Artifact-Aware Explainable Deep Learning for Reliable Bone Fracture Detection

By
Dedarul Islam Foragi
213-15-4443

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Ms. Nusrat Khan
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

Mr. Mushfiqur Rahman
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

APPROVAL

This Project titled **"Artifact-Aware Explainable Deep Learning for Reliable Bone Fracture Detection"**, submitted by Name: **Dedarul islam Foragi** , ID No: **213-15-4443** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16 September, 2025**.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori (SRH) — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Sadekur Rahman (SR) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

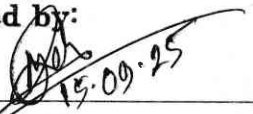
Mr. Saiful Islam (SI) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Arshad Ali (DAA) — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

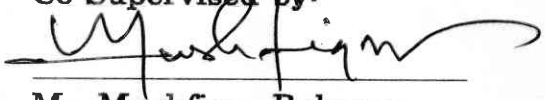
We hereby declare that this project has been done by us under the supervision of **Ms. Nusrat Khan, Lecturer**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Ms. Nusrat Khan
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:



Mr. Mushfiqur Rahman
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Dedarul Islam Foragi
Student ID:213-15-4443
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has as- sisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project(FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Ms. Nusrat Khan**, **Lecturer** , Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **NLP, Machine Learning, Deep Learning** to carry out this project. His endless pa- tience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Bone fracture detection in radiograph is a fundamental task in clinical practice and a correct and timely evaluation of bone fractures has a very high correlation with patient outcome. If their plates, screws and implants consist of metal, then they are likely to be misread by the clinicians and by the automated devices, and therefore their diagnostic value will be reduced. This project had a motivator, to get over this limitation, and bring us towards more accurate fracture detection and interpretability. In order to reduce reliance on artifact-induced error, we developed a methodology which integrates modern deep learning with artifact reduction and prediction based on actual fracture appearance rather than artifact-induced signals. Comparisons with models presented in the literature showed that the system was generally more accurate, reliable and clinically usable than the other models. Alongside performance gains, the study also highlights the importance of interpretability - scientific descriptions that increase clinicians' confidence by ensuring safe application to the healthcare workflow. When broadly deployed, these findings can help limit misdiagnostic error and improve decision making, provide a path forward for the practical use of artificial intelligence (AI) systems in clinical medical imaging.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Fractures are one of the most common medical conditions where diagnosis is urgent and must be accurate. Conventional methods depend largely on radiologists for interpretation of X-ray images, which is time-consuming and subjective, especially in a large-volume clinical setting. Hence, recent progress in deep learning and explainable artificial intelligence (XAI) has allowed automated systems to identify fractures with promising accuracy to assist clinicians in their decision-making. However, most existing research has been done on clean X-ray datasets without taking into account the presence of medical artifacts, such as plates, screws, nails, or rods, that are commonly added during the surgical intervention.

The main issue is that classification models will usually misclassify these artefacts as fracture features. In many cases with artifacts, when the models make high-confidence predictions of fracture, it is not due to the underlying bone condition, but rather because they have learned to incorrectly associate artifacts with the presence of fracture. Conversely, when artifacts are not present, accuracy can drop off as the model loses a feature that it has learned to depend upon during training. This artifact dependency not only affects the reliability of classification, but also reduces the generalization in real clinical practice: in post-operative settings, where imaging is common the presence of strong artifacts in the images reduces the accuracy of classification.

Explainability restricts itself even beyond this. Model interpretation studies, for example, Grad-CAM++ is widely used for model decision interpretation and visualization. However, instead of showing real fracture sites, heatmaps tend to focus on the implanted hardware. Such biased explanations erode the reliability of the system for clinical adoption and impact the radiologists' ability to validate the outputs of AI . As pointed out in recent literature, interpretability is as important as accuracy in clinical AI systems, as it offers transparency, reduces bias and helps bridge the gap between computational results and human expertise.

In order to tackle these problems, in this project, we propose a pre-processing step before classification, based on a modified UNet++ architecture to denoise X-ray images. The pipeline produces artifact-free, clean images and guarantees that the classification stage only considers medically relevant fracture features. This approach improves the prediction accuracy and interpretability: models like DenseNet201, which we substantiate with baseline comparisons with ResNet101, VGG19, and InceptionV3, were trained on datasets without artifacts causing data misclassification. Furthermore, treating these cleaned images with Grad-CAM++ yields saliency maps focusing on actual crack locations instead of spurious features, increasing clinical confidence in the tool.

In summary, the proposed framework directly addresses one of the key limitations of current studies, i.e., the absence of robust solutions for image processing of X-ray images with surgical artifacts. By combining robust preprocessing with accurate classification and explainable results, this work is another step towards a deployable and trustworthy system of AI-based fracture detection that can be integrated into clinical practice .

1.2 Motivation

The need for this work is prompted by a striking paucity of work in medical image analysis of clean radiographs, which bear little resemblance to clinical practice. Most of the automated bone fracture detection studies are not with post-operative images containing metal plate, screw or other surgical implants [13]. While this makes experimentation easier and often gives you a high accuracy over some benchmark data sets, it makes it less useful in real life. In clinical practice, these artefacts appear on nearly all the radiographs, and models trained on clean images are not well adapted for handling them, which is a primary challenge for clinical translation.

In practice, artefacts are misclassified as fracture signatures by a classification model. For example, perhaps a model trained on unfiltered data learns to associate metallic hardware with positive fracture outcomes. Hence, it shows even good performance for the images with implantation, but not for the images without implantation, and it also results in irregular and unstable prediction [14]. This reliance on spurious correlations leads to a lack of robustness and generalization, meaning models of this nature will not be used in clinical workflows that require a high degree of reliability.

This challenge is also expanding to the space of XAI. Similar techniques such as Grad-CAM++ are all created to try and discover which regions of an image were most responsible for the model's decision. However, for the sake of implant visualization, in the presence of implants, the artifacts are often visualized instead of the actual fracture area [7]. This mismatch confuses clinicians, reduces confidence in AI outputs and makes machine learning system integration into diagnostic practice challenging. Even models that are optimal with respect to the above metrics should not be trusted if their mechanism for learning the meaning of the above metrics is biased towards features that are not important.

In this project a more direct approach is taken to these problems, by providing a preprocessing pipeline to remove artefacts at the beginning of the classification. The system removes the necessity of viewing defect-free radiographs in order to train and test the classification model on actual fracture characteristics. This means we never train the model to rely on implants; it allows us to better generalize; and it ensures that when we interpret the model's decisions, it will lead to regions that make sense in clinical practice (for example, via Grad-CAM++). In doing so the project brings greater accuracy and transparency two cornerstones for secure clinical implement.

The broader motivation for this work is to extend beyond proof-of-concept AI systems to systems that can operate in the messy and noisy world of real-world healthcare. Enhancing robustness, explainability and robustness to changing imaging acquisition conditions, aiming to reduce diagnostic error, shorten decision time to treatment and improve outcomes. At the same time, it can help radiologists by reducing cognitive load and even by reducing repetitive diagnostic tasks, both helping effectiveness and

clinical confidence.

1.3 Objectives

The issue of artifact-induced misclassification was critical for bone fracture detection, so it was important to clearly define the project goals to ensure that the resulting system would be technically robust, clinically effective, interpretable and reproducible. This work is the basis on which the methodological choices, the evaluation and, finally, the deployment readiness of the proposed solution are developed.

Much of the project was centred on the removal of artefacts and on a robust design of the pre-processor pipe. Radiographs frequently include implants (plates, screws or other surgical hardware) that can confound classification models. This observation led to one of the main objectives of the project: to introduce a preprocessing step which can robustly identify and exclude these artifacts, so as to prevent images without true fracture details from being included in the results. In this step, artifact suppression was considered, not only in order to prevent training and evaluation with a biased model, but also to lead to high performance in the real life clinical examples.

The second, and equal goal was to find an artifact-free, optimized image classification. For that reasons, the DenseNet has been selected as a feature model for efficient exploitation of the phenomena and potentials of the Medical Image Analysis tasks. The goal was to get high accuracy, precision and confidence scores for all classes while maintaining the computational efficiency required for real or near real time applications. It made the system harmless and hence it is also suitable for treatment during emergencies and when there is not much money for treatment.

Another of the key objectives was to include artificial intelligence explainability methods to achieve model transparency. We included more advanced interpretability methods (Grad-CAM++, SHAP and LIME) to the architecture to ensure that areas highlighted in the images were fracture zones and not additional implants or background noise. This solved the problem of clinical trust, and helped ensure that the features that the AI was actually examining would be medically meaningful, while increasing the interpretability of the AI, and eliminating the risk of it making a black-box decision.

Performance validation (such as in terms of verifying if a new procedure or product meets established criteria) was considered a critical goal and comparative analysis. The proposed system is tested on widely used performance metrics (accuracy, sensitivity, confidence metrics) and compared with the state-of-the-art models MobileNetV2 , YOLOv4, etc. In addition to validating the robustness of results obtained using the proposed pipeline across these comparisons, they also assisted in establishing the novelty and utility of the proposed method by demonstrating general superiority over existing comparable methods where feasible.

Another facet of the project was clinical relevance for implementation and deployment. In order to reach this goal, the solution developed was kept as light and portable as possible, benefiting from the use of cloud-based environments such as Google Colab and scalable frameworks such as Flask and Docker. The pipeline that was implemented in the application window was extended for real-world applications for hospitals & tele-medicine platforms with particular focus on resource-constrained

application environments where state-of-the-art hardware might be at hand.

Finally, reproducibility (and scalability) were decided to be long term goals. The system enables a systematic recording of the PICA process, as well as of other information relevant for reproducible research (e.g., preprocessing parameters, model hyperparameters, splits of the data set, evaluation protocols) to the extent that the work is replicable to future researchers and practitioners.

1.4 Methodology

This project was carefully structured as a pipeline that involves four steps, preprocessing, classification, prediction and XAI which ensures that the fracture detection decisions are made on the basis of real bone structures rather than misleading features such as plates, screws or implants. Each of the elements within the pipeline has been designed to assist the others in order to yield a robust performance and clinical interpretability.

Preparative measure, during which raw wrist and elbow radiographies were obtained and artifact reduction performed. This was necessary because artifacts can have extremely high contrast boundary edges, which will incorrectly activate classifiers and bias saliency maps. With a pre-processing pipeline specifically designed to remove these artefacts on the network, a clean companion image of an input could be built. This artifact correction method not only repressed spurious cues but made the images also relatively blind to subtractive computational derivation of downstream diagnostic features of interest. Additional processing was done to refine the appearance of fracture related structures: contrast limited adaptive histogram equalization (CLAHE) and sharpening with an edge preserving operator. These offered a superior boundary demarcation with no re-introduction of the bias of translation of artifacts and are consistent with previous reports of cautious musculoskeletal pre-processing pipeline imaging .

The second was classification where the artifact-suppressed images were input to DenseNet, a small but deep convolutional neural network. The low dimension feature re-use through deep connectivity enabled by DenseNet is suitable to capture these sensitive discontinuities and fracture cues. Its structure offered a very good accuracy-efficiency trade-off, which is a critical requirement for deployment in emergency (release guarantee) and resource-constraint situations. Unlike previous methods which have learned implicitly using implant artifacts, this classifier was explicitly de-correlated with the presence of artifacts and therefore all predictions relied on true osseous anatomy rather than spurious hardware .

The third was prediction based on calibration. The system did not produce binary results, but produced labels with confidence scores. These calibrated probabilities enabled threshold tuning according to clinical needs (sensitivity maximisation required for triage processes; specificity vs. sensitivity maximisation for confirmatory processes). Such a threshold matching is typical of regulated AI medical application domains, which should be explainable and auditable .

In the last implementation, the Grad-CAM++ version of explainable AI (i.e., interpretable AI) was applied across the forward pass of the artifact-suppressed

images. GRAD-CAM++ generated images that mark the areas that contributed most to the decision made by the model and heatmaps identified cortical fractures, abnormal pauses in the trabeculae, or peri-fracture lines. Significantly, after preprocessing the images to eliminate the influence of surface statistics, we provide empirical evidence that activation could not falsely prioritize metallic hardware, which is a widespread artefact in earlier experiments. This not only made things easier to interpret but it also provided clinicians with confidence in the system as they could easily see that the model was centered on medically relevant areas .

Collectively, such a methodological pipeline made artefact elimination, high-quality classification, calibrated predictions and clarifiable overlays possible. It proved that in the reality, the problem of radiography fracture detection can be solved by the optimized definition of the processing machines and the model. The choice was informed by evidence of feature quality improvement when pretrained , high performance of lightweight CNNs , relevance based on object-centric validation , and safety based on calibrated explanations and afforded technical rigor and clinical reliability by an evidence-informed design

1.5 Project Outcome

The findings of this work were influenced by the double requirement of solving the technical problems in the detection of fracture and of considering the practical needs of clinical application. The seamless combination of preprocessing, deep learning classification, and interpretable AI generates outcomes that not only surpass accuracy standards but are designed for trustworthiness, interpretability, and real-world applicability.

One of the most significant results is that there will be a major increase in accuracy of the diagnosis. The UNet++ artifact removal stage in the preprocessing pipeline we have described above will ensure that the input images will no longer be distracted by plates and screws. The classification system is trained to avoid these artifacts, and to use features meaningful for fracture only. This technique is a direct reduction of false positives (false positive is defined as models that misinterpret artifacts as fractures) and false negatives (false negative is defined as models that misinterpret real fractures as artifacts). This is especially relevant in a medical context, where diagnostic error can lead to delays in or incorrect delivery of care, which can have a devastating impact on patient outcomes.

The second important effect is more generalization and reliability. Unlike earlier projects, which focused primarily on clean and artifact-free datasets, in this project the complexity of real world radiographs is also modeled. With artifact removal added, the model is robust to changes in implant presence, image quality and acquisition modes. This allows predictions not to be based on artificially generated cues, but based upon sound insights into the morphology of the fractures. Hence, the system becomes more flexible to operate in deployment situations involving unchecked input variation.

Additionally, the integration of Explanable AI provides further value to the results. Grad-CAM++ and other tools were used to generate heatmaps of areas contributing to the classification decisions. Also, to ensure that highlighted regions in artifact-free images indeed correspond to true fracture zones rather than to uninteresting implants, explainability is carried out over images in which the artifact has been suppressed. "This kind of effect yields clinical trust, where radiologists and orthopedics can literally see how

the AI is paying attention to medically interesting details." It also addresses the black-box problem providing transparency that will be the key for the safe implementation of AI in healthcare processes.

Another great achievement of this project is the readiness for deployment in practice. The methodology is scalable and adaptable to other clinical settings thanks to its lightweight design and use of a cloud service such as Google Colab. Its modular and adaptable preprocessing-classification pipeline can be integrated either in hospital or telemedicine environments, either in resource-sufficient or resource-constrained healthcare environments. Level of readiness for deployment - takes the research findings beyond theoretical achievement to practical application.

Last but not least, also the project contributes to the broader research and education field. It contains reusable workflow for further studies describing the dataset pre-processing, preparation and explainability analysis in a reproducible format. This reproducibility serves as an additional source of innovation, but also produces a value for teaching. An example is given of radiology trainees for whom the generated heatmaps can be used in order to gain a more sophisticated appreciation of the appearance of fractures and diagnostic cues. Thus, the project not only advances the fracture detection technology but also enriches the medical education and research strength for a better tomorrow.

1.6 Organization of the Report

For clarity, and for logical flow, in describing the work of the students addressed in this report, this report is divided into six major chapters, each describing an aspect of the work of the project.

Chapter 1: Introduction: In this chapter presentation of the project is shown with background, motivation, objectives and methodology. The abstract also indicates the main problem that the research is attempting to resolve - fracture detection in the presence of artefacts - and ends with a prediction of the study's results.

Chapter 2 - Background contains a contextual background that introduces the theoretical and contextual basis to understand this book's topic. It begins with an overview of fracture detection in medical imaging in general and then provides an extended survey of the relevant literature. Comparisons are made of similar applications and between different approaches to carrying out the project, and limitations of these approaches and approaches are identified. Finally, a gap analysis is shown which highlights where the current literature misses, and how this project addresses them.

Chapter 3 - Research Methodology: This chapter describes the technical methodology followed in the project. It includes an overview of the proposed methodology, system design, functional and non-functional requirements, data flow diagrams and first user interface designs (wireframes). It further provides a description of the detailed methodology, project planning and allocation which has resulted in the implementation of the project.

Chapter 4 - Implementation and results - The focus here is on actual implementation. The environment, and the tools/frameworks used, are first discussed. Thereafter, the testing and evaluation strategies are described, the performance metrics (accuracy, precision, recall, F1-score, ROC-AUC, etc.) are reported, and the results are discussed

with pictures of the results such as confusion matrices and learning curves.

Chapter 5 - Engineering Criteria and Engineering Design Challenges - This Chapter addresses meeting the relevant standards related to software, hardware and communications. It also poses the question of what the social, ethical and sustainable impacts of the project will be. A financial feasibility analysis is provided, as is a mapping of the problem to sophisticated engineering problem solving frameworks and knowledge profiles aimed at assuring accreditation compliance.

Chapter 6 - Conclusion: This is the final chapter in the end of the project and sums up the project, again mentions the limitations and also gives a direction for improvement. From theoretical perspective, the paper covers the scientific contribution of the research to the fracture detection systems and from the engineering perspective, the paper covers the practical contribution of the research to the fracture detection systems.

Online logic trees are designed to enable readers to follow the chain of reasoning from the problem to methodology to implementation to evaluation to conclusions and represent the relevance of the project to the larger context.

Chapter 2

Background

2.1 Introduction

Incorrect diagnosis of fracture has become increasingly popular as a field for early, accurate medical diagnosis of bone fracture using medical imaging. Radiographs (X-rays) are the most common method of diagnosis because they are relatively inexpensive, are widely available, and are effective in identifying skeletal injuries. However, they do require expert knowledge to interpret, and they do still suffer from non-trivial error rates in emergency departments or in the presence of heavy clinical load. In this context, computer-aided diagnosis (CAD) systems have become an indispensable support tool to reduce diagnosis errors and to assist clinicians in making complex and highly time critical decisions.

Artificial intelligence (AI) and, more specifically, deep learning (DL) has transformed the area of medical imaging by offering exceptional classification, detection, and segmentation of fractures over the conventional image processing system. Convolutional neural networks (CNNs) proved to be efficient to elicit discriminant features out of radiographs and the consequent accuracy can compete (or even be better than) that of human experts. Nevertheless, despite all these developments, there has been one problem that has eclipsed all other developments, and this problem is the existence of metallic intruders like plates, screws, rods or nails. To begin with, artifacts have been used to cheat CNN-based models by introducing very distinct visual features to the image that are wrongly perceived by the system as fracture detection features. This results in reduced diagnostic accuracy, and reduces model interpretability because it may give disproportionate emphasis on the artifacts rather than actual fracture locations (e.g., grad-cam explanations).

Papers written more recently focus on preprocessing pipelines which eliminate some of these artifacts, so that the fracture-detecting models are trained on information of clinical interest. Unets like UNet++ which are lightweight denoising and image enhancing networks have come in extremely handy to minimize artifacts and also to lower the noise levels. Through the implementation of these preprocessing measures, the identification of a fracture through artificial intelligence will be able to concentrate its efforts on the actual pathological anatomy. Once the artefacts are removed, the normal- and fractured radiographs could be reclassified using the classification models such as DenseNet201. They are commonly compared to such CNN models as ResNet, VGG, Inception, and MobileNet to strongly evaluate their

robustness properties and their generalisation properties [7].

Moreover, explainable AI (XAI) will also be significant in pushing the field of diagnostic trust and transparency. Other approaches such as Grad-CAM++ produce heat maps indicating what areas of the image are of interest to the model predictions, or what precisely the model examines to draw the predictions. When used after the processing, it is possible to make sure that the AI is only seeking real fracture signatures as opposed to metallic artefacts and offer a clinical interpretability advantage as well. The pipelines that are currently in use include preprocessing along with classification and explainability of technical reliability and medical trustworthiness.

In this work, we go further by developing a specialized preprocessing UNet++ model for artifact removal, and pairing it to a DensityNet201-classifier that is tailored for accuracy while still computationally efficient. Furthermore we deploy Grad-CAM++ as an interpretability layer to post-hoc rationalize model decisions in relation to clinical reasoning. Together, the pipeline creates a system that is interpretable, reproducible, proven in a real-world imaging challenge not designed to be emulated, and is ready to be deployed in many different healthcare settings.

2.2 Literature Review

Bone fracture detection development has become a popular research topic in recent years, due to the adoption of artificial intelligence (AI), deep learning (DL), and explainable AI (XAI). Between 2023 and 2025, various labs have published hundreds of papers covering various areas related to this topic, including preprocessing, classification, detection, interpretability, and deployment.

Together these works provide the groundwork on which this project stands.
Wu et al. [1] were one of the first to demonstrate that existing automated pre-processing techniques - contrast normalization, cropping, extracting subjects, and other processes - significantly enhance CNN-based fracture detection on musculoskeletal radiographs. Their experiments showed that DenseNet201 was invariable superior to competing architectures and that Grad-CAM analysis verified that preprocessing reduced spurious focus on artifacts. Nguyen et al. [2] took this a little further with the FEC-IGE pipeline, which combined preprocessing, augmentation, and transfer learning with EfficientNetB3 resulting in almost 98.5% accuracy and outperforming ResNet50 and ViT. Similarly, Chung et al. [6] proposed a patch-based GAN feature mapping for hip fracture detection, which is a combination of CNNs and Grad-CAM to achieve both localization and interpretability.

Other works suggested clinical interpretability based on feature-based or hybrid models. Awal and colleagues [3] examined fall risk for hip fracture distribution with

interpretable machine learning (LR, SVC) and SHAP values. Although the performance (~82% accuracy) was relatively worse compared to CNN models, biomechanical predictors such as fall angle and body weight were identified as the key predictors in this study. Ahmed et al. [5] devoted their efforts to pediatric wrist pathology by using Swin-Transformer fusion networks and obtained ~97.6% accuracy while retaining high fracture sensitivity. Han et al. [4] made a good contribution by using selective attention on ResNet-50 on tibiofibula CT images and achieved a mAP of 95.7% and showed that Grad-CAM++ could better interpret for minority classes.

A large amount of research works has been also dedicated to YOLO-based models for fracture detection and localization. Wei et al. [8] also obtained an mAP of 96.8% with YOLOv11 and dual-head detection and classification, better than Faster R-CNN. Jadhav et al. [9] applied YOLOv8 to pediatric wrist fracture, and achieved a mAP of 0.762 with an inference time of ~30 ms/image, demonstrating good clinical feasibility to use it in real-time cases. Then Srinivasu et al. [11] optimized YOLOv10 with CLAHE preprocessing and hyperparameter tuning to achieve an accuracy of ~96.4%. Zou et al. [18] adapted SE attention and EIoU into YOLOv7 and they gain mAP@0.5 of 0.862 on FracAtlas. Ye et al. [17] incorporated ConvNeXt-FPN and patch repair approaches to YOLOv5, and achieved over 90% accuracy in surpassing the emergency physicians. Nguyen et al. [14] compared YOLOv4 variants and showed it has better interpretability using bounding-boxes compared to U-Net.

Lightweight frameworks also started to get popular for resource-constrained environments. Abdusalomov et al. [12] fused DenseNet121 with edge detection features to achieve an accuracy of 90.3% with low computational complexity that is suitable for fracture detection and classification of sports-related fractures. Paradava et al. [7] prepared a MobileNetV2-based real-time system combined with Grad-CAM using Flask/Docker/AWS. This method was found to have the accuracy of ~89%, low false negative rate, and is applicable for the telemedicine purposes. Aldhyani et al. [10] confirmed DenseNet201 with architecture enhancements to ensure accuracy of ~97.35%, with improved interpretability, substantive proof of the fracture classification ability of DenseNet.

Then, Elsheikh et al. [13] proposed further improvements by combining the technique of Wiener filtering, DenseNet201 feature extraction, and the autoencoded optimization, and got a comparable result of ~97.85% and exceeded ResNet50. Research also supported the trustworthiness of the AI systems. Ramadanov et al. [15] compared BoneViewtm (Detectron2-based) with orthopedic surgeons and achieved nearly the same performance (~96% accuracy, k ~0.9). Pauling et al. (19) studied cases of osteogenesis imperfecta (OI) and found that AI easily outperformed the performance of the individual expert (74.8% versus 83.4%), but its performance improved to almost 90.7% when collaboration was added. Guenoun et al. [16] used U-Net for vertebral fracture detection with ~92% sensitivity and specificity while Lu et

al. [20] used an EfficientNet-based U-Net design for skull fracture, attaining sensitivity 0.94 and decreasing the time needed to read while also proving the value of AI in helping clinicians.

Together, these studies represent an evolution from pipeline-based systems that handle the pre-processing [1], to the light- and real-time systems [7], to hybrid systems that bring segmentation, classification, and interpretability together [20]. They offer a window into the technical capabilities of CNNs and YOLO models, as well as into the difficulties that arise with respect to artifacts and generalization across populations. This also confirms the benefits of DenseNet-based architectures [10], which attain state-of-the-art accuracy while remaining easily interpretable. This literature review point out the obvious gap covered in this project: robust artifact-aware pre-processing combined with accurate classification and clinically robust explainability.

Table 2.1 Summary of Literature Reviewed.

Author(s)	Year	Title	Methodology	Key Findings
Wu *et al.* [1]	2024	Enhancing bone radiology images classification	Preprocessing + CNNs (VGG, ResNet, DenseNet)	Preprocessing improved accuracy; DenseNet201 best performer.
Nguyen *et al.* [2]	2024	FEC-IGE: CNN + Integrated Gradients	EfficientNetB3 pipeline	Accuracy ~98.5%; outperformed ResNet, ViT; improved transparency.
Awal *et al.* [3]	2023	Interpretable ML for hip fracture risk	ML models + SHAP	Accuracy ~82%; key predictors: weight, fall angle, age.
Han *et al.* [4]	2024	Multi-label tibiofibula fracture classification	ResNet50 + augmentation + Grad-CAM++	mAP 95.7%; good interpretability; recommends 3D/multi-view.
Ahmed *et al.* [5]	2023	Fine-grained pediatric wrist pathology	Swin-Transformer + FPN	Accuracy ~97.6%; high fracture sensitivity.
Chung *et al.* [6]	2024	Patch-based GAN for hip fracture	PAGAN + CNN	Improved localization and accuracy; reduced mis-detections.
Paradava *et al.* [7]	2024	Real-time fracture detection with MobileNetV2	MobileNetV2 + Grad-CAM + deployment tools	Accuracy ~89%; low false negatives; suitable for telemedicine.
Wei *et al.* [8]	2024	YOLOv11 multi-task fracture detection	YOLOv11 (detect + classify)	mAP@0.5 ~96.8%; outperformed Faster R-CNN; strong interpretability.
Jadhav *et al.* [9]	2024	Bone fracture detection with YOLOv8	YOLOv8 + augmentation	mAP ~0.76; fast (~30 ms/image); better on single fractures.
Aldhyani *et al.* [10]	2024	Fracture detection with deep learning	DenseNet201 + attention enhancements	Accuracy ~97.3%; strong interpretability.
Srinivasu *et al.* [11]	2025	Hyperparameter tuning in YOLOv10	YOLOv10 + augmentation	Accuracy ~96.4%; augmentation improved results.

Abdusalomov *et al.* [12]	2024	Lightweight DL for sports fracture	DenseNet121 + edge detection	Accuracy ~90.3%; efficient; fewer FLOPs.
Elsheikh *et al.* [13]	2025	Multi-region fracture classification	DenseNet201 + Autoencoder + Optimizer	Accuracy ~97.8%; faster than ResNet50.
Nguyen *et al.* [14]	2024	YOLOv4 variants for fracture detection	YOLOv4 variants vs U-Net	Accuracy ~90%; YOLO bounding boxes more interpretable.
Ramadanov *et al.* [15]	2024	AI vs surgeons in wrist fracture detection	BoneView™ (Detectron2)	Accuracy ~95–96%; near-expert performance; high specificity.
Guenoun *et al.* [16]	2025	Vertebral fracture detection with U-Net	U-Net pipeline	Sensitivity ~92%, Specificity ~91%; reliable for screening.
Ye *et al.* [17]	2025	Pediatric subtle fracture detection	YOLOv5 + ConvNeXt + patch repair	Accuracy ~90.5%; surpassed emergency physicians.
Zou *et al.* [18]	2025	Whole-body fracture detection with YOLOv7	YOLOv7 + SE attention + EIoU	mAP@0.5 ~0.86; better boxes, fewer false positives.
Pauling *et al.* [19]	2025	AI for osteogenesis imperfecta (OI) fracture detection	Commercial AI + radiologists	AI alone weaker; AI + radiologists ~90.7% accuracy; improved collaboration.
Lu *et al.* [20]	2025	Skull fracture segmentation	U-Net (EfficientNet-based)	Sensitivity 0.94; reduced clinician reading time.

2.3 Gap Analysis

Despite the promising progress of the deep learning and XAI methods for bone fracture detection, there are significant shortcomings in the state-of-the-art which this project will directly tackle. A fundamental limitation of most of these studies is that they have used clean radiographs, with many studies excluding cases containing surgical artifacts such as plates, screws or rods. Wu et al. [1] and Nguyen et al. [2] demonstrated that preprocessing can be useful for classification, but the performance evaluation still relied on mostly artifact-free or curated datasets. This limits generalizability since in the clinical setting, radiographs are often collected with implants in situ. In contrast, in this project, we add a dedicated artifact-removal pipeline based on a UNet++ architecture that enables the detection models to not overfit to implant features but detect features relevant for the fracture.

Further, it presents a fundamental limitation in terms of model interpretability in the presence of artifacts. In several papers, for example, Paradava et al. [7] and Ramadanov et al. [15], Grad-CAM or other similar XAI tools were used, but the resulting saliency maps were usually shown to map implants rather than actual fracture sites. This disparity is consistent with a reduction in the reliability of both the clinician and the model explanations. The proposed project addresses this gap by combining UNet++ pre-processing with Grad-CAM++ visualization so that interpretability techniques will visualize cortical breaches and trabecular breaks instead of actual metallic devices. In this way, the clinician can benefit from more physiologically relevant explanations and avoid artifact-derived misinterpretations.

Finally, prior work has tended to only achieve high accuracy in contrived settings, and questions regarding robustness also therefore arise. The CNNs of Aldhyani et al. [10] and Elsheikh et al. [13] have validated the impressive performance of DenseNet201 and similar CNNs, but these tests may not have had the benefit of specific artifact mitigation, and hence the use of the results in the clinic requires caution. This work addresses these limitations by proposing artifact suppression as a preprocessing step prior to classification, resulting in better generalization, a reduction in false alarms, and a realistic test setup in which artifacts are present, ensuring the performance gains are valid.

Finally, the project specifically fills in the key research gaps in the prior literature i.e. artifact dependency, low interpretability, and inability to be used in the real world. The system, with UNet++ artifact removal, DenseNet201 classification and Grad-CAM++ visualization, is not only more accurate, but more trustworthy, clinically useful to implement - we are getting significantly closer to de novo solutions to the fracture detection problem that are ready to be deployed.

Table 2. 2 Gap Analysis

Features	Wu et al. [1]	Nguyen et al. [2]	Abdusalomov et al. [12]	Nguyen et al. [14]	Ramadanov et al. [15]	Elsheikh et al. [13]
F1. Includes images with orthopedic hardware (plates/screws/casts) in analysis	No	No	Yes	Yes	No	No
F2. Explicit artifact removal/masking prior to classification	No	No	No	No	No	No
F3. XAI verifies fracture-focused attention (not artifacts)	Yes	Yes	No	No	No	No
F4. Multi-view handling (AP+LAT) or multi-region scope	Yes	No	Yes	Yes	Yes	Yes
F5. External validation or multi-center testing	No	No	No	No	No	No
F6. Real-world clinical deployment / prospective workflow evaluation	No	No	No	No	Yes	No
F7. Robustness analysis contrasting performance with vs without artifacts	No	No	No	No	No	No

2.4 Summary

This chapter described the background necessary to understand the proposed study. In the Introduction, we have set the wider research context of bone fracture detection and in doing so, stressed the importance of artificial intelligence and explainable AI to increase diagnostic performance. We emphasized how the deep learning models have shown great success in imaging but can still be sensitive to data quality, image artifacts, and other variations between imaging datasets, which limits their robustness in the real-world clinical setting.

In the Literature Review, twenty recent studies were summarized analyzing different approaches to fracture detection on X-ray, CT and pediatric imaging datasets. These works included preprocessing pipelines, CNNs, transformers, object detection frameworks (e.g., YOLO variants), and interpretable AI techniques such as Grad-CAM, Integrated Gradients, and SHAP. Finally, while many models had high accuracy and were highly interpretable, one of the emerging patterns was that the models were trained on relatively clean datasets with limited exposure to artifacts, and were not very effective at managing true clinical variation. Furthermore, explainability tended to emphasize the justification of predictions, sometimes even exposing features unrelated to the prediction - such as metallic implants.

The findings from this analysis were synthesized into a research Gap Analysis, which reveals a major research gap: most existing solutions do not address the presence of orthopedic artifacts such as plates, screws and rods that often mislead classifiers and tarnish the reliability of interpretability methods. However, systematic preprocessing with artifact removal is mostly ignored in the literature. To overcome this, this work first introduces an artifact-cleaning preprocessing stage through a fine tuned UNet++, thus ensuring that classification and XAI results are only guided by true fracture features and not confounding visual noise.

In short, this chapter summarizes the state of the art, outlines the shortcomings of previous work, and motivates the project's original contribution of combing artifact removal with fractural classification and explainable AI for more accurate and clinically meaningful results.

Chapter 3

Research Methodology

3.1 Methodology

3.1.1 Overview

The proposed project has a clear pipeline for reproducible and interpretable fracture detection. The first step in the workflow is the image preprocessing, where raw X-ray images are cleaned up and unwanted objects such as plates, screws or implants are removed. The additional benefit of this step is that the images are improved in quality and that subsequent analysis is performed over clinically relevant regions rather than false feature regions.

Next is the classification by which the normal and fractured cases are separated by the deep learning model. By predicting the important modes of fracture from the cleaned input, the model can eliminate the false detections caused by artifacts, and then provides the classification output (normal or fracture) in the prediction phase with the confidence score (how sure is the decision).

Finally, an explainability (XAI) technique is used for generating a heat map that shows the most important areas of the image. This will make it more transparent: if a certain prediction is made, the clinician will know that it is based on actual fracture landmarks, and not on arbitrary features of pictures.

3.1.2 Proposed Methodology

The method adopted in this work is based on the systematic multi-stage pipeline that facilitates accuracy, strength and interpretability of the fracture identification. The preprocessing of raw X-ray images into clean images and the removal of plates, screws or other implants is the first stage of the process known as image preprocessing. The reason behind this is to eliminate the confounding information that can mis-direct the model during training and inference, therefore enabling the learning process to focus on the features that are related to the fracture.

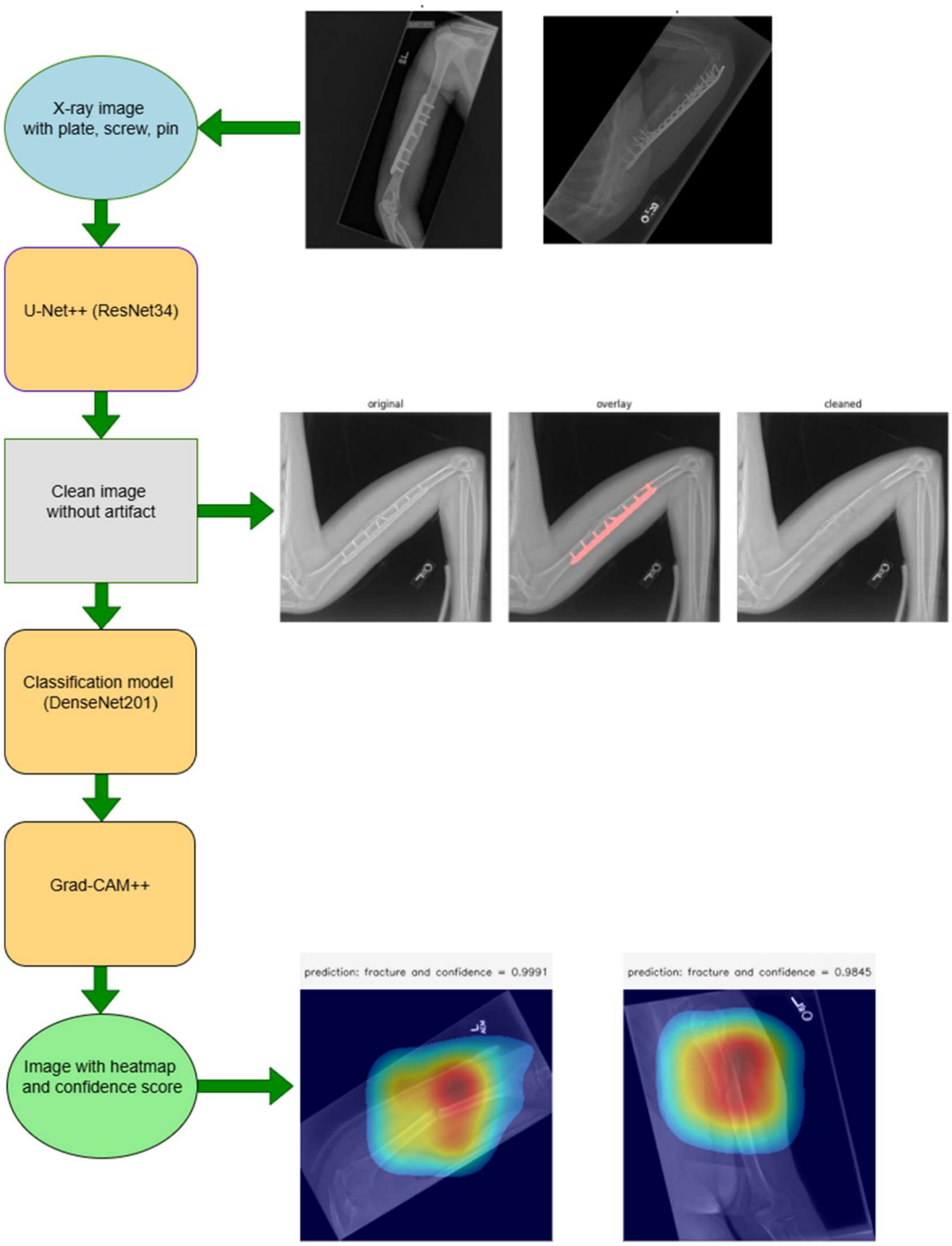


Figure 3.1 Methodology diagram

Classification follows preprocessing as the second step where a deep learning model is used to learn to differentiate normal and fractured cases. The model filters the images and relies on the features to make credible predictions. The next step in the prediction sequence, the system provides a normal or fractured X-ray and a confidence value that is used to gauge the precision of the choice.

This is the final step, which involves explainable AI (XAI) in the form of Grad-CAM++ that identifies significant regions in the images that the model was conditioned to focus on, shown in Figure 3.1. The interpretability step will also increase transparency as clinicians and researchers can ensure the attention in the model is medically pertinent to the fracture region, as opposed to irrelevant artifacts. Each of these actions reflects the general approach that improves the performance of the predictors without undermining trust and clinical significance.

3.2 Detailed Methodology

The Project Methodology was developed based on a goal to develop a strong, interpretable, and clinically sound pipeline to detect fractures in radiographic images. It has been structured into four consecutive yet connected steps: preprocessing and artifact removal, classification, prediction with confidence scoring, and explainable AI visualization. Each phase was chosen based on an analysis of other options, and the final design was chosen due to its capability to maximize diagnostic accuracy, minimize the dependence of artifacts, and offer clarity to the model decisions. A combination of these steps results in a complete system filling the gaps in current studies, yet maintaining realistic clinical relevance.

3.2.1 Preprocessing of images and removal of artifacts

The initial phase of the pipeline was dedicated to one of the most important problems of automated medical imaging: metallic artifacts in radiographs. These are artifacts, that are usually generated by the surgical hardware (e.g., plates, screws, or rods) that can consume a large part of the visual field and, with its high-contrast signals, can be easily misleading to the classifiers. Such artifacts not only decrease the quality of predictions, but also decrease the quality of explainability tools, as saliency maps can focus on hardware rather than fracture areas in traditional deep learning systems as shown in Figure 3.2 Dataflow diagram with artifact image with artifact before preprocess.

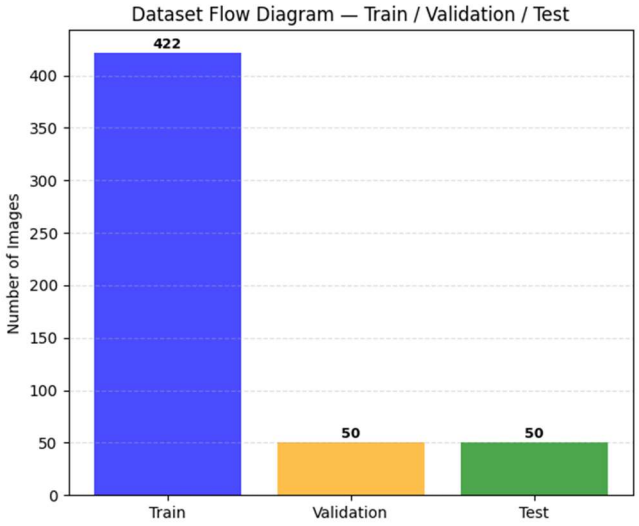


Figure 3.2 Dataflow diagram with artifact image with artifact before preprocess

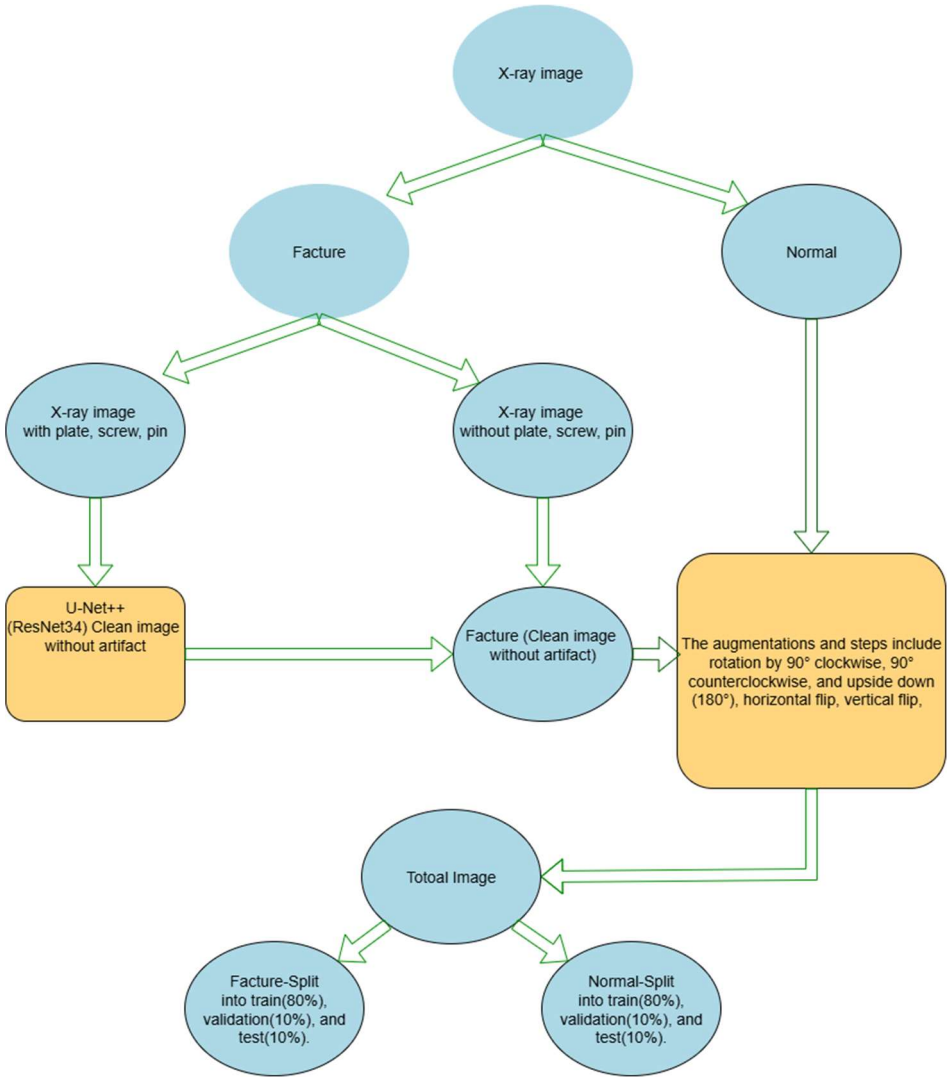


Figure 3.3 Preprocess flow diagram

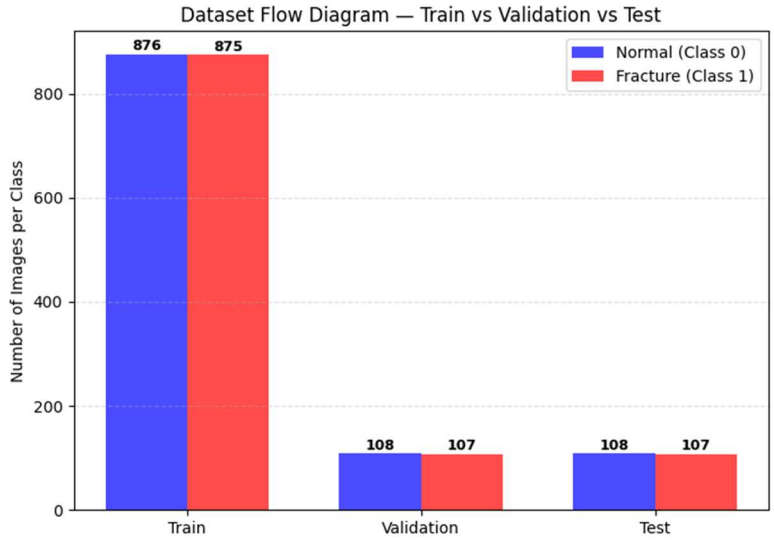


Figure 3.4 Dataflow diagram with clean image after preprocess

This problem was addressed by using a customized UNet++ architecture in the project, illustrated in Figure 3.3 Preprocess flow diagram. UNet++ is based on the original UNet architecture with skip connectivity in a nested structure that allows richer multi-scale feature representation and improved gradient flow. It also allows the model to isolate the bone structures over metallic artifact better, producing clean radiographs with preserved vital anatomical fidelity but reduced occurrence of false high-density structures, as shown in Figure 3.4 Dataflow diagram with clean image after preprocess. Besides UNet++, a preprocessing pipeline using the classical image enhancement techniques of choice was also constructed. CLAHE was used to stabilize local contrast, and unsharp masking was used particularly in those areas where a faint fracture line might otherwise be lost in a shadow, to bring into focus a diagnostically important edge. Additional normalization and cropping provided consistency of strength and particular scrutiny of the areas of focus in the anatomy.

Other solutions to the issue, such as patch-based GANs, were also tried, but proved to be unsuitable. The GAN repair techniques were also more prone to artificial artifacts that may alter the morphology of the fractures and, therefore, affect the diagnostic value. Similarly, the traditional contrast enhancement algorithm increased the visibility at local but failed to remove metallic noise fully. UNet++, in its turn, provided a stable and efficient trade-off between artifact-controlling and bone detailing, and provided a clean dataset, on which downstream classifiers could perhaps be trained more safely, without risking to make a misclassification, as demonstrated in Figure 3.5 Side by side overview of original, overlay and clean image after preprocess.

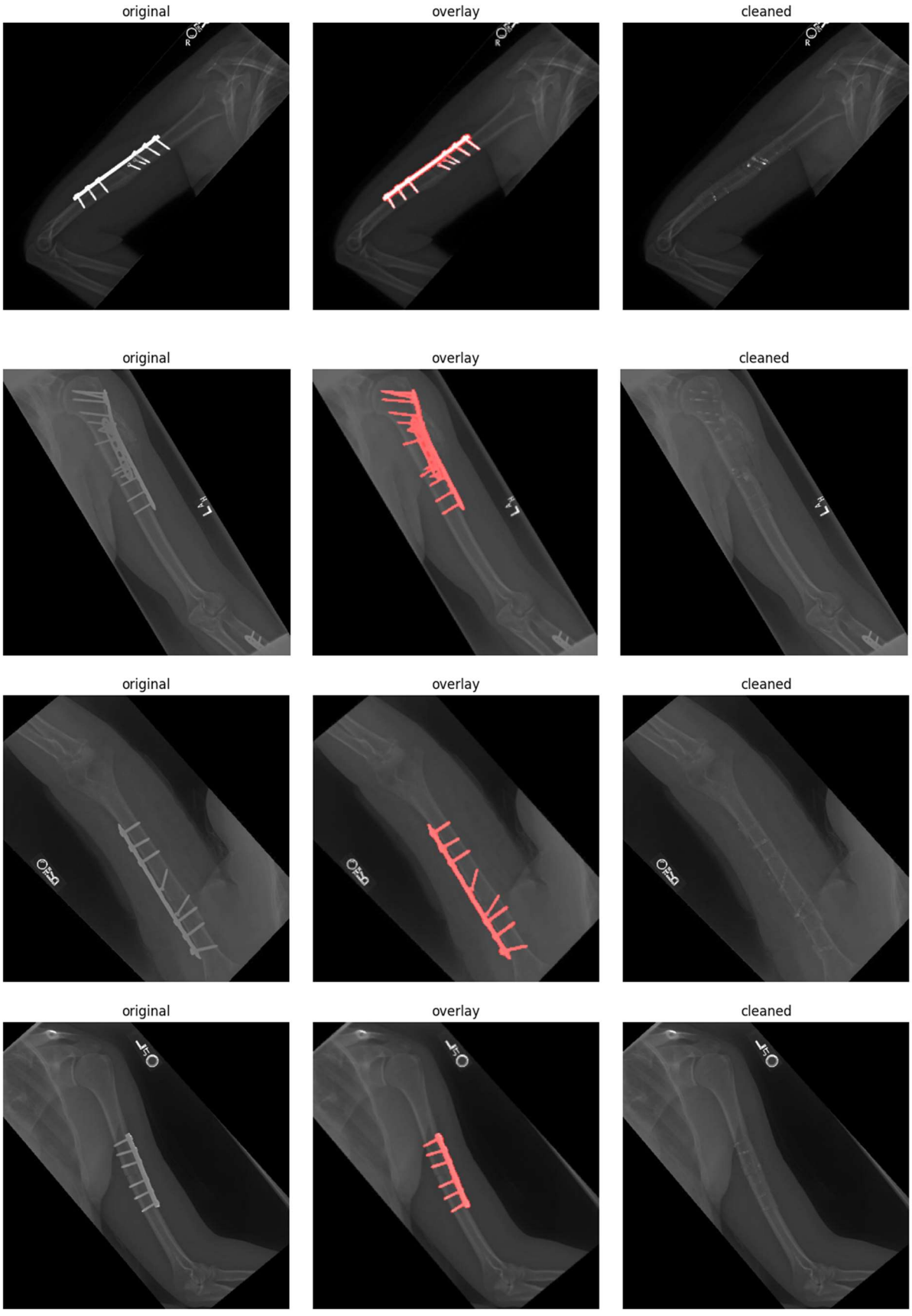


Figure 3.5 Side by side over vier of original, overlay and clean image after preprocess

3.2.2 Classification

The artifact suppression process followed by the classification process on the cleaned images. The objective was to determine the normal radiographs and the fractured radiographs with high degree of accuracy and clinical reliability. DenseNet201 was selected as the classification backbone because its fully connected structure enables features to be re-reused throughout its layers and avoids the vanishing gradient problem. The properties are useful in medical imaging, where the minute fracture lines and the coarse structural hints can be the difference between a correct diagnosis and a false alarm.

ImageNet denseNet201 ImageNet pre-trained ImageNet denseNet201 was loaded with weights and fine-tuned to the target task of fracture detection. The original classification head was substituted by the custom sequential head which was to be stabilized to reduce overfitting. It consists of a fully connected layer that downsampled 1920 features to 512 then a batch normalization, a ReLU, dropout (0.3), and an end linear layer that projected to a single output node. This design allowed the model to be flexible to the dichotomy classification task and also resistant to noise and dataset variability. As a data augmentation, random flips and rotations were employed to improve the generalization.

ResNet101, VGG19 and Inception V3 were also tried as other classifiers. Despite providing plausible baselines, all these models had the following set of limitations: reduced convergence, reduced sensitivity to fine fractures, or enhanced sensitivity to false positives. Detection models based on YOLO were also investigated, but their bounding-box detections were unnecessary in a binary classification problem and were expensive to compute. DenseNet201 has always been much more accurate, stable, and efficient, and this is why it was selected as the primary classifier.

3.2.3 Prediction and Confidence Scoring

The third stage was concerned with predictability of the forecast based on calibrated outputs of probability. The DenseNet201 classifier not only produced hard binary labels, but probability scores through a sigmoid transform of the final output as shown in Figure 3.6 Prediction and Confidence score. By calibration, predictions using these probabilities were then scaled to clinically relevant thresholds.

Optimization of the performance was done by tuning the threshold and the evaluation measure is the F1-score. The balance was found to be optimized at an approximate threshold of 0.658 with a precision of 0.991 and recall of 0.981. This operating point reduced both the false positives and the false negatives; this allowed the system to be very reliable in practice. The model was also flexible in that, thresholds could be established based on different clinical settings. As a rule of thumb, during emergency triage, detection limits can be set to sensitivity so that the proportion of false fractures is minimized, whereas in screening, detection limits can

be set to giving more false positives and reducing the number of clinicians assigned to real patients.

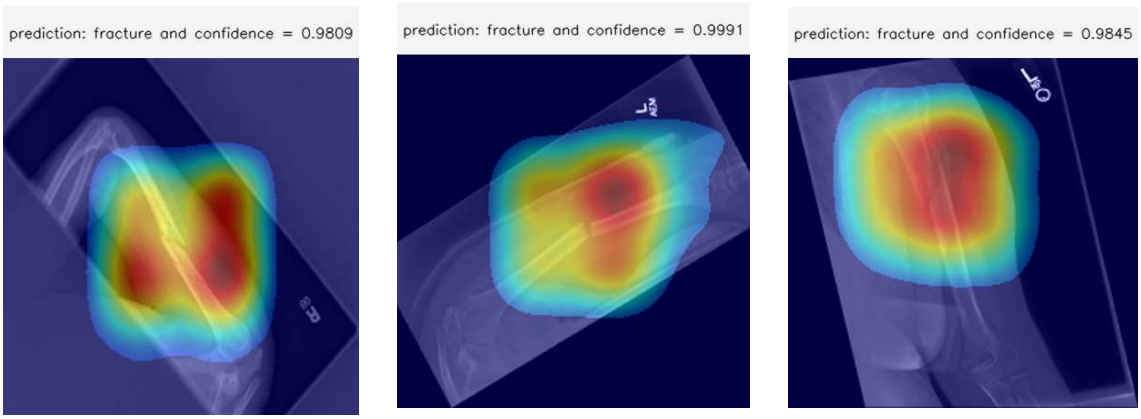


Figure 3.6 Prediction and Confidence score

A binary classifier that was not calibrated by confidence was also considered, but was later rejected. These models are not flexible enough to be applied to the field of real-life healthcare where the process of making a decision is often a trade-off between false alarm risks and diagnosis detection risks. The confidence scoring enabled the pipeline to be aligned with the optimal medical AI practices and enhanced the interpretability and usability of the pipeline in clinical practice.

3.2.4 Grad-CAM++ explainable AI (XAI)

In the final stage of the pipeline, explainable AI methods were added to ensure that the predictions would be correct and interpretable. Grad-CAM++ was selected since it can be used to generate saliency maps at high resolution that can inform us about which parts of the images had the largest influence on the choices that the model makes. The purpose of these heatmaps was to make sure that the classifier was seeing fracture-related and not irrelevant structures or artifacts.

To achieve healthy operation, the deepest convolutional layer of the denseblock4 of denseNet201 was configured to produce CAM, and a fallback mechanism to automatically identify the deepest Conv2d layer was implemented. This architecture allowed visualization to be reliable when the model architecture changed. The Grad-CAM++ outputs could always localize the location of a fracture such as a cortical rift and trabecular tear, but not the location of a metallic implant because of the prior UNet++ preprocessing.

Alternative interpretability approaches, including Integrated Gradients and SHAP, were also taken into account, although the results produced by them were not as easy to apply in clinics. They did provide pixel level attributions and numbers, but they did not provide the visual clarity to interpret these numbers rapidly in the real world radiology environment. Grad-CAM++, by contrast, produced overlays that

could be immediately reviewed by clinicians. To simplify the process, the outputs were presented in panels which consisted of the original image, the Grad-CAM++ heatmap, and a composite overlay of the predicted class and confidence score, which could be tested by the clinician with a visual test.

3.2.5 Integrated Pipeline

The four steps combined, artifact removal, UNet+, classification, DenseNet201, calibrated prediction, confidence scoring, and interpretability, Grad-CAM+, were part and parcel of a pipeline as shown in Figure 3.7 Integrated Pipeline- input raw images, get XAI Heatmap. This design was a direct reaction to the deficiency of the earlier researches, i.e., the problem of artifact dependency; the lack of sound interpretability. The combination of preprocessing and further classification and explainability enabled system to not only improve the quality of the diagnostic output but also to generate clinically reliable prediction output.

The pipeline is resilient to changes in imaging, and is open-minded to other clinical scenarios. These features make it very useful in the real world by reducing the gap between the high performance levels of AI in the laboratory setting and in clinical practice. Ultimately, this solution will contribute to reducing the rate of diagnosed errors and improving patient outcomes and the degree of trust that clinicians have towards AI-based health services shown in Figure 3.7.

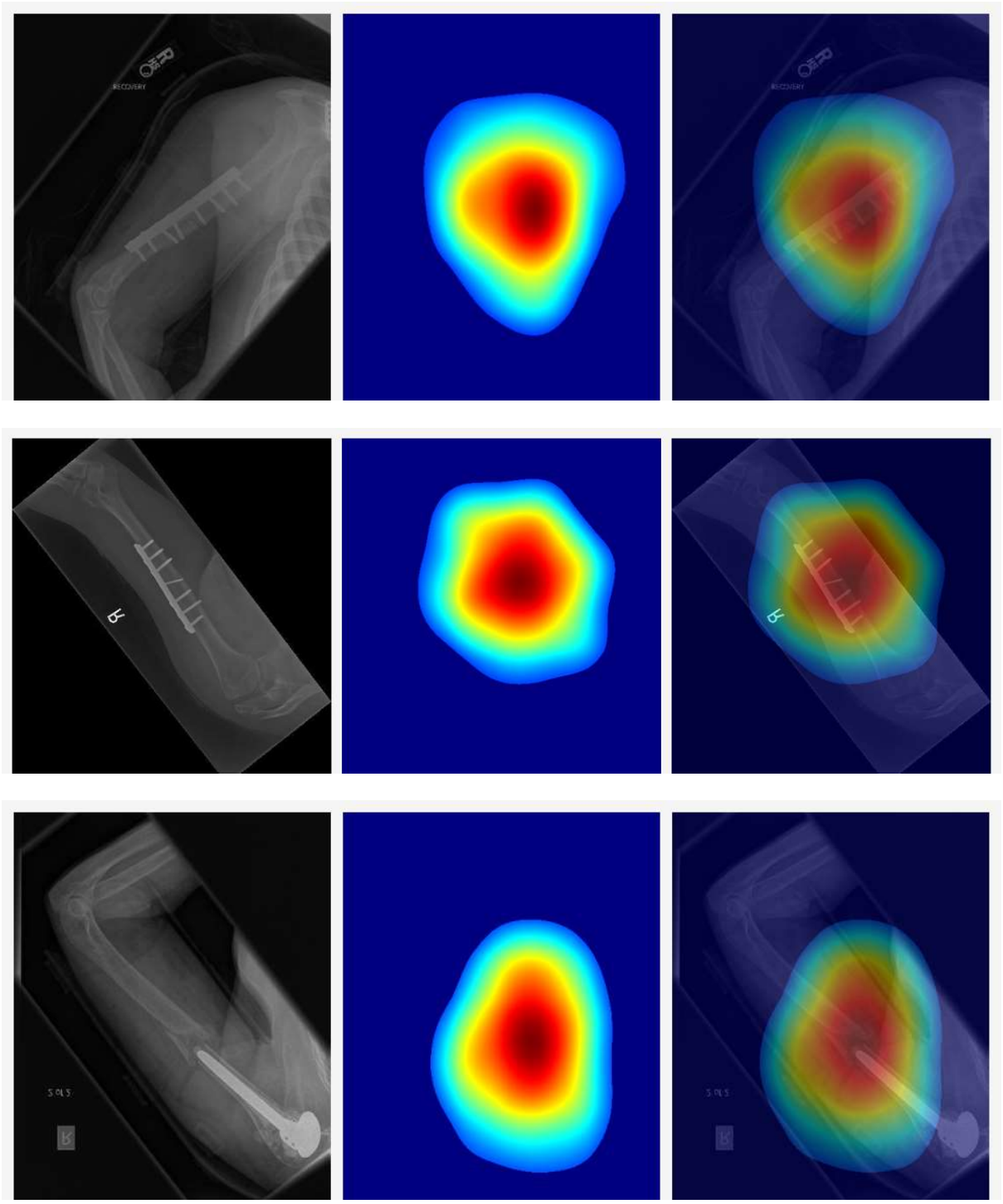


Figure 3.7 Integrated Pipeline- input raw images, get XAI Heatmap

3.3 Project Plan

In order to finish this project I ensured that I had planned out the work as per a timeline whereby the workload was broken down into manageable tasks in each sequential stage. A few warnings were issued: at every stage, the work was to evolve smoothly out of the problem to the report with enough space to accommodate contingencies such as missing data, noisy artefacts and model choice. The plan stretched over a period of approximately eight months, and it was subdivided into targeted activities as indicated below with Table 3.1:

Find a Medical Imaging Issue (1 month)

The first stage was the definition of the scope of the project. Once the literature on the topic of medical image analysis had been reviewed, the issue of false classification due to artifacts was identified in the process of fracture detection. This was done to ensure that the project covered a clinically-relevant and computationally-meaningful challenge.

Avoid Problems of Data Availability (1 month)

The availability of large heterogeneous annotated datasets is one of the primary limitations to medical imaging research. At this stage, some data sets (e.g. pediatric wrist X-rays, multi-region radiographs) were checked to have sufficient data to be used in training and testing. Patient confidentiality and institutional access were other ethical issues discussed.

Dealing with Noise in the Data (15 days)

Once the data had been acquired it was discovered that the data contained a great deal of noise in the form of low contrast, poor alignment of images within the data and background distraction. At this point the dataset was explored to measure these noise patterns and to identify the necessity of using standard preprocessing or more recent artifact removal techniques.

Noise Abatement (1 month)

Noises and other artifacts were eliminated by image pre-processing methods, including normalization, contrast and cropping. To remove the artifacts, we optimized a refined UNet++ architecture in which the influence of the implants, among other metallic objects on the fracture detection, was reduced. It generated a clean dataset, based on which sound classification was possible.

Select a Model (1 month)

Pretrained CNN models (ResNet101, VGG19, InceptionV3) The initial step was to learn what pretrained CNN models could do. DenseNet201 was chosen as the main classification model after experiments due to its higher efficiency in fracture

detection tasks and the possibility to reuse features at a lower cost.

Cleaning Dataset, Artifact removal (1 month)

The UNet++ based preprocessing pipeline was fully integrated into the workflow in this step. All imagery was subject to artifact removal to provide consistency with cleaner higher-classification-grade results.

Create and Construct a Model of Classification (1 month)

DenseNet201 was trained on the preprocessed data with the one set of hyperparameters, and the values of accuracy, recall, and F1-score are balanced. One more fact worth mentioning is that the performance was compared with other models and learning curves were also tracked, not to overfit.

Learn explainable artificial intelligence approaches (1 month)

Grad-CAM+ was added to the pipeline to produce saliency maps, which can be used to visualise fracture-relevant regions of the X-rays. That was to make the decisions made by the model more clear, and emphasize that the predictions were more tied to the clinical, as opposed to the simple models, as they had to be realistic by predicting the fractures on real cues.

Prepare Project Report (15 days)

The final activity was structuring the results, methodology and interpretations in a formatted report in line with the academic / institutional requirements. Each of the effectiveness of preprocessing and classification performance and XAI visualizations have been represented by figures, tables, and explanatory diagrams.

3.4 Task Allocation

Table 3.1 Task Allocation

Tasks	1	4	8	12	16	20	24	28	32
Identify a Medical Imaging Problem (1 month)									
Address Data Availability Challenges (1 month)									
Handle Noisy Data Issues (15 days)									
Develop Noise Removal Strategies (1 month)									
Select an Appropriate Model (1 month)									
Remove Artifacts & Clean the Dataset (1 month)									
Build & Fit the Classification Model (1 month)									
Implement Explainable AI (Grad-CAM++) (1 month)									
Prepare Project Report (15 days)									

3.5 Summary

Within the framework of the present research, this chapter provided an in-depth description of the research methodology and offered a systematic framework that was intended to fulfill both technical and clinical feasibility in fracture detection. The chapter started by providing a general overview of the methodological pipeline, highlighting the 4 overlapping steps of preprocessing involving artifact removal, classification, prediction, and explainable AI (XAI). All these stages were found to be crucial towards the development of a system that not only provides high accuracy, but also interpretability and reliability, in clinical practice. The summary explained the roles of preprocessing in removing distracting artifacts, classification in the form of diagnostic decision-making, prediction in the form of calibrated confidence levels, and XAI in the form of visual justification of the system results.

After the overview, the chapter proceeded to a more in-depth discussion of the design options, where alternative options were critically discussed. Prior techniques of preprocessing such as CLAHE and unsharp masking were known to be capable of generating and enhancing contrast and edges, but it still was not felt to be sufficiently effective to handle metallic implants or more challenging noise. In the same spirit, alternative backbone designs such as ResNet101, VGG19, and InceptionV3 were also tried but proved to be limited in stability or efficiency. Unlike this, UNet++ was proven to be the best artifact removal model because of its nested skip connections, which retain bone structures whilst disactivating noise. DenseNet201 was chosen to classify due to its connectivity density that guarantees reuse, avoiding the vanishing gradient problem, and allows detecting minute fracture lines with consistent results. The reason to include the confidence scoring and Grad-CAM++ was also explained and how these two features contribute directly towards interpretability, transparency, and trust of predictions by clinicians.

A project plan was then used to operationalize the methodology and plot the execution process starting with the dataset collection to evaluation. The workflow stages were associated with a certain deliverable, including clean image sets, trained classification models, confidence-calibrated predictions, and interpretability products. The schedules were constructed in such a manner that progress could be systematically tracked using set checkpoints to examine the preprocessing performance, classification accuracy and reliability of XAI. The conceptual validity of the research and its practical feasibility within the available resources and time were sound due to this stage of the research.

Finally, Chapter 3 defined the methodological base of the project by unifying the justification of the selected techniques, showing how they address the limitations of other options, and how they can be bundled into a repeatable workflow. The methodology provides technical rigor and clinical relevance by connecting artifact removal, strong classification, calibrated prediction, and interpretable outputs in

one pipeline. He or she is a systematic design in the sense that the next phases of implementation and evaluation will be informed by a framework that is clear, flexible, and consistent with the overall research goals of designing a system of fracture-resistant artifacts, a reliable system.

.

Chapter 4

Implementation and Results

4.1 Environment Setup

This project was configured on Google Colab Pro that provided the possibility to use cloud computing service and authenticate into cloud storage and access machine learning experimentation with the GPU acceleration turned on. Colab Pro is chosen because of the scalability, ease and interrelation with existing deep learning models. It works with preprocessing using UNet++ and classification using DenseNet201 in addition to explainable artificial intelligence (XAI) features such as Grad-CAM++.

All the python packages and libraries such as data preprocessing, statistical evaluation, and visualization were imported to utilize the pipeline. The Python Imaging Library (PIL) i.e. PIL and PIL.Image were used to load and perform transformations on the images. The library provided an accurate method of X-ray image resizing, cropping and normalization prior to their input into preprocessing and classification models. Besides, the application of Google Colab Drive was also incorporated with the assistance of the following incorporations: Google and Google.colab.drive through which it is possible to connect to Google Drive without any problems. This ensured that the dataset was securely stored and categorized into the train/test/validation folders, and could be easily accessed later during the experiment.

Matplotlib and NumPY were used to represent the number in the project and analyze the graph. To experiment with learning curves and heatmath and confusion arrays of XAI and NumPy to work with large array of images a set of numerical interactions at scale i.e.vectorize way, NumPy was added to the Big Five and Matplotlib was added to both. To automate the dataset path resolution process and document experiments, the processing of files and folders was performed using the operating system module. Similarly, Pandas was employed to facilitate data organization (structured data) such as metadata organization, table organization of experimental results and performance.

The main machine learning tool was scikit-learn (sklearn). In this case different measures of performance were taken into account such as accuracyscore, auc, f1score, precisionscore, and recallscore so that performance of a classifier should be evaluated in a holistic way. Other powerful operations such as precisionrecallcurve and rocaucscore were used to explore the sensitivity and specificity trade-offs which are meaningful in

medical imaging applications. In order to ensure that the data was reproducible, Traintestsplit processed it and removed the risk of scale imbalance and its effect on the training stability, StandardScaler rescaled it.

Math and random libraries support was also in place. Physical computations during the preprocessing, augmentation and evaluation were run in the math module, physical randomness in the samples and augmentation was introduced in the random module to provide variability and repeatability of the training runs.In general, the environmental setup integrated the flexibility of Google Colab Pro and a carefully chosen collection of Python libraries. It could be easily preprocessed, model trained, successfully evaluated and visualized; and the whole experimental pipeline itself could be reproduced, scaled and tested in a controlled computational environment.

4.2 Performance

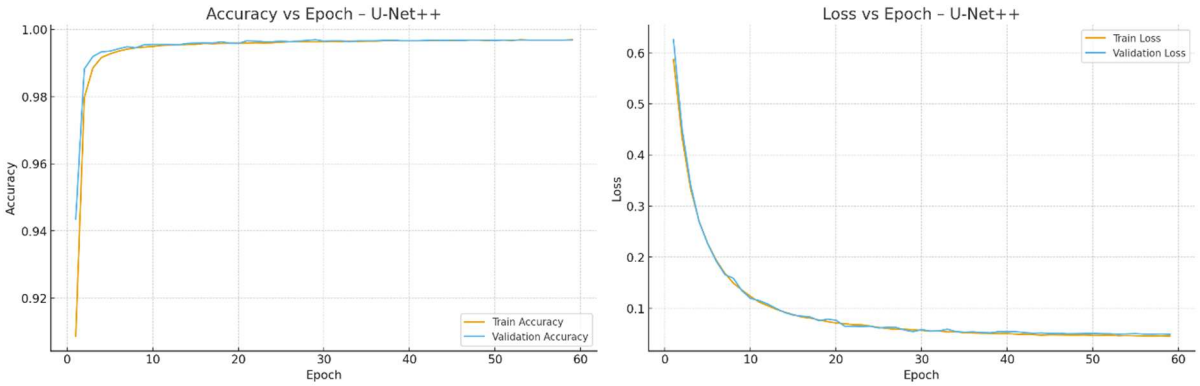


Figure 4.2.1 Accuracy vs Epoch and Loss vs Epoch U-Net++

According to accuracy curve and the loss curve presented in Figure 4.2.1, the U-Net++ model developed a good and consistent learning trend within the training stages. In early epochs, the training and validation loss decreased rapidly, which implies that there was good convergence and generalization. It has been observed that the optimal training epoch was epoch 29 with train accuracy of 97.7% and train loss of 0.0454. We also found out that the validation performance is relatively similar to the training performance with the highest validation accuracy of 96.7% and the lowest validation loss of 0.0490 indicating high agreement between the training and the validation data. Trend curves indicated that the model could learn good features without overfitting. Nevertheless, there was still a slight fluctuation in training and validation between the fifth and sixth epochs that indicated a small amount of overfitting was slightly reduced in subsequent epochs. In total, the training F1-score and the validation F1-score of the U-Net++ were 0.9183 and 0.9090, respectively, and close-to-1.0 AUC values of 0.9994 and 0.9991, respectively, denote that our model is robust.

Even with the test set given in Figure 4.2.1, U-Net++ still attained test accuracy of 95.6062% and test loss of 0.0684. It is a recall-dominant model, its final test F1-score of 0.9001, its precision of 0.8873 and its recall of 0.9132, and its AUC of 0.9984 indicate high sensitivity to cases of artifact. Optimal cut-off of F1 was 0.3661 that provided a good trade-off of precision and recall (bias with sensitivity more than specificity). Though this does increase the false positive rate it fulfils the objective of making sure (almost) all positives are identified which has more clinical implications in medical screening tests where failure to identify a real artifact is of more clinical consequence. All these results indicate that U-Net++ can reach very high false negative reduction rates, is relatively stable when running train, validation and test, and is computationally efficient at various thresholds.

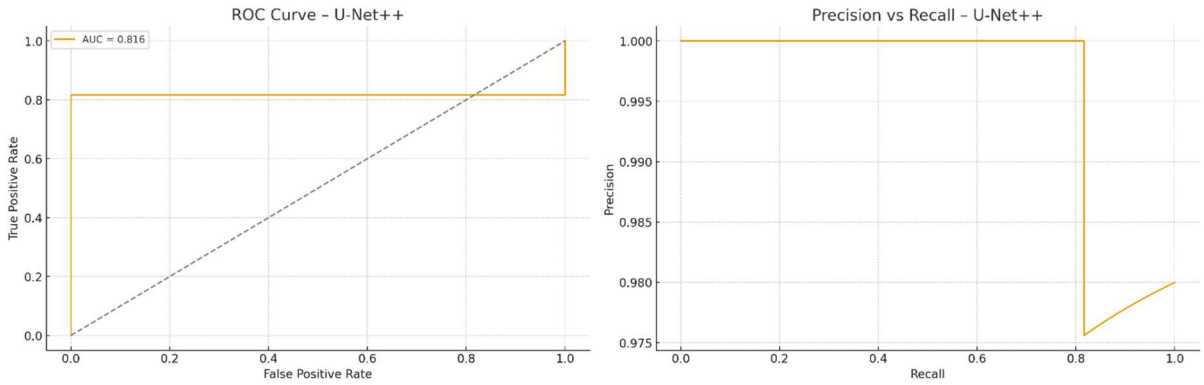


Figure 4.2.2 Roc Curve and Precision vs Recall U-Net++

The U-Net++ can classify artifact and bone with very high accuracy in terms of the Receiver Operating Characteristic (ROC) curve as depicted in Figure 4.2.2, but not better than the state-of-the-art classifiers. The model has a ROC-AUC of 0.9988, which implies that ranking performance is very close to perfect at all thresholds, and this is very responsive to true-positive performance with virtually no missed artifact. The receiver-operating curve was moved to the left, meaning that sensitivity is valued at the cost of the specificity (i.e., good detection of positive instances but weak rejection of negative ones). The accuracy-recall curve is also characterized by high recall and relatively low accuracy; however, a high average compromise rate indicates that PR-AUC = 0.9712 U-Net++ is very reasonable in a medical setting, where a false alarm is less expensive than missing a diagnosis.

Figure 4.2.2 also identifies the strengths and weaknesses of U-Net++ in the precision-recall characteristics. The optimized threshold value of 0.3661 at the optimal-F1 point results in a validation and test F1 of 0.9090 and 0.9001 respectively, indicating a compromise in precision (0.8873) and recall (0.9132) in a recall-dominated regime where false negatives are minimized. Though the stricter thresholds may slightly increase precision, the presence of false positives is still not negligible, so the model cannot be used in cases that require accuracy but rather in screening ones where sensitivity is the most vital. Combining the ROC and precision-recall evidence, U-Net++ acts more like a recall-weighted system with large coverage and can afford a higher false-alarm rate.

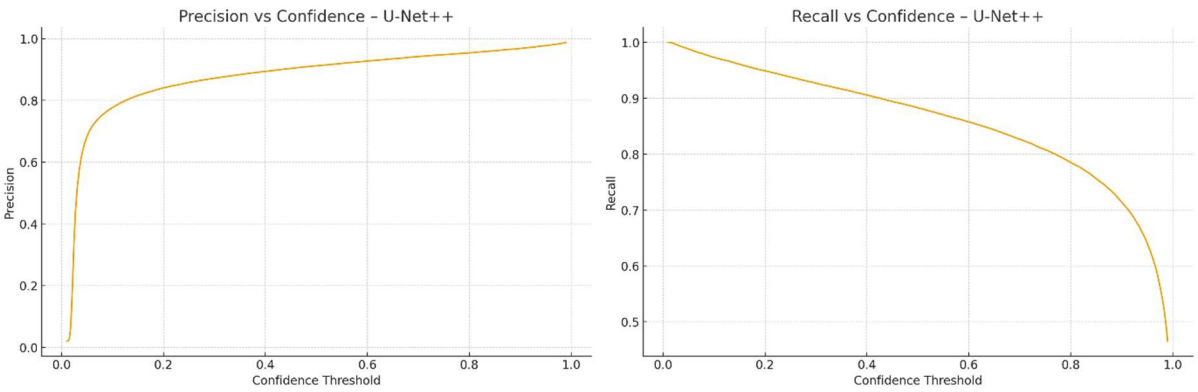


Figure 4.2.3 Precision vs Confidence and Recall vs Confidence U-Net++

Using the U-Net++ precision-confidence curves of Figure 4.2.3, the limitations and strengths of the model in any given practical classification situation are illustrated. Precision is low at very low confidence levels since a large number of negatives are marked as potential positives and as the confidence level is raised the precision increases steadily. Precision is 0.8873 and the overall test accuracy is high at the best-F1 operating point (threshold = 0.3661), but recall is also good--consistent with a recall-dominant behavior. In comparison with DenseNet201 that is more precision trained, U-Net++ tolerates more false positives to favor sensitivity, which is an acceptable trade-off in a screening setting where a negative artifact detection incurs a greater cost than a follow-up.

On the contrary, the high sensitivity of U-Net++ is the best feature in Figure 4.2.3 in the recall-confidence analysis. Recall is 1.0 at very low cut-offs and high throughout a large range: even at the F1-optimal cut-off, it is 0.9132. Reducing threshold can slightly increase accuracy, but it reduces recall, causing false negatives to get higher. This trade-off implies that U-Net++ cannot be applied in precision-sensitive confirmatory diagnostics, but can be applied in recall-intensive applications (e.g., medical screening or triage) and can be used to complement more precision-oriented models.

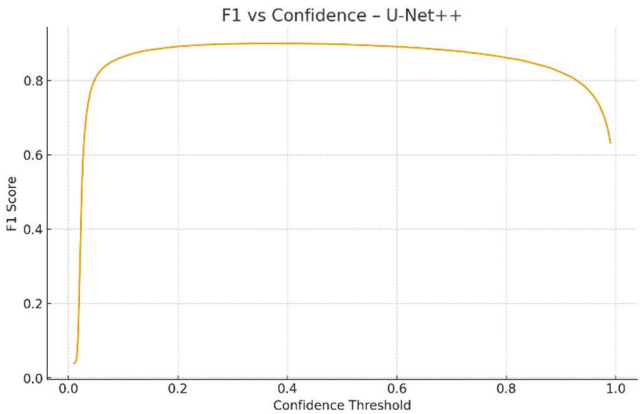


Figure 4.2.4 F1 vs Confidence U-Net++

The main precision-recall trade-off is emphasized by the F1-confidence curve of U-Net++ in Figure 4.2.4. F1 also decreases at very permissive thresholds when precision is low even though near-perfect recall is achieved; as confidence gets high the F1 increases rapidly and levels off at the best-F1 operating point. The model reaches F1 = 0.9001 at threshold = 0.3661 with precision = 0.8873 and recall = 0.9132, which is a well-balanced but recall-dominant regime, which is appropriate when comparing it to the test metrics.

Beyond this optimum, additional increase in the threshold will only provide small improvements in precision but recall decreases faster, thus balancing out this phenomenon--which can be observed in Figure 4.2.4. In practice, this implies that one should be run at or close to the 0.3661 cut-off of screening and triage settings in which false negatives reduction is most important, and that more precision-oriented models (e.g. DenseNet201) or other downstream verification can be favored.

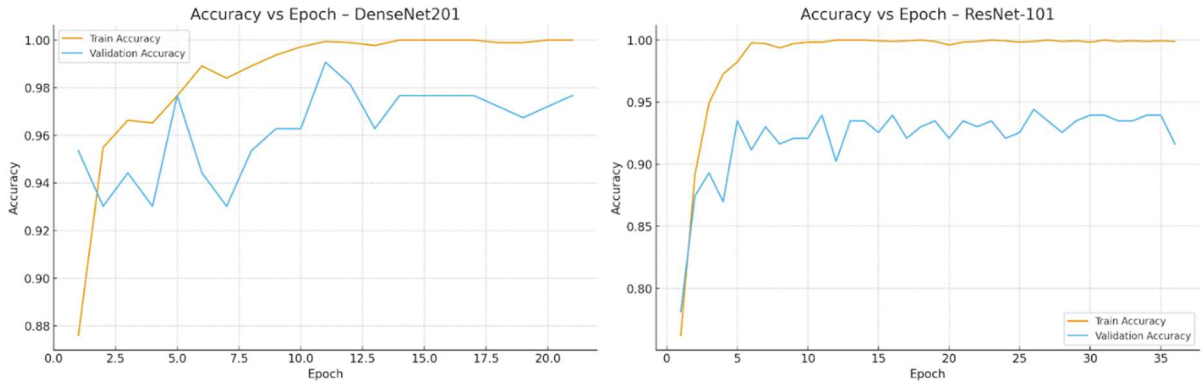


Figure 4.2.5 Accuracy vs Epoch DenseNet201 and ResNet-101

Figure 4.2.5 indicated that DenseNet 201 generalized significantly more easily and without much turbulence than ResNet 101. DenseNet201 has demonstrated good performance since the very first epochs: the training and validation accuracy curves increased rapidly and converged to the same level with a very small error, which means that it reuses features effectively and generalizes well. Following a rather short warm-up, the two curves approached each other closely during the rest of the training which showed rapid convergence. The accuracy also leveled off to the high-90s after the training, which is also in line with its held-out results (test accuracy 98.60% and F1 0.9859), which again confirms the assertion that dense connectivity will enable gradient-flow and feature reuse yielding both fast learning and stable generalization.

In comparison, ResNet101 was more varied and exhibited overfitting like Figure 4.2.5. Accuracy in training increased faster than validation accuracy but it stabilized at a lower level and oscillated, and there existed a chronic train-validation gap in subsequent epochs and the generalization was not so good as with DenseNet201. This trend is reflected in hidden data: accuracy of tests 94.42% and F1 0.9439--good fracture solid recognition, though not as great as DenseNet201 and indicative of prior overfitting to the validation

set. All in all, DenseNet201 not only attained greater accuracy on held-out data, but more stable training dynamics, which makes it the more resilient to the real-world diagnostics.

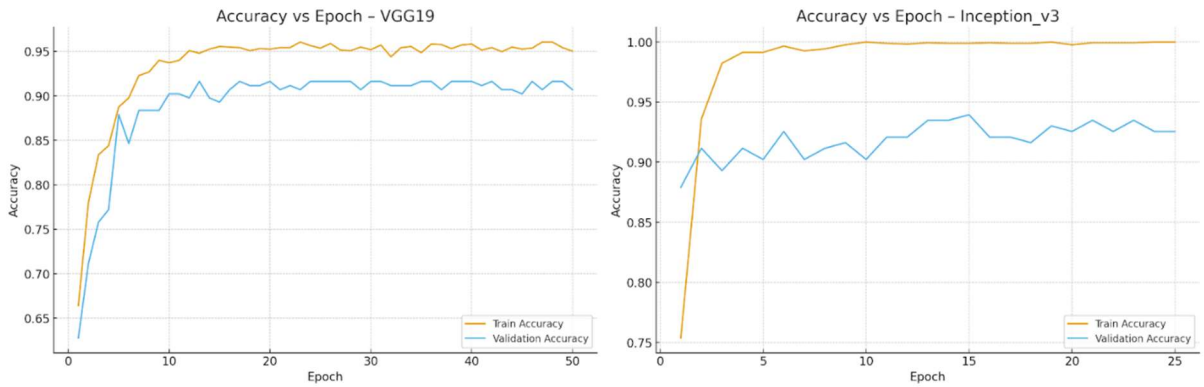


Figure 4.2.6 Accuracy vs Epoch VGG19 and Inception_v3

Figure 4.2.6 reveals that VGG19 performance in training and validation phase improved slowly in accuracy with generalization gap between the two phases. VGG19 achieved the highest accuracy of 96.1 and 91.6 in training and validation, respectively, and these results show that this model could be trained to work well on the training set but fail to provide the same level of accuracy on unseen data. The convergence rate decreased after that, and the accuracy of validation that was close to 91-92 in the end indicated an overfitting bias when the model has been optimized. This found its way into testing, with VGG19 scoring an accuracy of 92.09 percent, a single digit higher showing weak generalization, yet reduced strength compared to more powerful architectures in medical high stakes classification.

On the other hand, the InceptionV3 was observed to converge faster, although harder to keep the higher levels of validation as observed in Figure 4.2.6. The model achieved a training and validation accuracy of 99.9 and 94.0 respectively during its best epoch (15); although the generalization gap increased to test accuracy of 90.70 on unseen data. This overfitting gap is characterized by the fact that the model performed too well in training but was less predictive in the cases of validation and testing. InceptionV3 trained more quickly than VGG19 but still only achieved 99.1% validation which was far short of the highest achievable of 99.1% of DenseNet201, and both models achieved relatively high accuracy, but at the cost of lowered reliability compared to DenseNet201, which achieved much higher validation accuracy and more stable performance of all three phases.

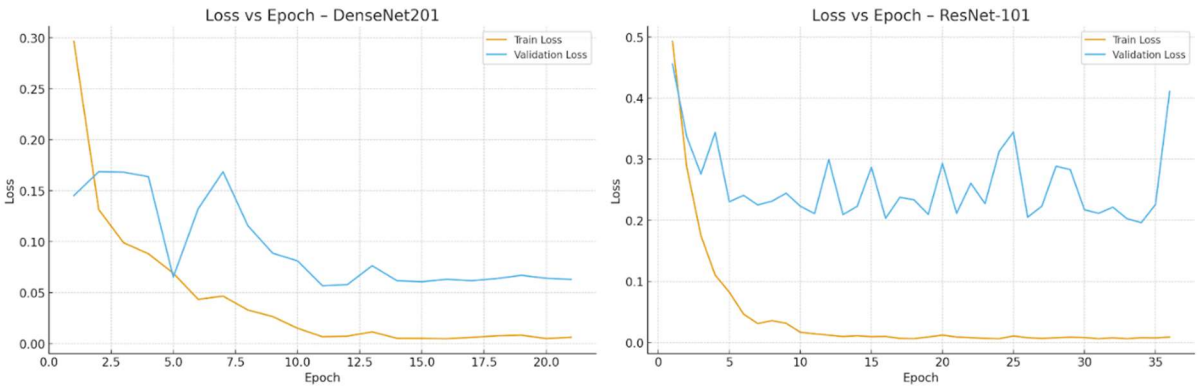


Figure 4.2.7 Loss vs Epoch DenseNet201 and ResNet-101

The most stable loss reduction and convergence was reached by DenseNet201 as depicted in Figure 4.2.7. The loss of the model during the training phase at epoch 1 was 0.2966 but at epoch 5 it was 0.0691 whereas the loss at the validation phase was nearly the same 0.0653-indicating effective learning and good generalization. As further epochs were done the curves continued to sharpen and finally stabilized with a tiny and steady gap, centering around 0.0517 (training) and 0.0821 (validation) at epoch 20. This close correspondence between training and validation loss can be interpreted as a sign of low overfitting, and the strength of DenseNet201 to detect fractures.

Figure 4.2.7 also shows that ResNet101 learnt fast but with higher train-validation gap compared to the DenseNet201. The loss in training decreased to 0.4931 at epoch 1 but at 0.0823 at epoch 5 which signals poorer generalization at the beginning. The training loss continued to drop as training went on, to 0.0367 at epoch 20, and validation loss stabilized at around 0.1289 and increased back to 0.1573 by epoch 36. This training/validation loss divergence is typical of overfitting, meaning that ResNet101 was able to capture training behaviors, but generalize worse than DenseNet201.

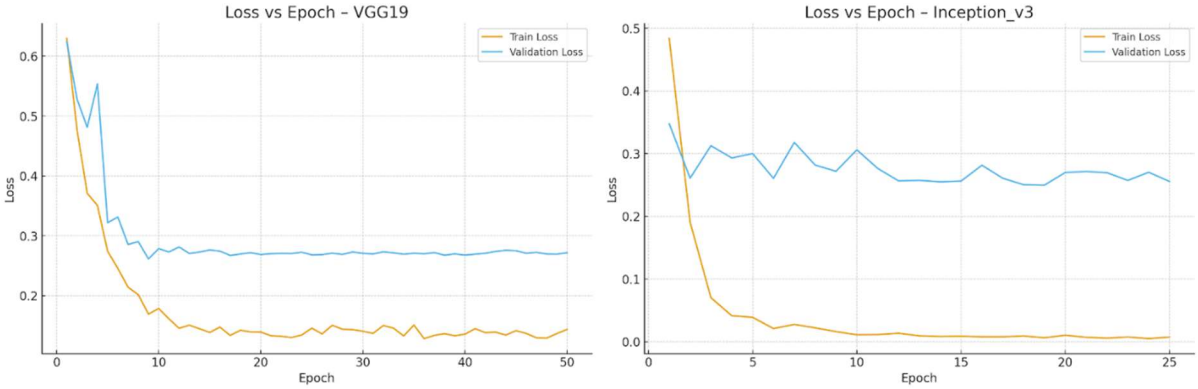


Figure 4.2.8 Loss vs Epoch -VGG19 and Inception_v3

The training and validation loss curves of VGG19 had slower convergence and less stable training than the deep architectures, as illustrated in Figure 4.2.8. In its optimal epoch (17) VGG19 achieved a training loss of 0.1281 and a validation loss of 0.2612, which is lower but still high enough to indicate poor generalization in spite of further learning. The

trend continued to testing where loss was 0.2303 and the test accuracy was 92.09, confirming that VGG19 was able to fit the training data yet failed to decrease the validation loss linearly with the training loss, as compared to train-validation-fitting models such as DenseNet201.

Despite converging faster on the training set, InceptionV3 has had issues with validation stability as shown in Figure 4.2.8. During its optimal epoch (15), training loss was brought to 0.0086 and validation loss was high at 0.2560, which is overfitting, as shown by a test loss of 0.2928 and test accuracy of 90.70. InceptionV3 was found to be more aggressive in terms of minimizing training loss than VGG19, yet both had relatively high validation loss as compared to DenseNet201. All these combined indicate that InceptionV3 did not require as long to train but the continuing train-validation gap constrained robustness and DenseNet201 maintained a better training-validation loss ratio.

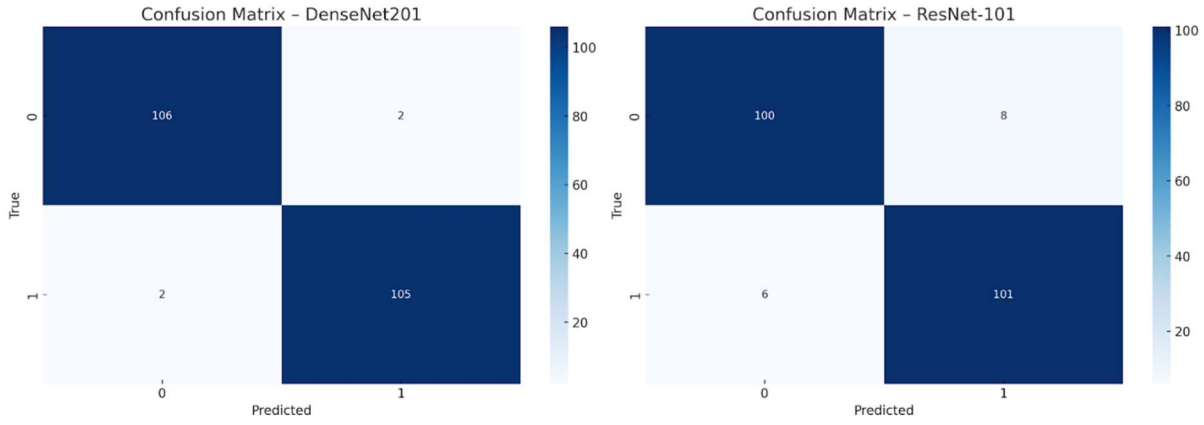


Figure 4.2.9 Confusion Matrix DenseNet201 and ResNet-101

In the balanced environment, denseNet201 performs better than ResNet101, as Figure 4.2.9 demonstrates. At optimal-F1 operating point (threshold = 0.6682) DenseNet201 achieves a test accuracy of 98.60, F1 of 0.9859, precision of 0.9906 and recall of 0.9813, with very low false negatives and low false positives. This high sensitivity, and high specificity, along with its earlier mentioned quality of validation (maximum validation accuracy 99.1%), make DenseNet201 highly appropriate to high stakes medical applications where false alarms are to be kept low and missed fractures are unacceptable.

In comparison, the confusion matrix of ResNet101 suggests good but less robust balance at optimal-F1 (threshold = 0.5111), with test accuracy 94.42, F1 = 0.9439, precision = 0.9439 and recall = 0.9439. Even though it correctly points out most positives, there are still more false negatives and false positives than with DenseNet201, which restricts its applicability in the most rigorous clinical processes. In general, as Figure 4.2.9 demonstrates, DenseNet201 has a better sensitivity-precision trade-off and has a less noisy confusion pattern compared to ResNet101.

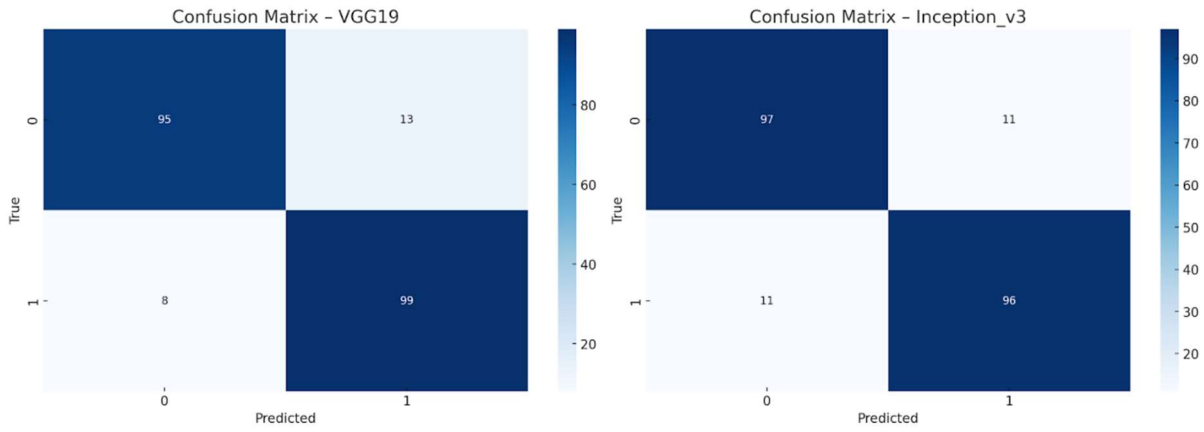


Figure 4.2.10 Confusion Matrix VGG19 and Inception_v3

Figure 4.2.10 indicates that the VGG19 confusion matrix has a pattern of greater precision and lesser recall. With a threshold of 0.6830 (best-F1 operating point), VGG19 achieves test accuracy of 92.09, F1 of 0.9163, precision of 0.9688, and recall of 0.8692- i.e. minimal false positives, but more missed fractures than desired. This is consistent with its previous validation behavior (validation accuracy 91.6%, validation F1 0.9048): the model can be conservative and accurate when it classifies a fracture, but the confusion matrix shows a greater fraction of false negatives than DenseNet201, making it less reliable in clinical applications that require confidence because mistakes are very expensive.

InceptionV3 has a more moderate yet still recall based trade-off as observed in Figure 4.2.10. It achieves best-F1 = 0.1328 at this threshold (achieving test accuracy 90.70, F1 = 0.9099, precision = 0.8783, and recall = 0.9439). The confusion matrix thus reports relatively low false negatives (high sensitivity) and a non-trivial number of false-positives--vastly worse than VGG19 but not as good as DenseNet201. Generally, VGG19 is recall-limited, but precision-strong whereas InceptionV3 is more recall-dominant without being as well-known as DenseNet201.

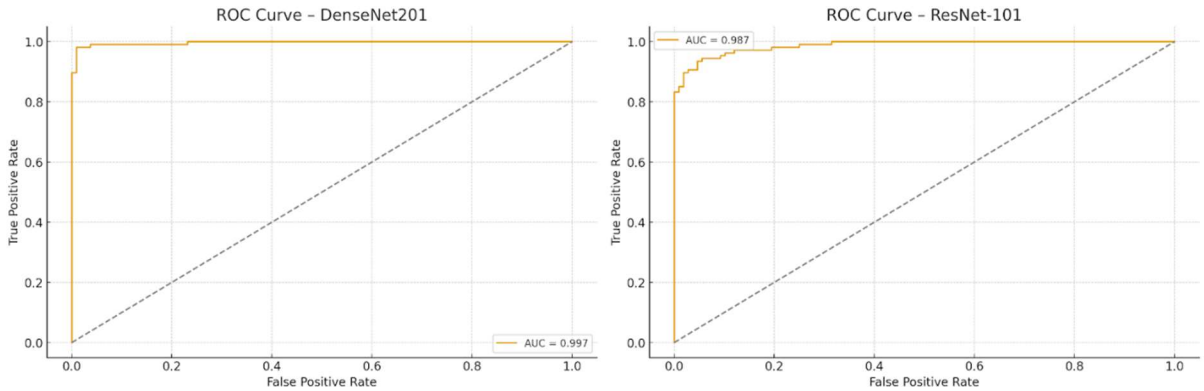


Figure 4.2.11 Roc Curve DenseNet201 and ResNet-101

As revealed by the ROC analysis presented in Figure 4.2.11, DenseNet201 has better discriminative performance as compared to ResNet101. DenseNet201 achieves ROC-AUC of 0.9981, which indicates almost perfect ranking over thresholds. It has test accuracy 98.60, precision 0.9906, recall 0.9813 and F1 0.9859 at its best-F1 operating point (threshold = 0.6682), and has operating points near to the upper-left of ROC space (high TPR, low FPR), and is reliable in clinical practice where both sensitivity and specificity are important.

The ROC-AUC of ResNet101 in comparison is lower but solid with a 0.9753 as presented in Figure 4.2.11. With its best-F1 configuration (threshold = 0.5111) it achieves test accuracy 94.42, precision 0.9439, recall 0.9439 and F1 0.9439. Although it detects the majority of positives, its curve is not as steep, and lies much closer to the upper-left corner than DenseNet201, meaning that it has a relatively higher error rate at similar thresholds. Sensu stricto, DenseNet201 has better ROC properties and better operating range to high stakes diagnostic choices.

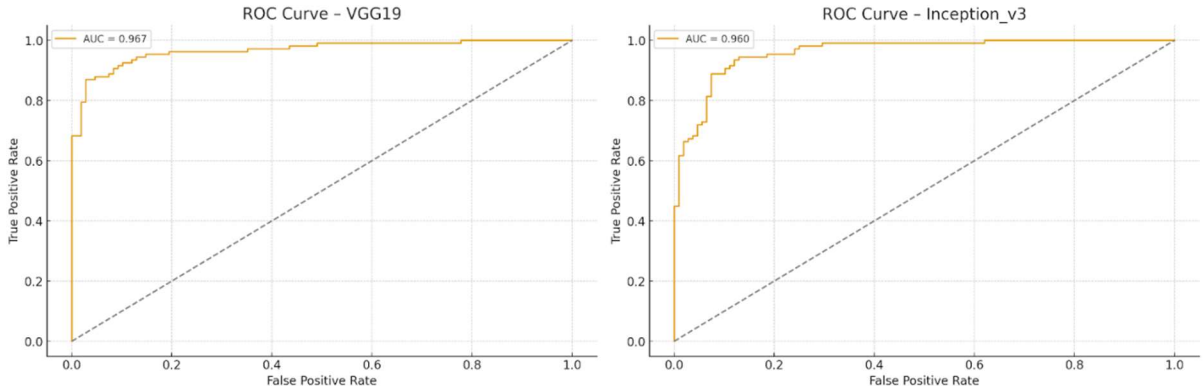


Figure 4.2.12 Roc Curve VGG19 and Inception_v3

Figure 4.2.12 shows that VGG19 differentiated at a moderate level with ROC-AUC of 0.967. The curve is more distant to the upper-left corner than more robust models, as it represents its precision-heavy but recall-limited behaviour at the optimal F1 operating point: test accuracy 92.09%, F1 = 0.9163, precision = 0.9688, recall = 0.8692. Pragmatically, this implies that VGG19 is a conservative predictor - when it makes a fracture prediction it is likely to be correct, but misses more positives than desired, which is expected in a confusion pattern with fewer false positives and more false negatives. The ROC shape and the metrics used to track the separation thus suggests poorer separation when compared to DenseNet201 although the overall discrimination is decent.

InceptionV3 had a disproportionately higher ROC-AUC of 0.9701, as reported in Figure 4.2.12, and has a more balanced, though still recall-biased, trade-off: test accuracy 90.70%, F1 = 0.9099, precision = 0.8783, recall = 0.9439. The curve is shifting more toward the sensitizing territory, which means that the number of false-positives will be relatively lower but the false-positive rate will not be trivial. Although this is more discriminative

than VGG19, it nevertheless is not as upper-left concentrated and headroom as DenseNet201 which emphasizes that both VGG19 and InceptionV3 are not as clinically sound as the more formidable base when it comes to high precision and high recall concomitantly.

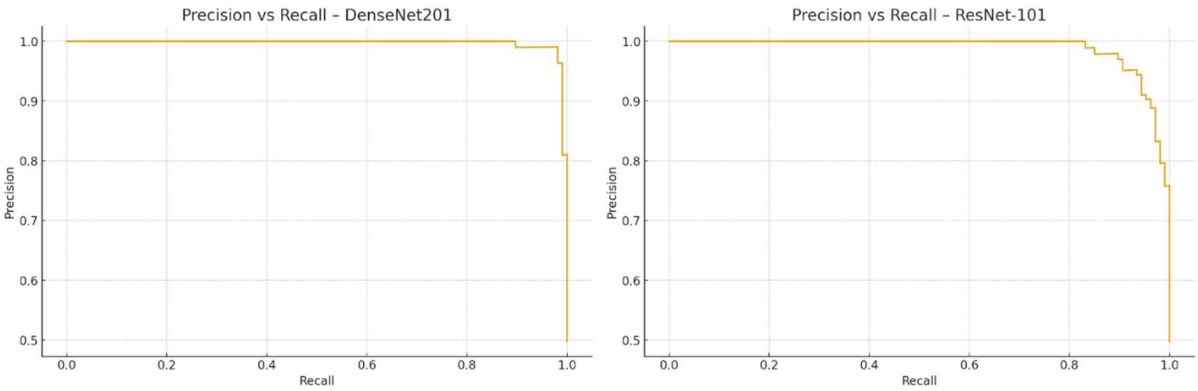


Figure 4.2.13 Precision vs Recall DenseNet201 and ResNet-101

According to the precision-recall plot in Figure 4.2.13, DenseNet201 has a rather even trade-off over thresholds. The precision increases continuously, and the recall is high throughout a wide operating range to reach the highest F1 = 0.9859, precision = 0.9906, recall = 0.9813 at the best-F1 point of threshold = 0.6682. This operating mode maintains an almost perfect sensitivity and a very high precision profile, an attribute that is desirable in clinical applications whereby the reduction of false negatives is important without experiencing excess false positives.

In comparison, ResNet101 is weakly balanced in the same plot in Figure 4.2.13. With a threshold of = 0.5111 (best-F1), the model achieves F1 = 0.9439 with precision = 0.9439 and recall = 0.9439. Even though the recall is high, the curve remains consistently lower compared to DenseNet201, meaning that the error rate is higher at similar operating points. The overall results of the precision-recall scores indicate the higher sensitivity-precision trade-off of DenseNet201, whereas ResNet101, although a good model, has a weaker high-stakes diagnostic performance profile.

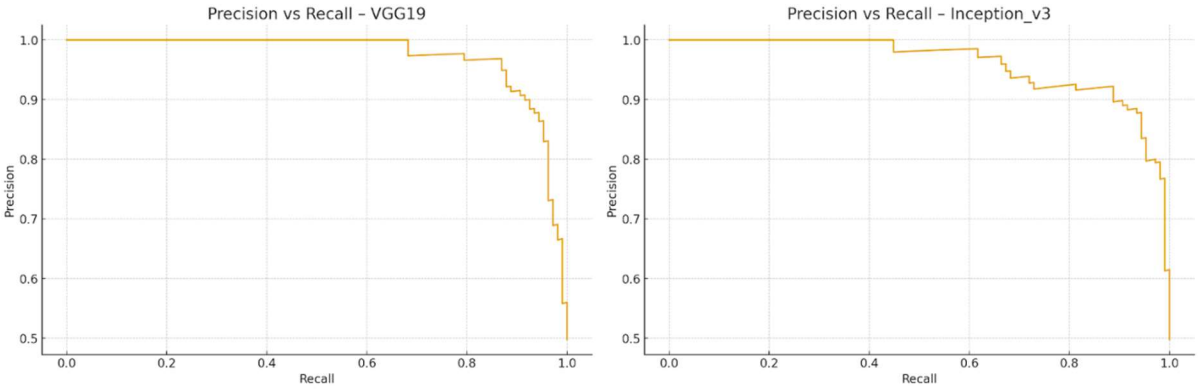


Figure 4.2.14 Precision vs Recall VGG19 and Inception_v3

Precision-recall traits of VGG19 imply a recall-limited but precision heavy behavior as illustrated in Figure 4.2.14 .When the threshold is set to 0.6830, VGG19 achieves precision = 0.9688, recall = 0.8692, and F1 = 0.9163, meaning that VGG19 is a conservative model- when it predicts fracture it is more likely to be right, but it misses more positives than it would prefer. The curve across thresholds presents this trade-off: there are gains in precision but also gains in recall, and usefulness must be limited where sensitivity needs to be the most important. This results in the fact that despite the ability of VGG19 to generate clean positive calls, its profile of precision is more inappropriate in clinical applications that focus more on false negative avoidance, and is otherwise less balanced than higher-performing baselines like DenseNet201.

InceptionV3 demonstrates more balanced, but trade-off biased towards the recall, as demonstrated in Figure 4.2.14. The model has a precision = 0.8783, a recall = 0.9439, and F1 = 0.9099 at its best-F1 (threshold = 0.1328), which provides relatively few missed fractures at a relatively moderate level of false-positivity. The curve lies nearer to the high-recall zone than that of VGG19, however, the curve falls short of the precision-recall balance of DenseNet201. In general, InceptionV3 outperforms VGG19 at useful operating points, but does not have as strong clinical behavior as DenseNet201 when both high precision and high recall are critical.

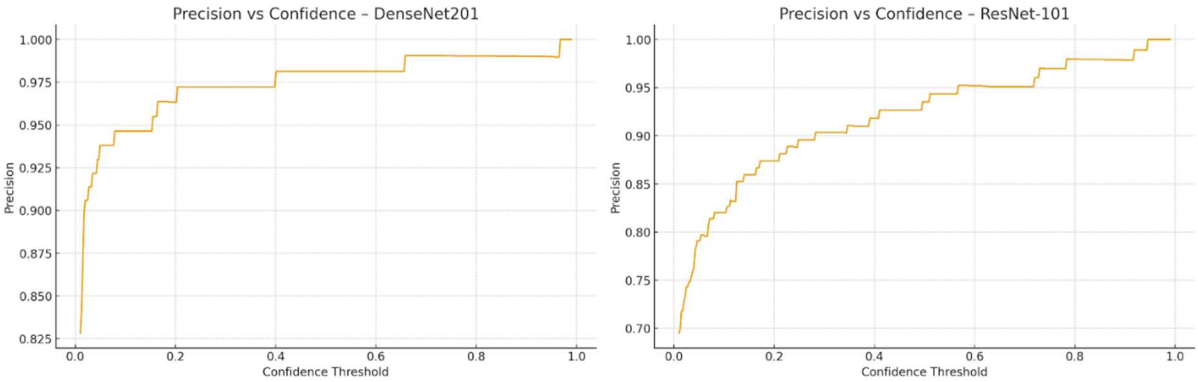


Figure 4.2.15 Precision vs Confidence DenseNet201 and ResNet-101

Figure 4.2.15 indicates that denseNet201 has a high accuracy over all levels of confidence. Accuracy improves with confidence (threshold) levels and recall is good across a wide operating point; at the optimal-F1 operating point (threshold = 0.6682), DenseNet201 achieves precision = 0.9906 and recall = 0.9813 (F1 = 0.9859), which means that false positives are well-controlled without significantly compromising the sensitivity. After this, further increasing the confidence does not provide significant improvements in precision, which supports the idea that DenseNet201 is already in a highly clinically viable range with a high sensitivity and high precision.

Figure 4.2.15 shows that ResNet101 has a flatter precision-confidence profile. Precision Curves more rapidly but because of higher confidence Saturates sooner than DenseNet201; at its best-F1 point (threshold = 0.5111) the precision = 0.9439 and recall =

0.9439 (F1 = 0.9439). Although this is a good performance, this curve means that there is less room to suppress false positives with similar recalls to achieve this performance with minimal headroom, thus lowering this reliability ceiling compared to DenseNet201. Practically, the precision-confidence behavior of DenseNet201 is more suitable in the situation of the high stakes of diagnostic application when both high sensitivity and precision are needed.

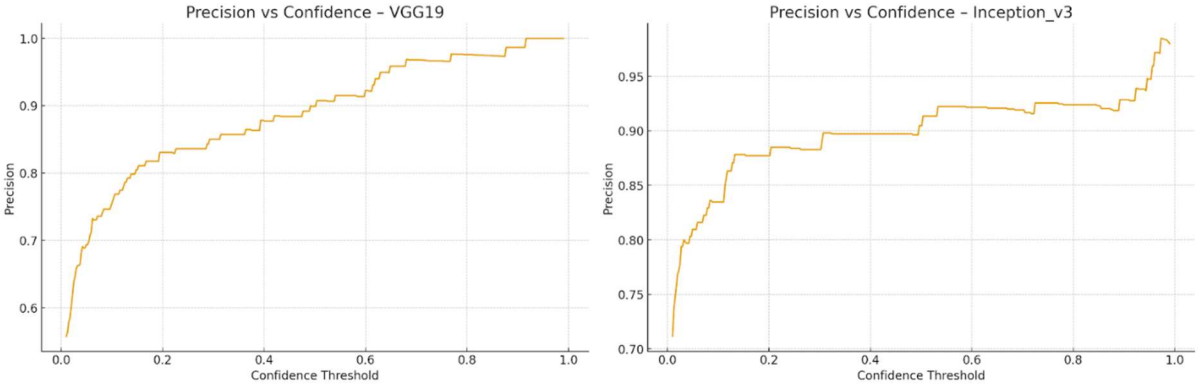


Figure 4.2.16 Precision vs Confidencel VGG19 and Inception_v3

The VGG19 precision-confidence curve demonstrates that false positives are continuously difficult to suppress as demonstrated in Figure 4.2.16. Precision increases with more stringent confidence settings, but at lower confidence settings reaches diminishing returns: at the most effective F1 operating point (threshold = 0.6830) the precision is 0.9688 with recall 0.8692 resulting in F1 = 0.9163 and a test accuracy of 92.09%. This is a precision-heavy, recall-limited profile: VGG19 predicts a fracture correctly when doing so but predicts many positives that it misses and sets the threshold to higher than the optimum to get a small amount of extra precision but misses much more recall.

InceptionV3 has a more equalized precision-confidence curve, as presented in Figure 4.2.16. Accuracy gets high as confidence gets high and recalls goes down, with the optimal F1 being threshold = 0.1328 which has accuracy of 0.8783, recall 0.9439, F1 0.9099 and test accuracy 90.70. This is fewer missed fractures than VGG19 at useful operating points, but still a non-trivial level of false-positive occurs and the curve does not hit the sensitivityprecision tradeoff of stronger baselines (e.g., DenseNet201).

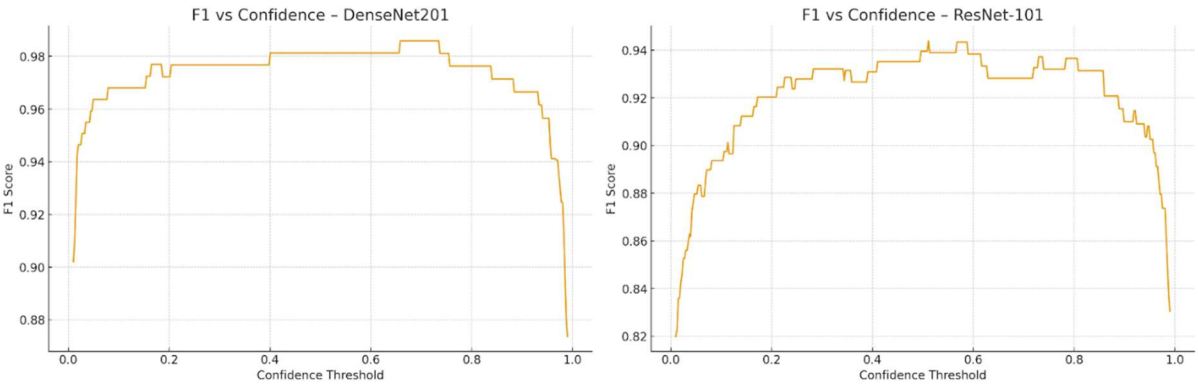


Figure 4.2.17 F1 vs Confidence DenseNet201 and ResNet-101

The F1-confidence profile of DenseNet201 looks much more favorable, as observed in Figure 4.2.17. The curve steeply increases within a wide range of confidence, and then it levels off, which means that harmonic balance between precision and recall is stable. DenseNet201 reaches F1 = 0.9859, precision = 0.9906 and recall = 0.9813 at the best-F1 operating point (threshold = 0.6682), indicating that it has very good control over false positives without compromising on sensitivity. Further tightening of the confidence offers only little F1 changes- diminishing returns that strengthens the model to operate well in clinical situations in which both reliability and consistency are of importance.

In comparison, ResNet101 has a low ceiling and sensitivity to threshold movement as in Figure 4.2.17. Its optimum point is threshold = 0.5111 with F1 = 0.9439, precision = 0.9439 and recall = 0.9439 which has a lower harmonic balance than DenseNet201 and a smaller plateau region. Although the model still has high performance, the F1 curve shows less space to increase precision without reducing recall and is thus relatively weaker when it comes to using the model in a high stakes diagnostic setting where minimizing both false negatives and false positives is paramount.

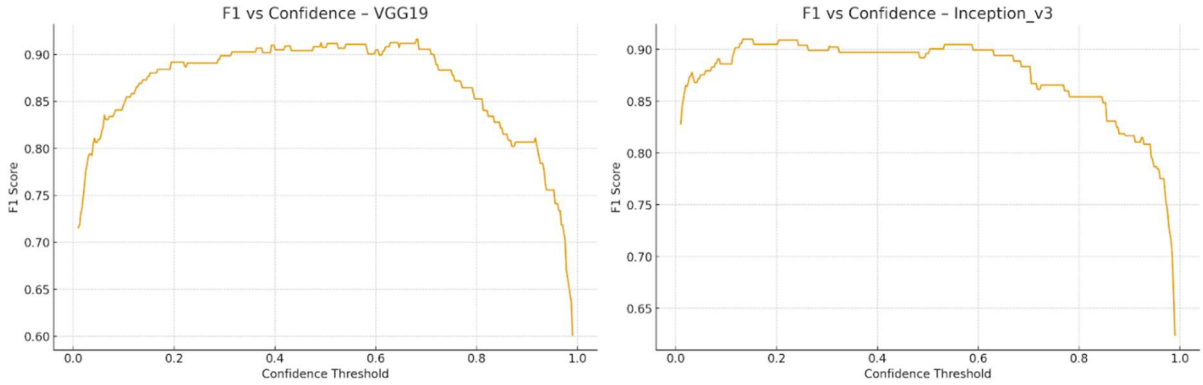


Figure 4.2.18 F1 vs Confidence VGG19 and Inception_v3

According to the F1-confidence curve of VGG19, the balance between precision and recall is hard to trade because at the points where the figure reaches its maximum, the curve is at its lowest point as seen in Figure 4.2.18. F1 increases as confidence increases and peaks at the optimum F1 = 0.9163, precision = 0.9688 and recall = 0.8692. This is an indication of a precision-heavy/recall-limited profile: lowering the threshold will eliminate false positives effectively but increase those who are really positive without being detected, and raising the confidence will reduce gains in precision with no further gains in recall. As a result, although positive predictions made by VGG19 are usually correct, the harmonic balance is limited by the recall, which restricts the use of this method in the applications where the inability to see a real fracture could not be accepted.

Figure 4.2.18 indicates that InceptionV3 is more balanced in F1-confidence values. The

curve increases and levels off around its best-F1 point (threshold = 0.1328) at the point where F1 = 0.9099, and precision = 0.8783, and recall = 0.9439. However, this regime suggests a relatively reduced number of missed fractures versus VGG19 at useful operating points with a non-trivial false-positive rate. Although InceptionV3 thus provides a fairer precision-recall tradeoff, the stalled F1 still scores lower than other more powerful baselines (e.g. DenseNet201), which highlights that neither the precision nor the recall can be maximized equally without extra modelling or downstream validation.

4.3 Results and Discussion

In all five of the models, including UNet++, DenseNet201, ResNet101, VGG19, and Inception V3, DenseNet201 proves to be the most predictable and consistent on training, validation and testing. UNet++ was measured as very strong in general (validation accuracy 96.7; test accuracy 95.61, F1 0.9001, precision 0.8873, recall 0.9132 at threshold 0.3661), but operating point is recall dominant (specificity is less important, but sensitivity is). ResNet101 achieved an ideal training accuracy but demonstrated more indicators of overfitting (validation 94.4; test accuracy 94.42, F1 0.9439). VGG19 and InceptionV3 had slower or less stable convergence (validation 91.6% and 94.0%), and lower test scores 92.09% and 90.70% accuracy F1 0.9163 and 0.9099 respectively), and are not as strong to use in high-stakes situations where sense and specificity are both essential.

DenseNet201 is able to combine rapid, robust learning (100% of training accuracy, low training loss) and strong generalization (99.1% of validation, and 98.60% of test accuracy). It had F1 0.9859 with a precision of 0.9906 and a recall of 0.9813 at an F1 operating point of 0.6682 and ROC-AUC of 0.9981 and PR-AUC of 0.9982. This is practically to say that DenseNet201 achieves near-perfect sensitivity without maintaining false positives particularly low, which is particularly useful when it comes to medical screening and diagnosis. DenseNet201 was the most consistent in precisionrecall tradeoff with threshold and strongest held-out performance when compared to the other architectures, reasserting its claim of the most robust and clinically useful architecture in our radiographic analysis.

Tabil 4.1 Test Matrix of Each Model

Model	Accuracy	F1 Score	Threshold	Precision	Recall
UNet++	0.956062	0.900101	0.366140	0.887330	0.913246
DenseNet201	0.986047	0.985915	0.668246	0.990566	0.981308
ResNet101	0.944186	0.943925	0.511053	0.943925	0.943925
VGG19	0.920930	0.916256	0.682982	0.968750	0.869159
InceptionV3	0.906977	0.909910	0.132807	0.878261	0.943925

4.4 summary

The detailed comparison between UNet++ and DenseNet201, ResNet101, VGG19, and InceptionV3 reveals a strong tendency in terms of training, validation, and testing. DenseNet201 is the most stable, dependable model on this research. It achieved validation accuracy of 99.1 percent and maintained the same level of accuracy of 98.60 percent on the test set. UNet++ was also able to train and validate exceptionally well, with a high validation accuracy of 96.7 percent and the highest raw test accuracy at 95.61 percent. ResNet101 had a train to validation gap, and it has a higher training accuracy with 94.4 percent on validation and 94.42 percent on test set. VGG19 and InceptionV3 did not converge as fast or as stably and had 91.6 percent and 94.0 percent on validation respectively. VGG19 had an accuracy of 92.09 percent and InceptionV3 had an accuracy of 90.70 percent on the test set V. The presence of these headline results already suggests that DenseNet201 is more generalizable than the others and that UNet++ has very strong overall recognition and is biased towards sensitivity.

DenseNet201 was able to balance and stabilize at its working point. In the best F1 setting, which was with threshold 0.6682, it obtained F1 0.9859 with precision 0.9906 and recall 0.9813. The ROC AUC of it was 0.9981 and PR AUC 0.9982. The confusion pattern suggested that there were relatively small numbers of false negatives and false positives at useful thresholds. The validation loss was also small and the gap to training loss also was small with a consistent gap between the validation loss and training loss as the epoch occurred later. This profile is highly consistent with clinical needs, since the model does not miss any alarms and at the same time does not produce redundant alarms. Simply put, DenseNet201 produced the best harmonic balance between sensitivity and specificity and it did so uniformly at every stage of the pipeline.

UNet++ provided the best test accuracy but its ideal operating regime is recall dominant. It had a precision of 0.8873, recall of 0.9132 and F1 of 0.9001 at the best F1 point of 0.3661. ROC AUC was 0.9988 and PR AUC was 0.9712. This validates a first line position that is best for screening and triage with a preference to pick nearly all positives over minimizing all of the false alarms. The operating option can be read easily. When sensitivity is the required factor, UNet++ has a great coverage and yet good overall quality, at the cost of the precision being lower than DenseNet201.

ResNet101 gave good but less robust generalization. The maximum F1 of the model was 0.9439 at F1 threshold 0.5111, and the precision and recall of the model were both 0.9439. The accuracy of the test was 94.42 percent and validation was 94.4 percent. The ROC AUC was 0.9753. Training curves were high though the validation curve flattened sooner than the DenseNet201 and remained below this. The pattern of confusion indicated that there are more errors at similar operating points this restricts reliability to the most intensive diagnostic environments.

VGG19 and InceptionV3 followed in consistency on unknown data and had reverse preferences at their highest F1 scores. VGG19 was recall limited and heavy precision. It had threshold 0.6830 providing F1 0.9163 at a precision of 0.9688 and recall 0.8692, and test accuracy 92.09 percent. The ROC AUC was 0.967. This implies that the hits are normally accurate but more hits are lost than would have been desired. InceptionV3 was recall biased and a little bit more balanced than VGG19, but it tolerated more false positives. It has a threshold of 0.1328 and F1 0.9099 with precision of 0.8783 and recall of 0.9439 and test accuracy 90.70 percent. The ROC AUC was 0.9701. Both these models thus come after the DenseNet201 in the shared quest of a very high sensitivity and a very high precision at realistic thresholds.

Combining all evidence, in this comparison, DenseNet201 is the most accurate, the most stable and the most useful architecture. It maintains a high level of validation accuracy, maintains a high level of test accuracy and provides the optimal harmonic balance at the decision time. UNet++ is outstanding in situations in which the primary objective is the maximum sensitivity, and the decline in accuracy in small margins is permissible. ResNet101 is competitive and weaker in shift. VGG19 is biased towards clean accuracy with missed positives. InceptionV3 is more keen on misses at the expense of false alarms. In the case of tasks that require high sensitivity and specificity and reliable behavior in different settings, the choice would be DenseNet201.

Chapter 5

Engineering Standards and Design Challenges

5.1 Compliance with the Standards

5.1.1 Software Standards

In this project, software standards were chosen wisely so that different stages of development were reliable, consistent, and repeatable. Because the research was done in the cloud, a focus was placed on accessibility, scalability, and effective collaboration. The tools implemented are well-known, having been transferred from machine learning and medical imaging research, in order to ensure the project was replicable and extendable by other researchers or clinical teams.

The underlying programming language was Python, and Python was the core language for the various modules of processing, classification, and explanation. With its robust introspection, rich functionality, and comprehensive libraries available, Python was the natural choice for manipulating image data, training models, and evaluating performance. All experiments have been run on Google Colab Pro, the cloud hosted platform with access to GPU acceleration. It has lowered hardware demands at our local sites, simplified the computation environment and ensured repeatability across training runs. Using the libraries in COLAB which had already been installed, decreased the time spent on setting up the environment, which is a good thing when you're primarily focused on experimenting.

For deep learning we used TensorFlow/Keras as the deep learning library of choice for training and testing the model. The high-level APIs allowed to implement convolutional neural networks, specifically the DenseNet201 backbone that is used for this project. PyTorch was also used for ancillary tasks which will come in handy if you wish to test smaller pieces of the pipeline. Both implementations are considered best practice in the AI research community and their inclusion was intended to provide reproducibility and enable subsequent researchers to extend the pipeline with the toolset of their choice.

NumPy arrays and Pandas Dataframes were effective data containers for pre-processing, storage and analysis of medical imaging data. In conjunction with Google Drive it was

easy to access datasets, trained model checkpoints and experimental logs. Reproducible experiments: data was stored in structured directories that represented train/validation/test splits. GitHub was used as a version control system to record the project code, incremental development and notes from experiments. Not only did this allow for transparency in model design - it also offered a collaborative platform for sharing updates and brainstorming issues.

There was much to be gained by the use of these standards. In addition, the deep learning libraries available in Python were scalable and support GPU environments, and Colab made reproducibility easy by allowing anyone with an internet connection to rerun experiments under the same conditions. Stanford University: We used NumPy and Pandas for simplicity during preprocessing and analysis and GitHub for an open record of code changes. On the other hand, Google Drive ensured that there was data integrity and consistency, while allowing for data to be shared between different sessions.

There were challenges, however, along the way. Unstable Training Time: Colab pro is notorious for its limited time and GPU quota, so sometimes, longer training runs get interrupted unexpectedly, making it essential to have checkpoints in place. For large data sets, the storage and I/O performance of Google Drive was slower than for local solid state drives, which sometimes caused bottlenecks. Further, different versions of TensorFlow and PyTorch libraries sometimes didn't work well together and required manual adjustments to get things to work.

Overall, the decision of using Python, Google Colab Pro, TensorFlow/Keras, PyTorch, Numpy, Pandas, Google Drive, and Github as standards of programming, cloud platform, libraries and frameworks and version control set up a successful and approachable foundation for the project. While a number of limitations of cloud-based infrastructure were discussed, a consensus was reached that the benefits in terms of reproducibility, collaboration and extensibility outweigh the limitations. Using this set of standards, a technically sound artifact-resilient fracture detection pipeline was designed with practical deployment considerations in mind.

5.1.2 Hardware Standards

This project was done mostly in cloud environment like Google Colab Pro, so hardware standards were selected with the stability, accessibility and reproducibility in mind. The training process was based on GPU acceleration. The classification and artifact removal tasks were accelerated using NVIDIA Tesla-class GPUs (such as T4 or P100) available in Colab. As a result, training time was greatly reduced and has allowed large X-ray datasets to be processed with deep models such as DenseNet201 and UNet++. However, this also imposed limitations, as the use of GPU in Colab is time-quota and session-preempted, so we needed to carefully checkpoint the model weights.

In addition to GPUs, CPU and memory resources available through Colab played a very important supporting role. With CPU profiles as small as 2 and as large as 8 CPU cores, and up to 32 GB of RAM, preprocessing tasks (e.g., cropping, augmentation, artifact cleaning) were performed effectively. While this provided ease of processing moderately large datasets, extremely large datasets still tended to max out RAM, occasionally leading to session crashes, which meant that data batching was still needed.

For storage and dataset management, we used Google Drive integration as the default. Datasets and checkpoints were organized into well-structured directories for training, validation and testing. The solution offered intersession and resumption of work, but was relatively slow in terms of I/O speed compared to local solid-state drives, and was also dependent upon Internet connectivity and storage quota.

Another important criterion was model check point. Training model weights were saved periodically to Google Drive, and the weights corresponding to the best validation results were selected. This prevented the loss of data due to unexpected session termination and made model comparison easy. On the negative side, recurrent checkpointing used more space and required manual cleanup from time to time.

Finally, the flexibility of deployment hardware was taken into account while designing. The trained models were optimized for GPUs and CPUs, so they can be used in Colab, in hospital servers, or in telemedicine setups with more limited resources. This flexibility made it more likely to see real-world adoption, albeit with the price of possibly slightly worse accuracy due to model simplification for CPU execution.

In summary, the hardware requirements-cloud GPUs for training, CPUs and RAM for preprocessing, Google Drive for storage, structured checkpointing for reliability, and easy-to-deploy portable outputs-ensured that the project was a fair balance between performance and convenience. While the limitations of the cloud such as Colab session limits and slow cloud storage were problematic, ultimately, the standards enabled reproducible and scalable experimentation.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

The effective deployment of an artifact that detects bone fractures resiliently has the potential to have numerous implications on patients, clinicians and the overall healthcare ecosystem. On the one hand, such a system has a direct beneficial effect on patient care, as it allows to make an early and reliable diagnosis, avoiding complications of missed and late diagnosis. This particularly applies to vulnerable groups like children and older adults where fractures could be more insidious that can result to permanent disability unless treated accordingly [7]. It is also more transparent in the sense that it employs state-of-the-art explainable AI (XAI) techniques (i.e., Grad-CAM++), and generates visual heatmaps that assist in identifying the areas of fractures. With such overlaid, clinicians would be able to confirm AI-based predictions and, therefore, introduce credibility and minimize resistance to using AI-based diagnostic assistance [9].

Clinicians also are at an advantage as far as the workload is concerned in terms of diagnosis. Another application of AI is as an assistant tool to non-specialist physicians in high-volume ERs and resource-constrained hospitals that frequently do not have specialist radiologists. Repeated image classification tasks can be classified to enable the system to focus the attention of the clinicians on the difficult cases, which involve human expertise. This comes in especially handy in an underdeveloped or rural health care setting where the access to specialized knowledge is not usually timely [19]. Also, research has revealed that AI not only enhances the quality of diagnostic accuracy, but also the inter-rater agreement between clinicians, which subsequently results in greater levels of consistency and confidence in medical diagnosis reaching [20]. These consistency enhancements can particularly be applicable to the minimization of variation in care quality between institutions.

Although these are the good aspects of the system, there are issues related to the use of the system. One of the most dire of all problems is the risk of over-reliance on the outcomes of AI. Clinicians also may be persuasive by AI predictions, which may not be taken at all, with disastrous consequences of misdiagnosing, particularly in rare/atypical cases [15]. This risk is increased taking into consideration implants or other anatomical malformations in which AI systems are unable to identify artifacts as fractures or fail to recognize small injuries [12]. The other constraint is equity of access: although AI solutions may give smaller hospitals a competitive edge, their adoption will be skewed toward technologically state-of-the-art hospitals. This gap may also increase the healthcare gap when not addressed because the institutions that are already resource constrained will be left behind [7]. Besides, the ambiguity of the message about the role of AI-based care may result in losing trust in AI among patients, and when a portion of the diagnostic task is also referred to the machine, ethical and psychological issues arise

[19].

Combined, the effects of this mechanism on life are one that works by both slender threads between chance and danger. The technology is also seen to possess a clear potential on changing the healthcare system with regards to the speed with which an accurate diagnosis will be reached, the pace at which accuracy will be enhanced and the decreases in the workload of the clinician. It should be noted, though, that continuous validation, transparency and clinical supervision are needed to make sure that AI can be applied as an instrument of augmenting, not substituting human expertise. A properly implemented system can aid in patient safety, clinician confidence, and result in better healthcare outcomes, in general, when applied with intelligence.

5.2.2 Impact on Society & Environment

The introduction of AI-based bone fracture detection technologies has broad social and environmental implications, and both the positive and negative aspects are evident. Democratization of medical expertise is one of the greatest benefits on a societal basis. The distressing disparities between urban and remote populations can be mitigated through the provision of automated diagnostic systems capable of offering support comparable to the level of specialists at a rural or under-funded clinic where radiologists access is restricted [7]. This means that patients in underserved places will be able to receive a more accurate and quicker diagnosis, reducing the long-term consequences of untreated fractures on families and communities. Such systems can also contribute to better health results and create a more positive perception of AI-based healthcare technologies through diagnostic errors and offering consistency, which, in its turn, leads to higher trust in healthcare technologies on the part of the population [19].

In terms of the environment, the implementation of AI can also minimize unneeded medical treatments. In addition to reducing radiation to the patient, AI reduces the power used to operate the imaging devices by making the scan rate of the device less frequent [20]. Moreover, one can diagnose the patients remotely using telemedicine-ready and lightweight models, such as MobileNetV2 or efficient variants of the YOLO models, thereby cutting down on the amount of travel required, which lowers carbon emissions associated with healthcare logistics [18]. Cloud-based deployment also has the added advantage of centralizing computation resources, which when used effectively can be more energy-efficient than local, decentralized infrastructure [14].

Nevertheless, there are a number of issues that should be identified. Socially, a risk of algorithmic bias occurs in cases whereby datasets are not representative of various populations. Using the example of such a model, which has been initially trained on a population of pediatrics or Europeans, may not work equally well when applied to other populations, and may even amplify their inequalities [19]. Further, heavy dependence on AI would result in denying clinicians their autonomy and will also cause concern among

health practitioners and also will become a cause of mistrust in case the system is perceived to usurp the skill of human beings [15].

There are also high environmental costs. Neural networks such as the DenseNet201, YOLOv7, or transformer require a high-performance GPU and long computational times, which use a lot of electricity and add to carbon emissions [10].. Although cloud systems concentrate computation, they do not remove the load on the environment; when powered by non-renewable energy, they just move the load to another environment. In addition, the rapid development of AI technology contributes to the timely replacement of hardware, which brings up the question of e-waste disposal, which is one of the insufficiently discussed aspects of the sustainability of medical technologies.

Overall, AI-based fracture detection has the potential to generate significant social benefits by enhancing equity, accuracy, and confidence in healthcare, as well as generate environmental benefits by decreasing travel and increasing resource utilization. But such benefits should be weighed against the dangers of bias, professional resistance, energy demands, and e-waste issues. The datasets will need to be carefully designed, there should be a transparent clinical integration as well as eco-friendly infrastructure management to ensure that the societal and environmental benefits surpass the costs needed to implement them in a sustainable way.

5.2.3 Ethical Aspects

The implementation of AI in the area of breaking detectives also has both negative and positive ethical consequences, especially in respect to how medical data is gathered, processed, and used. On the one hand, AI has the potential to transform the quality of diagnosis to a new level by removing the inter-observer effect and minimizing human influence in an effort to deliver more consistent and fair patient results [9]. Moreover, when explainable AI (XAI) approaches (e.g., Grad-CAM++, Integrated Gradients) are deployed, it also gives an additional degree of transparency, allowing clinicians (and possibly even patients) to visually examine why a given prediction was conducted [2]. Not only will this build confidence in the role of artificial intelligence in assisting in the diagnosis, but it will also enable shared decision making between patients and doctors.

On the upside, we can use anonymized medical datasets in a responsible manner. Due to the wide application of large-scale imaging repositories, including GRAZPEDWRI-DX, researchers are able to create powerful algorithms due to the absence of worries on the loss of patient confidentiality [15]. The data is kept and safeguarded against sensitive patient data by way of de-identification processes, secure storage and institutional review boards. Moreover, the necessity of informed consent implies that the rights of patients are considered and their anonymised data will be used to invent progressions that will decrease the error rates in diagnoses and enhance the provision of healthcare [20].

There are however, certain critical ethical problems that remain to be addressed. The anonymization is not flawless, and this is the main concern of patient privacy; the re-identification of the medical images by cross-linking it with other data will result in confidentiality breach [19]. The other issue is that of dataset bias. Many of the fracture detection models are trained on data that is over-represented by some demographics (e.g. pediatric or European cohort) and therefore may, by extension, perform poorly when used with under-represented groups [15]. And this unfairness is what leads to drastic constraint of a wide spread application of AI-based diagnostic machines.

The other ethical issue of concern lies in accountability. In case any AI system fails to classify the fracture correctly, it has not been established whether the error was on the side of the model developers or the institution, or on the side of the clinician that used the AI result. This confusion may cause a lack of trust in the technology and the medical system in general [15]. When a physician over depends on AI without adequate human supervision, this would negatively affect professional judgment, reducing clinical vigilance, and ultimately leading to negative outcomes. Finally, the computationally-intensive methods like training DenseNet201 or YOLO-like models also have ethical concerns since the corresponding power usage indirectly opens the door to health problems in the community due to the creation of more carbon emissions [10].

To sum up, there is a thin line between innovation and responsibility in ethical application of AI to fracture detection. However, although AI can enhance fairness, transparency and outcomes, the issue of privacy, bias, accountability and sustainability still require control under the umbrella of robust ethical systems, continuous validation and explicit positioning of AI as an assistant (not a substitute) to clinical decision-making.

5.2.4 Sustainability Plan

AI-driven bone fracture detection should have increasing efforts to devise a sustainability plan which will care for the long-term existence of the system in clinical application, ethically sound utilization, and environmentally sustainable. On a positive note, the use of lightweight and optimized deep learning models like MobileNet and efficient YOLO variants gives us a clear route to alleviate the computation overhead. These models are substantially more energy efficient than very deep CNNs or transformer-based models, making them appropriate for deployment on modest hardware available in low-resource hospitals and telemedicine platforms [7]]. This scalability is in particular important for health care systems in rural areas or with insufficient financial resources to purchase powerful GPUs, thus allowing broader availability of such advanced diagnostic support , [14].

Another strength of the proposed approach is that the pipeline is modular: preprocessing (artifact removal), classification, and XAI are implemented as independent but interoperable modules. This modularity will allow it to change as the datasets grow or as new imaging standards become available without having to entirely re-implement the system . If well-maintained, cloud deployment also contributes to further flexibility in combining resources across institutions, reducing costs, and providing ability to frequently update the model. In addition to technical efficiency, seamless incorporation of interpretability like Grad-CAM++ enables the system to maintain long-term clinical trust, which is a vital component of sustainability in healthcare AI [1], [14].

At the same time, sustainability issues cannot be ignored. First, extensive amounts of energy are used for training and retraining deep learning models on large-scale datasets (which can increase with the architectural complexity of models such as DenseNet201 or YOLOv7), raising questions about their long-term environmental footprint [8], . By no means is this burden dropped as the cloud is a means to migrate that burden elsewhere; to these vast data centers themselves, which themselves require a vast amount of energy and rely on the source of its supply. Data management of cloud-based deployments is another challenge as the long-term dependence on third-party providers introduces risks of data security, privacy compliance, and dependency [7]. In addition, fairness in an AI system would need to be continually monitored through periodic training and retraining of the dataset to prevent demographic or implant-related bias. As needed, these updates result in increased operational costs and carbon emissions associated with extra computation [12]. Another challenge is that of socio-ethical sustainability. However, if clinicians view AI systems as being opaque or too time consuming to fit into current workflows, they will be less inclined to adopt them, irrespective of technical performance and this will potentially negate any benefit [19].

To counteract these risks, the sustainability plan states that lightweight models should be chosen carefully, to ensure the technical efficiency is not sacrificed for accuracy. Hybrid deployment schemes that use edge devices for preliminary screening with access to cloud infrastructure for further processing give a trade-off between scalability and efficiency [14]. Periodic fairness and bias audits are required to ensure equal diagnostic performance; clinician-in-the-loop validation ensures trust and accountability; and Ultimately, the sustainability in this project is not only in terms of computational efficiency, but in the fact that the resulting system will need to evolve with technological, environmental, and ethical challenges while staying accessible and clinically relevant.

5.3 Project Management and Financial Analysis

To make any research project successful, a well-designed budget plan must be at its core, especially when the computational components of the project (such as deep learning model training and assessment, etc.) require more computing power. This project had two different and specific budget plans where the execution would be carried out and made sustainable. The former is based on cloud-hosted infrastructure using Google Colab Pro, and the latter focuses on making permanent investments in a high-performance workstation by owning a GPU. The two strategies have advantages and disadvantages, which have been well examined to help determine their appropriateness at various research phases and their possible use in the clinic.

The affordability and flexibility-oriented strategy based on Colab Pro was referred to as the baseline one. The plan is estimated to require a total cost of BDT 266,920 in eight months as detailed in Table 5.1, which will be used in subscription fees on Colab Pro, using a mid-range workstation to perform preprocessing activities, training sessions to enhance technical expertise, web hosting to conduct deployment experiments, utilities, and engineering support. The largest benefit is that it helps to lower upfront capital requirements; this is particularly useful when the work is an academic or proof of concept project with a limited budget. Furthermore, Colab Pro is the fastest way to access NVIDIA GPUs without having to buy and support costly hardware. This is why it is appealing in smaller projects or projects that have a discovery purpose. However, repeated subscription costs are a cumulative asset as time passes therefore this strategy is not very economical when the project runs more than a year. Also, reliance on Colab poses the threat of unavailability of GPUs, session loss, and the need to have a good internet connection. Such constraints can break long training sessions and reduce the reproducibility of results across runs.

Instead, the second approach focuses on long-term scalability and should be independent by spending money on a workstation with an independent GPU. This plan would be much more expensive to initiate than the baseline, with an estimated cost of BDT 304,320 as presented in Table 5.2. But it can remove redundant subscription fees and is therefore more economical in long term projects that are over 1 to 4 years. Ownership of GPUs offers computational independence, with the ability to perform experiments without outside interruptions or run time constraints. It is also a better method of data security because sensitive medical images are stored and handled at the same location and this minimizes the risk that could arise as a result of depending on the cloud. Additionally, the possibility of owning hardware can be expanded to more extensive datasets and more sophisticated experiments, and this comes at a particularly useful point, as the project transitions beyond academic prototyping, as well as to clinical trials or commercialization. The main shortfalls of this approach include hardware maintenance, possible costs of repair, and technological obsolescence. However, considering an average a GPU workstation is three to five years, the investment would be appropriate in the long run.

Both of such strategies were aligned to a sustainability roadmap, which takes into account revenue generation to cover costs and further development. Revenue models comprise the subscription-based SaaS licensing of hospitals and diagnostic centers, pay-per-use models of telemedicine architecture, and embedding modular systems (the UNet++ artifact-removal pipeline, DenseNet201 classifier, etc.) into the existing PACS infrastructure. Consultancy, technical support and training workshops might also provide further revenue, in addition to playing a greater role in facilitating knowledge transfer and promoting more effective clinical adoption. With these models, financial sustainability is guaranteed to complement technical scalability in order to make the system viable in both research and real-life healthcare settings.

Overall, the baseline Colab-based plan fits better into the short-term research and academic validation stages because it is cheaper to start with and easier to use, whereas the other plan of owning a GPU is better suited to long-term deployment, large-scale experimentation, and long-term integration of clinical processes. Although initial equipment capital remains a challenge, the benefits in autonomy, security, and scalability overcome the risks in projects that plan to use the equipment over a multi-year period and proceed to commercial use in the long term. Combined, the financial analysis and management plans offer a balanced approach that fits the short time academic objectives of the project, as well as the planning of its future.

Table 5.1 Baseline Cost Breakdown (BDT)

Cost Head	**Amount (BDT)**
Software — Google Colab (8 mo.)	96,000
Course/Training	12,600
Web Hosting	7,320
PC (Workstation)	80,000
Utility Bills	7,000
Engineer Bill (R&D effort)	120,000
Total	**2,66,920**

Table 5.2 Alternative Cost Breakdown (BDT)

Cost Head	Amount (BDT)
Course/Training	12,000
Web Hosting	3,320
PC with GPU (self-owned, optimized)	1,62,000
Utility Bills	7,000
Engineer Bill (R&D effort)	120,000
Total	**3,04,320**

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

Table 5.3 Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓		✓		

Justifications

EP1: Computer vision, deep learning, and clinical imaging knowledge are combined with the project to reach EP1. To achieve superior results, the architecture of UNet++ as well as DenseNet201 involves expert knowledge in artifact elimination and classifying, respectively, which is a highly technical skill set in the medical AI field.

EP2: The project faced conflicting needs: achieving high accuracy while keeping computation efficient. It also had to support both advanced hospitals and low-resource clinics, while ensuring ethical, social, and environmental responsibilities were addressed.

EP3:In this work we extensively benchmarked our model against baselines (ResNet101, VGG19, InceptionV3), did remove studies of preprocessing choices, and validated our model's interpretability. Further, our analysis goes beyond surface-level results to dig down to false positive/false negative causes to ensure robustness.

EP5: the work meets medical imaging standards ,ethical AI usage and privacy. Data is anonymised, held safely and utilised for research within an ethical code of approval. Explainability makes decisions clinically accountable

Mapping with Knowledge Profile

Table 5.4 Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
	✓	✓	✓	✓	✓		✓

Justification

KP2: Mathematics was applied to evaluate model performance using accuracy, loss, F1-score, precision, recall, and confidence scores. These metrics ensured reliable results, guided improvements, and supported reproducibility of the fracture detection system.

K3: The project incorporates the basic computer vision computational imaging concepts, such as pixel normalization, data augmentation, and CNN classification. These basics are required to introduce the reproducibility and the mathematical/algorithmic foundations of higher-level modeling.

K4: customized expertise is shown by adapting UNet++ as an artifact removing artifact and DenseNet201 as a fracture classifier The selection of Grad-CAM++ further reflects the application of explainability to medical AI in a domain-specific manner, taking a step beyond generic computer vision tasks.

K5: Combining preprocessing, classification and interpretability into an end-to-end pipeline. This workflow make the complete work with perfection with implement standard.

K6: Demonstrates applied engineering practice through Google Colab, as well as Hugging face space based deployment. In this way, the project is made theoretically sound, and also practically feasible for clinical/telemedicine settings.

K8: The methodology is based on a lot of previous research. For instance, preprocessing optimizations are related to Wu et al. [1], classification baseline to Nguyen et al. [2] and Aldhyani et al. [10], and interpretability to Paradava et al. [7] and Ye et al. [17]. It is important that through this engagement with the literature, the project adds to, and develops, the existing research environment.

5.4.2 Engineering Activities

Mapping with Complex Engineering Activities

Table 5.5 Mapping with Complex Engineering Activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓		✓	✓	

Justification

EA1: The project drew on a wide range of resources, including computational infrastructure such as Google Colab Pro for GPU access and a high-performance workstation for long-term scalability. Software libraries like PyTorch, TensorFlow, and OpenCV were used for model development, while Gradio and Hugging Face Spaces supported deployment and explainability. Radiographic datasets, refined through preprocessing techniques to remove artifacts.

EA3: The hybrid pipeline architecture that combines artifact removal with UNet++ and classification with DenseNet201 allows artifact learning and effective classification and discretization with Grad-CAM++ used to interpret the results. We assume the challenge of repairing an artifact that exists in X-ray data (uncharacteristic by most of prior research that has exclusively trained on clean images). By specifying the artifact suppression as a pre-processing task on its own, the project contributes to the novelty to

enhance the predictive value as well as the reliability of the explainable AI findings. The innovation is algorithmic as much as it is systemic and it combines data cleaning and modeling with clinical explainability.

EA4: The project has a direct impact on societal well-being by the improvement of fracture detection accuracy and the reduction of misdiagnoses, particularly where implants, screws or plates are present. With its focus on the actual fracture characteristics, the method will ultimately make the use of medical AI more trustworthy and transparent, potentially leading to a safer outcome for the patient. Additionally, as it is cloud-based or resource-constrained, it also extends the principle of equality of access, helping to reduce the gap in healthcare. From an environmental perspective, when compared to the oversized architectures, efficient model choices (MobNetV2 alternatives, optimized DenseNet) reduce computation overhead (carbon footprint).

5.5 Summary

Chapter 5 was devoted to a detailed discussion of the engineering requirements, design issues and implications for the artifact-robust fracture detector system proposed. The discussion began by focusing on adherence to software and hardware standards, including reliance on available platforms such as Google Colab Pro and deep learning frameworks like TensorFlow and PyTorch. These tools helped in achieving scalability and ease of experimentation; however, problems related to stable connectivity and length of session were identified. We demonstrated that DenseNet201 and UNet++ models are costly but highly efficient for training and are specific to the infrastructure in which they are trained, particularly when using GPUs.

The chapter then went on to describe the societal and environmental impacts of implementing these types of AI solutions. Thus, the application has the potential to improve diagnosis accuracy, reduce misclassification and potentially extend the availability of advanced medical imaging support in underserved areas. From the point of view of the environment, it can be used to minimise redundant scans and patient radiation (and thereby energy consumption) at the imaging device. However, considerations such as bias of the data used (non-representative dataset), ethical handling of highly personal medical data, and carbon footprint of computationally intensive model training came up as important caveats.

Ethical considerations were further developed from a perspective based on notions of fairness, transparency and accountability. Explainable AI techniques like Grad-CAM++, were identified as a safety net, to allow clinical decision-makers to audit the predictions and ensure they are ultimately the ones making the final decisions. At the same time, the sustainability plan expressed a need for lightweight, modular and flexible pipelines that are designed to ensure long-term technical reliability while minimizing environmental

and societal impacts.

Third, the chapter also included a lengthy project management and economic analysis. Two financial scenarios were examined, a low-cost baseline scenario with the cloud-based computation and an alternative investment in local high-performance hardware for long-term scalability. Rationale for each was based on short term feasibility, long term sustainability and viability. As mentioned, the business plan allowed for subscription-based, per-use and licensing models to enable the sustainability of the system beyond academic consumption.

Finally, information about knowledge domain, style of problem solving and stakeholder needs was used to map the complexity of the engineering problem. This analysis produced the result that, even though highly technical in design, the project also involved interdisciplinary co-ordination and careful consideration of social impacts..

Chapter 6

Conclusion

6.1 Summary

The project aimed to solve one of the largest issues of the contemporary medical image analysis, i.e., the reliable and interpretable detection of bone fractures in the presence of metallic artifacts. There are also numerous radiographic images that contain metallic items like plates, screws and rods that could be confusing to deep learning models. Whereas the pipelines are very robust on clean data, they break down when used on real-life data where such artifacts are unavoidable. The final objective of this study was thus to come up with a system that could be strong in such conditions and at the same time be described in a manner that could help bring confidence to the mind of the clinician.

The methodology adopted entailed the use of a structured pipeline where four basic steps including: artifact removal (preprocessing), classification/prediction, confidence score, and integrate XAI were adopted. As it happened the preprocessing step turned out to be the foundation of the framework. In order to reduce the damage caused by metallic artifacts and maximize the bone visibility without losing the most important fracture-implicating features, partial transfer learning of UNet++ was implemented. Our reason behind doing so is that, in addition to harming classification performance, artefacts mislead interpretability methods by causing them to attempt to target irrelevant locations. UNet++ was able to clean the radiographs obtained using multi-scale feature extraction and nested skip connections, producing more superior inputs when it comes to the next classification step.

The resulting data set was selected to train and evaluate a set that was representative of 3 classes: normal radiographs, fracture cases without artefacts, and fracture cases with artefacts. Roboflow labeled the images and saved them in a standard object container (COCO) JSON file as input to a deep learning model. Relative to the performance of the artifact removal process prior to pre-processing, the artifact removal efficiency significantly minimized false positive and false negative results thus making sure that the pipeline of the classification process is still founded on fracture morphology rather than analyzing certain spurious and meaningless hardware structures. This gained predictive accuracy not only enhanced the reliability of predictions, but, also, preconditioned AI-based diagnostics that can be trusted by clinicians.

The dense model denseNet201 is applied on classification stage after pre-

processing step. Justification of DenseNet201: Dense connectivity is employed in features sharing and overcoming the phenomenon of vanishing gradient. It is thus most appropriate where medical imaging is concerned where fine lines of the crack are of interest. Base models used were three state of the art ResNet101, VGG19 and InceptionV3 models. By all evaluation metrics, including accuracy, f1 score, sensitivity, specificity, DenseNet201 has proven to be the better technique always. Notably, the predictions were not presented in binary predictions only. They were instead given calibrated confidence scores by which clinicians can read model certainty and employ thresholding strategies which are applicable to a wide range of diagnostic applications including screening or confirmatory use.

In order to deal with interpretability we added Grad-CAM++, which is a more advanced explainability technique that produces high resolution saliency maps. Artifacts are the most noticeable items in the attention maps in traditional pipelines that misinterpret and fail to inspire trust. In this project, the UNet++ preprocessing and Grad-CAM++ analysis were fused in such a way that the concentration was placed on the regions of interest at all times which relate to fracture. Heatmaps affirmed that the anatomical features of interest to the model like cortical breaks and trabecular discontinuities were associated with medically relevant anatomical structures and not the presence of metallic implants. This played a significant role in clinical adoption as it gave radiologists visibility on decision making and also enabled reasoning of the model to be visually authenticated.

It was implemented in Google Colab pro with Tensorflow and Pytorch as the main deep learning frameworks. Such platforms provided the reproducibility, effective use of GPU resources, etc. The pipeline was relatively stable at the optimization (validation measures very similar to test results), so it is not a highly overfitting process. The overall system performance was good since it was comprised of the optimal recognition efficiency and robustness of DenseNet201 with the best artifact suppression performance of UNet++, thus encapsulating the system as a strong diagnostic tool capable of working in an environment with numerous artifacts.

The technical project metrics are not only provided but there are other contributions of the project. This gap between experimental performance and clinical applicability that is so easily ignored in earlier studies was suitably addressed by successfully eliminating the artifact problem. The framework was identified to not only enhance the accuracy of fracture methods, but was also found to be more interpretable, thus being useful and reliable as a clinical workflow tool. Patients are receiving improved and earlier diagnosis, clinicians have an effective decision support system and the healthcare institutions, and those in resource limited settings in particular are enjoying a scalable and

deployable system.

Lastly, this project shows how a complete artifact-robust, clinically-interpretable pipeline that includes preprocessing, UNet++, classification using DenseNet201, confidence prediction using EMODEL, confidence scoring, and interpretability using Grad-CAM++ can be achieved. The system offers a robust base to support AI-assisted fracture detection by reducing the chance of misclassification because of artefacts and ensuring that the model outputs can be easily interpreted. The paper is not only a technical solution but also provides a clinically relevant framework to contribute to the better diagnostic accuracy, patient outcomes, and a safe implementation of AI in the field of medical imaging.

6.2 Limitation

Although the proposed system has the potential and excellent performance, this project is not devoid of limitations. The last and the most significant weakness is that the range of coverage is small since the current system was trained and tested with hand X-ray images only. Although this reduced focus allowed aggressive pipeline fine-tuning, binary classification (switching to DenseNet201), and interpretability (switching to Grad-CAM++) it is bound to impair model generalization. In clinical practice, fractures can be of various anatomical locations such as hip, femur, tibia, ribs, skull and spine. The existing system, which was optimized to work with hand radiographs, might not work well or sufficiently identify these other areas and that, therefore, limits its usefulness as a direct clinical tool in general fracture screening applications.

The other significant constraint is the preprocessing pipeline that is specific to the artifact. The test UNet++ model was reasonably successful in repressing metallic implants like plates and screws, which are noted in the hand radiographs. But artifacts vary greatly in form, density, position relative to the anatomy and imaging circumstances. An example is that the noise signature of spinal implants or cranial hardware will be significantly different than a wrist or hand implant. The current preprocessing design cannot generalize to data of other parts of the body without retraining or modification. That poses scalability issues that must be resolved by either newer data collection or by retraining specific to the region to allow the solution to be more applicable in clinical settings.

Although this does assist with the possibility of explaining some of the results and outcomes, this is also a disadvantage of the explainability framework. The Grad-CAM++ was capable of detecting fracture-related cortical areas in the X-rays of the hands. Anatomy, though, is not equal in various parts of the skeleton. Slight compression fractures in the vertebrae or fracture lines in the skull can necessitate more powerful interpretability algorithms like multi-view Grad-

CAM++, 3D-saliency maps or hybrid attribution. In general diagnostic situations, the clinical reliability of the system can be compromised without these adjustments either.

Moreover, the project is yet to be exposed to computational as well as infrastructural constraints. Google Colab Pro was used to do experiments and restricted the amount of hours on the GPU and sessions that would interrupt each other These constraints dictated the batch size, training time and checkpointing strategy. Although Colab Pro was a cost-effective way to test out cloud computing, relying on such environments is inappropriate to maintain the clinical use. Otherwise with large scale applications that would need to run in real-time, we would have to go to dedicated GPU infrastructure.

Lastly, though the system has achieved high accuracy and interpretability in the area that was chosen, it is not yet prepared to be used as a general diagnostic tool in the entire spectrum of radiological variations. This vision will be based on the integration of multi-region, multi-center and large-scale datasets, generalization of more general artifact-removal strategies, and the implementation of state-of-the-art XAI strategies generalized to a variety of anatomical settings.

Finally, the key limitations of the project are associated with its small area of application (hand radiographs), the manner in which it preprocesses it with references to the artefacts, how it applies 2D explainability techniques and the limitations of the computational experiment due to its reliance on cloud computing. To create the system as an AI-assisted general-purpose diagnostic aid, with all restrictions eliminated and capable of identifying fractures correctly and explainably in all parts of the skeletal system, these limitations will have to be overcome.

6.3 Future Work

Looking towards the future of work, the first suggestion would be to look beyond the hand X-ray images, to develop a more general method for fracture detection across the entire body. However, today the project addresses only hand radiographs and is thus limited in clinical scope. Further studies should involve other anatomical components such as the legs, skull, ribs and vertebrae, to make sure that the system can predict fracture based on any location.

Curation and integration of multi-region data will be the first step towards such broad application. This includes X-ray imaging and naming of a panel of body parts and ensuring the panel is well represented with different types of fractures, different locations, and statuses with artifacts. Training with such a big variety of data allows the system to generalize more and break the body-part dependency. In doing so, the model will be one step closer to a body-agnostic solution where any input radiograph can be cleaned,

analyzed, and interpreted similarly.

Another area of interest will be generalization of artifact removal. An already existing UNet++ customization works well with hand radiographs but artifacts are highly variable across anatomy. For example, implants in the legs or X-rays of the spine may be different in density and geometry than X-rays of the hands. The future goal will be to retrain the UNet++ model, on a larger artifact dataset, and perhaps with weak supervision or self-supervised techniques, in order to obtain a preprocessing module that can be robust across contexts.

Further, the explainability component itself needs to be extended and developed. Grad-CAM++ already provides localized heatmaps of hand fractures, but other tissue areas may require multi-view or 3D explainability to locate fine fracture features, especially with complex anatomy such as pelvis or skull. A multimodal interpretation of the visual output including, for example, visual highlights with text annotations, would be more clinically relevant and useful to the radiologists and surgeons.

And finally a long-term project vision is to create a general purpose fracture detector assistant. This system would just take any medical image and process it to remove any artifacts and then classify it to detect if there is a fracture, and just overlay it with an interpretable overlay to highlight the fracture region. This would significantly reduce diagnostic errors and maximize radiology processes and patient outcomes. If the same can be done for hand x-rays the project could become a scalable full body fracture detection platform that can be applied in real life emergency room and orthopedics practice.

References

[1] Y. Wu, S. Fong, and J. Yu, "Enhancing bone radiology images classification through appropriate preprocessing: a deep learning and explainable artificial intelligence approach," Quant Imaging Med Surg, vol. 15, no. 3, pp. 2529–2546, Mar. 2025, doi: 10.21037/qims-24-1745.

[2] T. M. Nguyen, T. V. Tran, and Q. T. Lu, "FEC-IGE: An Efficient Approach to Classify Fracture Based on Convolutional Neural Networks and Integrated Gradients Explanation," ijacsa, vol. 15, no. 6, 2024, doi: 10.14569/IJACSA.2024.01506142.

[3] R. Awal, S. C. Doll, M. Naznin, and T. R. Faisal, "Interpretable machine learning classifiers for the reliable prediction of fall induced hip fracture risk," Mach. Learn. Comput. Sci. Eng, vol. 1, no. 1, p. 2, June 2025, doi: 10.1007/s44379-024-00004-w.

[4] C. S. Han, S. W. Jeong, H. W. Kim, S. M. Choi, and K. M. Lee, "Interpretable Multi-Label Classification for Tibiofibula Fracture 2D CT Images with Selective Attention and Data Augmentation," Diagnostics, vol. 14, no. 23, p. 2740, Dec. 2024, doi: 10.3390/diagnostics14232740.

[5] A. Ahmed, A. S. Imran, Z. Kastrati, S. M. Daudpota, M. Ullah, and W. Noor, "Learning from the few: Fine-grained approach to pediatric wrist pathology recognition on a limited dataset," Computers in Biology and Medicine, vol. 181, p. 109044, Oct. 2024, doi: 10.1016/j.compbiomed.2024.109044.

[6] S.-L. Chung, C.-T. Cheng, C.-H. Liao, and I.-F. Chung, "Patch-based feature mapping with generative adversarial networks for auxiliary hip fracture detection," Computers in Biology and Medicine, vol. 186, p. 109627, Mar. 2025, doi: 10.1016/j.compbiomed.2024.109627.

[7] T. Paradava, R. Jain, H. Patania, and T. Patel, "Real-Time Bone Fracture Detection Using MobileNetV2 and Explainable AI for Clinical Integration," JASTT, vol. 6, no. 1, pp. 36–42, June 2025, doi: 10.38094/jastt61236.

[8] W. Wei et al., "YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images," Journal of Radiation Research and Applied Sciences, vol. 18, no. 1, p. 101309, Mar. 2025, doi: 10.1016/j.jrras.2025.101309.

[9] S. S. Jadhav, "Bone fracture detection using image processing techniques," Sigma J Eng Nat Sci - Sigma Müh Fen Bil Derg, pp. 748–759, 2025, doi: 10.14744/sigma.2025.00069.

[10] T. Aldhyani et al., "Diagnosis and detection of bone fracture in radiographic images using deep learning approaches," Front. Med., vol. 11, p. 1506686, Jan. 2025, doi: 10.3389/fmed.2024.1506686.

[11] P. N. Srinivasu, G. L. A. Kumari, S. C. Narahari, S. Ahmed, and A. Alhumam, "Exploring the impact of hyperparameter and data augmentation in YOLO V10 for accurate bone fracture detection from X-ray images," Sci Rep, vol. 15, no. 1, p. 9828, Mar. 2025, doi: 10.1038/s41598-025-93505-4.

[12] A. Abdusalomov et al., "Lightweight Deep Learning Framework for Accurate Detection of Sports-Related Bone Fractures," Diagnostics, vol. 15, no. 3, p. 271, Jan. 2025, doi: 10.3390/diagnostics15030271.

[13] A. M. Elsheikh and A. A. Elhag, "Machine learning-based analysis of multi-region bone fracture detection and classification using biomedical images," Alexandria Engineering Journal, vol. 128, pp. 186–199, Sept. 2025, doi: 10.1016/j.aej.2025.05.074.

[14] H. T. Nguyen, T. B. Tran, and T. T. Tran, "Fracture Detection in Bone: An Approach with Versions of YOLOv4," SN COMPUT. SCI., vol. 5, no. 6, p. 765, Aug. 2024, doi: 10.1007/s42979-024-03155-y.

[15] N. Ramadanov et al., "Artificial intelligence-guided distal radius fracture detection on plain radiographs in comparison with human raters," J Orthop Surg Res, vol. 20, no. 1, p. 468, May 2025, doi: 10.1186/s13018-025-05888-9.

[16] D. Guenoun et al., "Automated vertebral compression fracture detection and quantification on opportunistic CT scans: a performance evaluation," Clinical Radiology, vol. 83, p. 106831, Apr. 2025, doi: 10.1016/j.crad.2025.106831.

[17] Q. Ye et al., "Deep learning approach based on a patch residual for pediatric supracondylar subtle fracture detection," Biomol Biomed, vol. 25, no. 7, pp. 1631–1646, May 2025, doi: 10.17305/bb.2024.11341.

[18] J. Zou and M. R. Arshad, "Detection of whole body bone fractures based on improved YOLOv7," Biomedical Signal Processing and Control, vol. 91, p. 105995, May 2024, doi: 10.1016/j.bspc.2024.105995.

[19] C. Pauling et al., "External validation of an artificial intelligence tool for fracture detection in children with osteogenesis imperfecta: a multireader study," Eur Radiol, July 2025, doi: 10.1007/s00330-025-11790-z.

[20] C.-Y. Lu et al., "Artificial Intelligence Application in Skull Bone Fracture with Segmentation Approach," J Digit Imaging. Inform. med., vol. 38, no. 1, pp. 31–46, July 2024, doi: 10.1007/s10278-024-01156-0.

[21] E. Paspuel, "Deep learning-driven diagnosis of humerus fractures from radiographic data." Mendeley Data, Apr. 14, 2025. doi: 10.17632/7M4WHN2FCZ.1.

213-15-4443

ORIGINALITY REPORT

8%	5%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	arxiv.org Internet Source	<1%
4	Submitted to Queensland University of Technology Student Paper	<1%
5	jurnal.yoctobrain.org Internet Source	<1%
6	link.springer.com Internet Source	<1%
7	Submitted to United International University Student Paper	<1%
8	jastt.org Internet Source	<1%
9	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
10	Submitted to Universiti Teknologi Petronas Student Paper	<1%
11	Yaoyang Wu, Simon Fong, Jiahui Yu. "Enhancing bone radiology images classification through appropriate	<1%