Interpretability for the Automated Diagnosis of Musculoskeletal Radiographs

Nisha Balaji

Abstract

Despite tremendous developments in performance, diagnostic machine learning algorithms are widely regarded as "black boxes." To break down this perception, I propose an investigation of interpretable models: systems that substantiate their conclusions with "justifications." Our models achieved AUROC scores of 0.890, 0.832, 891, 0.832, 0.898, 0.838, 0.897 for the anatomical regions of elbow, finger, forearm, hand, humerus, shoulder, and wrist, respectively. Through the perspective of musculoskeletal imaging, this work explores two different venues for interpretability: image gradient saliency mapping (IGSM) and local interpretable model-agnostic explanations (LIME). I trained an Inception-v3 neural network to detect abnormal radiographs. The different LIME and IGSM images provided insight into the model's understanding of different regions of images in addition to divulging how the different interpretable methodologies interacted with the model. The implications of such developments in interpretability within healthcare range from auditing models to garnering patient and physician trust.

1 Introduction

Since its debut into medical technology, artificial intelligence has become increasingly prevalent in the medical environments. It has aided in the development of new drugs, management of vast amounts of patient data, clinical prognosis, and even assistance in surgery [5]. One of its most prominent uses is within diagnostics: machine learning models can be used to detect abnormal conditions within medical slides. In recent years, artificial intelligence algorithms have developed to be increasingly competitive to - and in certain cases even more accurate than - practicing radiologists [17]. Despite such progress, a survey revealed that 97% of referring physicians trust human radiologists, in contrast to a mere 33% that trust diagnostic artificial intelligence models [28].

In this study, I consider a solution to close the disparity of trust with an investigation of interpretable models. Such models are capable of "justifying" their conclusions, and can thereby break down the perception of machine learning algorithms as "black box systems."

I specifically investigated interpretability in musculoskeletal radiographs. Musculoskeletal disorders are among the most prevalent, occurring in between 20 to 33% of the world's population. With the exception of cancerous diseases, musculoskeletal conditions are the primary cause of chronic pain [15]. Radiographs are vital to the diagnosis of musculoskeletal abnormalities and can provide crucial information regarding the type and location of an abnormality [26].

Medical imaging is crucial to the diagnosis of other disorders beyond musculoskeletal abnormalities [15]. However, radiologists continue to be overworked in our medical systems, with 45% of radiologists enduring burnout [12]. Increasing the presence of medical artificial intelligence technologies in radiology can substantially minimize radiologist workloads. Previous studies have shown that radiologists have both increased specificity and sensitivity when utilizing the aid of machine learning algorithms as second opinions [23]. In addition, in medical emergency setting where time is often an extremely important factor, artificial intelligence models can assist in efficient diagnosis and triaging [13].

In this work, I explore Image Gradient Saliency Maps (IGSMs) and Local Interpretable Model-Agnostic Explanations (LIMEs) as possible methods to construct model visualizations of musculoskeletal radiographs [22, 20]. Upon utilizing transfer learning to construct an Inception-v3 model finetuned to the task of identifying musculoskeletal disorders from the MURA dataset, I applied both image explainability techniques to the algorithm.

Previous research in medical interpretable technologies have explored different methods in model visualization. Object-detection neural networks have been utilized to detect symptoms of COVID-19 in chest computed tomography scans with the use of YOLO-v3 network, where bounding boxes are implemented to localize regions of an image slides indicative of COVID-19 [27]. Such studies require annotated datasets with the bounding boxes. With respect to classification models, the study that introduced the MURA dataset utilized Class Activation Maps to differentiate the portions of the radiograph that represented a musculoskeletal abnormality [19]. The LIME approach to model visualizations has been previously utilized in the classification of lymph node metastases [6].

2 Purpose

1. Fine-tune a pre-trained Inception-v3 models to classify musculoskeletal radiographs as either "normal" or "abnormal."
2. Utilize Image Gradient Saliency Maps (IGSMs) and Local Interpretable Model-Agnostics Explanations (LIMEs) to extract regions of an image most indicative of abnormality according to the fine-tuned model.

3 Methods

3.1 Dataset

The MURA dataset was used for the purposes of this study [19]. The dataset encompasses of 40,561 X-rays of seven different anatomical regions: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each anatomical region contains a set of studies; each study is labeled as either "0" or "1" for "normal" and "abnormal" respectively. There are 36,808 images in the training set and 3,197 images in the validation set. The abnormalities documented in this dataset includes fractures, hardware, degenerative joint disease, lesions and subluxations among others.

3.2 Convolutional Neural Network

3.2.1 Image Processing

Prior to training the neural network, all of the radiographs were pre-processed in accordance with the Inception-v3 architecture guidelines [4]. Center cropping to 312 by 312 images was performed to remove the variance in image dimensions. Images were resized to 299 by 299 pixels since the Inception-v3 model only accepts images of that size. In addition, due to the redundancy of color within radiographs, images were converted to three-channel grayscale. Normalization of the pixel values was implemented to systemize pixel values into a standardized range.

3.2.2 Transfer Learning

The Inception-v3 model is 48-layer convolutional neural network pre-trained for the ImageNet database [24]. The Inception-v3 architecture was chosen since it has among the lowest Top 1% and Top 5% error rates out of all pre-trained models [1]. The model was fine-tuned by unfreezing the parameters of the model; upon doing so, all weights were recalculated in accordance to the provided dataset [10].

The pre-trained model was separately fine-tuned for each of the seven different anatomical regions in the MURA dataset to prevent variations in morphological region. The completed transfer learning therefore yielded seven different sets of results: one for each anatomical region.

Binary cross entropy with logits loss and a stochastic gradient descent optimizer were utilized.

In regard to model training, an early-stopping algorithm was implemented. If the validation accuracy of the training model did not improve for four consecutive epochs, training was halted. Each of the seven fine-tuned models trained in less than 23 minutes on Google Colab Pro GPUs.

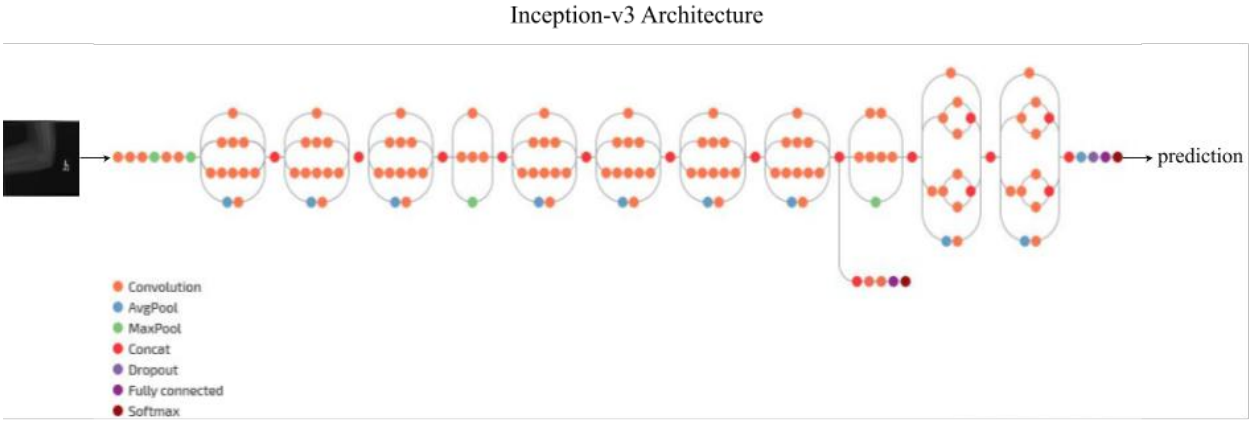


Fig. 1: Transfer learning was implemented with Inception-v3 to provide abnormality prediction when given an input radiograph. [3]

3.3 Interpretability

Model image visualizations are used to determine which regions of an image an algorithm perceives as most important in its final classification. These "influential regions" are denoted by superpixels.

3.3.1 Image Gradient Saliency Maps

To construct IGSMs, I first retrieve the output class score for the given radiograph [20]. Upon doing so, the gradient of the score is extracted in reference to the input image pixels. If the gradient for a specific pixel has a relatively large absolute value, it can then be concluded that even minute changes to those pixels would substantially contribute to the class scores.

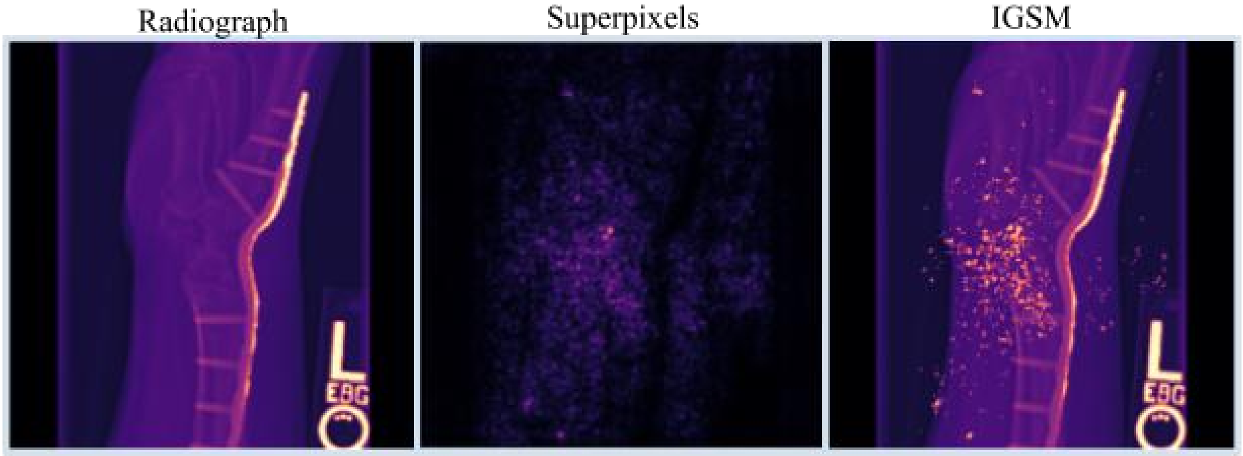


Fig. 2: The original image's model visualization is shown in the heatmap. This is further processed by selectively choosing the brightest superpixels to appear in the final saliency map.

When translated to a saliency map, larger gradients are represented as brighter superpixels. The superpixels are then transferred onto the original radiograph for improved visualization. To prevent an over-congestion of superpixels, the superpixels with brightness in the lowest 25th percentile (typically, black and dark red superpixels) are discarded. By doing so, the final saliency maps effectively isolate the superpixels with the largest gradients while ensuring that the maps remain intelligible and without an over-congestion of superpixels.

3.3.2 Local Interpretable Model-Agnostic Explanations

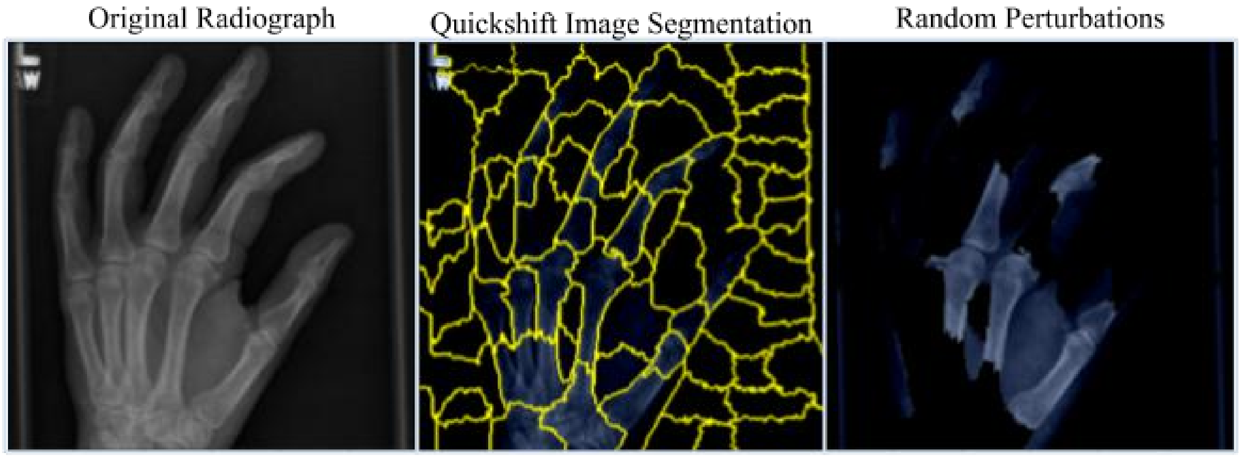


Fig. 3: Upon implementing the quickshift algorithm, the segmented superpixels are represented with yellow borders. By randomly toggling specific superpixels on and off, I are able to construct a random perturbation of the radiograph.

The LIME construction methodology involved the generating of image perturbations [2, 22]. Image perturbations are specific, induced changes in an image that can impact the algorithm's interactions with the image [18].

Prior to perturbation generation, radiographs undergo quickshift image segmentation, a specific image segmentation algorithm that operates locally on spatial and color dimensions [7, 16]. The image segmentation produces a set of superpixels, that are then randomly toggled between "on" and "off" states, which can be represented by random perturbation vectors [2]. I chose to generate 96 different random perturbation vectors. The perturbation vectors are represented such that each index represents a superpixel, and if the value within the index is "1" the superpixel is "on," and if the value is "0" the superpixel is turned "off". For each of the perturbed images, the cosine distance metric is used to compute the similarity between each perturbation and the original image. The cosine distances are then converted to weights in the range (0, 1) to determine the importance of each perturbation in accordance to its difference from the original image. Additionally, abnormality predictions are retrieved for each of the different image perturbations with the fine-tuned Inception-v3 models.

After generating perturbations, finding weights, and computing abnormality predictions, I construct a linear model that finds class predictions when given perturbation vectors and weights for each vector. Upon constructing the linear regression, I extract the coefficients for each superpixel from the perturbations vector. The coefficients represent the relative influence of each superpixel in the model interpretation.

In the final LIME, the 10 superpixels with the highest coefficients appear on a separate image.

It should be noted however, that LIMEs are far more accurate when a greater amount of pertubations are generated. In our instance, only 96 random perturbations are generated due to limitations in RAM availability. An increased number of generated perturbations would most likely improve LIMEs due to the availability of a larger set of samples when constructing the linear model; therefore allowing for more informed predictions for superpixel coefficients.

4 Results

4.1 Classification Performance

The performance metrics for each of the fine-tuned models are included in Table 1. The AUROC graphical representations and confusion matrices for each of the anatomical regions can be found in appendices A and B respectively. The Area Under the Receiving Operating Curve (AUROC) is defined as a metric

Anatomical Region	AUROC	Sensitivity	Specificity
Elbow	0.890	0.770	0.894
Finger	0.832	0.761	0.799
Forearm	0.891	0.656	0.933
Hand	0.832	0.619	0.915
Humerus	0.898	0.850	0.865
Shoulder	0.838	0.705	0.856
Wrist	0.897	0.695	0.956

Tab. 1: Metrics for the Fine-tuned Inception-v3 Models Classification Performance by Anatomical Region

for how well a model distinguishes between classes [14]. Sensitivity is defined as a medical metric for ability to identify disorder (abnormality), and specificity is the ability to identify lack of disorder (normality) [25]. The ROC curves and confusion matrices for each anatomical region's fine-tuned model can be found in appendices A and B, respectively.

4.2 Image Gradient Saliency Maps

The IGSM localization was varied across different anatomical regions and different abnormality types. Several characteristics differed across the IGSMs including the density of superpixels in the image, the existence of several clusters of superpixels, the number of superpixels on the image, and whether or not superpixels highlighted blackspace in radiographs. There tends to be a spectrum in terms of abnormality detection with the saliency images: some IGSMs contained superpixels in arbitrary regions irrelevant to disorder location.

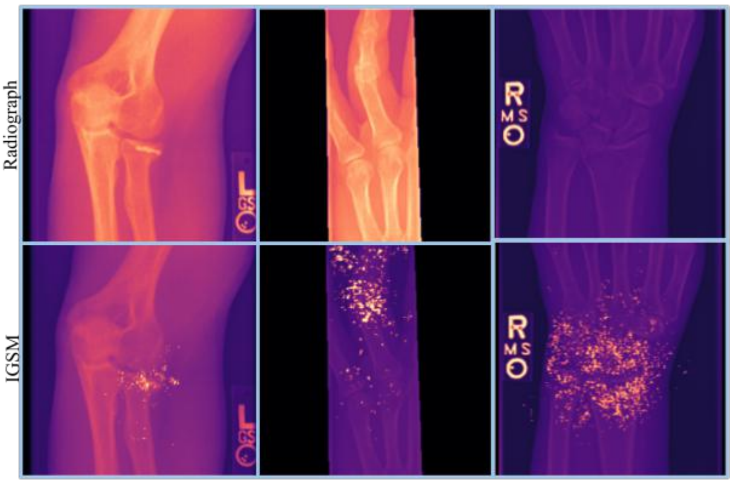


Fig. 4: IGSMs of an elbow, finger, and wrist radiograph.

4.3 LIME

The LIMEs also had a wide range of representations. More specifically, the variances in the LIMEs include the number of distinct segments highlighted, and the extent to which abnormalities were highlighted. Similar to IGSMs, there is a spectrum in terms of localizing abnormalities and not all LIMEs describe disorder regions as influential in model classification.

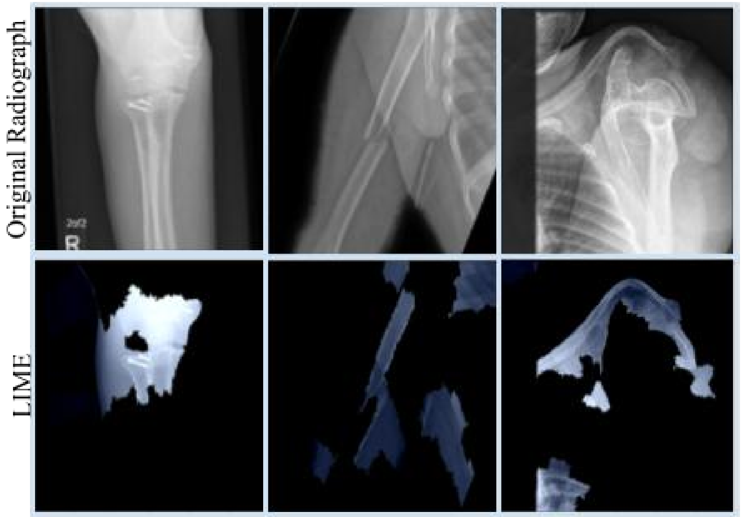


Fig. 5: LIMEs of an elbow, humerus, and shoulder radiograph.

4.4 Comparing LIME and Image Gradient Saliency Maps

The LIME and IGSMs displayed different behaviors in terms of constructing model interpretation images. For identical radiographs, it was not uncommon to find the LIMEs highlighting an entirely different region than that of the corresponding IGSM. For instance, in the leftmost column of Figure 6, note that while the LIME localizes the hardware as the most significant region of the model's interpretation, the IGSM contains no particular clustering of superpixels in an area; the superpixels are scattered over gray space and regions of the wrist. In addition, the LIMEs typically had multiple disconnected superpixel groups whereas the IGSMs usually did not have more than one distinct cluster of superpixels. In the IGSMs of the center and rightmost column in Figure 6, only a single cluster of superpixels is noticeable - a cluster near the fracture and a cluster near the base of the thumb respectively; however, the LIMEs of identical radiographs have at least two and at least four distinct parts respectively.

5 Discussion and Considerations for Future Investigations

While the AUROC scores of the different models are much higher than 0.5 (the no discrimination AUROC score), improvements can be made to the classification models. In particular an exploration of different model architectures can prove useful: different pre-trained models or customized models with different numbers of layers and different functions at various parts of the neural network.

With respect to the saliency imaging, the discrepancies between the localizations of LIMEs and IGSMs can be attributed to the vastly different methods of interpretation that both techniques apply. However, it would be expected in real-world clinical setting that both methods output similar model explanations; such inconsistencies should be rectified. In addition, the localization of black-space divulges that the Inception-v3 does not always consider relevant regions of an image when classifying radiographs. To improve the performance of the saliency imaging, an integration of local rule-based explanations, which utilize genetic algorithms and derivation counterfactual rules, can prove beneficial [9]. Influence functions, statistical tools used to evaluate the influence of removing an observation, are also relevant to the study of interpretable models [11].

Other avenues to explore within interpretability include linguistic explanations for diagnostic choices can offer further clarity into the model's decision making process [21]. Another possible course of exploration would be composition of example and counterexample groups for specified images. These groups can serve as comparisons to the original input radiograph [21].

Interpretability holds monumental implications in real-world settings [21]. In the context of regulation, interpretability can serve as a qualitative metric for model evaluations. Auditing models on the

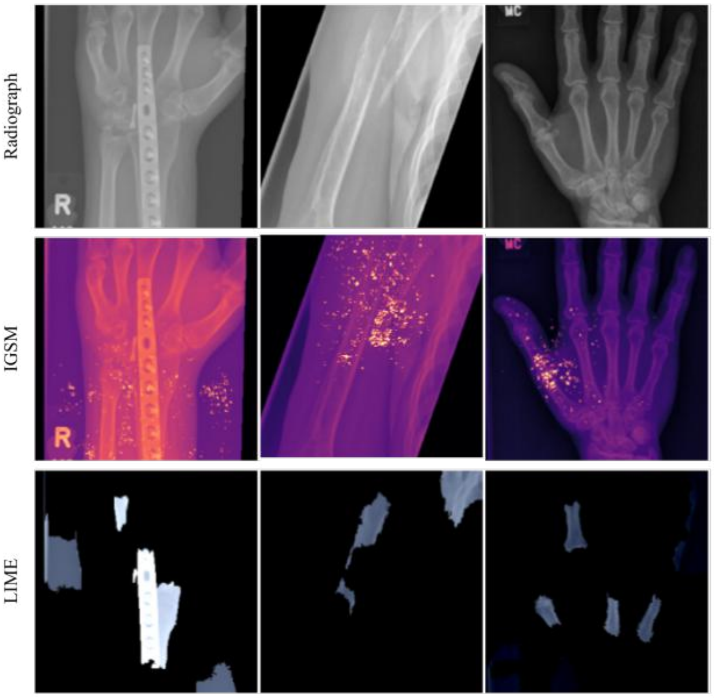


Fig. 6: A comparison of LIME and IGSMs.

basis of their internal mechanisms can establish quality assurance. In addition, interpretable models garner trust within physicians and patients. An increased confidence in diagnostic models can augment their use within clinics, thereby streamlining workflow and improving patient outcomes while remaining feasible and affordable [8].

6 Conclusion

This work proposes a method to construct interpretable models for diagnosing abnormality in musculoskeletal radiographs. Interpretable models hold immense potential for several different applications within the healthcare system. In addition to acting as a qualitative measure of model performance, interpretability instills trust into physicians and patients. It breaks down the conception of machine learning algorithms as "black box" systems. Furthering the precision of these methods and exploring other venues within interpretability can augment the use of artificial intelligence diagnostics in clinics. Such an integration would inevitably improve patient outcomes, while being fiscally affordable and productivity-wise efficient.

References

[1] Torchvision.models. 2019.

[2] Cristian Arteaga. Interpretable machine learning for image classification with lime. 2019.

[3] Semyon Boyko. Using modified inception v3 cnn for video processing and video classification. 2019.

[4] Sasank Chilamkurthy. Writing custom datasets, dataloaders and transforms. 2017.

[5] Sam Daley. Surgical robots, new medicines and better care: 32 examples of ai in healthcare., July 2019.

[6] Iam Palatnik de Sousa, Marley Maria Bernardes Rebuzzi Vellasco, and Eduardo Costa da Silva. Local interpretable model-agnostic explanations for classification of lymph node metastases. 2019.

[7] Brian Fulkerson and Andrea Vedaldi. Quick shift image segmentation. 2018.

[8] GE Healthcare. Three benefits to deploying artificial intelligence in radiology workflows, August 2019.

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018.

[10] Nathan Inkawhich. Finetuning torchvision models. 2017.

[11] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2017.

[12] Matt Morrill and Catherine Kosse. Are your radiologists overworked? here are 2 ways to delegate their workload. 2019.

[13] Supratik K. Moulik, Nina Kotter, and Elliot K. Fishman. Applications of artificial intelligence in the emergency department. 2020.

[14] Sarang Narkhede. Understanding auc - roc curve. 2018.

[15] World Health Organization. Musculoskeletal conditions, November 2019.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Jasmine Pennic. Ai vs. humans: Ai solution beats stanford radiologists in chest x-ray diagnostics competition. 2019.

[18] Prem Prakash. Perturbation theory in deep neural network (dnn) training. 2020.

[19] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Yi Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *CoRR*, abs/1712.06957, 2017.

[20] Aditya Rastogi. Visualizing neural networks using saliency maps in pytorch. 2020.

[21] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology Artificial intelligence*, May 2020.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier, 2016.

[23] Melisande Rouger. Ai can reduce reading times and improve reader performance in breast and chest. 2019.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

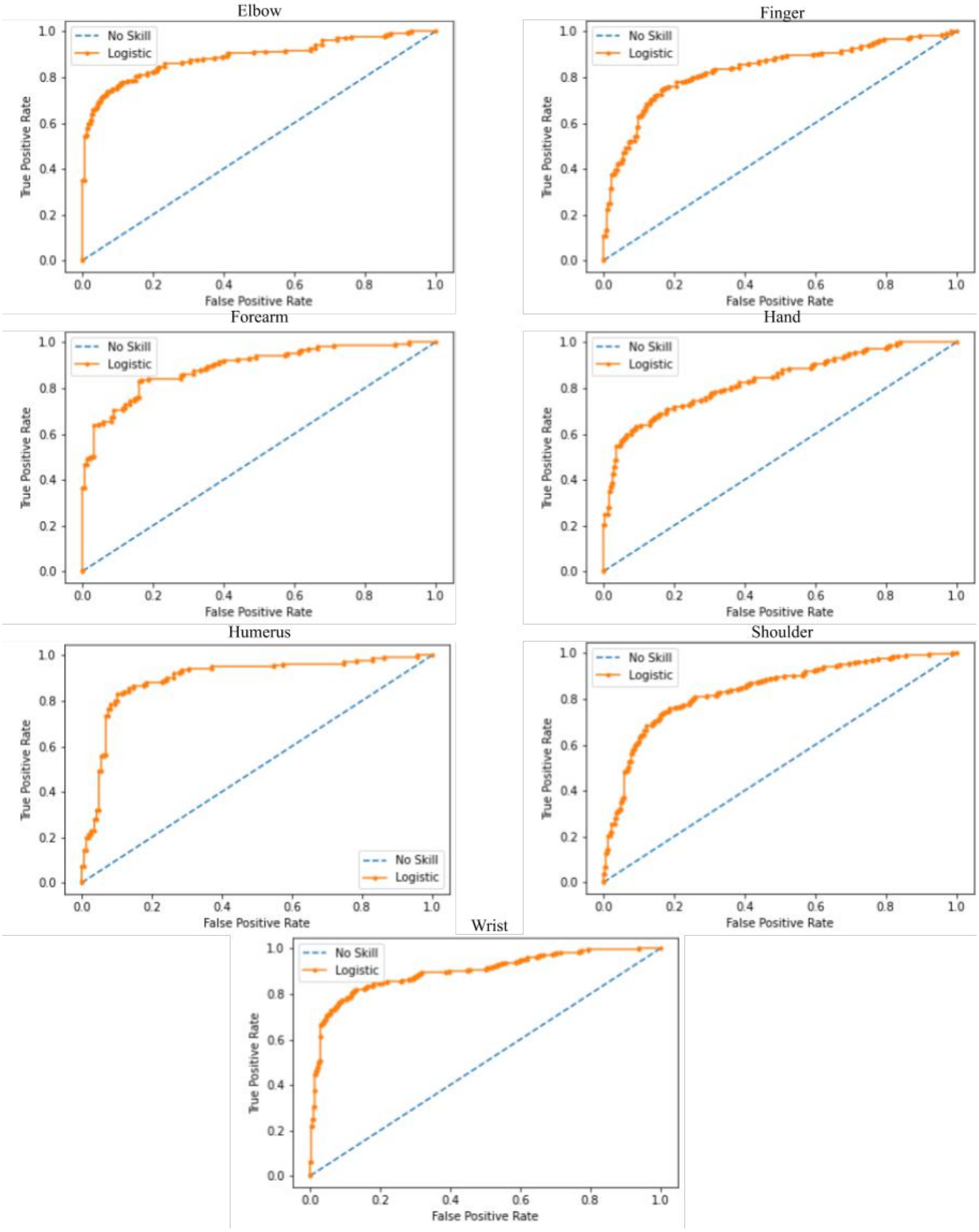
[25] Robert Trevethan. Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. 2017.

[26] Alexandra Villa-Forte. Tests for musculoskeletal disorders, March 2020.

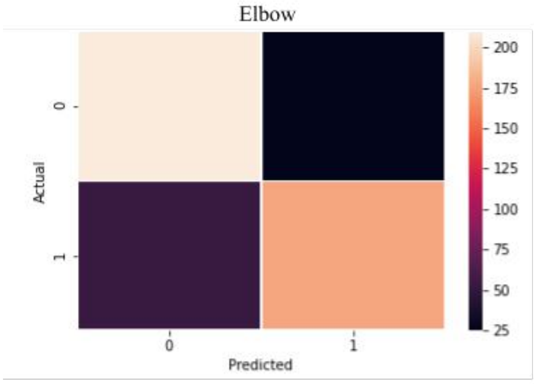
[27] Anmol Warman, Pranav Warman, Ayushman Sharma, Puja Parikh, Roshan Warman, Narayan Viswanadhan, Lu Chen, Subhra Mohapatra, Shyam Mohapatra, and Guillermo Sapiro. Interpretable artificial intelligence for covid-19 diagnosis from chest ct reveals specificity of ground-glass opacities. *medRxiv*, 2020.

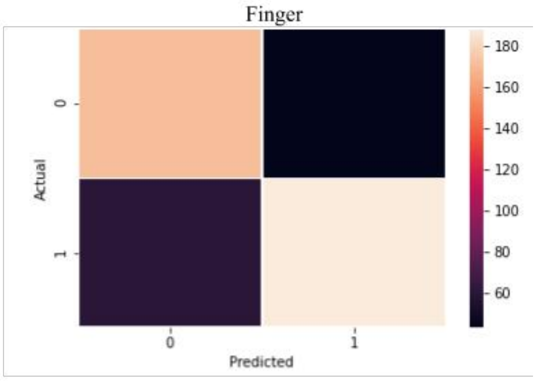
[28] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang "Anthony" Chen. Chexplain. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr 2020.

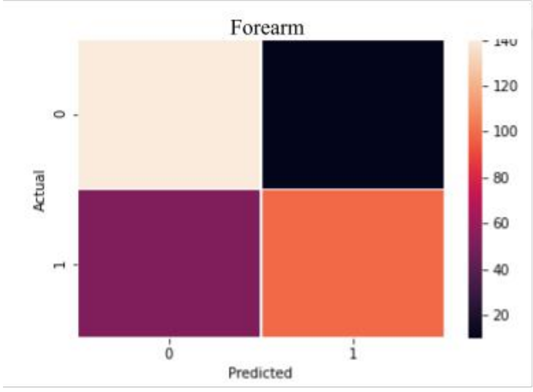
A ROC Curves for Fine-tuned Inception-v3 Models

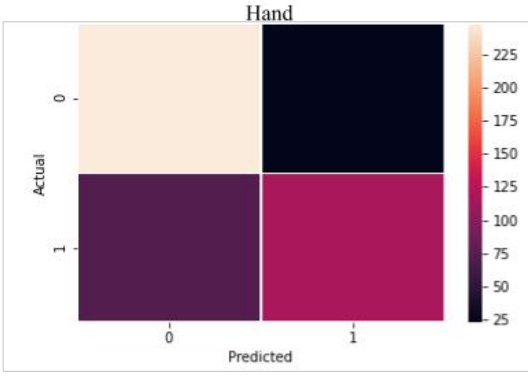


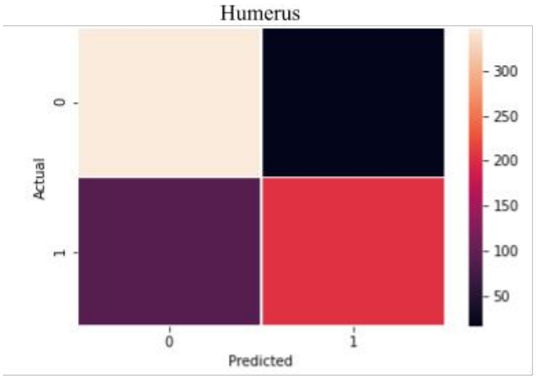
B Confusion Matrices for Fine-tuned Inception-v3 Models

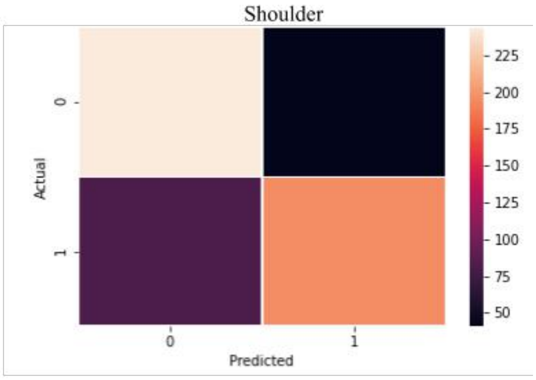

Elbow
Actual
Predicted

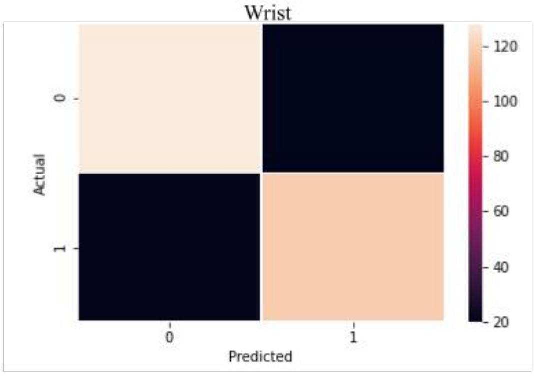

Finger
Actual
Predicted


Forearm
Actual
Predicted


Hand
Actual
Predicted


Humerus
Actual
Predicted


Shoulder
Actual
Predicted


Wrist
Actual
Predicted