A Novel Ensemble Approach for Rib Fracture Detection and Visualization using CNNs and Grad-CAM

Ann. Ital. Chir., 2025 96, 1: 86–97
https://doi.org/10.62713/aic.3666

Ling Wu[1], Hongyu Chen[1], Puxu Li[1], Kai Yang[2]

[1]Thoracic Surgery Department, Ningbo No.9 Hospital, 315020 Ningbo, Zhejiang, China
[2]Medical Department, Ningbo No.9 Hospital, 315020 Ningbo, Zhejiang, China

AIM: This study aimed to develop a reliable and efficient system for predicting and locating rib fractures in medical images using an ensemble of convolutional neural networks (CNNs).
METHODS: We employed five CNN architectures—Visual Geometry Group Network 16 (VGG16), Densely Connected Convolutional Network 169 (DenseNet169), Inception Version 4 (Inception V4), Efficient Network B7 (EfficientNet-B7), and Residual Network Next 50 layers (ResNeXt-50)—trained on a dataset of 840 grayscale computed tomography (CT) scan images in .jpg format collected from 42 patients at a local hospital. The images were categorized into two groups representing healed and fresh fractures. The ensemble model was designed to improve predictive accuracy and robustness, utilizing techniques like gradient-weighted class activation mapping (Grad-CAM) for visualization of fracture locations.
RESULTS: The ensemble model achieved an accuracy of 0.96, area under the curve (AUC) of 0.97, recall of 0.97, and F1 score of 0.96. Grad-CAM visualizations could effectively locate rib fractures, providing crucial assistance in diagnostics.
CONCLUSIONS: The ensemble model demonstrates high accuracy and robustness in fracture detection, underscoring its potential for enhancing diagnostic processes in clinical settings. Despite limitations such as the small dataset size and lack of diverse demographic representation, the results are promising for future clinical application.

Keywords: rib fractures; machine learning; CT images; predictive modeling; Grad-CAM visualization

Introduction

Rib Fracture Background

Rib fractures, resulting from the partial or complete disruption of one or more ribs in the thoracic cage, are clinically significant due to their implications for pain, respiratory compromise, and potential complications such as pulmonary or visceral injury [1,2]. These fractures can occur from blunt trauma, such as falls or accidents, or through pathological processes that weaken the bone, including osteoporosis [3]. As the ribs, along with the sternum and spine, serve as a protective barrier for vital organs like the heart and lungs, even subtle rib injuries necessitate careful assessment and prompt intervention [4]. The diagnostic workup for rib fractures involves several imaging modalities [5]. Conventional radiography is often the first-line investigative strategy by virtue of its affordability and widespread availability; however, this approach has limited sensitivity especially in detecting minor fractures [6]. Conversely, computed tomography (CT) scans provide high-resolution, three-dimensional images that are superior for identifying subtle or complex fracture patterns [7]. Ultrasound offers a non-invasive, radiation-free option that is especially significant for the pediatric population or for repeated assessments [8]. While magnetic resonance imaging (MRI) is not a routine choice for rib fracture diagnosis, it can be useful in assessing accompanying soft tissue injuries or in specific patient populations such as pregnant women [9].

Machine Learning in Rib Fracture Diagnosis

Machine learning applications in rib fracture diagnosis offer numerous benefits over conventional approaches [10]. These include enhanced diagnostic accuracy, improved efficiency, and objectivity in interpretation, potential for real-time image analysis, early intervention, and reduced workload for healthcare professionals. Among machine learning techniques, convolutional neural networks (CNNs) have shown considerable promise in fracture detection, owing to their ability to process large datasets and learn from complex patterns and features, thereby achieving accuracy rates superior to human interpreters [11]. Automation in rib fracture detection not only alleviates the diagnostic burden on radiologists but also ensures consistent and objective evaluations, minimizing inter-observer variability and

Submitted: 19 August 2024 Revised: 27 September 2024 Accepted: 28 October 2024 Published: 10 January 2025
Correspondence to: Kai Yang, Medical Department, Ningbo No.9 Hospital, 315020 Ningbo, Zhejiang, China (e-mail: bennyyang1983@aliyun.com).

reducing diagnostic errors. This in turn facilitates evidence-based clinical decision-making [12]. Additionally, machine learning algorithms enable prompt fracture identification, permitting timely interventions such as pain management, monitoring for complications, and treatment planning. The possibility of real-time image analysis by machine learning algorithms is particularly beneficial in emergencies, enabling rapid assessments and informed decisions [13].

Within the context of machine learning applications in rib fracture diagnosis, several studies have explored diverse methodologies and their potential contributions. Singh *et al.* [14] conducted a comprehensive review of the principles, applications, and limitations of machine learning in various imaging domains including thoracic radiology, highlighting its transformative potential. Bukkuri *et al.* [15] demonstrated the potential of a machine learning algorithm for medical image analysis by devising a topological invariant classifier. Silva *et al.* [16] mirrored Singh *et al.*'s work [14], offering insights into machine learning applications and limitations in similar imaging domains. Daghigh *et al.* [17] utilized decision tree regressor and adaptive boosting regressor machine learning methods for heat deflection temperature predictions. This approach emphasizes the distinctive application of these techniques in the context of their study [18]. Balcıoğlu and Seçkin [18] integrated experimental methods, finite element analysis, and machine learning algorithms to investigate the fracture behavior of polymer composites under different loading conditions. Huang *et al.* [19] introduced a semi-supervised learning framework to overcome the scarcity of precisely delineated labels for rib fracture tasks. Wang *et al.* [20] developed machine learning approaches, demonstrating their utility through the analysis and prediction of Mode-I fracture toughness of rocks. Meng *et al.* [21] presented a heterogeneous neural network for rib fracture detection and classification, consisting of a cascaded feature pyramid network and a classification network. Zhang *et al.* [22] investigated an algorithm combining nnU-Net and DenseNet for automated rib fracture recognition. Niiya *et al.* [23] evaluated the clinical utility of an AI-assisted CT diagnosis technology for rib fractures. These studies demonstrate the application of machine learning technologies across various fields, particularly in medical imaging and materials science. Collectively, they underscore the robust utility of machine learning in diverse domains, enhancing the accuracy and efficiency of analyses. They not only highlight the advancements in algorithm development but also emphasize the practical impact of these technologies in addressing complex real-world problems. The integration of experimental and computational methods in these studies paves the way for the development of innovative solutions that can significantly influence both scientific research and industrial applications.

The Necessity and Innovation of This Research

This research significantly advances the field of rib fracture detection by employing a strategic combination of innovative methodologies, which leverages the strengths of five advanced convolutional neural network (CNN) architectures—Visual Geometry Group Network 16 (VGG16), Densely Connected Convolutional Network 169 (DenseNet169), Efficient Network B7 (EfficientNet-B7), Inception Version 4 (Inception V4), and Residual Network Next 50 layers (ResNeXt-50). These models were selected due to their proven robust performance on various benchmark datasets in image classification, which ensures their reliability in medical image analysis. By integrating these diverse architectures, each known for unique feature extraction capabilities, our study creates a fusion model that is not only more robust and accurate but also adaptable to different rib fracture patterns captured in a unique dataset collected from a local hospital.

The versatility of our approach is further enhanced by incorporating gradient-weighted class activation mapping (Grad-CAM) technology, which allows for the generation of heat maps pinpointing the exact location of rib fractures in CT images. This dual-function system is capable of both high-accuracy classification and precise lesion identification, addressing the pressing need for versatile, efficient, and interpretable tools in medical diagnostics. The choice of these specific CNNs also brings scalability and efficiency, essential for processing the large datasets and high-resolution images typical of medical settings. Furthermore, the advanced visualization capabilities of models like Inception V4 and DenseNet169 improve the localization of fractures, making our method a valuable addition to diagnostic processes. By amalgamating models that interpret data through different lenses, our integrated ensemble approach significantly pushes forward the accuracy and generalizability of rib fracture predictions, representing a substantial contribution to the medical field.

Materials and Methods

Data Collection

Comprehensive datasets of 42 patients employed in this study were collected from Ningbo No.9 Hospital, which is a major medical center recognized for its advanced imaging and trauma care facilities. The data comprise a total of 840 grayscale CT scan images in .jpg format, each with a resolution of 1024 × 1024 pixels. The images were obtained following the hospital's established protocol for whole-body trauma CT scans. The acquisition procedure involved scanning patients in the supine position, with the imaging field extending from the top of the head to the upper thigh, ensuring comprehensive coverage of the thoracic region.

Patients included in this study were selected based on the following criteria:

Inclusion criteria: Patients eligible for inclusion in the study were those who underwent CT imaging for trauma assessment at Ningbo No.9 Hospital during the past three years and were diagnosed with rib fractures based on CT findings by a board-certified radiologist.

Exclusion criteria: This study excluded patients with a history of skeletal diseases affecting bone integrity, such as osteogenesis imperfecta or severe osteoporosis, due to the potential for these conditions to alter the typical imaging features of rib fractures. Additionally, patients with CT scans of insufficient quality for diagnostic purposes, including scans with artifacts that obscure the clarity of bone structures, were also excluded.

These criteria ensure that the study focuses on a population relevant to typical clinical cases encountered in trauma settings, excluding cases where pre-existing conditions or poor image quality could confound the results.

The dataset consisted of 840 CT images collected from 42 patients, with each patient contributing 20 images. To facilitate meaningful analysis, the datasets were meticulously categorized into two distinct categories: negative (21 patients) and positive (21 patients) groups. The negative group consisted of images representing healed fractures, including older fractures that displayed signs of remodeling or calcification. In contrast, the positive group encompassed images of fresh rib fractures characterized by the presence of acute cortical discontinuity, hematoma, or displacement.

The classification of CT images into “negative” (healed fractures) and “positive” (acute fractures) categories was based on established diagnostic criteria derived from clinical practice and radiological literature [7]. In particular, images for the positive group were scored based on the presence of any signs of acute cortical discontinuity, visible hematoma, or clear displacement of the rib segments, whereas those for the negative group were based on the evidence of a healed fracture, characterized by bone remodeling or calcification without signs of recent injury. These criteria were applied by a panel of three radiology specialists who independently reviewed each image during categorization, ensuring consistency and accuracy in classification.

To minimize potential biases and enhance the reliability of our dataset, image classification was conducted by three radiology residents who have relevant experience, and have been trained in recognizing rib fractures. Each resident independently reviewed and categorized the CT scan images based on their observations. By employing multiple reviewers with expertise in rib fractures, we ensured a comprehensive evaluation of each image, bolstered inter-rater agreement, and mitigated the risk of misclassification. Their independent assessments were then cross-validated to confirm the final classification, reducing the risks for observer bias and enhancing the dataset’s accuracy.

Research Flowchart

The proposed machine learning system for rib fracture prediction was designed and constructed in accordance with a series of well-structured stages, as shown in Fig. 1. Initially, CT data depicting rib fractures, including both healed and fresh fractures, were collected with an emphasis on imaging features specific to rib fractures.

The dataset, consisting of 840 CT images from 42 patients, was divided into a training set and a validation set using a stratified random sampling approach to ensure that both sets were representative of the overall dataset. This method helped maintain a consistent distribution of cases (healed and acute fractures) in both sets. The training set comprised 80% of the total dataset, equating to 672 CT images, which were used to train the models, allowing them to learn and adapt to the complex patterns associated with rib fracture detection. The validation set consisted of the remaining 20% of the dataset, equivalent to 168 CT images, which were utilized to validate the models’ performance and ensure their accuracy and generalizability before final testing. Training the models is the next crucial step. In this study, five distinct deep learning models were considered: VGG16, DenseNet169, EfficientNet-B7, Inception V4, and ResNeXt-50. Each model underwent thorough training on the training set, learning from the data’s complex patterns and features to develop an understanding of the rib fractures’ imaging characteristics. The goal was to fine-tune the models for optimal performance in rib fracture prediction. Once trained, the models were tested on the unseen test set, examining their generalization ability and performance across multi-center and multi-parameter datasets. The evaluation entails measuring various performance metrics such as accuracy, precision, and recall to assess the models’ efficacy in rib fracture prediction.

To further enhance the system’s predictive capabilities, we employed an ensemble model leveraging the VotingClassifier from the sklearn library,which is part of scikit-learn, an open-source machine learning library for Python (Scikit-learn is developed by a community of contributors and is distributed under the Berkeley Software Distribution (BSD) license). By aggregating the predictions of the five trained models, the ensemble model enhances the robustness and stability of the rib fracture prediction, capitalizing on the strengths of the individual models. Incorporated with Grad-CAM for model interpretability, this system visualizes regions in CT images critical to rib fracture prediction, providing insights to aid with the decision-making process in fracture detection. By highlighting key image features, Grad-CAM makes the models’ predictions more transparent and understandable.

Data Processing

To safeguard patient privacy and comply with ethical standards, we implemented data anonymization by removing any personally identifiable information, such as patient

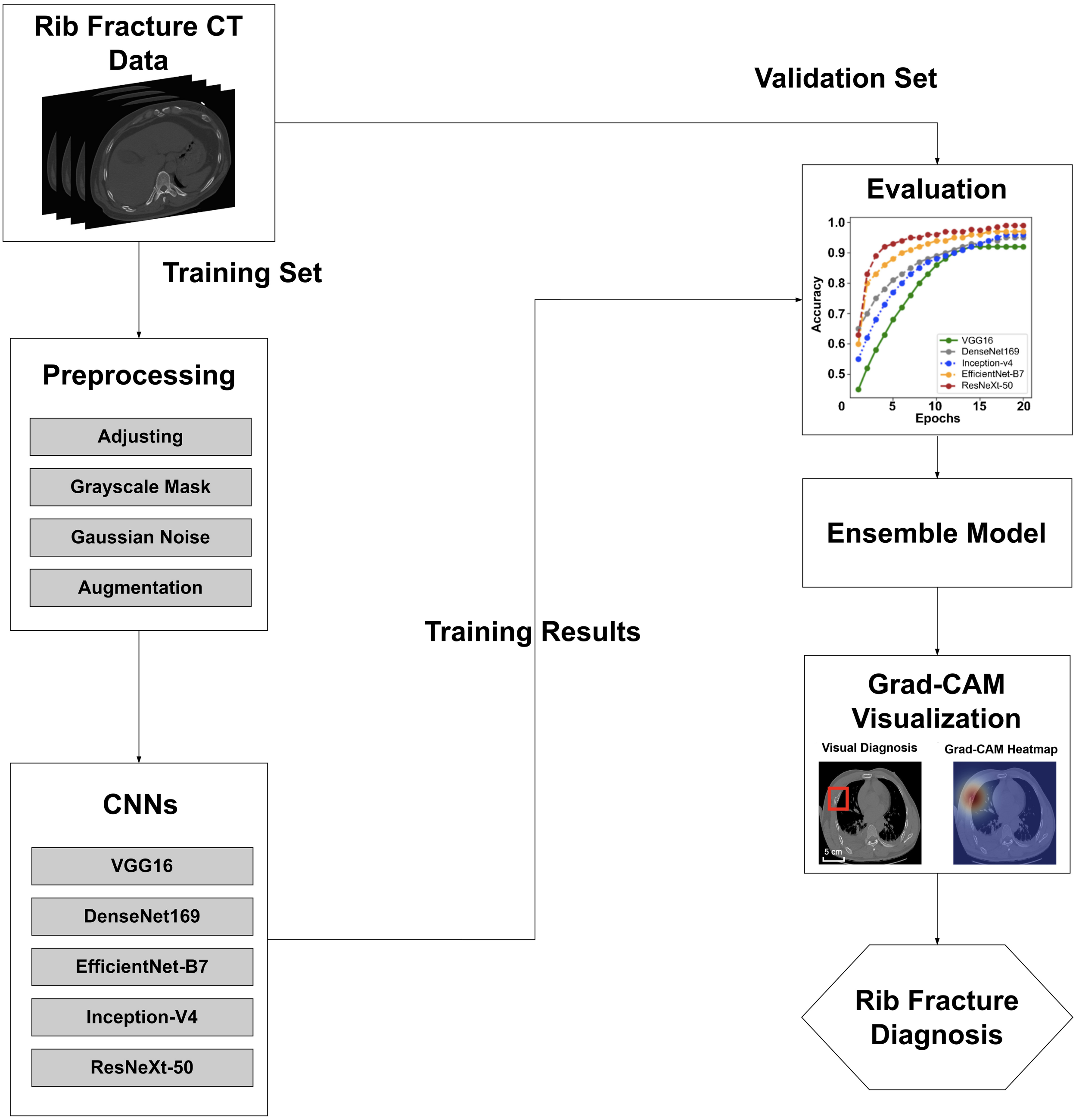

Fig. 1. Flowchart of the proposed deep learning system for rib fracture prediction and interpretation. This figure was created by the authors using Microsoft Office PowerPoint (Microsoft 16.0.14701.20210, Microsoft Office, Redmond, WA, USA). The red box indicates the location of rib fractures identified by the radiologist. Abbreviations: CNNs, convolutional neural networks; CT, computed tomography; Grad-CAM, gradient-weighted class activation mapping; VGG16, Visual Geometry Group Network 16; DenseNet169, Densely Connected Convolutional Network 169; EfficientNet-B7, Efficient Network B7; Inception V4, Inception Version 4; ResNeXt-50, Residual Network Next 50 layers.

names or identification numbers, from the medical images. Anonymization techniques included erasing metadata and blurring sensitive regions. Next, we cropped the target area within the images, as rib fractures are localized in specific regions. This step eliminated unnecessary background noise and irrelevant anatomical structures, allowing the models to focus on the crucial features associated with rib fractures. The cropped images underwent normalization, scaling pixel values to a range between 0 and 1 to enable consistent computation across the dataset. The images were resized to a standard dimension while preserving the aspect ratio to ensure compatibility with the chosen machine learning algorithms. To emphasize rib fractures, grayscale masks were applied to the cropped images. These masks were generated by thresholding the original images and converting them to grayscale, assigning higher pixel intensities to rib fractures and lower intensities to the surrounding regions, enhancing target areas' visibility. Gaussian noise was introduced to the preprocessed images to augment the dataset and increase its diversity. This noise, which follows a Gaussian distribution, mimics the noise typically found in medical images. By adding controlled levels of Gaussian noise, we made the models more robust and better equipped to handle noisy input data. Additional data augmentation techniques were applied, including Random Fog, Random Contrast, and Random Rotation. These transformations added variation to the dataset, reducing overfitting and improving the models' generalization ability. After preprocessing, radiologists provided initial approval of the resulting images, ensuring that essential features and properties for accurate rib fracture analysis were retained. Radiologists' expertise and experience were crucial for achieving consensus on the quality and suitability of preprocessed images. Fig. 2 presents a visual representation of original image and preprocessed image, and Fig. 3 depicts the outcome of the data augmentation process.

Reference Standard for Rib Fracture Detection Using Machine Learning

The reference standard for this research was derived from annotated CT reports of rib fractures, each of which had been rigorously reviewed and approved by a seasoned, board-certified radiologist specializing in emergency radiology at a level-one trauma center. These CT reports functioned as the benchmark against which the effectiveness of our deep learning/machine learning algorithms was measured. The reference standard included a thorough evaluation of rib fractures, factoring in details such as their location, extent, and other relevant attributes as described in the reports. The establishment of the reference standard involved careful validation and assessment of the reports to confirm their dependability and precision. The reference standard was further bolstered by the radiologist's extensive expertise and long-standing experience in emergency radiology, lending substantial credibility to the foundation for the performance evaluation of the proposed deep learning/machine learning models.

Training Using Pretrained CNN Models

In our study, we employed several advanced CNNs to evaluate their effectiveness in detecting rib fractures from CT images. These models included the VGG16, DenseNet169, Inception V4, EfficientNet-B7, and ResNeXt-50. Each model brings unique strengths to our ensemble approach, allowing for comprehensive analysis and improved prediction accuracy. We fed the models with a curated collection of preprocessed and annotated CT scans. Data augmentation methods, such as rotation and flipping, were employed to bolster the models' robustness and generalization capabilities.

VGG16 consists of 13 convolutional layers followed by 3 fully connected layers, enabling the model to learn intricate data representations. Its architecture prioritizes depth over width by stacking multiple convolutional layers sequentially, a design choice that has proven effective in enhancing performance across various image classification benchmarks. The model's success lies in its ability to capture spatial hierarchies of features, making it highly suitable for tasks such as object detection and image recognition [24]. DenseNet169 incorporated densely connected layers with a growth rate of 32, and used batch normalization, Rectified Linear Unit (ReLU) activation, and dropout regularization at a rate of 0.2 [25]. Inception V4 employed the inception module with a mixture of 1×1, 3×3, and 5×5 filters in its deep architecture. It utilized batch normalization, ReLU activation, and dropout regularization at a rate of 0.3 [26]. EfficientNet-B7, scaled with a compound coefficient of 1.4, featured depthwise separable convolutions, squeeze-and-excitation blocks, and the Swish activation function. It incorporated batch normalization and dropout regularization at a rate of 0.4 [27]. ResNeXt-50, an extension of the ResNet architecture, used residual blocks with a cardinality of 32. It featured batch normalization, ReLU activation, and dropout regularization at a rate of 0.5 [28].

For training, we used the Adam optimizer with a learning rate of 0.001, a batch size of 32, and a weight decay of 10^{-5}. We trained each model for 50 epochs and employed early stopping with a patience of 5 epochs to prevent overfitting. We employed 10-fold cross-validation to assess the models' performance. The dataset was partitioned into ten equal subsets, with each subset serving as the test set once. We averaged the performance metrics across the ten iterations for a robust evaluation of the models' performance.

Statistical Evaluation

This study employed a suite of statistical metrics to assess the performance of machine learning models in rib fracture prediction. The overall correctness of predictions, combining both fracture and non-fracture cases, is represented by accuracy. The F1 score, a harmonized measure of precision and recall, provides an overall performance metric for rib fracture identification. Receiver operating characteristic (ROC) curves illustrate model performance across different classification thresholds, with the area under the curve (AUC) summarizing this performance. These metrics furnish a thorough evaluation of the models' predictive powers and inform future model refinement.

Ensemble Model

In this study, an ensemble model was constructed to enhance the robustness and versatility of the rib fracture prediction system. The ensemble model was built by combining the predictions of the five individual CNN models, namely VGG16, DenseNet169, Inception V4, EfficientNet-B7, and ResNeXt-50. Each model offers unique strengths

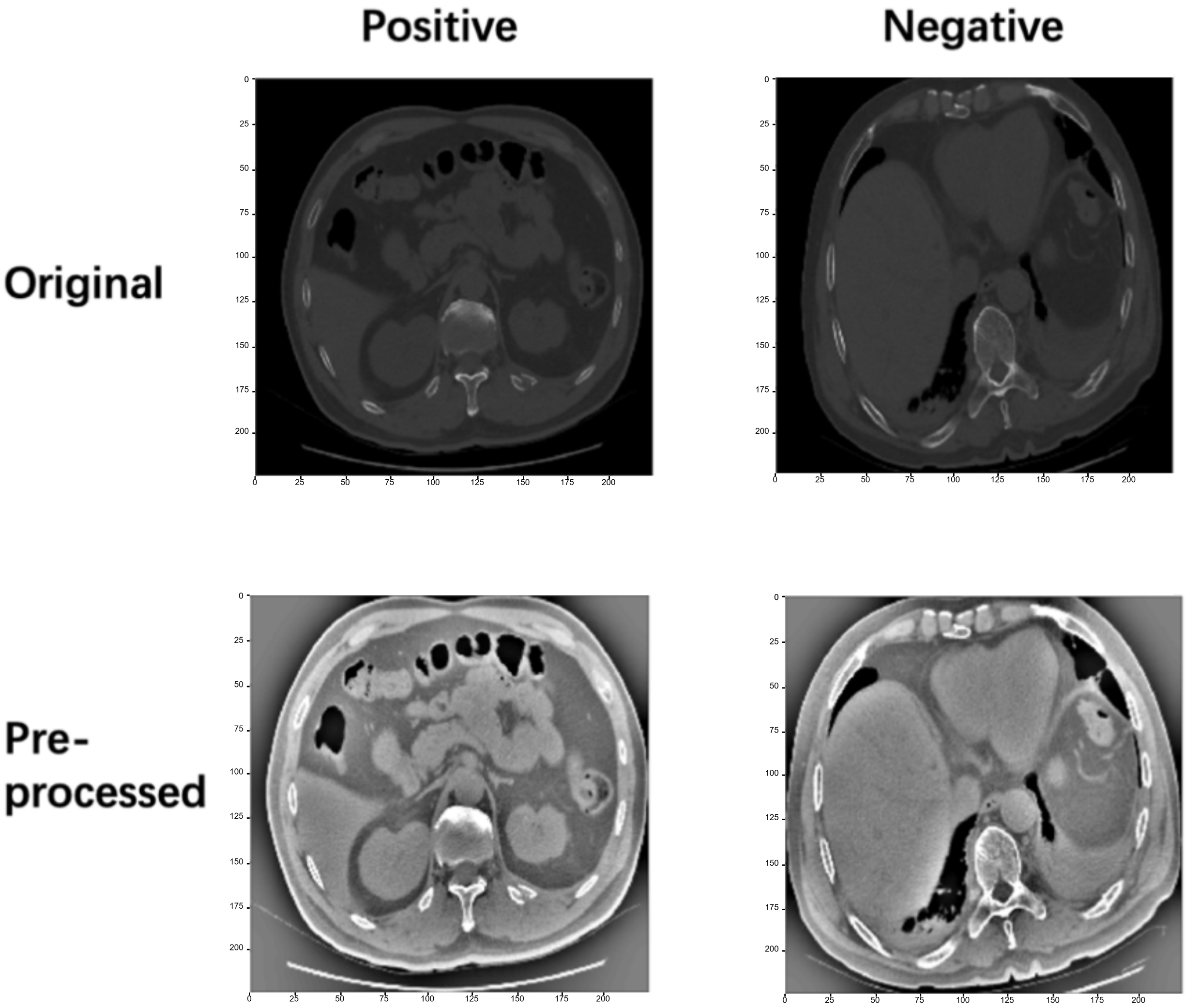

Fig. 2. Comparison of original and preprocessed CT images for rib fracture detection. The x-axis and y-axis represent the pixel coordinates of the images. The comparison showcases the outcome of enhanced fracture visibility following the implementation of image processing techniques.

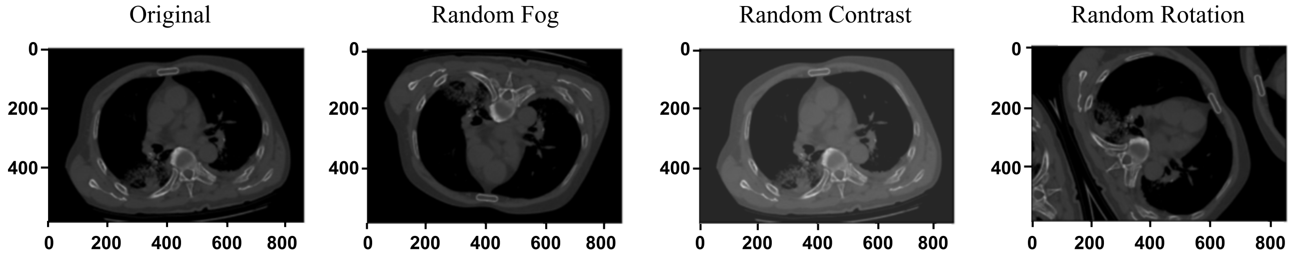

Fig. 3. Effects of data augmentation on preprocessed CT images. This figure displays examples of CT images before and after data augmentation techniques were applied, highlighting how various augmentations such as Random Rotation and Random Contrast adjustment alter the appearance of the image. The x-axis and y-axis denote the pixel coordinates.

and exhibits different learning patterns. By aggregating their outputs, the ensemble model leveraged the skills of these individual models and increased the overall prediction accuracy. The VotingClassifier from the scikit-learn library was used for the implementation of the ensemble model. This classifier collected the predictions from each of the five CNN models and applied a majority voting approach to determine the final prediction. In the event of a tie, the ensemble model follows a predefined rule or randomly selects one of the tied outcomes as the final prediction. The ensemble model was expected to outperform any individual model in terms of prediction accuracy, generalizability, and robustness. This is due to the fact that the ensemble approach combines the insights and learnings from multiple models, helping to mitigate any weaknesses or biases that a single model might possess.

Grad-CAM Visualization

The Grad-CAM is a visualization technique used in this research to provide insights into the decision-making processes of the ensemble model [29]. The technique works by highlighting the regions in an input image that have the most influence on the model's predictions, making it an invaluable tool for understanding and interpreting the predictions made by the ensemble model. This technique is particularly useful for identifying and localizing rib fractures in CT images, allowing medical practitioners to gain a clearer understanding of the fractures and facilitating better-informed decisions.

In the context of our research, Grad-CAM was applied to the ensemble model, which combined the predictions of the five individual CNN models. The Grad-CAM method focuses on the convolutional layers of the CNN models, as these layers are responsible for extracting spatial and visual features from the input images. By examining the activations of the convolutional layers, Grad-CAM identifies the critical regions in the images that contribute significantly to the model's decisions. The Grad-CAM process is shown in Fig. 4. First, the target convolutional layer is identified: In this research, Grad-CAM was connected to one of the final convolutional layers of the CNN models within the ensemble. This target layer was chosen because it contained high-level feature maps that were closely related to the model's predictions. Second, the gradient of the output class is computed with respect to the target layer: Grad-CAM calculated the gradient of the predicted class score with respect to the feature maps of the target convolutional layer. These gradients represented the importance of each feature map in the final prediction. Lastly, the class activation mapping (CAM) is generated: Grad-CAM generated the CAM by taking the weighted sum of the target layer's feature maps using the computed gradients as weights. The CAM was then upscaled to the size of the input image, highlighting the influential regions.

Results

Baseline Characteristics of Patients

This study included a sample of 42 patients, comprising 24 males and 18 females, with an age range from 18 to 65 years. The median age of these patients was 37 years. The disease status at the time of imaging included acute rib fractures due to blunt trauma (100% of cases), with 12 patients (28.6%) also exhibiting minor associated injuries such as soft tissue damage. All patients provided informed consent for the use of their anonymized images in this research study.

Model Evaluation Metrics

Table 1 presents the performance evaluation of the five CNN models, which were assessed using the validation set for predicting rib fractures: VGG16, DenseNet169, Inception V4, EfficientNet-B7, and ResNeXt-50. The models were evaluated using four metrics: accuracy, AUC, recall, and F1 score. The choice to focus on validation set results is to emphasize the models' ability to generalize to new, unseen data, which is crucial for clinical applications. For accuracy, ResNeXt-50 showed the highest performance, achieving an accuracy of 0.99. EfficientNet-B7 followed closely with an accuracy of 0.97. Inception V4 and DenseNet169 reported accuracies of 0.96 and 0.95, respectively, while VGG16 obtained an accuracy of 0.92. In terms of AUC, ResNeXt-50 also achieved the highest value of 0.98. EfficientNet-B7 recorded an AUC of 0.96. Inception V4 and DenseNet169 attained AUC values of 0.94 and 0.93, respectively, and VGG16 reported the lowest AUC of 0.90. Regarding recall, ResNeXt-50 again outperformed the other models with a recall of 0.98. EfficientNet-B7 displayed a recall of 0.95, while Inception V4 and DenseNet169 showed recall values of 0.92 and 0.90, respectively. VGG16 had the lowest recall at 0.85. Regarding the F1 score, ResNeXt-50 led the pack with a score of 0.97. EfficientNet-B7 achieved an F1 score of 0.94. Inception V4 and DenseNet169 attained F1 scores of 0.90 and 0.88, respectively, while VGG16 registered the lowest F1 score of 0.82.

Model Convergence Analysis

Fig. 5 illustrates the model convergence of the five CNN models over 20 epochs. The models considered include VGG16, DenseNet169, Inception V4, EfficientNet-B7, and ResNeXt-50. VGG16 started with an accuracy of 0.2 at epoch 1 and increased steadily to reach 0.88 by epoch 12. The accuracy remained consistent from epoch 12 onwards. DenseNet169 showed rapid convergence, starting at an accuracy of 0.65 at epoch 1 and reaching 0.91 by epoch 12. The accuracy remained s around 0.95 for subsequent epochs. Inception V4 demonstrated an initial accuracy of 0.55 at epoch 1, which increased to 0.88 by epoch 12 and further improved to 0.94 in the subsequent epochs. EfficientNet-B7 started with an accuracy of 0.6 at epoch 1 and exhibited a rapid increase in accuracy, reaching 0.97 by epoch 16. The accuracy surpassed 0.95 for subsequent epochs. ResNeXt-50 demonstrated a significant improvement in accuracy, starting at 0.63 at epoch 1 and achieving an accuracy of 0.99 by epoch 19. The accuracy converged quickly and remained at this high level for the remaining

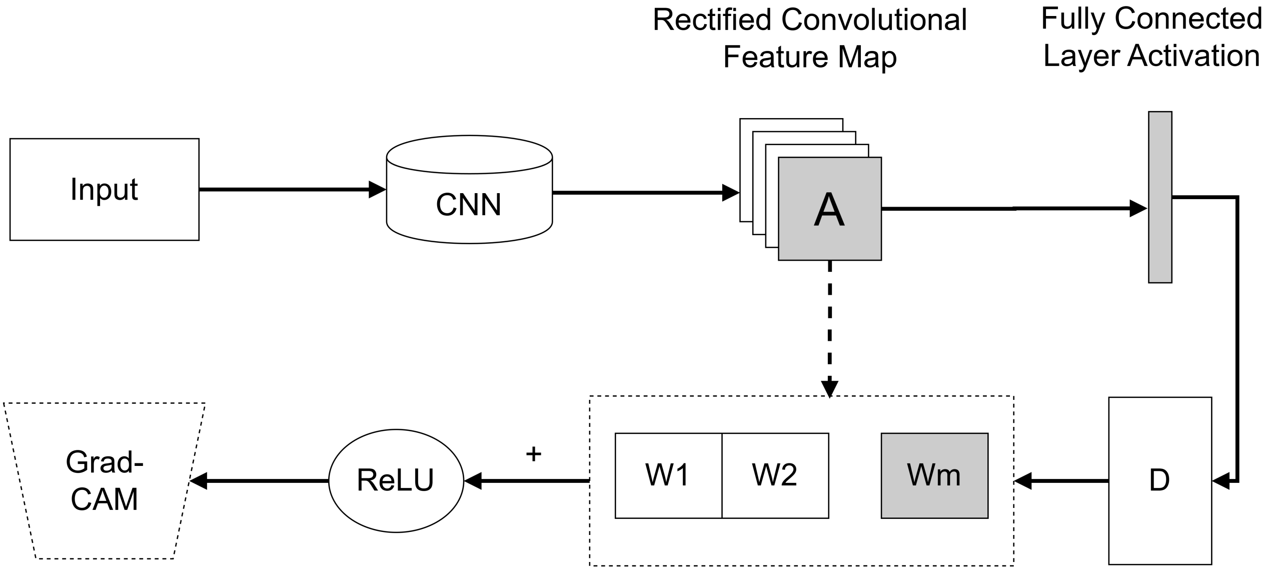
Rectified Convolutional Feature Map
Fully Connected Layer Activation
Input
CNN
A
Grad-CAM
ReLU
+
W1
W2
Wm
D

Fig. 4. Grad-CAM visualization process for identifying influential regions in rib fracture predictions. This figure was created by the authors using Microsoft Office PowerPoint (Microsoft 16.0.14701.20210, Microsoft Office, Redmond, WA, USA). A: represents the rectified convolutional feature map, which is the output of the CNN after activation functions such as ReLU. These feature maps encode spatial and semantic information from the input image. D: denotes the fully connected layer activation, which is the output of the fully connected neural network layer. This represents the final decision or classification logits before applying a softmax or other activation for prediction. W1, W2, ..., Wm: represent the weights of the fully connected layer corresponding to each feature map in A. These weights are used to compute the Grad-CAM by combining gradients and feature maps to localize discriminative regions of the input image. Abbreviations: CNN, convolutional neural network; ReLU, Rectified Linear Unit.

Table 1. Performance metrics of CNN models for rib fracture prediction.

Metrics	VGG16	DenseNet169	Inception V4	EfficientNet-B7	ResNeXt-50
Accuracy	0.92	0.95	0.96	0.97	0.99
AUC	0.90	0.93	0.94	0.96	0.98
Recall	0.85	0.90	0.92	0.95	0.98
F1 score	0.82	0.88	0.90	0.94	0.97

Abbreviations: AUC, area under the curve; DenseNet169, Densely Connected Convolutional Network 169; Inception V4, Inception Version 4; EfficientNet-B7, Efficient Network B7; ResNeXt-50, Residual Network Next 50 layers.

epochs. In summary, among the five models, ResNeXt-50 achieved the highest accuracy of 0.99, followed by EfficientNet-B7 at 0.97. Inception V4 and DenseNet169 recorded accuracies of 0.96 and 0.95, respectively, while VGG16 achieved an accuracy of 0.92.

Performance of Ensemble Model

The ensemble model, which combined the predictions of the five CNN models, was evaluated using the same metrics employed in assessing the individual models. The ensemble model achieved an accuracy of 0.96, an AUC of 0.97, a recall of 0.97, and an F1 score of 0.96. These results demonstrated that the ensemble model exhibited robust performance, a sign of effective maximization of the individual models' strengths for improving predictive capability.

Visualization of Rib Fractures with Grad-CAM

Fig. 6 showcases the heatmaps of rib fractures generated using the Grad-CAM technique, visualizing the in the images produced by the ensemble model. The heatmap displays areas of high and low model response in the image. In the heatmap, red regions represent high model response, indicating the areas with the most significant contribution to the model's decision, while blue regions indicate low model response. In this study, the location of rib fractures in the images was determined by radiologists through direct inspection of the images. Their diagnosis is marked with a red box in the images for comparison (Fig. 6). Upon examining the Grad-CAM heatmaps alongside the visual diagnosis images, we found that the regions of highest model response in the heatmap overlapped with the actual location of the rib fractures, as identified by the radiologists.

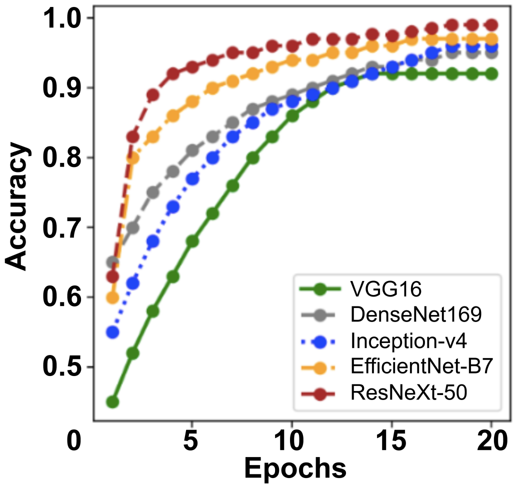

Fig. 5. Comparison of accurate convergence rates of five CNN models on CT images of rib fractures. Abbreviations: CT, computed tomography; CNN, convolutional neural network.

This finding underlines the effectiveness of the Grad-CAM technique in highlighting the regions of interest, aiding in the interpretation of the model's predictions and providing valuable insights into the areas that contribute to the detection of rib fractures.

Discussion

This study demonstrated that CNNs are highly effective in predicting rib fractures. The effectiveness of five pre-trained CNN models was evaluated, with ResNeXt-50 emerging as the top-performing model in our study. This model demonstrated impressive results, achieving an accuracy of 0.99, an AUC of 0.98, recall of 0.98, and an F1 score of 0.97. These metrics collectively indicate the robust performance of ResNeXt-50 in identifying and classifying rib fractures. This can be attributed to the unique architecture of the ResNeXt-50 model, which leverages residual connections and a large number of feature channels. The model utilizes "cardinality" as a dimension, adding multiple parallel residual blocks to increase network capacity and improve performance [30].

Our study expanded on the initial success of individual CNN models by developing an ensemble model that integrates the strengths of all five CNNs. The ensemble model demonstrated an accuracy of 0.96, an AUC of 0.97, recall of 0.97, and an F1 score of 0.96. These metrics affirm the effectiveness of the ensemble approach in consolidating the predictive capabilities of multiple models, offering a more versatile and accurate diagnostic tool. While the ensemble model's accuracy is slightly lower than that of the ResNeXt-50 model alone, it benefits from the aggregated expertise of various models, resulting in a more comprehensive and robust system [31].

While ResNeXt-50 showed the best individual performance, the ensemble model integrated the diverse strengths of all five CNN models, enhancing the overall robustness and reducing the risk of overfitting to specific patterns present in the training data. In clinical practice, the variability in imaging due to different machines, settings, and patient demographics can introduce complexities that a single model might not handle as effectively as an ensemble. For instance, in scenarios where image quality is compromised or where fracture presentations are atypical, the ensemble approach can leverage the collective strengths of all individual models to maintain high accuracy.

Additionally, to address the concerns about the ensemble model's added complexity, our experiments indicated that the ensemble model consistently performed better under conditions of limited data and increased class imbalance, which are common challenges in medical imaging contexts. This suggests that the ensemble model is not only more versatile but also more capable of adapting to diverse clinical environments, such as different imaging machines, settings, or patient demographics.

The incorporation of the Grad-CAM technique further enhanced the ensemble model's utility by providing interpretable visualizations that indicate the regions of interest in the medical images. Our study showed that the Grad-CAM heatmaps overlapped with the actual rib fracture locations identified by radiologists, demonstrating the model's ability to accurately localize rib fractures. This feature significantly adds to the clinical utility of the model, aiding radiologists in the diagnostic process.

A similar approach has been adopted by Koh *et al.* [32], who used a deep learning model to diagnose COVID-19 and highlight significant features in chest X-ray images that correlate with the presence of the virus. They utilized the Grad-CAM technique for visualization, allowing for rapid and accurate diagnosis, especially in settings where traditional testing methods may be lacking. While their approach shows promising results, our study extends this concept by employing an ensemble of CNN models, offering a more robust and versatile diagnostic system. The use of Grad-CAM, which is an advanced version of CAM, further enhances the accuracy and interpretability of the localization results.

The success in leveraging machine learning for rib fracture diagnosis, as demonstrated in this study, has significant clinical implications. Early and accurate diagnosis of rib fractures is crucial for patient management and treatment planning. By employing our ensemble model, healthcare professionals can achieve more timely and precise diagnoses, leading to better patient outcomes. Moreover, the Grad-CAM visualizations can serve as a valuable tool for radiologists to quickly identify rib fractures in medical images, aiding their decision-making process. The ensemble model's ability to produce accurate and interpretable results holds great potential for integration into clinical workflows, supporting radiologists' works and improving the quality of care.

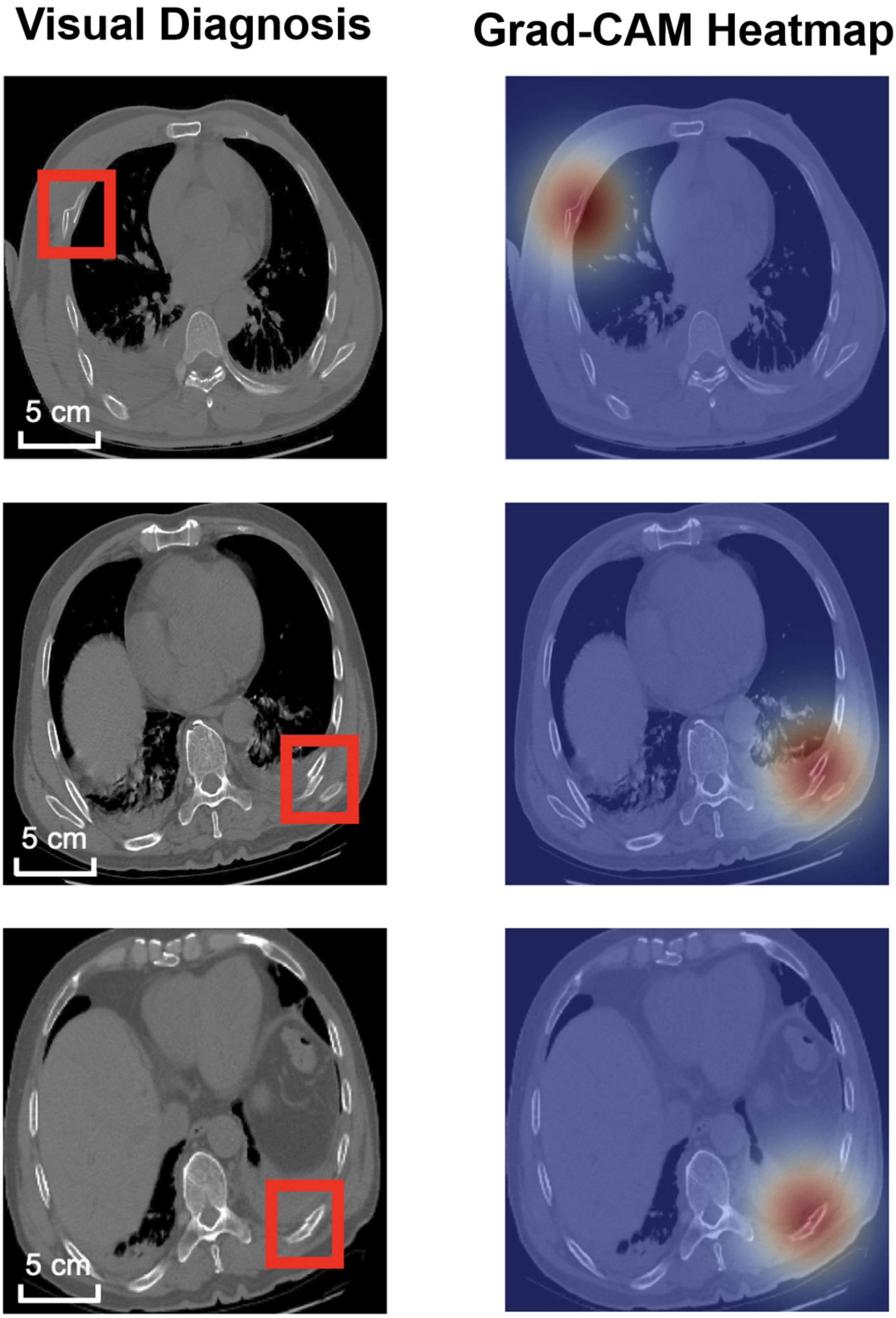

Fig. 6. Grad-CAM heatmap visualization of rib fractures in ensemble model. Highlighting rib fracture locations identified by radiologists with red boxes.

While our study presents promising results, several limitations warrant discussion. The dataset utilized was derived from a single institution, potentially subjecting to biases related to patient demographics, imaging equipment, and operator expertise. Such biases could limit the generalizability of the model to broader populations or different clinical settings. For instance, variations in imaging protocols and machine calibrations across different hospitals might result in images that differ significantly from those in our training set. Furthermore, the diverse variations related to rib fractures, including nuances in fracture size, location, and the presence of concurrent injuries, pose additional challenges for generalization. Thus, the performance of our model in real-world scenarios may differ, especially in underrepresented groups. Addressing these biases would require a multi-centric approach to data collection, to ensure compilation of a more representative dataset that includes varied patient demographics and imaging conditions. Another potential bias stems from the retrospective nature of our data collection, which could influence the selection of images and inadvertently affect the model's training. Prospective studies could help validate our findings and mitigate the risk of selection bias. The generalization of our model is further challenged by the inherent variability in clinical interpretations of rib fractures. Differences in radiologist experience and diagnostic practices may affect the ground truth used for training our models, complicating efforts to achieve consistent performance across diverse clinical environments. Future research should focus on addressing these limitations by using more diverse datasets and conducting clinical validation studies to ensure the models' applicability in real-world clinical scenarios.

Conclusions

In conclusion, our study demonstrates that the ensemble model, despite its seemingly complex setup compared to single models like ResNeXt-50, offers significant advantages in terms of robustness and generalizability across diverse clinical scenarios. This approach proves essential for real-world applications where data variability and unpredictable factors are common. By leveraging multiple CNN architectures, the ensemble model not only addresses overfitting but also enhances the reliability of fracture detection, which is critical for clinical decision-making. Our findings advocate for further research and development in this area, emphasizing the potential of ensemble models in advancing medical imaging diagnostics.

Availability of Data and Materials

The datasets used or analyzed during the current study are available from the corresponding author upon reasonable request.

Author Contributions

All authors contributed to this paper. KY, LW, HYC, PXL: investigation, methodology, data collection, and analysis; LW: writing—original draft; HYC: formal analysis; KY: conceptualization, supervision, writing—reviewing. All authors have been involved in revising it critically for important intellectual content. All authors gave final approval of the version to be published. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

The ethical considerations pertaining to this research have been rigorously examined and approved by the Ethics Committee of Ningbo No.9 Hospital (No.2023LIK04). The study involving the utilization of medical images and data adheres to the highest standards of ethical conduct and patient confidentiality. The approval from the Ethics Committee underscores our commitment to upholding the welfare and rights of all individuals involved in this study. We hereby declare that informed consent was obtained from the human subjects involved in this study for the publication of their images. The purpose and nature of the study, as well as the potential risks and benefits, were explained to the participants prior to obtaining their consent. This study was in accordance with the Declaration of Helsinki.

Acknowledgment

Not applicable.

Funding

This study was supported by Zhejiang Province Medical and Health Science and Technology Plan Project (No. 2024XY140).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Ziegler DW, Agarwal NN. The morbidity and mortality of rib fractures. The Journal of Trauma. 1994; 37: 975–979.

[2] Karmakar MK, Ho AMH. Acute pain management of patients with multiple fractured ribs. The Journal of Trauma. 2003; 54: 615–625.

[3] Sözen T, Özışık L, Başaran NÇ. An overview and management of osteoporosis. European Journal of Rheumatology. 2017; 4: 46–56.

[4] Lovell NC. Trauma analysis in paleopathology. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists. 1997; 104: 139–170.

[5] Koh DM, Burke S, Davies N, Padley SPG. Transthoracic US of the chest: clinical uses and applications. Radiographics: a Review Publication of the Radiological Society of North America, Inc. 2002; 22: e1.

[6] Mathis G. Thoraxsonography–Part I: Chest wall and pleura. Ultrasound in Medicine & Biology. 1997; 23: 1131–1139.

[7] Sharma N, Aggarwal LM. Automated medical image segmentation techniques. Journal of Medical Physics. 2010; 35: 3–14.

[8] Yarmolenko PS, Eranki A, Partanen A, Celik H, Kim A, Oetgen M, *et al*. Technical aspects of osteoid osteoma ablation in children using MR-guided high intensity focussed ultrasound. International Journal of Hyperthermia: the Official Journal of European Society for Hyperthermic Oncology, North American Hyperthermia Group. 2018; 34: 49–58.

[9] Raptis CA, Mellnick VM, Raptis DA, Kitchin D, Fowler KJ, Lubner

M, *et al*. Imaging of trauma in the pregnant patient. Radiographics: a Review Publication of the Radiological Society of North America, Inc. 2014; 34: 748–763.

[10] Jin L, Yang J, Kuang K, Ni B, Gao Y, Sun Y, *et al*. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. EBioMedicine. 2020; 62: 103106.

[11] Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging: JMRI. 2019; 49: 939–954.

[12] Desai KT, Befano B, Xue Z, Kelly H, Campos NG, Egemen D, *et al*. The development of "automated visual evaluation" for cervical cancer screening: The promise and challenges in adapting deep-learning for clinical testing: Interdisciplinary principles of automated visual evaluation in cervical screening. International Journal of Cancer. 2022; 150: 741–752.

[13] Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database: the Journal of Biological Databases and Curation. 2020; 2020: baaa010.

[14] Singh R, Homayounieh F, Vining R, Digumarthy SR, Kalra MKR, Singh F, *et al*. The value in artificial intelligence. Value-based Radiology: A Practical Approach. 2020; 35–49.

[15] Bukkuri A, Andor N, Darcy IK. Applications of topological data analysis in oncology. Frontiers in Artificial Intelligence. 2021; 4: 659037.

[16] Silva CF, von Stackelberg O, Kauczor HU. Value-based Radiology: A Practical Approach. Springer: Cham, Switzerland. 2019.

[17] Daghigh V, Lacy TE, Daghigh H, Gu G, Baghaei KT, Horstemeyer MF, *et al*. Machine learning predictions on fracture toughness of multiscale bio-nano-composites. Journal of Reinforced Plastics and Composites. 2020; 39: 587–598.

[18] Balcıoğlu HE, Seçkin AÇ. Comparison of machine learning methods and finite element analysis on the fracture behavior of polymer composites. Archive of Applied Mechanics. 2021; 91: 223–239.

[19] Huang YJ, Liu W, Wang X, Fang Q, Wang R, Wang Y, *et al*. Rectifying Supporting Regions With Mixed and Active Supervision for Rib Fracture Recognition. IEEE Transactions on Medical Imaging. 2020; 39: 3843–3854.

[20] Wang YT, Zhang X, Liu XS. Machine learning approaches to rock fracture mechanics problems: Mode-I fracture toughness determination. Engineering Fracture Mechanics. 2021; 253: 107890.

[21] Meng XH, Wu DJ, Wang Z, Ma XL, Dong XM, Liu AE, *et al*. A fully automated rib fracture detection system on chest CT images and its impact on radiologist performance. Skeletal Radiology. 2021; 50: 1821–1828.

[22] Zhang J, Li Z, Yan S, Cao H, Liu J, Wei D. An Algorithm for Automatic Rib Fracture Recognition Combined with nnU-Net and DenseNet. Evidence-based Complementary and Alternative Medicine: ECAM. 2022; 2022: 5841451.

[23] Niiya A, Murakami K, Kobayashi R, Sekimoto A, Saeki M, Toyofuku K, *et al*. Development of an artificial intelligence-assisted computed tomography diagnosis technology for rib fracture and evaluation of its clinical usefulness. Scientific Reports. 2022; 12: 8363.

[24] Handayani VW, Yudianto A, MAR MS, Rulaningtyas R, Rasyad Caesarardhi M. Classification of Indonesian adult forensic gender using cephalometric radiography with VGG16 and VGG19: a Preliminary research. Acta Odontologica Scandinavica. 2024; 83: 308–316.

[25] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 4700–4708.

[26] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; 2818–2826.

[27] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. PMLR. 2019; 6105–6114.

[28] Brock A, De S, Smith SL, Simonyan K. High-performance large-scale image recognition without normalization. International Conference on Machine Learning. PMLR. 2021; 1059–1071.

[29] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision. 2020; 128: 336–359.

[30] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5987–5995). IEEE (Institute of Electrical and Electronics Engineers): Honolulu, HI, USA. 2017.

[31] Solorzano L, Robertson S, Acs B, Hartman J, Rantalainen M. Ensemble-based deep learning improves detection of invasive breast cancer in routine histopathology images. Heliyon. 2024; 10: e32892.

[32] Koh SJT, Nafea M, Nugroho H. Towards edge devices implementation: deep learning model with visualization for COVID-19 prediction from chest X-ray. Advances in Computational Intelligence. 2022; 2: 33.