X-Care: Suite of Rehabilitation Tools

Dev Shah
University of Waterloo
d73shah@uwaterloo.ca

Hargun Mujral
University of Waterloo
hmujral@uwaterloo.ca

Ali Al-Hamadani
University of Waterloo
a2alhama@uwaterloo.ca

Helen Li
University of Waterloo
h273li@uwaterloo.ca

Guruprasanna Suresh
University of Waterloo
grajukan@uwaterloo.ca

Mukund Chettiar
University of Waterloo
msayeega@uwaterloo.ca

***Abstract*—This paper introduces a novel diagnostic tool that combines Supervised Learning in image classification, and Retrieval-Augmented Generation (RAG) using LLMs, to support the detection and diagnosis of injuries based on X-ray images for patients. By applying transfer learning with specialized datasets on DenseNet and EfficientNet convolutional neural networks (CNNs), which were pretrained on the ImageNet dataset, the tool approaches expert-level accuracy in detecting and classifying various skeletal fractures. Then, the RAG-powered rehabilitation agent allows for informative plans and recommendations, using clinically verified documents from a comprehensive medical database (PubMed). The fusion of these technologies offers significant potential to improve patient care by providing rapid and sophisticated analyses, in an effort to bridge the gap between diagnostic imaging and consultation with a professional to enable a quicker recovery and reduce patient uncertainty.**

I. Introduction

There is an urgent necessity to advance the quality and speed of X-ray diagnosis. As it currently stands, there is an average delay of 14-days [Vancouver Coastal Health, 2024] between performing an X-ray scan and an appointment with a specialist to discuss results. This gap not only hinders the diagnostic process but also leaves patients in a state of uncertainty and anxiety while awaiting further consultation with healthcare specialists. And considering that up to 25% of X-rays are rejected or must be repeated [Little et al., 2017], this issue is exacerbated. Thus we hope to tackle the problem of the absence of immediate rehabilitative guidance for patients throughout prevalent X-ray procedures.

A. Motivation

Conventional methods of interpreting X-ray images are often hampered by speed and variability of human analysis, which can yield inconsistent diagnostic outcomes [Kuo et al., 2022]. By employing Supervised Learning and Retrieval-Augmented Generation (RAG), we strive for diagnostic precision and consistency surpassing traditional human-led evaluations.

Our aspirations, however, extend beyond diagnostics. Leveraging AI to scrutinize X-ray data, we aim to offer patients personalized rehabilitation plans and provide immediate, informed responses to their queries. This approach is designed to kickstart recovery ahead of specialist consultations, providing vital, timely advice when it is most needed [Sharma, 2023].

The fundamental objective is to elevate patient outcomes through rapid, precise diagnostics and early intervention strategies. Through this project, we underscore our commitment to the application of state-of-the-art AI technologies within the healthcare field, contributing to diagnosis and facilitating a seamless recovery journey.

II. Related Work

Multiple studies have highlighted the efficacy of artificial intelligence in the realm of radiography-based fracture detection. A comprehensive review by a team of researchers in England, encompassing 42 studies, revealed comparable diagnostic performance between AI and clinicians, with both exhibiting a sensitivity of 91-92% [Kuo et al., 2022]. Notably, the BoneView model released by startup Gleamer in France, demonstrated proficiency to a degree higher than musculoskeletal radiologists in detecting fractures across diverse anatomical regions [Oppenheimer et al., 2023]. Their approach reduced missed fractures by 29% and enhanced sensitivity by 16%, illustrating the promise of AI in the field of orthopaedic radiology [Guermazi et al., 2022].

Presently, the most common methodology involves training AI models using deep learning networks. A study demonstrated the effectiveness of re-training a pre-existing CNN model, specifically an Inception v3 model, which achieved a binary classification model with 90% sensitivity and 88% specificity in fracture detection [Kim and MacKinnon, 2018]. Additionally, a group of doctors and researchers from Sweden utilized openly available deep learning networks to create AI fracture detectors that performed similarly to senior orthopaedic surgeons when analyzing medical images [Olc, 2017].

Although numerous works have shown the effectiveness of employing pre-trained deep-learning networks for the task of classifying fractures, integrating AI into real-world systems to benefit patients beyond mere fracture detection persists to be an issue. Novel techniques using Large Language Models (LLMs), such as RAG, have proven to reduce hallucination and provide accurate information when summarizing radiology papers, extracting relevant information and returning deterministic data [Lewis et al., 2021]. This approach can provide patients with recent and precise information extending from

the training data in the LLM, and when used effectively toward creating rehabilitation plans, can allow for the initiation of treatment without delays [Sharma, 2023].

III. Methodology

A. X-Ray Classification & Recognition: Computer Vision

We approach the problem of X-ray recognition and interpretation by dividing it into two parts: The presence of a fracture (Binary Classification) and fracture's location within the X-Ray scan.

a) Research & Plan: The initial stage involved researching the problems and devising a plan. We opted to use a DenseNet model for binary classification, with the aim of identifying the presence or absence of a fracture using the FracAtlas Dataset. An object detection algorithm, such as Faster R-CNN, was incorporated for detecting the fracture's location.

b) Plan Update: Upon implementation, we faced challenges integrating Faster R-CNN with DenseNet and discovered limitations in the FracAtlas dataset. Consequently, we adjusted our strategy to assign dedicated models for each step and sought a new dataset for more comprehensive body part classification.

B. Step 1: Presence of a Fracture

1) Model Information: The model for detecting the presence of fractures was trained on the FracAtlas dataset, comprising 4,083 annotated X-ray images. We utilized a DenseNet architecture for its efficiency in feature extraction, crucial for the binary classification task at hand. The classification was binary, with labels 1 (fractured) and 0 (not fractured). The dataset was divided into training, validation, and testing sets with a ratio of 70:15:15, ensuring a balanced approach to model training and evaluation.

2) Testing Results: The initial reference model's hyperparameters, configurations, and accuracy are as follows (Table I):

TABLE I
Comparison of Model Parameters and Performance Before and After Tuning

Parameter	Initial	Tuned
#(Epochs)	5	3
Batch Size	64	64
Loss Function	BCE Loss	BCE Loss
Optimizer	SGD	Adam
Learning Rate	0.001	0.001
Validation Accuracy	74.56%	88.40%
Test Accuracy	71.84%	86.60%

BCE Loss: Binary Cross Entropy Loss.

We chose to initiate testing with this model and its configurations. The validation accuracy achieved 74.56% across epochs 1-5. While this is a good start, there is room for improvement through hyperparameter testing. It can be inferred from these tests that the number of epochs does not significantly impact validation accuracy. After fine-tuning the hyperparameters, the test accuracy greatly increased to 86.6%. The finalized model is as shown in Table I.

C. Step 2: Fracture Location (Multi-class Classification)

In this step, we focus on the multi-class classification of fractures based on their location within X-ray images. Identifying an optimal dataset for this purpose required a comprehensive review of available resources. Below, we detail our evaluation of several datasets considered for this task.

1) Datasets for Multi-class Classification: Our search for appropriate datasets yielded three promising candidates, each with distinct features and potential applications in fracture location classification.

a) UNIFESP X-Ray Body Part Classification Dataset: This dataset consists of 2,481 X-ray images across 22 body parts, annotated in a multilabel format, making it ideal for detailed classification tasks in our second step.

b) VinDr-BodyPartXR: Derived from DICOM scans, it includes general labels such as "abdominal," "pediatric," and "adult." Its broad categorizations, however, lacked the specificity needed for precise classification in our study.

c) MURA (Musculoskeletal Radiographs): A large collection focusing on upper body parts, including 7 specific areas. While extensive, its scope was too narrow for the comprehensive classification goals of our project.

2) Model Information: The selected dataset, UNIFESP X-Ray Body Part Classification, consists of 2481 images, narrowed down to 1606 images after preprocessing for clearer label distinction. We employed a DenseNet architecture tailored for a 22-class classification task, representing various body parts. The training and testing were conducted on an 80:20 split.

3) Data Preprocessing: Upon examination, it was discovered that only the training set of the UNIFESP dataset had labels. Thus, we could only use the images from the training set, totalling 1738 images. The labels, represented as numerical values ranging from 0 to 21 for the 22 body parts, were found under the "Target" column in the training CSV file. To streamline the dataset and make it easier to work with, images with multiple corresponding body part labels were removed, resulting in a finalized dataset of 1606 images. Then, we divided this dataset into an 80:20 ratio for training and testing, respectively.

4) Testing Results: In the initial trial, the model was trained with configurations summarized in Table II.

TABLE II
Model Configuration and Performance

Parameter	Initial	Tuned
#(Epochs)	9	5
Batch Size	64	32
Optimizer	SGD	Adam
Learning Rate	0.001	0.0001
Validation Accuracy	96.18%	99.38%
Test Accuracy	90.06%	91.30%

To minimize overfitting and improve the test accuracy, we visually inspected the dataset for duplicate images. Although all images were unique, many shared similarities in terms of backgrounds and bone placement, potentially contributing

to the risk of overfitting. Consequently, we concluded that training the model with a reduced number of epochs (3-5) would be more appropriate.

Next, ten different model configurations with 3-5 epochs were tested, and the highest scoring model (ID #9) was selected as the final "tuned" result. The outcomes of the 10 models are presented in Table III.

TABLE III
OUTCOMES OF THE 10 MODEL CONFIGURATIONS

ID	E	BS	Opt	LR	WD	TAcc	TeAcc
1	5	32	SGD	0.001	0.001	96.11%	89.44%
2	3	64	Adam	0.0001	0.001	95.95%	88.51%
3	5	64	RMS	0.01	0.001	46.18%	40.68%
4	3	32	RMS	0.001	0.01	45.87%	45.96%
5	5	64	SGD	0.0001	0.01	50.23%	47.83%
6	3	32	Adam	0.01	0.01	45.95%	45.03%
7	5	32	RMS	0.001	0.1	44.63%	46.89%
8	3	64	SGD	0.0001	0.1	50.23%	48.76%
9	5	32	Adam	0.0001	0.1	99.38%	91.30%
10	3	32	SGD	0.01	0.1	44.55%	47.20%

E: Epochs, BS: Batch Size, Opt: Optimizer, LR: Learning Rate, WD: Weight Decay, TAcc: Train Accuracy, TeAcc: Test Accuracy.

Among the initial ten models, three exhibited high accuracy ($> 70\%$), while the remaining models yielded accuracy below 50%. To ensure consistent weight initialization across all models, the top-performing models (Models 1, 2, and 9) were reassessed after fixing the seed.

In general, regarding our specific task, the batch size seemed the least relevant, as successful models were achieved with varying batch sizes. Optimizers such as SGD and Adam performed well, alongside effective learning rates of 0.001 and 0.0001, as well as weight decays of 0.001 and 0.1.

To confirm these inferred conclusions, we tested two more models (11 and 12), as shown in Table IV.

TABLE IV
PERFORMANCE OF ADDITIONAL MODELS

ID	E	BS	Opt	LR	WD	TAcc	TeAcc
11	5	32	Adam	0.01	0.1	45.40%	43.79%
12	3	32	Adagrad	0.0001	0.1	79.67%	73.29%

E: Epochs, BS: Batch Size, Opt: Optimizer, LR: Learning Rate, WD: Weight Decay, TAcc: Train Accuracy, TeAcc: Test Accuracy.

Model 11 was derived from the highest accuracy model (Model 9) to investigate the hypothesis that a learning rate of 0.01 would adversely affect its accuracy. The hypothesis proved correct, as Model 11 exhibited poor results. Additionally, Model 12 underwent tuning with the same hyperparameter configurations as Model 9, but with Adagrad as its optimizer. The outcomes indicate that Adagrad performed less effectively than Adam. Therefore, Adam remains the optimal optimizer for this task.

D. Patient Diagnosis: RAG

To generate an accurate diagnosis for the provided X-ray, a Retrieval Augmented Generation (RAG) pipeline was developed. This pipeline integrates two embedding models: the open-source HuggingFace model and OpenAI's text-embedding-002-ada model. These models were selected to balance cost and performance, with the HuggingFace model serving as a cost-effective alternative to the more expensive OpenAI model. Additionally, the pipeline utilizes two Large Language Models (LLMs): Cohere's prompting LLM and OpenAI's GPT-4, providing alternatives for cost efficiency and accuracy.

The RAG approach was preferred over fine-tuning models like GPT-4 due to its demonstrated accuracy with smaller datasets and the impracticality of annotating each prompt for GPT-4 fine-tuning. In the pre-processing stage, medical diagnosis documents from PubMed and Springer were converted into vector embeddings using Cohere's Embedding model.

Upon classification of the user's X-Ray image and embedding of the fracture location and other patient information, this data is used to query a vector database. Relevant documents are then retrieved to provide context for the GPT-4 LLM, which generates the diagnosis response.

E. Full Stack

The interface for user interaction with the system was developed using a React frontend and a Flask backend. The X-Ray diagnosis architecture facilitates user input through three distinct steps:

1) Step 1: Fracture Identification: The user uploads an X-Ray image, which is analyzed by the CV model to detect the presence of a fracture. Depending on the outcome, the system either stops with a notification of no issues or proceeds after identifying the fracture type, allowing user corrections if necessary.

2) Step 2: Location Identification: This step involves the CV model pinpointing the fracture's location, visually indicated by a labeled box. Users can correct the identified location if they believe there's an error.

3) Step 3: Diagnosis and Treatment Suggestions: Finally, users receive a diagnosis summary and a suggested treatment plan, generated by the RAG pipeline. They have the option to request a diagnosis regeneration for additional considerations.

IV. RESULTS AND DISCUSSION

In this section, present your results in the form of tables, diagrams, and plots. Discuss your results and the significance of each. Continue the discussion to explain what you learned and discovered over the course of your project. How did your model perform? What insights did it generate and how did you interpret it? Did your methods work as expected? Are there advantages (or disadvantages) to the methods you used?

TABLE V
EXAMPLE TABLE SHOWING THE RESULTS OF AN EXPERIMENT.

Model	Accuracy	Precision	Recall	F1
CNN	97.78%	82.32%	88.66%	90.61%
SVM	86.43%	78.41%	67.43%	55.21%
RNN	79.21%	94.13%	80.03%	75.79%

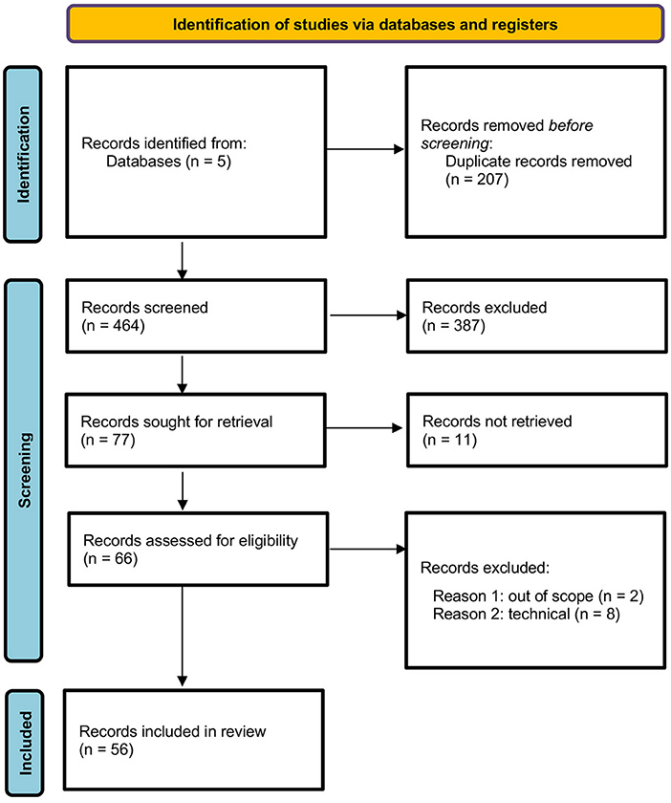

Fig. 1. Example of an image in a single column. Ethics.

Below is an example of a table to display results. You can add columns and rows. Describe what information is presented in the table in the caption. Make sure to describe the significance of the information presented in each table and figure. Also, here is a useful website for generating tables.

A. Ethical Considerations

In this section write about the possible Ethical Considerations you thought of when writing the paper. It is important to think about the possible implications of your research.

V. Conclusion

In this section wrap up your project and give a summary of the work that was done. What were the overall results and contributions? What are the impacts of this research on other people?

VI. Future Work

In this section write about the possible future works your team would do on this topic or related topics. Describe what your next steps would be in its development, what challenges remain, and what's most important in your opinion to work on next.

VII. Limitations

Talk about any limitations you experienced during your project here related to resources, time, or any other contraints.

VIII. Acknowledgements

If any indivduals or organizations helped you with your project or your paper, make sure you acknowledge them here.

IX. Appendix

Use this section to add additional figures and tables.

References

[Olc, 2017] (2017). Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthopaedica*, 88(6):581–586. PMID: 28681679.

[Guermazi et al., 2022] Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M., Lacave, E., Rahimi, H., Pourchot, A., Parisien, R. L., Merritt, A. C., Comeau, D., Regnard, N.-E., and Hayashi, D. (2022). Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, 302(3):627–636. PMID: 34931859.

[Kim and MacKinnon, 2018] Kim, D. and MacKinnon, T. (2018). Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, 73(5):439–445.

[Kuo et al., 2022] Kuo, R. Y., Harrison, C., Curran, T.-A., Jones, B., Freethy, A., Cussons, D., Stewart, M., Collins, G. S., and Furniss, D. (2022). Artificial intelligence in fracture detection: A systematic review and meta-analysis. *Radiology*, 304(1):50–62. PMID: 35348381.

[Lewis et al., 2021] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

[Little et al., 2017] Little, K., Reiser, I., Liu, L., Kinsey, T., Sánchez, A., Haas, K., Mallory, F., Froman, C., and Lu, Z. (2017). Unified database for rejected image analysis across multiple vendors in radiography. *J Am Coll Radiol*, 14(2):208–216. Epub 2016 Sep 20. PMID: 27663061.

[Oppenheimer et al., 2023] Oppenheimer, J., Lüken, S., Hamm, B., and Niehues, S. (2023). A prospective approach to integration of ai fracture detection software in radiographs into clinical workflow. *Life (Basel)*, 13(1):223. PMID: 36676172; PMCID: PMC9864518.

[Sharma, 2023] Sharma, S. (2023). Artificial intelligence for fracture diagnosis in orthopedic x-rays: current developments and future potential. *SICOT J*, 9:21. Epub 2023 Jul 6. PMID: 37409882; PMCID: PMC10324466.

[Vancouver Coastal Health, 2024] Vancouver Coastal Health (2024). X-Ray. https://www.vch.ca/en/service/x-ray#short-description--6916. Online; accessed 29 January 2024.