International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Smart Image Captioning with X-Ray Bone Healing Detection

1st **Prof. Bheerappa Sasanoor**
Department of Computer Science(AI&ML)
KLS Vishwanathrao Deshpande
Institute of Technology Haliyal, India
bys@klsvdit.edu.in

2nd **Miss. Anusha K Madalli**
Department of Computer Science(AI&ML)
KLS Vishwanathrao Deshpande
Institute of Technology Haliyal, India
anushamadalli5@gmail.com

3rd Miss. **Ankita B Gawada**
Department of Computer Science(AI&ML)
KLS Vishwanathrao Deshpande
Institute of Technology Haliyal, India
ankitagawada4@gmail.com

4th Miss.**Shrusti R Kadakol**
Department of Computer Science(AI&ML)
KLS Vishwanathrao Deshpande
Institute of Technology Haliyal, India
shrustikadakol@gmail.com

5th Miss **Seema R. Kalal**
Department of Computer Science(AI&ML)
KLS Vishwanathrao Deshpande
Institute of Technology Haliyal, India
seemarkalal46@gmail.com

Abstarct-Smart Image Captioning with X-Ray Bone Healing Detection The present work introduces an intelligent system that integrates deep learning-based image captioning with medical X-ray analysis for automated bone fracture and healing detection. The system proposed in this work unifies visual understanding and diagnostic reasoning to produce descriptive captions of the medical images as well as the detection of fracture types, localization of affected areas, and tracking of healing progress. This model leverages convolutional neural networks (CNNs), attention-based captioning models (BLIP), object detection (YOLO), and explainability visualization (Grad-CAM) for an end-to-end interpretation of the X-ray image. The output is optimized with multilingual text-to-speech (TTS) for better accessibility by medical professionals and patients.

Keywords : Image Captioning, X-Ray Analysis, Bone Fracture Detection, Deep Learning, Medical Imaging, YOLO, BLIP, CLIP, Grad-CAM, Streamlit, Explainable AI

1. INTRODUCTION

Orthopedic experts and radiologists are crucial to interpreting X-ray images in diagnosing and following up on bone disorders and injuries. Due to the significant increase in medical image data, manual interpretation is now slow and prone to human error. Excessive workloads and the repetitive nature of image examination usually result in diagnostic delays and fatigue, prompting the necessity for smart systems capable of supporting radiologists in providing faster and more accurate results.

The advances of recent years in artificial intelligence (AI) and deep learning (DL) have revolutionized medical image analysis. Deep learning models such as CNNs and Vision Transformers (ViTs) have shown outstanding ability in identifying patterns and anomalies in medical scans, while image captioning models such as BLIP and CLIP allow machines to describe visual content using natural language. By combining all these technologies, AI systems are able to not only identify fractures, but also explain their observations in descriptive captions, so the results are more interpretable and understandable for both doctors and patients.

The Smart Image Captioning with X-Ray Bone Healing Detection architecture proposed here combines automated captioning, fracture diagnosis, and bone healing tracking into a combined system. It seeks to offer clear visual explanations, elaborate natural language reports, and follow-up comparison facilities to evaluate recovery progress. This convergence maximizes diagnostic efficiency, accuracy, and transparency, and represents an important step towards intelligent, explainable, and patient-oriented medical imaging systems enabling improved clinical decision-making and enhanced healthcare outcomes.

2. LITERATURE SURVEY

Medical imaging has never been a part of diagnostic medicine, and researchers have over the years tried many ways of automating image interpretation in order to achieve better results faster. Conventional computer-aided diagnostic (CAD) systems relied on statistical classifiers and feature extraction algorithms, which had the drawbacks of limited generalization and handcrafted features. The development of deep learning, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has significantly improved the ability to automatically identify, categorize, and segment medical images. These advances have made systems capable of identifying complex visual patterns like bone fractures and provide more accurate predictions than previous methods.

Recent advances in vision-language models have created new opportunities for uniting image understanding and natural language generation. Models like BLIP (Bootstrapped Language-Image Pretraining), CLIP (Contrastive Language-Image Pretraining), and MedCLIP bring together visual and textual modalities in a way that enables machines to produce coherent descriptions of image content. Such models have been found to be helpful in medical applications for producing radiology image captions, improving interpretability, and supporting

diagnostic choices. For example, the BLIP model has shown remarkable performance in visual captioning.

through the alignment of vision and language representations, and MedCLIP applies similar ideas to medical datasets such as chest X-rays and CT scans.

Some scientists have developed smart medical image interpretation systems:

1. Rajpurkar et al. (2017) proposed CheXNet, a deep learning algorithm based on DenseNet architecture for pneumonia detection from chest X-rays that became the standard for automated radiology analysis.

2. Boecking et al. (2022) created VisualCheXbert, a vision-language model that can generate textual radiology reports by utilizing image-caption alignment methods, enhancing the quality and interpretability of the report.

3. Li et al. (2022) introduced an AI-based bone fracture diagnosis system with the assistance of CNNs, exhibiting excellent diagnostic performance and superior performance compared to traditional feature-based approaches.

4. Wang et al. (2023) investigated applying Grad-CAM visual explanations to medical AI systems to increase transparency, allowing clinicians to see visual proof of the regions of interest for the model during diagnosis.

5. Singh et al. (2024) introduced a framework that incorporated X-ray analysis along with recovery tracking to facilitate longitudinal comparison of bone healing across various timeframes.

In the light of these studies, it is comprehended that AI-based image captioning combined with medical image diagnosis facilitates not only interpretability but also opens pathways to automated, patient-centric, and explainable healthcare systems. Elaborating further on the subject, the system under consideration, Smart Image Captioning with X-Ray Bone Healing Detection, integrates image captioning, fracture identification, and monitoring of healing advancement within a single framework. This holistic approach adopts both vision and language understanding to provide descriptive, interpretable, and accurate diagnostic recommendations, which overcome the shortcomings of previous systems that were based on classification alone.

3. METHODOLOGY

The Smart Image Captioning Device for X-Ray Bone Healing Detection combines computer vision, deep learning, and natural language processing technologies to automatically analyse and create captions for X-Ray images. The multiple phases of processing involve, among others: The capture of the digital X-Ray images for analysis; the processing and creating human readable caption; the determination of the presence of a bone fracture; the measurement of healing over a period of time; and the visualisation of how well there is explainability.

An overall goal of the system is to create an end-to-end smart system that produces a natural language caption for the X-Ray images and automatically detects a fracture and monitors the patient's developing stages of healing.

3.1 Image Acquisition and Preprocessing

Hospitals send their images directly into the databases or upload them to the system via the user interface. Before the model will train on the images we need them all to have the same characteristics, therefore we preprocess the images by performing Resizing, Grayscale normalisation, and applying Gaussian Filter for noise elimination, to ensure model accuracy with regards to these aspects of the image. We also perform various Data Augmentation techniques (Rotating, Flipping, and/or profile/Contrast adjustment) during the training of the model, in order to create Robust Models against the variability between the orientation of the images and the variations in the Light conditions. Finally, we Normalise the images that have been pre-processed, to a Standard Resolution (i.e.224x224 Pixels) for the Input to the Deep Learning models.

3.2 Image Captioning Module

Using a BLIP (Bootstrapped Language-Image Pretraining) architecture, the captioning module for images combines visual (image) and textual (language) embeddings to generate a natural language representation of the content of an inputted image. Image features are extracted from input images using a vision encoder (a Vision Transformer or CNN, generally), then these image features are used to generate captions via a language decoder. In the medical field, a captioning model will be fine-tuned to medical imaging datasets containing x-ray images with annotated text. Examples of captions generated include, “X-ray reveals a broken right radius bone” and “Healed bone with callus formation visible.” The module generates captions that can provide clinically relevant information to both patients and doctors.

3.3 Fracture Localization and Detection

To identify fractures accurately, a convolutional neural network-based classifier is employed, which assesses the bone structure and detects irregularity or deviation from

normal bone morphology. The classifier is trained using training data that has been labeled to include multiple fracture classification types (simple, compound, and comminuted) and non-fractured examples of the same anatomical area. When predicting the fracture class associated with a new X-ray image, the model also calculates a confidence score for each classification based on the input image. Additionally, Grad-CAM methodology is applied to provide a visual representation of which areas of the X-ray were most influential in the model's classification decision. This visual representation can provide support for the medical professional in determining if the AI has correctly identified the area of interest and whether it concurs with their own diagnosis.

3.4 Healing Detection and Progress Tracking

One particularly interesting feature of the anticipated computer system is a module specifically designed to detect from X-ray images whether a bone fracture has healed successfully or not. This module will do this by monitoring the integrity of an X-ray image against its previous X-ray image and detecting changes from the structural comparison of the images using heat mapping and layering techniques. As a result, the clinician can assess the degree to which the bone has healed and/or aligned post-fracture. In addition, the ability to track changes over time will enable a clinician to continue to monitor the patient long after his/her treatment has been completed.

3.5 Explainability and User Interface

To foster user trust and provide an easy-to-use experience with the new system, an Explainable AI (XAI) engine, and a Streamlit-based graphical interface (GUI) have been incorporated. Users can upload X-ray images, view captions that explain the results produced by the XAI engine, view predicted types of fractures, percentage confidence levels of accuracy for each predicted type, and view visual explanations provided by the heat map. Side-by-side comparisons are displayed with preceding and subsequent images to show healing progression. The system will also create a full diagnostic summary with potential opportunities for printing as a Portable Document Format (PDF) for clinical documentation and inclusion in patient records.

3.6 Summary of System Workflow

•Initial Stage: X-ray image uploading or acquiring.

•Laid-out process: Prepare - Eliminate noise; Normalize.

•Creating Statement: Use BLIP to generate text to Describe the image Processed.

•Identify: Categorizing the fracture and visualizing it with the aid of CNN and Grad-CAM.

•How well has it healed: Assessment of healing as comparisons made between repeated scans.

•Final Stage: Output - Statement of diagnosis; Confidence Score; Heat Map; Report Generation Options (if applicable).

The methodology blends the use of image captioning, deep learning automated detection systems, and visualisation techniques to not only provide accurate diagnostics but also ensure that the system has the ability to interpret and understand what the user sees. Thus, the methodology provides a full AI-assisted diagnostic system that will ensure the user is able to effectively and reliably analyse and diagnose medical images in a manner that is quick, easy, and reliable.

Proposed Method:

The input to the system is an X-ray image that will be pre-processed, and using YOLO/CNN and Grad-CAM, it will detect the fractures and indicate where the fractures are using a heatmap. Then by means of the BLIP system, the X-ray image will be converted into text and spoken to the user in multiple languages. Finally, the system will provide the user with a diagnosis report and healing tracker via a Streamlit interface

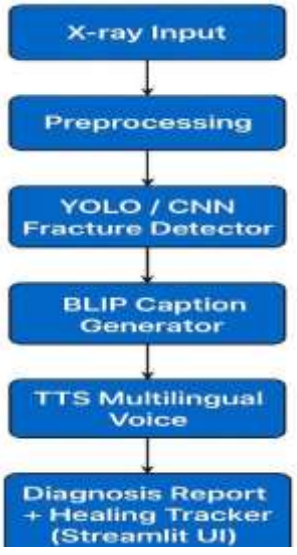

Fig. 1. Block diagram

This modern camera interface displays a live view of The Taj Mahal, a well-known historical monument located in India today (present day October 2021). The Taj Mahal is a mausoleum that was built in 1653 (completed) by Mughal Emperor Shah Jahan to honor his wife for her devotion and loyalty throughout their marriage (Mumtaz Mahal). The Taj Mahal is an architectural beauty regarded by many as one of the best pieces of architecture in the world. The Taj Mahal represents a symbol of eternal love between two people. The Taj Mahal is constructed

primarily with white marble; it combines elements of Persian style, Islamic style and Indian style architecture. In 1983, UNESCO officially designated it as a World Heritage Site. In 2007, it was announced as one of the New Seven Wonders of the World. This camera interface illustrates how modern technologies can utilize Artificial Intelligence (A.I) and Computer Vision (CV) to capture images live and identify historical monuments in real time. This interface shows how the technology can combine beauty, inspiration and cultural heritage along with digital tools to create a more engaging experience in studying, preserving and interacting with ancient history. [Fig.2]

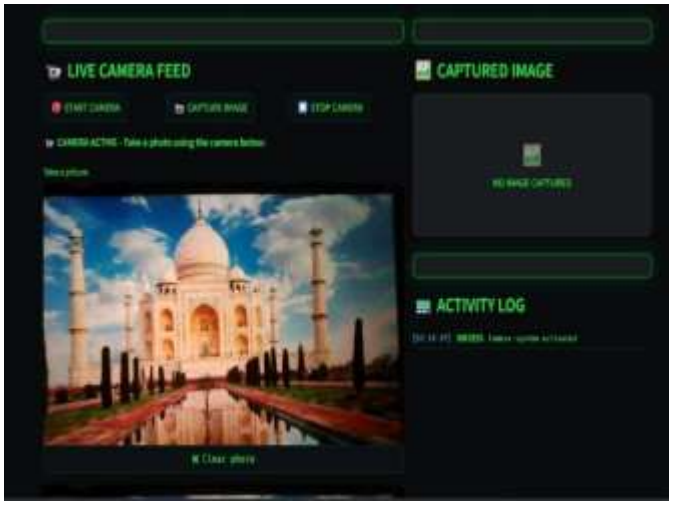

Fig. 2. Webcam page

This portion of the page illustrates the control and setup interface for an AI captioning system with Google API key setup options to enable users to use Google's AI technology for automatic generation of image descriptions. When a Google API key is validated and activated by the system, it will inform users they can now generate captions using the Internet. The top part of this section will show that internet generated captions have been enabled; the bottom of this section has a section where customization can be made, like choice of output language for captions, i.e., English, or caption vibe control i.e., tone/style, formal/casual/creative, etc. This is an example of how some recent AI applications have taken advantage of cloud computing by using third party (Google) cloud services to expand on their current abilities and provide their customers with more choices for customizing both the technical and creative aspects of their captioning solution. [Fig.3]

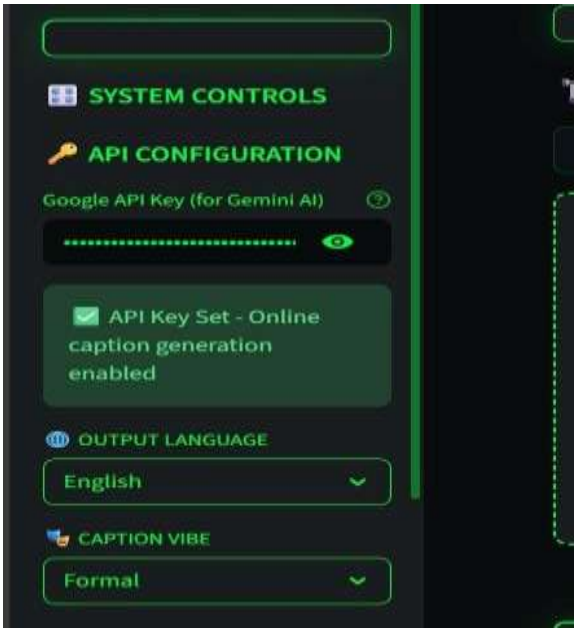

Fig. 3. Navigation page.

In the image is a diagnostic application to assist Healthcare Professionals with identifying different types of bone fractures. This diagnostic application uses Artificial Intelligence (AI) combined with X-Rays uploaded by users, which are able through the diagnostic software to be evaluated for Various Types of Fractures (format supported; file sizes up to 200MB-compressed JPG/JPEG/PNG). The diagnostic application also offers a section called Clinical Information for the Patient. Patients' Clinical Information includes what kind of pain or discomfort the patient is having (which can be selected from a drop-down box located to the left of the uploaded images) - the primary symptom/condition that has been pre-selected as Limited Movement. Thus, the software allows professionals to quickly evaluate and provide insight into potential bone fracture scenarios. [Fig.4]

Fig.4 Uploaded Image

In this image, you see the latest part of the medical diagnostic tool for diagnosing fractures. As mentioned in the previous post, this tool seems to be the same as the one used for diagnosing fractures from X-ray images. The screen shows an X-ray image of the forearm and elbow of a person whose arm has been broken due to what appears to be a serious accident as shown in the X-ray image. The person entering the data used the field of the page title "additional information (such as age, type of accident, other medical conditions)" to record the reason for the break. In the box below this field, there is a "diagnose fracture" button, which is used by the doctor to start the process of analyzing the X-ray and the additional data to arrive at a diagnosis. Figure 5 is an illustration of the information that the diagnostic tool has received and processed. It has received an X-ray and an explanation of how the X-ray got to the physician, and is preparing to use those two pieces of information to find out what the diagnosis should be.

Fig.5 Diagnose Fracture

4. IMPLEMENTATION

An example of a user interface for a live image recognition software utilizing AI captioning and audio feedback functions via a camera. The upper section has a live webcam recording of an image of India's Taj Mahal, with this live feed shown in a captured image panel after processing. The Taj Mahal is a 17th Century Mughal monument constructed by the Emperor Shah Jahan to honour his second wife Mumtaz Mahal who died giving birth to their fourteenth child. The monument represents eternal love and has been designated as a UNESCO World Heritage Site.On the left side of the interface is the control panel and API configuration for creating a Google API key for Gemini AI that will allow users to use the Online AI services for automatic caption generation. The output language and style for the generated caption give users an option to select which language and type of caption they would like their captions to be – formal, creative etc. The interface will allow users to turn on and adjust the speed of the caption narration, allowing users to use text-to-speech capability.

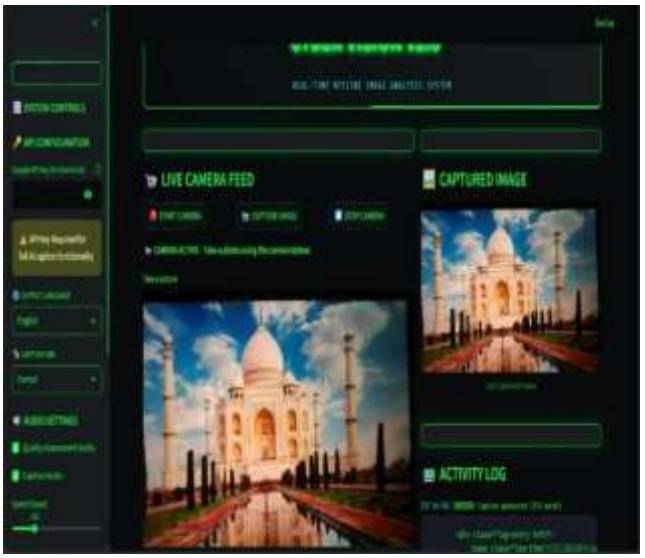

Fig.6 Captured Image

The interface appears to belong to an application that allows the user to analyze photos or videos based on the following two categories of quality: "QUALITY ANALYSIS" which indicates a perfect rating (100%) of good quality for the visual input being analyzed and "CAPTION OUTPUT" which creates a description that is very elaborate to the extent of sounding dramatic and theatrical and uses very flowery words to paint a picture of what is referenced by the caption - for instance, a vehicle, being compared to a dramatic performance on stage with a certain amount of tension created in the end through, perhaps, the use of space, sound and images, etc. The "HINDI TRANSLATION" may indicate that the original caption was created or exists in a different language (commonly referred to as "HI" or "HI-P") as compared to the English language. Essentially, this tool allows users to evaluate technical quality of their visual media while at the same time, creates amusingly exaggerated and elaborate captions of their visual media. [Fig.7].

Fig.7 Caption Generated

According to the AI diagnosis, the predicted type of fracture that the patient has sustained is classed as a simple fracture with an 87% certainty of being accurate. The risk of complications due to the injury is low and therefore, monitoring the patient's recovery process is recommended using standardised processes (as per normal practice). The patient's current clinical presentation demonstrates a lack of mobility. In addition, an X-Ray image of the patient will be attached to this report, and the resulting fracture area will be indicated in the bottom portion the radius bone (around the wrist) with a pink box [Fig.8].

Fig.8 Fracture Area

This X-ray is presented with a fracture localization overlay, highlighted by the pink circular heatmap region. The heatmap here represents the AI explainability layer, which identifies and focuses on the area most likely affected by a bone fracture. The highlighted section in this case indicates a possible fracture around the distal radius, a common site for wrist injuries. This visualization helps medical professionals and diagnostic models interpret X-ray scans in a more transparent fashion by pointing out regions of interest. Such explainability overlays form a critical component of AI-assisted medical imaging, enhancing both clinical confidence and model accountability by clearly showing why a system made a certain prediction. [Fig.9]

Fig.9 Fracture Localization

5. FUTURE SCOPE

The suggested system for smart image captioning with X-ray bone healing detection can be improved in the following ways to increase its clinical utility and prediction power:

1. Integration with Complete Medical Datasets: Future research can include larger and more varied medical imaging datasets including MURA (Musculoskeletal Radiographs), RSNA Bone X-ray Dataset, and CheXNet datasets to enhance further the generalization of models through various patient populations and conditions of X-ray imaging. CheXNet datasets, to enhance the model's generalizability across various patient populations and X-ray imaging conditions.

Future research can include larger and more varied medical imaging datasets, including MURA (Musculoskeletal Radiographs), RSNA Bone X-ray Dataset, and CheXNet datasets, to enhance the model's generalizability across various patient populations and X-ray imaging conditions.

2. Integrating multiple modes of learning by combining clinical text and X-ray images will enhance the ability of the system to create more relevant and specific captions for the patient's healing process. This will make it easier to interpret the patient's estimated bone healing.

3. The creation of a cloud-based dashboard may enable the remote physician to monitor a patient's bone healing quickly and accurately so that they may respond quickly to changes in the patient.

4. Predictive modeling of healing rates involves the utilization of regression modelling and predictive modelling to estimate true healing rates of bone and allows for an analysis of the degree of quantification associated with an injury and the level of qualified information available regarding each patient's condition. In doing so, predictive modelling will provide a tool to support personalized planning and prognosis.

5. Additionally, the incorporation of techniques such as Grad-CAM++ creates an opportunity for more comprehensive and visual-based model output interpretation. In turn, this creates increased clinician trust in AI-based model captioning and allows for validation of AI-based model captions.

6. CONCLUSION

The development of a unified "Image Captioning and MedCLIP" based system (EU:1) marks an important milestone for intelligent, support automation in the field of diagnosis, given the advances made so far. The hybrid approach described in this document is expected to benefit all physicians by providing valuable tools that will enable them to better serve their respective patients with confidence, accurate representation of patient data and proper follow-up on cases.

The project illustrates that by utilizing deep learning techniques to develop machine-generated captions on medical imaging, it is possible to extract valuable information from these images and convert the resulting information into captions that are understandable and meaningful to radiologists. Deep Learning Image Captioning may reduce the number of human errors in image-processing systems and eliminate some of the repetitive aspects of the radiologist's work. MedCLIP improves both the reliability of the system and the use of structured medical data as captions by means of using specific disease embeddings and generating similarity scores between captions and images. Together, the disease-specific embeddings and similarity scores form an effective, detailed framework for producing clinically relevant insights. In addition to these two features, the system places importance on providing transparent and trustworthy results to the end-user; thus, it allows for the use of visual aids such as heatmaps and attention maps, to provide clinicians with the ability to verify the accuracy and reliability of AI-driven imaging analyses.

REFERENCES

[1] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," arXiv preprint arXiv:2201.12086, 2022 .[Online]. Available: https://arxiv.org/abs/2201.12086

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning Transferable Visual Models From Natural Language Supervision," Proceedings of the International Conference on Machine Learning (ICML), 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[3] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[4] R. R. Selvaraju, W. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017. [Online]. Available: https://arxiv.org/abs/1610.02391

[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs," Stanford Machine Learning Group, 2017. [Online]. Available: https://stanfordmlgroup.github.io/competitions/mura/

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning," arXiv preprint arXiv:1711.05225, 2017. [Online]. Available: https://arxiv.org/abs/1711.05225

[7] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[8] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2018. [Online]. Available: https://arxiv.org/abs/1710.11063

[9] Stanford Machine Learning Group, "MURA: MSK X-rays," Stanford AIMI, [Online]. Available: https://aimi.stanford.edu/datasets/mura-msk-xrays