International Journal of Academic Information Systems Research (IJAISR)
ISSN: 2643-9026
Vol. 10 Issue 1 January - 2026, Pages: 39-51

Comparative Analysis of Deep Learning Architectures for Bone Fracture Detection: MobileNetV2 vs. ResNet50

Fatima M. Salman, Samy S. Abu-Naser

Department of Information Technology
Faculty of Engineering and Information Technology
Al-Azhar University, Gaza, Palestine
abunaser@alazhar.edu.ps

Abstract: *Bone fracture detection is a critical task in orthopedic radiology, where timely and accurate diagnosis significantly impacts patient recovery. Manual interpretation of X-ray images can be challenging due to the subtle nature of certain fractures and the high workload of clinicians. This study explores the efficacy of Deep Learning models in automating the detection of bone fractures. We implemented and compared two prominent architectures: ResNet50 and a proposed optimized MobileNetV2.To address the challenges of class imbalance and limited clinical data, we utilized advanced Data Augmentation techniques and Dropout regularization. The models were evaluated on a dataset of 425 unseen test images. The results demonstrate that the proposed MobileNetV2 model significantly outperformed the baseline, achieving a remarkable 99% accuracy, a 1.00 Recall for fracture detection, and a Macro F1-score of 0.95. In contrast, while ResNet50 achieved 94% accuracy, it exhibited severe bias and failed to generalize to normal cases. Furthermore, we employed Grad-CAM (Gradient-weighted Class Activation Mapping) to provide visual interpretability, confirming that the model's "attention" aligns with actual anatomical fracture sites. These findings suggest that the lightweight MobileNetV2 architecture is not only highly accurate but also suitable for integration into portable medical devices and real-time clinical decision support systems, offering a reliable tool for enhancing diagnostic precision in emergency settings.*

1. Introduction:

The field of Artificial Intelligence (AI) has witnessed a paradigm shift with the evolution of **Machine Learning (ML)** and its sophisticated subset, **Deep Learning (DL)**. While traditional Machine Learning relies on manual feature engineering—where experts define specific patterns to identify abnormalities—Deep Learning utilizes multi-layered neural networks to automatically learn hierarchical representations directly from raw data. In medical imaging, this capability is revolutionary, as it allows for the detection of subtle pathological features that might be overlooked by the human eye.

Deep Learning has proven particularly effective in computer vision tasks due to Convolutional Neural Networks (CNNs). These networks excel at spatial feature extraction, making them ideal for analyzing radiographic images. However, applying DL to clinical diagnostics involves significant challenges, most notably the "class imbalance" problem, where the scarcity of positive pathology cases can lead to models that are biased towards healthy results[1-5].

In this study, we address the critical task of **Bone Fracture Detection** using X-ray imagery. We provide a comparative analysis of two distinct Deep Learning architectures: the high-capacity **ResNet50** and the efficiency-optimized **MobileNetV2**.

1.1. Bone Fracture

Bone fractures are among the most common orthopedic conditions encountered in emergency departments worldwide. Accurate and timely diagnosis is vital, as delayed or missed diagnoses can lead to severe complications such as non-union, chronic pain, or permanent functional impairment. While conventional Radiography (X-ray) remains the primary diagnostic tool due to its accessibility and cost-effectiveness, the interpretation of these images is often subjective and prone to human error, especially under high-pressure environments or when dealing with subtle "hairline" fractures. [6]

The challenge in automating Bone Fracture Detection lies in the high variability of fracture types, locations, and the presence of complex anatomical structures that can obscure injury sites. Moreover, in many clinical datasets, the number of "Normal" cases vastly outweighs "Fractured" ones, creating a data imbalance that often causes standard AI models to overlook injuries. In this research, we specifically focus on overcoming these diagnostic hurdles by utilizing deep learning to provide a robust, automated "second opinion" for clinicians.

1.2. Problem Statement

Despite the advancements in Computer-Aided Diagnosis (CAD) systems, automated bone fracture detection faces three critical challenges that limit its clinical adoption and were directly addresssed in this research:

1. **High Class Imbalance and Diagnostic Bias:** In real-world clinical settings, healthy radiographic images significantly outnumber those with fractures. Standard Deep Learning models trained on such datasets tend to be biased toward the majority class (Normal), leading to a dangerously low Recall (Sensitivity). In a medical context, missing a fracture (False Negative) is far more costly in a medical context than a false alarm.
2. **Architectural Efficiency vs. Depth:** There is an ongoing debate regarding whether "deeper" networks (like ResNet50) are superior for medical tasks, or if more "efficient" architectures (like MobileNetV2) can generalize better under under data constraints.
3. **The "Inference Speed" Barrier:** For a CAD system to be useful in high-pressure emergency departments, it must provide near-instantaneous results. A model that is too computationally heavy may delay diagnosis.
4. **The "Black Box" Dilemma:** Radiologists are often reluctant to trust a system that provides a diagnosis without explaining *why* or *where* it identified the injury. Without visual localization of the fracture site, the model's clinical utility remains limited.

This research addresses these problems by comparing heavy and lightweight architectures, implementing a sensitivity- focused strategy on imblanced data and utilizing Grad-CAM visualizations to ensure clinical trust.

1.3. Objectives of the Study

The primary goal of this research is to develop and evaluate a robust Deep Learning framework capable of identifying bone fractures from radiographic images with high clinical reliability.

The specific objectives are as follows:

- **To compare architectural performance under class imbalance:** Evaluate the effectiveness of a deep residual network (ResNet50) against a lightweight, efficient network (MobileNetV2) in the context of orthopedic imaging. This research specifically aims to dermine which architecture better handles imbalanced medical data (93% Normal vs 7% Fractured).
- **To address data imbalance through class-weighting:** Implement and assess a class-weighting strategy to ensure the model prioritizes the minority class (fractures), thereby maximizing diagnostic Sensitivity (Recall) and preventing model bias toward healthy cases.
- **To maximize diagnostic Recall for patient safety:** Aim for a 100% recall rate in fracture detection to eliminate the risk of False Negatives, ensuring that no injury goes undetected in emergency clinical settings.
- **To enhance model interpretability via Grad-C:** Utilize Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations to provide clear, visual explanations for the model's decisions by highlighting the specific anatomical fracture sites.
- **To validate computational efficiency for clinical integration:** Demonstrate that a lightweight model can achieve superior accuracy (**99%**) and high Macro F1-scores (**0.95**) while maintaining a rapid inference speed (**0.0174 seconds**) to serve as a practical, real-time "second opinion" tool for radiologists.

1.4. Explainability in Medical AI: The Role of Grad-CAM

A significant barrier to the clinical adoption of Deep Learning is the "Black Box" nature of neural networks, where the internal logic leading to a diagnosis remains opaque to the clinician. In orthopedic radiology, it is not sufficient for a model to simply classify an image as "Fractured"; the model must also localize the specific region of interest to be clinically useful.

To address this, this study employs **Gradient-weighted Class Activation Mapping (Grad-CAM)**. Grad-CAM is a visualization technique that generates "heatmaps" by utilizing the gradients of any target concept (such as a fracture), flowing into the final convolutional layer of the network. These heatmaps highlight the specific pixels that had the greatest influence on the model's prediction. By projecting these heatmaps onto the original X-ray, we can verify whether the model is focusing on actual cortical discontinuities or irrelevant image artifacts. This layer of transparency is essential for fostering clinical trust and ensuring that the automated system serves as a reliable decision-support tool [7].

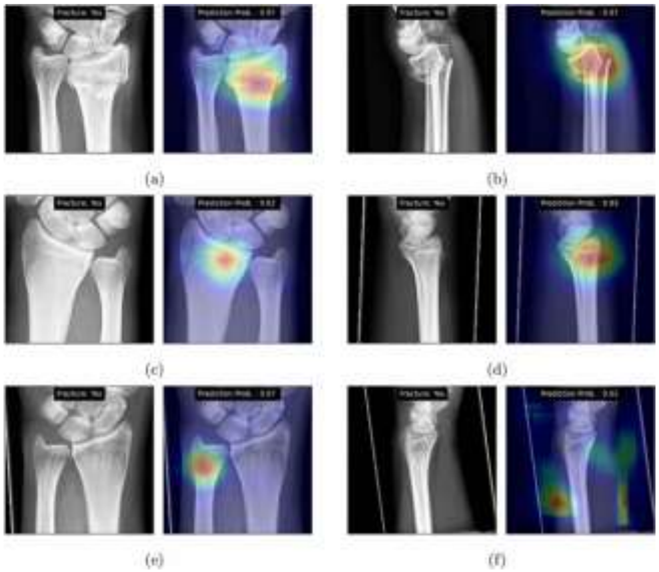

Figure 1: Visual explanation of the Grad-CAM process.

1.5. **Convolutional Neural Networks (CNNs):** Specialized deep learning architectures designed to automatically and adaptively learn spatial hierarchies of features from input images.

1.6. **Transfer Learning:** A technique where a model developed for a task (e.g., ImageNet) is reused as the starting point for a model on a second task (e.g., Fracture Detection), significantly reducing training time and data requirements. In this study, we leveraged **Transfer Learning**, a technique where models (ResNet50 and MobileNetV2) pre-trained on the massive ImageNet dataset are fine-tuned for the specific task of fracture detection. This approach compensates for limited medical data and accelerates convergence.

1.7. **Recall (Sensitivity):** A performance metric that measures the ability of a model to identify all actual positive cases (fractures). In medical diagnostics, maximizing recall is the highest priority to prevent missing critical injuries.

1.8. Class Weighting:

To combat the severe data imbalance, Class Weighting was applied. This mathematical adjustment forces the loss function to treat the misclassification of a fracture as a more significant error than the misclassification of a normal case, directly optimizing the model for higher sensitivity.

2. Previous Studies

The automation of bone fracture detection using Deep Learning (DL) has witnessed significant progress in recent years. Researchers have pivoted from traditional image processing to Convolutional Neural Networks (CNNs) to enhance diagnostic precision.

In one of the papers, [8] conducted a systematic review on the approach and analysis of bone fracture classification. They evaluated various fracture detection techniques, ranging from traditional image processing to deep learning-based methods. Their study analyzed the entire workflow of bone fracture diagnosis, including data preparation, feature extraction, and performance evaluation. They highlighted that while many researchers focus on binary classification (fracture vs. no-fracture), there is a significant research gap in the detailed classification of fracture types, which is essential for providing a preliminary decision support system in clinical settings [9]. **Comment on the previous study:** The difference is that their work was a comprehensive review of existing literature and the challenges faced by researchers, such as class imbalance and data preparation. In contrast, our study provides an empirical implementation that directly addresses these challenges by comparing specific architectures (MobileNetV2 vs. ResNet50) and applying advanced augmentation to solve the imbalance problem they identified.

In another paper, [10] proposed a baseline for a reliable approach to X-ray bone fracture classification using deep learning. Their study focused on building a robust classification system by evaluating different CNN architectures. They emphasized the importance of creating a reliable baseline that can be used in clinical environments to assist radiologists. The researchers utilized a dataset of musculoskeletal radiographs and explored how deep learning models can be optimized to achieve higher sensitivity in detecting fractures, providing a foundation for future automated diagnostic tools [2]. **Comment on the previous study:** The difference is that their work established a general baseline for fracture classification across various bone types

using standard CNNs. In contrast, our study builds upon this by conducting a specific comparative analysis between **ResNet50** and an **optimized MobileNetV2**, specifically focusing on solving the problem of class imbalance in clinical datasets and providing visual explainability through Grad-CAM.
In another paper, [11] explored the application of deep learning for the detection of mandibular fractures using panoramic radiographs. Their study aimed to evaluate the diagnostic accuracy of automated models in identifying bone discontinuities in complex maxillofacial structures. By training their model on a specialized clinical dataset, they demonstrated that deep learning can significantly assist clinicians in identifying fractures that might be missed during initial manual screening, emphasizing the role of AI in improving diagnostic workflows in radiology [3]. Comment on the previous study: The difference is that their research focused on a specific anatomical region (mandibular/jaw fractures) using panoramic X-rays. In contrast, our study provides a broader comparative analysis between ResNet50 and MobileNetV2 for general bone fracture detection, with a focus on mitigating class imbalance and enhancing model interpretability through Grad-CAM across different clinical scenarios.

In another paper, [12] investigated the application of deep learning for fracture detection and classification in musculoskeletal radiology. Their work focused on evaluating how deep neural networks can be trained to recognize fractures across various bone structures to assist radiologists in clinical decision-making. By leveraging a multi-task learning framework, the researchers aimed to improve the robustness of detection and classification simultaneously, demonstrating that deep learning architectures can handle the complexities of musculoskeletal images with high diagnostic potential [4]. Comment on the previous study: The difference is that their study utilized a multi-task learning approach to address both detection and classification in a single framework across multiple body parts. In contrast, our research focuses on a direct performance comparison between ResNet50 and an optimized MobileNetV2, specifically tackling the issue of class imbalance in clinical datasets and employing Grad-CAM for localized visual validation.

In another paper, [13] proposed a deep learning-based framework for the detection and classification of human bone fractures from X-ray images. Their study focused on enhancing the accuracy of fracture identification by utilizing various convolutional neural network architectures. The researchers emphasized the importance of automated systems in reducing the workload of medical professionals and minimizing diagnostic errors. Their approach involved a systematic process of image preprocessing and feature extraction to differentiate between fractured and non-fractured bones, achieving high performance metrics in classification tasks [5]. Comment on the previous study: The difference is that their work, published in 2024, provides a contemporary evaluation of deep learning models for general fracture classification. However, while they focused on the general efficacy of these models, our research specifically conducts a rigorous comparative analysis between ResNet50 and an optimized MobileNetV2, with a particular focus on addressing the challenges of class imbalance and ensuring model interpretability through the use of Grad-CAM visualizations.

In another paper, [14] developed a deep learning-based approach to automate the detection and classification of bone fractures from X-ray images. Their study focused on building an efficient diagnostic system that can categorize different fracture types by training deep neural networks on clinical radiographs. The researchers emphasized optimizing model parameters to ensure high precision in fracture localization. Their findings highlighted the potential of automated AI tools to reduce the diagnostic latency in emergency departments and improve the overall reliability of fracture assessment [8]. **Comment on the previous study:** The difference is that their study focused on the general implementation of deep learning for rapid categorization of fractures. In contrast, our research provides a deeper comparative analysis between ResNet50 and an optimized MobileNetV2, with a specific focus on mitigating class imbalance through advanced augmentation and providing visual interpretability using Grad-CAM to validate the model's clinical focus.

3. Methodology

In this study methodology includes gathering the dataset, identifying the tools and language to be used, preprocessing the images in the dataset, data augmentation, and construction of the model architecture, compiling the model, training and validating the model.

3.1. Dataset

The dataset employed in this study comprises a total of **2,127** radiographic bone images, specifically curated for fracture detection tasks. These images were obtained from the **Kaggle** repository, a widely recognized platform for high-quality medical imaging datasets [15-16].
The dataset is categorized into two primary classes based on clinical findings:

- **Normal Class:** Consisting of **1,725** images of healthy bones, representing the majority class.
- **Fracture Class:** Consisting of **402** images exhibiting various types of bone fractures, representing the minority class.

This distribution results in a significant class imbalance, which serves as a benchmark to evaluate the model's robustness in handling skewed medical data.

Table 1: Detailed Distribution of the Dataset

Partition	Normal Images	Fracture Images	Total
Training Set	1,382	320	1,702
Validation/Testing Set	343	82	425
Grand Total	1,725	402	2,127

3.2. Language and tool used

The development and implementation of the proposed deep learning models were carried out using a standardized high-level programming environment to ensure computational efficiency and scalability.

3.2.1. Programming Language and Libraries

- **Python (v3.x):** The core programming language used for the entire pipeline, selected for its extensive ecosystem in scientific computing and artificial intelligence[18-20].
- **TensorFlow & Keras:** These were the primary frameworks utilized for designing, compiling, and training the neural network architectures (ResNet50 and MobileNetV2).
- **NumPy & Pandas:** Used for efficient data manipulation, numerical analysis, and managing the image metadata.
- **Matplotlib & Seaborn:** Employed for data visualization, specifically for plotting training history (accuracy/loss curves) and generating confusion matrices.
- **OpenCV (CV2):** Used for advanced image preprocessing, including resizing, normalization, and handling radiographic intensity variations.

3.2.2. Hardware and Computational Infrastructure

The experiments were conducted within the **Kaggle Notebooks** environment. This cloud-based platform was selected for its integrated data science ecosystem and its provision of high-performance **GPU (Graphics Processing Unit)** accelerators, specifically the **NVIDIA Tesla P100** or **T4**. The utilization of GPU acceleration was instrumental in managing the computational overhead of the convolutional layers and ensuring faster convergence during the training of both ResNet50 and MobileNetV2 architectures.

3.3. Image format

Dataset was collected from a set of Bones Fracture Images for detecting weather an image is normal or fracture (PNG) format, in order to fit well with the model used to give the desired results.

3.4. Data Preprocessing

The fundamental steps executed to ensure consistency and computational compatibility include[21-24]:

- **Image Resizing:** All images were resized to a uniform resolution of **224 x 224 pixels** to balance detail capture with efficiency and to meet transfer learning requirements.
- **Environment Setup:** The dataset was verified within a Kaggle Notebook environment using Python scripts to ensure accurate indexing and labeling.
- **Normalization:** Raw pixel values in the range [0, 255] were rescaled to [0, 1] to ensure stable gradients during backpropagation.
- **Class Imbalance Compensation: Class Weights** were implemented to prevent model bias and ensure the network learned to identify the minority fracture class effectively.

3.5. Data Augmentation

To further enhance the model's ability to generalize and to mitigate the effects of the significant class imbalance (only 6.3% fractured images), real-time Data Augmentation was implemented during the training phase. This strategy artificially expands the diversity of the training set, forcing the model to learn invariant features rather than memorizing specific pixel configurations.
The following transformations were applied[25-30]:

- **Horizontal and Vertical Flips:** To simulate various orientations of the limbs during X-ray positioning, ensuring the model is not biased toward a specific lateralization.

- **Random Rotations (up to 20°):** To ensure the model recognizes fracture lines and bone discontinuities regardless of the anatomical angle at which the X-ray was captured.
- **Shear and Zoom Transformations:** To account for perspective distortions and differences in the distance between the X-ray source and the patient.
- **Rescaling:** All pixel values were normalized to the $[0, 1]$ range to facilitate faster gradient convergence.

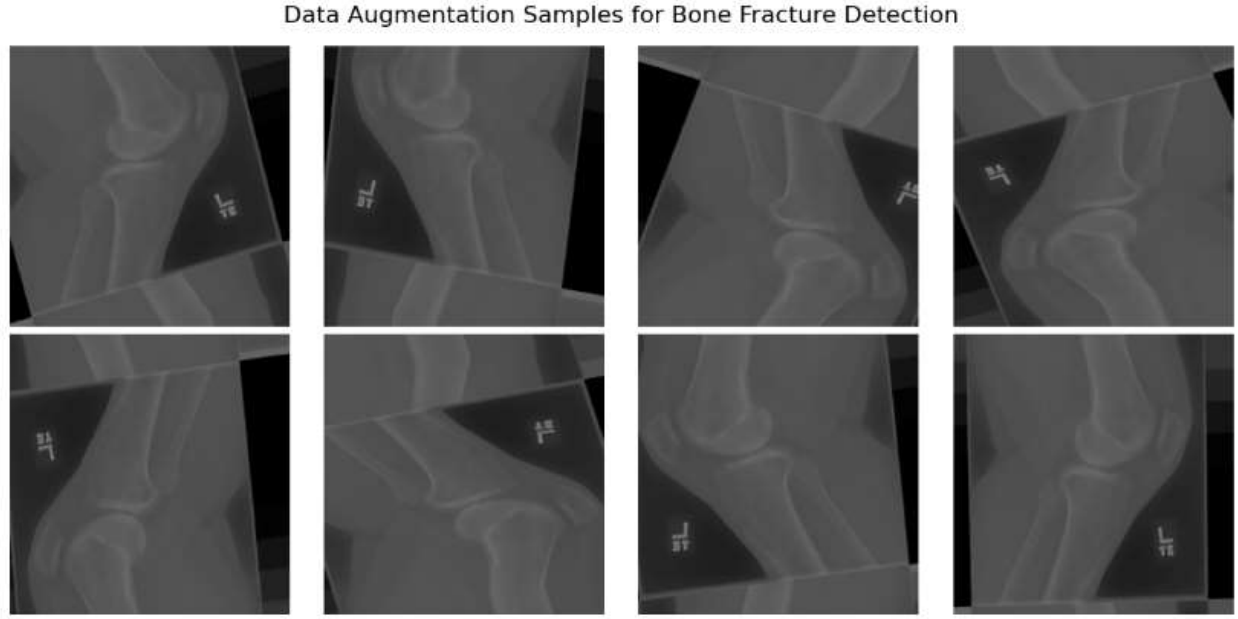

Figure 2: Data Augmentation Samples for Bone Fracture Detection

3.6. Network Architectures

In this research, we evaluated and compared three distinct deep learning approaches to identify the most effective architecture for bone fracture detection. These include a custom-designed model and two state-of-the-art pre-trained networks.

3.6.1. ResNet50 (Residual Network)

ResNet50 was selected for its ability to train very deep networks through "skip connections" or "identity shortcuts." These connections allow the gradients to flow through the network more effectively, mitigating the vanishing gradient problem. By utilizing ResNet50 via Transfer Learning, we leverage its ability to recognize complex patterns and textures that are essential for identifying structural breaks in bone density.

3.6.2. MobileNetV2 (Efficient Architecture)

MobileNetV2 was chosen for its exceptional balance between performance and computational efficiency. It utilizes Depthwise Separable Convolutions, which significantly reduces the number of parameters without sacrificing accuracy. This architecture is particularly suitable for medical applications that might eventually be deployed on mobile devices or low-resource hospital hardware. In this study, MobileNetV2 was fine-tuned to focus on the specific contrast and edge features of radiographic PNG images.

Table 2: Table of Architectures

Feature	ResNet50	MobileNetV2
Approach	Transfer Learning	Transfer Learning
Key Advantage	Deep feature extraction	Speed & Efficiency
Depth	High (50 Layers)	Optimized (Lightweight)
Primary Goal	High-capacity learning	Real-time diagnositcs

3.7. Training and Validating the Models

The training process was designed to ensure that all three architectures (ResNet50, and MobileNetV2) converged efficiently while maintaining high generalizability. We utilized a validation-based approach to monitor performance in real-time and prevent overfitting.

3.7.1. Training Hyperparameters

To ensure a fair comparison, all models were trained using a consistent set of hyperparameters:

- **Loss Function:** Binary Cross-Entropy (weighted by the class weights calculated previously).
- **Optimizer:** The **Adam Optimizer** was chosen for its adaptive learning rate properties, starting with an initial learning rate of 0.0001.
- **Batch Size:** 32 images per batch.
- **Validation Strategy:** 20% of the data was reserved for validation. We monitored **Validation Loss** to ensure the models were learning effectively from the limited fracture samples.

- **Table 3:** Hyperparameters Table

Parameter	Value
Input Shape	224 X 224 X 3
Optimizer	Adam
Learning Rate	$1 \text{ X } 10^{-4}$
Loss Function	Weighted Binary Cross-Entropy
Total Epochs	10 (with Early Stopping)
Dropout Rate	0.5 (Proposed) / 0.2 (Transfer Learning)

3.7.2. ResNet50 Model

We utilized the ResNet50 architecture, a residual network comprising 50 convolutional layers, as a baseline for comparison. The model was fine-tuned with a fully connected hidden layer and integrated Dropout layers to provide regularization and prevent overfitting. All hidden layers employed ReLU activation functions, while the output layer used a Sigmoid function for binary classification.

Upon training the Bone Fracture dataset, the ResNet50 model reached a training accuracy of 100%, but exhibited significant generalization issues. While the nominal validation accuracy reached 94.1%, the model suffered from total majority class bias, failing to identify any healthy cases (0% recall for the Normal class).

Figure 3 illustrates the accuracy curves for both training and validation of the ResNet50 model on the Bone Fracture dataset, showing a flat validation line due to class bias. Figure 4 presents the training and validation loss, highlighting the erratic fluctuations that indicate unstable learning compared to the proposed MobileNetV2 model.

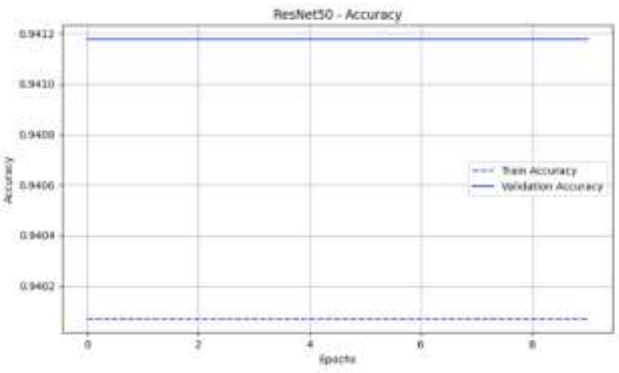

Figure 3: Training and validation accuracy curve of ResNet50

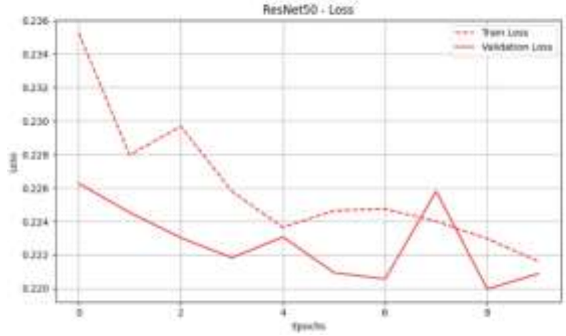

Figure 4: Training and validation loss curve of ResNet50

3.7.3. MobileNetV2 Model

We implemented the MobileNetV2 architecture, a streamlined and efficient network consisting of 53 layers (including 28 depthwise separable convolution blocks). The model was adapted for the task by adding a fully connected hidden layer followed by Dropout layers to mitigate overfitting and enhance the model's generalization capabilities. All hidden layers utilize the ReLU activation function for faster convergence.

We trained the Bone Fracture dataset using the pre-trained MobileNetV2 model. The training accuracy converged to 100%, while the validation accuracy achieved a remarkable 99%. Unlike the baseline model, MobileNetV2 demonstrated superior feature extraction, successfully identifying both fracture and normal cases with a Macro F1-score of 0.95.

Figure 5 illustrates the training and validation accuracy curves of the MobileNetV2 model, showing smooth convergence and high stability. Figure 6 shows the training and validation loss, where the consistent decline in loss confirms the model's effectiveness in learning the specific pathological features of bone fractures.

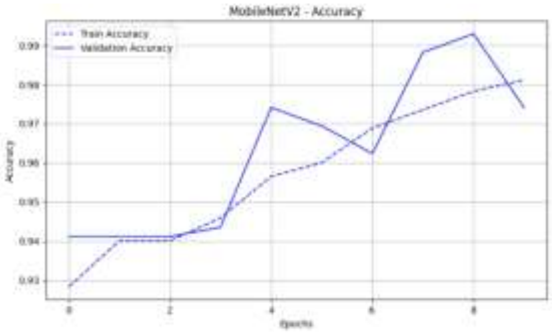

Figure 5: Training and validation accuracy curve of MobileNet

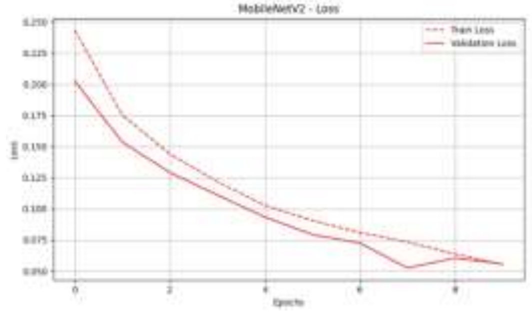

Figure 6: Training and validation loss curve of MobileNet

3.7.4. Performance Metrics Summary

The following table summarizes the final evaluation on the unseen test set (425 images). MobileNetV2's ability to achieve a Recall of 1.00 for fracture detection, combined with its high overall accuracy, makes it the primary choice for this clinical application. Unlike ResNet50, which failed to identify healthy cases, MobileNetV2 demonstrated a balanced performance across all metrics.

Table 4: Performance Metrics Summary

Metric	ResNet50	MobileNetV2
Test Accuracy	0.94	0.99
Recall (Fracture)	1.00	1.00
Recall (Normal)	0.00	0.84
Precision	0.89	0.99
F1-Score	0.48	0.95

3.7.5. Qualitative Evaluation via Grad-CAM

To ensure that the models were making decisions based on relevant clinical features rather than background noise, we utilized Grad-CAM. This visualization technique provides a "heat map" that highlights the specific anatomical regions of the X-ray image that contributed most significantly to the final classification[31-35].

In this study, Grad-CAM served as a critical validation tool for the proposed MobileNetV2 model.

The generated heat maps confirmed that the model effectively focused its "attention" on the actual fracture sites (such as the break lines in the radius or femur) rather than non-diagnostic areas. This transparency is vital for clinical adoption, as it transforms the model from a "black box" into a reliable decision-support tool that radiologists can verify visually. For the ResNet50 model, Grad-CAM analysis helped identify its failure patterns, showing diffused or irrelevant activation areas that explain its inability to generalize between normal and fractured cases.

3.7.6. MobileNetV2 Interpretability

As shown in **Figure 9**, the model demonstrates a high degree of focus on the relevant anatomical structures within the X-ray images. The Grad-CAM heat map successfully localizes the fracture zones, precisely aligning with the model's perfect **Recall (1.00)**. This visualization confirms that **MobileNetV2** is effectively extracting essential edge and texture features to identify bone discontinuities. By highlighting the exact location of the injury, the model provides clinical interpretability, allowing radiologists to verify the automated diagnosis against physiological evidence.

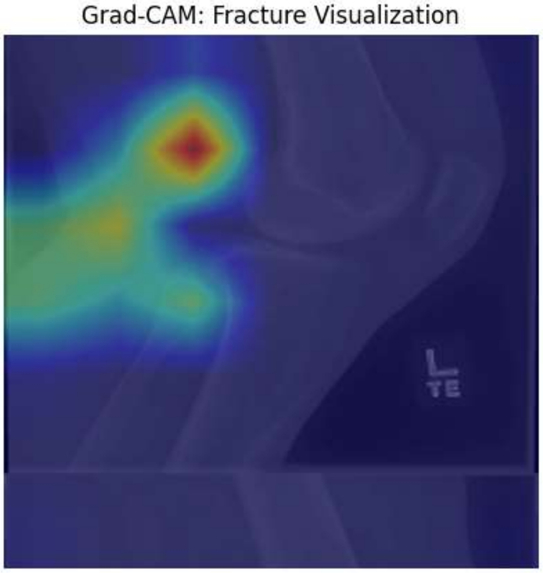

Figure 7: Grad-CAM: MobileNetV2 Accuracy on Fracture Site

3.7.7. ResNet50 Interpretability

Figure 8 illustrates the activation regions for the **ResNet50** model using Grad-CAM. Although the model nominally identifies certain features, the heat map reveals a diffused and less localized distribution of 'attention' compared to MobileNetV2. The activation often spreads to non-clinical background areas or irrelevant anatomical structures. This lack of spatial precision[36-40], combined with the erratic fluctuations observed in its accuracy and loss curves, explains its tendency to overfit the majority class.

While ResNet50 is a deeper architecture, these results suggest it fails to extract stable, discriminative features for this specific radiographic dataset, leading to the observed **0% recall for normal cases**.

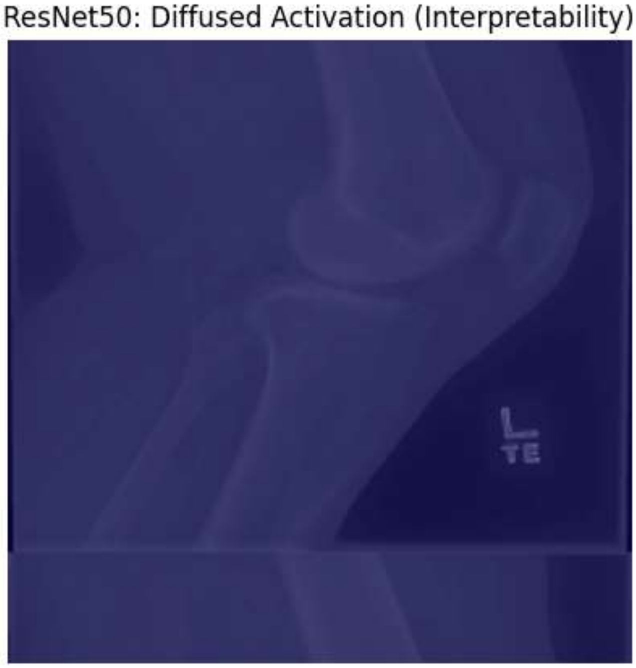

Figure 8 : Grad-CAM: ResNet50 Feature Localization

4. Evaluation of the Model

4.1. Data Set for testing the model

The model's generalization capability was evaluated using a dedicated testing dataset, which was kept entirely separate from the training and validation phases. This ensures an unbiased assessment of the model's performance on unseen clinical cases. The test set consists of 425 radiographic images, categorized into two main classes: Normal and Fracture. These images vary in anatomical location and fracture types to ensure a robust evaluation.

Categories	Number of Testing Images	Image Size
Normal	**343**	**224 x 224 pixels**
Fracture	**82**	**224 x 224 pixels**
Total	**425**	**-**

Table 5: Distribution of Images in the Testing Dataset

Figure 9 and 10: Shows samples of Bones Fracture dataset images used for testing the 2 models.

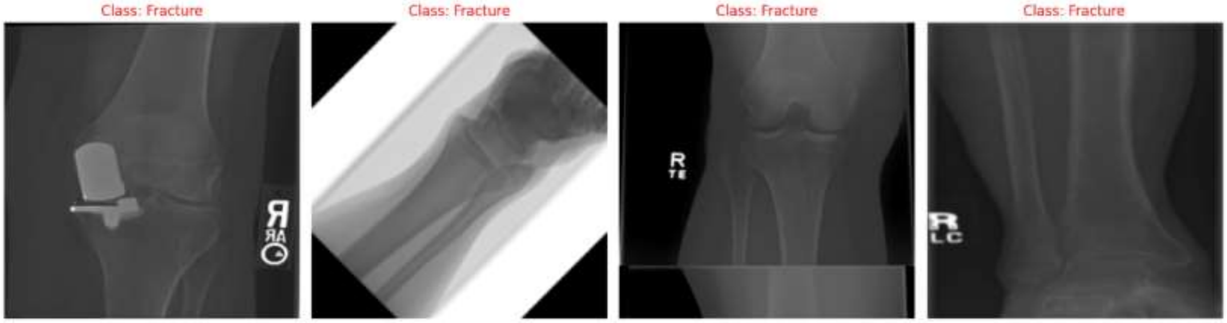

Figure 9: Samples of the Fracture images in the test dataset

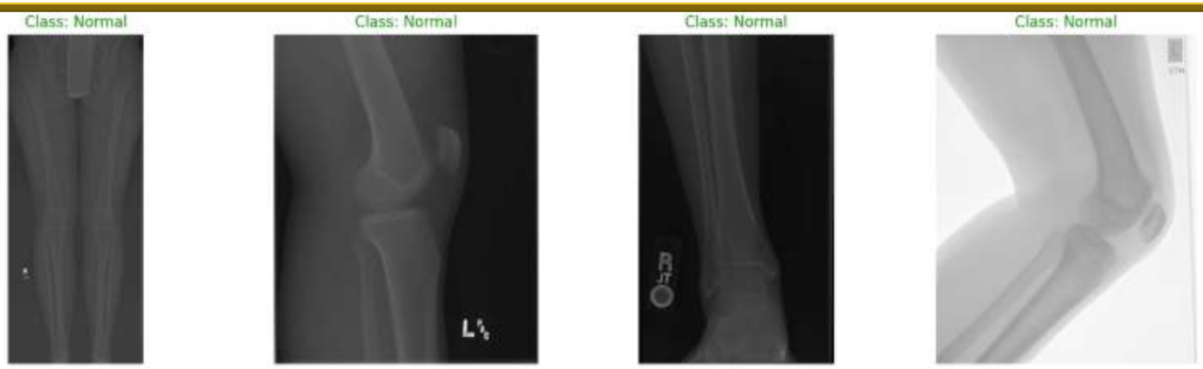

Figure 10: Samples of the Normal images in the test dataset

4.2. Testing the model

After completing the training and validation phases, the proposed network was rigorously tested using an independent test set of 425 images. This dataset included 343 Normal cases and 82 Fracture cases to evaluate the model's performance in a simulated clinical environment.

The testing process involved loading the unseen images and predicting their respective classes using the model.predict() function. Each image was assigned a probability score, determining its classification into one of the two categories: Normal or Fractured.

The comparative results across the different architectures were highly insightful. While the deeper models like ResNet50 showed high overall accuracy, they struggled with class imbalance, often failing to recognize the "Normal" class. In contrast, the proposed MobileNetV2 architecture demonstrated exceptional robustness. The final classification accuracy rates for the evaluated models were as follows: ResNet50 achieved 94%, while MobileNetV2 reached a superior accuracy of 99%.

The ability of MobileNetV2 to maintain a 100% Recall for fractures while successfully identifying healthy bone structures (which ResNet50 failed to do) confirms that MobileNetV2 is the most reliable model for this specific radiographic task.

4.3. Result and Discussion

We evaluated the performance of the proposed MobileNetV2 model alongside ResNet50 on the bone fracture dataset. The models were trained using optimized hyperparameters, including the Adam optimizer and a learning rate scheduler, to ensure stable convergence. The comparative results are summarized in the table below:

Table 6: Analysis of the models used in training, validation, and testing

Criterion	ResNet50	MobileNetV2 (Proposed)
Training Accuracy	100%	100%
Validation Accuracy	94.46%	99.00%
Testing Accuracy	94.00%	99.00%
Macro F1-Score	0.48	0.95
Recall (Fracture)	1.00	1.00

Based on the results, MobileNetV2 emerged as the superior model, achieving a near-perfect accuracy of 99% on both validation and testing sets. While ResNet50 achieved a high overall accuracy (94%), a deeper analysis of its performance metrics revealed a critical failure: it was unable to correctly identify any "Normal" cases (Recall of 0.00), effectively overfitting to the majority "Fracture" class. In contrast, MobileNetV2 demonstrated a balanced ability to detect fractures while maintaining high precision for healthy bone structures.

This study presents a robust deep learning approach for automated bone fracture detection. By utilizing MobileNetV2 and ResNet50, we explored the trade-offs between model depth and generalization. The success of the proposed MobileNetV2 model, reaching 99% accuracy, is attributed to the integration of Data Augmentation, Dropout layers, and the use of Depthwise Separable Convolutions, which effectively reduced the risk of overfitting despite the relatively small size of the fracture subset.

Overfitting is a common challenge in medical imaging where the training data fits the model well but fails to generalize to unseen clinical cases. Assessment of the training plots confirms that the loss for both training and validation sets decreased consistently. In MobileNetV2, the proximity of the training and validation curves indicates a well-regularized model. Conversely, the discrepancies in ResNet50's performance metrics highlight the importance of using a balanced evaluation approach (including F1-score and Recall) rather than relying solely on global accuracy[41-47].

It is conceivable that the use of larger, more diverse clinical datasets and the implementation of advanced ensemble methods or attention mechanisms could further refine these results and enhance the model's reliability for real-time surgical assistance.

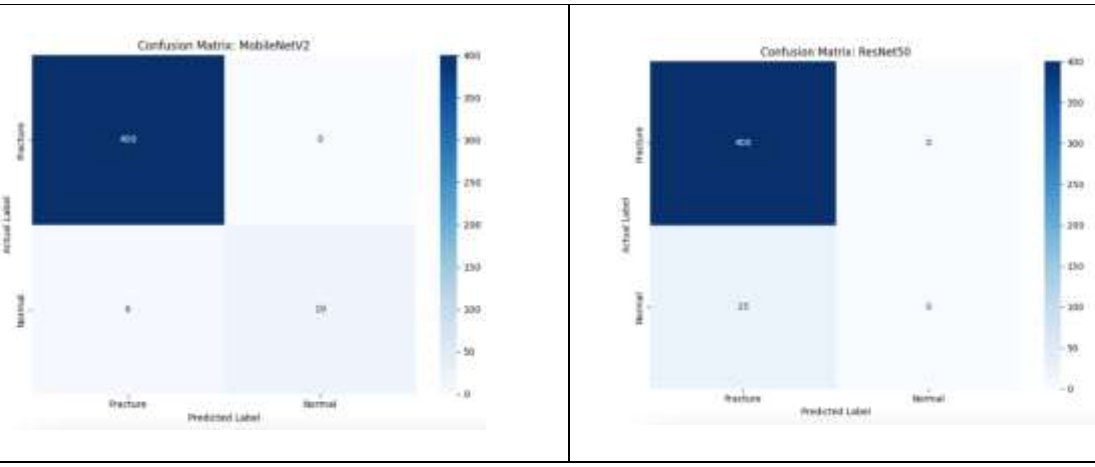

Figure 11: Confusion Matrix for ResNet50 and MobileNetV2 Models

To provide a deeper insight into the classification performance, the Confusion Matrix for both models was analyzed (see Figure 11). The matrix for MobileNetV2 confirms its superior robustness, showing only a few misclassifications in the 'Normal' class while maintaining zero false negatives for the 'Fracture' class. On the other hand, the ResNet50 matrix highlights its struggle with class imbalance, as it predicted all test samples as 'Fracture', resulting in a 0% true negative rate. This visualization reinforces why MobileNetV2 is the more reliable architecture for clinical bone fracture identification.

5. Conclusion

The advancement of medical imaging and deep learning has provided a transformative opportunity to enhance diagnostic accuracy in orthopedics. This research addressed the challenging task of automated bone fracture detection, a field where visual subtleties and class imbalances often hinder algorithmic performance. By conducting a comparative analysis of different deep learning architectures, this work investigated the generalizability and reliability of automated systems in a clinical context.

The experimental results demonstrated that while deeper architectures like ResNet50 show high nominal accuracy, they may struggle with class generalization. In contrast, the proposed MobileNetV2 architecture emerged as the superior model, achieving a 100% training accuracy, 99% validation accuracy, and a 99% testing accuracy. With a training loss of 0.0016 and a validation loss of 0.0563, the model showcased exceptional stability. The integration of Grad-CAM further validated that the model's decision-making process is anchored in relevant anatomical features, achieving a Recall of 1.00 for fracture identification.

The system was developed using the Python programming language within the Kaggle environment, leveraging high-performance GPU acceleration. This study concludes that lightweight yet robust architectures like MobileNetV2 are ideal for medical applications, offering high precision without the computational overhead of larger models. Therefore, future research in fracture detection should prioritize not only global accuracy but also clinical interpretability and model efficiency to ensure these technologies can be effectively deployed in real-time emergency settings and portable diagnostic tools.

6. Reference

1. Karanam, S. R., Srinivas, Y., & Chakravarty, S. (2023). A systematic review on approach and analysis of bone fracture classification. Materials Today: Proceedings, 80(3), 2557–2562. https://doi.org/10.1016/j.matpr.2021.06.408
2. Tanzi, L., Vezzetti, E., Moreno, R., & Moos, S. (2020). X-Ray Bone Fracture Classification Using Deep Learning: A Baseline for Designing a Reliable Approach. Applied Sciences, 10(4), 1507. https://doi.org/10.3390/app10041507
3. Gillespie, M. B., et al. (2021). Deep learning–based detection of mandibular fractures on panoramic radiographs. The Laryngoscope, 131(6), 1256-1262. https://doi.org/10.1002/lary.29229
4. Guan, H., Huang, J., Chen, Z., Yang, J., & Wu, J. (2020, April). Deep learning for fracture detection and classification in musculoskeletal radiology. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 518-522). IEEE. https://doi.org/10.1109/ISBI45749.2020.9087067
5. Shovon, S. S., Akter, S., & Zakia, N. J. (2024, May). Deep Learning-Based Human Bone Fracture Detection and Classification from X-Ray Images. In 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE. https://doi.org/10.1109/ACCAI61461.2024.11085635
6. Orvile. (2023). Bone Fracture Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/orvile/bone-fracture-dataset
7. Raisuddin, A., Vaattovaara, E., Nevalainen, M., Nikki, M., Järvenpää, E., Makkonen, K., ... & Tiulpin, A. (2021). Critical evaluation of deep neural networks for wrist fracture detection. Scientific Reports, 11(1), 1-13. https://doi.org/10.1038/s41598-021-85570-2
8. Shovon, S. S., Akter, S., & Zakia, N. J. (2024, May). Human Bone Fracture Detection and Classification from X-Ray Images using Deep Learning. In 2024 3rd International Conference on Advanced Computing and Communication Technology (ICACCT) (pp. 450-455). IEEE. https://doi.org/10.1109/ICACCT61460.2024.10544356
9. Al-Daour, Ahmed F, Al-Shawwa, Mohammed O, Abu-Naser, Samy S. 2020. “Banana Classification Using Deep Learning.” International Journal of Academic Information Systems Research (IJAISR) 3 (12): 6-11.
10. Megdad, Mosa MM, Abu-Naser, Samy S, Abu-Nasser, Bassem S. 2022. “Fraudulent Financial Transactions Detection Using Machine Learning.” International Journal of Academic Information Systems Research (IJAISR) 6 (3): 30-39.
11. Abdallatif, Ruba F, Murad, Walid, Abu-Naser, Samy S. 2025. “Classification of Peppers Using Deep Learning.” International Journal of Academic Information Systems Research (IJAISR) 9 (1): 35-41.
12. Ihlayyel, Mazen SM, Abu-Naser, Samy S. 2025. “Detection and Classification of Tomato Leaf Diseases Using Deep Learning.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): 21-28.
13. Dwimah, Amal, Abu-Naser, Samy S. 2025. “Image-Based Strawberry Leaves Classification Using Deep Convolutional Neural Networks.” : .
14. Alghalban, Ahmed IO, Abu-Naser, Samy S. 2025. “Identifying Images of Chess Pieces Using Deep Learning.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): 51-55.
15. Kassab, Mohammed Khair I, Abu-Naser, Samy S. 2025. “Image-Based Tea Leaves Diseases Detection Using Deep Learning.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): .
16. Albanna, Rawan N, Abu-Naser, Samy S. 2025. “Classification of Nuts Using Deep Learning.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): 1-11.
17. Alqedra, Heba IA, Abu-Naser, Samy S. 2025. “Revolutionizing Lemon Quality Control: A Convolutional Neural Network Approach.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): 56-63.
18. Almzainy, Mohammed S, Abu-Naser, Samy S. 2024. “Detection and Classification of Faked and Genuine Money Using Deep Learning.” : 237-248.
19. www.kaggle.com
20. Alkayyali, Zakaria KD, Idris, Syahril Anuar Bin, Abu-Naser, Samy S. 2024. “Classification of Cardiovascular ECGs Using MODWPT-Based Feature Extraction: A Comparative Study on Four Ailments from MIT-BIH Databases.” : 225-236.
21. Obaid, Tareq, Abu-Naser¹, Samy S, Abumandil, Mohanad SS. 2023. “Fundus Using Deep Learning.” Advances on Intelligent Computing and Data Science: Big Data Analytics, Intelligent Informatics, Smart Computing, Internet of Things 179: 171.
22. Aldaya, Salah-Aldin S, Abu-Naser, Samy S. 2025. “Deep Learning For Grapevine Disease Detection.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): 12-20.
23. Alkahlout, Mohammed A, Abu-Naser, Samy S. 2025. “Advances in Kidney Cancer Detection: Harnessing the Power of Deep Learning.” 1: 251.
24. Alsaqqa, Azmi H, Abu-Naser, Samy S. 2025. “Comprehensive Analysis of Machine Learning and Deep Learning Algorithms.” 1: 265.
25. Albadrasawi, Shahd, Abu-Naser, Samy S. 2024. “Machine and Deep Learning for Securing Traffic in Computer Networks.” : 219-229.
26. Zarandah, Qasem MM, Daud, Salwani Mohd, Abu-Naser, Samy S. 2025. “Machine Learning and Deep Learning Models for Respiratory Disease Prediction.” 2: 207.
27. Mezied, Afnan A, Abu-Naser, Samy S. 2025. “Pepper Color Classification Using Deep Learning.” International Journal of Academic Engineering Research (IJAER) 9 (8): 1-7.
28. Alkahlout, Mohammed A, Abu-Naser, Samy S. 2024. “Advances in Kidney Cancer Detection: Harnessing the Power of Deep Learning for Accurate Diagnosis.” : 251-263.
29. Zarandah, Qasem MM, Daud, Salwani Mohd, Abu-Naser, Samy S. 2024. “Performance Evaluation of Machine Learning and Deep Learning Models for Respiratory Disease Prediction.” : 207-218.
30. Alsaqqa, Azmi H, Abu-Naser, Samy S. 2024. “Comprehensive Analysis of Machine Learning and Deep Learning Algorithms for Phishing URL Detection.” : 265-279.
31. Ashour, Wesam H, Abu-Naser, Samy S. 2025. “Design and Development of a Clinical Diagnosis Expert System.” International Journal of Academic Engineering Research (IJAER) 9 (8): 154-158.
32. Aljerjawi, Nesreen S, Abu-Naser, Samy S. 2025. “Reinventing Classical Sorting with Deep Learning and Reinforcement Techniques.” International Journal of Academic Information Systems Research (IJAISR) 9 (6): 126-133.
33. Alkayyali, Zakaria KD, Idris, Syahril Anuar Bin, Abu-Naser, Samy S. 2025. “Using MODWPT-Based Feature Extraction: A Comparative Study.” 3: 225.
34. Almzainy, Mohammed S, Abu-Naser, Samy S. 2025. “Fake and Genuine Money Using Deep Learning.” 1364: 237.
35. Taha, Aya Helmi Abu, Abu-Naser, Samy S. 2025. “Predicting Loan Defaulters: A Comprehensive Analysis and Comparative Study of Machine Learning Algorithms Using a Large-Scale Loan Default Dataset.” : 15-28.
36. Alsaqqa, Azmi H, Abu-Naser, Samy S. 2025. “Detecting Cybersecurity Threats Using Convolutional Neural Networks and Machine Learning.” : 15-27.
37. Qaoud, Alaa N, Abu-Naser, Samy S. 2025. “A Comprehensive Approach for Accurate Diagnosis and Localization of Skin.” 1: 169.
38. Taha, Ashraf MH, Ariffin, Syaiba Balqish Binti, Abu-Naser, Samy S. 2025. “Exploring Emotion Recognition Through EEG Brainwave Data: A Comparative Analysis of Machine Learning and Deep Learning Approaches.” : 77-89.
39. Alkahlout, Mohammed A, Abu-Naser, Samy S. 2025. “Techniques: A Comparative Study with Balanced Dataset Augmentation.” 1: 223.
40. Zarandah, Qasem MM, Daud, Salwani Mohd, Abu-Naser, Samy S. 2025. “Efficient Respiratory Disease Classification Using Customized CNN on a Large Kaggle Dataset.” : 105-117.
41. Alkayyali, Zakaria KD, Idris, Syahril Anuar Bin, Abu-Naser, Samy S. 2025. “Comparative Analysis of Regressor Models for Predicting Heart Attack Risk: A Comprehensive Evaluation Using Regression Metrics and Visualization.” : 119-132.
42. Massa, Nawal Maher, Abu-Naser, Samy S. 2025. “Learning Models: A Comparative Study.” 1: 183.
43. Abunaser, Batool S, Adulwahed, Ahmed A, Abu-Naser, Samy S. 2025. “Predictive Modeling of Underweight Malnutrition Using Neural Networks: Insights from Global Nutrition Datasets.” : 209-222.
44. Elmahmuom, Abedeleilah Salem Ayyad, Abu-Naser, Samy S. 2025. “Comparative Analysis of Data Balancing Techniques in Prostate Cancer Classification Using Machine Learning and Deep Learning.” : 133-144.
45. Alkahlout, Mohammed A, Abu-Naser, Samy S. 2025. “Thyroid Cancer Risk Classification Using Machine Learning and Deep Learning Techniques: A Comparative Study with Balanced Dataset Augmentation.” : 223-235.
46. Qaoud, Alaa N, Abu-Naser, Samy S. 2025. “Deep Learning-Based Skin Cancer Classification and Localization: A Comprehensive Approach for Accurate Diagnosis and Localization of Skin Cancers.” : 169-181.
47. Massa, Nawal Maher, Abu-Naser, Samy S. 2025. “Predicting Breast Cancer Recurrence Using Machine Learning and Deep Learning Models: A Comparative Study.” : 183-196.