Interpretability for the Automated Diagnosis of Musculoskeletal Radiographs

Nisha Balaji
Dougherty Valley High School; San Ramon, United States of America
Email: nishabalaji2003@gmail.com

Abstract – Despite tremendous developments in performance, diagnostic machine learning algorithms are widely regarded as "black boxes." To break down this perception, we propose an investigation of interpretable models: systems that substantiate their conclusions with "justifications." We explore explainable models through the perspective of the diagnosis of musculoskeletal disorders which are among the most prevalent in the world. Two different avenues are pursued for interpretability: saliency imaging techniques to visually localize irregularity region within a radiograph and a clustering of abnormally classified radiographs to isolate different causes of abnormality into separate clusters. As for the model itself, three DenseNet-169 models are fine-tuned to various extents and tested on classification accuracy. The top-bottom (all parameters) fine-tuned model achieves the best AUROC of 0.865, and the saliency maps effectively localize the cause of irregularity within bone X-rays. The clustering algorithm makes substantial associations based on properties such as anatomical region, image orientation, and hardware type. The implications of such developments in interpretability within healthcare range from auditing models to garnering patient and physician trust.

Key Words – computer vision, automated diagnosis, explainable models, saliency imaging

INTRODUCTION

In an era imbued with the advent of technology, instrumentalizing machine learning in the field of medical imaging has become an increasingly prevalent proposition. Within medicine, artificial intelligence has aided in the development of new drugs, managing vast amounts of patient data, clinical prognosis, and even assistance in surgery [2]. One of its most prominent uses is to diagnose abnormal conditions within medical slides (radiographs and magnetic resonance images). Musculoskeletal disorders are extremely widespread, 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 [8]. Radiographs are vital to the diagnosis of musculoskeletal abnormalities [17].

In recent years, artificial intelligence diagnostic models have been at par with human radiologists in identifying disorders. In certain cases, such as chest X-ray diagnostics, machine learning models surpass capabilities of human radiologists [11]. Despite such progress, a survey revealed 97% of referring physicians trust human radiologists, in contrast to a mere 33% that trust diagnostic artificial intelligence models [18]. This illustrates the perception of machine learning systems as "black boxes"–its inner workings remain a mystery and its categorizations dubious due to a lack of interpretability.

Without substantial evidence to rationalize their claims, even stunningly accurate models can prove insufficient for real-world use. Machine learning models must be fashioned to explicitly portray the reasoning behind their diagnostic decisions.

In this work, we propose a method to construct interpretable models which "justify" their conclusions. Upon fine-tuning a DenseNet-169 model for musculoskeletal image diagnosis, two relatively distinct methods are utilized to make further deductions about the bone X-rays. Saliency maps are constructed to isolate regions of the image most indicative of irregularity.

In addition, an unsupervised learning algorithm is applied to cluster abnormally classified images based on disorder type. For instance, all anomalies caused by degenerative joint disease would be categorized into a different cluster than those caused by lesions. Such models can be massively contributory to discovering the cause of irregularity: it can be inferred that the abnormality of an image can be attributed to the general cause of abnormality within its cluster. Furthermore, this study provides a comparison of transfer learning DenseNet-169 models and their ability to accurately classify images and construct saliency maps. To summarize, we augment the applicability of artificial intelligence in medical imaging by making predictions more accessible and justified.

RELATED WORKS

The utilization of artificial intelligence for image classification in medicine has advanced tremendously. The conjunction of computer vision and machine learning has a variety of medical applications–for instance, clinical prognosis. A model to determine the effectiveness of neoadjuvant chemotherapy was constructed to replace an invasive procedure [7]. Machine learning also has a

widespread use in diagnostics. In a study conducted in 2015, an Elastic Net classifier was utilized to categorize brain tumor slides as either glioblastoma multiforme or lower grade gliomas [1].

In both forenamed studies, algorithms are created to assign image slides to a binary classification; however, both models do not provide insight as to how the algorithm arrived at a particular decision–the constraint that the following study contends with. A mere designation to a category is inadequate to completely assess the contents of an image. Medical classification models should be designed to provide explanations for their findings.

An interpretable model has already been constructed to identify articles within photographs from the ImageNet database. The model accomplishes this feat through subtle changes in convolutional neural network architecture [19]. The modified convolutional neural network places an emphasis on filters activated by singular parts within distinct object categories.

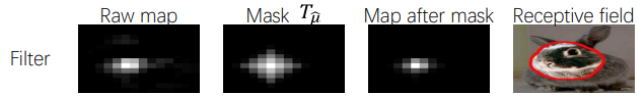

FIGURE 1: The visualizations aid in communicating which parts of the image are most influential in model classification [19].

There is an abundance of applications for interpretability in medicine [15]. Interpretability can be used as metric for auditing models. In addition, interpretable models can provide quality assurance and instill trust into their users.

PURPOSE

1. Fine-tune a pre-trained DenseNet-169 model to classify musculoskeletal radiographs as either "normal" or "abnormal."
2. Accentuate regions of the image slides most influential in the model's diagnosis by using saliency mapping.
3. Cluster resemblant "abnormal" images together and thereby isolate similar causes of abnormality together.

METHODS

I. Dataset

The MURA dataset is utilized for the purposes of this study [13]. The dataset includes 40,561 X-rays of seven anatomic regions organized by study and 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 anatomic regions are elbow, finger, forearm, hand, humerus, shoulder, and wrist.

II. Logistic Regression

As a baseline comparison model, a logistic regression is implemented [10]. Prior to model training, images are processed with a different procedure than that of the convolutional neural network: after center cropping images into 224 by 224 pixels, local binary patterns are extracted from the image. Local binary patterns are image texture descriptors that improve feature detection in certain types of models, such as logistic regressions [12].

In addition, a dimensional reduction with principal components analysis (PCA) is performed on flattened local binary pattern arrays. After PCA, 500 principal components remain. The logistic regression model is trained on these components.

III. Convolutional Neural Network

Prior to training the neural network, the radiographs are processed. First, images are center cropped to 312 by 312 images and then resized into 224 by 224 images. The center-cropping is performed to remove the variance in image dimensions. Images are resized to 224 by 224 pixels since the Dense-Net 169 model only accepts images of that size. In addition, due to the redundancy of color within radiographs, images are converted to three-channel grayscale. In addition, normalization of the pixel values is implemented to systemize pixel values into a standardized range. This processing aids in making the data accessible to fine-tuning the convolutional neural network.

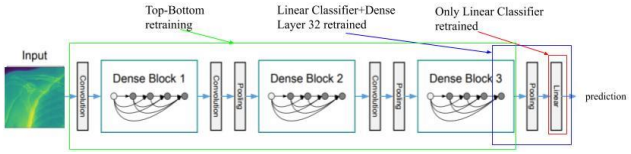

FIGURE 2: The regions within the labeled rectangles represent the parts of the model that are fine-tuned for each of the three model variations.

The DenseNet-169 model is a 169-layer convolutional neural network pretrained for the ImageNet dataset [7, 4]. Three variations of the fine-tuning algorithm are implemented for purposes of comparison:

1. Only the parameters of the linear classifier in the last layer is retrained

2. Both the parameters of the last dense layer (Dense Layer 32) in the third dense block and the last linear classifier is retrained

3. Every parameter of the model, top to bottom, is retrained (top-bottom model)

Binary cross entropy loss function is utilized in addition to stochastic gradient descent with learning rate 0.001 and momentum 0.9. The model calculates and returns the probability that a specified radiograph belongs in the "abnormal" class. If this probability exceeds 0.5, the image is classified as "abnormal." Otherwise it is placed into the "normal" category.

In regard to model training, an early-stopping algorithm is implemented. If the validation accuracy of the model does not improve for four consecutive epochs, the model halts training. Early-stopping and transfer learning are instrumentalized to decrease training time: the model trained in approximately 40 minutes. The model was trained on Google Colab Pro GPUs, and on the training to validation split of 36,808 to 3,197, respectively.

IV. Saliency Imaging

Saliency imaging isolates regions of abnormality within radiographs. The distribution of pixels in the saliency map allows for evaluators to visualize the model's interpretation of the original image.

Saliency mapping is applied to radiographs classified as "abnormal." Upon classification, the gradients of the score for each pixel in the input image are extracted. These weights are represented as superpixels in the saliency map. Brighter super-pixels associate with more important regions [10].

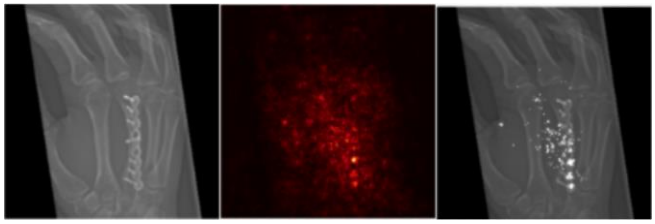

FIGURE 3: The original image's model visualization is shown in the heatmap. This is further processed by selectively choosing the brightest superpixels to appear in the final saliency map.

The saliency maps are then blended into the original radiograph for ease of visualizing. To prevent congested final images, superpixels in the bottom 25th percentile of brightness are discarded. When transferred to the final blended image, saliency map pixels are brightened even more to distinctly appear within the background radiograph.

V. Clustering

A K-means clustering algorithm is performed on images classified as abnormal by the top-bottom model. The abnormally classified images are taken from both the training and validation set to augment the size of the clustering dataset. The model only considers images *classified* as abnormal for consistence with real-world settings where "gold standard" data labels are nonexistent, and the model's classification is the basis for evaluation.

The algorithm clusters features extracted from the top-bottom model prior to the linear classifier layer. Principal components analysis algorithm (PCA) is performed to reduce the dimensionality of the features. After the PCA, a K-means clustering algorithm is executed on the components, with K=4. This specific K-value is chosen since the MURA dataset contains four general categories of abnormalities: fractures, hardware, degenerative joint disease, miscellaneous (typically lesions and subluxations).

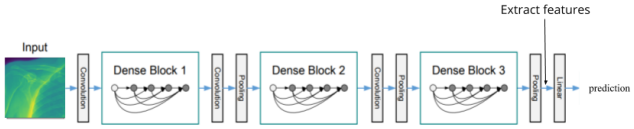

FIGURE 4: Clustering is performed on features extracted from the top-bottom fine-tuned model directly prior to the final linear layer.

In addition to a general clustering of all abnormally classified images (the general set), the clustering algorithm was also implemented within specific anatomic regions.

RESULTS

I. Classification Performance

TABLE 1: Metrics for the top-bottom model in classification performance

F1-Score	ROC-AUC	Recall	Precision
0.775	0.865	0.871	0.698

The performance metrics for the top-bottom model are included in Table 1. The model also attains a validation set accuracy of 80.638%. The proportion of false positives is higher than that of false negatives.

Sensitivity is defined as a medical metric for ability to identify disorder (abnormality), and specificity is the ability to identify lack of disorder (normality) [16]. The sensitivity of the model is 0.871 while it attains a specificity of 0.766.

TABLE 2: Comparison of ROC-AUC scores for all three fine-tuned models

Linear	Dense32 + Linear	Top-Bottom
0.806	0.783	0.865

As shown in Table 2, the three fine-tuned models had varying performances, with the top-bottom fine-tuned model attaining the best ROC-AUC score of 0.865. The fine-tuned models that are exclusively retrained at the last layers typically suggested probabilities that tended toward the threshold of 0.5, while the top-bottom trained model calculated more extreme probabilities nearing either 1 or 0. The logistic regression performed the poorest with a ROC-AUC score of 0.576.

II. Saliency Imaging

Saliency mapping tends to provide reasonable justifications for abnormality cause. In the specific case of localizing hardware abnormalities, the saliency maps are especially successful. Typically, superpixels are concentrated at the site of irregularity. As the distance from the region of abnormality increases, the density of the superpixels decreases.

However, in certain cases the saliency maps provide more ambiguous results for cause of abnormality. In these instances, gray space or irrelevant anatomical regions are accented. In others, a dispersed set of superpixels appear without a notable area of concentration.

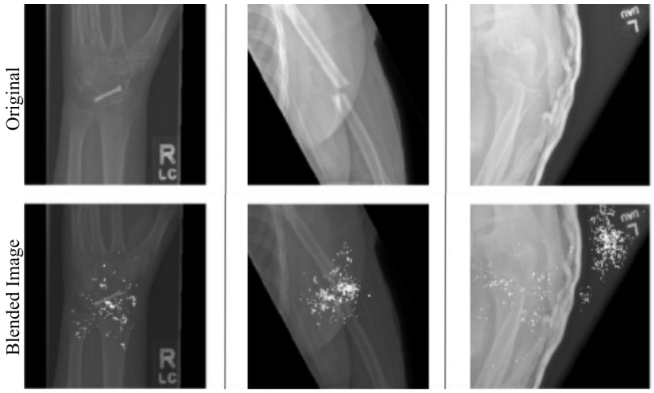

FIGURE 5: In Column A, the original image contains hardware in the wrist region which the blended saliency map accurately isolates. In Column B, the original image contains a fracture in the humerus which is also isolated by the saliency map. In Column C, the saliency map highlights arbitrary gray space in the radiograph, which is irrelevant to the task of abnormality detection.

The top-bottom model localizes abnormality cause most accurately. Isolating black space or irrelevant regions of the image are more common in the less comprehensively retrained models.

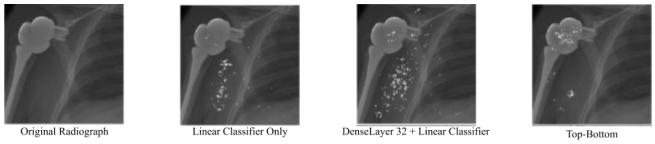

FIGURE 6: The top-bottom model effectively isolates the hardware abnormality. The other fine-tuned models highlight irrelevant portions of the image.

III. Clustering

The clustering algorithm does not separate radiographs based on the four primary abnormalities within MURA (as forenamed: fractures, hardware, degenerative joint disease, and a miscellaneous category with substantial amounts of lesions and subluxations). It tends to differentiate based on some properties irrelevant to diagnosis such as radiograph orientation and anatomical region of the radiograph. One diagnosis-based property that the clusters appear to make associations are the hardware abnormalities.

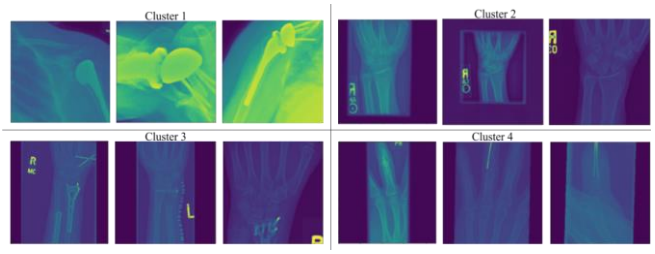

FIGURE 7: Upon categorizing a general set of abnormally classified images, notable patterns found include in Cluster 1) an extensive amount of shoulder socket hardware appears, in Cluster 2) wrist images without hardware abnormalities appear, in Cluster 3) a disproportionate number of wrist images with hardware plates appear, in Cluster 4) an example of clustering based on anatomical location is shown as a large majority of finger region images are observed in this cluster. The above images are samples of radiographs that appeared frequently in the specified clusters

As noted in Figure 7, the clustering of the general set of abnormally classified images appears to make substantial correlations based on hardware type and anatomical position. It is also notable that a disproportionate quantity of shoulder-socket hardware appears within the same cluster.

In the case of clustering within specific anatomical regions, there was an absence of abnormality-oriented clustering. Instead the algorithm's clustering entirely tends to center more around positional properties such as orientation and anatomical region. For example, in exclusive clustering of the forearm region, a separation of horizontally-oriented forearms versus vertically-oriented forearms appear.

DISCUSSION

The top-bottom model's performance exemplifies the benefits of implementing transfer learning. Pre-trained models can be fine-tuned with short training durations, in addition to attaining acceptable accuracy. The superior results of the top-bottom model in both classification accuracy and saliency imaging demonstrates that models that are more extensively fine-tuned perform better.

Saliency mapping provides insight into the decision-making process of the convolutional neural network. By accurately localizing abnormalities, the machine learning model justifies its reasoning, and its predictions can be accepted as well-informed. However, in cases where the saliency maps provide accents of irrelevant regions of the radiograph, the trustworthiness of the classification is questionable. Such discrepancies must be rectified. Architectural changes and the integration of local rule-extraction methods [4] in the convolutional neural network can combat such deviations.

The clustering model seems effective at distinguishing abnormalities caused by hardware presence. It can be inferred that such an association occurs specifically for hardware abnormalities than other types of abnormalities since hardware is especially semblant in radiographs. To

extend such differentiations to other types of abnormalities, changes to the architecture of the convolutional neural network can be explored, like with saliency mapping. Since the algorithm clustered based upon features extracted from the fine-tuned model, the algorithm's differentiations based upon orientation and anatomical region divulges the neural network extracts a substantial number of features based on these non-abnormality detection properties. To control for differences in anatomical region, the neural network can be trained separately for each anatomical region. By doing so, features that are more oriented toward abnormality detection rather than positional properties may be extracted. This would be far more useful than model visualizations of orientation-based properties, as such characteristics of radiographs do not require specialized medical training to find.

Interpretability holds monumental implications in real-world settings [15]. In the context of regulation, interpretability can serve as a qualitative metric for model evaluations. Auditing models on the 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 [3].

CONSIDERATIONS FOR FUTURE INVESTIGATIONS

To improve existing performance in saliency imaging and clustering, architectural changes to the convolutional neural network can be explored. An integration of local rule-based explanations, which utilize genetic algorithms and derivation counterfactual rules, can also prove beneficial [4]. Influence functions, statistical tools used to evaluate the influence of removing an observation, are also relevant to the study of interpretable models [6].

Linguistic explanations for diagnostic choices can offer further clarity into the model's decision-making process [15]. 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 [15].

CONCLUSION

This work proposes a method to construct interpretable models for diagnosing abnormality in musculoskeletal radiographs. We implement transfer learning with the DenseNet-169 model to detect irregularities within medical slides. Interpretability is then explored through two distinct lenses: saliency imaging and clustering. Saliency imaging provides reasonable visual justifications for the final model classification. K-means clustering succeeds in differentiating properties such as anatomical region, radiograph orientation, and hardware type. 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.

Interpretability holds monumental implications in real-world settings [11]. In the context of regulation, interpretability can serve as a qualitative metric for model evaluations. Auditing models on the 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 [3].

ACKNOWLEDGEMENTS

The author would like to thank the Summer STEM Institute staff and mentor Allison Tam for their continual guidance during this research.

REFERENCES

[1] Jocelyn Barker, Assaf Hoogi, Adrien Depeursinge, and Daniel Rubin. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. Medical Image Analysis, 30,12 2015.

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

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

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

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

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

[7] Subramani Mani, Yukun Chen, Lori R. Arlinghaus, Xia Li, A. Bapsi Chakravarthy, Sandeep R. Bhava, E. Brian Welch, Mia A. Levy, and Thomas E. Yankeelov. Early pre-diction response of breast tumors to neoadjuvant chemotherapy using quantitative mri and machine learning. AMIA ... Annual Symposium proceedings. AMIA Symposium,2011, 868–877., 2011.

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

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary De-Vito, Zeming Lin,

Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-sos, D. Cournapeau, M. Burcher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research,12:2825–2830, 2011.11

[11] Jasmine Pennic. Ai vs. humans: Ai solution beats Stanford radiologists in chest x-ray di-agnostics competition. 2019.

[12] Matti Pietik¨ainen. Local binary patterns. 2010.

[13] 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: To-wards radiologist-level abnormality detection in musculoskeletal radiographs. CoRR, abs /1712.06957, 2017.

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

[15] Mauricio Reyes, Raphael Meier, S`ergio 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.

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

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

[18] 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, Apr2020.

[19] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In 2018 IEEE Conference on Com-puter Vision and Pattern Recognition, CVPR2018, Salt Lake City, UT, USA, June 18-22,2018, pages 8827–8836. IEEE Computer So-ciety, 2018