Real Time Orthopaedic Trauma Radiographic Assessment Using Hybrid Large Language Model (LLM) Neural Network Architecture

Nicholas Alexander Vandal, Mohammed Shayan Abdullah[1], Jaret McGraw Karnuta[2], Stephen Richard Barchick, Derek James Donegan, Samir Mehta[3]
[1]The University of Pennsylvania, [2]Hospital of the University of Pennsylvania, [3]University of Pennsylvania

INTRODUCTION:
The advent of transformer-based Large Language Models (LLMs), exemplified by ChatGPT has revolutionized natural language processing, overcame previous limitations and offers new pathways for leveraging artificial intelligence (AI) in healthcare despite being fraught with privacy concerns and overall lack of technical input from orthopedic surgeons. Current open source LLMs lack the ability to accurately assess radiographic parameters in patients sustaining orthopedic trauma. Our study introduces a composite orthopaedic trauma multimodal model LLM paired with a custom vision classifier trained on fracture parameters. We maximally employed fine-tuning and transfer learning on locally controlled graphics processing units (GPUs) to classify and provide suggested management of traumatic fractures in real time, with an initial focus on the distal radius.

METHODS:
This study was conducted with Institutional Review Board (IRB) approval. A retrospective cohort of distal radius fractures treated surgically at a single academic health system were collected over a twenty-five-year period (1998 to 2023). A subset of 100 patient charts with accessible plain film radiographs were annotated with clinical and radiographic attributes, and the deidentified text of the orthopedic consult note or clinic note, radiologist reads, and operative reports were collected for inputs into our language model.Using open-source machine learning libraries PyTorch and fastai, we trained two models, 1) a multilabel vision classifier using traditional convolutional neural net architecture, and 2) an image aware LLM, whose input includes the text representation of label probabilities from the vision classifier. The multilabel classifier incorporates discrete probability estimates of fracture patterns occurring within each radiograph using parameters that orthopaedic surgeons understand such as dorsal displacement, or comminution. Our most performant visual model was paired with a PubMed and GPT-4 guided fine-tuned multimodal LLM (derived from the LLaVA architecture and training methodology). This hybrid approach of LLM plus traditional vision classifier provides a more comprehensive analysis of the fracture.

Vision Classifier Training: The vision classifier pretraining involved two stages. The first stage utilized large-scale grayscale pretraining, adapting pretraining code to train multiple classifiers over the entire 14 million ImageNet dataset with single-channel grayscale transformations. The second stage involved merging and cleaning the following preexisting fracture databases, GRaZPEDWRI-DX, FracAtlas, MURA, LERA, Joint Fracture Dataset, and Bone Break Classification Image Dataset, to create a combined dataset of 66,163 radiographs with 76 possible tags. This dataset was then used to train multiple neural network architectures (Resnet, TResNet, Convnext) while also employing commonly used techniques machine learning to improve the generalization of models (data augmentation and masked binary cross-entropy).

Language Model Training: The language model training followed the methodology of LLaVA-Med which is a vision capable LLM (LLaVA architecture) trained for the biomedical domain. Model weights were initialized from LLaVA-Med and fine-tuning of the LLaVA-Ortho model was done in two phases: concept alignment and instruction tuning. Orthopaedic concept alignment data with 41,482 orthopedic trauma-specific image-text pairs downloaded from PubMed Central articles, using the actual figure captions as expected outputs.Instruction tuning employed GPT-4 to generate plausible conversations regarding a specific image PubMed image based on its caption and mentions within the article text. The final phase of training was accomplished using Low-Rank Adaptation (LoRA) technique on locally controlled GPUs on our custom dataset to avoid HIPAA and privacy concerns of cloud-based AI providers.

Statistical Analysis was performed using on the vision classifier Multilabel Accuracy and F1 scores. This metric measures the accuracy of the model for each image. It is calculated with a threshold (t=0.5) to determine the presence of each label. The F1 score is a harmonic mean of precision and recall of the classifier, providing a single metric that balances both

RESULTS: We developed an initial orthopedic-specific multimodal LLM, which employs an additional domain-specific visual classifier and is capable of inference and training without exposure of HIPAA data to cloud-based services. Quantitative assessment of LLMs is challenging and evaluation of our ortho-LLaVA model is ongoing, but qualitative experiments demonstrate improved fluency of orthopaedic and radiographic terminology. The visual pretraining achieved significant performance improvements, recovering 95% of the performance of color-trained models on grayscale datasets.

The combined fracture database models demonstrated high validation metrics, with the best-performing model achieving a multilabel accuracy of 0.988 and F1 score 0.938 indicating robust performance. The model demonstrated impressive performance in test sets, achieving an average F1 score of 0.81 for six major pathological findings internal distal radius test set.

DISCUSSION AND CONCLUSION: This approach represents a significant step forward in applying AI technologies to orthopedic surgery, promising enhanced diagnostic and treatment strategies within a privacy-conscious framework. The study demonstrates the feasibility of developing a locally controlled, fine-tuned multimodal AI model for orthopedic trauma, with future plans to expand the model to a broader range of traumatic injuries.

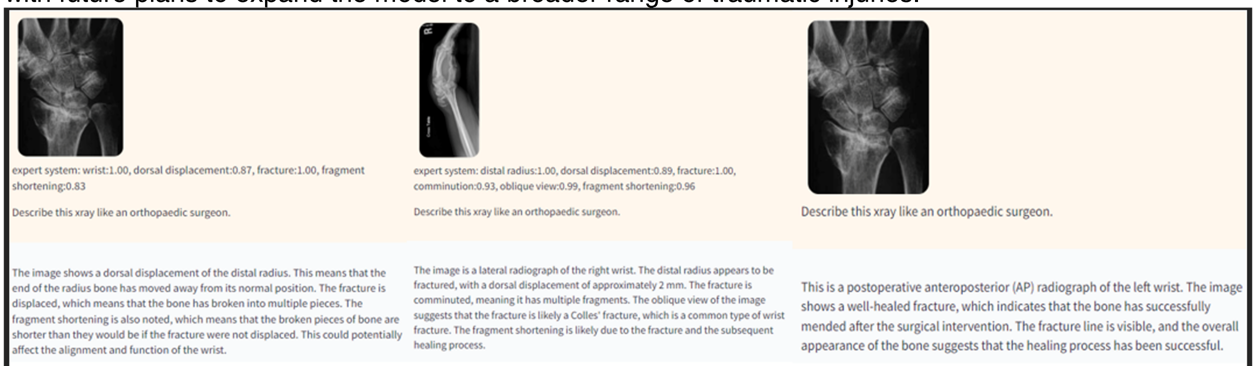

Figure 1. Radiographic evaluation of distal radius fractures using developed hybrid LLM + expert vision classifier model