectures are plotted in Fig. 6. Our model outperformed the other models with an AUC score of 0.965.

TABLE V. ACCURACY FOR KL GRADES USING CLASSIC CNN ARCHITECTURES WITH TRANSFER LEARNING (MCC IN PARENTHESES).

KL Grade	Custom Model	VGG	ResNet	DenseNet
KL=0	93.5%	97.8%	99.7%	95.6%
KL=1	26.5%	0.3%	0.0%	4.9%
KL=2	79.7%	33.6%	1.1%	58.3%
KL=3	68.5%	31.0%	0.0%	54.2%
KL=4	16.4%	14.8%	0.0%	32.8%
Overall	82.7% (0.62)	75.4% (0.35)	71.3% (0.04)	78.5% (0.48)

TABLE VI. CONFUSION MATRIX FOR NON-OA/OA USING 2-CATEGORY INPUT FOR THE VALIDATION SET.

Act\Pred	Non-OA	OA	Total	Acc
Non-OA	5126	154	5280	97.1%
OA	292	824	1116	73.8%
Total	5418	978	6396	93.0%

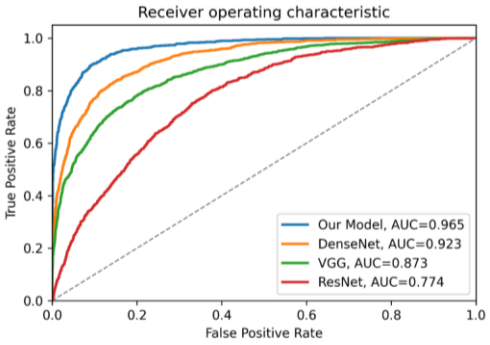

Fig. 6. ROC curves for non-OA/OA classification using the validation set.

F. Hand OA

In the previous sections, individual joints were evaluated to determine if they had OA. Having overall hand OA is determined by the presence of OA in more than one finger. Fig. 7 shows a hand without hand OA. The joints can all be seen to have good spacing and are graded with KL ≤ 1. Fig. 8 shows a hand with hand OA. Joints marked in red have a grade of KL=4 with orange indicating KL=3 and yellow meaning KL=2. Joints marked in green are KL=0. The example hand had no joints with KL=1. It can be seen that there is OA (KL ≥ 2) on more than one finger. Using the joint classification model and noting the location of the joints we were able to evaluate our model in classifying hand OA. The results in Table VII show the performance for OA classification at the hand level.

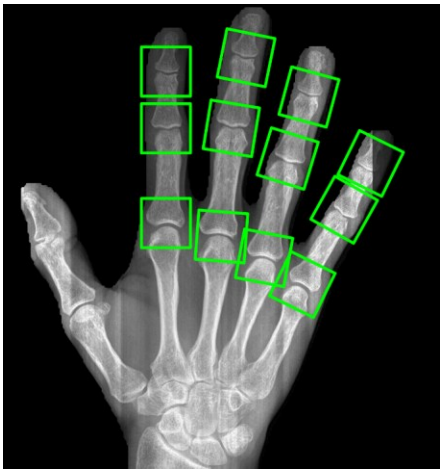

Fig. 7. Hand without OA. Joints can be seen to have good spacing.

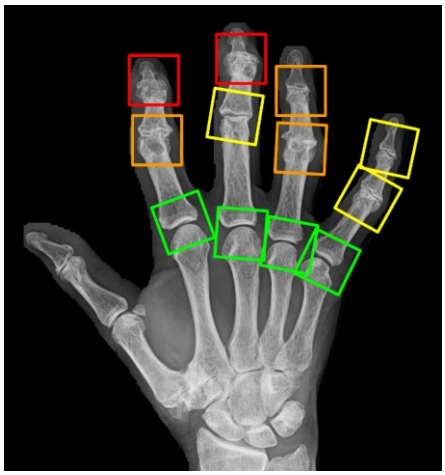

Fig. 8. Hand with OA due to the presence of OA on more than one finger. Joints marked in red are graded KL=4. Orange are KL=3. Yellow are KL=2. Green are KL=0.

TABLE VII. CONFUSION MATRIX FOR HAND NON-OA/OA USING POST PROCESSED 2-CATEGORY VALIDATION SET.

Act\Pred	Non-OA	OA	Total	Acc
Non-OA	286	24	310	92.3%
OA	41	182	223	81.6%
Total	327	206	533	87.8%

G. *Additional Metrics and Post-processed Results From 5-category Model*

We'd like to evaluate the joint OA/non-OA classification model and hand OA/non-OA classification model using more evaluation metrics in this section. In addition, for each classification task, besides training a new model as we did above, we can also post-process the results from the 5-category classification model using threshold KL=2, i.e., KL $\leq$ 1 is non-OA and KL $\geq$ 2 is OA. By post-processing the confusion matrix of the validation set from Table II into 2-categories, we can evaluate the model's performance for classifying non-OA and OA, at both joint level and hand level. Table VIII shows additional metrics including precision, recall, and F1 score. For individual joint classification of OA, the post-processed 5-category classification model had a better recall (sensitivity) over the 2-category model, while the overall accuracy is similar. Since sensitivity is a particularly important metric for medical-related decision-making systems, we'd like to bring the different performances of different models into the audience's attention. A high sensitivity or recall means the model is less likely to miss a positive case, while a high precision means the model is less likely to generate a false positive. Precision was higher when using the 2-category model on the individual joints. Similarly, for hand OA, although having the same overall accuracy, post-processed 5-category classification model had higher recall while 2-category model had higher precision.

TABLE VIII. ADDITIONAL METRICS FOR JOINT-LEVEL AND HAND-LEVEL OA CLASSIFICATION USING THE VALIDATION SET.

Method	Precision	Recall	F1	MCC	Acc
Joint OA classification using post-processed 5-category model	0.75	0.86	0.80	0.76	92.5%
Joint OA classification using 2-category model	0.84	0.74	0.79	0.75	93.0%
Hand OA classification using post-processed 5-category model	0.81	0.92	0.86	0.76	87.8%
Hand OA classification using post-processed 2-category model	0.88	0.82	0.85	0.75	87.8%

H. *Testing Set*

A testing set was set aside to the end of the study and was not seen by the models during the training process. The results are presented in Table IX. The performance of the proposed model is consistent with the validation set seen in Table II indicating the good generalizability of the model. The overall performance on the testing set was an accuracy of 82.7% with an MCC of 0.61 for 5 KL category classification at the joint level.

We continue to test the OA/non-OA model at joint level with the results presented in Table X. The ROC curve for the testing set can be seen in Fig. 9, with an AUC score of 0.966. The corresponding results on the validation set are in Table VI. Finally, we evaluated the hand level classification on the testing set through post-processing the results from Table IX (5-category model) and post-processing the results from Table X (2-category model). The hand level OA/non-OA classification performance is presented in Tables XI and XII. Again, a similar performance of the validation set was presented by the testing set.

TABLE IX. CONFUSION MATRIX FOR KL GRADES USING THE TESTING SET.

Act / Pred	P0	P1	P2	P3	P4	Total	Acc
A0	4313	188	80	0	1	4582	94.1%
A1	321	184	200	1	0	706	26.1%
A2	65	118	678	34	0	895	75.8%
A3	1	0	45	105	4	155	67.7%
A4	1	0	2	44	11	58	19.0%
Total	4701	490	1005	184	16	6396	82.7%

A0 denotes category is actually KL=0 while P0 means predicted as KL=0 by the model.

TABLE X. CONFUSION MATRIX FOR NON-OA/OA USING 2-CATEGORY INPUT FOR THE TESTING SET.

Act\Pred	Non-OA	OA	Total	Acc
Non-OA	5154	134	5288	97.5%
OA	322	786	1108	70.9%

Act\Pred	Non-OA	OA	Total	Acc
Total	5476	920	6396	92.9%

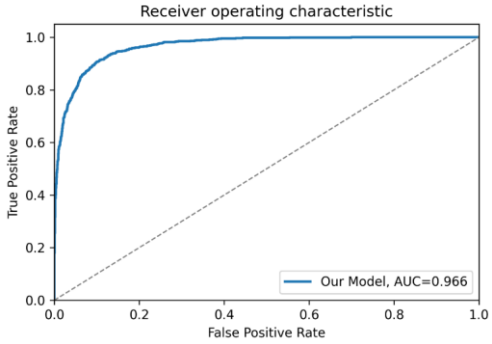

Fig. 9. ROC curve for non-OA/OA classification using the testing set.

TABLE XI. CONFUSION MATRIX FOR HAND NON-OA/OA USING POST PROCESSED 5-CATEGORY TESTING SET.

Act\Pred	Non-OA	OA	Total	Acc
Non-OA	275	46	321	85.7%
OA	15	197	212	92.9%
Total	290	243	533	88.6%

TABLE XII. CONFUSION MATRIX FOR HAND NON-OA/OA USING POST PROCESSED 2-CATEGORY TESTING SET.

Act\Pred	Non-OA	OA	Total	Acc
Non-OA	299	22	321	93.1%
OA	39	173	212	81.6%
Total	338	195	533	88.6%

IV. CONCLUSION

In this paper, we proposed a machine-learning-based method to classify the KL grade of individual hand joints as well as non-OA/OA for the whole hand. Our custom CNN model outperformed classical CNN architectures using transfer learning. Our best performance for 5-KL-category joint classification using the testing set was an accuracy of 82.7% with an MCC of 0.61. For 2-category OA/non-OA joint classification, the accuracy was 92.9%, with an MCC of 0.74 and an AUC score of 0.966. After post-processing the joint level classification results, hand level OA/non-OA classification using post-processed 5-category and 2-category both had an overall accuracy of 88.6%. Future work includes reducing misclassification of boundary categories KL=1 and KL=2 as well as further exploring better solutions for the imbalanced dataset.

REFERENCES

[1] Lawrence, R. C., Helmick, C. G., Arnett, F. C., Deyo, R. A., Felson, D. T., Giannini, E. H., ... & Wolfe, F. (1998). Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *41*(5), 778-799.

[2] Turkiewicz, A., Petersson, I. F., Björk, J., Hawker, G., Dahlberg, L. E., Lohmander, L. S., & Englund, M. (2014). Current and future impact of osteoarthritis on health care: a population-based study with projections to year 2032. *Osteoarthritis and cartilage*, *22*(11), 1826-1832.

[3] Hunter, D. J., McDougall, J. J., & Keefe, F. J. (2009). The symptoms of osteoarthritis and the genesis of pain. *Medical Clinics of North America*, *93*(1), 83-100.

[4] Hunter, D. J., March, L., & Chew, M. (2020). Osteoarthritis in 2020 and beyond: a Lancet Commission. *The Lancet*, *396*(10264), 1711-1712.

[5] Safiri, S., Kolahi, A. A., Smith, E., Hill, C., Bettampadi, D., Mansournia, M. A., ... & Cross, M. (2020). Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the global burden of disease study 2017. *Annals of the rheumatic diseases*, *79*(6), 819-828.

[6] Karsdal, M. A., Michaelis, M., Ladel, C., Siebuhr, A. S., Bihlet, A. R., Andersen, J. R., ... & Kraus, V. B. (2016). Disease-modifying treatments for osteoarthritis (DMOADs) of the knee and hip: lessons learned from failures and opportunities for the future. *Osteoarthritis and cartilage*, *24*(12), 2013-2021.

[7] Jack Farr, I. I., Miller, L. E., & Block, J. E. (2013). Quality of life in patients with knee osteoarthritis: a commentary on nonsurgical and surgical treatments. *The open orthopaedics journal*, *7*, 619.

[8] Lespasio, M. J., Sultan, A. A., Piuzzi, N. S., Khlopas, A., Husni, M. E., Muschler, G. F., & Mont, M. A. (2018). Hip osteoarthritis: a primer. *The Permanente Journal*, *22*.

[9] Gabay, O., & Gabay, C. (2013). Hand osteoarthritis: new insights. *Joint Bone Spine*, *80*(2), 130-134

[10] Kalichman, L., Cohen, Z., Kobyliansky, E., & Livshits, G. (2004). Patterns of joint distribution in hand osteoarthritis: contribution of age, sex, and handedness. *American Journal of Human Biology: The Official Journal of the Human Biology Association*, *16*(2), 125-134.

[11] Chaisson, C. E., Zhang, Y., McAlindon, T. E., Hannan, M. T., Aliabadi, P., Naimark, A., ... & Felson, D. T. (1997). Radiographic hand osteoarthritis: incidence, patterns, and influence of pre-existing disease in a population based sample. *The Journal of Rheumatology*, *24*(7), 1337-1343.

[12] Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., & Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, *8*(1), 1-10.

[13] Smolen, J. S., van der Heijde, D. M., Keystone, E. C., van Vollenhoven, R. F., Goldring, M. B., Guérette, B., ... & Landewé, R. B. (2013). Association of joint space narrowing with impairment of physical function and work ability in patients with early rheumatoid arthritis: protection beyond disease control by adalimumab plus methotrexate. *Annals of the rheumatic diseases*, *72*(7), 1156-1162.

[14] Kellgren, J. H., & Lawrence, J. (1957). Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, *16*(4), 494.

[15] Visser, A. W., Bøyesen, P., Haugen, I. K., Schoones, J. W., Van der Heijde, D. M., Rosendaal, F. R., & Kloppenburg, M. (2014). Radiographic scoring methods in hand osteoarthritis–a systematic literature search and descriptive review. *Osteoarthritis and cartilage*, *22*(10), 1710-1723.

[16] Du, Y., Shan, J., & Zhang, M. (2017, November). Knee osteoarthritis prediction on MR images using cartilage damage index and machine learning methods. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 671-677). IEEE.

[17] Yoo, T. K., Kim, D. W., Choi, S. B., Oh, E., & Park, J. S. (2016). Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. *PloS one*, *11*(2), e0148724.

[18] Duryea, J., Jiang, Y., Zakharevich, M., & Genant, H. K. (2000). Neural network based algorithm to quantify joint space width in joints of the hand for arthritis assessment. *Medical physics*, *27*(5), 1185-1194.

[19] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1-6). Ieee.

[20] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[21] Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, *6*(1), 1-18.

[22] Haque, K. F., Haque, F. F., Gandy, L., & Abdelgawad, A. (2020, August). Automatic detection of COVID-19 from chest X-ray images with convolutional neural networks. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 125-130). IEEE.

[23] Sa, R., Owens, W., Wiegand, R., Studin, M., Capoferri, D., Barooha, K., ... & Chaudhary, V. (2017, July). Intervertebral disc detection in X-ray images using faster R-CNN. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 564-567). IEEE.

[24] Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?. *arXiv preprint arXiv:1511.06348*.

[25] Antony, J., McGuinness, K., Moran, K., & O'Connor, N. E. (2017, July). Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International conference on machine learning and data mining in pattern recognition* (pp. 376-390). Springer, Cham.

[26] Chen, P., Gao, L., Shi, X., Allen, K., & Yang, L. (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, *75*, 84-92.

[27] Zhang, B., Tan, J., Cho, K., Chang, G., & Deniz, C. M. (2020, April). Attention-based cnn for kl grade classification: Data from the osteoarthritis initiative. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (pp. 731-735). IEEE.

[28] Xue, Y., Zhang, R., Deng, Y., Chen, K., & Jiang, T. (2017). A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PloS one*, *12*(6), e0178992.

[29] Sharon, H., Elamvazuthi, I., Lu, C. K., Parasuraman, S., & Natarajan, E. (2019, October). Classification of rheumatoid arthritis using machine learning algorithms. In *2019 IEEE Student Conference on Research and Development (SCOReD)* (pp. 245-250). IEEE.

[30] Üreten, K., Erbay, H., & Maraş, H. H. (2020). Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. *Clinical rheumatology*, *39*(4), 969-974.

[31] Üreten, K., Erbay, H., & Maraş, H. H. (2020). Detection of hand osteoarthritis from hand radiographs using convolutional neural networks with transfer learning. *Turkish Journal of Electrical Engineering & Computer Sciences*, *28*(5).

[32] OAI. (n.d.). Retrieved June 25, 2021, from https://nda.nih.gov/oai

[33] Davis, J. E., Schaefer, L. F., McAlindon, T. E., Eaton, C. B., Roberts, M. B., Haugen, I. K., ... & Driban, J. B. (2019). Characteristics of accelerated hand osteoarthritis: data from the osteoarthritis initiative. *The Journal of rheumatology*, *46*(4), 422-428.

[34] Schaefer, L. F., McAlindon, T. E., Eaton, C. B., Roberts, M. B., Haugen, I. K., Smith, S. E., ... & Driban, J. B. (2018). The associations between radiographic hand osteoarthritis definitions and hand pain: data from the osteoarthritis initiative. *Rheumatology international*, *38*(3), 403-413.

[35] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[36] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

[37] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412-424.

[38] Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational biology and chemistry*, *28*(5-6), 367-374.

[39] Scikit-learn. (n.d.). Retrieved July 23, 2021, from https://scikit-learn.org/stable/modules/model_evaluation.html#matthews-corrcoef

[40] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[41] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[42] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).