2020 24th International Computer Science and Engineering Conference (ICSEC) | 978-1-6654-0339-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICSEC51790.2020.9375275

Abnormality Detection in Musculoskeletal Radiographs using EfficientNets

Kasemsit Teeyapan

Department of Computer Engineering, Faculty of Engineering, Chiang Mai University
Biomedical Engineering Institute, Chiang Mai University
Chiang Mai, Thailand
kasemsit.t@cmu.ac.th

***Abstract*—Abnormality detection in musculoskeletal radiographs, a regular task for radiologists, requires both experiences and efforts. To increase the number of radiographs interpreted each day, this paper presents cost-efficient deep learning models based on ensembles of EfficientNet architectures to help automate the detection process. We investigate the transfer learning performance of ImageNet pre-trained checkpoints on the musculoskeletal radiograph (MURA) dataset which is very different from the ImageNet dataset. The experimental results show that, the ImageNet pre-trained checkpoints have to be retrained on the entire MURA training set, before being trained on a specific study type. The performance of the EfficientNet-based models is shown to be superior to three baseline models. In particular, EfficientNet-B3 not only achieved the overall Cohen's Kappa score of 0.717, compared to the scores 0.680, 0.688, and 0.712 for MobileNetV2, DenseNet-169, and Xception, respectively, but also being better in term of efficiency.**

***Index Terms*—EfficientNet, deep learning, transfer learning, musculoskeletal radiographs, MURA**

I. Introduction

Musculoskeletal conditions, a general term used to describe a group of conditions that affects joints, bones, muscles, and the spine, are recognized as major health burdens on individuals. According to World Health Organization (WHO), the conditions are so prevalent affecting at least one billion people worldwide with more than 30 million yearly emergency department visits [1].

Conditions, such as fractures, are diagnosed through radiographic studies. Typically, one study, which includes multiple radiographic images, is examined as normal or abnormal. Such a classification task is critical and requires experienced healthcare professionals. However, radiograph interpretation is laborious, and there may be no experienced radiologist available during off-hour times. In addition, due to radiologist's fatigue, the radiographic studies can be prone to human errors. As a result, automatic tools for assisting healthcare professionals are utmost crucial. At the very least, identifying radiographs as normal can help patients avoid further unnecessary diagnostic procedures. Abnormal cases can be later submitted to relevant healthcare professionals for further investigation.

In recent years, the advancement in the field of artificial intelligence, particularly deep learning, has gained major attention as a potential tool to create an automatic system for detecting abnormality in musculoskeletal radiographs [2]. Huge efforts have been devoted to creating large-scale public radiograph datasets, such as the MURA dataset [1] which are the datasets of radiographic studies. The dataset, as a result, further encourages the progress of AI applications in radiology.

In this paper, the main problem is on detecting an abnormality in musculoskeletal radiographic studies of upper extremity, as either normal or abnormal. We train and validate the state-of-the-art model architecture called "EfficientNet" [3] on the MURA dataset. In particular, EfficientNet-B0 to EfficientNet-B3 will be considered. As EfficientNet-B3 can achieve the top-1 accuracy of 81.1% on the ImageNet dataset [4], we investigate whether the ImageNet pre-trained checkpoints can be efficiently transferred to learning the radiographs in the MURA dataset which are completely different types of images.

Seven types of studies in the dataset, i.e. elbow, finger, forearm, hand, humerus, shoulder, and wrist, regardless of views, are separately trained. For one study of a patient, the predicted probabilities of all views are averaged. We ensemble five EfficientNets to determine the abnormality of the study. The performance of the model is reported in terms of classification accuracy and Cohen's kappa score [5]. The experimental results show that the ensembles of EfficientNets achieved the performance on the validation set higher than three baseline ensemble models, which include the ensemble of DenseNet-169 [6], the ensemble of Xception [7], and the ensemble of MobileNetV2 [8].

This paper is organized as follows. Section II provides an overview of deep learning for abnormality detection in musculoskeletal radiographs. An overview of the MURA dataset is presented in Section III. The proposed methodology using EfficientNets with transfer learning is described in Section IV, followed by the experimental setups in Section V. Section VI presents the experimental results and discussion, and finally, the concluding remarks are provided in Section VII.

II. Related Works

The research on using deep learning to detect an abnormality in an X-ray image of upper extremity is largely enabled due to the advent of the MURA dataset by Rajpurkar et al. [1] where they also proposed a baseline model based on DenseNet-169, pre-trained on the ImageNet dataset. A committee of three radiologists was used to create the gold stan-

978-0-7381-1218-3/20/$31.00 ©2020 IEEE

dard. Three additional radiologists as well as their DenseNet-169 model ensemble were compared using Cohen's kappa score, which shows that the performance of their baseline model is still below the worst radiologist.

Since then, many research studies have conducted further experiments on the MURA dataset using various network architectures. For example, D. Dias [9] tested two transfer learning strategies for Inception-v3, VGG-19, SqueezeNet, DenseNet-121 [6], and ResNet-152. The reported results on the validation set show that SqueezeNet achieved the highest performance, with much lower number of parameters than its competitors. However, the results are lower than the DenseNet-169 baseline.

Other works which investigated transfer learning for the MURA dataset include N. Harini et al. [10] who experimented with Inception-v3, Xception [7], VGG-19, DenseNet169, and MobileNet on only finger, wrist, and shoulder images. Similarly, G. Chada [11] considered only finger and humerus data for training DenseNet-169, DenseNet-201, and InceptionResNetV2.

In [12], an ensemble model consisting of VGG-19 and ResNet were proposed and tested on four types of studies, elbow, wrist, finger, and humerus. For the case that the type of radiographs is unknown, H. El-Saadawy et al. [13] designed a hierarchical classifier consisting of two stages of MobileNets where the first stage has one network for categorizing a bone type and the second stage consists of local classifiers to determine whether the input is normal or abnormal.

S. Panda and M. Jangid [14] worked on evaluating six different optimizers on their own CNN architecture with the MURA dataset and concluded that Adam [15] was the best optimizer according to their experiments. Although the MURA dataset is designed to be a classification problem, B. Guan et al. [16] further annotated arm fractures in the dataset by adding bounding boxes so that they could localize fractures using their proposed object detection algorithm.

The works on the MURA dataset in the literature are usually evaluated on the validation set, as the test set is not publicly available. The DenseNet-169 ensemble baseline [1] was reported on the test set with the overall Kappa score of 0.705 for all types of studies. However, the performance from other subsequent works was reported under different metrics, or the validation set performance was compared to the test set performance from the DenseNet-169 baseline. Two other studies which reported high performance include [17] who applied a capsule network, consisting of two convolutional capsule layers, to achieve the average Kappa score of 0.801. The other study which proposed a multi-scale convolutional neural network with fully connected graph convolutional network [18] also achieved the average Kappa score of 0.836.

III. MURA DATASET

The MURA dataset [1] is a large musculoskeletal radiographic image dataset of upper extremity, prepared by the Stanford ML group using the de-identified data from Stanford Hospital. It consists of totally 14,863 studies from 12,173 patients. Each study was manually labeled as normal or abnormal by board-certified radiologists at the time of the clinical radiographic interpretation between 2001 and 2012.

TABLE I
MURA DATASET

Study type	Label	Training set		Validation set	
		Studies	Images	Studies	Images
Elbow	Normal	1094	2925	92	235
	Abnormal	660	2006	66	230
Finger	Normal	1280	3138	92	214
	Abnormal	655	1968	83	247
Forearm	Normal	590	1164	69	150
	Abnormal	287	661	64	151
Hand	Normal	1497	4059	101	271
	Abnormal	521	1484	66	189
Humerus	Normal	321	673	68	148
	Abnormal	271	599	67	140
Shoulder	Normal	1364	4211	99	285
	Abnormal	1457	4168	95	278
Wrist	Normal	2134	5765	140	364
	Abnormal	1326	3987	97	295
All types	Normal	8280	21935	661	1667
	Abnormal	5177	14873	538	1530
	Total	13457	36808	1199	3197

There are seven types of studies in the dataset, i.e. elbow, finger, forearm, hand, humerus, shoulder, and wrist. The images in each study vary in width and height. The total number of images in the dataset is 40,561 images. The dataset were split into training set (13,457 studies), validation set (1,199 studies), and test set (207 studies) without overlapping in patients among the sets. The details of the dataset grouped by labels for the training and validation sets are shown in Table I.

IV. EFFICIENTNETS AND TRANSFER LEARNING METHODOLOGY

In this section, we overview the original EfficientNet architecture and the transfer learning architecture for the MURA dataset

A. *EfficientNets*

The baseline network presented in the MURA paper [1] was based on the DenseNet-169 architecture [6] which was a very powerful model at the time. However, more recently, a more cost-efficient family of models called EfficientNets was proposed by M. Tan and Q. Le [3]. These models have demonstrated high performance on the ImageNet dataset with fewer parameters and FLOPS than other state-of-the-art methods, such as GPipe [19].

The core concept behind the EfficientNets is the compound scaling method which uniformly scales the baseline network by three factors: network width, network depth, and image resolution. This idea makes EfficientNets differ from other common scaling methods which scale up depth, width, and resolution, separately. The baseline network, named EfficientNet-B0, primarily consists of mobile inverted bottleneck MBConv block [8] and the squeeze-and-excitation optimization [20]. This network was initially created by a multi-objective neural architecture search that optimized accuracy and FLOPS. After that, it was scaled up to produce larger architectures, where the authors of the paper presented up to seven scaled architectures named EfficientNet-B1 to EfficientNet-B7.

TABLE II
ARCHITECTURE USED IN TRANSFER LEARNING

	Layer
1	Base model
2	Global average pooling
3	Batch normalization layer
4	Dropout with dropout rate 0.5
5	Dense layer with one neuron using sigmoid activation function and Glorot normal initializer

TABLE III
NUMBER OF PARAMETERS AND INPUT RESOLUTIONS OF THE TRANSFER LEARNING MODELS UNDER VARIOUS BASE ARCHITECTURES SORTED BY THE IMAGENET TOP-1 ACCURACY REPORTED IN THE LITERATURE.

Base architecture	ImageNet top-1 accuracy	No. of parameters	Input resolution
MobileNetV2	74.7%	2,264,385	300 × 300
DenseNet-169	76.2%	12,651,201	300 × 300
EfficientNet-B0	76.3%	4,055,972	224 × 224
EfficientNet-B1	78.8%	6,581,640	240 × 240
Xception	79.0%	20,871,721	300 × 300
EfficientNet-B2	79.8%	7,775,610	260 × 260
EfficientNet-B3	81.1%	10,791,216	300 × 300

B. Transfer learning

Transfer learning [21] is the problem of applying a model trained on one task to a different but related task. In our case, instead of training a model on the MURA dataset from scratch, we will use the model trained on the ImageNet dataset as a starting point.

According to [3], EfficientNets do not only performed well when trained on the ImageNet dataset. Their ImageNet pre-trained models can also efficiently learn widely used transfer learning datasets, such as CIFAR-10 and Food-101, achieving new state-of-the-art accuracies for 5 out of 8 datasets [3]. Yet, the imaging modality of these datasets completely differs from radiological imaging. Hence, this leads to evaluating the transfer learning performance on the upper extremity radiographs in the MURA dataset in Section VI.

To make the EfficientNet architecture suitable for two-class classification, in this paper, the output layer of EfficientNets is replaced with a one-unit dense layer and the sigmoid activation function. We also modifies the top layers by adding global average pooling, batch normalization [22], and dropout [23] layers. The resulted architecture is presented in Table II where the base model is an EfficientNet model. In fact, the base model can be other models where, in this paper, we also test DenseNet-169, Xception, and MobileNetV2 for comparison.

The number of parameters of the transfer learning networks depends on the type of base model, as summarized in Table III, where EfficientNet-B0 and EfficientNet-B3 are about the same size as MobileNetV2 and DenseNet-169, respectively, in terms of the number of parameters.

V. EXPERIMENTAL SETUPS

A. Preprocessing

The transfer learning architecture in Table II are trained separately on each type of studies in an end-to-end fashion. The data preprocessing steps are kept minimal. In both training and validation sets, each radiographic image is resized and padded to a target width and height, depending on the choice of architecture, as listed in Table III where the input dimensions of EfficientNets follow the Keras implementation of EfficientNets. We apply zero-centering with respect to the mean of ImageNet dataset, without scaling. Unlike the preprocessing steps in other literature such as [12], [18], and [16], we do not perform manual image processing, for example, by morphological operation, smoothing, or contrast/brightness adjustments.

B. Data augmentation

Since the number of images in the training set for each type of studies is rather small, data augmentation is performed on the training set by a random horizontal flip to increase data variations. Although, in [1], the random image rotation of 30 degrees was implemented, [16] argues that, according to their experimental results on the arm radiographs in the MURA dataset, the random rotation is ineffective but horizontal flipping could be helpful for training. Therefore, we limit the data augmentation to only the random horizontal flipping.

C. Model training

In the MURA dataset, the radiographic data are labeled at the study level; however, during training, each image is considered to have its own label. We train one model for one study type. The base model of the architecture in Table II is initialized with the ImageNet pre-trained weights, while the output unit is initialized by Glorot normal initialization. Other training setups are partially similar to [1], i.e. the loss function is defined as the weighted binary cross entropy. The mini-batch size is set to 8 and the optimizer is Adam with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 10^{-4} and is reduced by a factor of 10, each time the training loss has stopped improving. Furthermore, early stopping is also applied by monitoring the validation Kappa score. The training loop will stop if the score does not increase for more than five consecutive epochs. After the last epoch in training a model is finished, the model with the highest validation Kappa score, of the image level, is saved.

D. Test time and model ensemble

As mentioned in Section V-C, one type of study has one model, trained at image level. However, for an image, it is possible that the abnormality is unnoticeable, even though the image belongs to the study labeled as abnormal. This is because the image may be captured at the angle where the abnormality cannot be seen. Therefore, at test time, the abnormality probabilities of all the images in one study, outputted from one model, are averaged by the arithmetic mean. Any studies with the average probability larger than 0.5 are considered abnormal.

According to [24], ensembles of smaller models can result in both higher accuracy and efficiency than a single large model. In fact, model ensemble is considered one type of model scaling methods. For example, ensembles of EfficientNet-B3 may perform better than EfficientNet-B4. As a result, the training process in Section V-C is independently performed for 5 times to obtain 5 models. The predicted abnormality probabilities of one study by each of 5 models are then ensembled by averaging using the arithmetic mean. A study is interpreted as abnormal if the ensemble average is larger than 0.5.

TABLE IV
VALIDATION KAPPA SCORES OF EFFICIENTNET-B3 TRAINED WITH ONLY IMAGENET PRE-TRAINED CHECKPOINT VS. WITH BOTH IMAGENET AND MURA PRETRAINED CHECKPOINTS.

Study type	ImageNet pretrained	ImageNet+MURA pretrained
Elbow	0.7341	**0.7468**
Finger	0.6542	**0.7584**
Forearm	0.6657	**0.7421**
Hand	0.5720	**0.6001**
Humerus	0.7629	**0.7927**
Shoulder	**0.6591**	0.6177
Wrist	0.7469	**0.7758**
Overall	0.6890	**0.7205**

E. Evaluation metrics

Both accuracy and Cohen's kappa score [5] are used as evaluation metrics. Cohen's kappa score is a more robust metric than accuracy. It measures an inter-rater agreement. The kappa score of 0 implies a random agreement among raters, while the kappa score of 1 implies a complete agreement. The definition of Cohen's kappa κ for two raters is

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the probability of agreement, equivalent to the accuracy, and p_e is the the probability of random agreement between two raters.

VI. EXPERIMENTAL RESULTS

Two sets of experiments were conducted on a single Nvidia RTX 2080Ti 11GB GPU using TensorFlow 2.3. The first experiment shows that training ImageNet pre-trained checkpoints on the whole MURA training data, before training on each study type, helps boost performance. Then, the second experiment evaluates the performance of EfficientNet ensembles.

A. Experiment: MURA pre-trained checkpoints

To evaluation the transfer learning performance of the EfficientNets whether their knowledge from the ImageNet dataset can efficiently help learn to detect an abnormality in a radiographic study, we first independently trained five EfficientNet-B3 models as described in Section V-D, for each type of studies. ImageNet pre-trained checkpoints are used as initial conditions. Since there are 7 types of study, we end up with 35 models, supporting the entire upper extremity. By following the training strategies in Section V-C, the resulted Kappa scores on the validation set on each study type are shown in Table IV, with the overall Kappa score of 0.6890. However, the overall score is rather low compared to the result on the test set ($\kappa = 0.705$) from DenseNet-169 in [1].

We suspect that the poor performance was due to the small sample size of the training data. As a result, we modifies the training scheme as follows. After the model initialization by the ImageNet pre-trained checkpoint, the model is then trained on the whole MURA training set (all study types) using the training configurations according to Section V-C. After the training terminates, the model checkpoint (called the MURA pre-trained checkpoint) with the highest validation Kappa score is restored and used as the initial checkpoint for training each individual study type. We found that using the MURA pre-trained checkpoint as an immediate step in training a study type largely improves the performance as shown in Table IV with the overall $\kappa = 0.7205$. As a result, the experimental results in the rest of this section will follow the training scheme of using ImageNet+MURA pre-trained checkpoints.

B. Experiment: Model ensemble

The base models that will be evaluated in this experiment are EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, and EfficientNet-B3. Moreover, we also experiment with the three baseline models, MobileNetV2 [8], DenseNet-169 [6], and Xception [7], for comparison. Our DenseNet-169 differs from [1] in that the input size is 300×300 pixels, whereas [1] used the size of 320×320. In addition, our DenseNet-169 is evaluated on the validation set. Therefore, for all models to be comparable, we also train DenseNet-169 in our settings. The experimental results for all the models are presented in Table V for the baseline models and Table VI for the EfficientNets.

The results in Tables V and VI are obtained by performing Section V-D three times for each type study where the average Kappa score and the standard deviation of the study level among three ensemble models are reported in the tables. The two tables also highlight the highest Kappa score among all models for each study type in bold, across both tables.

C. Discussion

A practical guideline in transfer learning normally suggests to freeze some layers or parameters during an initial phase of training. As the first step, we tried freezing the entire base model and trained only the top layers. However, this strategy did not yield a good result. We also experimented with unfreezing some MBConv blocks in the EfficientNet architecture. However, we observed no significant difference in the performance. Therefore, all models in this paper were trained without freezing any parameters. This idea is also supported by the experimental results in [9]. The rationale behind this is that the radiographic images in the MURA dataset are largely different from the images in the ImageNet dataset; therefore, most parameters are needed to be fine-tuned.

Moreover, due to the small number of images in each study type, the ImageNet pre-trained checkpoints do not efficiently learn a new domain of radiographic images. Consequently, Section VI-A experimentally shows that it is also necessary to retrain the ImageNet checkpoints with the entire MURA training data, prior to training with one study type, in order to improve the performance. This is rational because all radiographic images in the dataset share the same abstract features. Therefore, a model trained on elbow images, for example, can still benefit from the other types of studies such as wrist images.

In Section VI-C, we further built the ensemble models to evaluate whether the transfer learning architectures based on EfficientNets can be more efficient than based on the baseline models. The experimental results in Tables V and VI show that the performance in each study type is quite varied. However, it can be observed that EfficientNet-B1, EfficientNet-B2, and EfficientNet-B3 together achieved the

TABLE V

Validation accuracy and Cohen's Kappa from the baseline models with the standard deviations in parentheses.

Study type	MobileNetV2		DenseNet-169		Xception	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Elbow	0.873 (0.006)	0.733 (0.014)	0.861 (0.017)	0.707 (0.037)	0.859 (0.010)	0.705 (0.022)
Finger	0.815 (0.013)	0.627 (0.026)	0.838 (0.009)	0.673 (0.017)	0.870 (0.014)	**0.739 (0.029)**
Forearm	0.837 (0.017)	0.671 (0.035)	0.850 (0.008)	0.696 (0.015)	0.845 (0.004)	0.686 (0.009)
Hand	0.818 (0.012)	0.599 (0.030)	0.804 (0.015)	0.578 (0.032)	0.826 (0.012)	0.620 (0.023)
Humerus	0.901 (0.011)	0.802 (0.023)	0.911 (0.000)	**0.822 (0.000)**	0.911 (0.000)	**0.822 (0.000)**
Shoulder	0.797 (0.018)	0.593 (0.037)	0.794 (0.015)	0.587 (0.031)	0.811 (0.013)	0.621 (0.026)
Wrist	0.873 (0.011)	0.730 (0.023)	0.883 (0.009)	0.752 (0.019)	0.893 (0.006)	0.773 (0.014)
Overall	0.844 (0.002)	0.680 (0.004)	0.848 (0.003)	0.688 (0.007)	0.859 (0.005)	0.712 (0.009)

TABLE VI

Validation accuracy and Cohen's Kappa from the EfficientNets with the standard deviations in parentheses.

Study type	EfficientNet-B0		EfficientNet-B1		EfficientNet-B2		EfficientNet-B3	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Elbow	0.882 (0.019)	0.753 (0.042)	0.890 (0.004)	**0.772 (0.007)**	0.878 (0.010)	0.744 (0.020)	0.871 (0.007)	0.730 (0.015)
Finger	0.832 (0.003)	0.663 (0.006)	0.842 (0.009)	0.682 (0.017)	0.840 (0.006)	0.677 (0.011)	0.865 (0.017)	0.728 (0.035)
Forearm	0.845 (0.019)	0.687 (0.038)	0.845 (0.004)	0.687 (0.008)	0.852 (0.009)	0.701 (0.018)	0.855 (0.016)	**0.706 (0.032)**
Hand	0.794 (0.018)	0.560 (0.035)	0.824 (0.007)	0.622 (0.014)	0.826 (0.006)	**0.625 (0.011)**	0.812 (0.003)	0.595 (0.009)
Humerus	0.901 (0.004)	0.803 (0.009)	0.899 (0.004)	0.798 (0.009)	0.911 (0.000)	**0.822 (0.000)**	0.904 (0.013)	0.807 (0.026)
Shoulder	0.789 (0.010)	0.576 (0.021)	0.801 (0.015)	0.601 (0.030)	0.797 (0.016)	0.594 (0.031)	0.827 (0.015)	**0.652 (0.030)**
Wrist	0.892 (0.010)	0.772 (0.021)	0.902 (0.014)	**0.792 (0.029)**	0.897 (0.006)	0.783 (0.013)	0.895 (0.000)	0.777 (0.001)
Overall	0.847 (0.000)	0.689 (0.000)	0.858 (0.006)	0.710 (0.012)	0.857 (0.003)	0.707 (0.006)	0.861 (0.002)	**0.717 (0.003)**

highest performance 6 out of 7 study types, i.e. all types except for the finger study in which Xception received the highest Kappa score.

Moreover, for the best overall Kappa score, we observe that the ranks in an ascending order are as follows: MobileNetV2, DenseNet-169, EfficientNet-B0, EfficientNet-B2, EfficientNet-B1, Xception, and EfficientNet-B3. This order of performance is in line with the ImageNet's top-1 accuracies reported in the literature, according to Table III, except that in our case, EfficientNet-B2 is ranked lower than EfficientNet-B1 and Xception.

MobileNetV2 has the lowest number of parameters among all models and it also achieved the lowest overall Kappa score. EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, and EfficientNet-B3 are preferred over DenseNet-169 since the number of parameters are much lower, while still providing the higher performance. Likewise, EfficientNet-B3 is superior to Xception in terms of both the number of parameters and the performance. It is worth noting that, our DenseNet-169 baseline has the overall Kappa score of 0.688 which is lower than the score reported in [1]. This is possibly due to many factors, such as different input size, validation/test set, and optimization settings. However, as the training and validation schemes in our experiments are the same for all models, we conclude that the EfficientNets are better than the DenseNet-169, not only on the ImageNet, but also on the MURA dataset.

Finally, to understand how a trained model weighed regions of an image to make a prediction, we implemented class activation mappings (CAM) [25] to visualize the decision heat maps. A visualization example for an EfficientNet-B3 is given in Fig. 1 for a study of finger radiographs with two views. The study is labeled as abnormal, and the model predicted the abnormality probabilities of 0.715 and 0.957 for each view, suggesting that the study has the average probability of 0.836 which is considered abnormal. From the figure, it is clear that the heat maps concentrate around the top joint of the index finger. After reaching out to an experienced radiologist for a comment, we were confirmed that the abnormality is slightly above the top joint. Therefore, the model did weigh its prediction around the right region of the images.

VII. Conclusion

In this paper, we investigated deep transfer learning for detecting abnormalities in the radiographic images of the upper extremity. The experimental results showed that the mobile-size architectures, called EfficientNets, which are trained on the ImageNet dataset, can be efficiently transferred their knowledge to the radiographic image dataset, MURA. Since the number of images in each study type is limited, it is beneficial to pre-training the model on the entire MURA dataset as a starting point for training a specific study type. The transfer learning models of EfficientNets were ensembled and shown to achieve higher Cohen's Kappa scores than the MobileNetV2 ensemble and the DenseNet-169 ensemble, while also having fewer parameters. Moreover, the best model in the experiments was EfficientNet-B3 which was also superior to Xception in both effectiveness and efficiency.

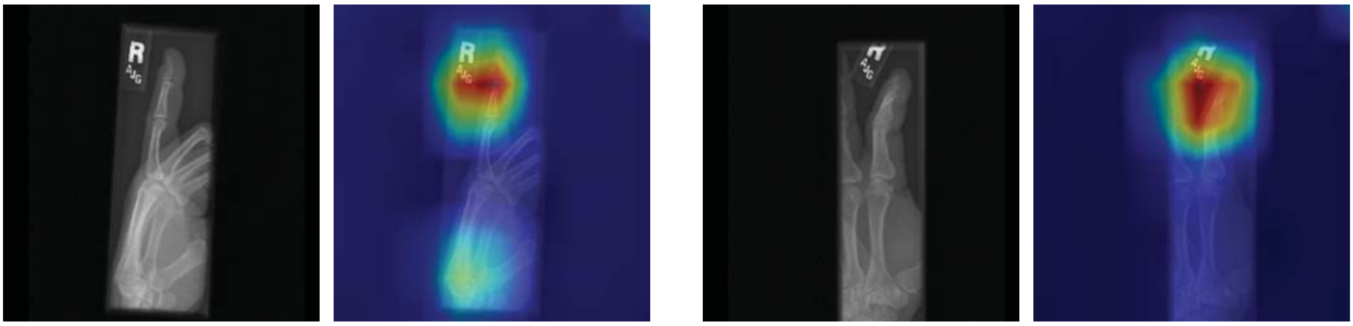
Acknowledgment

This work was financially supported by the Murata Science Foundation (MSF), 2019. We also thank Kamolsin Teeyapun, M.D., for the valuable comments on the radiographs.

References

[1] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs," in *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, Amsterdam, The Netherlands, jul 2018.

[2] P. Chea and J. C. Mandell, "Current applications and future directions of deep learning in musculoskeletal radiology," *Skeletal Radiology*, vol. 49, no. 2, pp. 183–197, feb 2020.



(a) Predicted abnormality probability of 0.715

(b) Predicted abnormality probability of 0.957

Fig. 1. A study of finger radiographs with two views: (a) and (b). For each view, the radiograph (left) is compared with the heat map (right) produced from EfficientNet-B3. The label of the study is abnormal and the model also predicted as abnormal at both the image and study levels.

[3] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning, PMLR*, vol. 97, may 2019, pp. 6105–6114.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2009, pp. 248–255.

[5] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jul 2017, pp. 2261–2269.

[7] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jul 2017, pp. 1800–1807.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, dec 2018, pp. 4510–4520.

[9] D. D. A. Dias, “Musculoskeletal Abnormality Detection on X-Ray Using Transfer Learning,” Master’s thesis, Universitat Pompeu Fabra, 2019.

[10] N. Harini, B. Ramji, S. Sriram, V. Sowmya, and K. Soman, “Musculoskeletal radiographs classification using deep learning,” in *Deep Learning for Data Analytics*. Elsevier, jan 2020, pp. 79–98.

[11] G. Chada, “Machine Learning Models for Abnormality Detection in Musculoskeletal Radiographs,” *Reports*, vol. 2, no. 4, p. 26, 2019.

[12] T. C. Mondol, H. Iqbal, and M. Hashem, “Deep CNN-Based Ensemble CADx Model for Musculoskeletal Abnormality Detection from Radiographs,” in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, sep 2019, pp. 392–397.

[13] H. El-Saadawy, M. Tantawi, H. A. Shedeed, and M. F. Tolba, “A Two-Stage Method for Bone X-Rays Abnormality Detection Using MobileNet Network,” in *Advances in Intelligent Systems and Computing*. Springer, 2020, vol. 1153 AISC, pp. 372–380.

[14] S. Panda and M. Jangid, “Improving the Model Performance of Deep Convolutional Neural Network in MURA Dataset,” in *Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, vol 141*, 2020, pp. 531–541.

[15] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *the 3rd International Conference for Learning Representations*, San Diego, dec 2015.

[16] B. Guan, G. Zhang, J. Yao, X. Wang, and M. Wang, “Arm fracture detection in X-rays based on improved deep convolutional neural network,” *Computers & Electrical Engineering*, vol. 81, p. 106530, jan 2020.

[17] A. F. Saif, C. Shahnaz, W. P. Zhu, and M. O. Ahmad, “Abnormality Detection in Musculoskeletal Radiographs Using Capsule Network,” *IEEE Access*, vol. 7, pp. 81 494–81 503, 2019.

[18] S. Liang and Y. Gu, “Towards Robust and Accurate Detection of Abnormalities in Musculoskeletal Radiographs with a Multi-Network Model,” *Sensors*, vol. 20, no. 11, p. 3153, jun 2020.

[19] Y. Huang, Y. Cheng, A. Bapna, O. Firat, M. X. Chen, D. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu, and Z. Chen, “GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism,” in *Advances in neural information processing systems*, 2019, pp. 103–112.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, sep 2017.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[22] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, feb 2015, pp. 448–456.

[23] R. S. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] D. Kondratyuk, M. Tan, M. Brown, and B. Gong, “When Ensembling Smaller Models is More Efficient than Single Large Models,” may 2020. [Online]. Available: http://arxiv.org/abs/2005.00570

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2016, pp. 2921–2929.