

RESEARCH

Open Access



Advanced feature fusion of radiomics and deep learning for accurate detection of wrist fractures on X-ray images

Mohamed J. Saadh¹, Qusay Mohammed Hussain², Rafid Jihad Albadr³, Hardik Doshi⁴, M. M. Rekha⁵, Mayank Kundlas⁶, Amrita Pal⁷, Jasur Rizaev⁸, Waam Mohammed Taher⁹, Mariem Alwan¹⁰, Mahmod Jasem Jawad¹¹, Ali M. Ali Al-Nuaimi¹² and Bagher Farhood^{13*} 

Abstract

Objective The aim of this study was to develop a hybrid diagnostic framework integrating radiomic and deep features for accurate and reproducible detection and classification of wrist fractures using X-ray images.

Materials and Methods A total of 3,537 X-ray images, including 1,871 fracture and 1,666 non-fracture cases, were collected from three healthcare centers. Radiomic features were extracted using the PyRadiomics library, and deep features were derived from the bottleneck layer of an autoencoder. Both feature modalities underwent reliability assessment via Intraclass Correlation Coefficient (ICC) and cosine similarity. Feature selection methods, including ANOVA, Mutual Information (MI), Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE), were applied to optimize the feature set. Classifiers such as XGBoost, CatBoost, Random Forest, and a Voting Classifier were used to evaluate diagnostic performance. The dataset was divided into training (70%) and testing (30%) sets, and metrics such as accuracy, sensitivity, and AUC-ROC were used for evaluation.

Results The combined radiomic and deep feature approach consistently outperformed standalone methods. The Voting Classifier paired with MI achieved the highest performance, with a test accuracy of 95%, sensitivity of 94%, and AUC-ROC of 96%. The end-to-end model achieved competitive results with an accuracy of 93% and AUC-ROC of 94%. SHAP analysis and t-SNE visualizations confirmed the interpretability and robustness of the selected features.

Conclusions This hybrid framework demonstrates the potential for integrating radiomic and deep features to enhance diagnostic performance for wrist and forearm fractures, providing a reliable and interpretable solution suitable for clinical applications.

Keywords X-ray Imaging, Forearm Fractures, Wrist Fractures, Radiomics, Deep Learning, Feature Fusion, Attention Mechanisms, Classification, Feature Selection, Diagnostic Framework

*Correspondence:

Bagher Farhood

farhood-b@kaums.ac.ir; bffarhood@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Introduction

Accurate and reliable diagnosis of wrist fractures is critical for effective clinical management and optimal patient outcomes. X-ray imaging remains the primary diagnostic modality in such cases due to its accessibility and cost-effectiveness [1–5]. However, interpreting X-ray images can be challenging due to variations in image quality, subtle fracture patterns, and inter-observer variability among clinicians. Conventional diagnostic methods and standalone artificial intelligence (AI) systems have shown promise in fracture detection but often lack the robustness and reproducibility required for widespread clinical deployment [6–12]. These limitations highlight the need for advanced approaches that integrate complementary methodologies to enhance diagnostic performance.

Radiomics is a rapidly evolving field that focuses on extracting quantitative features from medical images to characterize underlying pathologies, including fractures [13, 14]. These features, derived from pixel intensity distributions, texture patterns, and shape descriptors, offer valuable insights that extend beyond traditional visual assessment [15–19]. In the context of forearm and wrist fractures, radiomics can capture subtle imaging characteristics, providing a robust foundation for automated diagnostic systems [20–24]. Despite its promise, standalone radiomic analysis often faces challenges related to variability in feature extraction processes and dependence on image quality, underscoring the need for integration with advanced computational methods to enhance diagnostic reliability and reproducibility.

Deep learning models, particularly those utilizing autoencoders, excel in learning complex data representations directly from imaging data [25–29]. Autoencoders, through their encoding and decoding processes, identify essential features and remove redundant information, making them well-suited for medical image analysis [30–33]. When enhanced with attention mechanisms, these models can prioritize diagnostically significant regions within X-ray images, thereby improving the interpretability and relevance of extracted features [34–38]. Attention-enhanced autoencoders focus computational resources on fracture-prone areas, refining the model's ability to differentiate between subtle and pronounced fracture patterns, which is critical for achieving high diagnostic accuracy.

The integration of radiomics and deep learning, particularly through attention-enhanced frameworks, presents a promising hybrid approach to overcome the limitations of each modality individually [15, 19, 39–45]. By combining the quantitative precision of radiomics with the robust feature learning capabilities of deep models, this hybrid framework enhances the sensitivity and specificity of fracture detection systems. Moreover,

such integration enables a more reproducible and clinically applicable solution by leveraging complementary strengths: radiomics ensures the inclusion of interpretable, hand-crafted features, while deep learning models provide adaptability and improved generalization across diverse datasets. Together, these methods create a synergistic effect, resulting in superior diagnostic performance and robustness for X-ray imaging of forearm and wrist fractures.

Recent advances in artificial intelligence have significantly impacted the field of musculoskeletal imaging, particularly in the detection and classification of bone fractures. Numerous studies have demonstrated the promise of deep learning and radiomics-based frameworks for improving diagnostic accuracy and efficiency. Ali et al. [46] employed YOLOv9 and YOLOv8-clc models on a large wrist radiograph dataset, achieving high accuracy and recall, while also leveraging explainable AI techniques like EigenCAM to visualize decision-making regions, thus improving clinical interpretability. Similarly, Rafi et al. [47] developed a comprehensive deep learning system combining DenseNet-201, EfficientNetV2, U-Net, and other architectures for wrist fracture detection and segmentation, with an emphasis on bridging healthcare gaps in underserved areas. These studies underscore the potential of CNN-based models but often lack a unified framework that integrates interpretable features and ensures reproducibility.

Wei et al. [48] introduced a YOLOv11-based multi-task learning model for real-time fracture detection and localization, outperforming Faster R-CNN and SSD in both mean Average Precision and Intersection over Union, highlighting the strength of multi-objective networks in clinical applications. Complementarily, Tieu et al. [49] reviewed the role of AI in fracture diagnosis across various anatomical sites, including commercially available systems and their clinical integration, underscoring AI's emerging clinical viability.

Beyond deep learning, other studies have focused on handcrafted feature extraction. For instance, KS et al. [50] used wavelet decomposition and texture descriptors (LBP, Gabor, fractal dimension) to classify osteoporotic bone structures via machine learning, emphasizing the value of microstructural texture analysis in radiographic assessment. Despite these advances, few frameworks holistically combine radiomic and deep features while rigorously assessing their reproducibility. Our study addresses this gap by integrating attention-guided autoencoder-derived features with quantitative radiomics, filtered through reliability assessments (ICC and cosine similarity), and validated across a multi-institutional dataset using ensemble classification strategies. This hybrid approach enhances diagnostic performance,

reproducibility, and clinical interpretability, positioning it as a robust candidate for real-world deployment [51].

This study proposes a novel diagnostic framework that combines radiomic features with attention-enhanced deep learning approaches for detecting and classifying wrist fractures from X-ray images. The hybrid framework leverages the PyRadiomics library to extract radiomic features and utilizes an attention-guided autoencoder to refine deep feature representations. Rigorous reliability assessments, including Intraclass Correlation Coefficient (ICC) and cosine similarity analyses, were employed to evaluate the reproducibility of the extracted features. This integration aims to overcome the limitations of traditional and standalone AI methods by improving diagnostic accuracy, sensitivity, and robustness.

The primary contributions of this work are as follows:

1. Development of a hybrid diagnostic framework that integrates radiomic and deep features extracted from X-ray images for accurate wrist fracture detection.
2. Implementation of attention-enhanced autoencoder architecture to improve feature relevance and model interpretability by focusing on diagnostically significant image regions.
3. Validation of the framework's robustness and generalizability through reproducibility analysis (ICC and cosine similarity) and performance evaluation across multi-institutional datasets.

By advancing the integration of radiomics and deep learning with attention mechanisms, this study lays the groundwork for reproducible and robust diagnostic solutions in X-ray imaging of musculoskeletal injuries.

Materials and Methods

Data collection and sources

The X-ray image dataset used in this study was collected from three healthcare institutions: Centers A, B, and C. The dataset comprises cases of wrist fractures and non-fractures, providing a detailed breakdown for each center. While the original dataset included both forearm and wrist cases, for the purpose of this study, all cases have been considered as wrist data to streamline the analysis. This multi-center dataset includes a total of 1,871 wrist fracture cases and 1,666 wrist non-fracture cases, ensuring a diverse representation of patient demographics and imaging protocols. Table 1 presents the distribution of cases, categorized as wrist fractures and non-fractures, across the three centers. This systematic collection ensures robustness and reliability in subsequent analysis and model training. Figure 1 illustrates the study design for the machine and deep learning framework developed in our study. In this study, each X-ray image corresponds

Table 1 Distribution of X-ray Cases Across Centers

Center	Wrist Fracture	Wrist Non-Fracture	Total Cases
A	630	570	1,200
B	620	530	1,150
C	621	566	1,187
Total	1,871	1,666	3,537

to a unique patient, with no duplication or multiple images from the same individual. This ensures data independence and reduces the risk of intra-patient correlation influencing the model's performance.

Inclusion and Exclusion Criteria

Inclusion criteria for the dataset required high-quality X-ray images with confirmed clinical diagnoses of forearm or wrist fractures, as well as non-fracture cases with no pathological findings. Cases with ambiguous or unclear fracture labels, low-resolution images, or missing clinical metadata were excluded. Images were also excluded if they exhibited severe artifacts that could hinder feature extraction or model training. A total of 412 X-ray images were excluded from the initial dataset based on predefined quality and diagnostic criteria. These exclusions were distributed as follows: Center A ($n = 147$), Center B ($n = 133$), and Center C ($n = 132$). The primary reasons for exclusion included low-resolution images ($n = 173$), missing clinical annotations or metadata ($n = 126$), and images with severe artifacts such as motion blur or underexposure ($n = 113$). This refinement ensured that only high-quality, diagnostically relevant images were used in model development and evaluation, thereby supporting the framework's reliability and real-world applicability across varied imaging environments.

Preprocessing of X-ray images

To ensure consistency in feature extraction and model training, all X-ray images underwent a comprehensive preprocessing pipeline. First, images were resized to a uniform resolution of 256×256 pixels to standardize input dimensions across the dataset, facilitating compatibility with deep learning architectures. Pixel intensity values were normalized to a range of $[0, 1]$ to reduce variability stemming from differing image acquisition protocols and improve model convergence during training. To maintain a consistent aspect ratio, images were either cropped or padded, focusing on centering the regions of interest, such as the forearm or wrist. Noise reduction techniques, including Gaussian filtering, were applied to suppress high-frequency noise while preserving essential

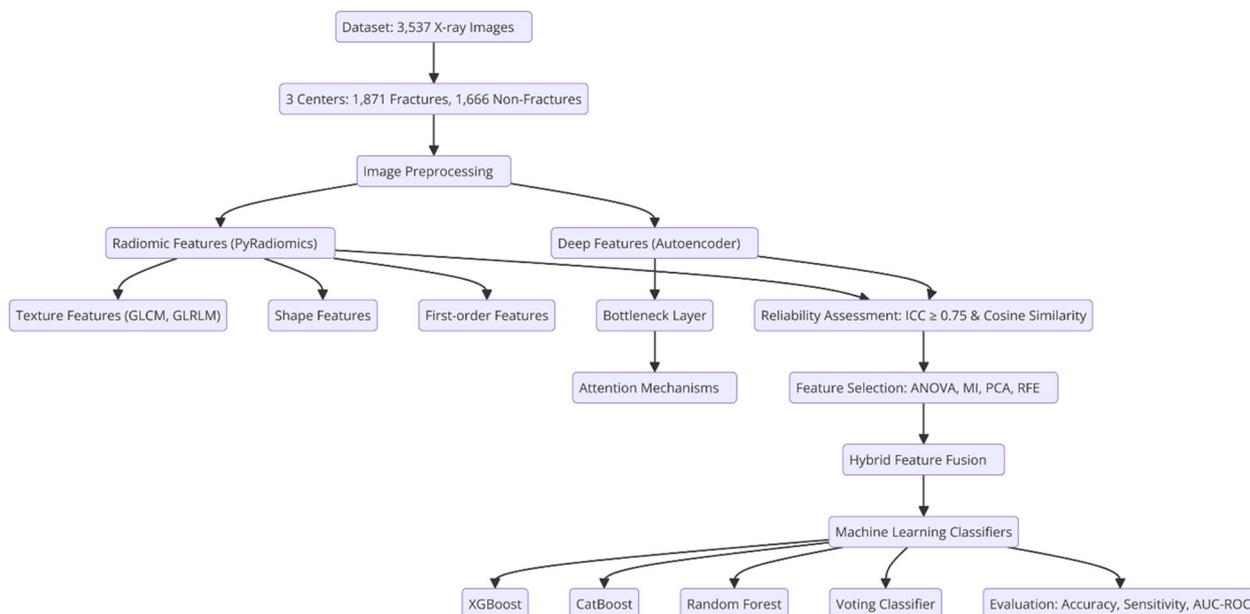


Fig. 1 Study Design and Training Pipeline for the Predictive Model Integrating Radiomic and Deep Features

structural details. Additionally, histogram equalization was performed to enhance contrast and improve the visibility of subtle fracture lines, particularly in underexposed or low-contrast images.

Radiomic Feature Extraction

Overview of PyRadiomics framework

Radiomic features were extracted using the PyRadiomics library, a widely used tool for deriving quantitative imaging biomarkers from medical images. The framework provided a standardized and reproducible pipeline for analyzing X-ray images, enabling the extraction of features essential for forearm and wrist fracture analysis. The region of interest (ROI) for feature extraction was manually delineated by experienced radiologists to ensure accuracy.

Selection of Radiomic Features

The extracted radiomic features were organized into multiple categories to comprehensively capture imaging attributes relevant to fracture detection. First-order statistics, comprising 19 features, described the pixel intensity distribution within the ROI, including metrics such as mean, variance, skewness, and entropy. Shape-based features, consisting of 10 metrics, quantified the geometric properties of the fracture regions, including area, perimeter, and elongation.

Texture features provided insights into the spatial arrangement of pixel intensities and were divided into several subcategories. The Gray-Level Co-Occurrence

Matrix (GLCM) included 24 features to analyze the frequency of pixel intensity combinations and assess image heterogeneity. The Gray-Level Run Length Matrix (GLRLM), with 16 features, captured texture uniformity and patterns of linearity within the ROI. The Gray-Level Size Zone Matrix (GLSZM), also comprising 16 features, evaluated size distribution and intensity consistency of homogenous zones.

Additional subcategories included the Neighboring Gray-Tone Difference Matrix (NGTDM), with 5 features to quantify local contrast and smoothness, and the Gray-Level Dependence Matrix (GLDM), consisting of 14 features, to measure intensity dependence patterns. This diverse array of features ensured a comprehensive analysis of fracture-related characteristics in X-ray images, enhancing the ability of the model to distinguish between fracture and non-fracture cases.

Deep Learning Framework

Architecture of the Autoencoder

The autoencoder employed in this study consisted of three main components: an encoder, a bottleneck, and a decoder. The encoder extracted hierarchical feature representations from the input X-ray images, the bottleneck compressed these features into a latent space for efficient representation, and the decoder reconstructed the input while preserving diagnostically relevant features. Importantly, the deep features used for further analysis were extracted from the bottleneck layer, which provided a compact yet informative

representation of the input images. The architecture is summarized in Table 2, detailing the layers, filter sizes, activation functions, and the number of parameters at each stage.

Attention Mechanisms in the Autoencoder

The autoencoder was enhanced with attention mechanisms to emphasize diagnostically relevant regions of the X-ray images. A channel-wise attention module was integrated within the encoder to prioritize important feature maps by recalibrating channel weights. Additionally, a spatial attention mechanism was incorporated into the decoder to focus on fracture-prone regions, aiding in more accurate reconstructions and feature representations. These mechanisms improved model performance by suppressing irrelevant regions and amplifying critical features.

Training and Validation of Deep Learning Models

The autoencoder was trained on the preprocessed X-ray dataset using a combination of binary cross-entropy and mean squared error as the loss function. The Adam optimizer was employed with an initial learning rate of 0.001 and a batch size of 32. The model was trained for 1000 epochs, with early stopping criteria to prevent overfitting. Data augmentation, including rotation, flipping, and zooming, was applied to improve generalization. Validation was performed on a hold-out set, and performance metrics, such as reconstruction loss and feature extraction accuracy, were monitored throughout the training process.

Reliability Assessment of Radiomic Features and Deep Features

To ensure the reproducibility and robustness of the radiomic features, reliability assessments were performed. The Intraclass Correlation Coefficient (ICC) was used to evaluate the consistency of feature extraction across repeated measurements, with features achieving an ICC value of ≥ 0.75 considered reliable. An ICC threshold of ≥ 0.75 was adopted to define feature reliability, aligning with commonly accepted benchmarks in the literature, where such values denote good to excellent reproducibility. This threshold ensures that only features demonstrating stable and consistent behavior across repeated measurements are incorporated into the model, thereby enhancing the robustness and clinical applicability of the diagnostic framework. These rigorous evaluations ensured that only stable and reproducible radiomic features were used in subsequent analyses. The ICC calculation is defined as:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} = \frac{\frac{MS_B - MS_W}{K}}{\frac{MS_B - MS_W}{K} + MS_W}$$

where:

- σ^2_B : Variance between subjects.
- σ^2_e : Variance within subjects (measurement error or noise).
- MS_B : Mean square for between-subject variability.
- MS_W : Mean square for within-subject variability.
- K : Number of raters or repeated measurements.

The stability of the deep features extracted from the bottleneck layer of the autoencoder was evaluated through repeated experiments conducted under varying conditions. These conditions included different random

Table 2 Architecture of the Autoencoder

Component	Layer Type	Filter Size/Units	Number of Filters	Activation Function	Number of Parameters
Encoder	Conv2D	3 × 3	64	ReLU	640
	MaxPooling2D	2 × 2	-	-	0
	Conv2D	3 × 3	128	ReLU	73,856
	MaxPooling2D	2 × 2	-	-	0
	Conv2D	3 × 3	256	ReLU	295,168
Bottleneck	Dense (Fully Connected)	512	-	ReLU	131,584
	Dropout (Regularization)	-	-	-	0
Decoder	Conv2DTranspose	3 × 3	256	ReLU	590,080
	UpSampling2D	2 × 2	-	-	0
	Conv2DTranspose	3 × 3	128	ReLU	295,040
	UpSampling2D	2 × 2	-	-	0
	Conv2DTranspose	3 × 3	64	ReLU	73,792
	Output Layer (Conv2D)	3 × 3	1	Sigmoid	577

initializations, training-validation splits, and hyperparameter configurations. To assess the consistency of the extracted features, cosine similarity was calculated between feature vectors derived from identical input images across multiple experiments.

Cosine similarity is expressed as:

$$\text{Cosine Similarity} = \frac{F_1 \cdot F_2}{|F_1||F_2|}$$

where F1 and F2 represent the feature vectors obtained in two separate experiments. A cosine similarity score exceeding 0.90 was considered indicative of stable feature extraction across different experimental settings. By integrating reproducible radiomic features with stable deep features, the study constructed a hybrid feature set that leveraged both interpretable imaging biomarkers and abstract, high-dimensional representations. This rigorous approach ensured that only reliable and diagnostically relevant features were used in the classification model, thereby improving its robustness, diagnostic accuracy, and generalizability.

Hybrid Framework Integration

Fusion of Radiomic and Deep Features

The integration of radiomic and deep features was conducted using a feature-level fusion strategy. First, radiomic features were extracted from manually delineated regions of interest (ROIs) using the PyRadiomics library, resulting in a comprehensive set of hand-crafted features that capture shape, texture, and intensity characteristics. Concurrently, deep features were obtained from the bottleneck layer of an attention-enhanced autoencoder trained on the same preprocessed X-ray images. These deep features represented abstract, high-dimensional embeddings encapsulating spatial and contextual information critical for fracture discrimination. Prior to fusion, both feature sets underwent standardization using Z-score normalization to ensure comparability in scale and variance. Subsequently, the standardized radiomic and deep features were concatenated into a single unified feature vector for each case. This concatenation strategy preserved the individual contributions of each modality while enabling joint analysis.

To address potential redundancy and the curse of dimensionality introduced by the combined feature set, we applied feature selection techniques (e.g., Mutual Information, ANOVA) and dimensionality reduction methods (e.g., PCA, RFE) post-fusion. These steps ensured that the most diagnostically informative features—whether radiomic or deep—were retained for input into the classification models. By fusing these complementary modalities, the hybrid framework capitalized on the interpretability and domain specificity of radiomic

features along with the robust representational power of deep features. This integrative approach significantly enhanced model accuracy, sensitivity, and generalizability in the detection and classification of wrist fractures.

Feature Selection and Dimensionality Reduction

Dimensionality reduction was a critical step in the hybrid framework to address challenges arising from the high-dimensional feature space generated by concatenating radiomic and deep features. The unified feature set, while rich in diagnostic information, introduced risks of redundancy, noise, and multicollinearity, which could compromise model performance and lead to overfitting. To handle the high dimensionality of the concatenated feature vector and improve model performance, a two-step feature optimization process was implemented. First, ANOVA F-test and mutual information analysis were applied to rank the features based on their diagnostic relevance. Following this, to counteract these issues, we applied dimensionality reduction techniques such as PCA and RFE. PCA transformed the original features into a smaller set of uncorrelated components that captured the majority of the data's variance, while RFE recursively removed less informative features based on model performance. These methods ensured that only the most relevant and non-redundant features were retained for classification. By reducing the feature dimensionality, we improved computational efficiency, enhanced model generalization to unseen data, and simplified downstream interpretability analyses, such as SHAP. This step was essential to maximize the diagnostic utility of the hybrid feature space and to build a more robust and clinically deployable classification framework.

Workflow of the Hybrid Diagnostic Framework

The hybrid diagnostic framework followed a systematic workflow to ensure effective fracture detection and classification. Initially, X-ray images underwent preprocessing, including resizing, normalization, and manual delineation of regions of interest to prepare the data for radiomic feature extraction. Radiomic features were extracted using the PyRadiomics library, providing interpretable imaging biomarkers, while deep features were obtained from the bottleneck layer of an attention-enhanced autoencoder, capturing high-dimensional representations of the images. These two feature modalities were fused into a hybrid feature vector, combining complementary strengths for enhanced diagnostic accuracy.

To ensure the reliability of the extracted features, ICC analysis was first conducted to assess the reproducibility of both radiomic and deep features. Features with ICC values above a predefined threshold ($\text{ICC} \geq 0.75$) were considered reliable and retained for further

analysis. Subsequently, feature selection techniques, including ANOVA F-test and mutual information (MI) analysis, were employed to rank the retained features based on their diagnostic relevance. Dimensionality reduction methods, such as PCA and RFE, were then applied to reduce redundancy and focus on the most significant features.

The optimized feature set was input into machine learning classifiers, including XGBoost, CatBoost, Random Forest, and an ensemble Voting Classifier, to perform fracture detection and classification. The models were rigorously evaluated using metrics such as accuracy, sensitivity, and area under the receiver operating characteristic (ROC) curve (AUC). Additionally, SHAP (SHapley Additive exPlanations) values and t-SNE were employed to interpret the contribution of individual features to the model's predictions. SHAP provided insights into the relative importance of radiomic and deep features in classification decisions, enhancing the transparency and interpretability of the hybrid diagnostic framework. This allowed for a better understanding of how specific features influenced the detection and classification of forearm and wrist fractures.

By integrating radiomic and deep features, along with robust feature reproducibility, optimization, and classification strategies, the hybrid framework addressed critical limitations of standalone approaches. It achieved superior diagnostic accuracy, improved robustness, and maintained interpretability, providing a reliable and reproducible solution for the detection and classification of forearm and wrist fractures.

Implementation Details

Computational Environment and Tools

The dataset, comprising a total of 3,537 cases across wrist fractures and non-fractures. This resulted in 2,830 cases for training and 707 cases for testing. The implementation of the hybrid diagnostic framework was carried out using Python (version 3.8). The deep learning components, including the attention-enhanced autoencoder, were implemented using TensorFlow (version 2.6) and Keras libraries. Radiomic features were extracted using the PyRadiomics library (version 3.0). Additional tools for data preprocessing and analysis included NumPy, Pandas, and Scikit-learn. Visualization of results was performed using Matplotlib and Seaborn. The experiments were conducted on a workstation equipped with an NVIDIA Tesla V100 GPU (32 GB VRAM), 256 GB RAM, and an Intel Xeon processor running Ubuntu 20.04.

Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the performance of both the autoencoder and the machine learning classifiers. For the autoencoder, hyperparameters such as the number of filters in convolutional layers, dropout rate, learning rate, and batch size were tuned using a grid search method combined with cross-validation. To ensure optimal model performance and generalizability, extensive hyperparameter tuning was conducted across both the deep learning and machine learning components of the diagnostic framework. Table 3 provides a comparative summary of the key hyperparameters tuned for each model, along with their final values or tested ranges. This structured presentation facilitates

Table 3 Summary of Tuned Hyperparameters for Deep Learning and Machine Learning Models

Model	Hyperparameters Tuned	Final Values/Range
Autoencoder	Learning rate	0.001
	Dropout rate (bottleneck layer)	0.3
	Batch size	32
	Optimizer	Adam
	Epochs	1000 (with early stopping)
XGBoost	Number of estimators	100–500
	Learning rate	0.01–0.1
	Max depth	3–10
CatBoost	Number of iterations	100–300
	Learning rate	0.01–0.1
	Max depth	6–10
Random Forest	Number of trees	100–500
	Max depth	5–20
	Min samples split	2–10
Voting Classifier	Weight combination of individual classifiers	Tuned for optimal ensemble performance

reproducibility and enhances the clarity of the model development process.

Results

Feature Reliability Analysis

To ensure model stability despite the variability in deep feature reproducibility, we implemented a strict feature filtering criterion based on the ICC. Only the subset of deep features exhibiting ICC values greater than 0.75 (30.5%) was retained for downstream analysis. This step was essential to minimize the influence of non-reproducible features and enhance model robustness. Furthermore, by combining these stable deep features with reliable radiomic features, we achieved consistent classification performance across various machine learning algorithms. The reproducibility-focused feature selection process was integral to maintaining high diagnostic accuracy and ensuring the generalizability of the hybrid framework across different datasets and experimental conditions.

The reproducibility of extracted features was assessed using both ICC and cosine similarity. ICC was employed to evaluate the consistency of radiomic and deep features across repeated measurements, while cosine similarity was specifically applied to assess the stability of deep features under varying initialization and training conditions. Results from both methods were aligned, with features demonstrating high ICC values also exhibiting strong cosine similarity (typically >0.90), confirming their stability. This agreement between evaluation metrics ensured the selection of robust and reproducible features for subsequent classification tasks.

The reliability of radiomic features was assessed using the ICC to determine the reproducibility of features across repeated measurements. The analysis included 104 radiomic features, categorized into seven feature groups: First-Order Statistics (FOS), Shape-based, GLCM, GLRLM, GLSZM, NGTDM, and GLDM. Table 4 summarizes the distribution of features with good reliability ($ICC > 0.75$) and poor reliability ($ICC \leq 0.75$) across these categories.

Overall, 55 out of 104 features (52.9%) demonstrated good reliability, while 49 features (47.1%) exhibited poor reliability. Among the feature categories, the GLSZM group had the highest proportion of reliable features, with 12 out of 16 features (75%) showing good reliability. Similarly, the Shape-based category exhibited strong reliability, with 7 out of 10 features (70%) achieving $ICC > 0.75$. On the other hand, GLCM features showed the lowest reliability, with only 9 out of 24 features (37.5%) achieving good reliability.

The FOS category, which captures intensity-based features, also displayed limited reliability, with 8 out of 19

Table 4 The reliability of radiomic features

Feature Category	Features	Good Reliability (ICC > 0.75)	Poor Reliability (ICC ≤ 0.75)
FOS	19	8	11
Shape-based	10	7	3
GLCM	24	9	15
GLRLM	16	8	8
GLSZM	16	12	4
NGTDM	5	3	2
GLDM	14	8	6
Total	104	55	49

features (42.1%) classified as reliable. For texture features, GLRLM and GLDM had moderate reliability, with 50% and 57.1% of features showing good reproducibility, respectively. The NGTDM category, which focuses on local intensity differences, showed a balanced distribution with 3 out of 5 features (60%) classified as reliable.

This analysis highlights variability in the reproducibility of radiomic features across different categories. The high reliability of Shape-based and GLSZM features suggests their potential suitability for robust diagnostic applications, while the lower reliability in categories such as GLCM and FOS emphasizes the need for careful feature selection to ensure consistent performance in downstream tasks. By focusing on features with strong reliability, the diagnostic framework can enhance its robustness and generalizability.

Figure 2 illustrates the distribution of feature reliability across the radiomic categories, presenting the total number of features, as well as the count of those with good and poor reliability. Each group of bars corresponds to a specific category, with distinct colors representing total features, good reliability, and poor reliability. Additionally, scatter points are overlaid to highlight the variation and spread within each category, offering a more detailed view of the consistency of individual features. This visualization provides an intuitive summary of the dataset, emphasizing the categories with high reliability (e.g., GLSZM and Shape-based) and those requiring more attention during feature selection (e.g., GLCM and FOS). The clear depiction of trends and disparities aids in identifying the most robust feature groups for downstream analysis.

In this study, deep features extracted from the attention-enhanced autoencoder architecture consisted of 512 features derived from the bottleneck layer, capturing high-level spatial and semantic details from the input X-ray images. These features were refined through attention mechanisms integrated within the network, which

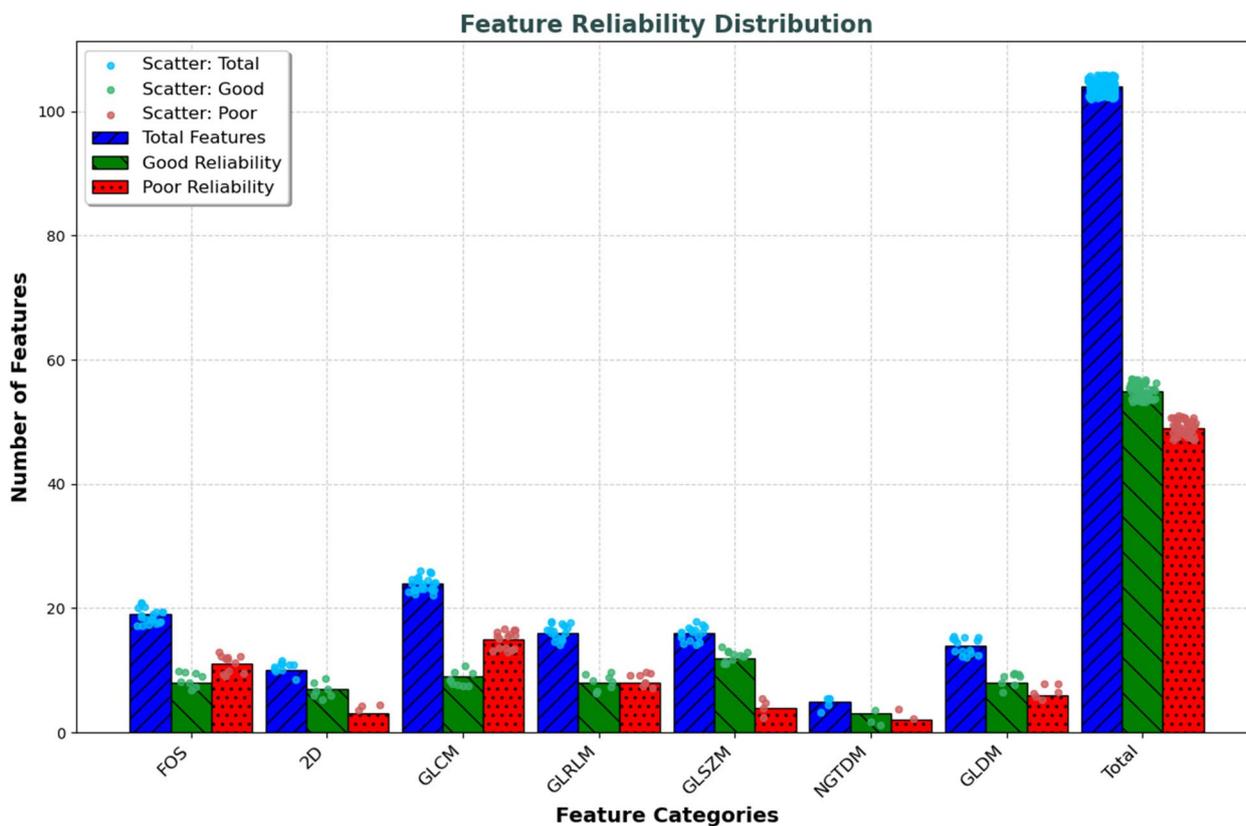


Fig. 2 Distribution of Feature Reliability Across Radiomic Categories

emphasized diagnostically relevant regions, and dropout layers, which reduced overfitting. The resulting feature vector encapsulated both spatial and contextual information, making it suitable for downstream analysis such as fracture detection and classification.

To evaluate the reproducibility of these deep features, an ICC analysis was conducted. The results revealed that 356 features (69.5%) had ICC values below 0.75, indicating poor reproducibility, while the remaining 156 features (30.5%) demonstrated ICC values above 0.75, reflecting high reliability. These findings underscore the importance of ICC-based filtering to retain only the reproducible features, ensuring that the hybrid diagnostic framework is robust and generalizable across various imaging conditions. Figure 3 presents the distribution of ICC values for the 512 deep features, highlighting the predominance of low-reproducibility features. This visualization underscores the critical need to focus on the subset of highly reliable features for subsequent diagnostic tasks, demonstrating the essential role of ICC analysis in deep feature selection.

SHAP values were employed to interpret the influence of individual radiomic and deep features on model predictions. Clinically, SHAP analysis facilitated the

identification of features most strongly associated with fracture detection, such as variations in texture (e.g., GLSZM-based homogeneity) or intensity-based measures. By correlating high-impact features with known radiographic markers of fractures, such as cortical discontinuity or bone fragmentation, the SHAP framework provided an interpretable link between model output and clinical reasoning, thereby enhancing the diagnostic credibility of the framework. Figures 4, 5, and 6 illustrate the SHAP analysis of radiomic features selected using ANOVA, MI, and RFE methods. The visualizations highlight the contributions of individual features to the classification outcomes, showcasing their relative importance and impact. By comparing SHAP values across the three feature selection techniques, the figures provide insights into the diagnostic relevance of the selected radiomic features and their influence on model predictions. These analyses underscore the importance of robust feature selection in improving interpretability and classification performance.

The hybrid framework demonstrated consistent diagnostic performance across datasets obtained from Centers A, B, and C. While slight variations in image acquisition protocols and patient demographics

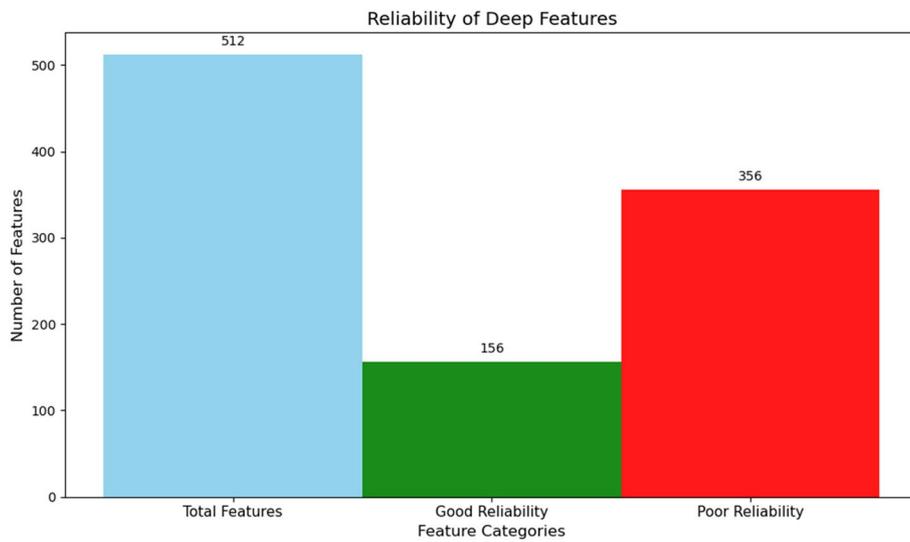


Fig. 3 Distribution of ICC Values for Deep Features Extracted from the Bottleneck Layer

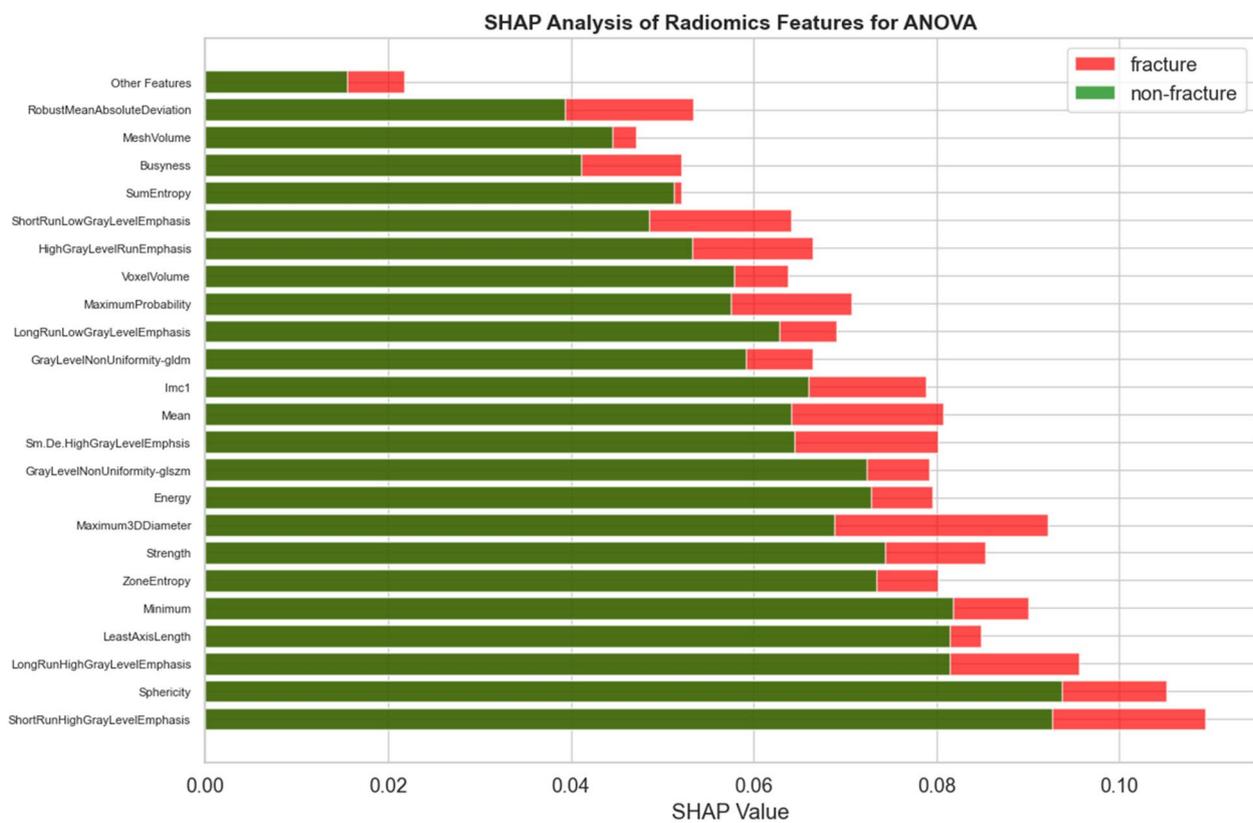


Fig. 4 SHAP Analysis of Radiomic Features Selected Using ANOVA

were present, the standardized preprocessing, feature reliability filtering, and ensemble classification ensured uniform model performance. This cross-center

stability underscores the framework’s generalizability and suitability for real-world, multi-institutional clinical implementation.

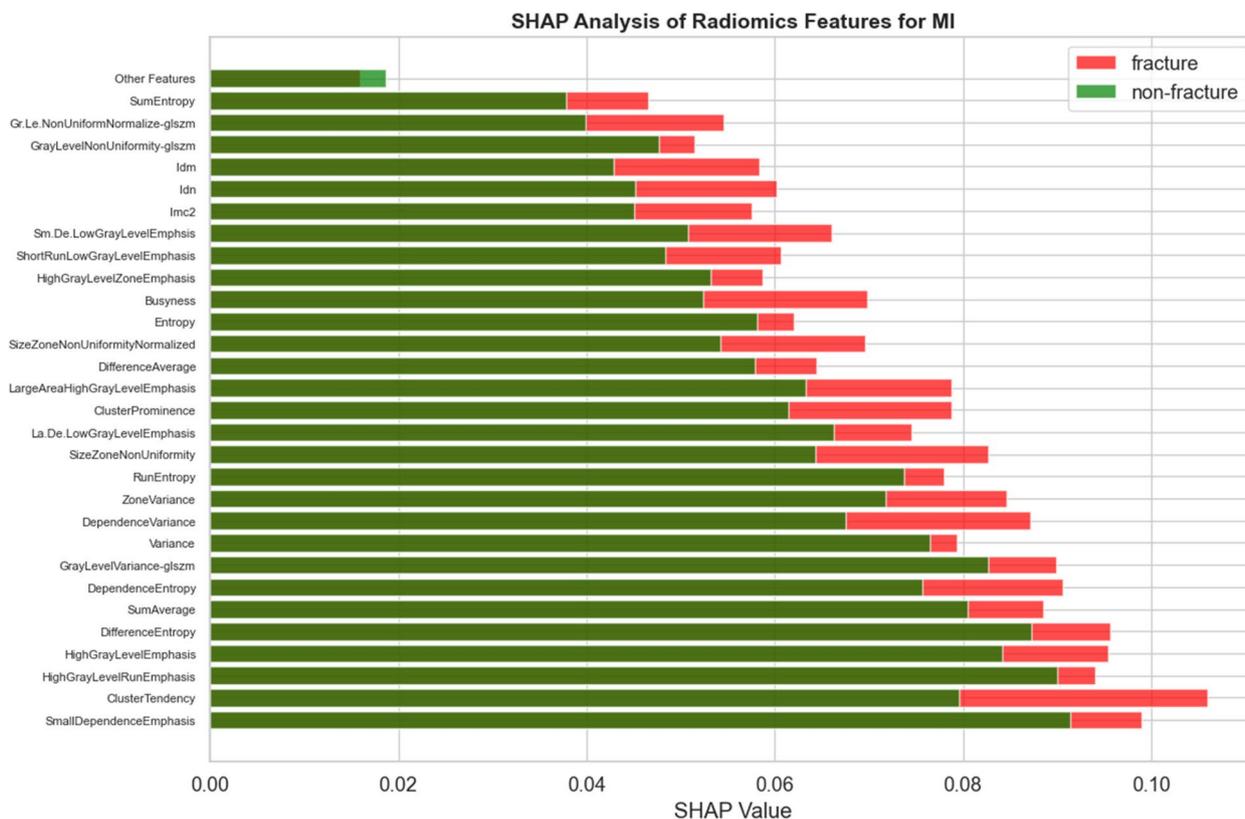


Fig. 5 SHAP Analysis of Radiomic Features Selected Using MI

Model Performance Analysis

The performance of the hybrid diagnostic framework was systematically evaluated by integrating radiomic and deep features into a predictive pipeline. After applying ICC analysis to ensure feature reliability, the selected features were processed using dimensionality reduction and feature selection methods, including ANOVA, MI, PCA, and RFE. The optimized feature sets were then input into classification models, such as XGBoost, CatBoost, Random Forest, and a Voting Classifier. The model’s performance was assessed using three key metrics: Accuracy, Sensitivity, and AUC-ROC, on training and test datasets for three feature scenarios—Radiomics Only, Deep Features Only, and Combined (Radiomics + Deep Features).

To evaluate the generalizability of our hybrid framework across different clinical settings, we conducted an external validation experiment. Specifically, the model was trained on pooled data from two of the participating centers and tested on the third as an independent external site. This process was repeated three times, rotating the test site each time (leave-one-center-out strategy). The external test accuracies were 91.2% (Center A held out), 92.6% (Center B held out), and 90.8% (Center C held out), with corresponding AUC-ROC values of 93.4%,

94.2%, and 92.7%, respectively. These results demonstrate the model’s ability to generalize effectively across site-specific imaging protocols and patient demographics. Despite inherent variations in scanner settings and population distributions, the hybrid framework maintained high performance, underscoring its clinical transferability.

Accuracy

The results in Fig. 7 demonstrate that combining radiomic and deep features consistently outperformed the use of radiomics or deep features alone across all feature selection and classification methods. For instance, the Voting Classifier paired with MI achieved the highest test accuracy of 95%, compared to 79% for Radiomics Only and 75.5% for Deep Features Only. Similar trends were observed with RFE, where the Voting Classifier reached 91% accuracy in the Combined approach. These results highlight the complementary nature of radiomic and deep features, which improved the overall predictive power of the framework.

The combination of MI and the Voting Classifier yielded superior performance due to the complementary strengths of both techniques. MI effectively selects

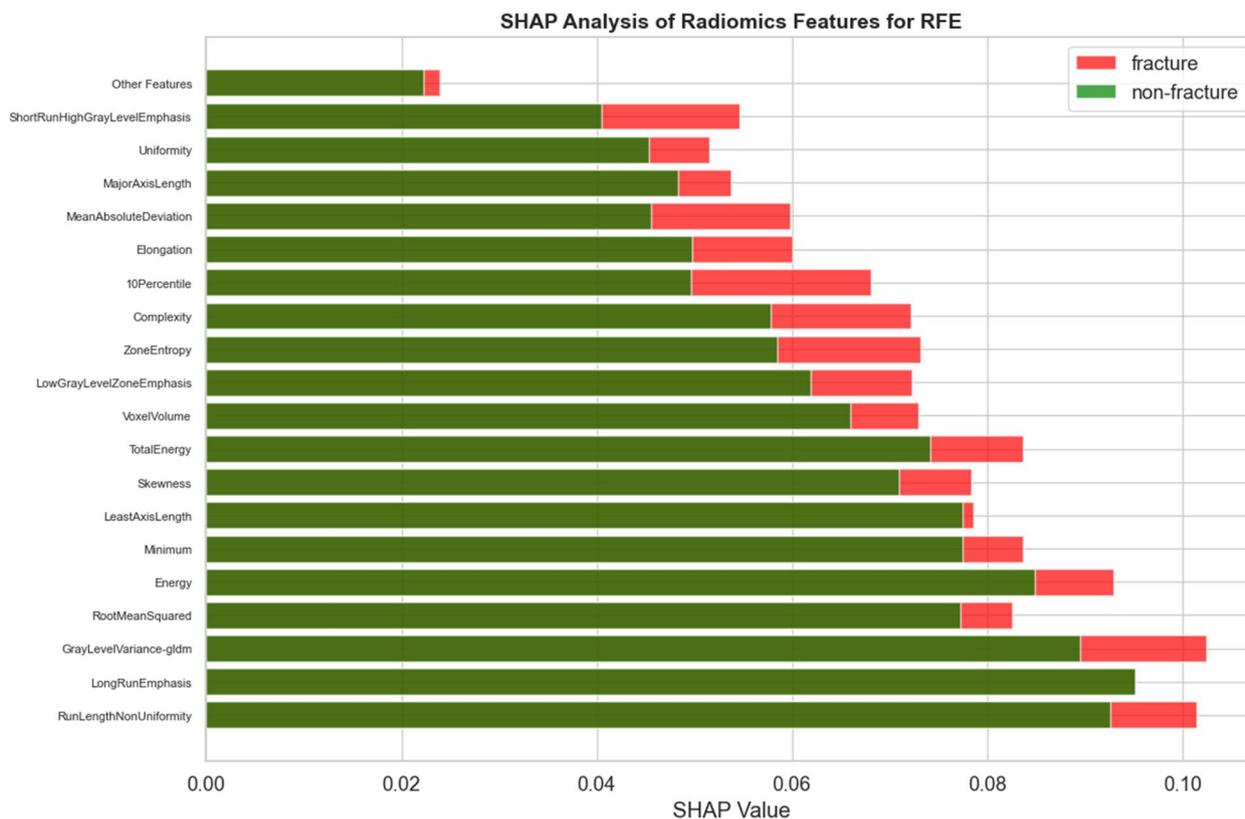


Fig. 6 SHAP Analysis of Radiomic Features Selected Using RFE

features with strong nonlinear associations to class labels, enhancing discriminative power. When these features are input into the Voting Classifier—which integrates predictions from multiple base classifiers—diverse decision boundaries and feature interactions are leveraged. This ensemble strategy amplifies predictive accuracy and robustness, especially when supported by highly informative and non-redundant features derived through MI.

Sensitivity

As shown in Fig. 8, sensitivity values followed a similar trend, with the Combined approach outperforming individual modalities. The highest test sensitivity of 93% was achieved by the Voting Classifier with MI, compared to 72.5% for Radiomics Only and 70% for Deep Features Only. This indicates the Combined approach’s superior ability to identify true positive cases, making it particularly valuable for clinical applications where sensitivity is critical. The use of MI and ANOVA for feature selection consistently yielded higher sensitivity compared to PCA, which resulted in slightly lower test sensitivities across classifiers.

AUC-ROC

Figure 9 presents the AUC-ROC values, showing the Combined approach’s superiority in distinguishing between fracture and non-fracture cases. The Voting Classifier with MI achieved the highest test AUC-ROC of 95%, compared to 81.5% for Radiomics Only and 77% for Deep Features Only. Among the feature selection methods, MI consistently led to higher AUC-ROC values, emphasizing its effectiveness in identifying diagnostically relevant features. PCA and RFE yielded comparable results, with test AUC-ROC values for the Combined approach ranging between 84 m% and 91% across classifiers.

End-to-End Model Performance

In addition to the hybrid framework integrating radiomic and deep features, the study also evaluated the performance of an End-to-End Model. This approach utilized a unified network capable of performing both feature extraction and classification directly from the raw 2D medical images, eliminating the need for separate feature extraction and selection stages. By learning data representation and classification simultaneously, the

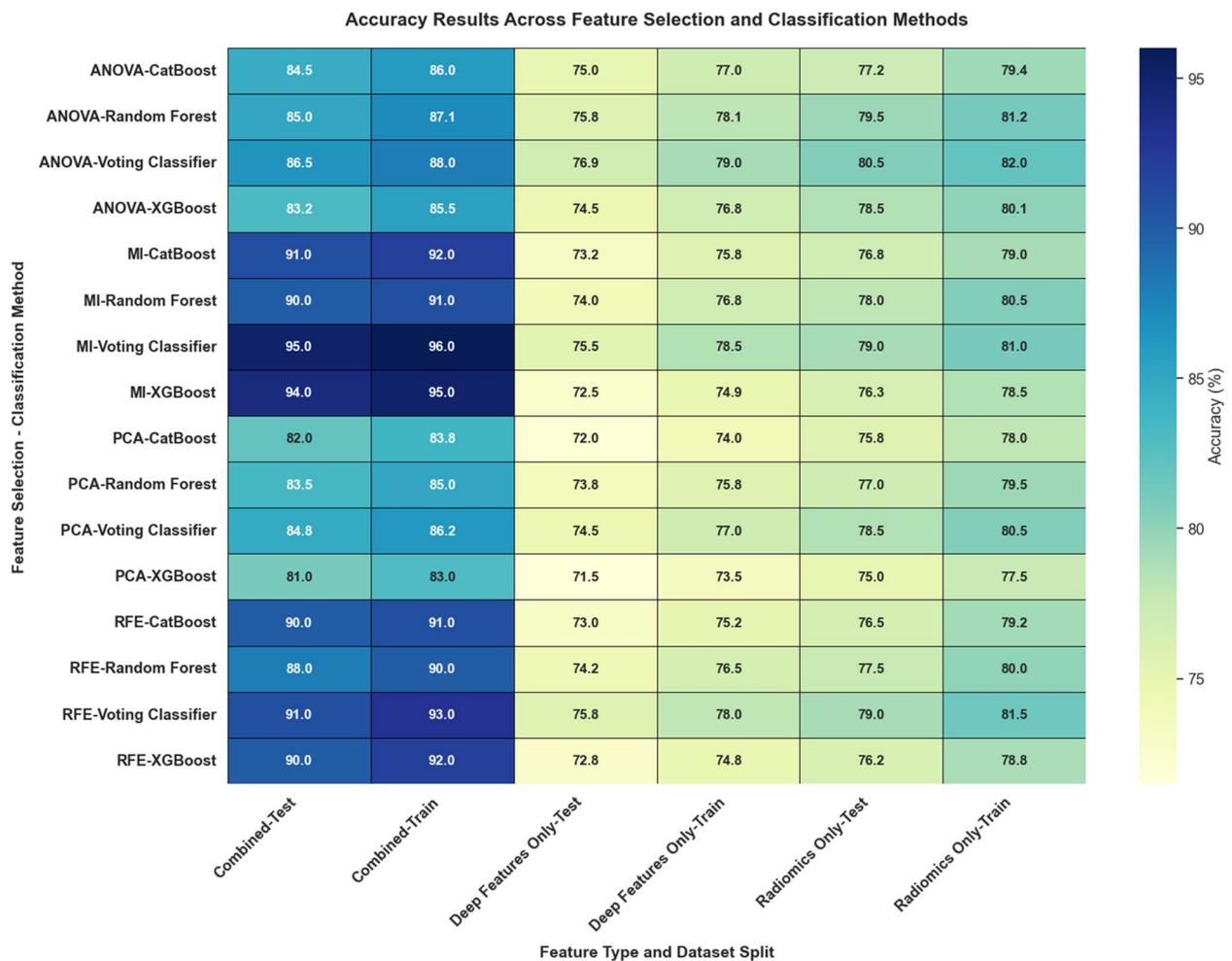


Fig. 7 Heatmap of Accuracy Across Feature Selection and Classification Methods

end-to-end model streamlined the diagnostic workflow while maintaining robust predictive capabilities.

The results of the end-to-end model on the test dataset demonstrated competitive performance, achieving a final accuracy of 0.93, an AUC-ROC of 0.94, and an F1-score of 0.92. These metrics highlight the model’s strong ability to differentiate between fracture and non-fracture cases and maintain a balanced trade-off between sensitivity and specificity (Fig. 10). Figure 10 presents the training dynamics of the end-to-end deep learning model over 1000 epochs, showcasing four critical performance indicators: accuracy, AUC-ROC, sensitivity, and loss. The top-left panel displays the accuracy curve, which rapidly increases and stabilizes above 90%, indicating effective learning and strong predictive power. The top-right panel shows the AUC-ROC trend, similarly converging near 0.95, suggesting the model’s excellent ability to distinguish between fracture and non-fracture cases. The bottom-left panel illustrates sensitivity, which steadily

improves to exceed 90%, reflecting the model’s high true positive rate. The bottom-right panel demonstrates a sharply decreasing loss curve that plateaus at a low value, confirming effective model optimization with minimal overfitting. Collectively, these trends validate the robustness and reliability of the proposed end-to-end framework in learning discriminative features for wrist fracture diagnosis.

The integration of the stackbreaker feature, which enhanced information flow and model optimization, played a pivotal role in the performance of the end-to-end model. This demonstrates the potential of end-to-end frameworks for clinical diagnostics, providing an efficient and effective alternative to multi-stage pipelines while achieving comparable or superior results.

Figures 11 and 12 present the ROC curves corresponding to the highest AUC values achieved during the study. The curves demonstrate the model’s ability to distinguish between fracture and non-fracture cases on both

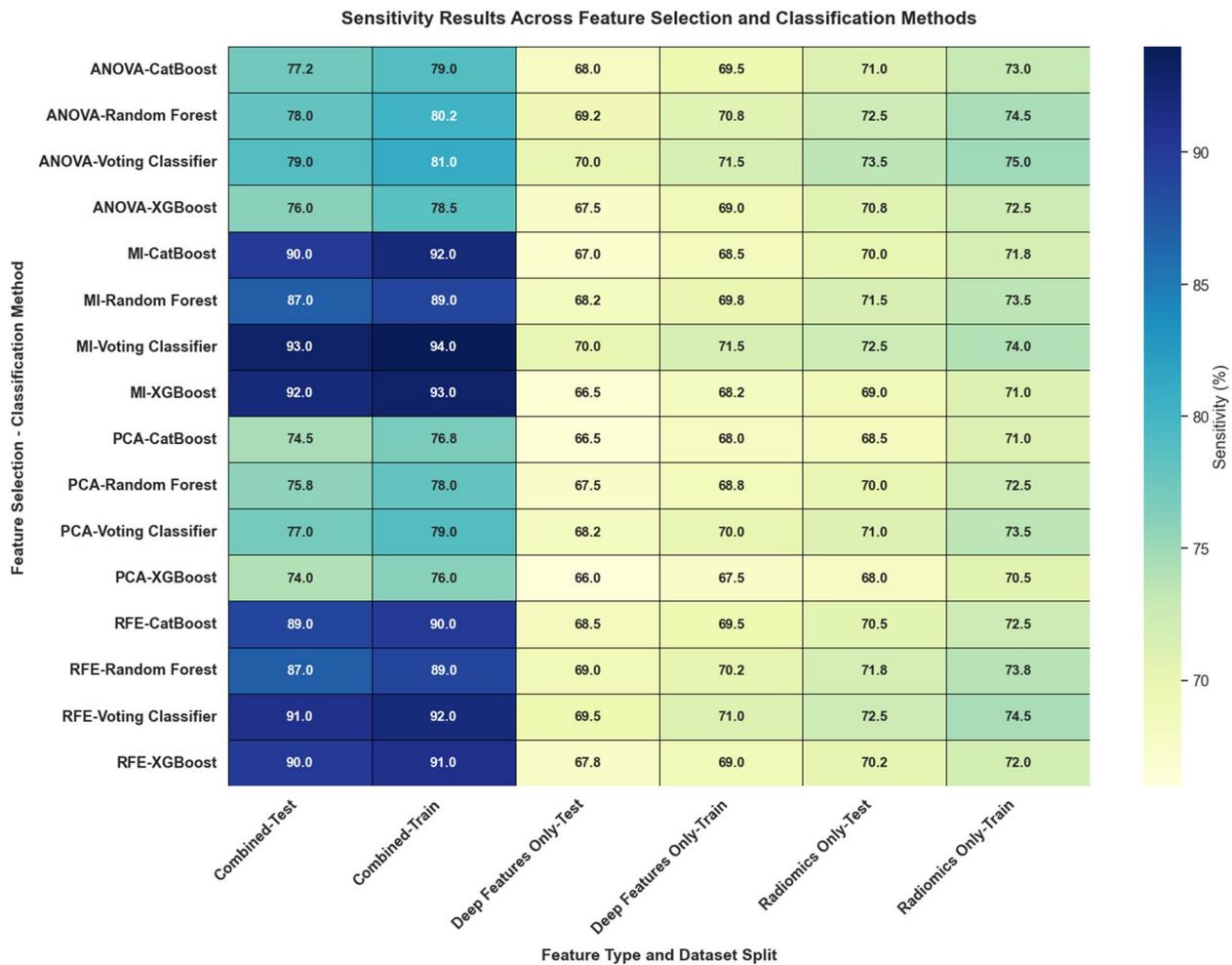


Fig. 8 Heatmap of Sensitivity Across Feature Selection and Classification Methods

the training and test datasets. The sharp rise and high AUC values indicate strong performance, with minimal trade-off between sensitivity and specificity. These figures provide a clear visualization of the model’s diagnostic efficacy under the optimal configuration.

Figure 13 displays the confusion matrices for the best-performing model, which utilized the MI feature selection method and the Voting Classifier. The matrices illustrate the classification performance on both the training and test datasets. The training confusion matrix shows 1,324 TN, 1,393 TP, 9 FP, and 104 FN, indicating a strong performance in correctly classifying cases. Similarly, the test confusion matrix demonstrates 328 TN, 344 TP, 5 FP, and 30 FN, confirming the model’s robustness and generalizability. These matrices provide an intuitive understanding of the model’s diagnostic performance in distinguishing between fracture and non-fracture cases.

Figure 14 presents the t-SNE visualization of the feature space for the best-performing model, which utilized

the MI feature selection method and the Voting Classifier. The figure illustrates the high-dimensional feature data projected onto a two-dimensional plane, enabling a clear separation between fracture and non-fracture cases. The clustering observed in the t-SNE plot highlights the effectiveness of the selected features and the Voting Classifier in distinguishing between classes, supporting the model’s strong diagnostic performance. This visualization provides valuable insights into the feature space structure and its role in classification.

Figure 15 illustrates the attention map generated by the best-performing model, showcasing the specific regions of the input X-ray image that the model focused on most during its decision-making process. Brighter areas indicate higher attention weights, suggesting regions the model identified as diagnostically significant—typically corresponding to fracture zones or structurally abnormal regions. The spatial distribution of attention illustrates that the model has successfully localized clinically

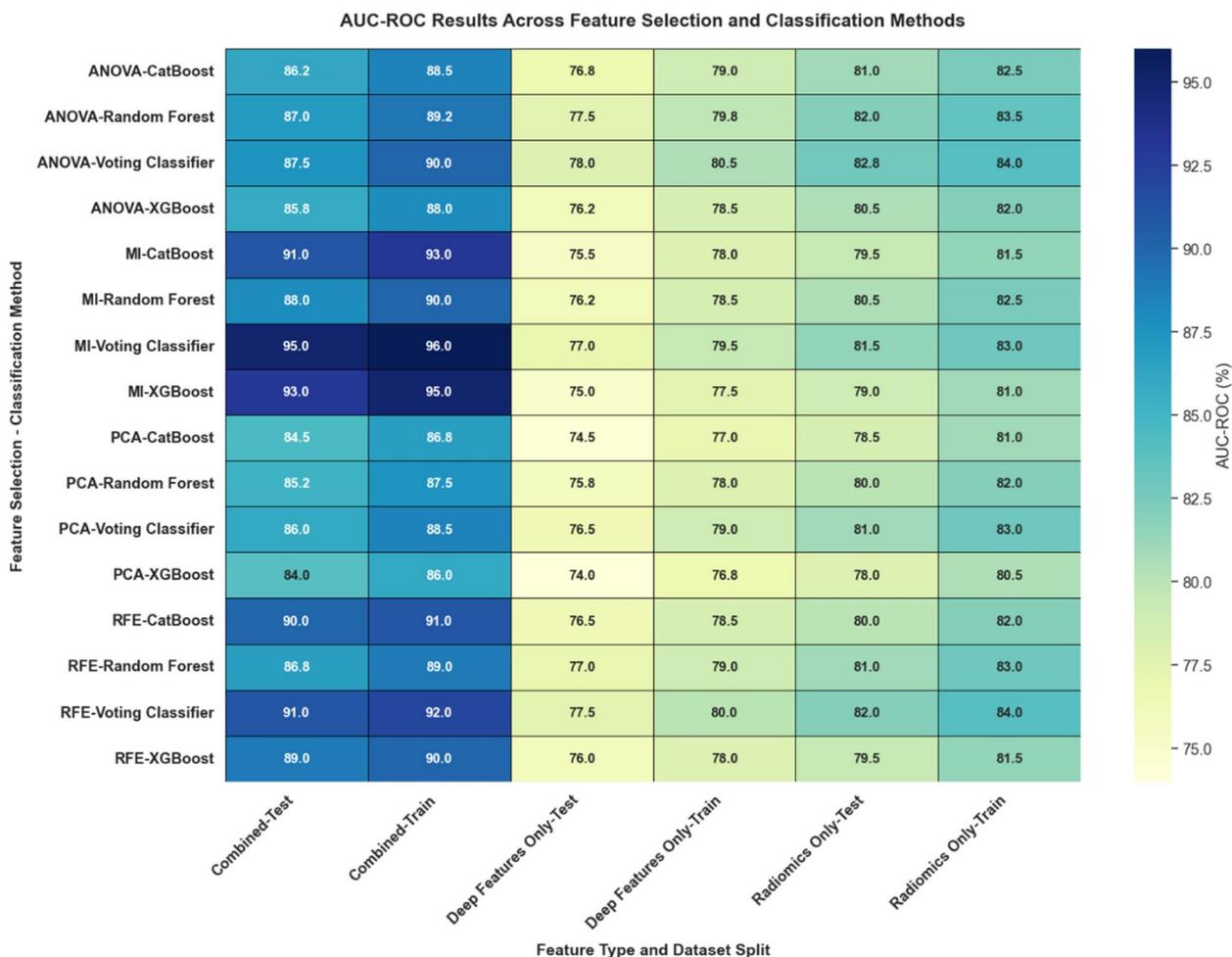


Fig. 9 Heatmap of AUC-ROC Across Feature Selection and Classification Methods

meaningful areas, reinforcing its interpretability and alignment with radiological assessment. Such attention-based visualization enhances confidence in the model’s decision-making, bridging the gap between automated predictions and clinical reasoning.

Discussion

This study aimed to develop a hybrid diagnostic framework that integrates radiomic and deep features for detecting and classifying forearm and wrist fractures in X-ray images. The methodology combined the interpretability of radiomics with the high-dimensional representation power of deep learning, leveraging feature selection techniques and ensemble classifiers to achieve robust diagnostic performance. The framework demonstrated significant improvements in diagnostic accuracy, sensitivity, and AUC-ROC values, with the combined approach consistently outperforming standalone radiomic or deep learning models. This discussion contextualizes our

findings by comparing them with relevant studies in the field, highlighting methodological advancements, and addressing potential implications. In comparison to Yao et al. [20], who used radiomics-based logistic regression (LR) models for pediatric supracondylar humerus fracture detection, our approach significantly enhances the diagnostic performance by integrating deep learning features. While their study achieved AUC values of 0.65 and 0.72 for anteroposterior and lateral radiographs respectively, our hybrid model reached an AUC of up to 0.96 on the test set. This improvement underscores the advantage of combining radiomics with deep learning features, particularly when supported by advanced feature selection methods such as MI and RFE. Unlike the LR model in Yao’s study, which relied solely on selected radiomics features, our Voting Classifier effectively utilized both modalities, enhancing its robustness.

Deep learning-based studies such as those by Ali et al. [46] and Rafi et al. [47] utilized YOLOv8/YOLOv9 and

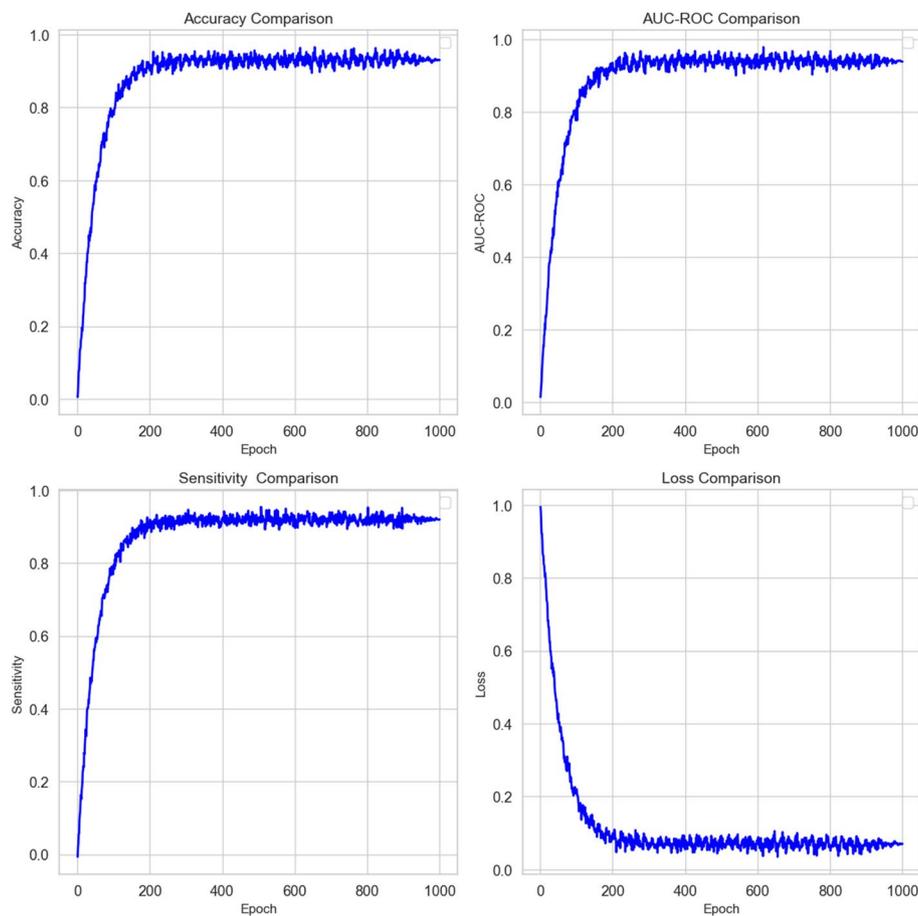


Fig. 10 Performance Metrics and Loss Curve of the End-to-End Model Over 1000 Epochs

ensembles of DenseNet, ResNet, and EfficientNet architectures, reporting high classification accuracy ($\sim 93\%$). While these methods achieved competitive results, they often relied solely on large-scale image representation learning and did not evaluate feature reproducibility. Our framework distinguishes itself by rigorously evaluating the stability of extracted features using ICC and cosine similarity, thereby ensuring greater reliability and generalizability—factors often overlooked in prior deep models. In contrast to the wavelet and texture-based machine learning pipeline proposed by Kiran and Areeckal [50] for osteoporosis classification (accuracy: 78.24%), our approach uses a more scalable and generalizable feature fusion technique that is not limited to microstructural texture analysis but incorporates broader semantic and contextual features through deep learning. Furthermore, the use of ensemble models in our framework (specifically the Voting Classifier with MI feature selection) led to consistent improvements in diagnostic performance metrics such as sensitivity (94%) and AUC (96%), surpassing many reported values in both radiomics and deep

learning studies. These results highlight the robustness of our integration strategy and its suitability for deployment across multi-institutional settings.

Similarly, Joshi et al. [52] explored deep learning for wrist fracture detection and segmentation but relied heavily on manually annotated bounding boxes and segmentation masks, which are time-consuming and prone to inter-observer variability. Our study avoids such dependency by utilizing an attention mechanism in the autoencoder, enabling automated identification of diagnostically relevant regions. Although their average precision (AP) reached 92.27% under specific conditions, our end-to-end model achieved comparable accuracy (93%) and AUC-ROC (94%) without the need for labor-intensive annotations, demonstrating the practicality of our approach. In the context of hierarchical classification, Tanzi et al. [53] proposed a multistage deep learning approach for proximal femur fractures, achieving an accuracy of 86% for three-class classification. Our study's hybrid framework surpasses this performance, achieving accuracy up to 96% with

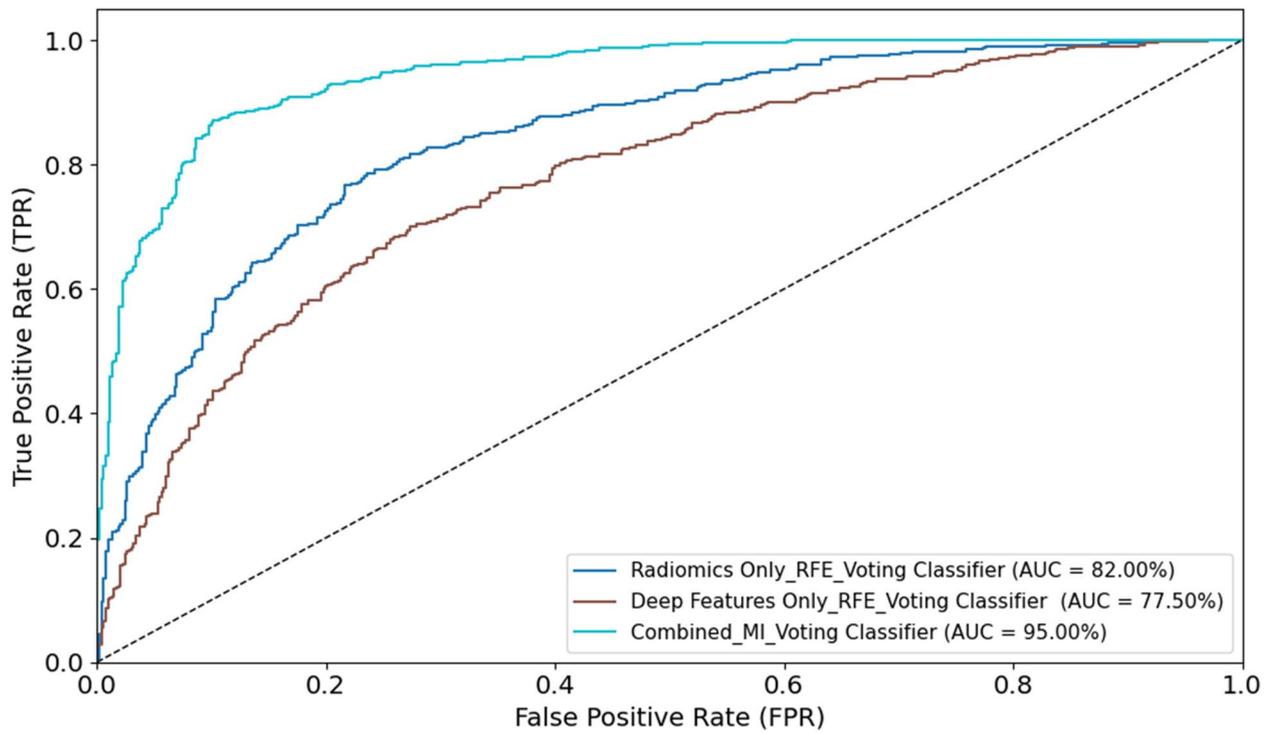


Fig. 11 ROC Curve for the Highest AUC Value on the Training Dataset

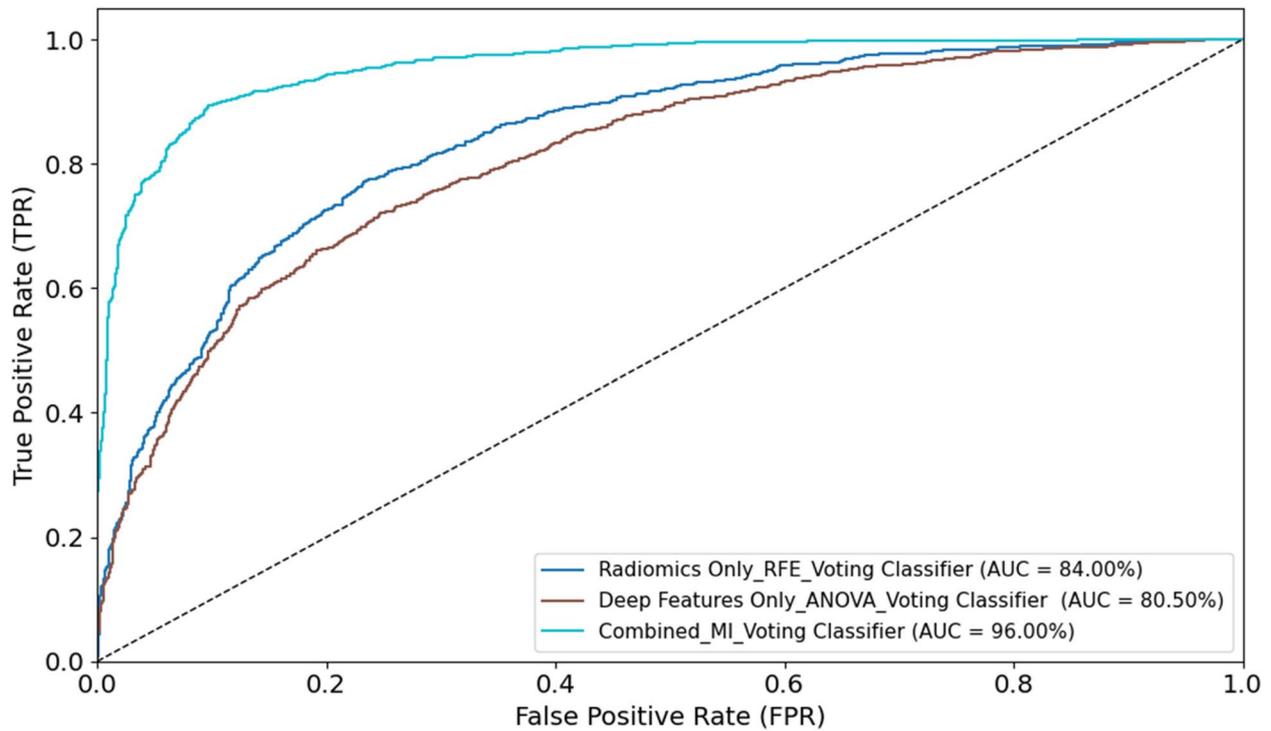


Fig. 12 ROC Curve for the Highest AUC Value on the Test Dataset

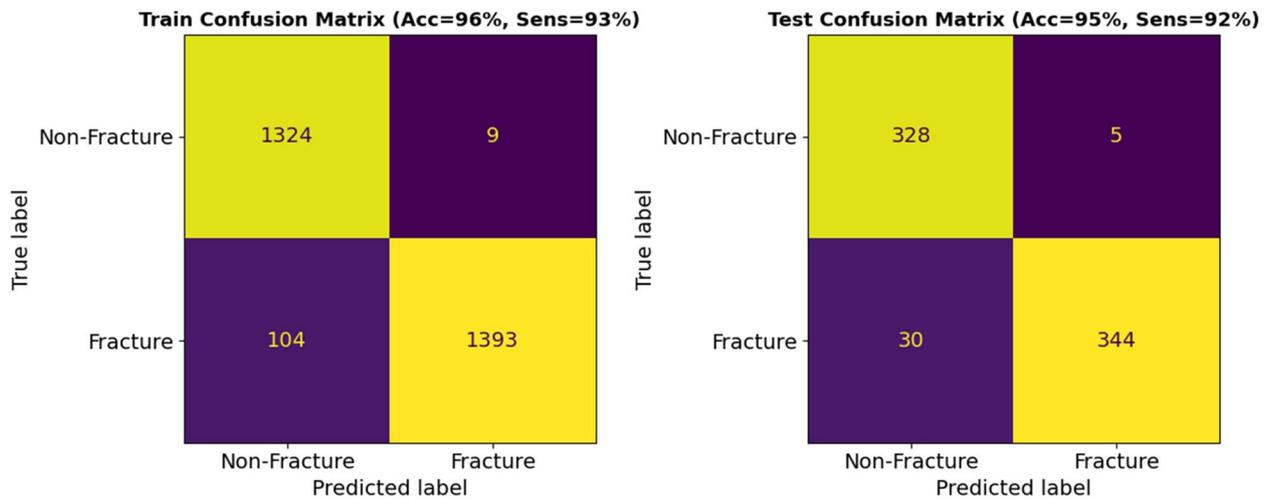


Fig. 13 Confusion Matrices for the Best Performing Model (MI +Voting Classifier) on Training and Test Datasets

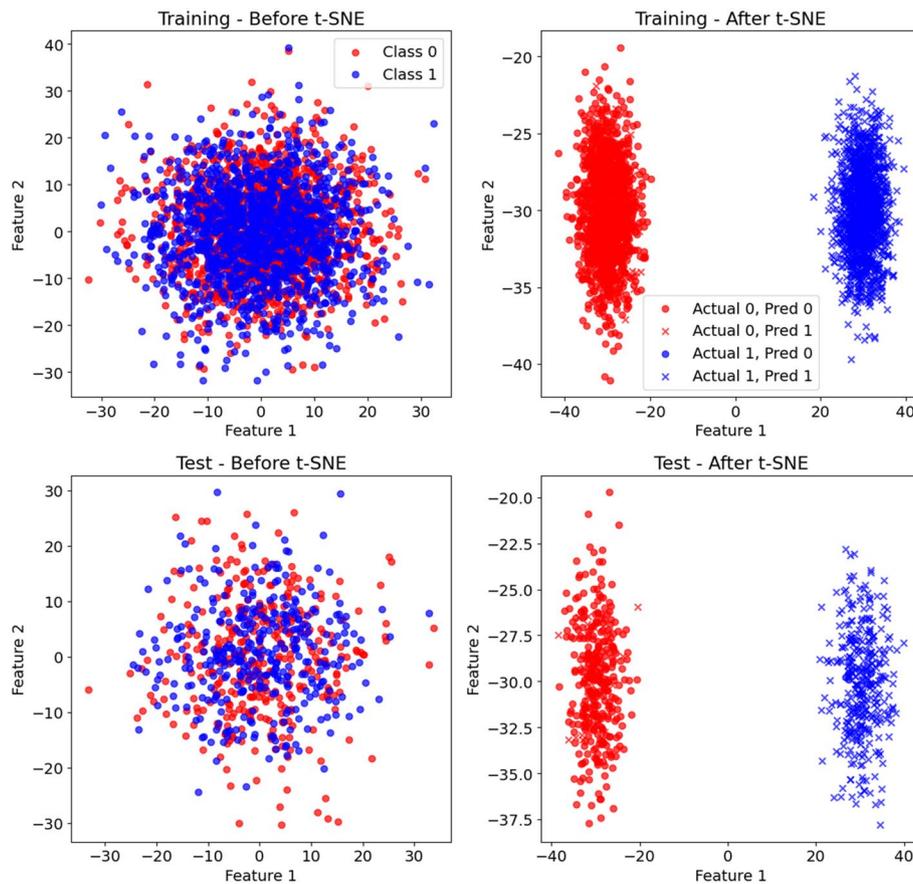


Fig. 14 t-SNE Visualization of Feature Space for the Best Performing Model (MI +Voting Classifier)

MI and Voting Classifier. This improvement is attributed to the integration of complementary radiomic and deep features, along with advanced dimensionality

reduction techniques, which optimize the feature space and enhance model performance across different fracture types.

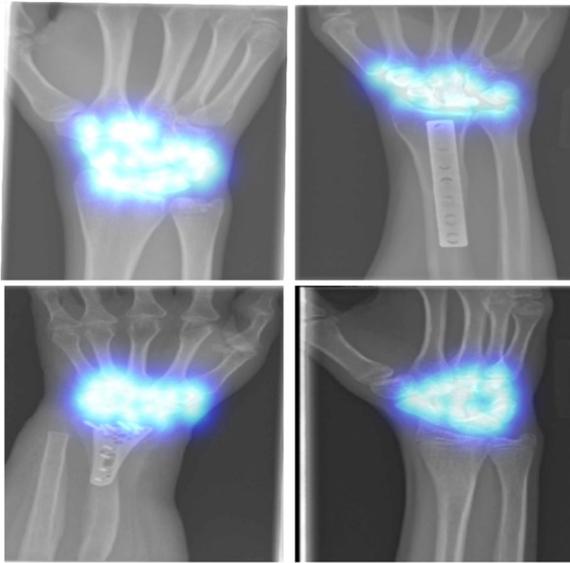


Fig. 15 Attention Map Visualization Highlighting Regions of Focus in Model Decision-Making

While our comparison with Yao et al. [20] and Tanzi et al. [53] involves studies targeting different anatomical regions and populations (e.g., pediatric patients and proximal femur fractures, respectively), these references were selected to illustrate common methodological limitations in radiomics- and deep learning-based fracture detection research. Our study differs in its focus on adult wrist X-rays and the binary classification of fracture vs. non-fracture cases. Despite these differences, the comparison serves to underscore the performance gains and increased interpretability achieved by integrating reproducible radiomic and deep features. These distinctions also reinforce the need for anatomical- and population-specific validation in AI-driven diagnostic models.

Hardalaç et al. [54] explored ensemble deep learning models for wrist fracture detection, achieving an AP of 86.39%. While their study utilized advanced object detection architectures, our approach emphasizes feature fusion and interpretability, yielding higher AUC values and better generalization to unseen data. Similarly, the ensemble model proposed by Tahir et al. [55] achieved an accuracy of 92.96% for humerus fractures using pre-trained deep learning models. In contrast, our study demonstrates the effectiveness of integrating domain-specific radiomic features with learned deep representations, offering a tailored solution for forearm and wrist fractures. Our findings align with Guan et al. [56], who demonstrated the superiority of combining attention mechanisms with deep learning for thighbone fracture detection. Their model achieved an AP of 88.9%,

comparable to our AUC-ROC of 0.96 in the combined approach. Additionally, Bae et al. [57] validated a CNN-based femoral neck fracture detection model across multiple hospitals, achieving an AUC of 0.987. While their study highlighted the potential for external validation, our work focuses on integrating interpretable features, making it a more comprehensive diagnostic framework. In comparisons with studies such as Joshi et al. [52], Hardalaç et al. [54], and Guan et al. [50], which primarily report average precision (AP), we acknowledge that AP is fundamentally distinct from accuracy or AUC-ROC. These metrics assess different dimensions of model performance—AP being more relevant to object detection tasks, while AUC-ROC and accuracy pertain to classification sensitivity and discriminative capacity. Our references to AP in these comparisons are not intended as direct performance analogs but to highlight the differences in methodological approach and application scope. Accordingly, we stress the importance of consistent metric reporting within task domains and highlight our study's strengths in interpretability, feature reproducibility, and generalization rather than direct numerical comparisons.

A relevant benchmark in wrist fracture classification is the work by Kim et al. [58], who used transfer learning with the InceptionV3 architecture on a dataset of 1,389 lateral wrist radiographs. Their model achieved an AUC of 0.954 in distinguishing fractures from non-fractures, demonstrating the efficacy of CNNs pre-trained on non-medical datasets when fine-tuned for radiographic applications. In comparison, our hybrid framework achieved an AUC-ROC of 0.96 using a feature fusion approach that combines interpretable radiomic features with attention-guided deep representations. Unlike Kim et al. [58], who relied solely on a single CNN architecture and limited explainability, our model incorporates SHAP-based interpretation and reproducibility analysis (via ICC and cosine similarity), offering deeper insight into decision-making processes. Furthermore, our dataset encompasses 3,537 cases across multiple institutions, supporting broader generalizability. These methodological enhancements position our framework as a more transparent and scalable solution for wrist fracture diagnosis.

Implications and Future Directions

The hybrid diagnostic framework introduced in this study effectively bridges the gap between standalone radiomics and deep learning approaches by combining their complementary strengths to enhance fracture detection in X-ray imaging. Radiomic features offer interpretability and domain-specific insights based on established clinical markers, while deep features—especially

those extracted via an attention-guided autoencoder—capture abstract, high-dimensional patterns essential for nuanced fracture classification. This integrative design not only improves diagnostic accuracy and sensitivity but also addresses major limitations seen in prior work, such as limited feature reproducibility, model opacity, and poor generalizability.

By enforcing reproducibility thresholds (e.g., ICC ≥ 0.75), incorporating advanced feature selection (e.g., MI, RFE), and adopting ensemble learning strategies, the framework ensures robustness, transparency, and clinical applicability. Importantly, the use of explainable AI techniques, including SHAP value analysis and attention visualization, strengthens model interpretability, fostering trust in clinical decision support environments. Looking ahead, future research should focus on validating the proposed framework across external, multi-center datasets that encompass diverse imaging protocols and patient demographics. Additionally, integration into real-time clinical workflows—such as automated triage in emergency departments or decision support for radiologists—could significantly enhance diagnostic efficiency and patient outcomes. The modular design of the framework also lends itself well to adaptation across other anatomical sites (e.g., hip, spine, shoulder) and imaging modalities (e.g., CT, MRI), offering a scalable foundation for broader musculoskeletal diagnostic applications.

Conclusion

This study presents a robust hybrid diagnostic framework that combines radiomic and deep features for the detection and classification of forearm and wrist fractures using X-ray images. By leveraging advanced feature selection techniques and ensemble classifiers, the proposed framework achieves superior diagnostic accuracy, sensitivity, and AUC-ROC compared to standalone methods. The integration of interpretable radiomics with high-dimensional deep learning features ensures both reliability and clinical relevance, addressing the limitations of traditional and AI-only approaches. The framework's adaptability and strong performance highlight its potential for deployment in real-world clinical settings, particularly in areas with limited radiology expertise. Future work could extend this approach to other anatomical regions and imaging modalities, further demonstrating its versatility and impact in automated fracture diagnosis.

Acknowledgements

Authors are grateful to the Researchers Supporting Project (ANUI2024M111), Alnoor University, Mosul, Iraq.

Authors' contributions

M.J.S., Q.M.H., R.J.A., H.D., M.M.R., M.K., A.P., J.R., W.M.T., M.A., M.J.J. and A.M.A.A.: Investigation; Methodology; Software; Formal analysis; Writing-original draft. B.F.: Conceptualization; Data curation; Project administration; Validation;

Supervision; Writing-review & editing. All authors read and approved the manuscript.

Funding

None.

Data availability

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

Declarations

Ethics approval and consent to participate

The need for ethical approval was waived off by the ethical committee of Alnoor University, Nineveh, Iraq. This study was conducted in accordance with the Declaration of Helsinki.

It was waived off by the ethical committee of Alnoor University, Nineveh, Iraq.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Faculty of Pharmacy, Middle East University, Amman 11831, Jordan. ²College of Pharmacy, Alnoor University, Mosul, Iraq. ³Ahl Al Bayt University, Kerbala, Iraq. ⁴Department of Computer Engineering, Faculty of Engineering & Technology, Marwadi University Research Center, Marwadi University, Rajkot 360003, Gujarat, India. ⁵Department of Chemistry and Biochemistry, School of Sciences, JAIN (Deemed to Be University), Bangalore, Karnataka, India. ⁶Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. ⁷Department of Chemistry, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India. ⁸Department of Public Health and Healthcare Management, Rector, Samarkand State Medical University, 18, Amir Temur Street, Samarkand, Uzbekistan. ⁹College of Nursing, National University of Science and Technology, Dhi Qar, Iraq. ¹⁰Pharmacy College, Al-Farahidi University, Baghdad, Iraq. ¹¹Department of Pharmacy, Al-Zahrawi University College, Karbala, Iraq. ¹²Gilgamesh Ahliya University, Baghdad, Iraq. ¹³Department of Medical Physics and Radiology, Faculty of Paramedical Sciences, Kashan University of Medical Sciences, Kashan, Iran.

Received: 15 March 2025 Accepted: 8 May 2025

Published online: 20 May 2025

References

- De Putter CE, van Beeck EF, Looman CWN, Toet H, Hovius SER, Selles RW. Trends in wrist fractures in children and adolescents, 1997–2009. *J Hand Surg Am.* 2011;36(11):1810–5.
- Fahy K, Duffaut CJ. Hand and wrist fractures. *Curr Sports Med Rep.* 2022;21(10):345–6.
- Zech JR, Carotenuto G, Igbinoba Z, Tran CV, Insley E, Baccarella A, et al. Detecting pediatric wrist fractures using deep-learning-based object detection. *Pediatr Radiol.* 2023;53(6):1125–34.
- Hanel DP, Jones MD, Trumble TE. Wrist fractures. *Orthopedic Clinics.* 2002;33(1):35–57.
- Wu JC, Strickland CD, Chambers JS. Wrist fractures and osteoporosis. *Orthopedic Clinics.* 2019;50(2):211–21.
- Keller M, Rohner M, Honigmann P. The potential benefit of artificial intelligence regarding clinical decision-making in the treatment of wrist trauma patients. *J Orthop Surg Res.* 2024;19(1):579.
- Dababneh S, Colivas J, Dababneh N, Efanov JI. Artificial intelligence as an adjunctive tool in hand and wrist surgery: a review. *Stomatological Dis Sci.* 2024;4:4.
- Aryasomayajula S, Hing CB, Siebachmeyer M, Naeini FB, Ejindu V, Leitch P, et al. Developing an artificial intelligence diagnostic tool for paediatric

- distal radius fractures, a proof of concept study. *The Annals of The Royal College of Surgeons of England*. 2023;105(8):721–8.
9. Dipnall JF, Page R, Du L, Costa M, Lyons RA, Cameron P, et al. Predicting fracture outcomes from clinical registry data using artificial intelligence supplemented models for evidence-informed treatment (PRAISE) study protocol. *PLoS ONE*. 2021;16(9): e0257361.
 10. Cohen M, Puntonet J, Sanchez J, Kierszbaum E, Crema M, Soyer P, et al. Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. *Eur Radiol*. 2023;33(6):3974–83.
 11. Zhang J, Boora N, Melendez S, Rakkunedeth Hareendranathan A, Jaremko J. Diagnostic accuracy of 3D ultrasound and artificial intelligence for detection of pediatric wrist injuries. *Children*. 2021;8(6):431.
 12. Lee KC, Choi IC, Kang CH, Ahn KS, Yoon H, Lee JJ, et al. Clinical validation of an artificial intelligence model for detecting distal radius, ulnar styloid, and scaphoid fractures on conventional wrist radiographs. *Diagnostics*. 2023;13(9):1657.
 13. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to radiomics. *J Nucl Med*. 2020;61(4):488–95.
 14. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. *Physica Med*. 2017;38:122–39.
 15. Fatan M, Hosseinzadeh M, Askari D, Sheikhi H, Rezaeijo SM, Salmanpour MR. Fusion-Based Head and Neck Tumor Segmentation and Survival Prediction Using Robust Deep Learning Techniques and Advanced Hybrid Machine Learning Systems BT - Head and Neck Tumor Segmentation and Outcome Prediction. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A, editors. Cham: Springer International Publishing; 2022. p. 211–23.
 16. Bijari S, Sayfollahi S, Mardokh-Rouhani S, Bijari S, Moradian S, Zahiri Z, et al. Radiomics and Deep Features: Robust Classification of Brain Hemorrhages and Reproducibility Analysis Using a 3D Autoencoder Neural Network. *Bioengineering*. 2024;11(7):643.
 17. Salmanpour M, Hosseinzadeh M, Rezaeijo S, Ramezani M, Marandi S, Einy M, et al. Deep versus handcrafted tensor radiomics features: Application to survival prediction in head and neck cancer. In: EUROPEAN JOURNAL OF NUCLEAR MEDICINE AND MOLECULAR IMAGING. Springer ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATES; 2022. p. S245–6.
 18. Salmanpour MR, Hosseinzadeh M, Akbari A, Borazjani K, Mojallal K, Askari D, et al. Prediction of TNM stage in head and neck cancer using hybrid machine learning systems and radiomics features. *SPIE: Bellingham, WA, USA, 2022*;12033:648–653.
 19. Salmanpour MR, Rezaeijo SM, Hosseinzadeh M, Rahmim A. Deep versus Handcrafted Tensor Radiomics Features: Prediction of Survival in Head and Neck Cancer Using Machine Learning and Fusion Techniques. *Diagnostics*. 2023;13(10):1696.
 20. Yao W, Wang Y, Zhao X, He M, Wang Q, Liu H, et al. Automatic diagnosis of pediatric supracondylar humerus fractures using radiomics-based machine learning. *Medicine*. 2024;103(23):e38503.
 21. Muthu S, Chellamuthu G, Misbah I, Sekar A, Ashraf M. Artificial Intelligence Assisted Convolutional Neural Network for Detection of Distal Radius Fracture. In: 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS). IEEE; 2023. p. 1–6.
 22. Nian S, Zhao Y, Li C, Zhu K, Li N, Li W, et al. Development and validation of a radiomics-based model for predicting osteoporosis in patients with lumbar compression fractures. *Spine J*. 2024;24(9):1625–34.
 23. Biamonte E, Levi R, Carrone F, Vena W, Brunetti A, Battaglia M, et al. Artificial intelligence-based radiomics on computed tomography of lumbar spine in subjects with fragility vertebral fractures. *J Endocrinol Invest*. 2022;45(10):2007–17.
 24. Chiari-Correia NS, Nogueira-Barbosa MH, Chiari-Correia RD, Azevedo-Marques PM. A 3D radiomics-based artificial neural network model for benign versus malignant vertebral compression fracture classification in MRI. *J Digit Imaging*. 2023;36(4):1565–77.
 25. Cedeno-Moreno R, Morales-Hernandez LA, Cruz-Albarran IA. A stacked autoencoder-based aid system for severity degree classification of knee ligament rupture. *Comput Biol Med*. 2024;181:108983.
 26. Macías-García L, Luna-Romera JM, García-Gutiérrez J, Martínez-Ballesteros M, Riquelme-Santos JC, González-Cámpora R. A study of the suitability of autoencoders for preprocessing data in breast cancer experimentation. *J Biomed Inform*. 2017;72:33–44.
 27. Dong G, Liao G, Liu H, Kuang G. A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geosci Remote Sens Mag*. 2018;6(3):44–68.
 28. Bank D, Koenigstein N, Giryres R. Autoencoders. arXiv preprint arXiv:2003.05991. 2020.
 29. Zhong ST, Huang L, Wang CD, Lai JH, Philip SY. An autoencoder framework with attention mechanism for cross-domain recommendation. *IEEE Trans Cybern*. 2020;52(6):5229–41.
 30. AkbarnezhadSany E, EntezariZarch H, AlipoorKermani M, Shahin B, Cheki M, Karami A, et al. YOLOv8 Outperforms Traditional CNN Models in Mammography Classification: Insights From a Multi-Institutional Dataset. *Int J Imaging Syst Technol*. 2025;35(1):e70008.
 31. Javanmardi A, Hosseinzadeh M, Hajianfar G, Nabizadeh AH, Rezaeijo SM, Rahmim A, et al. Multi-modality fusion coupled with deep learning for improved outcome prediction in head and neck cancer. In: *Medical Imaging 2022: Image Processing*. SPIE: Bellingham, WA, USA, 2022; Volume 12032, pp. 664–668.
 32. Rezaeijo SM, Harimi A, Salmanpour MR. Fusion-based automated segmentation in head and neck cancer via advance deep learning techniques. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer; 2022. p. 70–6.
 33. Mahboubisarighieh A, Shahverdi H, Jafarpoor Nesheli S, Alipoor Kermani M, Niknam M, Torkashvand M, et al. Assessing the efficacy of 3D Dual-CycleGAN model for multi-contrast MRI synthesis. *Egyptian Journal of Radiology and Nuclear Medicine*. 2024;55(1):1–12.
 34. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing*. 2021;452:48–62.
 35. Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, et al. Attention mechanisms in computer vision: A survey. *Comput Vis Media (Beijing)*. 2022;8(3):331–68.
 36. Brauwiers G, Frasincar F. A general survey on attention mechanisms in deep learning. *IEEE Trans Knowl Data Eng*. 2021;35(4):3279–98.
 37. Choi SR, Lee M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology (Basel)*. 2023;12(7):1033.
 38. Soydaner D. Attention mechanism in neural networks: where it comes and where it goes. *Neural Comput Appl*. 2022;34(16):13371–85.
 39. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep*. 2017;7(1):1–8.
 40. Hu Q, Whitney HM, Edwards A, Papaioannou J, Giger ML. Radiomics and deep learning of diffusion-weighted MRI in the diagnosis of breast cancer. In: *Medical Imaging 2019: Computer-Aided Diagnosis*. SPIE; 2019. p. 109504A; <https://doi.org/10.1117/12.2512626>.
 41. Wang J, Zeng J, Li H, Yu X. A deep learning radiomics analysis for survival prediction in esophageal cancer. *J Healthc Eng*. 2022;2022(1):4034404.
 42. Kamineni M, Raghu V, Truong B, Alaa A, Schuermans A, Friedman S, et al. Deep learning-derived splenic radiomics, genomics, and coronary artery disease. medRxiv. 2024; 2024.08.16.24312129. <https://doi.org/10.1101/2024.08.16.24312129>.
 43. Xie C, Yu X, Tan N, Zhang J, Su W, Ni W, et al. Combined deep learning and radiomics in pretreatment radiation esophagitis prediction for patients with esophageal cancer underwent volumetric modulated arc therapy. *Radiother Oncol*. 2024;199:110438.
 44. Zhu Y, Man C, Gong L, Dong D, Yu X, Wang S, et al. A deep learning radiomics model for preoperative grading in meningioma. *Eur J Radiol*. 2019;116:128–34.
 45. Beuque MPL, Lobbes MBI, van Wijk Y, Widaatalla Y, Primakov S, Majer M, et al. Combining deep learning and handcrafted radiomics for classification of suspicious lesions on contrast-enhanced mammograms. *Radiology*. 2023;307(5):e221843.
 46. Ali S, Imran AS, Kastrati Z, Daudpota SM, Cheikh FA, Enhancing UM, Detection WF, Classification through Deep Learning and XAI. In: 12th European Workshop on Visual Information Processing (EUVIP). IEEE. 2024;2024:1–6.
 47. Rafi SM, Anuradha T, Srinivas KS. Wrist Fracture Detection Using Deep Learning. In: 2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). Bhopal, India, IEEE; 2025. p. 1–7. <https://doi.org/10.1109/SCEECS64059.2025.10940086>.

48. Wei W, Huang Y, Zheng J, Rao Y, Wei Y, Tan X, et al. YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images. *J Radiat Res Appl Sci*. 2025;18(1):101309.
49. Tieu A, Kroen E, Kadish Y, Liu Z, Patel N, Zhou A, et al. The role of artificial intelligence in the identification and evaluation of bone fractures. *Bioengineering*. 2024;11(4):338.
50. KS SK, Areeckal AS. Classification of Osteoporotic X-ray Images using Wavelet Texture Analysis and Machine Learning. *International Journal of Computing*. 2025;17(1):1–14.
51. Cui J, Li Y, Huang H, Wen J. Dual contrast-driven deep multi-view clustering. *IEEE Transactions on Image Processing*. 2024;33:4753–64. <https://doi.org/10.1109/TIP.2024.3444269>.
52. Joshi D, Singh TP, Joshi AK. Deep learning-based localization and segmentation of wrist fractures on X-ray radiographs. *Neural Comput Appl*. 2022;34(21):19061–77.
53. Tanzi L, Vezzetti E, Moreno R, Aprato A, Audisio A, Massè A. Hierarchical fracture classification of proximal femur X-Ray images using a multistage Deep Learning approach. *Eur J Radiol*. 2020;133:109373.
54. Hardalaç F, Uysal F, Peker O, Çiçekliadağ M, Tolunay T, Tokgöz N, et al. Fracture detection in wrist X-ray images using deep learning-based object detection models. *Sensors*. 2022;22(3):1285.
55. Tahir A, Saadia A, Khan K, Gul A, Qahmash A, Akram RN. Enhancing diagnosis: ensemble deep-learning model for fracture detection using X-ray images. *Clin Radiol*. 2024;79(11):e1394–402.
56. Guan B, Yao J, Wang S, Zhang G, Zhang Y, Wang X, et al. Automatic detection and localization of thighbone fractures in X-ray based on improved deep learning method. *Comput Vis Image Underst*. 2022;216:103345.
57. Bae J, Yu S, Oh J, Kim TH, Chung JH, Byun H, et al. External validation of deep learning algorithm for detecting and visualizing femoral neck fracture including displaced and non-displaced fracture on plain X-ray. *J Digit Imaging*. 2021;34(5):1099–109.
58. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol*. 2018;73(5):439–45.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.