Diving Deep into Bone Anomalies on the FracAtlas Dataset using Deep Learning and Explainable AI

Filza Akhlaq[1], Subhan Ali[1], Ali Shariq Imran[1], Sher Muhammad Daudpota[2] and Zenun Kastrati[3]
[1]Dept. of Computer Science, Norwegian University of Science & Techology (NTNU), 2815 Gjøvik, Norway
[2]Dept. of Computer Science, Sukkur IBA University, Nisar Ahmed Siddiqui Rd, Sukkur, Sindh, Pakistan
[3]Dept. of Informatics, Linnaeus University, 351 95, Växjö, Sweden

2024 International Conference on Engineering & Computing Technologies (ICECT) | 979-8-3503-4971-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECT61618.2024.10581288

***Abstract*—Medical image analysis has undergone significant advancements with the integration of machine learning techniques, particularly in the realm of bone anomaly detection. The availability of recent datasets and the lack of benchmarking and explainability components provide numerous opportunities in this domain. This study proposes a benchmarking approach to a recently published FracAtlas dataset utilizing state-of-the-art deep-learning models coupled with explainable artificial intelligence (XAI) having two distinct modules. The first module involves the binary classification of fractures in different body parts and explains the decision-making process of the best-performing model using an XAI technique known as EigenCAM. EigenCAM generates heatmaps on every layer of the YOLOv8m model to explain how the model reached a conclusion and localizes the fracture using a heatmap. To verify the heatmap, we also detected fractures using the YOLOv8m detection model, which achieved a mAP@0.5 of 59.5%, outperforming the baseline results on this dataset. The second module involves a multi-class classification task to categorize images into one of the five anatomical regions. The best-performing model for binary classification is the YOLOv8m model, with an accuracy of 83.1%, whereas the best-performing model for multi-class classification is the YOLOv8s, achieving an accuracy of 96.2%.**

***Index Terms*—Fracture classification and detection, medical imaging, X-rays, Explainable AI, Deep learning**

I. Introduction

In recent years, the field of medical image analysis has observed notable progress, fueled by the intersection of machine learning and healthcare. Machine learning has various applications in healthcare, including accurately and quickly identifying bone anomalies using medical imaging datasets [1]. The increasing availability of high-resolution medical images results in the need for sophisticated classification models that can not only differentiate between fractured and non-fractured cases but also provide insights into the distinctive features of fractures.

This article presents a unique approach to bone fracture detection for the FracAtlas dataset [2], implementing a two-module image analysis pipeline. The first module involves a binary classification, where the algorithm distinguishes between fractured and non-fractured cases. The second module probes further into regional classification, differentiating fractures across five individual anatomical regions. The proposed approach refines the diagnostic process as well as allows for a more refined understanding of the fracture patterns, aiding in personalized treatment strategies.

In this study, we initially trained various state-of-the-art (SoTA) deep learning models for binary classification and selected the top-performing model. To comprehend how the model learns to classify an image as a fracture, we employed explainable Artificial Intelligence (XAI) on each layer of the best-performing model. XAI also facilitates the localization of fractures through generated heatmaps. To compare this, we additionally trained a model to detect bounding boxes of fractures within the dataset. For multi-class classification, we trained multiple SoTA models to classify images into five anatomical regions.

Although deep learning models excel in classification problems, they are black boxes and lack interpretability. Interpretability, explainability, and model trust are crucial in fields such as medicine, where human lives are at stake. This has led to the development of XAI. We have implemented EigenCAM to generate heatmaps on every convolutional layer, allowing us to understand how the model predicts an image as a fracture.

This study contributes to the growing body of medical image analysis literature and addresses the practical challenges associated with bone fracture diagnosis. Integrating a two-module pipeline and XAI techniques represents a promising step toward improving the reliability and interpretability of automated fracture classification systems. As we delve into the details of the methodology and experimental results, the potential impact of this approach on clinical decision-making becomes evident, providing a valuable tool for healthcare professionals in their pursuit of precise and well-informed fracture diagnoses. By bridging the gap between artificial intelligence and clinical practice, we aim to facilitate healthcare professionals with robust tools for accurate and interpretable fracture diagnosis, ultimately enhancing patient care and outcomes.

The following are the main research objectives of this study.

- To analyze the performance of deep learning models on fracture classification for multi-regional anatomical X-ray images.
- To visualize the working and insight of the models that perform the best using Explainable AI.
- To examine the performance of YOLOv8 detection models for fractures in different human body regions.

This work marks the first of its kind in the classification and localization of fractures throughout the entire body using XAI and detection models. Furthermore, we use SoTA models for multi-class classification to identify body parts within the

979-8-3503-4971-9/24/$31.00 ©2024 IEEE

FractAtlas dataset, thereby establishing a benchmark for future studies in this area. The rest of the paper is structured as follows. Section II presents the literature review. Section III discusses the methodology and explains in detail the two-module approach that was opted for this study. Section IV presents the dataset, followed by findings and their interpretation in Section V. Finally, section VI presents a conclusion and some future directions.

II. Literature Review

Recent advancements in deep learning and computer vision have streamlined the automated identification of fractures, making the process more efficient and accurate. In [3], the authors proposed a deep learning-based pipeline for detecting wrist fractures named DeepWrist. They used the SeresNet50 [4] model, which is pre-trained on ImageNet [5] for detecting fractures. In their work, the authors created their own three datasets from the Oulu University Hospital's (OUH) Picture Archiving and Communication System (PACS) and the Radiology Information System. The training dataset consisted of 1000 cases with distal radius fractures and 1000 wrist cases without fractures. They also removed images that had artifacts, resulting in 953 studies having fractures; in total, they used 1953 wrist studies for training. They annotated the images by themselves using the patients' radiology reports. To make results interpretable of the detection model, they generated the heatmaps using GradCAM [6] on the part of the radiograph highlighting the fractured area.

Lysdahlgaard [7] utilized heat maps as an XAI technique to detect abnormalities in wrist and elbow X-ray images. The author trained twenty transfer learning models such as VGG16, VGG19 [8], ResNet [9], DenseNet [10], InceptionV3 [11], Xception [12], etc., coupled with GradCAM XAI technique for generating heatmaps. These baseline models were trained on MURA-dataset [13] and evaluated the model's efficiency in recognizing regions of interest using the Dice Similarity Coefficient (DSC). The best model on this dataset was VGG16, with a test accuracy of 84%. Thian et al. [14] utilized the Faster Region-based Convolutional Network method (Fast R-CNN) [15] to detect fractures. They trained the Faster R-CNN on 7356 wrist radiographic studies with 14614 images.

In synthesizing the existing literature, our study identifies gaps and opportunities for further research in the realm of bone fracture classification. The proposed two-module model, coupled with XAI techniques, aims to address these gaps and contribute to the ongoing efforts to improve automated fracture diagnosis systems' accuracy, interpretability, and clinical relevance.

III. Methodology

In this study, we are proposing an approach comprising two modules, as illustrated in Figure 1. In the first module, we perform binary classification and localization along with XAI techniques on the FracAtlas dataset. The second module conducts a multi-class classification to classify the images into anatomical regions (Hand, Hip, Shoulder, Leg, and Mixed).

A. Binary Classification:

In this stage, we craft a binary classifier to classify the images as "Fractured" or "Non-fractured. The steps involved in the binary classification task are discussed below:

1) Data Preprocessing and Feature Engineering: In this stage, the following steps are performed on the dataset to ensure unbiased and efficient training.
   - Image Resizing: Since all the images in this dataset had different dimensions, we had to resize all the images to the same dimension. In this stage, we resized the images to 244×244 as this size is most commonly used in X-ray images.
   - Down-sampling: As this dataset is highly imbalanced, the non-fractured instances had to be down-sampled equal to the number of the fractured instances (716).
2) Model Training : We performed a set of experiments on the dataset utilizing existing state-of-the-art deep learning models, including recently introduced YOLO variants. The data split used for these models was 70 and 30 for training and testing, respectively. In total, nineteen deep-learning models and five YOLO models are trained and tested in this stage. For the deep learning models, a final layer with one node was added for binary classification. The YOLOv8m model outperformed all other models. Then, we applied EigenCAM to generate heatmaps on every layer to see how the model reached the prediction. We also used heatmaps to localize the fractured region in the affected images.
3) Localization: For the localization of fractures, two sets of data were used. First, the models were trained and validated on the replication split used by the authors to report their results (Table VIII) (Set 1), where the authors only trained the model on Fractured instances. The second approach is feeding the model with both fractured and non-fractured instances (Set 2). For Set 1, the results are reported in Table I on the validation split for the sake of comparing to the results reported in [2]. The results of the validation split of Set 2 are reported in Table III.

TABLE I: Localization results using various settings in models for Set 1.

Models/Metrics	Precision	Recall	mAP@0.5
YOLO8s	0.575	0.462	0.496
YOLO8s 100e imgsize=600	0.695	0.477	0.548
YOLO8s 73e imgsize=640	0.709	0.535	0.562
YOLO8x 100e imgsize=640	0.729	0.503	0.575
YOLO8x 225e imgsize=640	0.721	0.560	0.595

The best-performing model, as presented in Table I, is then applied to the four categories to determine the root cause of such performance metrics. These results are reported in Table II, the Hand class tops in the mAP@0.5, followed by Leg, Hip, and then Shoulder. This difference is due to the fact that the number of instances of a class is directly affecting its performance.

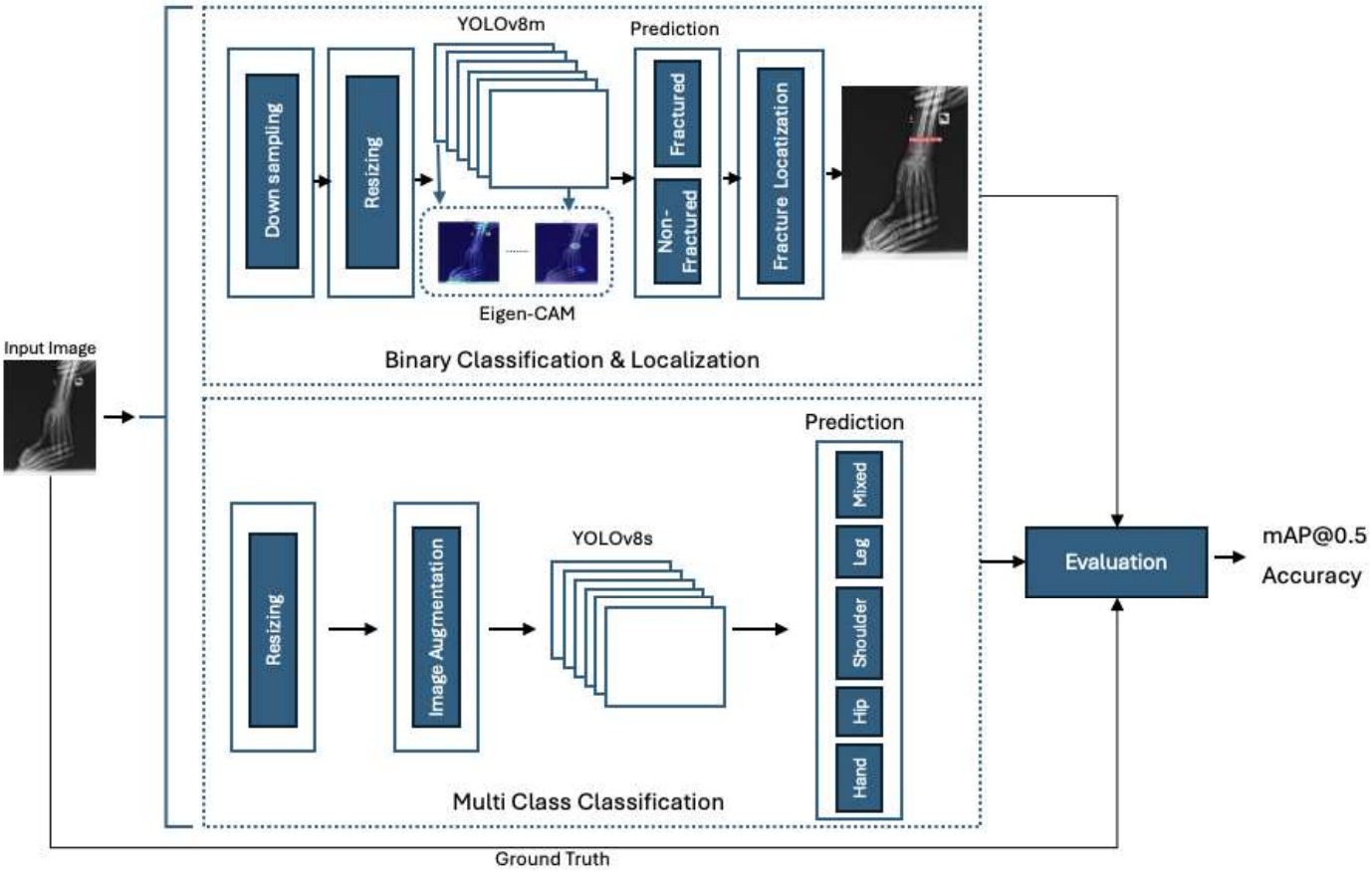

Fig. 1: A high-level architecture of the proposed approach.

TABLE II: Localization results of the best model on the four categories.

Categories (number of samples)	Precision	Recall	mAP@0.5
Hip (10)	0.758	0.273	0.316
Hand (49)	**0.883**	**0.589**	**0.665**
Shoulder (5)	0.141	0.200	0.083
Leg (34)	0.628	0.444	0.446

TABLE III: Localization results of various models on Set 2.

Models	Precision	Recall	mAP@0.5
YOLOv8s	0.526	0.785	0.502
YOLOv8m	0.515	0.808	0.538
YOLOv8l	0.507	0.804	0.513
YOLOv8x	0.514	0.770	0.477

B. Multiclass Regional Classification:

In this stage, the classifier aims to classify between the five regions of the X-ray images. The steps involved in the binary classification task are discussed below:

1) Data Pre-processing: To train the classifiers for this stage, the images are divided into five classes, i.e., Hand, Hip, Leg, Shoulder, and Mixed. Furthermore, the following steps were performed:
   - Image Resizing: All of the images are resized to a single size of (224, 224) for the deep learning models, and for YOLO models, the size of (640, 640) is used.
   - Image Augmentation: As mentioned earlier, the dataset is imbalanced in terms of regions; for example, Hand and Leg have many more instances than the other three regions. The images needed to be augmented to assure an unbiased training of the data. A total of 300 images from each class are chosen to be used. As the Shoulder and Hip classes only had 179 and 101 images respectively, these two classes needed augmentation. For augmentation, a rotation range of 40, a width and height shift range of 0.2, a shear and zoom range of 0.2, and a horizontal flip are applied to the training images randomly.
2) Model Training for Final Classification: For training a multi-class classifier, we also conducted the same type of experiments and trained multiple SoTA deep learning models for classifying images into five anatomical regions. Out of all models, the YOLOv8s model outperforms all and sets a benchmark accuracy of 95.2%. The hyperparameters used for YOLOv8s are listed in Table IV.

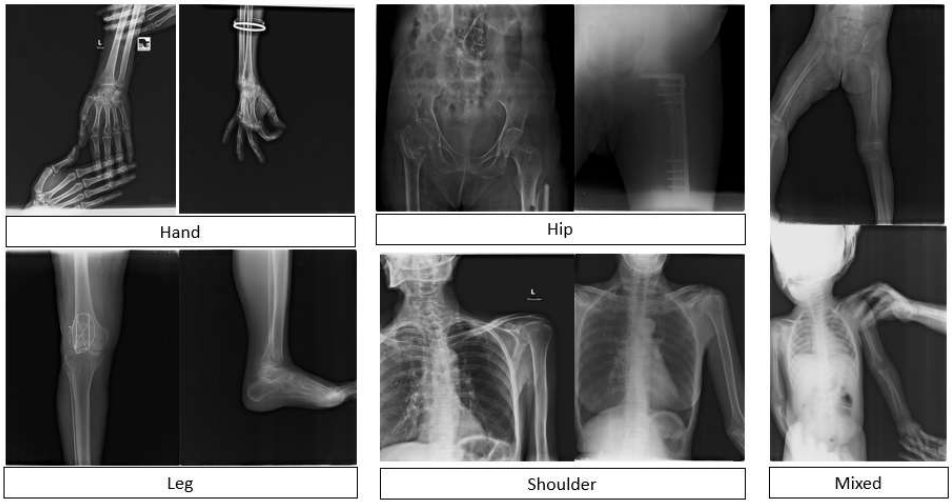

Fig. 2: Some instances from the dataset belonging to five region categories.

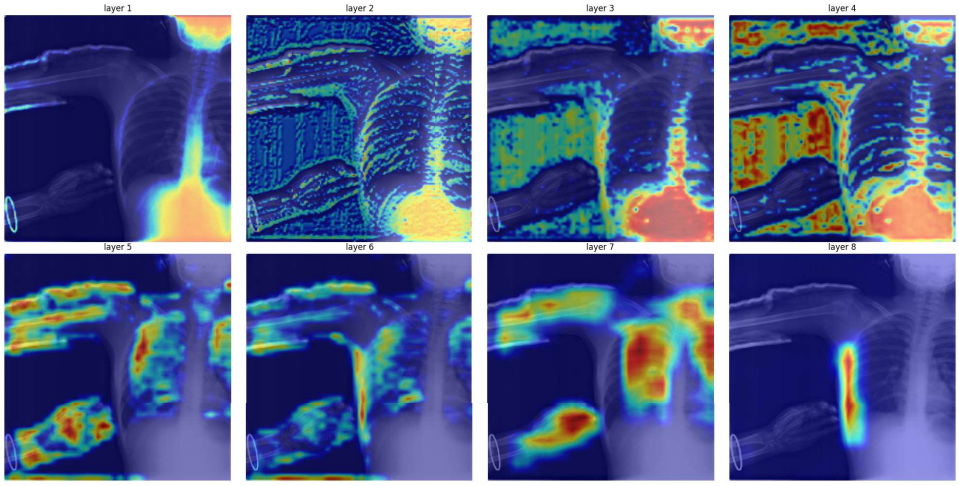

Fig. 3: The activation maps visualizing each layer of the YOLOv8m model for Binary Classifier on a fractured image from the Mixed class, produced using EigenCAM. Starting from layers 1 up to 8, EigenCAM visualizes what the YOLO model perceives in each layer.

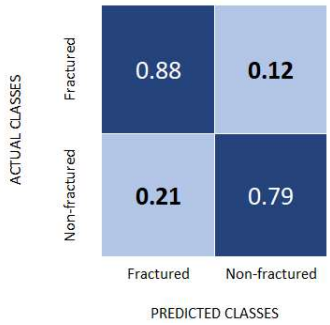

Fig. 4: The confusion matrix for the model chosen for Binary Classification, YOLOv8s.

TABLE IV: Training hyperparameters for the best models.

Parameters	Values for Stage I	Values for Stage II
Optimizer	AdamW	AdamW
Learning Rate	0.000714	0.000714
Image Size	640	640
Callback Epoch	53	143

IV. Dataset

The FracAtlas dataset [2], released in 2023, is a specialized collection for fracture classification, localization, and segmentation in musculoskeletal radiographs. It comprises 4,083 X-ray images, with a total of 922 fractures identified across the 717 fractured images. The dataset focuses on the hand, hip, leg, and shoulder regions and features a diverse age range of subjects, from 8 months to 78 years old, with a gender distribution of about 62% male and 38% female. The development of FracAtlas involved data collection from three hospitals, namely Lab-Aid Medical Center, Brahmanbaria, Anupam General Hospital and Diagnostic Center, Bogra and Prime Diagnostic Center, Barishal. Expert radiologists and an orthopedic surgeon performed data cleaning and meticulous

Fig. 5: The Confusion matrix for YOLOv8s in Stage 2.

TABLE V: Classification results for binary classifiers.

Models/Metrics	A	P	R	F1-score
Xception	0.60	0.58	0.67	0.61
VGG16	0.75	0.75	0.77	0.75
VGG19	0.73	0.69	0.81	0.74
ResNet50	0.75	0.73	0.78	0.75
ResNet50V2	0.48	0.23	0.02	0.04
ResNet101	0.73	0.71	0.79	0.74
ResNet101V2	0.56	0.55	0.70	0.61
ResNet152V2	0.55	0.59	0.30	0.39
InceptionV3	0.51	0.52	0.25	0.33
InceptionResNetV2	0.50	0.51	1.00	0.67
MobileNetV2	0.66	0.63	0.81	0.70
DenseNet121	0.64	0.61	0.78	0.68
DenseNet169	0.63	0.60	0.69	0.64
DenseNet201	0.67	0.62	0.83	0.71
NASNetMobile	0.59	0.59	0.72	0.64
NASNetLarge	0.57	0.58	0.47	0.51
EfficientNetB7	0.71	0.71	0.66	0.68
EfficientNetV2B3	0.68	0.63	0.87	0.72
EfficientNetV2L	0.72	0.71	0.77	0.73
YOLOv8n	0.80	0.74	0.85	0.79
YOLOv8s	0.82	0.79	0.84	0.81
YOLOv8m	**0.83**	**0.87**	**0.81**	**0.83**
YOLOv8l	0.81	0.88	0.78	0.82
YOLOv8x	0.76	0.79	0.74	0.76

annotation. The labeling process was thorough, ensuring the accuracy and reliability of the annotations. The only experiments performed on this dataset by the authors were to validate the credibility of the dataset. The results of their experiments on the validation split of the dataset are shown in Table VIII, where the models were only provided with "Fractured" instances. Figure 2 shows some of the dataset instances. The breakdown of this dataset into these regions based on fractured and non-fractured instances is mentioned in Table VII.

TABLE VI: The class-wise accuracy of YOLOv8s.

Region	**Accuracy**
Hand	0.95
Hip	0.94
Leg	0.97
Mixed	0.86
Shoulder	0.97

TABLE VII: The samples across five region categories.

Region	Fractured	Non-Fractured
Hand	379	902
Leg	212	1912
Mixed	105	293
Hip	10	169
Shoulder	10	91
Total	716	3367

TABLE VIII: The results of the experiments performed on the FracAtlas dataset in the baseline study.

Task	Model	Type	P	R	mAP@0.5
Localization	YOLO8s	Box	0.807	0.473	0.562
Segmentation	YOLO8s-seg	Box	0.718	0.607	0.627
		Mask	0.830	0.499	0.589

V. Results

A. Results of Binary Classification and Localization

Our exploration of multiple models provides a comprehensive understanding of their performance in our study, including several deep learning models as depicted in Table V, where A, P, and R stand for Accuracy, Precision, and Recall, respectively. It can be seen that the YOLOv8m model outperformed all the other models with an accuracy of 83%; hence, it was chosen as the binary classifier for our final model. Figure 4 shows the confusion matrix for this best model.

The best-performing model, YOLOv8m, is then fed to EigenCAM to visualize each layer of the model and analyze how the model classifies based on the visualization of each layer. Each layer's visualization is illustrated in Figure 3.

To compare the localization achieved through EigenCAM, we performed localization through different variants of the YOLO object detection model. The YOLOv8x model outperformed the baseline results set on this dataset by achieving mAP@0.5 of 59.5%. The comparison of these performances can be seen in Figure 6.

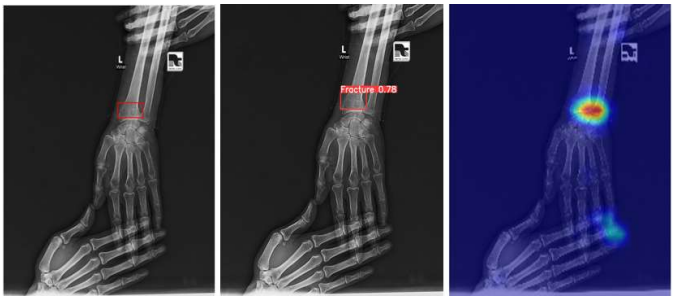

Fig. 6: This figure shows (from left to right) the ground truth provided with the dataset, the performance of our detection model, and the heatmap generated by EigenCAM on a fractured instance.

B. Results of Multi-class Classification

The performance of all the models considered for this stage is shown in Table IX, where A, P, and R stand for accuracy, precision, and recall, respectively. All these metrics are macro averages of the results of the five classes. After analyzing the results from all of the applied models for this stage, the YOLOv8s has outperformed all the other models, yielding an accuracy of 96.2%.

TABLE IX: Classification results for Multi-class classifiers.

Models/Metrics	A	P	R	F1-score
Xception	0.77	0.77	0.77	0.77
VGG16	0.94	0.94	0.94	0.94
VGG19	0.93	0.93	0.93	0.93
ResNet50	0.95	0.95	0.95	0.95
ResNet50V2	0.39	0.39	0.39	0.28
ResNet101	0.95	0.95	0.95	0.95
ResNet101V2	0.36	0.52	0.36	0.24
ResNet152V2	0.20	0.24	0.20	0.07
InceptionV3	0.71	0.71	0.71	0.71
InceptionResNetV2	0.20	0.04	0.20	0.07
MobileNetV2	0.84	0.84	0.84	0.84
DenseNet121	0.80	0.80	0.80	0.79
DenseNet169	0.79	0.80	0.79	0.79
DenseNet201	0.81	0.81	0.81	0.81
NASNetMobile	0.66	0.66	0.66	0.66
NASNetLarge	0.77	0.77	0.77	0.77
EfficientNetB7	0.95	0.95	0.95	0.95
EfficientNetV2B3	0.94	0.94	0.94	0.94
EfficientNetV2L	0.95	0.95	0.95	0.95
YOLOv8n	0.91	0.91	0.91	0.91
YOLOv8s	**0.96**	**0.96**	**0.96**	**0.96**
YOLOv8m	0.95	0.95	0.95	0.95
YOLOv8l	0.95	0.95	0.95	0.95
YOLOv8x	0.95	0.95	0.95	0.95

VI. Conclusion and Future Work

In this work, we conducted a binary and multi-class classification on the FracAtlas dataset for detecting fractures in different parts of the body. We trained different SoTA deep learning models for binary and multi-class classification and found the best-performing models based on classification accuracy. For binary classification, we also employed XAI on the best-performing model, i.e., YOLOv8m, to visualize how the model is reaching a prediction, which also resulted in localizing the fractures through heatmaps. At the classification layer of the model, the fractures are visualized.

To compare the visualizations provided by EigenCAM, we also trained the YOLOs object detection models for detecting bounding boxes of fractures. YOLOv8x object detection model also outperformed the previous benchmark for localization of fractures on this dataset with a mAP@0.5 of 59.5%. For binary and multi-class classification, we set a new benchmark accuracy of 83% and 96.2%, respectively.

In the future, Vision Transformers (ViT) can be applied for classification tasks on this dataset to achieve better results. Moreover, for the image augmentation techniques, future studies can experiment with some algorithms that learn from the data itself for performing augmentation. Additionally, we have used YOLOv8, a single-stage object detection model for fracture detection. In the future, two-stage fracture detection models such as Faster-RCNN, SSD, and SoTA transformers-based object detection models can be utilized for better fracture detection results. Finally, for the explainability of the performances of black-box models, more XAI techniques can be experimented with, such as Gradient-weighted Class Activation Mapping (Grad-CAM) based models, which are specifically catered to perform on medical images.

VII. Acknowledgement

This work was supported by the Curricula Development and Capacity Building in Applied Computer Science for Pakistani Higher Education Institutions (CONNECT) project, No. NORPART-2021/10502, funded by Diku.

References

[1] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review," *Computers in Biology and Medicine*, p. 107555, 2023.

[2] I. Abedeen, M. A. Rahman, F. Z. Prottyasha, T. Ahmed, T. M. Chowdhury, and S. Shatabda, "Fracatlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs," *Scientific Data*, vol. 10, no. 1, p. 521, 2023.

[3] A. M. Raisuddin, E. Vaattovaara, M. Nevalainen, M. Nikki, E. Järvenpää, K. Makkonen, P. Pinola, T. Palsio, A. Niemensivu, O. Tervonen *et al.*, "Critical evaluation of deep neural networks for wrist fracture detection," *Scientific reports*, vol. 11, no. 1, p. 6006, 2021.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[7] S. Lysdahlgaard, "Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs," *Radiography*, vol. 29, no. 6, pp. 1132–1138, 2023.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[13] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.

[14] Y. L. Thian, Y. Li, P. Jagmohan, D. Sia, V. E. Y. Chan, and R. T. Tan, "Convolutional neural networks for automated fracture detection and localization on wrist radiographs," *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180001, 2019.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.