Summer Internship Report

at

CARPL.AI

(April 15th to June 28th, 2024)

A Report

By

ESHIKA BINDAL

PGDM (Hospital and Health Management)

2023-2025

International Institute of Health Management Research, New Delhi

Acknowledgements

I would like to extend my heartfelt gratitude to several individuals and institutions for their invaluable support and guidance throughout my summer internship, which have culminated in the successful completion of this report.

First and foremost, I am deeply grateful to my mentors and supervisors at CARPL.ai, Mr. Rinku Yadav and Dr Suyash Khubchandani, for their unwavering support, insightful guidance, and continuous encouragement. Their dedication, extensive knowledge, and insightful feedback have been invaluable. Their expertise and mentorship have been instrumental in my learning and development during this internship.

I would also like to express my sincere appreciation to Dr Amit Kumar for his significant contribution to this project. His assistance with data analysis and his readiness to help me overcome various challenges have been critical in ensuring the successful completion of this report.

Additionally, I am immensely grateful to my mentor at IIHMR Delhi, Dr. Punit Yadav, for his continuous support and valuable advice, which have greatly enriched my internship experience. His support has not only helped me overcome obstacles but has also encouraged me to explore new opportunities.

Furthermore, I am thankful to my college, the International Institute of Health Management Research, New Delhi, for providing me with this excellent internship opportunity. The knowledge and skills I have gained during this period will undoubtedly be beneficial for my future endeavors.

Finally, I would like to acknowledge the support of my family, friends, and colleagues, whose encouragement and understanding have been a source of strength throughout this journey.

I am sincerely grateful to have had the opportunity to work with such inspiring individuals. Their belief in my abilities has motivated me to continually strive for excellence.

With warm regards and sincere gratitude,

Eshika Bindal

(Completion of Summer Internship from CARPL.ai Pvt. Ltd.)

The certificate is awarded to

ESHIKA BINDAL

In recognition of having successfully completed her
Internship in the department of

HCP Onboarding and Demand Generation

and has successfully completed her Project on

AI-assisted fracture detection in upper body MSK Imaging: Performance and Discordance Analysis

Date: 28/06/2024

Organization: CARPL.ai Pvt. Ltd.

She comes across as a committed, sincere & diligent person who has a strong drive & zeal for learning

We wish her all the best for future endeavors

Vaishali

CARPL.AI PVT. LTD. New Delhi

Organization Supervisor **Head-HR/Department Head**

Certificate of Approval

The Summer Internship Project of titled **"AI-Assisted Fracture Detection in Upper Body MSK Imaging: Performance and Discordance Analysis"** at **CARPL.ai Pvt. Ltd.** is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of Post Graduate Diploma in Health and Hospital Management for which it has been submitted.It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approve the report only for the purpose it is submitted.

Dr. Punit Yadav
Professor
IIHMR, Delhi

FEEDBACK FORM
(Organization Supervisor)

Name of the Student: ESHIKA BINDAL

Summer Internship Institution: CARPL.ai. Pvt. Ltd.

Area of Summer Internship: Demand Generation, Database Management, HCP onboarding and Marketing.

Attendance: 100%.

Objectives met:
1. Learnt about strategies of demand generation.
2. Learnt about AI ecosystem.
3. Learnt automation tools - Apollo.io
4. Prepared database for different strategies - ICP Readrouts.
5. Conferences and CARPL Awareness Programs.

Deliverables:
1. Database - ICP
   - Conferences (SIIM)
   - CARPL awareness Program
2. ROI analysis of Modalities in Australia
3. Analysis of AI performance at clinical site.

Strengths:
Database Management, Workload handling, Modality Research, Good at handling softwares such as Apollo.io, PowerBI, Excel, Powerpoint.

Suggestions for Improvement:
Learn more about AI tools and its applications in health care.

Rinku

Signature of the Officer-in-Charge (Internship)

Date: 28/6/2024
Place: New Delhi

FEEDBACK FORM
(IIHMR MENTOR)

Name of the Student: Eshika Bindal

Summer Internship Institution: Carpl. AI

Area of Summer Internship: HCP (Health Care Provider) onboarding and lead generation

Attendance: 100%

Objectives met: Yes.
AI assisted fracture detection in Upper Body MSK imaging: Performance and Discordance Analysis.

Deliverables: Lead generation
Research
Conference preparation

Strengths: Application of Research methodology and statistics in real world situation.

Suggestions for Improvement: Strength of Research methodology

Signature of the Officer-in-Charge (Internship)

Date: 30 July

Place: New Delhi

Table of Contents

Acronyms/Abbreviations

AI	Artificial Intelligence
JR	Junior Radiologist
SR	Senior Radiologist
CPU	Central Processing Unit
RIS	Radiology Information System
PACS	Picture Archiving and Communication System
HIS	Hospital Information System
MSK	Musculoskeletal
MRI	Magnetic Resonance Imaging
CT-scan	Computed Tomography scan
MRN	Medical Record Number
GT	Ground Truth
CNNs	Convolutional Neural Networks
PPV	Positive Predictive Value
NPV	Negative Predictive Value
FPR	False Positive Rate
FNR	False Negative Rate
FDR	False Discovery Rate
MCC	Matthews Correlation Coefficient
CAC	Coronary Artery Calcium
CXR	Chest X-ray

OBSERVATIONAL LEARNING

Section - 1: Introduction

CARPL.AI is designed to be a unifying platform for radiology automation, offering the world's first vendor-neutral testing and deployment platform specifically for medical imaging AI applications. It facilitates seamless integration of AI applications within the radiology ecosystem, connecting patients, radiologists, clinicians, and researchers. CARPL.AI acts as an intermediary, bridging the gap between healthcare providers and AI developers, thereby enhancing access, affordability, and the quality of medical care. The platform boasts the world's largest AI marketplace with over 100 apps, providing diverse tools to enhance clinical automation, precision medicine, and predictive medicine. By offering a unified interface to access, validate, test, and integrate AI algorithms into radiology workflows, CARPL.AI simplifies the process for renowned healthcare providers, AI researchers, industry teams, and startups globally. This dedication to advanced analytics tools ensures that clinicians can effortlessly adopt cutting-edge AI solutions to improve patient outcomes and operational efficiencies. Used by the world's leading healthcare providers, CARPL ensures that radiologists and other healthcare professionals can access a wide array of AI tools through a single, integrated platform. This streamlined approach not only enhances productivity but also drives innovation and improvements in medical care quality.

At its core, the CARPL platform connects various stages of the radiology workflow. Starting from the patient, it incorporates scanning, image analysis, radiologist interpretation, report generation, and clinical review. AI apps integrated into the platform enhance each of these stages, improving the overall accuracy and efficiency of radiological processes. The feedback loop embedded in the system allows continuous improvement and adaptation of AI algorithms, ensuring better performance and wider applications.

Vision

The vision of CARPL.AI is to transform the radiology landscape by making AI-driven solutions more accessible, affordable, and high-quality. The company aims to democratize AI in radiology, ensuring that cutting-edge technologies are available to all healthcare providers, thereby improving patient outcomes and operational efficiencies in medical imaging.

Objectives

CARPL.AI's primary objectives include:

- Enhancing Radiology Automation: By integrating advanced AI applications, CARPL.AI seeks to automate routine tasks, allowing radiologists to focus on more complex cases.
- Improving Accuracy and Efficiency: The platform aims to enhance the accuracy of diagnoses and the efficiency of radiological processes through robust AI solutions.
- Facilitating AI Development and Deployment: CARPL.AI provides a comprehensive ecosystem for the testing, monitoring, and deployment of AI applications, accelerating innovation in the field.

- Expanding Access and Affordability: By offering a vendor-neutral platform, CARPL.AI ensures that a wide range of AI tools are accessible to healthcare providers of all sizes, enhancing affordability and quality of care.

Key Features:

CARPL's comprehensive feature set supports a wide range of use-cases for its users:

- Creating "SUPER-RADIOLOGISTS": CARPL is transforming radiology by acting as an AI broker that makes radiologists exponentially more productive. By leveraging the power of the entire AI ecosystem through 30+ use cases, CARPL enables radiologists to achieve unprecedented efficiency and accuracy.
- World's Largest AI Marketplace: CARPL hosts over 110 applications from 50+ providers, covering all AI use cases in radiology.
- World's First AI Testing & Monitoring Platform: It offers a single user interface for AI validation, allowing for standardization and seamless comparison or ensembling of AI solutions.
- Deployment at Leading Clinical & Research Groups: CARPL is utilized by the world's leading healthcare providers and academic groups, supporting cutting-edge AI research and deployment.

CARPL's DEV-D Framework: A Two-Step Offering

CARPL.ai's DEV-D framework is designed to guide healthcare organizations in selecting and implementing AI tools for radiology. It focuses on making informed decisions about which AI to use and how to use it effectively, ensuring the successful integration and utilization of AI in clinical settings.

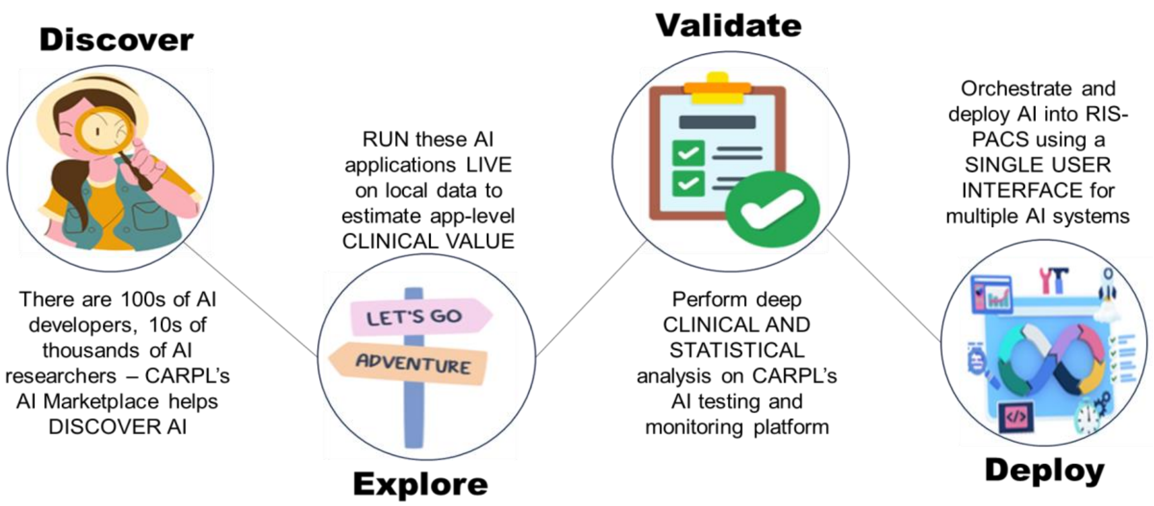

Fig 1: CARPL's DEV-D Framework

Section - 2: Mode of data collection

Data was collected from CARPL.ai

Section - 3: General findings on learning: Department-wise observations

CARPL.ai is organized into several departments, each playing a crucial role in developing and maintaining the platform's capabilities to enhance radiology AI deployments.

1. Technology and Engineering

This department is responsible for the development and maintenance of CARPL's platform. It includes software engineers, data scientists, and AI specialists who work on creating and refining the algorithms, ensuring interoperability with various healthcare systems, and maintaining the platform's infrastructure. The engineering team focuses on integrating AI applications seamlessly with radiology workflows and ensuring high performance and reliability of the system.

2. Clinical Solutions

Led by the Chief Medical Officer, the Clinical Affairs department ensures that all AI applications and tools are clinically validated and safe for use. This team comprises radiologists and other medical experts who test and validate AI solutions, ensuring they meet clinical standards and provide accurate, reliable results. They also work on optimizing clinical workflows and enhancing user experience for radiologists using the platform.

The CARPL modules and services also facilitate the attainment of regulatory approvals and compliance for AI models. Prior to commencing clinical studies on a medical device, obtaining regulatory and ethical permissions is essential. The FDA provides examples of AI/ML-based technologies applied in real-world scenarios, such as imaging systems that use algorithms for diagnostic purposes. Clinical trials are conducted by various organizations and AI companies to assess their real-world performance and obtain regulatory approval such as:

- Conformité European (CE mark) - Europe
- Food and Drug Administration (FDA or USFDA) - USA
- Therapeutic Goods Administration (TGA) -Australia
- Brazilian Health Regulatory Agency (Anvisa) - Brazil

3. Deployment

The Deployment Team at CARPL.ai is essential for integrating AI solutions into radiology departments, ensuring a seamless, efficient, and effective implementation of AI technologies. This team specializes in configuring and deploying AI applications through a single interface, thereby eliminating the need for separate integrations with RIS, PACS, and HIS systems. CARPL.ai offers flexible deployment options, including on-cloud, on-premises with CPU-only, and on-premises with both CPU and GPU resources, to meet the diverse infrastructure needs of healthcare providers. For custom deployments outside of CARPL's infrastructure, secure SSH access to the target machine is required to ensure a secure and tailored setup.

The deployment team's responsibilities extend beyond initial installation; they provide ongoing support and maintenance to ensure that AI applications continue to function optimally. This includes monitoring system performance, troubleshooting issues, and providing updates to keep the platform current with the latest AI advancements. By offering multiple deployment configurations and integrating AI through a unified interface, the deployment team addresses the common challenges AI companies face when implementing algorithms across different modalities in radiology departments. Their expertise ensures that healthcare providers can leverage AI to enhance diagnostic accuracy and patient outcomes without extensive modifications to their existing systems.

4. AI onboarding

The AI Onboarding Team at CARPL.ai plays a pivotal role in creating a comprehensive platform for medical imaging AI, encompassing capabilities such as dataset management, annotation, algorithmic integration, automated validation, testing, deployment, and IT integration. This team is responsible for establishing a strategic partner ecosystem, enabling new business models, co-sell strategies, research studies, key alliances, and managing the enterprise business and profitability of this ecosystem. They drive global AI developer relationships by developing business and operational opportunities with C-suite executives, facilitating the onboarding of AI models from around the world for various imaging modalities to implement in clinical deployments. The team's ultimate goal is to create the world's largest and most accessible AI imaging marketplace for healthcare providers, ensuring seamless integration of these AI models with the CARPL platform to enhance clinical workflows and outcomes.

5. Business Development and Sales

This department focuses on expanding CARPL.ai's market presence and building relationships with healthcare providers and AI developers. They are responsible for identifying new business opportunities, forging strategic partnerships, and driving sales. The team works to onboard new clients and ensure they are effectively utilizing the platform to improve their radiology services.

6. Marketing and Communications

This department handles CARPL.ai's branding, public relations, and marketing efforts. They work to promote the platform, communicate its benefits, and build a strong brand presence in the market. This includes creating content, managing social media, and organizing events to showcase CARPL's innovations and successes.

7. Operations and Administration

This department manages the financial health of the organization, including budgeting, accounting, and financial planning. They also handle administrative tasks to ensure smooth day-to-day operations. The finance team works to secure funding, manage investments, and ensure that the company remains financially viable.

Given the critical nature of healthcare and the use of AI, CARPL.ai has a dedicated team focused on regulatory compliance. This department ensures that all AI applications and the platform itself comply with relevant healthcare regulations and standards, such as those set by the FDA. They manage the regulatory approval process for new AI tools and ensure ongoing compliance with regulatory requirements.

CARPL.ai's structured approach with these specialized departments allows it to effectively address the complex needs of integrating AI into radiology, ensuring that the platform remains at the forefront of healthcare innovation.

Section - 4: Conclusive learning, limitations and suggestions for improvement

The internship offered a hands-on environment where I could apply the theoretical knowledge I acquired during my studies. Working alongside experienced professionals allowed me to enhance my communication skills significantly.

Throughout my internship, I was assigned multiple tasks and responsibilities which helped me refine my time management skills. I learned to prioritize tasks, set realistic deadlines, and efficiently allocate my time to ensure timely completion of projects.

Key Learnings:

1. Understanding the Radiology AI Landscape: I gained comprehensive insights into the integration of AI applications in radiology, focusing on how CARPL.AI facilitates this integration across various stages of the radiology workflow. Witnessed firsthand how AI can enhance efficiency and accuracy in radiological processes, supporting radiologists in their decision-making.
2. Market Research and Stakeholder Engagement: Conducted market research to identify potential accounts and stakeholders for CARPL.AI, which involved understanding market needs and competitor analysis. Developed skills in engaging with healthcare providers and AI developers, understanding their requirements and challenges in adopting AI solutions.
3. Observation Across Departments: Had exposure to various departments such as Technology and Engineering, Clinical Solutions, Deployment, AI Onboarding, and Regulatory Compliance. Learned about the roles each department plays in developing and maintaining the CARPL.AI platform, ensuring its functionality and compliance with healthcare standards.

   I was given the opportunity to network and collaborate with professionals from various departments. Engaging with individuals across different roles and levels of expertise broadened my perspective and allowed me to learn from their experiences. This dynamic nature of the work environment taught me the importance of adaptability and flexibility.

4. Insights into Healthcare Operations: Gained insights into healthcare operations management, particularly in the context of integrating AI into clinical settings. Explored regulatory requirements and compliance considerations critical for deploying AI solutions in healthcare environments.

Clinical Solutions Sessions at CARPL.AI

During my internship at CARPL.AI, I had the privilege of attending weekly Clinical Solutions sessions organized by the Clinical Affairs team. These sessions were designed to familiarize all interns and employees with key radiology topics and the various AI solutions available on the CARPL.AI platform. Each week, we delved into different applications, including Bone Age, Mammography, Cina Chest, MSK Fracture, Lunit Insight CXR, Oxipit, Housefield Unit, CAC, Claripy, Fatty Liver, and Lung Nodules. The primary objectives of these sessions were to provide a deep understanding of radiology concepts, educate participants on the capabilities of AI tools integrated into the platform, and share practical insights into the application of these tools in clinical settings.

One of the significant learnings from these sessions was gaining a comprehensive understanding of various radiology procedures and imaging techniques, which is crucial for anyone working in healthcare managementThis foundational knowledge was further enriched by detailed explanations of how different AI applications function and their roles in improving radiological diagnostics. We explored specific AI tools like Lunit Insight CXR for chest x-rays, Oxipit for automated radiology reporting, and Claripy for image clarity enhancement, understanding their technical workings and clinical benefits.

The sessions also highlighted the clinical impact of integrating AI into radiology workflows, including improvements in diagnostic accuracy, reduction in interpretation time, and enhancement of clinical decision-making. These insights were reinforced through real-world examples and case studies, illustrating the practical applications and positive outcomes of using these AI tools in clinical environments.

Limitaions:

- The time constraint might be a limitation, as it may have not provided sufficient duration to engage deeply into long-term projects.
- Integrating AI applications into existing healthcare IT infrastructures (RIS, PACS, HIS) proved more intricate than anticipated, potentially slowing down deployment timelines and increasing costs.
- Adoption rates among healthcare providers were hindered by insufficient training and support for end-users to effectively utilize AI tools within their daily workflows, potentially delaying full realization of benefits.

Proposed recommendations:

- Invest in developing standardized integration protocols and tools that simplify the deployment of AI applications with existing healthcare systems, reducing time and complexity.

- Strengthen partnerships with healthcare institutions to access more diverse and robust clinical data for AI validation, enhancing the reliability and confidence in AI-driven diagnostics.
- Implement a structured feedback mechanism from end-users to continuously improve AI algorithms and platform usability, ensuring alignment with evolving clinical needs and preferences.
- Develop comprehensive training programs tailored for healthcare professionals to increase their proficiency in using AI tools effectively, fostering quicker adoption and maximizing operational efficiencies.

PROJECT REPORT

AI-Assisted Fracture Detection in Upper Body MSK Imaging: Performance and Discordance Analysis

Section-1: Introduction

Radiologists play a critical role in interpreting MSK imaging studies, including X-rays, CT scans, and MRI scans. Their interpretations are pivotal for diagnosing fractures, assessing bone and joint conditions, and guiding treatment decisions [1]. However, the accuracy and consistency of these interpretations can vary significantly among radiologists due to factors such as experience, training background, and individual cognitive biases [2]. This variability can lead to discrepancies in diagnoses, potentially affecting patient outcomes and treatment efficacy.

JRs, especially those early in their careers, often face challenges related to limited clinical experience and exposure to complex cases [3]. This lack of exposure can impact their ability to accurately identify and interpret subtle or atypical findings in MSK imaging studies. On the other hand, SRs, while more experienced, may encounter challenges related to maintaining consistency and efficiency, particularly when faced with a high volume of cases [4].

Detecting fractures in MSK imaging remains a fundamental task in radiology. Fractures can range from obvious and straightforward to subtle and challenging to identify, depending on the location and severity [5]. Traditional methods rely heavily on radiologists' visual analysis and clinical judgment, which may lead to missed diagnoses or delayed treatment initiation in complex cases [6].

The challenges in fracture detection include differentiating between acute fractures, chronic injuries, and normal anatomical variants. Moreover, the interpretation of fractures can be influenced by imaging artifacts, patient positioning, and variations in imaging protocols across different healthcare facilities [3]. These factors underscore the need for standardized approaches and tools that can enhance diagnostic accuracy and reduce variability in fracture detection.

The disparity in diagnostic accuracy between JRs and SRs has been a subject of extensive research. Studies have shown that JRs may exhibit higher rates of misdiagnosis or uncertainty compared to their more experienced counterparts [1]. This discrepancy can be attributed to the learning curve associated with gaining clinical expertise and the ability to interpret complex imaging findings accurately.

Conversely, SRs may face challenges related to cognitive fatigue, which can affect decision-making and diagnostic accuracy, particularly when interpreting a large volume of studies over time [4]. Bridging the gap between JRs and SRs in terms of diagnostic accuracy is crucial for maintaining consistency in clinical practice and ensuring optimal patient care outcomes [3].

AI, particularly deep learning algorithms, has emerged as a transformative technology in medical imaging, including MSK radiology. AI has shown promise in enhancing diagnostic accuracy, reducing interpretation time, and providing decision support tools for radiologists [7]. AI models trained on large datasets can learn to recognize patterns and abnormalities in medical images, potentially outperforming human experts in certain diagnostic tasks [8].

In MSK imaging, AI applications range from fracture detection to assessment of bone density and joint abnormalities. For example, AI-driven algorithms have been developed to automatically detect osteoporotic fractures in vertebral bodies, demonstrating high sensitivity and specificity compared to traditional methods [6]. These advancements highlight AI's potential to assist radiologists by flagging suspicious findings, reducing oversight errors, and improving overall diagnostic confidence.

The comparison between AI and radiologists in MSK imaging holds significant implications for clinical practice. Firstly, evaluating the performance of AI against human experts can provide insights into AI's capability to standardize and improve diagnostic consistency [9]. Secondly, understanding how AI impacts the concordance between JRs and SRs can shed light on its role in bridging the gap in diagnostic accuracy across different levels of expertise [5].

The integration of AI into radiology practice is not without challenges, including regulatory considerations, data privacy concerns, and the need for continuous validation and improvement of AI algorithms [3]. However, the potential benefits of AI, such as reducing variability in diagnostic outcomes and enhancing patient care quality, underscore its relevance and importance in modern healthcare settings.

While AI offers promising opportunities to augment radiologists' capabilities in MSK imaging, further research is needed to validate its clinical efficacy, optimize integration strategies, and address existing challenges in implementation [7]. By leveraging AI's strengths and mitigating its limitations, healthcare providers can potentially improve diagnostic accuracy, streamline workflow efficiencies, and ultimately enhance patient outcomes in MSK radiology.

This study aims to address critical challenges and opportunities in upper body MSK radiology by evaluating the integration of AI to enhance diagnostic accuracy and consistency among radiologists. Also, it will assess AI's performance in fracture detection compared to radiologists and explore its impact on reducing variability between junior and senior radiologists, while also examining its potential benefits in clinical practice.

Research Question

1. How does the performance of AI compare with that of radiologists in detecting fractures in upper body MSK scans?
2. What is the impact of AI on the concordance and discordance between junior and senior radiologists in interpreting upper body MSK scans?

Specific Objectives

1. To evaluate the efficacy of the AI solution compared to radiologists for upper body MSK scans.

2. To analyze the concordance and discordance between junior and senior radiologists, with and without AI assistance.

Section 2: Mode of Data collection

The dataset was collected from a hospital in the US provided by Carpl.AI [10] and comprises MSK scan reports of upper body parts, including the shoulder, forearm, arm, hand, wrist, and elbow. Radiologists reviewed these scans and provided their reports, which were used to prepare the ground truth. An AI solution was also applied to these scans, generating results indicating the presence or absence of fractures. The collected data includes:

<u>For AI vs. Radiologists</u>

- **Radiologist Reports**: Written reports by radiologists indicating the presence or absence of fractures.
- **AI Results**: The AI solution's classification of each scan as either having a fracture, mentioned as suspicious finding or no fracture mentioned as no finding. These AI-generated results were then compared against the radiologist's reports to assess accuracy.

<u>For Concordance and Discordance Analysis</u>

- Scan Reports from JR: Interpretations provided by junior radiologists with and without AI assistance.
- Scan Reports from SR: Interpretations provided by senior radiologists with and without AI assistance.

Section 3: Data compilation, analysis and interpretation

1. Analysis of AI vs. Radiologist Performance

1.1 Data Preparation: The dataset was obtained in an Excel sheet format, containing 2279 scans of patients including scans of different upper body parts such as 568 for shoulder, 237 for arm, 259 for elbow, 284 for forearm, 521 for hand and 410 for wrist. The dataset has the following columns:

- <u>Group:</u> Indicates the body part category (e.g., shoulder, arm, forearm, elbow, hand, wrist).
- <u>Medical Record Number (MRN):</u> A unique identifier assigned to each patient within the hospital's records, ensuring accurate tracking and access to patient information across different departments and visits.
- <u>Accession Number:</u> A unique identifier assigned to each specific imaging study or scan within the hospital's system, ensuring precise matching to the patient's records and the specific body part being imaged.
- <u>Body Part</u>: Specifies the exact body part imaged in the scan.
- <u>Impression:</u> The radiologist's interpretation of the scan, indicating the presence or absence of a fracture.

- AI Result: The AI's classification of the scan, where 'suspicious finding' indicates a fracture and 'no finding' indicates no fracture.

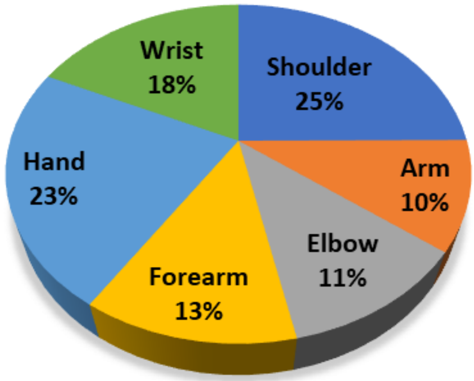

Fig 2: Distribution of radiographic scans in the dataset across different upper body

Ground Truth Marking: Each scan was reviewed and annotated with a ground truth label. If a fracture was present, a '1' was marked in the ground truth column; if no fracture was present, a '0' was marked.

Exclusion of Incomplete Data: Scans where the AI did not provide a result were excluded from the analysis to ensure a consistent comparison between AI results and ground truth.

1.2 Contingency Table Preparation

The data was sorted into groups based on the body part visualized (shoulder, arm, forearm, elbow, hand, wrist). Each result was compared, and the counts of TP, FP, TN, and FN were recorded for each body part.

- True Positives (TP): Cases where both AI and ground truth indicated a fracture.
- False Positives (FP): Cases where AI indicated a fracture but ground truth did not.
- True Negatives (TN): Cases where both AI and ground truth indicated no fracture.
- False Negatives (FN): Cases where AI indicated no fracture but ground truth did.

For each body part, a 2x2 contingency table was created to compare the AI results with the ground truth. The table included TP, FP, TN, and FN. Figure 3 represents the confusion matrix for total scans and for each upper body part.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig 3: Confusion matrix

Actual Value

Predicted Value

Shoulder	GT(+)	GT(-)
AI (+)	47	99
AI (-)	8	414

Arm	GT(+)	GT(-)
AI (+)	52	32
AI (-)	1	152

Elbow	GT(+)	GT(-)
AI (+)	47	89
AI (-)	1	122

Forearm	GT(+)	GT(-)
AI (+)	74	35
AI (-)	29	146

Hand	GT(+)	GT(-)
AI (+)	144	84
AI (-)	7	286

Wrist	GT(+)	GT(-)
AI (+)	131	79
AI (-)	6	194

Total	GT(+)	GT(-)
AI (+)	495	418
AI (-)	52	1314

Fig 4: Confusion matrices of radiographic scans in the dataset across different upper body regions.

1.3 Calculation of Performance Metrics

Performance metrics are quantitative measures used to evaluate the effectiveness and accuracy of a system or process. In this study, performance metrics help assess how well the AI system performs compared to the ground truth provided by radiologists. These metrics are derived from the contingency tables prepared for each body part, comparing the AI results to the ground truth provided by radiologists. The performance metrics used in this study include [11]:

i. Sensitivity: Indicates the AI's ability to correctly identify true fractures. High sensitivity means the AI system misses fewer fractures.

$$TPR = TP / (TP + FN)$$

ii. Specificity: Indicates the AI's ability to correctly identify true non-fractures. High specificity means the AI system produces fewer false positives.

$$SPC = TN / (FP + TN)$$

iii. PPV: Indicates the accuracy of the AI's positive fracture predictions. High PPV means that when the AI predicts a fracture, it is likely to be correct.

$$PPV = TP / (TP + FP)$$

iv. NPV: Indicates the accuracy of the AI's negative fracture predictions. High NPV means that when the AI predicts no fracture, it is likely to be correct.

$$NPV = TN / (TN + FN)$$

v. False Discovery Rate (FDR): The proportion of positive results predicted by the AI that are actually false positives. Helps in understanding the rate of false alarms. A lower FDR indicates higher reliability of positive predictions.

$$FDR = FP / (FP + TP)$$

vi. False Positive Rate (FPR): The proportion of actual negative cases that are incorrectly identified as positive by the AI. Measures the rate of incorrect positive predictions. A lower FPR indicates better performance in correctly identifying non-fractures.

$$FPR = FP / (FP + TN)$$

vii. False Negative Rate (FNR): The proportion of actual positive cases that are incorrectly identified as negative by the AI. Indicates how often the AI misses actual fractures. A lower FNR is desirable for ensuring that fractures are not overlooked.

$$FNR = FN / (FN + TP)$$

viii. F1 Score: The harmonic mean of precision (PPV) and recall (sensitivity). It balances the trade-off between precision and recall. Useful for imbalanced datasets where a high F1 score indicates a balance between precision and recall.

$$F1 = 2TP / (2TP + FP + FN)$$

ix. Matthews Correlation Coefficient (MCC): A measure of the quality of binary classifications. It takes into account true and false positives and negatives. Provides a balanced measure that can be used even if the classes are of very different sizes. An MCC value of 1 represents a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation.

TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

	Sensitivity	Specificity	PPV	NPV	FPR	FDR	FNR	F1 score	MCC
Shoulder	0.855	0.807	0.322	0.981	0.193	0.678	0.146	0.468	0.448
Arm	0.981	0.826	0.619	0.993	0.174	0.381	0.019	0.759	0.703
Elbow	0.979	0.578	0.346	0.992	0.422	0.654	0.021	0.511	0.434
Forearm	0.718	0.807	0.679	0.834	0.193	0.321	0.282	0.698	0.519
Hand	0.954	0.773	0.632	0.976	0.227	0.368	0.046	0.760	0.665
Wrist	0.956	0.711	0.624	0.970	0.300	0.376	0.044	0.755	0.623
Total	0.905	0.759	0.542	0.962	0.241	0.458	0.095	0.678	0.578

Fig 5: Performance matrices of radiographic scans across different upper body regions.

1.4 Interpretation

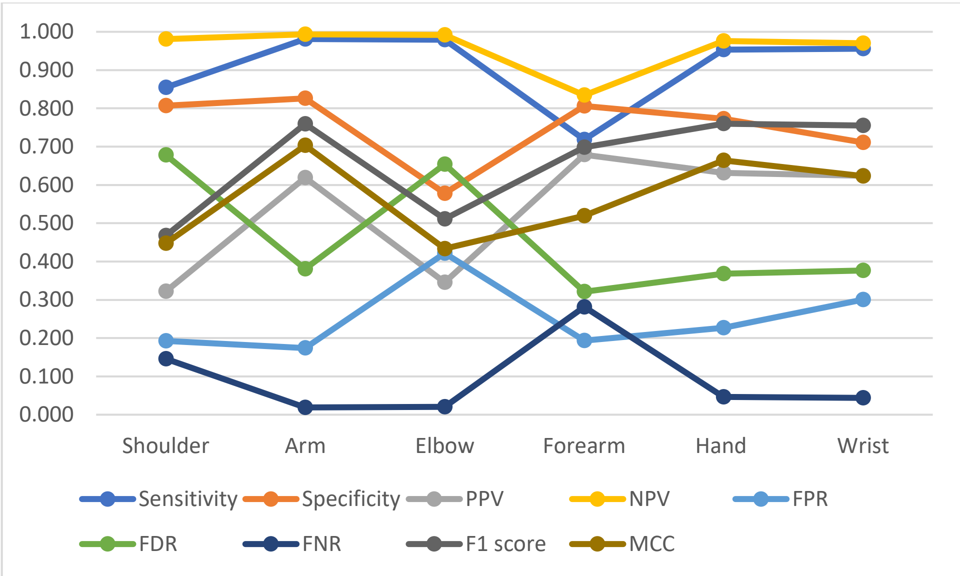

Fig 6: Graphical representation of performance matrices of radiographic scans across different upper body

The analysis of the AI solution's performance in detecting fractures in various upper body MSK scans reveals several key insights. The AI demonstrated high sensitivity and NPV across all regions, indicating its reliability in correctly identifying fracture cases and ruling out non-fracture cases. High sensitivity across most upper body MSK scan regions, particularly excelling in the arm (0.981) and showing consistent reliability in detecting true positive fracture cases. Specificity was more variable, with the arm performing well (0.826) and the elbow showing lower specificity (0.578), indicating a higher rate of false positives in the elbow region. The PPV was highest for the arm (0.619) and lowest for the shoulder (0.322), reflecting a need for improvement in the AI's confidence in predicting shoulder fractures.

NPV was consistently high, particularly for the arm (0.993), suggesting that the AI is reliable in ruling out fractures when it predicts none. FPR and FDR were higher for the elbow and shoulder, indicating more frequent incorrect fracture identifications in these regions. Conversely, the arm had the lowest rates, affirming robust performance in minimizing false positives. The high false positive rate in the elbow and shoulder region can be attributed to several factors, including anatomical complexity leading to more ambiguous imaging interpretations and the presence of osteophytes or calcifications that can mimic fracture lines. These challenges contribute to the AI's difficulty in accurately distinguishing between true fractures and anatomical variants or benign findings.

FNR was notably high for the forearm (0.282), suggesting that the AI might miss more fracture cases in this region. This could be attributed to several factors such as the region's complex anatomy (involving multiple bones and joints), variability in imaging quality, and potential technical limitations of the AI algorithm. Forearm fractures can be subtle, easily obscured by overlapping structures or artifacts, leading to missed detections.

F1 Score, which balance both precision and recall, was highest for the arm (0.759), indicating a reliable performance in diagnosing arm fractures. The MCC was highest for the arm (0.703) and lowest for the elbow (0.434), further supporting the superior performance of the AI in detecting arm fractures.

The variability in performance metrics across different body parts underscores the need for targeted improvements, especially for the elbow and forearm, to enhance the AI's diagnostic accuracy and consistency across all regions.

2. Concordance and Discordance Analysis

2.1 Data Preparation: The dataset was obtained in an Excel sheet format, containing 1171 reports of patients including 334 scans without AI and 837 scans with AI. The dataset contains the following columns: group, MRN, Accession number, body part, impression, JR interpretation with and without AI assistance and SR interpretation with and without AI assistance.

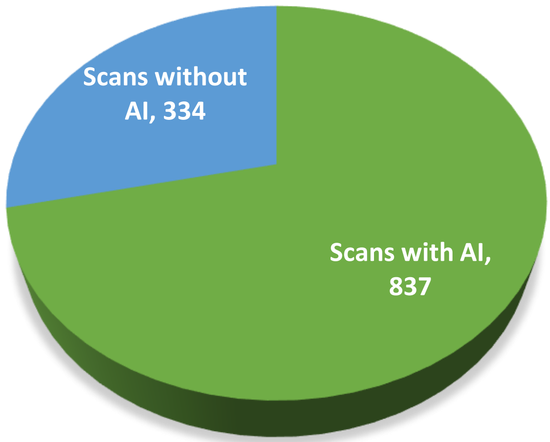

Fig 7: Total number of radiographic scans across different upper body regions with and without using AI

GT Marking: Each scan was reviewed and annotated with a ground truth label against each JR and SR column with and without AI. If a fracture was present, a '1' was marked in the ground truth column; if no fracture was present, a '0' was marked.

2.2 Data analysis

Each scan was evaluated for the presence or absence of fractures by both JR and SR, both with and without AI assistance. Two tables were prepared in MS excel based on whether AI was utilized during interpretation. For each scenario (with AI and without AI), GT labels were assigned after thorough review of both JR and SR reports.

Subsequently, the dataset was analyzed to determine concordance and discordance rates. Concordance was identified as in cases where both JR and SR provided identical interpretations (either both indicating a fracture or both indicating no fracture). Discordance was noted when JR and SR provided conflicting interpretations (one indicating a fracture while the other did not).

Upon analysis, the results indicated a higher concordance rate when AI was utilized compared to interpretations without AI. Specifically, with AI assistance, the concordance rate was found to be 96.3% (806 patients), with a discordance rate of 3.7% (31 patients). In contrast, without AI, the concordance rate was slightly lower at 94.32% (315 patients), with a discordance rate of 5.68% (19 patients).

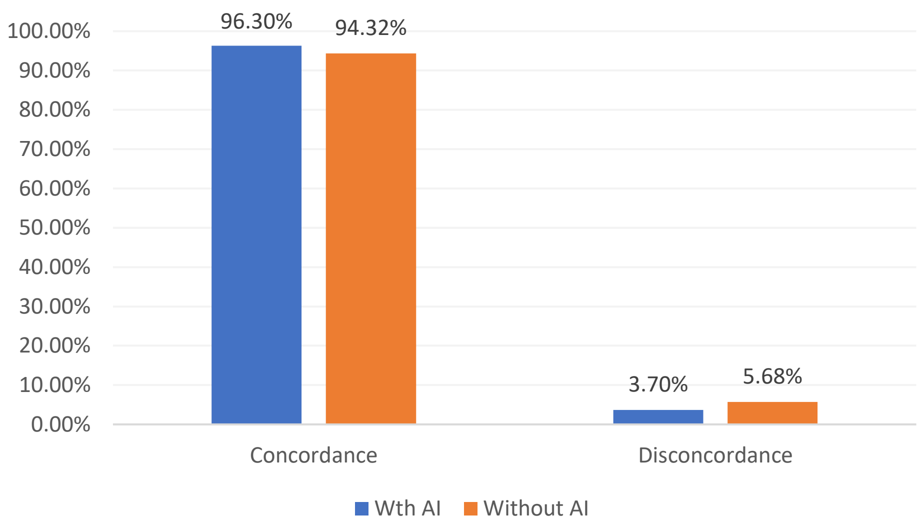

Fig 8: Concordance and discordance rates between JR and SR with and without using AI

To assess the statistical significance of these findings, a Chi-Square test was conducted. The Chi-Square test evaluates whether the observed differences in concordance and discordance rates between AI and non-AI scenarios are statistically significant. The calculations for the Chi-Square test involve determining the expected frequencies under the following hypothesis:

- Null Hypothesis (H_0): There is no significant difference in discordance rates between JR and SR with AI and without AI.
- Alternative Hypothesis (H_1): There is a significant difference in discordance rates between JR and SR with AI and without AI.

To facilitate the Chi-Square test, a contingency table was constructed as follows:

	Concordance	Disconcordance
Wth AI	806	31
Without AI	315	19

Fig 9: Confusion matrix for Chi-Square test

For α=0.05 and df=1, the critical value of Chi-Square is approximately 3.841.

With AI: χ^2_{AI}= 0.000362

Without AI: $\chi^2_{non\text{-}AI}$= 0.000003

2.3 Interpretation

Both calculated Chi-Square values are much smaller than the critical value of 3.841. Therefore, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in discordance rates between interpretations with and without AI assistance.

The study concludes that while AI improves concordance rates slightly, it does not impact discordance rates significantly. In other words, the presence or absence of AI during interpretation does not lead to significantly different levels of disagreement between the human interpretations. The Chi-Square test supports this conclusion by showing that the observed differences in discordance rates between AI and non-AI scenarios are likely due to random chance rather than a systematic effect of AI on the disagreement between JR and SR interpretations.

This interpretation underscores the importance of statistical testing to ensure robustness in comparing different methodologies or technologies in medical image interpretation, helping to inform decision-making regarding AI integration in clinical settings.

Section 4: Recommendations and Conclusion

Recommendations

- From the comprehensive analysis encompassing both the performance of AI in fracture detection compared to radiologists and its impact on the concordance and discordance between JR and SR, several recommendations emerge to enhance the clinical application and efficacy of AI in MSK imaging:
- Algorithm Refinement and Training: It is crucial to continue refining AI algorithms, especially for regions with lower performance metrics identified in the study, such as the elbow and forearm. This refinement should focus on improving sensitivity to detect subtle fractures and specificity to reduce false positives. Techniques like deep learning and CNNs could be explored further to enhance image interpretation accuracy, particularly in complex anatomical areas.

- Integration of Clinical Context: Enhancing AI models by integrating broader clinical context, including patient history and additional diagnostic findings can help mitigate diagnostic errors caused by imaging artifacts or anatomical variants that mimic fractures, thereby improving both sensitivity and specificity across all MSK regions.
- Continuous Validation and Improvement: Establish a framework for continuous validation of AI algorithms using real-world data and clinical feedback. This process should include regular updates and adaptations to ensure that AI systems evolve alongside advancements in medical imaging technology and clinical practices.
- Multidisciplinary Collaboration: Foster collaborative efforts between AI developers, radiologists, orthopedic specialists, and other healthcare professionals. This collaboration can help validate AI outputs, interpret complex cases, and refine algorithms based on diverse clinical insights and expertise.

Conclusion

The study provides valuable insights into the performance of AI in fracture detection across various upper body MSK regions and its impact on the interpretation concordance between JR and SR radiologists.

The comparative analysis of AI and radiologist revealed that AI demonstrates high sensitivity and NPV across most MSK regions, effectively identifying fractures and ruling out non-fracture cases. However, variability in specificity and PPV suggests areas for improvement, particularly in distinguishing between true fractures and anatomical variants, as observed in regions like the elbow and shoulder.

The study of discordance results between JR and SR found that AI contributes slightly to improved concordance rates between JR and SR interpretations. However, statistical analysis indicated that AI does not significantly influence discordance rates, suggesting that the presence or absence of AI during interpretation does not lead to markedly different levels of disagreement between radiologists.

The findings underscore the potential of AI to enhance diagnostic accuracy and consistency in MSK imaging, albeit with room for refinement. By addressing the recommendations outlined, healthcare providers can optimize the integration of AI into clinical workflows, improving overall patient care through more reliable fracture detection and interpretation.

In conclusion, while AI holds promise in revolutionizing MSK imaging, ongoing collaboration, algorithmic refinement, and validation efforts are essential to realize its full potential in clinical practice. By leveraging AI alongside clinical expertise, healthcare systems can strive towards more precise, efficient, and patient-centric diagnostic solutions in MSK healthcare.

References

[1] Brady AP. Radiology reporting—from Hemingway to HAL? Insights Imaging 2018;9:237–46. https://doi.org/10.1007/s13244-018-0596-3.

[2] Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal Radiol 2020;49:183–97. https://doi.org/10.1007/s00256-019-03284-z.

[3] Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. Npj Digit Med 2019;2:111. https://doi.org/10.1038/s41746-019-0189-7.

[4] Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284:574–82. https://doi.org/10.1148/radiol.2017162326.

[5] Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, et al. Deep learning in fracture detection: a narrative review. Acta Orthopaedica 2020;91:215–20. https://doi.org/10.1080/17453674.2019.1711323.

[6] Yoda T, Maki S, Furuya T, Yokota H, Matsumoto K, Takaoka H, et al. Automated Differentiation Between Osteoporotic Vertebral Fracture and Malignant Vertebral Fracture on MRI Using a Deep Convolutional Neural Network. Spine 2022;47:E347–52. https://doi.org/10.1097/BRS.0000000000004307.

[7] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8. https://doi.org/10.1038/nature21056.

[8] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[9] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning 2017. https://doi.org/10.48550/ARXIV.1711.05225.

[10] CARPL.ai - Radiology Automation Simplified n.d. https://carpl.ai (accessed June 20, 2024).

[11] Confusion Matrix - Online Calculator n.d. https://onlineconfusionmatrix.com/ (accessed June 20, 2024).

Eshika B ST

ORIGINALITY REPORT

6%	4%	3%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	d.docksci.com Internet Source	1 %
2	Submitted to Queen's University of Belfast Student Paper	<1 %
3	docplayer.net Internet Source	<1 %
4	Submitted to University of Durham Student Paper	<1 %
5	Senik Matinyan, Pavel Filipcik, Jan Pieter Abrahams. "Deep learning applications in protein crystallography", Acta Crystallographica Section A Foundations and Advances, 2024 Publication	<1 %
6	Submitted to Capital City College Group Student Paper	<1 %
7	Submitted to University of Waikato Student Paper	<1 %
8	Submitted to VinUniversity Student Paper	