Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Enhancing diagnostic accuracy in bone fracture detection: A comparative study of customized and pre-trained deep learning models on X-ray images

Abdulmajeed Alsufyani *

Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

ARTICLE INFO

Article history:
Received 15 November 2024
Received in revised form
25 April 2025
Accepted 30 April 2025

Keywords:
Bone fractures
X-ray images
Deep learning
Fracture detection
Medical imaging

ABSTRACT

This study examines the performance of several deep learning models for detecting bone fractures from X-ray images. Traditional radiological methods depend on manual interpretation, which can lead to mistakes. Deep learning provides a useful alternative by automating the process of fracture detection. In this research, five models were tested: one custom Convolutional Neural Network (CNN) and four pre-trained models — AlexNet, DenseNet121, ResNet152, and EfficientNetB3. The models were trained on a dataset containing 10,581 X-ray images, which were labeled as either fractured or non-fractured. The models' performance was measured using accuracy, precision, recall, and F1-score. Among these models, EfficientNetB3 achieved the best results, with 99.20% accuracy and perfect recall, showing its high potential for use in clinical practice. ResNet152 and the custom CNN also performed well, although with slightly lower accuracy. The findings of this study emphasize the value of using advanced deep learning architectures for medical image analysis.

1. Introduction

Bones are often viewed as immobile structures that provide physical reinforcement. The process of bone remodeling continues throughout an individual's life, governed primarily by physiological requirements. According to Cowan et al. (2020), newborns typically have 270 bones, which fuse to form roughly 206 bones in adulthood. These include the skull bones, vertebrae, rib cage, and upper and lower extremities. Anatomical changes in specific bones lead to the variety in their number. An organism's skeletal structure comprises calcium-rich connective tissue and bone-specific cells. Fractures can be caused by pressure on a bone or by certain conditions (Mohanty and Senapati, 2023). The conventional approach to diagnosing fractures mostly depends on radiologists' ability to visually study X-ray images to identify and categorize fractures (Sharma, 2023). Every year, a substantial number of people suffer fractures, necessitating a prompt and correct diagnosis to avoid long-term injury or death.

X-rays are widely utilized in diagnosing bone fractures due to their rapidity, affordability, and user-friendliness, making them one of the primary tools in medical imaging. Medical imaging is essential for diagnosing and treating various medical disorders, such as fractures in orthopedics. AI has rapidly attracted more attention as a means of improving medical imaging interpretation and raising diagnosis accuracy. Accurate fracture detection is essential for determining the most appropriate treatment and forecasting the result. Fracture detection and classification have extensively utilized traditional machine learning methods for pre-processing, feature extraction, and classification. Aso-Escario et al. (2019) identified the delayed detection of spine fractures as a significant public health hazard. Moreover, if a fracture is misdiagnosed or undiagnosed, it can lead to nonunion, malunion, or additional harm to the surrounding tissues and the fractured bone.

Traditional fracture detection comprises radiologists visually examining X-rays to identify and categorize fractures. However, many factors can make X-ray interpretation difficult. However, according to Sharma (2023), this approach has the potential to be time-consuming, based on personal judgment, and susceptible to mistakes, especially with complex fracture patterns or slight anomalies. Radiologists are often exhausted and make mistakes due to excessive workloads and tight deadlines. Patients, doctors, and radiologists can be harmed by

* Corresponding Author.
Email Address: a.s.alsufyani@tu.edu.sa
https://doi.org/10.21833/ijaas.2025.05.008
Corresponding author's ORCID profile:
https://orcid.org/0000-0001-6110-3642

incorrect fracture diagnosis. The study by Taylor-Phillips and Stinton (2019) indicated that radiologists have poor focus, eye fatigue, and fracture detection. This shows how weariness affects diagnosis accuracy. Emergency misdiagnosis can increase without a second opinion.

Medical imaging applications show deep learning (DL) efficacy. CNNs and RNNs, deep learning models, can represent imaging data hierarchically and independently recognize fracture patterns. CNN-based deep learning algorithms excel at picture recognition. This makes them ideal for medical picture interpretation. Sharma (2023) suggested training these models on large datasets with annotations to improve performance and fracture detection sensitivity and specificity. This helps identify subtle patterns and features humans may overlook in X-ray pictures, improving fracture detection accuracy and efficiency. In this context, several deep learning models will be evaluated for the classification of X-ray images into fractured and non-fractured. The aim is to improve comprehension of the theme and come up with suggestions that future researchers can use.

The purpose of this research is to evaluate the performance of different deep learning models in detecting bone fractures from X-ray images. Given the limitations of traditional radiological methods, which rely on manual interpretation and are prone to errors, deep learning offers a promising alternative for automating fracture detection. Five models were evaluated: a custom Convolutional Neural Network (CNN) and four pre-trained architectures — AlexNet, DenseNet121, ResNet152, and EfficientNetB3.

2. Literature review

Recent advances in deep learning have improved medical picture classification. According to several studies, DL models outperform conventional methods in this discipline. Tanzi et al. (2020) studied DL for X-ray bone fracture classification. Researchers analyzed and evaluated deep-learning research that categorized bone fractures. Tanzi et al. (2020) found that deep learning, particularly CNNs, can now diagnose bone fractures like humans.

Yadav et al. (2022) conducted their investigation alternatively. Yadav et al. (2022) proposed SFNet which uses a combination of machine learning and deep learning approaches to accurately detect bone fractures. Yadav et al. (2022) sought to achieve precise detection of bone fractures to accurately identify and assess the severity of the fractures. To evaluate their performance, the researchers conducted a comparative analysis of various CNNs: AlexNet, VGG16, ResNeXt, and MobileNetV2. Their hybrid model, which integrated edge detection approaches, demonstrated improved classification accuracy and computing efficiency performance. The study emphasized the significance of combining many modes of features to improve the diagnostic precision of deep learning models. Contrary to Tanzi et al. (2020) and Yadav et al. (2022) placed greater emphasis on the integration of conventional machine-learning methods.

Yadav et al. (2022) proposed a hybrid SFNet DL model that assists doctors in obtaining a second opinion on diagnosing fractures and healthy bones. Nevertheless, some obstacles impede the advancement of DL in this domain. In their study, Su et al. (2023) highlighted the obstacles that hinder the consistent advancement and comparison of approaches in the field. These challenges encompass the lack of specific standards for identifying, categorizing, identifying, and specifying tasks. The objective of Su's et al. (2023) research was to tackle these concerns. A comprehensive analysis and evaluation of 40 articles from reputable databases such as WOS, Scopus, and EI. The authors examined alternative CNNs and evaluated their effectiveness in various fracture detection tasks. The authors emphasized the drawbacks of conventional two-stage detectors compared to more sophisticated models such as Faster R-CNN, which exhibited superior precision in detecting fractures in various anatomical locations. The researchers concluded that although deep learning algorithms perform similarly to clinicians, their clinical application still struggles to establish reliability.

Jones et al. (2020) created a deep learning system to identify fractures in various body parts, attaining exceptional accuracy and AUC ratings. Their approach successfully identified fractures in many clinical contexts, including emergencies. The study provided empirical evidence that DL systems can significantly reduce diagnostic errors and improve patient outcomes by promptly and accurately identifying fractures. In a further study conducted in 2022, Hardalaç et al. (2022) enhanced fracture detection methods by developing five distinct ensemble models. These models were combined to create a single detection model known as 'wrist fracture detection-combo. The findings demonstrated that DL models can attain elevated levels of sensitivity and specificity, rendering them appropriate for clinical application in fracture identification.

Deep learning algorithms are now favored for medical image categorization because they can automatically extract essential features from raw images. ResNet50, VGG16, and Inception are notable among the different designs. See Tables 1 and 2.

Substantial advances in deep learning have dramatically altered medical imaging categorization, including noteworthy developments in bone fracture identification. CNNs, together with advanced models like ResNets and EfficientNets, witness wider use because of their phenomenal ability to extract features from images. Regardless of the recent breakthroughs, multiple fundamental problems continue to affect medical imaging datasets and decrease model interpretation abilities, rendering them more computationally adept. The extensive use of deep learning in clinical procedures will yield effective outcomes when resolving existing issues.

Table 1: Comparison of models and techniques

Reference	Models evaluated	Best performing model	Key findings
Tanzi et al. (2020)	VGG16, CaffeNet, Network-in-Network	VGG16	VGG16 outperformed other models; a high dataset required
Yadav et al. (2022)	AlexNet, VGG16, ResNeXt, MobileNetV2	SFNet + Canny	The hybrid model showed the highest accuracy; edge detection improved the performance
Su et al. (2023)	Faster R-CNN, Inception-ResNet	Faster R-CNN	Faster R-CNN provided the highest accuracy across various tasks
Jones et al. (2020)	Custom DL system	Custom DL system	High performance across multiple anatomical regions
Hardalaç et al. (2022)	Faster R-CNN, ResNeXt101, FPN	Faster R-CNN	High sensitivity and specificity; suitable for clinical use

Table 2: Class distribution fractured and non-fractured images across the training, validation, and test sets

Type	Sample count			
	Training set	Validation set	Test set	Total
Fractured	4606	337	238	5181
Non-fractured	4640	492	268	5400
Total	9246	829	506	10581

Although previous research provides significant insights, the literature analysis highlights the substantial gap in developing and evaluating DL models for fracture detection. An important obstacle is the limited availability of extensive, varied, and thoroughly analyzed datasets that are appropriate for training resilient models. Existing studies frequently use small, diverse datasets that may not effectively represent clinical imaging quality and anatomical variations. Interpretability, regulatory approval, and clinician acceptance must also be addressed when integrating deep learning models into healthcare workflows. Model predictions must be reliable and interpretable to obtain the trust of healthcare professionals and regulatory agencies. Standardized model training, validation, and evaluation processes are needed to compare techniques and ensure study repeatability. This research develops and validates deep learning models utilizing big datasets of varied patient groups and clinical settings to close these gaps. The present models still encounter multiple restrictions. Su et al. (2023) conducted a review of 40 studies about detecting skeletal fractures while showing that research used conflicting metrics to evaluate models and different standards for dataset construction. The authors report that standardization benchmarks are non-uniform across studies, thus hampering research comparisons. Hardalaç et al. (2022) developed ensemble models targeting wrist fracture detection, which proved that combining multiple models could yield more accurate classification results. They designed a framework demanding an array of computational capabilities that could restrict its use in medical environments.

3. Methodology

This section outlines the methodology employed in our study to detect bone fractures using deep learning models. We begin by describing the data set used for training and testing our models, followed by a discussion of the deep learning architectures evaluated. Next, we detail the metrics used to assess model performance and the training and testing procedures adopted. Fig. 1 illustrates a brief overview of the proposed methodology to identify bone fractures. By detailing these elements, we give a complete overview of the phases utilized to ensure the models were optimized and rigorously evaluated for their ability to accurately detect bone fractures from medical imaging data. The approach ensures a thorough and reproducible evaluation of the deep learning models in this critical healthcare application.

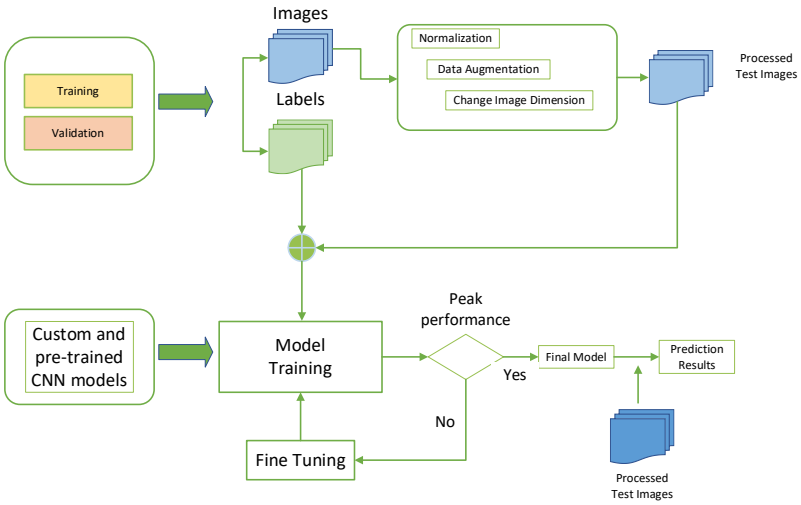

Fig. 1: Schematic representation of the proposed method for bone fracture detection (data preprocessing, model training, and validation)

3.1. Dataset description and preprocessing

The dataset that we used for training and testing the bone fracture detection models was sourced from the Kaggle repository (Rodrigo, 2024). It comprises X-ray images categorized into regions where fractures are present or absent. The dataset contains a variety of X-ray images from multiple bodily sections, with lower limbs, upper limbs, lumbar, hips, knees, etc., providing a comprehensive foundation for training deep learning models to identify fractures across different bones. The entire dataset is divided into three separate folders: training, testing, and validation. Each folder contains radiographic images categorized as either fractured or non-fractured. Fig. 2 showcases representative images from the dataset, featuring both fractured and non-fractured examples. In total, the dataset includes 10,581 X-ray images, divided into training (9246), validation (829), and testing (506) sets to evaluate model performance effectively. Tables 1 and 2 show the distribution of images belonging to fractured and non-fractured classes. The X-ray images varied in size, necessitating preprocessing steps to standardize the data prior to inputting it into the models. We applied standard image preprocessing methods, including resizing, normalization, and augmentation, to increase the superiority of the training data. Thus, preprocessing involved resizing the images to a unique dimension that is suitable for a specific deep learning model and normalizing (i.e., scaling) pixel values to fall within a specific range.

Fig. 2: Some sample fractured and non-fractured images from the dataset

Data normalization is an important preprocessing step before training a deep learning model. It ensures that all input features are on a similar scale. In deep learning, particularly when using gradient-based optimization methods like stochastic gradient descent (SGD), the performance and convergence of the model are strongly affected by the scale of the input data (Goodfellow, 2016). If the features are not scaled, variables with larger ranges could dominate the learning process, leading to slower convergence or a suboptimal model. Moreover, deep learning models often use activation functions like sigmoid or ReLU, which are sensitive to the magnitude of the input values (Nwankpa et al., 2018). For instance, if data is not scaled, gradients calculated for some features may be too large or too small, which may obstruct the model's capacity to identify optimal weights during training. Thus, scaling helps avoid issues like exploding or vanishing gradients, which occur when weights update inconsistently across layers. We used the technique division by maximum value where all pixel values are divided by 255 to normalize them to the range [0, 1]. These steps ensure that the model trains efficiently and generalizes better to unseen data. Additionally, we implemented data augmentation strategies to enhance the quality and quantity of our training data, which is particularly important for medical image classification problems. These techniques helped prevent overfitting and enhance model generalization. Table 3 outlines the augmentation methods used in this study. This preprocessing pipeline ensured that the models could generalize well across diverse X-ray images, improving their ability to accurately detect bone fractures. In summary, the following pre-processing steps were performed to ensure consistent input data quality and facilitate efficient training:

- Resizing: All X-ray images were resized to 224×224 pixels for compatibility with pre-trained models.
- Normalization: Pixel values were normalized to the range [0, 1] by dividing by 255, ensuring that input features were on a similar scale, thereby improving model convergence.
- Data Augmentation: To increase the diversity of training data and prevent overfitting, we applied augmentation techniques such as rotation (±40°), width and height shifts (±20%), zooming (±20%), shearing (±20%), and horizontal flipping.

3.2. Deep learning models

This section provides a summary of the deep-learning models that we assessed for bone fracture detection. Various architectures, ranging from traditional convolutional neural networks (CNNs) to

more advanced models like ResNet152 and EfficientNetB3, were selected due to their demonstrated success in medical image analysis tasks. Every single model brings unique strengths, such as deeper layers, efficient feature extraction, or reduced parameter complexity, making them suitable for handling the complexities of medical imaging. Additionally, specific modifications and customizations were applied to tailor these models for our dataset and improve their performance in detecting fractures. Below, we provide a brief description of each model used in our study.

Table 3: Image augmentation settings

Method	Amount/value
Width shift	0.2
Height shift	0.2
Rotation range	40
Shear range	0.2
Zoom range	0.2
Horizontal flip	true
Fill mode	"nearest"

3.2.1. Custom CNN model

Fig. 3 demonstrates the suggested Convolutional Neural Network (CNN) architecture designed to detect bone fractures, which is composed of multiple essential layers that systematically extract and enhance features from input X-ray images measuring 128x128 pixels. The architecture of the network initiates with three convolutional layers, each employing Rectified Linear Unit (ReLU) activation functions, which facilitate the incorporation of non-linearity, thereby enabling the network to acquire more intricate forms of information. The first convolutional layer applies 32 filters, each measuring 3x3 pixels, to the input image. The second layer uses 64 filters of size 3x3 on the output of the previous layer. It captures more complex features at a finer scale. The third layer employs 128 filters, each measuring 3x3 pixels. As we go deeper into the network, we increase the number of filters to capture even more abstract and high-level features. Following each convolutional operation, max pooling is employed to reduce the spatial dimensions and computational load while preserving important information. Following the feature extraction layers, a flattening layer transforms the two-dimensional feature maps into a one-dimensional vector. This vector is subsequently fed into a fully connected dense layer containing 128 neurons to acquire complex features at a higher level. During training, a dropout layer is used to reduce overfitting by disabling half of the neurons randomly. The final layer of the model consists of two neurons and uses the softmax activation function to determine the likelihood of a fracture or non-fracture.

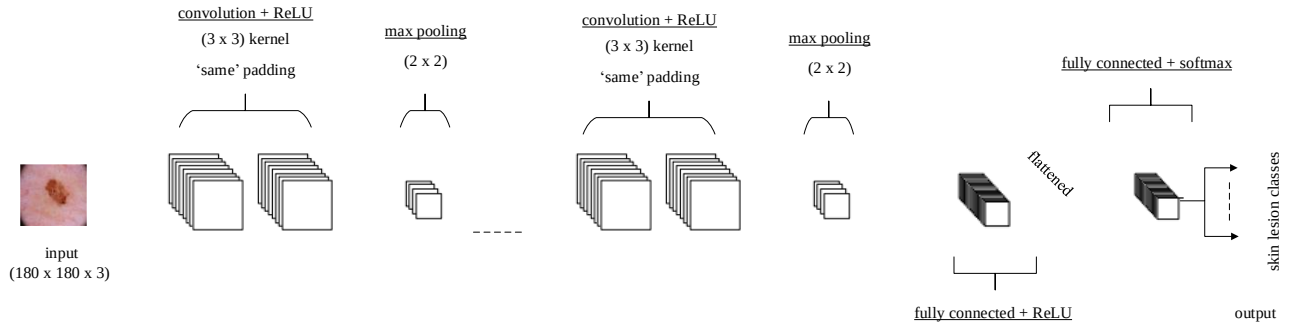

Fig. 3: Proposed CNN architecture for bone fracture detection

3.2.2. Pre-trained CNN models

The AlexNet (Krizhevsky et al., 2017) model, implemented for this study, follows a deep learning architecture constructed for image classification. It works with images that are 227x227 pixels and have three color channels (red, green, and blue). The network starts with a wide receptive field in its initial convolutional layer, using 96 filters of size 11x11 with a stride of 4. This is followed by batch normalization and max pooling, which help decrease the spatial dimensions and reduce the computational complexity. The second convolutional layer uses 256 filters of size 5x5 pixels to detect larger-scale patterns in the image. After this layer, batch normalization and max pooling are applied to stabilize the training process and reduce the size of the feature maps. The third, fourth, and fifth convolutional layers use 384, 384, and 256 filters, respectively, with a size of 3x3 pixels. These layers continue to extract features from the image, focusing on finer details as we progress through the network. Max pooling is applied after the fifth layer to reduce the feature map size and prepare the output for the subsequent layers.

After the convolutional layers extract features from the image, these features are flattened into a one-dimensional vector. This reshaping allows the extracted information to be processed by the fully connected layers. These layers take the flattened features and analyze them to make a final prediction about the content of the image. Each layer has 4096 neurons and ReLU activation, which means there are 4096 interconnected nodes in each layer. These layers help the network learn more complex and abstract features from the input data. To avoid overfitting, dropout is applied after each dense layer, where 50% of the neurons are randomly deactivated during training. This technique forces the network to rely on multiple neurons for learning, improving generalization by reducing the model's dependence on any single set of features. The final output layer

employs a softmax activation function, which converts the network's outputs into probabilities for each class. In this case, it predicts the likelihood that an input belongs to one of two categories—fractured or non-fractured. This architecture, known for its success in large-scale image classification, is well-suited to handle the complex task of detecting bone fractures in medical images.

The second pre-trained model we used is called DenseNet121 (Huang et al., 2017). It's a type of deep learning model that uses a special architecture known as Dense Convolutional Network (DenseNet). It is known for its efficient feature reuse and compact structure, where each layer collects inputs from all prior layers. This densely connected pattern helps in reducing the number of parameters and promoting feature propagation, which is particularly useful in medical imaging tasks that require extracting delicate features. DenseNet121 is composed of several dense blocks, each made up of multiple convolutional layers. Within these blocks, every layer is connected to all preceding layers, allowing for improved feature propagation. After each dense block, there is a transition layer that utilizes pooling to reduce the spatial dimensions of the feature maps. DenseNet121 has been pre-trained on the ImageNet dataset, enabling it to learn robust, general-purpose features that can be adapted for specialized tasks such as bone fracture detection.

In this study, DenseNet121 is specifically fine-tuned to enhance its performance in detecting bone fractures. The pre-trained DenseNet121 model we used was originally trained on the ImageNet dataset. When we loaded this model, we excluded the top layer, which is responsible for making final predictions. The input images are resized to 224x224 pixels to match the input requirements of DenseNet121. After the pre-trained DenseNet121 model processes the input image, the output is a set of feature maps. These feature maps represent different aspects of the image. To reduce the dimensionality of these feature maps and convert them into a single vector, a technique called global average pooling is used. After the global average pooling, a fully connected layer with 512 neurons is added to the network. This layer combines the features extracted from the image into a higher-level representation.

To prevent the model from overfitting, a dropout layer is used. This layer randomly deactivates 50% of the neurons during training. This forces the remaining neurons to learn to perform their tasks without relying too heavily on the deactivated neurons, making the model more adaptable and less likely to memorize the training data. Finally, a dense layer containing a single neuron is implemented with a sigmoid activation function to facilitate binary classification.

This layer predicts whether the input X-ray image indicates the presence of a fracture or not, producing an output that represents the probability of fracture occurrence. To optimize the model's performance for this task, the convolutional layers of the DenseNet121 base model are kept fixed, preventing any updates to their weights during the training process. This approach helps preserve the valuable features learned from the initial training, allowing the model to concentrate on refining the subsequent layers specifically for the fracture detection task. This approach allows only the newly added layers to be trained on the bone fracture dataset. It takes advantage of DenseNet121's robust feature extraction abilities while specifically adapting the model to excel at the current task.

Our third pre-trained model is ResNet-152 (He et al., 2016), which is a very deep convolutional neural network (CNN) architecture based on the concept of residual learning. ResNet-152 comprises 152 layers and was developed to address the vanishing gradient issue commonly encountered in deep networks. Utilizing residual connections enables the training of much deeper models while maintaining effective learning, improving performance without suffering from the degradation problems typically associated with very deep architectures. The key feature of ResNet (Residual Networks) is the use of residual blocks, where the input to a layer is added to the output of a few stacked layers (skip connections). Skip connections help gradients move smoothly through the network while training, which makes the process more efficient and allows for improved performance in deeper neural networks.

For the bone fracture detection task, ResNet152 is fine-tuned after loading with its pre-trained weights. The base model is used without its top classification layer, so the network can be adapted to the specific requirements of our binary classification problem. The input images are resized to 224x224 pixels to match the input size expected by ResNet152. Once features are extracted from the base ResNet-152 model, global average pooling is employed to shrink the dimensionality of the feature maps and flatten them into a vector. This vector flows through a densely linked layer featuring 1024 neurons, employing the ReLU activation function, which empowers the model to grasp complex patterns within the dataset. To prevent the model from overfitting, a dropout layer is used. This layer randomly deactivates 50% of the neurons throughout the training process. The final layer constitutes a densely connected layer featuring a single neuron that employs a sigmoid activation function, thereby generating the likelihood of a fracture as depicted in the specified X-ray image.

In this approach, unlike partial fine-tuning, all layers of the ResNet152 base model are unfrozen, allowing the entire network to be retrained. This enables the model to update the weights of both the pre-trained ResNet layers and the layers added on top to specialize in bone fracture detection. This full fine-tuning permits the model to obtain subtle, domain-specific patterns present in the medical X-ray images while benefiting from the powerful feature extraction capabilities learned from ImageNet.

The final pre-trained model that we used is EfficientNetB3 (Tan and Le, 2019), part of the EfficientNet family, which scales a model's depth, width, and resolution equally through a compound scaling method. This architecture is designed to achieve better accuracy with fewer parameters compared to traditional models, making it highly efficient for both computational cost and performance. EfficientNetB3 uses an input image size of 300x300 pixels and has fewer parameters than other networks while delivering strong results on image classification tasks. It is pre-trained on the ImageNet dataset, allowing it to leverage the rich, general-purpose features learned during large-scale image classification.

This study involves fine-tuning the EfficientNetB3 model specifically for identifying bone fractures. The base model, pre-trained on ImageNet, is loaded without its top classification layers, making it adaptable for the binary classification of fractured versus non-fractured bones. Input images are resized to 300x300 pixels to align with the input size required by EfficientNetB3. To fine-tune the model for this specific task, the last 20 layers of the EfficientNetB3 model are unfrozen, allowing these layers to be trained on the bone fracture dataset. By selectively unfreezing certain layers, the model can adapt to task-specific patterns while preserving the strong feature extraction abilities gained from pre-training on ImageNet in the earlier layers.

After the base model's results are generated, global average pooling is utilized to convert the feature maps into a singular vector. This vector is then sent through a fully connected layer consisting of 1024 neurons that employ ReLU activation. This layer assists the model in understanding complex patterns within the data. Ultimately, a dense layer consists of a single neuron that utilizes sigmoid activation to determine the likelihood of a fracture in the X-ray image and categorize it as either fractured or non-fractured. By freezing the majority of the EfficientNetB3 model and only retraining the top layers, we ensure that the model is computationally efficient while effectively learning the specific features necessary for bone fracture detection.

3.2.3. Summary of model configurations

In summary, we evaluated a range of deep learning architectures, including a custom CNN and pre-trained models such as EfficientNetB3, ResNet152, AlexNet, and DenseNet121. Specific configurations include:

- Input Shape: We have used an input image size of 224×224 for pre-trained models and 128×128 for custom CNN.
- Activation Functions: ReLU was used in all intermediate layers due to its computational simplicity and efficiency in mitigating the vanishing gradient problem. ReLU is particularly advantageous as it introduces non-linearity without saturating, thereby allowing faster convergence. The final dense layer used a sigmoid activation function, as the task involved binary classification. The sigmoid maps the output to a probability range [0, 1], making it suitable for predicting the presence or absence of fractures.
- Pooling Mechanisms: Global average pooling (GAP) was utilized in pre-trained models, while max-pooling was employed in the custom CNN. GAP computes the average feature map values for each channel, significantly reducing the spatial dimensions while preserving channel-specific information. It also prevents overfitting by reducing the number of parameters compared to fully connected layers. On the other hand, max pooling was applied after each convolutional block to down-sample feature maps. Max pooling selects the maximum value within a defined kernel (e.g., 2×2), enabling the extraction of the most prominent features while reducing spatial dimensions.
- Training Parameters: All models were trained for 20 epochs with a batch size of 32 using the binary cross-entropy loss function.

3.3. Evaluation metrics

Deep learning models were assessed for their ability to detect bone fractures using various metrics like accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified images (fractured or non-fractured) in relation to the overall number of images analyzed. It offers a comprehensive overview of the model's efficiency across the complete dataset. Precision is the measure of accurately predicted fracture cases relative to all images categorized as fractures. This is particularly important when false positives (non-fractures classified as fractures) are costly, such as in medical diagnosis. Recall, also known as sensitivity, pertains to the rate at which the model correctly identifies actual fractures.

High recall is critical in medical tasks to minimize missed fractures (false negatives), ensuring that patients with fractures are correctly diagnosed. The F1-Score is a metric that integrates both precision and recall, particularly beneficial when dealing with imbalanced datasets where there are fewer instances of fractures compared to non-fractured cases. It provides a balance between precision and recall.

For bone fracture detection, these metrics are crucial. Accuracy alone might not be enough, as it could be misleading in an imbalanced dataset where non-fractured cases dominate. Precision and recall become more important, as precision ensures fewer false positives, and recall ensures that most actual fractures are detected. The F1-score is useful for evaluating models in this context because it finds a middle ground between accuracy and completeness. These metrics together ensure that the model is reliable and minimizes diagnostic errors, crucial for patient care.

3.4. Training and testing

The deep learning models were trained to detect bone fractures using TensorFlow and the Keras functional API, with free GPU resources provided by Google Colab for computational efficiency. We trained the models for a duration of 20 epochs utilizing a batch size of 32 to ensure a balanced trade-off between computational load and gradient updates. A summary of our model configurations, including hyperparameters, can be found in Table 4. Key hyperparameters and training configurations are as follows:

- Optimizer: The study utilized the Adam optimizer set at a learning rate of 0.0001. Adam is a popular choice due to its adaptive learning rate and capability to manage sparse gradients often encountered in image data.
- Loss Function: We chose the binary_crossentropy loss function because our classification task involves distinguishing between fractures and non-fractures, which are binary outcomes. This loss function is well-suited for models predicting probabilities in binary classification problems.
- Evaluation Metric Used for Model Training: The primary metric tracked during training was accurate, which provides insight into how well the model is learning to classify fractured and non-fractured X-ray images.

Table 4: Listing of model settings and hyperparameters

Parameters	Value
Batch size	32
Epochs	20
Optimizer	Adam
Learning rate	0.0001
Loss function	Binary cross entropy
Pooling	Max-pooling (custom CNN) Global Average (pre-trained CNNs)
Activation	sigmoid (last dense layer)
Pre-trained CNN weights	ImageNet

The hyperparameter tuning process involved adjusting critical parameters to achieve optimal model performance. The following hyperparameters were tuned during the training process:

- Batch Size: We experimented with batch sizes of 16, 32, and 64, ultimately selecting 32 for a balance between convergence speed and memory efficiency.
- Learning Rate: The learning rate varied in the range of 1×10^{-5} to 1×10^{-3} using a grid search. The optimal learning rate of 1×10^{-4} was selected based on validation performance.
- Dropout Rate: A dropout rate of 0.5 was chosen to prevent overfitting, based on experiments with values ranging from 0.3 to 0.7.

To validate the model during training, a portion of the dataset was set aside as the validation set as mentioned in Tables 1 and 2. After every epoch, the model's effectiveness was assessed on this dataset to check for overfitting and evaluate how well it would perform on new data.

After the models were trained, they were tested on a separate set of X-ray images. These images were not used at all during the training or validation phases. To ensure a comprehensive evaluation of the model's ability to detect bone fractures, they were assessed using multiple performance metrics. These metrics included accuracy, precision, recall, and F1-score. To calculate the performance metrics, the model's predictions for the test set were compared to the actual labels of those images.

By testing the models on a separate test set, we were able to evaluate how well they generalized to new, unseen data. This is important because it ensures that the models are practical and can be used in real-world medical settings. By using a variety of metrics, we were able to comprehensively evaluate the model's ability to accurately identify fractures and get around misclassifications.

4. Results and discussions

This section of the paper presents the results of the deep-learning models used to detect bone fractures. It also provides a detailed analysis of these results. In total, five models were evaluated: one custom Convolutional Neural Network (CNN) and four pre-trained models (AlexNet, DenseNet121, ResNet152, and EfficientNetB3), all fine-tuned for this specific task. The performance metrics that we used to assess the models offer insights into the models' capacity to accurately identify fractures, minimize false positives, and generalize effectively. We aim to provide a detailed comparison of the strengths and weaknesses of each model and discuss their effectiveness in detecting bone fractures from X-ray images. Subsequently, we also interpret the reasons behind the varying performances of the models, considering factors such as architecture complexity and training efficiency. In addition to discussing the technical performance of the models, we will also explore how these findings can be applied in a real-world medical setting. This includes evaluating the potential of these models to assist in medical diagnosis. Finally, we will address the weaknesses of this study and propose avenues for further study to enhance the application of deep learning in fracture detection.

4.1. Performance analysis

As shown in Table 5, EfficientNetB3 outperformed the other models by a significant margin, with an accuracy of 99.20%. EfficientNetB3's excellence is due to its revolutionary compound scaling strategy, which balances network depth, width, and resolution, optimizing performance while retaining computational efficiency. The model's exceptional performance, achieving near-perfect precision and recall for both fractured and non-fractured cases, demonstrates its competence in dealing with the inherent difficulties of medical

imaging data. Its efficient scaling method allows for better performance with fewer parameters, making it highly accurate and resource-efficient. The confusion matrix for EfficientNetB3, as shown in Table 6, demonstrates few misclassifications, supporting its resilience and dependability for clinical applications. The absence of false negatives (FN = 0) indicates that the EfficientNetB3 model successfully detected every fracture in the dataset. This is crucial in medical diagnosis since missing a fracture (FN) can lead to severe clinical consequences. With only 4 false positives, the model shows high precision, meaning it rarely misclassifies healthy cases as fractures. This reduces the likelihood of unnecessary interventions or further diagnostics. Despite its high performance, EfficientNetB3's recall of 100% suggests that it may slightly be overfitted with the training data, though its high F1-score mitigates this concern.

Table 5: Performance results obtained from various models for the detection of bone fracture using test dataset

Model	Accuracy	Precision	Recall	F1-score
CNN	98.02%	97.60%	98.36%	97.98%
AlexNet	89.72%	86.78%	95.15%	90.73%
DenseNet-121	91.10%	92.43%	90.67%	91.54%
ResNet-152	98.22%	98.50%	98.10%	98.30%
EfficientNetB3	99.20%	98.53%	100%	99.26%

Table 6: Confusion matrix for the studied models using test dataset having non-fractured and fractured samples

Model used	TN	TP	FP	FN
CNN	252	243	5	4
AlexNet	199	255	39	13
DenseNet-121	218	243	20	25
ResNet-152	234	263	4	5
EfficientNetB3	234	268	4	0

ResNet-152 and CNN also performed well, with accuracies of 98.22% and 98.02%, respectively. ResNet-152's architecture, which includes a deep residual learning framework, efficiently addresses the vanishing gradient problem, allowing the model to retain and transfer knowledge across its layers. This architectural advantage most certainly contributed to its excellent accuracy, making it a viable candidate for applications requiring detailed feature extraction, such as bone fracture identification. The low number of false positives (FP) shows the model's ability to avoid unnecessary diagnoses of fractures in healthy individuals. The presence of 5 false negatives (FN) indicates that the model missed detecting some fractures, which slightly impacts the recall. This can be critical in medical applications since undetected fractures may lead to improper patient treatment. The low FN count still indicates strong overall detection performance. However, the model is computationally expensive and may require more resources and time to train, which could limit its accessibility for some applications. Similarly, the CNN model's simple architecture, which had three convolutional layers followed by dense layers, was beneficial in producing competitive results. The low number of false positives (FP) and false negatives (FN) demonstrates that the custom CNN has strong generalization capabilities. In particular, the high precision shows the model's ability to minimize unnecessary fracture diagnoses, while high recall indicates its ability to detect most fractures accurately. While the model performs well, it is slightly outperformed by more complex models like ResNet-152 and EfficientNetB3, indicating that further optimization could potentially improve performance.

DenseNet121, despite its reputation for efficient feature reuse and high generalization capabilities, attained an accuracy of 91.10%. While this model performed consistently, as seen by balanced precision, recall, and F1 scores, it did not achieve the accuracy of ResNet-152 or EfficientNetB3. The marginally inferior performance of DenseNet121 could be attributed to its complicated connectivity design, which, while useful in some situations, may not have provided a major advantage in this application. The relatively high number of false negatives (FN) and false positives (FP) indicates that the model struggled more with distinguishing between fractured and non-fractured cases compared to other models, leading to lower precision and recall. The relatively high number of false classifications (both FP and FN) suggests that the model may be overfitting, especially given the complex nature of DenseNet-121. Overfitting could prevent the model from generalizing well to unseen data.

AlexNet, with an accuracy of 89.72%, performed the least effectively of the models examined. Although AlexNet was innovative when it was first introduced, its relatively shallow architecture and limited depth and capacity make it less capable of catching the subtle patterns found in medical imaging datasets than more contemporary systems. The decreased recall for the fractured class in AlexNet suggests a higher likelihood of false negatives, which is especially concerning in clinical settings where missing a fracture could have catastrophic consequences. Furthermore, a high false positive rate (FP=39) means that many non-fractured cases were incorrectly classified as fractures. This could lead to unnecessary medical interventions, such as further diagnostic imaging or unnecessary treatment. This impacts the precision of the model negatively, as it tends to over-predict fractures.

When comparing these models, EfficientNetB3's excellent accuracy, together with its balanced performance across other evaluation parameters, makes it the best model for the task of detecting bone fractures. Its compound scaling technique is very advantageous, allowing the model to scale adequately without disproportionately increasing computational costs, resulting in not just accuracy but also efficiency.

ResNet-152, while somewhat less accurate than EfficientNetB3, is still a highly competitive option. Its use of residual connections enables the creation of deeper networks while avoiding the degradation

problem, making it an excellent choice for applications needing in-depth feature extraction.

CNN's performance, which closely matches that of ResNet-152, demonstrates that simpler designs can still produce great results, especially when combined with appropriate training procedures and sufficient data. However, for clinical applications where even minor improvements in accuracy are crucial, the sophisticated topologies of ResNet-152 and EfficientNetB3 are preferred.

DenseNet121's performance, while respectable, suggests that its dense connection pattern, which is intended to increase gradient flow and feature reuse, may not provide significant advantages over other models in this application. Finally, AlexNet's performance highlights the importance of deeper and more sophisticated models in modern medical imaging tasks, where the capacity to collect fine-grained information is critical.

EfficientNetB3 is the best model for the Bone Fracture Multi-Region X-ray Dataset, with the highest accuracy and most balanced metrics. Its architectural advances allow it to excel at this complicated task while being computationally efficient. Given the importance of precise diagnosis in clinical settings, EfficientNetB3's performance makes it the best option for bone fracture detection. However, ResNet-152 and CNN are also viable options, especially in cases where computational resources are limited. The study's findings highlight the necessity of using advanced and well-optimized models for medical imaging tasks, where even little increases in accuracy might have huge therapeutic effects.

To evaluate the statistical significance of the performance differences between the best-performing model, EfficientNetB3, and other CNN models, we employed a paired t-test. This method compares the paired performance metrics by calculating the t-statistic (Goodfellow, 2016) and testing it against a student's t-distribution with a specified degree of freedom. If the resulting *p*-value is below the commonly accepted significance threshold of 5%, the null hypothesis—stating that there is no significant difference between the model performances—is rejected, confirming a significant difference. Conversely, a *p*-value above the threshold indicates that the null hypothesis cannot be rejected, suggesting similar performance between the models. This statistical analysis was applied to compare the metrics of EfficientNetB3 with each CNN model, as presented in Table 7. P-values below the 5% threshold are highlighted, indicating statistically significant differences. The results provide strong evidence that the performance improvements of EfficientNetB3, particularly in terms of metrics like accuracy, precision, and recall, are statistically significant compared to other custom CNN and transfer learning models.

Table 7: Results of the paired t-test (t-statistic and p-value) comparing the performance of the best ensemble model with other base CNN models

Models	Accuracy	Precision	Recall	F1-score
CNN	(-6.53, 5.32e-05)	(-5.77, 0.0001)	(-7.73, 1.45e-05)	(-5.29, 0.0002)
AlexNet	(-36.6, 2.07e-11)	(-38.16, 1.44e-11)	(-14.86, 6.10e-08)	(-22.84, 1.84e-09)
DenseNet-121	(-68.27, 7.83e-14)	(-21.64, 2.25e-09)	(-28.79, 1.78e-10)	(-1.89, 0.045)
ResNet-152	(-7.87, 1.24e-05)	(-0.81, 0.21)	(-5.84, 0.0001)	(-3.04, 0.006)

Fig. 4 depicts the training and validation curves for accuracy and loss using the custom CNN model, and the best-performing pre-trained model EfficientNetB3 with fine-tuning. Furthermore, to provide a comprehensive comparison of our proposed EfficientNetB3 model's performance for bone fracture detection with existing works in the literature, we present results from several studies that have employed different deep learning models. These works focus on detecting fractures using X-ray images, with performance metrics such as accuracy, precision, recall, and F1-score as key indicators. Table 7 summarizes the results from existing literature compared with our study.

Our EfficientNetB3 model outperformed other models in the study, achieving an impressive accuracy of 99.20%. This highlights the effectiveness of the EfficientNet architecture and its ability to accurately detect bone fractures in X-ray images. The precision of 98.53% and perfect recall of 100% indicate that the model excels at correctly identifying fractures while minimizing false positives and false negatives. When comparing our results to those of existing studies, our model consistently outperforms them across all metrics. For example, Gan et al. (2019) used an InceptionV4 model and achieved an accuracy of 93.0%, which is 6.2% lower than our EfficientNetB3 model. Similarly, Nishiyama et al. (2021) applied a custom CNN model and reported an accuracy of 84.5%, demonstrating a noticeable gap in performance. The YOLO-based deep learning model, evaluated by Son et al. (2021), shows the weakest recall value in comparison, with a precision of 97.5%, highlighting the limitations of the models used by the authors compared to modern fine-tuned models like EfficientNetB3. Furthermore, while U-Net and ResNet-50 from Wang et al. (2022) provided relatively competitive results with 96.40% accuracy and 97.60% recall, our proposed model still demonstrated a significant improvement. The comparison of our model with existing methods confirms its cutting-edge performance in bone fracture detection. This suggests that our model has the potential to be a valuable tool in clinical settings, assisting healthcare professionals in accurately diagnosing fractures. By providing more accurate and reliable predictions, it has the potential to enhance diagnostic workflows, reduce human errors, and improve patient care.

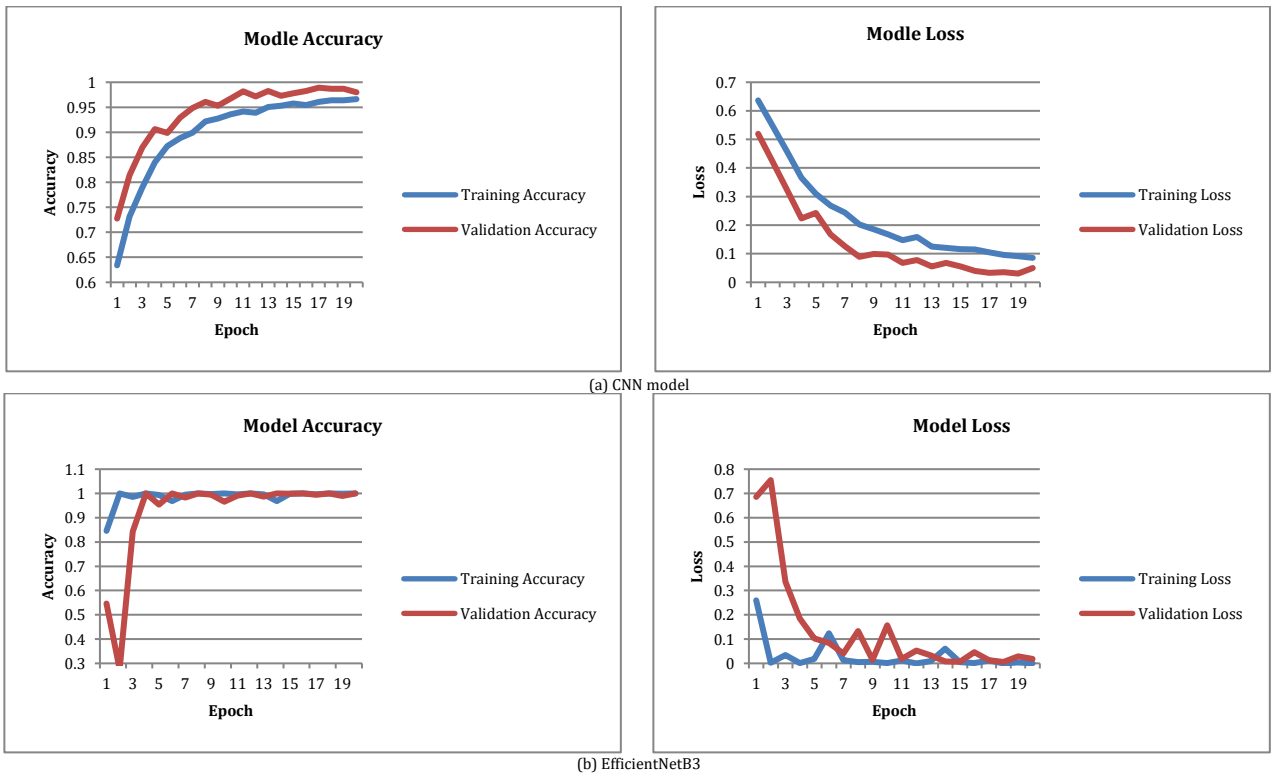

Fig. 4: Training and validation curves for accuracy and loss using: (a) the custom CNN model; (b) the best performing pre-trained model EfficientNetB3 with fine-tuning

4.2. Discussions

4.2.1. Interpretation of results

The performance of the deep learning models varied significantly, demonstrating the importance of model selection and architecture for bone fracture detection. The custom CNN, while showing promising results, was outperformed by some of the pre-trained models, suggesting that leveraging pre-trained knowledge from large-scale datasets can be beneficial.

The results indicate that model architecture, complexity, and optimization strategies significantly influenced performance. EfficientNetB3 outperformed the other models due to its scalable architecture, which balances depth, width, and resolution, allowing it to extract detailed and complex features from X-ray images. Its superior performance, particularly in recall (100%) and F1-score (99.26%), highlights its ability to capture subtle patterns in the dataset without overfitting. Its ability to perfectly detect all fractures with very few false positives demonstrates its potential for real-world deployment in healthcare settings. ResNet-152 also performed exceptionally well due to its deep residual architecture, which mitigates the vanishing gradient problem by allowing information to flow across layers via skip connections. Its higher layer depth enables better feature extraction, which is why it outperformed simpler models like the custom CNN and AlexNet. However, the presence of a few false negatives shows there is room for improvement in recall. Despite these few errors, the model's effectiveness in minimizing false positives and its overall robust performance makes it a valuable tool for clinical use, though further refinements could enhance its fracture detection capabilities.

The custom CNN showed competitive performance, especially in terms of precision (97.60%) and recall (98.36%), likely due to its design being tailored specifically to the dataset. Achieving such performance with only three convolutional layers highlights the strength of well-designed, relatively simple CNN architectures for specific tasks like bone fracture detection. This result underlines the model's robustness and practicality, making it an effective tool for fracture detection even in resource-constrained environments. However, it lacked the deeper architecture of models like ResNet-152, limiting its ability to capture more complex features. AlexNet struggled relative to the others, likely because its architecture was designed for a broader image classification task and lacked the depth and refinement of more modern architectures. Despite decent recall (95.15%), its lower precision (86.78%) indicates that it is prone to false positives, which may arise from an inability to fully capture the details of fracture patterns in medical imaging data. Given that AlexNet is a simpler model, it may not be well-suited to the complexity of bone fracture detection. It might have benefitted from more advanced architectures like ResNet or EfficientNet, which tend to capture more intricate features in medical images. DenseNet-121 achieved respectable performance, leveraging its dense connectivity to enhance feature reuse. However, its moderate complexity led to marginally lower scores compared

to ResNet-152 and EfficientNetB3, as it may not have been able to capture the same level of detail in the X-ray images.

The compound scaling approach of EfficientNet has recently received interest because it achieves balanced accuracy and model size. The authors of Wang et al. (2022) combined U-Net with ResNet-50 for CT image fracture detection, which proved effective for both sensitivity and specificity. This study demonstrates that successful medical use of the framework is reliant on fine-tuning procedures shaping pre-trained networks. The authors stated that deep learning models with high performance need explanation techniques to establish credibility among medical professionals.

4.2.2. Implications for clinical practice

The results of this investigation have various implications for clinical practice and future research. The outstanding performance of **EfficientNetB3** and ResNet-152 suggests that deep architectures with optimized parameter scaling are well-suited for bone fracture detection. This has significant implications for clinical practice, as accurate and reliable models can reduce diagnostic errors, enabling faster and more accurate treatment decisions. The high recall achieved by EfficientNetB3 implies that it could be particularly useful in clinical settings where missing fractures (false negatives) can have serious consequences.

The strong performance of X-ray image analysis by CNN-based models like ResNet and EfficientNet presents several obstacles when it comes to model generalization across different datasets. Studies commonly face limitations because their datasets contain a restricted number of samples while being unbalanced, which produces biased classification outcomes. The resolution of these problems depends on acquiring larger and more diverse datasets that require enhancement through preprocessing methods. The process of integrating AI diagnosis equipment into healthcare practice demands regulatory validation and confirmation by conducting research at multiple facilities.

This research reinforces the potential of deep learning models in the effective diagnosis of fractures characterized by minimal errors and quicker workflow in radiological processes. The EfficientNetB3 model achieved 99.20% accuracy in medical imaging tasks because of its compound model scaling method. The high recall value of 100% indicates the model's ability to identify all fractures while ensuring no cases of missed diagnoses in emergency medicine contexts. Additional improvements need to take place to enhance precision because the current rate of false positives leads to unnecessary follow-up examinations.

AI-based diagnostic tools need to integrate smoothly with medical imaging systems that currently exist for clinical practice deployment. Deep learning models function as decision-support tools that radiologists can use to obtain additional opinions when reading scans during busy periods of operation. Jones et al.'s (2020) study reveals that using AI systems enhances fracture detection proficiency, adding to the radiologists' professional expertise and empowering them to handle complex medical interpretations.

For future research, these findings highlight the importance of leveraging advanced architectures that can effectively scale based on available data and hardware. The effectiveness of transfer learning demonstrated in this research also suggests that pre-trained models, fine-tuned on medical datasets, could reduce the need for large annotated medical datasets, which are often difficult to obtain.

The choice of model should be carefully considered based on factors such as accuracy, computational efficiency, and interpretability. The effectiveness of deep learning models greatly depends on the quality and volume of the training data available. Efforts should be made to collect and annotate large and diverse datasets. Creating deep learning models that can be easily interpreted is crucial for establishing credibility and comprehending the reasoning behind the models' predictions. Future studies need to concentrate on creating methods that clarify the reasoning behind the choices made by these models.

4.2.3. Limitations

There are several limitations in this study. The limited size of the dataset used in this study may restrict the model's ability to accurately predict bone fractures in a broader population. A larger dataset would likely improve the model's generalizability and enhance its clinical applicability. Although data augmentation was employed to artificially increase the dataset size, this might not accurately represent the full spectrum of data observed in the real world. Additionally, models such as AlexNet and DenseNet-121 may have underperformed due to their lack of architectural complexity or misalignment with the specific task of fracture detection. Lastly, the study may be affected by class imbalance, where the distribution of fractured vs. non-fractured images could introduce biases that affect model performance.

The AI-supported system used for diagnosing fractures faces various obstacles that need to be mitigated for worldwide adoption for clinical purposes. The problem of insufficient data variety poses a major challenge. This study used images from a single repository as its dataset, although the dataset fails to demonstrate a complete representation of actual clinical images. The future of research requires multiple medical facility image collections that integrate images from various medical equipment to strengthen model performance.

Deep learning prediction systems face difficulties because their results are not easily understandable to medical personnel. Medical professionals need explainable AI tools to understand the decision

processes of their models. The use of Grad-CAM and attention map methods should be implemented to display which X-ray image regions most influence the classification determination. AI-dependent fracture diagnosis techniques garner increased acceptability from healthcare professionals due to their augmented interpretation abilities.

The outstanding performance of EfficientNetB3 is difficult to use in real-time applications because of its high computational requirements. Increased optimization of edge computing models or the creation of lighter models like MobileNet-based frameworks will enhance their deployment practicality in clinical settings with limited resources.

4.2.4. Future work

Our goal is to overcome some of these limitations by using larger and more varied datasets to enhance the models' ability to generalize. Furthermore, exploring ensemble techniques, which combine predictions from multiple models, could enhance performance further by leveraging the strengths of various architectures. We also plan to explore the interpretability of these models, particularly in clinical settings. Explainable AI techniques could be applied to highlight which areas of the X-ray images are most indicative of fractures, offering insights to radiologists and improving trust in the AI-driven decision-making process. Combining X-ray images with other modalities, such as clinical data or patient history, can potentially improve the accuracy of bone fracture detection. Finally, optimizing deep learning models for deployment in real-world clinical environments remains an important area of future work. This includes focusing on resource efficiency, reducing inference time, and ensuring compatibility with clinical workflows.

5. Conclusions

In this study, we evaluated several deep-learning models for detecting bone fractures in X-ray images, including a custom CNN model and four pre-trained models (AlexNet, DenseNet121, ResNet152, and EfficientNetB3). Our analysis revealed that EfficientNetB3 outperformed the other models by a significant margin, achieving an accuracy of 99.20%. This demonstrates its superior ability to generalize to the fracture detection problem, making it a promising candidate for real-world clinical applications. The custom CNN model also exhibited strong performance, while AlexNet and DenseNet121 showed comparatively lower results, likely due to differences in model architecture and complexity.

These findings emphasize the effectiveness of deep learning, particularly fine-tuned pre-trained models, in medical imaging tasks such as fracture detection. The results hold potential for improving clinical practice by enabling more accurate and timely diagnoses, ultimately aiding radiologists and healthcare professionals in their decision-making processes. In summary, our study highlights the significance of using advanced deep-learning models for fracture detection and paves the way for future research focused on refining these methods and exploring their broader clinical applications.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Aso-Escario J, Sebastián C, Aso-Vizán A, Martínez-Quiñones JV, Consolini F, and Arregui R (2019). Delay in diagnosis of thoracolumbar fractures. Orthopedic Reviews, 11(2): 7774. https://doi.org/10.4081/or.2019.7774 **PMid:31210909 PMCid:PMC6551460**

Cowan PT, Launico MV, and Kahai P (2020). Anatomy, bones. StatPearls Publishing, Treasure Island, USA.

Gan K, Xu D, Lin Y, Shen Y, Zhang T, Hu K, Zhou K, Bi M, Pan L, Wu W, and Liu Y (2019). Artificial intelligence detection of distal radius fractures: A comparison between the convolutional neural network and professional assessments. Acta Orthopaedica, 90(4): 394–400. https://doi.org/10.1080/17453674.2019.1600125 **PMid:30942136 PMCid:PMC6718190**

Goodfellow I (2016). Deep learning. MIT Press, Cambridge, USA.

Hardalaç F, Uysal F, Peker O, Çiçeklidağ M, Tolunay T, Tokgöz N, Kutbay U, Demirciler B, and Mert F (2022). Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors, 22(3): 1285. https://doi.org/10.3390/s22031285 **PMid:35162030 PMCid:PMC8838335**

He K, Zhang X, Ren S, and Sun J (2016). Deep residual learning for image recognition. In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, USA: 770–778. https://doi.org/10.1109/CVPR.2016.90 **PMid:26180094**

Huang G, Liu Z, Van Der Maaten L, and Weinberger KQ (2017). Densely connected convolutional networks. In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Honolulu, USA: 4700-4708. https://doi.org/10.1109/CVPR.2017.243 **PMCid:PMC5598342**

Jones RM, Sharma A, Hotchkiss R, Sperling JW, Hamburger J, Ledig C, O'Toole R, and Gardner M et al. (2020). Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digital Medicine, 3(1): 144. https://doi.org/10.1038/s41746-020-00352-w **PMid:33145440 PMCid:PMC7599208**

Krizhevsky A, Sutskever I, and Hinton GE (2017). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25: 1097–1105.

Mohanty S and Senapati MR (2023). Fracture detection from X-ray images using different machine learning techniques. In the 1st International Conference on Circuits, Power and Intelligent Systems, IEEE, Bhubaneswar, India: 1-6. https://doi.org/10.1109/CCPIS59145.2023.10291652

Nishiyama M, Ishibashi K, Ariji Y, Fukuda M, Nishiyama W, Umemura M, Katsumata A, Fujita H, and Ariji E (2021). Performance of deep learning models constructed using panoramic radiographs from two hospitals to diagnose

fractures of the mandibular condyle. Dentomaxillofacial Radiology, 50(7): 20200611.
https://doi.org/10.1259/dmfr.20200611
PMid:33769840 PMCid:PMC8474128

Nwankpa C, Ijomah W, Gachagan A, and Marshall S (2018). Activation functions: Comparison of trends in practice and research for deep learning. Arxiv Preprint Arxiv:1811.03378.
https://doi.org/10.48550/arXiv.1811.03378

Rodrigo M (2024). Bone fracture multi-region x-ray data. Available online at:
https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data

Sharma S (2023). Artificial intelligence for fracture diagnosis in orthopedic X-rays: Current developments and future potential. SICOT Journal, 9: 21.
https://doi.org/10.1051/sicotj/2023018
PMid:37409882 PMCid:PMC10324466

Son DM, Yoon YA, Kwon HJ, An CH, and Lee SH (2021). Automatic detection of mandibular fractures in panoramic radiographs using deep learning. Diagnostics, 11(6): 933.
https://doi.org/10.3390/diagnostics11060933
PMid:34067462 PMCid:PMC8224557

Su Z, Adam A, Nasrudin MF, Ayob M, and Punganan G (2023). Skeletal fracture detection with deep learning: A comprehensive review. Diagnostics, 13(20): 3245.
https://doi.org/10.3390/diagnostics13203245
PMid:37892066 PMCid:PMC10606060

Tan M and Le Q (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In the International Conference on Machine Learning, PMLR, Long Beach, USA: 6105-6114.

Tanzi L, Vezzetti E, Moreno R, and 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

Taylor-Phillips S and Stinton C (2019). Fatigue in radiology: A fertile area for future research. The British Journal of Radiology, 92(1099): 20190043.
https://doi.org/10.1259/bjr.20190043
PMid:30933540 PMCid:PMC6636274

Wang X, Xu Z, Tong Y, Xia L, Jie B, Ding P, Bai H, Zhang Y, and He Y (2022). Detection and classification of mandibular fracture on CT scan using deep convolutional neural network. Clinical Oral Investigations, 26(6): 4593-4601.
https://doi.org/10.1007/s00784-022-04427-8
PMid:35218428

Yadav DP, Sharma A, Athithan S, Bhola A, Sharma B, and Dhaou IB (2022). Hybrid SFNet model for bone fracture detection and classification using ML/DL. Sensors, 22(15): 5823.
https://doi.org/10.3390/s22155823
PMid:35957380 PMCid:PMC9371081