MSc Data Science AUEB

Deep Learning

STANFORD MURA (X-RAY) CLASSIFICATION COMPETITION

Professor: Themos Stafylakis

Gerasimos Kazantzis | f3352406

Abstract

Fracture detection in musculoskeletal radiographs is a critical task in medical diagnostics, requiring both high precision and expert interpretation. In this study, we explore the application of deep learning for automatic binary classification of radiographic images into fractured and non-fractured cases. The primary objective is accurate fracture detection, while body part identification is incorporated as an auxiliary task in selected models to provide anatomical context and improve overall performance.

We first implement a custom Convolutional Neural Network (CNN) as a baseline and then evaluate three state-of-the-art architectures: ResNet50, DenseNet169, and EfficientNet-B0. All models are adapted for grayscale input and trained in a multi-task learning framework, simultaneously predicting fracture presence and anatomical region.

Experiments are conducted on the MURA dataset, one of the largest publicly available collections of musculoskeletal radiographs. Results show that multi-task learning consistently enhances fracture classification, with EfficientNet-B0 achieving the best performance in terms of accuracy, recall, and robustness. The inclusion of anatomical context is particularly effective in reducing errors for body parts with similar visual characteristics. Overall, this work demonstrates the value of modern deep learning approaches and auxiliary supervision in improving medical image analysis.

Table of Contents

1. Introduction....3

2. Problem Description and Dataset Analysis....3

3. Methodology and Model Architectures....3

3.1 Methodological Overview....4

3.2 Custom CNN....4

3.3 ResNet50 with Multi-Task Learning....4

3.4 DenseNet169 with Multi-Task Learning....5

3.5 EfficientNet (B0) with Multi-Task Learning....6

4. Experimental Results and Evaluation....7

4.1 Evaluation Metrics....7

4.2 Custom CNN Results....7

4.3 ResNet50 Results....7

4.4 DenseNet169 Results....8

4.5 EfficientNet Results....8

4.6 Comparative Analysis....8

4.7 Generalization to External Dataset....9

5. Conclusion and Future Work....9

5.1 Conclusion....9

5.2 Future Work....10

6. References....11

1. Introduction

Musculoskeletal radiographs are among the most commonly used diagnostic tools in clinical settings for detecting bone fractures and related abnormalities. Despite their widespread use, interpreting these X-ray images remains a complex and error-prone task, often requiring significant medical expertise and time. In recent years, deep learning has shown considerable promise in automating medical image analysis and assisting clinicians in achieving faster and more accurate diagnoses.

In this project, we investigate the use of deep learning models to automate the classification of radiographic images from the MURA (Musculoskeletal Radiographs) dataset. The initial approach focused solely on binary fracture classification using a custom-built convolutional neural network (CNN). While this model provided a reasonable baseline, its performance in terms of accuracy, recall, and overall generalization was limited.

To address these limitations, we extended the problem to a multi-task learning framework. In addition to detecting fractures, the updated models also predict the anatomical body part (e.g., elbow, wrist, etc.) shown in the image. This auxiliary task acts as an inductive bias that encourages the model to learn more discriminative and generalizable features, ultimately improving performance on the primary task of fracture detection.

We implemented and compared four different architectures: a custom CNN, ResNet50, EfficientNet, and DenseNet169. All models were modified to accept grayscale input and adapted to perform both fracture classification and body part identification. Among the evaluated models, the EfficientNet-based architecture yields the best results in terms of validation accuracy and loss.

2. Problem Description and Dataset Analysis

The main objective of this project is to build a deep learning model that can automatically detect fractures in musculoskeletal radiographs. Since accurate fracture classification is a critical task in clinical diagnosis, the goal is to develop a model that achieves high accuracy, recall, and generalization on unseen data.

Initially, the task was framed as a binary classification problem: given a radiograph, the model must predict whether a fracture is present or not. However, after experimenting with a custom CNN model, it became clear that performance could be further improved. To this end, the problem was extended to a multi-task learning setup, where the model not only predicts the presence of a fracture but also classifies the body part shown in the image. This auxiliary task helps provide anatomical context, which encourages the model to learn more meaningful visual features for fracture detection.

The dataset used in this project is the MURA (Musculoskeletal Radiographs) dataset, which consists of over 40,000 X-ray images from seven different body parts: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each image is labeled as either “positive” (fracture) or “negative” (no fracture). The dataset is imbalanced, with some body parts appearing more frequently than others and with varying proportions of fracture cases across body parts — as illustrated in Figure 1.

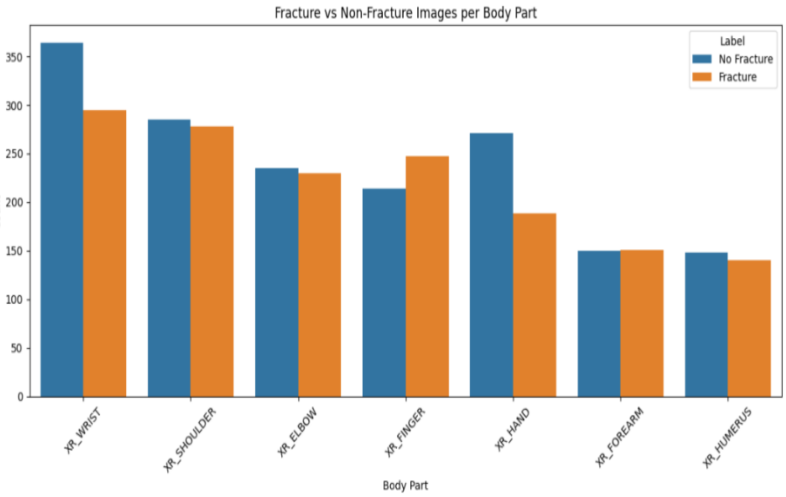

Figure 1: *Distribution of fracture and non-fracture cases across body parts in the MURA dataset.*

To structure the data for multi-task learning, each image is paired with:

- A binary label for **fracture classification**.
- A categorical label indicating the **body part** (mapped to an integer index).

All images were converted to grayscale and resized to a fixed resolution to match the input requirements of the deep learning models. The dataset was divided into three subsets: 29,446 images for training, 7,362 for validation, and 3,197 for testing. The training set was used to fit the models, the validation set guided early stopping and hyperparameter tuning, while the test set was held out for final evaluation and generalization assessment.

3. Methodology and Model Architectures

This section describes the overall methodology followed throughout the project and presents the different model architectures used for fracture classification on the MURA dataset. The evolution of the modeling approach reflects a progression from a simple custom CNN to more sophisticated multi-task models based on pretrained convolutional neural networks.

3.1 Methodological Overview

The project began with a baseline **binary classification** setup using a custom-built CNN trained from scratch to predict the presence of fractures in grayscale radiographs. While this model offered a reasonable starting point, its limited capacity to generalize prompted a methodological shift.

To enhance performance, the task was reformulated as a **multi-task learning** problem. Instead of predicting only fractures, the model also learned to classify the **body part** shown in the radiograph. This auxiliary task encourages the model to learn more anatomically relevant features, improving performance on the primary fracture detection task.

All models were trained using grayscale images resized to 224×224 pixels. Data augmentations such as random horizontal flips and rotations were applied to the training set. The loss function used was a combination of **cross-entropy losses** for both tasks, and optimization was performed using the **Adam** optimizer with appropriate learning rate scheduling and early stopping based on validation loss.

3.2 Custom CNN

The first model implemented was a custom convolutional neural network (CNN), built from scratch without the use of pretrained weights. It consists of three convolutional blocks, each followed by a ReLU activation function and max pooling layer. The final convolutional feature maps are flattened and concatenated with an embedding vector representing the body part index. This embedding allows the model to leverage additional anatomical context for fracture classification. The combined features are passed through two fully connected layers with dropout regularization before producing the final binary output for fracture prediction. While this architecture is relatively simple compared to the others, it serves as a strong baseline and demonstrates the benefit of incorporating body part information even in basic models.

See Figure 2 for the Custom CNN architecture.

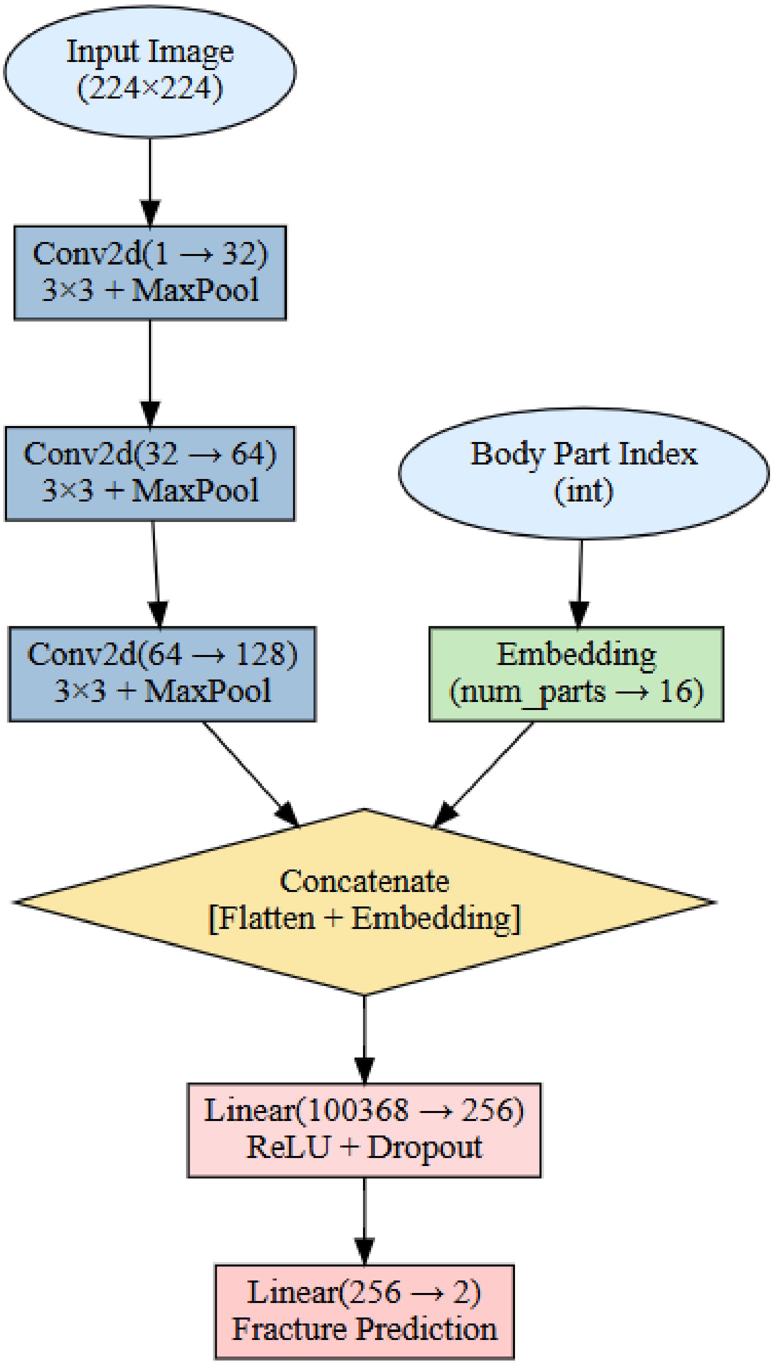
Input Image (224×224)
Conv2d(1 → 32) 3×3 + MaxPool
Conv2d(32 → 64) 3×3 + MaxPool
Conv2d(64 → 128) 3×3 + MaxPool
Body Part Index (int)
Embedding (num_parts → 16)
Concatenate [Flatten + Embedding]
Linear(100368 → 256) ReLU + Dropout
Linear(256 → 2) Fracture Prediction

Figure 2: *Custom CNN Architecture*

3.3 ResNet50 with Multi-Task Learning

The second model implemented was based on the ResNet50 architecture, a widely used deep convolutional neural network known for its deep residual learning capabilities. To adapt ResNet50 for the MURA challenge, we modified the input layer to accept single-channel grayscale images and used pretrained weights from ImageNet to leverage transfer learning. All layers except the final block (layer4) were frozen to retain general visual features, while allowing fine-tuning on domain-specific radiographic patterns.

Following the convolutional backbone, a global average pooling operation was applied to extract a compact 2048-dimensional feature vector. For the body part classification task, this feature vector was passed through a fully connected layer with ReLU activation and dropout, followed by a linear output layer predicting one of the seven anatomical body parts.

For the fracture classification task, the model used a multi-task learning setup. An embedding layer was introduced to encode the body part index into a 32-dimensional vector. This embedding was concatenated with the pooled ResNet features (2048 + 32 = 2080), and the combined vector was passed through a sequence of fully connected layers with batch normalization, ReLU activations, and dropout regularization. The final layer outputted a binary prediction indicating the presence or absence of a fracture.

This architecture benefited from both the pretrained ResNet backbone and the auxiliary body part prediction task, which provided anatomical context and improved the model's ability to distinguish between fractured and non-fractured regions. The use of multi-task learning proved especially effective in improving recall and generalization compared to the baseline CNN model.

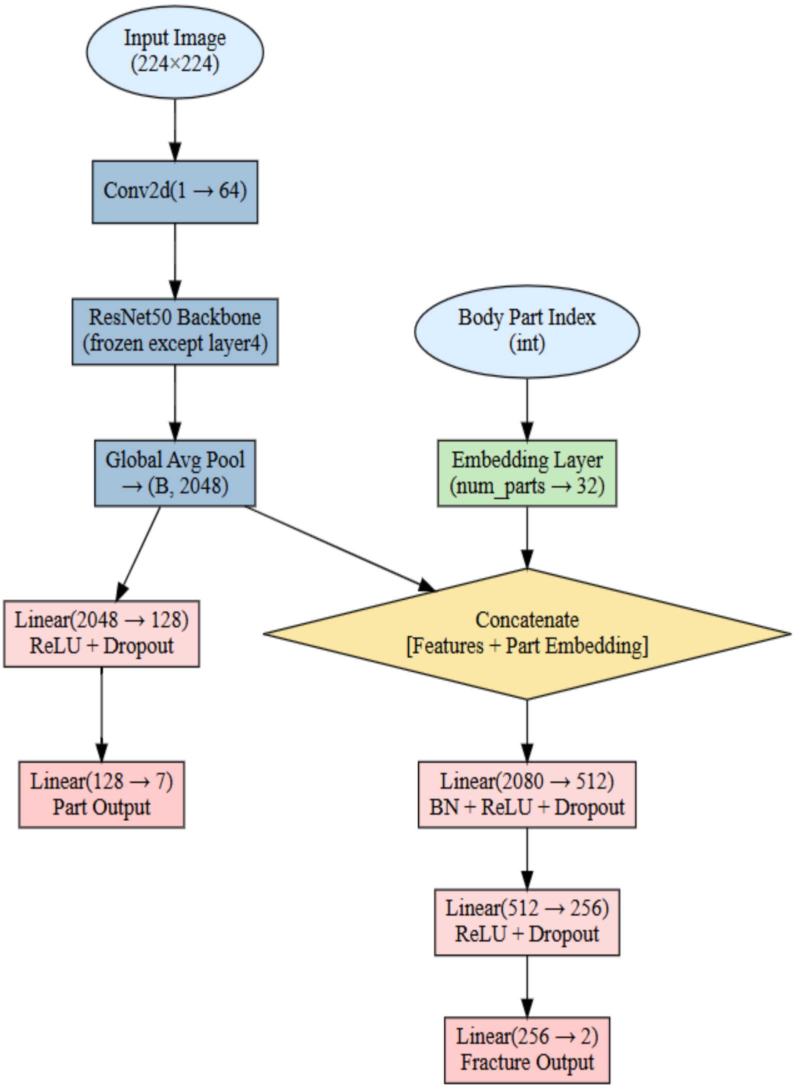

Figure 3: *ResNet50 Multi-Task model Architecture*

3.4 DenseNet169 with Multi-Task Learning

The third model explored was based on the **DenseNet169** architecture, a densely connected convolutional network known for promoting feature reuse and alleviating the vanishing gradient problem. Similar to the ResNet-based model, the DenseNet169 was used as a feature extractor, with its classification layers removed. The model received grayscale images and produced a 1664-dimensional feature vector using a global adaptive average pooling layer.

To incorporate contextual anatomical information, an **embedding layer** was used to encode the body part index into a 32-dimensional vector. This embedding was concatenated with the extracted image features, forming a 1696-dimensional joint representation. This representation was then passed through a fracture classification head consisting of three fully connected layers with batch normalization, ReLU activations, and dropout regularization.

In parallel, a secondary head was added to predict the body part from the raw image features. This consisted of a linear layer that transformed the 1664-dimensional feature vector to a 128-dimensional hidden representation, followed by a final classification layer. Both tasks were trained jointly in a **multi-task learning** setup to encourage the model to learn anatomically relevant features.

This architecture achieved **stronger generalization and recall** than previous models, suggesting that the increased capacity of DenseNet169 and the incorporation of body part context contributed significantly to its performance on the fracture detection task.

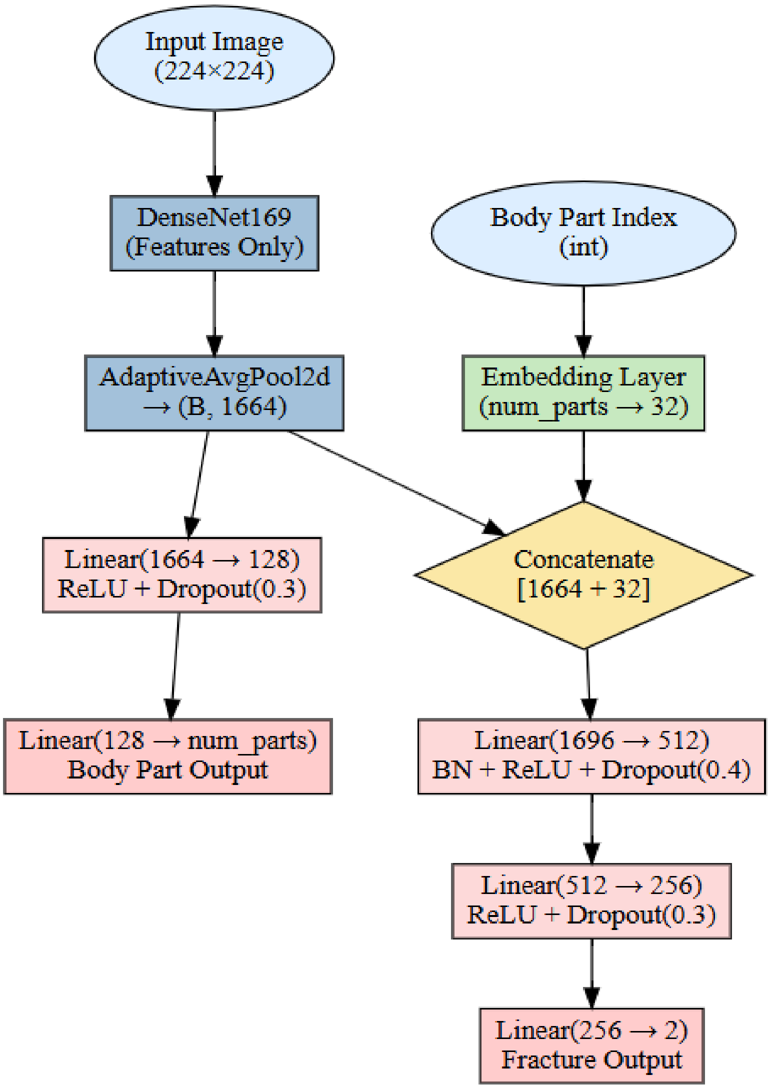

Figure 4: *DenseNet169 Multi-Task model Architecture*

3.5 EfficientNet (B0) with Multi-Task Learning

The final and best-performing model was built upon EfficientNet-B0, a convolutional neural network known for its balance between model size and performance through compound scaling of depth, width, and resolution. EfficientNet-B0 was used as a frozen feature extractor, outputting a 1280-dimensional feature vector via an adaptive average pooling layer after processing the 224×224 grayscale input images.

To enrich the model with anatomical context, a 32-dimensional body part embedding was generated from the body part index and concatenated with the extracted image features, resulting in a joint 1312-dimensional representation. This combined vector was passed through a three-layer fracture classification head consisting of fully connected layers with ReLU activations and dropout, culminating in a final binary output for fracture prediction.

In parallel, the model included a body part classification head, where the image features were processed through two fully connected layers to produce predictions for the anatomical category. This auxiliary task provided additional supervision that helped guide the model to learn more spatially aware and discriminative features.

Among all the tested architectures, the EfficientNet-based model achieved the highest scores in accuracy, F1, and ROC-AUC, highlighting the strength of compound scaling and the benefits of multi-task learning for medical image analysis.

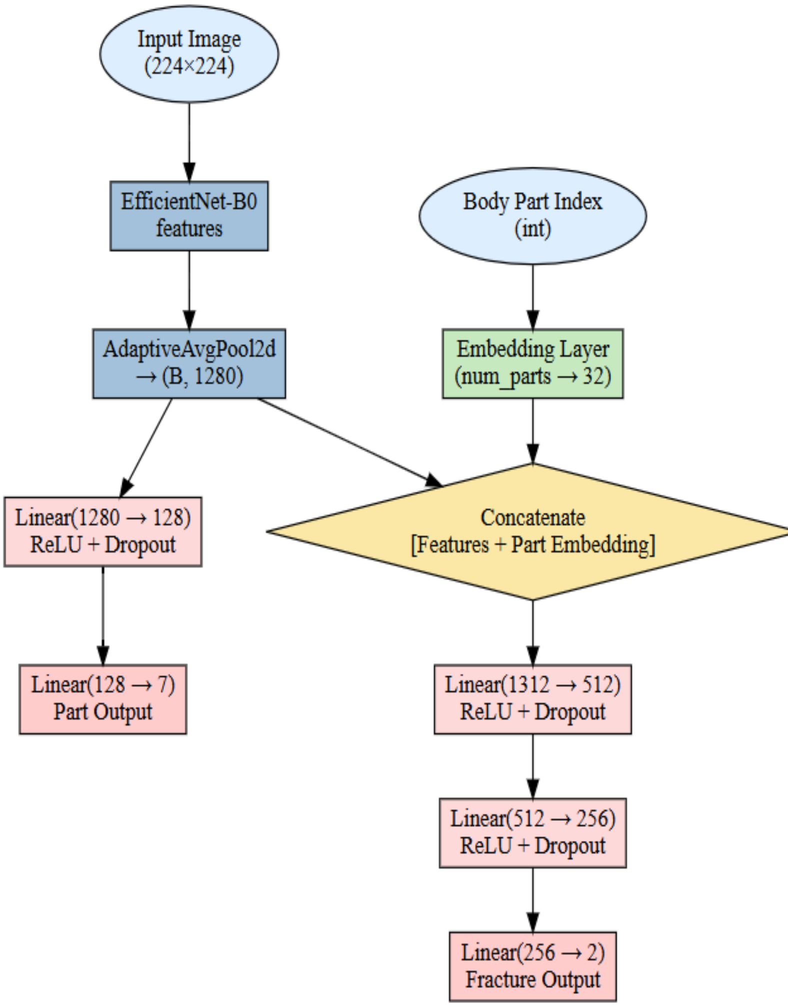

Figure 5: *EfficientNet-B0 Multi-Task model Architecture*

4. Experimental Results and Evaluation

4.1 Evaluation Metrics

To thoroughly evaluate the performance of the different models on the fracture classification task, we employed a variety of well-established classification metrics:

- **Accuracy**: The proportion of correctly classified samples out of the total samples. While useful, it can be misleading on imbalanced datasets.
- **Precision**: Measures the proportion of predicted positives that are actually positive. In this context, it tells us how often a predicted fracture is indeed a fracture.
- **Recall (Sensitivity)**: Indicates the proportion of actual positives that are correctly identified. High recall is critical in medical applications to minimize missed fracture cases.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced view when class distributions are skewed.
- **ROC AUC (Receiver Operating Characteristic - Area Under Curve)**: Evaluates the trade-off between true positive and false positive rates, summarizing classifier performance across thresholds.
- **Cohen's Kappa Score**: Measures agreement between predicted and true labels, adjusted for the agreement that could occur by chance. This is especially informative for imbalanced classification problems.

For the **body part classification** (auxiliary task), we report:

- **Accuracy**
- **Per-class precision, recall, and F1-score**, to ensure that performance is consistent across all seven body part classes.

4.2 Custom CNN Results

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-score	Support
0	0.6511	0.7612	0.7019	1667
1	0.6811	0.5556	0.6120	1530
Accuracy			0.6628	3197
Macro Avg	0.6661	0.6584	0.6569	3197
Weighted Avg	0.6655	0.6628	0.6588	3197

Table 1: Custom CNN Classification Report

Despite being the simplest architecture in this study, the custom CNN achieved a test accuracy of **66.28%**, serving as a strong baseline. It performed better on the **non-fractured class (class 0)**, achieving a recall of **76.12%**, but struggled more with detecting fractures (**class 1**, recall **55.56%**), which is typical for imbalanced medical datasets. The relatively low F1-score for the fracture class (0.6120) highlights the model's limited sensitivity, suggesting it may miss many actual fracture cases. Nonetheless, the **macro-averaged F1-score of 65.69%** indicates balanced learning across classes despite the model's simplicity.

4.3 ResNet50 Results

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-score	Support
0	0.6967	0.9244	0.7945	1667
1	0.8721	0.5614	0.6831	1530
Accuracy			0.7507	3197
Macro Avg	0.7844	0.7429	0.7388	3197
Weighted Avg	0.7806	0.7507	0.7412	3197

Table 2: ResNet50 Fracture Classification Report

Roc AUC Score: 0.8237

Cohen's Kappa Score: 0.4930

Class	Precision	Recall	F1-score	Support
Elbow	0.9131	0.9720	0.9417	465
Finger	0.9411	0.9696	0.9551	461
Forearm	0.9101	0.8405	0.8739	301
Hand	0.9649	0.9565	0.9607	460
Hummerus	0.9810	0.8958	0.9365	288
Shoulder	0.9807	0.9911	0.9859	563
Wrist	0.9667	0.9697	0.9682	659
Accuracy			0.9531	3197
Macro Avg	0.9511	0.9422	0.9560	3197
Weighted Avg	0.9534	0.9531	0.9527	3197

Table 3: ResNet50 Body Part Classification Report

The ResNet50-based model showed a clear improvement over the custom CNN, achieving an overall fracture classification accuracy of 75.07%. The model demonstrated a high recall of 92.44% for the non-fracture class, but its performance dropped for the fracture class, with a recall of 56.14%. This indicates the model is strong at identifying healthy cases but still struggles with correctly detecting all fractures. Nevertheless, its ROC AUC score of 0.8237 and Cohen's Kappa of 0.493 suggest more robust discrimination and reliability compared to the baseline.

The model's auxiliary task — predicting the body part — performed remarkably well, achieving an accuracy of 95.31%. All body part classes exceeded 87% F1-score, with the shoulder and wrist showing near-perfect results. This strong performance in anatomical classification reinforces the benefit of multi-task learning by providing valuable context for fracture prediction.

4.4 DenseNet169 Results

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-score	Support
0	0.7515	0.8980	0.8183	1667
1	0.8589	0.6765	0.7569	1530
Accuracy			0.7920	3197
Macro Avg	0.8052	0.7872	0.7876	3197
Weighted Avg	0.8029	0.7920	0.7889	3197

Table 4: DenseNet169 Fracture Classification Report

ROC AUC Score: 0.8540

Cohen's Kappa Score: 0.5795

Class	Precision	Recall	F1-score	Support
Elbow	0.9197	0.9849	0.9512	465
Finger	0.9781	0.9675	0.9727	461
Forearm	0.9446	0.8505	0.8951	301
Hand	0.9636	0.9739	0.9686	460
Hummerus	0.9849	0.9062	0.9439	288
Shoulder	0.9824	0.9911	0.9867	563
Wrist	0.9599	0.9818	0.9707	659
Accuracy			0.9615	3197
Macro Avg	0.9619	0.9509	0.9556	3197
Weighted Avg	0.9620	0.9615	0.9612	3197

Table 5: DenseNet169 Body Part Classification Report

The DenseNet169 model delivered a solid performance in both fracture and body part classification tasks. For fracture detection, it achieved an overall accuracy of 79.20%, with a macro-averaged F1-score of 78.76% and a ROC AUC score of 0.8540, demonstrating balanced capability in distinguishing between fractured and non-fractured cases. Notably, its recall for the non-fracture class (89.80%) remained high, while the fracture class (67.65%) showed improvement compared to the ResNet50.

On the body part classification task, the model excelled with an accuracy of 96.15%, surpassing previous architectures. Most body parts achieved F1-scores above 96%, and the macro-averaged F1-score reached 95.56%, highlighting DenseNet169's capacity to extract rich, discriminative features from musculoskeletal images. This confirms its effectiveness in leveraging anatomical context to assist fracture classification in a multi-task setting.

4.5 EfficientNet Results

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-score	Support
0	0.7665	0.9040	0.8296	1667
1	0.8700	0.7000	0.7758	1530
Accuracy			0.8064	3197
Macro Avg	0.8183	0.8020	0.8027	3197
Weighted Avg	0.8161	0.8064	0.8039	3197

Table 6: EfficientNet Fracture Classification Report

ROC AUC Score: 0.8709

Cohen's Kappa Score: 0.6089

Class	Precision	Recall	F1-score	Support
Elbow	0.9502	0.9849	0.9673	465
Finger	0.9867	0.9631	0.9748	461
Forearm	0.9664	0.8605	0.9104	301
Hand	0.9555	0.9804	0.9678	460
Hummerus	0.9716	0.9514	0.9614	288
Shoulder	0.9808	1.000	0.9903	563
Wrist	0.9656	0.9803	0.9729	659
Accuracy			0.9681	3197
Macro Avg	0.9681	0.9601	0.9635	3197
Weighted Avg	0.9683	0.9681	0.9678	3197

Table 7: EfficientNet Body Part Classification Report

4.6 Comparative Analysis

To evaluate and compare the performance of all models, we summarize their key classification metrics in the tables below. These include **accuracy**, **macro-averaged F1-score**, **ROC AUC score**, and **Cohen's Kappa** for the fracture classification task. Additionally, we report **accuracy and macro F1-score** for the auxiliary body part classification task.

Model	Accuracy	Macro F1	ROC AUC	Cohen's Kappa
Custom CNN	66.28%	65.69%	0.7272	-
ResNet50	75.07%	73.88%	0.8237	0.4931
DenseNet169	79.20%	78.76%	0.8540	0.5795
EfficientNet-B0	80.64%	80.27%	0.8710	0.6089

Table 8: Fracture Classification – Model Comparison

Model	Accuracy	Macro F1
ResNet50	95.31%	94.60%
DenseNet169	96.15%	95.56%
EfficientNet-B0	96.81%	96.35%

Table 9: Body Part Classification – Model Comparison

From the comparative results, it's evident that **EfficientNet-B0** outperforms all other models in both fracture and body part classification tasks. It achieved the highest **accuracy (80.64%)**, **macro F1-score (80.27%)**, and **ROC AUC (0.8710)** on the fracture task, along with the best **Cohen's Kappa score (0.6089)** — indicating strong inter-annotator agreement with the ground truth labels.

Among backbone networks, **DenseNet169** consistently outperformed **ResNet50**, demonstrating better generalization and feature extraction, especially on the minority (fracture) class. While the **Custom CNN** provided a useful baseline, it lagged significantly behind the pretrained models.

In the auxiliary body part classification task, all pretrained models showed excellent performance, but **EfficientNet-B0** again led with **96.81% accuracy** and **96.35% macro F1-score**.

These results validate the importance of transfer learning using powerful pretrained backbones and confirm that multi-task learning, when combined with deep feature extractors, significantly improves both diagnostic accuracy and anatomical understanding in musculoskeletal radiograph classification.

4.7 Generalization to External Dataset

To evaluate the generalization capabilities of the best-performing model, **EfficientNet-B0**, we tested it on an **external X-ray dataset** derived from a different clinical setting and dataset challenge. This dataset included **829 radiographs**, annotated for the presence or absence of fractures, and was not part of the MURA dataset used during training.

Despite the domain shift, the model achieved:

Class	Precision	Recall	F1-score	Support
0	0.72	0.98	0.83	492
1	0.94	0.44	0.60	337
Accuracy			0.76	829
Macro Avg	0.83	0.71	0.72	829
Weighted Avg	0.81	0.76	0.74	829

Table 10: EfficientNet-B0 Generalization Performance on External X-ray Dataset

$$\begin{bmatrix} 482 & 10 \\ 188 & 149 \end{bmatrix}$$

Table 11: Confusion Matrix – EfficientNet-B0 Evaluation on External Dataset

The **confusion matrix** highlights that while the model excels at identifying non-fractured cases, it struggles with correctly detecting fractures, missing 188 of the 337 fractured samples. This suggests that the model may have **overfit to the original MURA distribution** and struggles with **data heterogeneity** in real-world clinical data.

These results underscore the importance of **domain adaptation** and **cross-institutional training data** to improve robustness and reduce bias. They also affirm that while the model exhibits strong potential, further fine-tuning or transfer learning may be required before clinical deployment.

5. Conclusion and Future Work

5.1 Conclusion

In this project, we developed and evaluated deep learning models for automated fracture detection in musculoskeletal radiographs using the MURA dataset. Initially formulated as a binary classification task, the problem was later extended to a multi-task learning setting that also included body part classification. This reformulation proved beneficial, as providing anatomical context led to improved performance in the primary task of fracture detection.

We experimented with four different model architectures:

- A Custom CNN built from scratch, serving as a lightweight baseline.
- A ResNet50-based model, leveraging pretrained weights for stronger feature extraction.
- A DenseNet169 model, which improved performance further through dense connections and feature reuse.
- An EfficientNet-B0 model, which ultimately achieved the best overall performance in both tasks.

Across all models, the inclusion of body part information via an embedding layer consistently enhanced fracture classification. EfficientNet-B0 achieved the highest accuracy (80.6%), ROC AUC score (0.871), and Cohen's Kappa (0.609), demonstrating the strength of compound scaling in vision tasks. Classification reports also revealed that class 1 (fracture) was consistently harder to detect than class 0 (non-fracture), with lower recall values across all models — highlighting the ongoing challenge of reducing false negatives in medical imaging.

To evaluate model robustness, we tested the EfficientNet-B0 architecture on an external X-ray dataset from a different challenge. Despite the domain shift, the model maintained an accuracy of 76% and high precision, especially for the non-fractured class. This experiment suggests that EfficientNet-B0 exhibits strong generalization capabilities, though performance for detecting fractures declined due to reduced recall — a valuable insight for future deployment considerations.

5.2 Future Work

Although the models achieved promising results, several directions could further improve performance and robustness:

- **Incorporate clinical metadata** such as age, sex, or patient history, to provide additional context to the model.
- **Use attention mechanisms** (e.g., self-attention or transformers) to help the model focus on fracture-prone regions.
- **Visual explanation tools** like Grad-CAM could be integrated for model interpretability and better error analysis.
- **Train on external datasets** or test on real hospital images to evaluate generalization beyond MURA.
- **Address class imbalance** using focal loss, oversampling, or data synthesis techniques like GANs.
- **Explore ensemble methods**, combining multiple architectures to benefit from diverse feature representations.

This work lays a solid foundation for future improvements and highlights the potential of deep learning in assisting radiologists with fracture diagnosis in a reliable and scalable manner.

6. References

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. In *Advances in Neural Information Processing Systems* (NeurIPS), 25.

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely Connected Convolutional Networks*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708.

- Tan, M., & Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6105–6114.

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. In *Advances in Neural Information Processing Systems* (NeurIPS), 32.

- Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., ... & Ng, A. Y. (2017). *MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs*. arXiv preprint arXiv:1712.06957.

- B. Madushani Rodrigo. (2023). *Fracture Multi-Region X-ray Data* [Dataset]. Kaggle. https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data

- Ruder, S. (2017). *An Overview of Multi-Task Learning in Deep Neural Networks*. arXiv preprint arXiv:1706.05098.

- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). *ImageNet: A Large-Scale Hierarchical Image Database*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255.