Hand Osteoarhritis Detection using Transfer Learning

Zhiheng Chang
School of Computing and DataScience
Wentworth Institute of Technology
Boston, United States
Changz@wit.edu

Ming Zhang
School of Computing and DataScience
Wentworth Institute of Technology
Boston, United States
Zhangm1@wit.edu

Osteoarthritis is one of the most common form of arthritis, affecting millions of people worldwide. It is caused by the protective cartilage that cushions the ends of the bones wears down over time. The diagnosis of Osteoarthritis is determined by the joint space width between the bones. The process of measuring the width is really time consuming and financial costly. While many modern solutions to Osteoarthritis are mainly focusing on Knee and Hip, this paper will introduce a computer-based solution to Hand Osteoarthritis using deep learning specifically transfer learning along with different dataset options, data augmentation strategies, and fine-tuning. Our model overall achieves 77.7% accuracy with AUC score of 0.786.

Keywords— Hand Osteoarthritis, Machine Learning, Deep Learning, Transfer Learning, Convolution Neural Network, Image Recognition

I. INTRODUCTION

Osteoarthritis (OA) is the deterioration of the protective cartilage between the bones that can cause severe joint pain and is a major cause of disability. However, this common skeletal disease is one of the most difficult to detect and prevent in its early stages. In the United States alone, OA has been the leading cause of total knee and hip replacements. The knee is the most common site for this disease. OA affects over 32.5 million US adults. Given the aging population and the obesity epidemic, the number of people with the disease will rise, which in turn will significantly increase the cost to the public health system.

Currently, the progression of OA is considered to be unstoppable. Doctors can only rely on visual assessment of 2D X-rays, coupled with knowledge of the patient's condition, to diagnose the condition. This is a very time-consuming and resource-intensive process, and prolongs the wait for an accurate diagnosis and deployment of a precise treatment plan. Since there is no cure for OA, treatment is focused on symptom relief and functional recovery. This means that the quicker and more accurate the initial diagnosis, the better the OA control outcome.

Machine learning (ML) is one of the most popular field in computer science and engineering by virtue of its power of pattern reorganization. Some tasks that are relatively difficult for humans can be handled by computers with ease. In recent years, applications of machine learning in the medical field have begun to be explored. The use of machine learning to assist in detection, diagnosis, and medical image interpretation has greatly helped physicians to reduce the time and effort required [1].

There are existing researches and applications that use ML to automate the process of OA diagnosis using deep learning, however, they are mainly focused on Knee OA (KOA) and Hip OA (HPOA). Kokkotis et al. proposed a model using feature selection methodology with genetic algorithms to predict various risk factors of KOA [2]. Abedin et al. use convolution neural network (CNN) to predict the KOA severity level [3]. Xue et al. use VGG-16 on Hip x-ray to predict diagnosis decision of HPOA [4]. Machine learning solutions for hand OA (HDOA), oppositely, are few and far between.

This paper will bring ML into HDOA diagnosis to help physicians to diagnose HDOA by building a deep learning (DL) model that takes hand joint X-ray images as input. However, the main challenge is like many existing medical research projects which are: unbalanced dataset with lack of medical image data. Although existing applications mentioned previously overcome this challenge by using data augmentation and other algorithms to generate more data, it is not really helpful in this situation where the data is highly unbalanced. Additionally, DL models recognize image feature by using CNN which needs substantial amount of annotated data to yield accurate predictions. Transfer learning (TL) will be one of the methods we used to address these issues. TL uses existing knowledge (source domain) to learn new knowledge (target domain). The core of it is to find the similarity between the source domain and the target domain. Since it is too expensive to learn the target domain directly from scratch, source domain that is similar to the target domain can be used to help learn the new knowledge more efficiently. Many research of medical image recognition use TL achieving higher result comparing to using single deep learning method of small or limited dataset [5].

Besides TL, other methods: balancing the dataset, increasing data variance, data augmentation, cropping data, and TL with fine tuning are being experiment on in order to obtain the best result as possible. Each of the method will be explained in more detail in Method, and the experiment result and analysis will be in the Result and Evaluation, followed by Conclusion and Future Work.

II. DATA

The data is obtained from Tufts Medical Center. There is total 3,591 hand x-ray images. Each hand is segmented into 12 joint images. The thumb is excluded from the segmentation. Three joints for index finger, middle finger, ring finger and pinky are: Distal interphalangeal joint (DIP), Proximal interphalangeal joints (PIP), and Metacarpophalangeal joint (MCP). After the segmentation, there are total 41,061 joints images, each has an imaging size of 180 pixels by 180 pixels. A Kellgren and Lawrence (KL) Score which is commonly used for classifying

TABLE I. DATA DISTRIBUTION GROUPED BY JOINT TYPE AND KL SCORE

KL	DIP2	DIP3	DIP4	DIP5	PIP2	PIP3	PIP4	PIP5	MCP2	MCP3	MCP4	MCP5
0	1911	2251	2516	1897	2736	2368	2182	2576	3145	3142	3522	3548
1	447	408	386	580	359	588	847	461	228	263	46	28
2	996	744	556	869	461	556	501	494	181	155	18	14
3	173	129	98	197	26	62	39	41	23	25	2	0
4	63	53	29	31	8	14	15	10	13	4	1	0

the severity of OA ranging from 0 to 4 is associated with each joint. A high score means severe, and low score means none or minimal.

Table 1 shows the data distribution among 3 joint types. Number (2-5) after the joint types is corresponding to the finger: index finger to pinky. The data are highly unbalanced: with over 90% of data are being labeled as 0 or 1. Severity of 2 (MCP4 and MCP5), 3, and 4 are significantly insufficient. The worst is that for MCP5, severity score of 3 and 4 are none.

However, the hand x-rays are not uniformed and came with various quality. Thus, certain number of the segmented joint images with poor quality which are not suitable for training a valid model since even human will not be able to learn from these data are manually removed from the dataset. After the removal, the dataset now contains total of 38,133 joint images: 11,997 DIP, 12,867 PIP, and 13,269 MCP.

III. METHODS

Most DL models which solve real world complex problem requires a vast amount of data to be trained on. To obtain those labeled data for supervised learning could be difficult, especially in medical problem. To increase the training data, each joint data group into 2 categories: NOA (0) of KL score from 0-1, and OA (1) of KL score from 2-4. There are 9,319 NOA and 2,678 OA in DIP. PIP: 11,189 NOA and 1,678 OA. Lastly, 12,966 NOA and 303 OA in MCP. Since DIP contains the greatest number of OA class, the experiment will be mainly focus on the detecting OA on DIP joint. Nevertheless, after categorizing the data into 2 categories, the data are still insufficient for training a good DL model. To conquer this DL weakness, TL is used to transfers the knowledge of a pre-trained model that is trained on a massive dataset to a new model of handling OA detection to create a better model. One of the most famous and largest datasets is the ImageNet dataset. It contains millions of images and over 1000 categories. Ideally, the target task should be as relative as possible to the domain source, so a pretrain model that is based on medical dataset is most suitable to our case; however, it is very difficult to obtain such dataset and model.

TABLE II. MODEL VAL_ACCURACY AND TRAINING TIME COMPARASION

MODEL	VAL_ACCURACY	TIME PER EPOCH
RESNET50	86.0%	804s
EFFICIENTNET	85.6%	1577s
MOBILENET	82.8%	190s
VGG16	81.0%	687s
DENSENET	75.0%	1217s
INCEPTIONV3	64.0%	202s

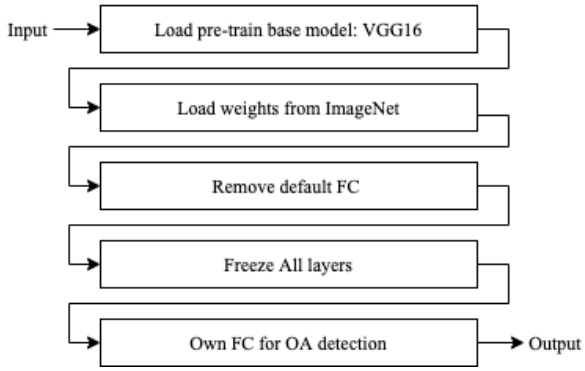

Fig. 1. OA detection using transfer learning.

In this experiment, A pretrain model of modern CNN architecture: VGG16 is selected as the base model based on the performance comparison. 6 popular available CNN architecture on ImageNet as pretrain models are ran on the DIP dataset with a simple full connection (FC): a flatten layer followed by 1 neuron as output. The comparison result is shown in table 2. Models being compared are: InceptionV3, DenseNet, VGG16, MobileNet, EfficientNet, and ResNet50. Among all 6 models, ResNet50 achieved the highest validation accuracy of 86.0% and InceptionV3 yield the lowest validation accuracy of 64.0%. For training time in second, MobileNet has the least training time per epoch averaging of 190 second on 8 core CPU intel i9, and EfficientNet take the longest training time per epoch average of 1,577 second. The reason of selecting VGG16 as the base model is because of its excellent balance of accuracy and computation complexity: the architecture is relatively light weight and carries out an average validation accuracy that leaves rooms for future improvements. To perform all the experiments efficiently, NIVIDA GPU RTX 2070 SUPER is set up for model training. With VGG16, each epoch only takes average of 10 second for batch size of 128.

The figure 1 shows the workflow of OA detection using transfer learning. By removing its original FC and freeze all the layers in the pre-trained base model to prevent altering the knowledge from the domain source and insert our own FC for predicting NOA or OA to complete the target task. Our FC contains one hidden layer with 512 neurons and Relu activation function, followed by 1 output neurons using Sigmoid activation function. The model is compiled using Binary cross-entropy loss function and Adam optimizer. Batch size remains the same as during the model selection 128. Each model in below methods is trained 300 epochs with early stopping.

A. Balancing the data

The DIP unbalanced dataset (UNBAL) contains total of 9,319 NOA and 2,678 OA. To create a balanced data set, N number of images are randomly selected from NOA class where N equals to the number of samples in OA. This leads to a balanced dataset with 2,678 NOA and 2,678 OA. With the same ration of 80% 10%, and 10% train test validation split, the balance dataset (BAL) now contains 2378 samples of each class for training, 150 samples of each class for both validation and testing set.

Fig. 2. Applying data augmentation on the balanced DIP dataset

B. Full Dataset

Another approach that may affect model performance is use all the available data of all the joints instead of only the DIP data. By combine all the joint images from different joint type, an unbalanced dataset contains all the joints (UNBAL_A) are created. UNBAL_A contains total of 33,474 NOA and 4659 OA. Apply previous method to also obtain a balanced dataset (BAL_A) which contains 4659 samples for each class.

C. Data augmentation

Data augmentation is widely used in DL when there is limited data. Data augmentation is simply of making minor alterations to the existing dataset such as flips, translations, and rotations. There are many different data augmentation strategies available, however, not all methods are useful in some case. For example, applying a width shift method, only part of the joint will be captured within the original image size. For our model, only vertical flip, zoom with range 0.2, and random rotation with range from -10 to 10 degrees are used on the BAL dataset. Examples of data augmentation are shown in Figure 2. The Left most image is the original joint image, second and third images are rotated image, the fourth image is zoomed image, and the fifth image is the vertical flipped image. The dynamic data augmentation is done through the Keras ImageDataGenerator API while training.

D. Cropping data

HDOA is usually determined by the KL score, which in turn is determined by the joint space width (JSW) which is the distance between upper bone and lower bone of a joint. A small JSW will result in a high KL score, and another way around. Therefore, for HDOA, we are only examining the center part of the joint image where the JSW is located, and the rest part of the image can be simply discarded since it carries less or no

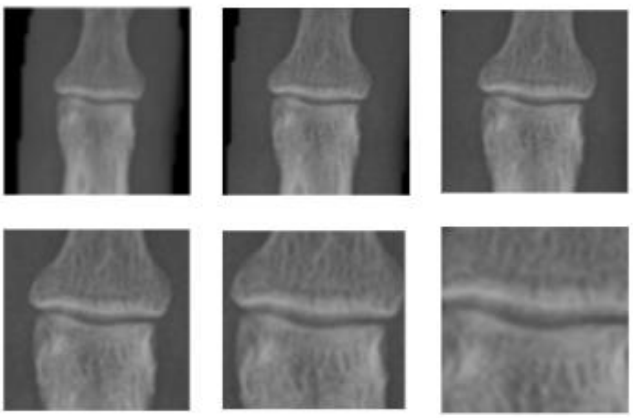

Fig. 3. Different center crop size on the balanced DIP dataset (The top row shows crop size of {10, 20, 30}, and the bottom row shows crop size of {40, 50, 60}.)

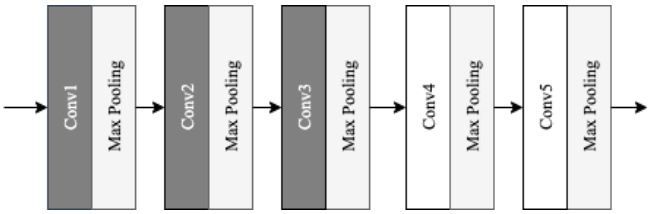

Fig. 4. VGG16 pretrain model with fine-tuning

information. In the experiment, we center crop each sample in BAL dataset by 10, 20, 30, 40, 50 and 60 pixels from each side, resulting squared image sizes of 160, 140, 120, 100, 80, and 60 pixels displayed in figure 3.

E. Enable fine-tuning in TL

A Typical CNN has multiple convolution blocks, each of the convolution block extract different level of features. The head convolution blocks extract general features, while tail blocks extract complex and precise feature. Previously, all the layers in the pre-trained VGG16 model with weights load from the ImageNet are frozen to transfer the knowledge from the domain source. However, with fine-tuning, the model is first initialized with the weights from the ImageNet and adapt with target domain during the model training. We enable fine-tuning by unfreeze the last 2 convolution blocks of VGG16 base model and trained with the BAL dataset to adapt to extract complex and precise features that are relative to HDOA. As shown in figure 4, the first three convolution blocks remain frozen during the training, while the last two blocks are untethered.

IV. Result and Evaluation

The model first train on the UBAL dataset yield an 88.48% of validation accuracy within 50 epochs. We set this as the base performance of the model, and other results returned by different method mentioned in the previous section are compared with this base line as shown in table 3. For center crop method, since all the crop size return the similar result, the validation accuracy and other evaluation metrices are being averaged. Model trained with UNBAL_A achieved the highest validation accuracy of 92.58%. The lowest validation is 81.5% by model trained with center crop method. It also takes the greatest number of epochs to train. Surprisingly, Model with fine-tuning takes the least number of epochs to train.

TABLE III. MODEL VAL_ACCURACY AND # OF EPOCH COMPARASION

MODEL	VAL_ACC	# OF EPOCHS
UNBAL	88.48%	50
BAL	83.70%	39
UNBAL_A	92.58%	87
BAL_A	85.67%	51
DATA AUGMENTATION	84.00%	51
CENTER CROP	81.57%	123
FINE-TUNING	84.67%	37

TABLE IV. MODEL EVALUATION METRICES SUMMARY

MODEL	ACC	TPR	PPV	F1	AUC
UNBAL		0.55	0.31	0.40	0.706
BAL	66.6%	0.70	0.55	0.61	0.760
UNBAL_A		0.44	0.03	0.06	0.540
BAL_A	68.0%	0.70	0.62	0.66	0.718
DATA AUGMENATON	68.7%	0.72	0.66	0.66	0.752
CENTER CROP	65.5%	0.64	0.71	0.68	0.679
FINE-TUNING	77.7%	0.78	0.78	0.78	0.786

Each model has also been evaluated on the test datasets. The results are shown in table 4. The evaluation metrices are accuracy (ACC), recall (TPR), precision (PPV), F1 score (F1), and area of AUC-ROC curve (AUC). For models that are evaluated on unbalanced test data, the accuracy (ACC) is left as blank since it is not a convincible measurement of models' overall performance. Receiver Operator Characteristic (ROC) curve better demonstration of model performance in binary classification problems. It is a probability curve that plots TPR against false positive rate (FPR) at various threshold. Therefore, the Area Under the Curve (AUC) which measures the ability of a classifier to distinguish between NOA and OA classes is chose as the determinant.

The figure 5 displays the plot of ROC curves of all models. The initial model that is trained on UNBAL dataset (vgg16) achieve a base line AUC of 0.706. By making the dataset balanced between NOA and OA (vgg16_balanced), AUC score improved significantly from 0.706 to 0.760. Including all joints in the training set, however, UNBAL_A (vgg16_unbalancedAll) gives the lowest AUC of 0.540 which indicates that the model is not able to distinguish between the two classes: it is simply predicting most of the test cases as NOA since it is the majority class. The AUC of model using BAL_A (vgg16_balancedAll), on the other hand, is slightly higher than the initial model.

Applying data augmentation method on the BAL dataset, the model (vgg16_balanced_data_augmentation) achieves AUC of 0.752 which improved upon the UNBAL but does not exceed

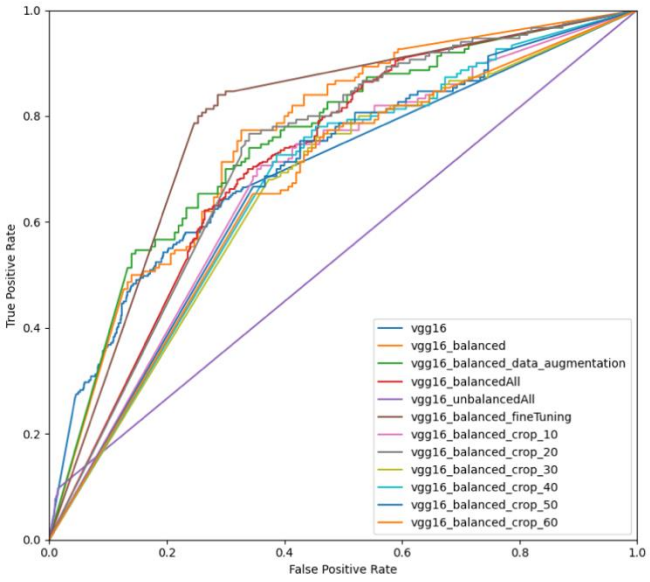

Fig. 5. ROC Curve of all models

the BAL model performance. One of the reasons could be the data augmentation method using strategies: flip, zoom, and rotation serve as regularization method which prevent model from overfitting the data. Center crops (vgg16_balanced_crop) are similar to the data augmentation where the average AUC is even lower than the initial model: only 0.679. Getting such low result possibly due to important information are lost during the image cropping.

Model with fine-tuning (vgg16_balanced_fineTuning) yield the highest result with AUC of 0.786. F1 score of 0.78 indicates the balance between TPR and PPV compared to second highest result of 0.61 returned by BAL model. Since the test sets are the same and balanced compares to the BAL model, comparison of the ACC can also be considered. The model with fine-tuning is around 10% higher than BAL model. Thus, the Model with fine-tuning is able to classify NOA and OA much correctly.

V. CONCLUSION

The DL model build using transfer learning with fine-tuning, base model using VGG16, and trained with a balance dataset offers the best overall result: 77.7% ACC and 0.786 AUC. Other method that are expected to give better results such as using full dataset, applying data augmentation strategies, and center crop are being proven to have minimal or no improvement at all. However, due to insufficient time and computational constraints, other trails of these methods and some other methods could not be put into practice. This will be discussed more as future works in the next section.

VI. FUTURE WORK

In this paper, we explored using deep TL to predict HDOA and came up with a model that achieves 77.7% ACC using VGG16 as base model. Other modern architecture could be experimented on such as ResNet and EfficientNet to potentially obtain a finer result. Additionally, different combination of different data augmentation strategies may also impact the final result. Different crop sizes on different joint and the number of layers that can be tuned in fine-tuning can also get varied overall performance of the model. Furthermore, the model with the best result can be acquired by combining this method in different ways.

VII. RESOURCE

The open-source code of this project and the results of the experiments can be found at GitHub. The public access link is: https://github.com/ZhihengChang/HandOAClassification.

VIII. RESOURCE

[1] B. F. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging." *Radiographics*, vol. 37, pp. 505-515, 2017

[2] C. Kokkotis, S. Moustakidis, D. Tsaopoulos, V. Baltzopoulos, and G. Giakas, "Identifying robust risk factors for knee osteoarthritis progression: An evolutionary machine learning approach." *Healthcare switzerland*, vol. 9, issue. 3, 2021

[3] J. Abedin, J. Antony, K. McGuinness, K. Moran, N. O'Connor, D. Rebholz-Schuhmann, and J. Newell "Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images." *Scientific Reports*, vol. 9, issue. 1, 2019

[4] R. Gebre, J. Hirvasniemi, R. Heijden, I. Lantto, S. Saarakkala, J. Leppilahti, and T. Jämsä, “Detecting hip osteoarthritis on clinical CT: a deep learning application based on 2-D summation images derived from CT.” *Osteoporosis International*, 2021

[5] S. Kundu, BG. Ashinsky, M. Bouhrara, EB. Dam, S. Demehri, M. Shifat-E-Rabbi, RG. Spencer, et al. “Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning.” *Proc Natl Acad Sci USA*, 2020 Oct 6

[6] H. Tang, X. Cen. “A Survey of Transfer Learning Applied in Medical Image Recognition.” *2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Advances in Electrical Engineering and Computer Applications (AEECA), 2021 IEEE International Conference*, August 2021