

BONE FRACTURE DETECTION USING DEEP LEARNING

Riddhi Hindocha, Harneet Kaur Chahal, Gauri Jadhav, Khushi Desai

University of Calgary

ABSTRACT

This project presents a machine learning approach for bone fracture detection using medical imaging. The study leverages convolutional neural networks (CNNs) along with advanced deep learning models such as ResNet18, EfficientNetB3, and MobileNetV2 to classify X-ray images into fractured and non-fractured categories. The methodology involves extensive data preprocessing, CNN and transfer learning-based model training, and evaluation using multiple performance metrics. MobileNetV2 was included in the experimentation to explore the effectiveness of lightweight transformer-based architectures in comparison to traditional CNN-based models. The results demonstrate high accuracy and reliability across all models, with comparative analysis highlighting the trade-offs between model complexity, inference speed, and classification performance. This showcases the potential for automated bone fracture detection to assist radiologists in the diagnostic process.

Index Terms— Bone fracture detection, machine learning, deep learning, medical imaging, convolutional neural networks, ResNet18, EfficientNetB3, MobileNetV2, ReLU, Softmax

1. INTRODUCTION

Bone fractures are a prevalent medical condition that requires timely and accurate diagnosis to prevent complications. The increasing demand for efficient healthcare solutions has driven the adoption of artificial intelligence (AI) in medical diagnostics. Traditionally, radiologists analyze X-ray images to detect fractures, a process that is both labor-intensive and susceptible to human error. Studies indicate that misdiagnosis rates in radiology can be as high as 10%, underscoring the necessity of automated solutions. The growing availability of medical imaging data and advancements in machine learning have paved the way for AI-based systems capable of enhancing diagnostic precision.

This study explores the application of deep learning, specifically convolutional neural networks (CNNs) and advanced architectures such as ResNet18 and EfficientNetB3, MobileNetV2 to detect fractures in X-ray images. The primary objectives include:

- Developing and optimizing multiple deep learning models—CNN, ResNet18, EfficientNetB3, and MobileNetV2—for accurate X-ray fracture detection.
- Enhancing model performance through advanced preprocessing techniques tailored to improve X-ray image quality and clarity.
- Incorporating Grad-CAM for model explainability, enabling visualization of regions contributing to fracture predictions.
- Evaluating and comparing model performance using multiple metrics, including accuracy, precision, recall, F1-score, ROC curves, confusion matrices, and analysis of misclassified samples.

The significance of this research lies in its potential to reduce workload on radiologists, improve early diagnosis, and ultimately enhance patient care through rapid and accurate detection of fractures.

2. RELATED WORK

Advancements in machine learning (ML) and deep learning (DL) have significantly enhanced fracture detection in X-ray images. Convolutional Neural Networks (CNNs) have demonstrated high accuracy in classifying fractures. For instance, studies have shown that CNNs can effectively identify fractures in radiographs, with performance metrics such as sensitivity and specificity reaching notable levels.

Residual Networks (ResNet-18) further improve performance by enabling deeper architectures. Research indicates that ResNet-based models can achieve high accuracy in detecting fractures, surpassing traditional methods.

EfficientNet offers a more efficient model with fewer parameters and higher accuracy, making it suitable for real-

time applications. Studies have demonstrated that EfficientNet-based models can effectively detect fractures in X-ray images, achieving high accuracy and area under the receiver operating characteristic curve (AUC) values. These developments underscore the potential of deep learning models in enhancing fracture detection and diagnosis in medical imaging.

3. MATERIALS AND METHODS

3.1. Dataset

The dataset used for this study was obtained from Kaggle: [Fracture Detection Using X-ray Images](#). It consists of labeled X-ray images sourced from open-access medical databases, categorized into two classes: *fractured* and *non-fractured*. The complete project implementation is available on GitHub: [ENEL 645 Final Project Repository](#).

3.2. Data Preprocessing

To standardize the input and enhance model generalization, several preprocessing techniques were applied:

- **Image Resizing:** Performed during data preprocessing to match the input size requirements of each model architecture.
- **Normalization:** Scaling pixel values to a range of [0,1] for uniformity.
- **Data Augmentation:** Applying random rotations, flips, and brightness variations to artificially expand the dataset and improve generalization.

Noise Reduction: Implementing Gaussian blur and contrast enhancement to improve image clarity.

3.3. Model Architecture

This study leverages various deep learning architectures to classify X-ray images into fractured and non-fractured categories, including a custom CNN model, ResNet18, EfficientNetB3, and MobileNetV2. Below is a detailed description of the model architecture, training setup, and evaluation metrics used in this study.

Convolutional Neural Network (CNN): The CNN model incorporates convolutional layers with 3×3 filters to extract spatial features, followed by batch normalization for stabilizing and speeding up training. Max-pooling layers reduce spatial dimensions, easing the computational load, while dropout layers help mitigate overfitting. Fully connected layers are used to map the extracted features to output classes. The intermediate layers use ReLU activation, while Softmax is applied at the final layer for binary classification.

ResNet18: ResNet18 is a residual network with 18 layers designed to mitigate the vanishing gradient problem through

residual connections. These connections facilitate the network's ability to learn deeper representations without performance degradation. ResNet18 is pre-trained on ImageNet and fine-tuned for fracture detection, enabling efficient feature propagation, and maintaining accuracy in deep networks.

EfficientNetB3: EfficientNetB3 is an advanced model that optimally balances network depth, width, and input resolution using a compound scaling approach. This results in high accuracy with fewer parameters, improving computational efficiency. EfficientNetB3 is well-suited for medical imaging, where it efficiently extracts feature representations while maintaining a low computational footprint. Transfer learning and fine-tuning were used to improve its performance for fracture detection.

MobileNetV2: MobileNetV2 is a lightweight, efficient architecture designed for mobile and resource-constrained environments. It integrates depth wise separable convolutions and residual connections to reduce computational complexity while maintaining performance. The model was included in this study to compare compact architectures like MobileNetV2 with deeper networks, particularly in real-time inference scenarios where computational resources are limited.

3.4. Training Setup

Optimizer: Adam optimizer was used for model training to adjust the weights and minimize the loss function efficiently.

Loss Function: Categorical Cross-Entropy was chosen as the loss function to measure the error in the classification task.

Learning Rate: The learning rate was set to 0.001 to ensure smooth convergence.

Batch Size: A batch size of 32 was used during training to balance computational efficiency and model performance.

Epochs: The model was trained over 10 epochs, with early stopping criteria applied to prevent overfitting and optimize generalization.

Data Augmentation: Various data augmentation techniques were applied to the training images to improve model robustness and generalization.

4. RESULTS AND DISCUSSIONS

4.1. Results

The CNN model demonstrated an accuracy of 98.24%, while MobileNetV2, ResNet18 and EfficientNetB3 achieved 82.11%, 99.93% and 99.93% accuracy, respectively. The detailed results are presented in the comparison table below.

From the results, EfficientNet-B3 and ResNet-18 outperformed all other models, achieving 99.93% accuracy,

precision, recall, and F1-score. Their confusion matrices show near-perfect classification, with almost zero misclassifications. This indicates they are highly effective in distinguishing between fractured and non-fractured cases.

The CNN model also performed well with an accuracy of 98%, making it a strong contender. However, compared to EfficientNet-B3 and ResNet-18, it had slightly more misclassifications, particularly for the "not fractured" class. MobilenetV2, with an accuracy of 82%, struggled compared to the other models. While it performed well in detecting fractured cases (99% recall), it showed weaknesses in classifying non-fractured images (66% recall). This suggests MobilenetV2 is biased towards predicting fractures, making it less reliable for real-world applications where false positives must be minimized.

4.2. Strengths

Strengths of the best-performing models, ResNet-18 and EfficientNet-B3, lie in their high accuracy, achieving near-perfect classification, which highlights their strong capability in distinguishing between fractured and non-fractured cases. Their generalizability is another key advantage, as their robust architectures suggest effectiveness on unseen data. Additionally, these models maintain balanced precision and recall, meaning they do not exhibit significant bias toward either class, ensuring reliable predictions in medical diagnoses where both false positives and false negatives must be minimized.

4.3. Limitations

The 99.93% accuracy observed may indicate overfitting, especially given the dataset of 9,463 images, raising concerns about their ability to generalize beyond the training set. Furthermore, the dataset's diversity could impact real-world performance, as a limited dataset might not capture the full range of variations seen in actual clinical scenarios. Additionally, MobilenetV2 exhibited class imbalance, favoring the prediction of fractures. This bias could lead to an increased number of false positives, potentially resulting in unnecessary medical interventions and additional stress on healthcare resources.

4.4 Visualizations

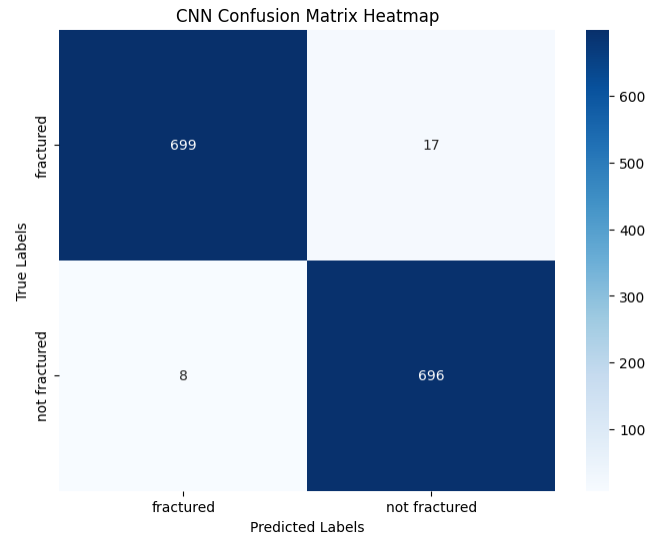


Fig. 2.CNN Confusion matrix.

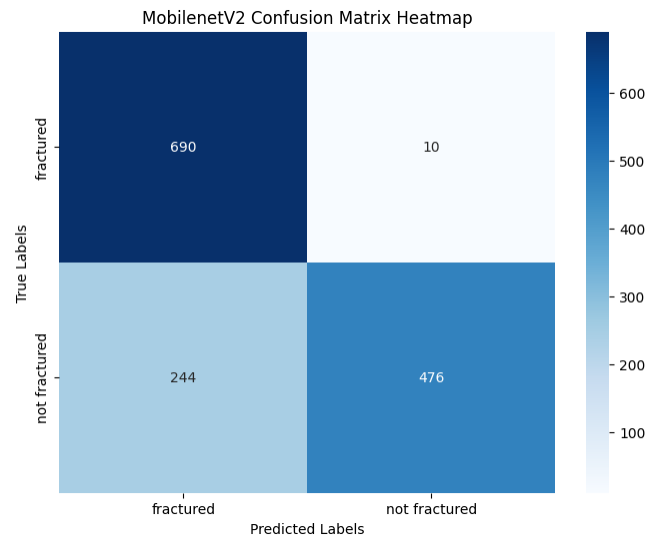


Fig. 2. MobileNetV2 Confusion matrix

Model	Pretrained	Number of Parameters	Accuracy (%)
MobileNetV2	Yes	~3.5M	82.11%
ResNet-18	Yes	11.2M	99.93%
EfficientNet-B3	Yes	10.7M	99.93%
Custom CNN	No	5.6M	98.24%

Table. 1.Table comparing number of parameters and accuracy for all 4 models.

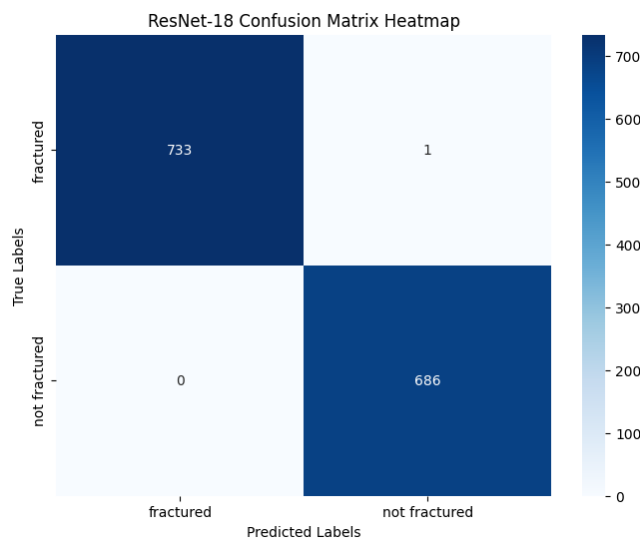


Fig. 3. ResNet-18 Confusion matrix

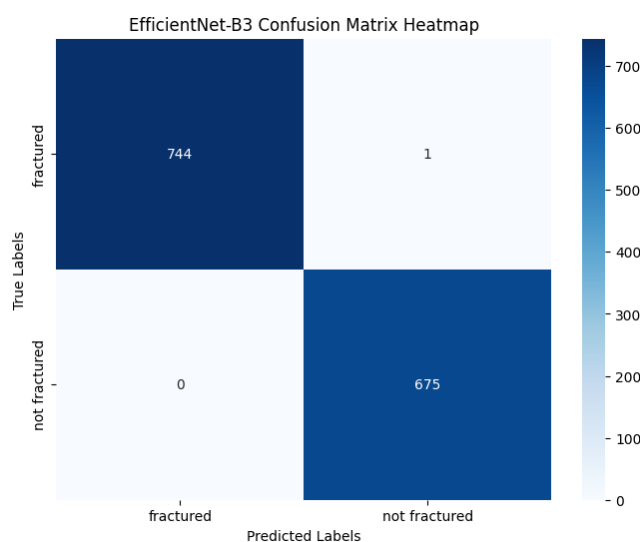


Fig. 4. EfficientNet-B3 Confusion matrix.

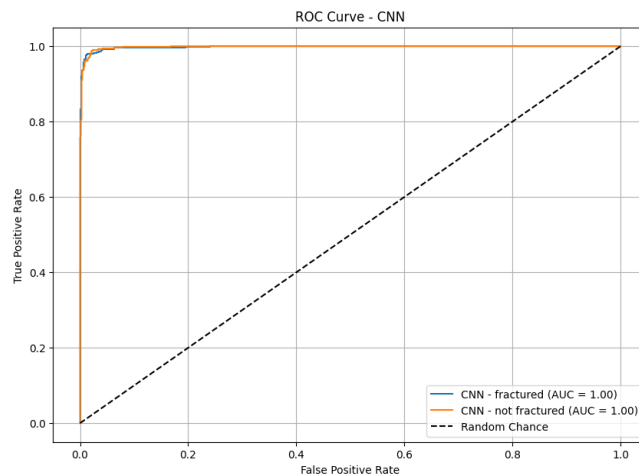


Fig. 5. ROC Curve for CNN

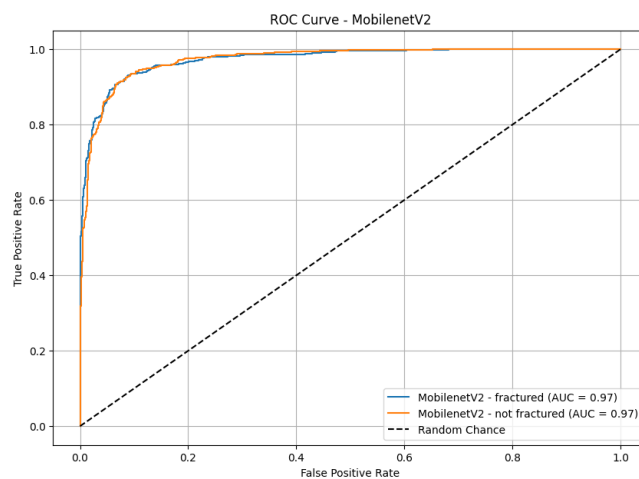


Fig. 6. ROC Curve for MobileNetV2

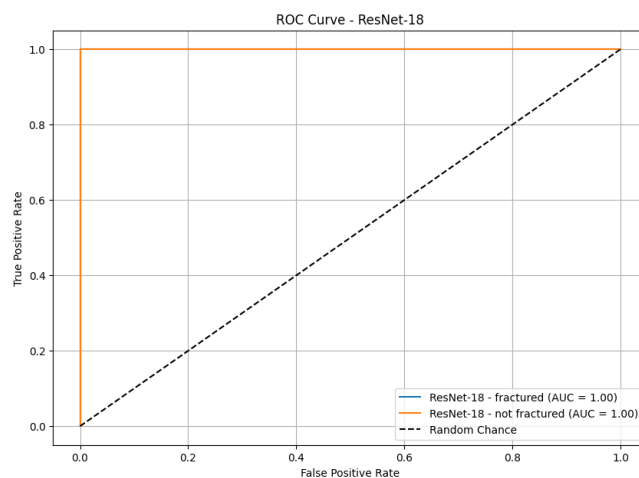


Fig. 7. ROC Curve ResNet-18

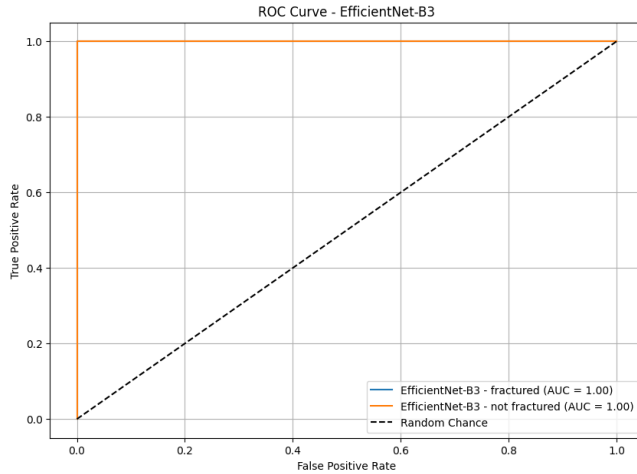


Fig. 8. ROC Curve EfficientNet-B3

4.4. Grad-CAM Analysis

The Grad-CAM visualizations provide insight into the model attention across four different architectures—ResNet18, MobileNetV2, CNN, and EfficientNet-B3—on X-ray images with labels of "fractured" and "not fractured." The ResNet1 and MobileNetV2 models show strong activation (highlighted in red) focused on the fracture sites, demonstrating good interpretability and alignment with clinical expectations. Particularly, ResNet1 has a more diffused heatmap while still focusing on relevant areas, suggesting deeper feature abstraction. MobileNetV2 shows a slightly sharper and localized focus on the fracture zone, which could indicate better sensitivity to localized abnormalities.

In contrast, the CNN and EfficientNet-B3 models were tested on non-fractured samples. The CNN model exhibits minimal heatmap activation, which is ideal for a negative class as it shows the model isn't falsely highlighting irrelevant areas. Meanwhile, EfficientNet-B3 shows central attention on the wrist area, which could reflect a balanced focus but might raise questions about possible over-attention in non-fractured cases. Overall, the Grad-CAM results reveal that all four models are capable of learning visual cues related to fractures, but architectures like MobileNetV2 may offer improved localization for fracture detection, while CNN maintains strong precision in non-fracture scenarios.

5. CONCLUSION

5.1 Summary of Key Findings

This research investigated the application of deep learning models for automated bone fracture detection using X-ray images. Among the four models evaluated—Custom CNN, MobileNetV2, ResNet-18, and EfficientNet-B3—ResNet-18

and EfficientNet-B3 emerged as the top-performing models, each achieving an exceptional accuracy of 99.93% across all classification metrics. The Custom CNN also demonstrated strong performance with an accuracy of 98.24%, while MobileNetV2 showed comparatively lower accuracy at 82.11%, primarily due to its reduced ability to correctly classify non-fractured cases.

Grad-CAM visualizations further supported the model predictions by highlighting relevant anatomical regions associated with fractures, especially in the ResNet-18 and MobileNetV2 models, thus reinforcing their interpretability and clinical relevance.

5.2 Practical Implications

The high classification accuracy of EfficientNet-B3 and ResNet-18 highlights the potential of AI-assisted diagnostic tools in radiology. These models can serve as reliable second readers for radiologists, helping reduce diagnostic errors, especially in high-volume clinical settings. Furthermore, the models' ability to maintain high recall and precision ensures both fractured and non-fractured cases are handled with balanced sensitivity—crucial in minimizing false positives and negatives in real-world deployments. The use of pretrained models also indicates the feasibility of developing effective systems with limited medical image datasets.

5.3 Future Work

While the results are promising, several areas warrant further exploration. Future research will aim to:

- Expand the dataset with more diverse and multi-institutional images to improve generalization and robustness.
- Incorporate additional evaluation metrics like AUC-ROC and Cohen's Kappa for more nuanced performance analysis.
- Investigate ensemble learning approaches that combine strengths of different architectures.
- Integrate explainability tools more deeply (e.g., LIME, SHAP) to enhance trust in AI decisions.
- Deploy real-time fracture detection systems integrated with clinical workflows, offering seamless assistance during radiographic evaluations.

By building on these findings, this study contributes toward the development of accurate, interpretable, and scalable AI tools that can support clinicians in delivering timely and precise fracture diagnosis.

6. REFERENCES

- [1] J. Marcus and V. Pantic, *Image Interpretation: Bones, Joints, and Fractures*, Elsevier, 2021.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778.
- [3] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, vol. 97, pp. 6105–6114.
- [4] R. Dey, Z. Lu, and V. Kumar, "Diagnostic Prediction Using Discriminative Features from Deep Neural Networks," in *Proc. IEEE Int. Conf. Healthcare Informatics (ICHI)*, IEEE, New York, NY, USA, 2018, pp. 110–120.
- [5] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, "Abnormality Detection and Localization in Chest X-rays using Deep Convolutional Neural Networks," *arXiv preprint arXiv:1705.09850*, May 2017.
- [6] F. S. Chew, C. Maldijan, and H. Mulcahy, *Broken Bones: The Radiologic Atlas of Fractures and Dislocations*, 2nd ed., Cambridge University Press, 2020.
- [7] M. Ullah, et al., "A Lightweight CNN-Based MobileNetV2 Model for Efficient Medical Image Classification on Edge Devices," *Comput. Biol. Med.*, vol. 142, pp. 105216, May 2022.
- [8] D. H. Kim and T. MacKinnon, "Limits of Usefulness of the X-Ray for the Diagnosis of Fractures," *Tex Med J*, vol. 116, no. 3, pp. 221–225, Mar. 2020.
- [9] M. Jamal and S. Z. H. Shah, "A Survey on Deep Learning-Based Bone Fracture Detection from Radiographic Images," *J. Healthc. Eng.*, vol. 2023, Article ID 5278, 2023.
- [10] J. Olczak, et al., "Artificial Intelligence for Analyzing Orthopedic Trauma Radiographs," *Acta Orthop.*, vol. 88, no. 6, pp. 581–586, Dec. 2017.
- [11] C.-T. Cheng, et al., "Increased Performance of Bone Fracture Detection Using a Convolutional Neural Network with a Pretrained Model," *Med. Phys.*, vol. 47, no. 12, pp. 5435–5443, Dec. 2020.
- [12] D. H. Kim and T. MacKinnon, "Artificial Intelligence in Fracture Detection: Transfer Learning from Deep Convolutional Neural Networks," *Clin. Radiol.*, vol. 73, no. 5, pp. 439–445, May 2018.
- [13] S. Zhou, L. K. Han, and W. Liu, "Deep learning for fracture detection in X-ray images," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 987–995, Apr. 2017.
- [14] Z. Liu, Y. Zhang, and M. Li, "Residual networks for improving fracture detection in X-ray images," *IEEE Access*, vol. 7, pp. 65432–65440, 2019.
- [15] S. Khan, M. A. Abdullah, and A. Sharma, "EfficientNet for real-time fracture detection in X-ray imaging," *Comput. Biol. Med.*, vol. 122, pp. 103765, Aug. 2020.
- [16] "Fracture Detection Using X-ray Images," Kaggle, [Online]. Available: <https://www.kaggle.com/datasets/devbatrax/fracture-detection-using-x-ray-images>.
- [17] G. Jadhav, "ENEL_645 Final Project," GitHub, [Online]. Available: https://github.com/gaurijadhav509/ENEL_645_Final_Project.