Bone Reports 25 (2025) 101845

Contents lists available at ScienceDirect

Bone Reports

journal homepage: www.elsevier.com/locate/bonr

Artificial intelligence in pediatric osteopenia diagnosis: evaluating deep network classification and model interpretability using wrist X-rays☆

Chelsea E. Harris [a], Lingling Liu [a], Luiz Almeida [b], Carolina Kassick [a], Sokratis Makrogiannis [a,*]

[a] Division of Physics, Engineering, Mathematics, and Computer Science, Delaware State University, 1200 N. Dupont Hwy., Dover, 19901, DE, USA
[b] Department of Orthopaedic Surgery, Duke University, 2080 Duke University Road, Durham, 27710, NC, USA

ARTICLE INFO

Keywords:
Osteopenia prediction
X-ray images
Deep learning
Explainable AI

ABSTRACT

Osteopenia is a bone disorder that causes low bone density and affects millions of people worldwide. Diagnosis of this condition is commonly achieved through clinical assessment of bone mineral density (BMD). State of the art machine learning (ML) techniques, such as convolutional neural networks (CNNs) and transformer models, have gained increasing popularity in medicine. In this work, we employ six deep networks for osteopenia vs. healthy bone classification using X-ray imaging from the pediatric wrist dataset GRAZPEDWRI-DX. We apply two explainable AI techniques to analyze and interpret visual explanations for network decisions. Experimental results show that deep networks are able to effectively learn osteopenic and healthy bone features, achieving high classification accuracy rates. Among the six evaluated networks, DenseNet201 with transfer learning yielded the top classification accuracy at 95.2 %. Furthermore, visual explanations of CNN decisions provide valuable insight into the blackbox inner workings and present interpretable results. Our evaluation of deep network classification results highlights their capability to accurately differentiate between osteopenic and healthy bones in pediatric wrist X-rays. The combination of high classification accuracy and interpretable visual explanations underscores the promise of incorporating machine learning techniques into clinical workflows for the early and accurate diagnosis of osteopenia.

1. Introduction

Osteopenia, otherwise known as low or decreased bone density, is a disorder that affects millions of people worldwide. It was estimated by the year 2020, the prevalence of osteopenia among Americans would reach over 47 million (Karaguzel and Holick, 2010). According to the World Health Organization (WHO), by bone densitometry, osteopenia diagnosis is determined by a T-score between -1 and -2.5 (Kanis et al., 1994). More severe low bone density is categorized as osteoporosis (Bartl and Frisch, 2009). As osteopenia or osteoporosis worsens, there is an increased risk for skeletal fractures. The diagnostic difference between osteopenia and osteoporosis is BMD.

Osteopenia may be caused by lifestyle and genetic predispositions such as lack of physical activity, calcium deficiency, and vitamin deficiency. Optimal bone formation in childhood and adolescence is essential in the prevention of osteoporosis later in life. Approximately 50 % of the calcium in the adult skeleton is deposited during the early years of 13-17. After age 30, there is a natural gradual reduction of bone mass. Early screening and treatment of osteopenia can prevent the development of osteoporosis and future skeletal fractures (Karaguzel and Holick, 2010).

The emergence of machine learning and especially deep learning techniques has driven significant progress in disease detection and diagnosis using medical imaging modalities (Litjens et al., 2017; Shin et al., 2016). These advances have also translated into the field of osteoporosis and osteopenia diagnosis (Smets et al., 2021; Makrogiannis and Zheng, 2021; Pisani et al., 2013; Zheng and Makrogiannis, 2016; Nasser et al., 2017; Zheng et al., 2020; Yousfi et al., 2020; Su et al., 2020; Erzen et al., 2023; Wang et al., 2023; Mikulić et al., 2024; Kumar et al., 2022). Conventional machine learning techniques frequently included a feature engineering stage followed by a classification stage (Pisani et al., 2013; Zheng and Makrogiannis, 2016; Nasser et al., 2017; Zheng et al., 2020; Yousfi et al., 2020). Deep learning-based techniques typically employ end-to-end architectures for the classification task (Erzen et al., 2023; Wang et al., 2023; Mikulić et al., 2024; Kumar et al., 2022), while hybrid approaches combine deep and hand-crafted features for

☆ This article is part of a Special issue entitled: 'AI/ML in bone research' published in Bone Reports.

* Corresponding author.
E-mail address: smakrogiannis@desu.edu (S. Makrogiannis).

https://doi.org/10.1016/j.bonr.2025.101845
Received 16 August 2024; Received in revised form 11 April 2025; Accepted 21 April 2025
Available online 23 April 2025

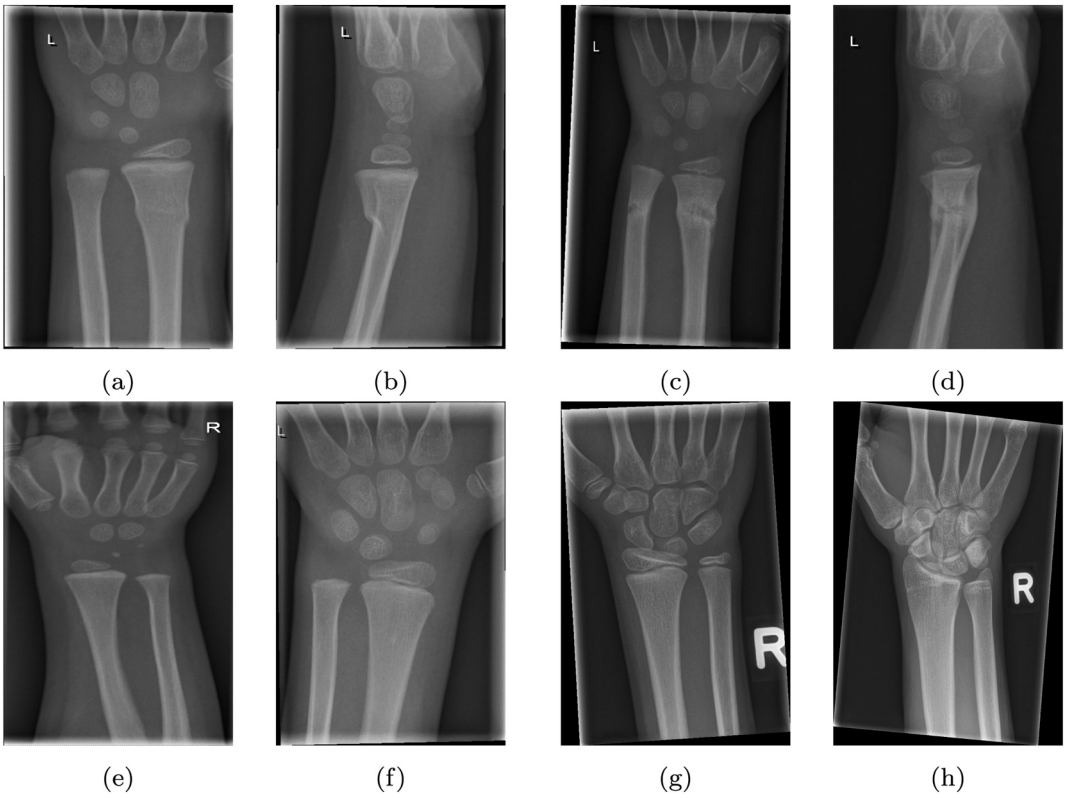

Fig. 1. Non-osteopenia subject X-ray with visible fractures from the GRAZPEDWRI-DX dataset in (a) anteroposterior view and (b) lateral view. Osteopenia X-ray example with visible fracture in (c) anteroposterior and (d) lateral views. Healthy subject X-ray examples in the anteroposterior view from different age groups: (e) 0-4 years, (f) 5-9 years, (g) 10-14 years, and (h) 15-19 years.

classification (Su et al., 2020). While deep learning techniques may yield disease diagnosis results of very high accuracy, their clinical use is hampered by the difficulty in explaining and interpreting the decisions made by deep networks. To address this limitation, the artificial intelligence (AI) research community has shown significant interest in explainable AI (XAI) techniques such as Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al., 2017), Local Interpretable Model-agnostic Explanations (LIME), and other XAI techniques (Mikulić et al., 2024; Linardatos et al., 2021; Fernandes et al., 2024).

1.1. Related works

In a related work published in (Mikulić et al., 2024), the authors evaluated the effect of occluding confounding variables within pediatric wrist radiographs on CNN performance. Image regions surrounding confounding variables, such as metal, fracture, and text annotations were occluded by changing the image pixel intensities to zero in these areas. Trained radiologists compared the visual explanation outputs for models trained on occluded and those trained on non-occluded data; and preferred the interpretability results of the models trained on occluded data. The trade-off of the occluded data use was a decrease in CNN performance, while offering more trustworthy insights into the model's predictions. In a recent paper by Wang et al., a Siamese-based network model was used to detect osteopenia in digital wrist x-rays (Wang et al., 2023). Their method uses a Siamese network, which processes pairs of input images produced from segmenting the ulna and radius bone regions from wrist X-rays. The extracted features were concatenated and passed through fully connected and softmax layers to predict osteopenia. To enhance the feature extraction, convolutional block attention modules were integrated into the deep learning pipeline, focusing on important parts of the images. A deep learning architecture incorporating convolutional identical blocks with skip connections, namely the OsteoNet, was proposed in (Kumar et al., 2022) et al. to classify knee X-ray imaging into healthy and ostoporosis states. The OsteoNet was trained from scratch on a small set of X-ray images of known labels and then evaluated on a new, unseen testing dataset that was not used during the training process. Their model outperformed pretrained state-of-the-art models such ResNet50-v2 and MobileNet-v2, with a significantly lower training time. Their proposed model reached 82.61 % accuracy on the unseen test set of X-ray images. An analysis of the GradCAM results on their proposed network revealed that the accuracy of their network is in part due to the network's focus on the bone regions in the image to learn features and make predictions. The authors in (Su et al., 2020) fused deep features extracted from CNNs with hand-crafted texture features of bone X-ray images for osteoporosis diagnosis. A support vector machine (SVM) classifier was used to distinguish osteoporotic versus normal bone tissue features. Their research showed that feature fusion with feature selection by relevancy improves classification performance.

Our research addresses an important aspect in the application of deep learning for the classification of osteopenia. We specifically focus on the field of pediatric osteopenia diagnosis, which has been less explored compared to adult populations. While previous studies, such as those by Zhang et al. (Zhang et al., 2020) and Sato et al. (Sato et al., 2022), have utilized convolutional neural networks for similar tasks, our work distinguishes itself by specifically targeting the pediatric demographic. Children exhibit different skeletal characteristics than adults (Ferjani et al., 2024), which necessitate tailored approaches for accurate diagnosis. Furthermore, our methods demonstrate superior sensitivity and accuracy in osteopenia classification compared to what has been reported in other studies (Wang et al., 2023; Kumar et al., 2022; Zhang et al., 2020; Sato et al., 2022). Additionally, we incorporate explainable AI techniques and present medical expert ratings of the XAI results as a valuable contribution to the field. This not only enhances the

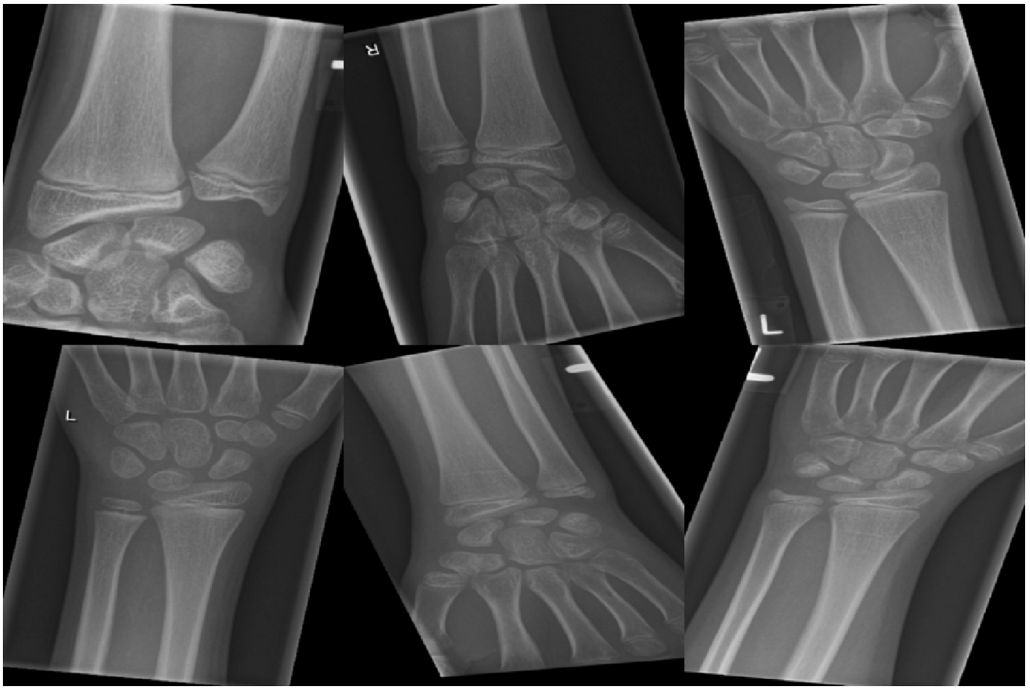

Fig. 2. Data augmentation examples of random rotations and reflections on X-ray images from the training set.

interpretability of the model's predictions, but also bridges the gap between algorithmic outputs and clinical insights. We believe that these aspects of our study present a clear advancement over existing literature and give a valuable contribution to the ongoing discourse in the field of AI-assisted diagnosis, particularly concerning pediatric health.

In this work, we develop and evaluate a deep learning framework for diagnosis of pediatric osteopenia in wrist x-rays. We apply fine-tuning to pre-trained network models, and we employ XAI techniques to offer insight in the decisions made by multiple deep network models. We also compare the performances of deep network models with fine-tuning to performances of other machine learning techniques that use features extracted from the deep network models as their input. Cross-validation results show that end-to-end deep learning techniques yield higher classification accuracy and F1 scores than the baseline techniques. The visual explanations reveal the areas in the x-rays that contribute to the predictions made by the CNNs.

2. Materials and methods

2.1. Dataset

The GRAZPEDWRI-DX dataset is a publicly available dataset of wrist radiographs of 6091 pediatric patients of ages 0-19 years containing 10,643 studies and 20,327 images. The X-ray images were acquired between the years of 2008 and 2018 at the University Hospital Graz, Department of Pediatric Surgery. The images were converted from DICOM to 16-bit grayscale PNG format while maintaining the complete grayscale spectrum to ensure full pixel resolution (Nagy et al., 2022). To ensure accurate study assessments and annotations, experienced radiologists reviewed each study at least twice. This dataset includes several X-ray views, including anteroposterior (AP) and lateral projections. Examples of X-rays from the dataset presenting visible fractures, lateral and anteroposterior views, and subjects of different age groups are shown in Fig. 1.

2.2. Transfer learning on pre-trained deep network models

Transfer learning is a widely used machine learning technique that uses the knowledge from a model trained for one task to serve as the base model for a new task. Utilizing the network weights of a pre-trained model is especially useful, where limited training data is available for the target task. To pre-process the data, we resized the images according to the pre-trained network input dimensions. We normalized the image intensity range from 0 to 65,535 to 0-255 to align with the ImageNet data distribution used during the deep networks' pre-training. Previous works that studied transfer learning techniques for medical image analysis (Tajbakhsh et al., 2016; Morid et al., 2021), indicated that transfer learning of deep networks previously trained on natural imaging databases is an effective approach to extending the analysis to medical imaging data. To further improve model performance and generalization, we applied geometric augmentations, including random rotations between -22.5 and 22.5 degrees and random vertical and horizontal reflections, to both the training and testing sets. These augmentations are performed on the mini-batch of each training iteration to reduce network overfitting related to arm positioning. The augmentation ratio used in this study was 1:1, thus maintaining class and age representation of the training dataset. Examples of data augmentations of X-rays in our training set are displayed in Fig. 2.

We developed an osteopenia vs. healthy bone classification framework applying transfer learning to five convolutional neural networks. These networks were initially trained on the extensive ImageNet dataset of natural images. For our experiments, we replaced the pre-trained fully connected layer of each network with a module including one average pooling layer and two fully connected layers, which we then trained on the X-ray images. The five CNNs used in this study are DenseNet201 (Huang et al., 2017), InceptionV3 (Szegedy et al., 2016), Inception-ResNetV2 (Szegedy et al., 2017), ResNet101 (He et al., 2016), and Xception (Chollet, 2017). We selected these five network models because they have sufficient depth to achieve high accuracy, and because they employ diverse types of network modules such as inception, residual connections, depthwise separable convolutions, and feedforward layer connections. All networks were trained using the categorical entropy loss function and the Adam optimizer with a learning rate of 10^{-3}. We employed Bayesian hyperparameter optimization (Snoek et al., 2012) to fine-tune the number of epochs (2-80) and batch size (8-64) hyperparameters for the training process. The selection

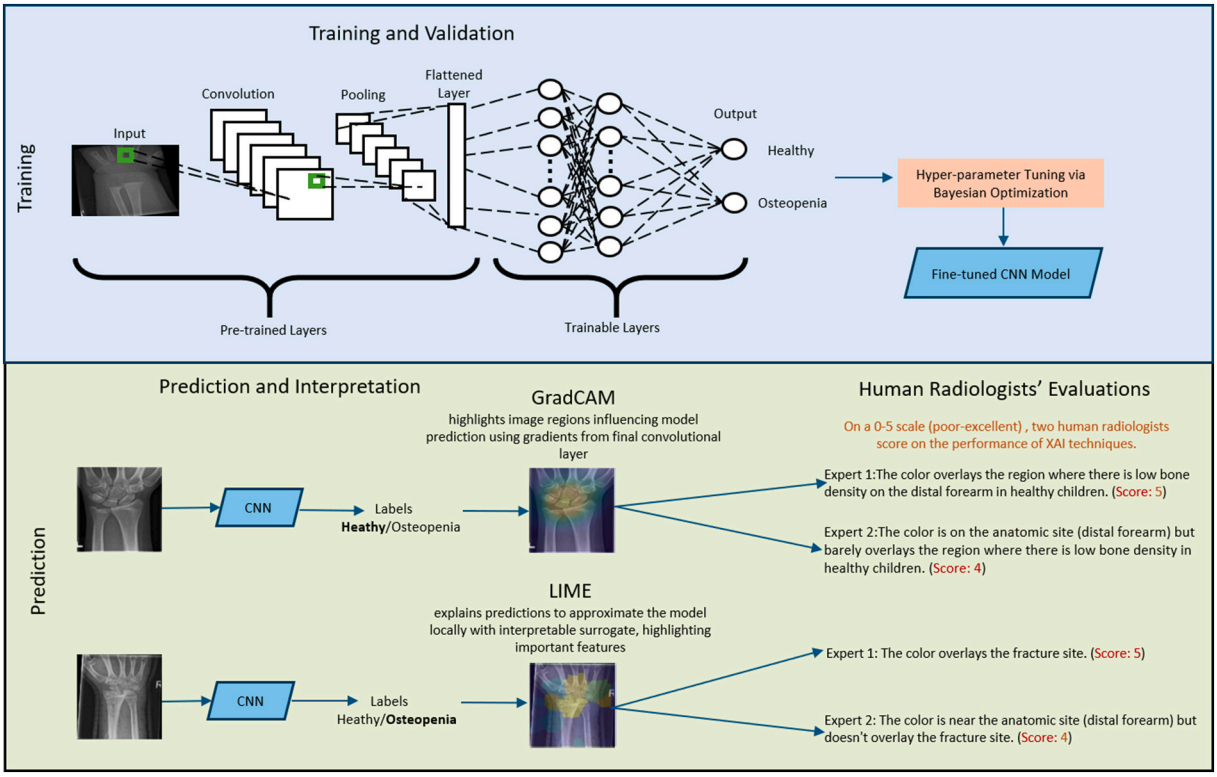

Fig. 3. Transfer learning and model interpretability workflows.

Table 1
Number of images per class for training and testing groups.

	Osteopenic subject images	Healthy subject images
Training set	812	3913
Testing set	307	1685

of a hyperparameter tuning method was carefully considered. While traditional algorithms for global optimum solutions for hyperparameters, like grid search, are commonly used, these search strategies can be very time costly as they make a complete search over the hyperparameter space defined. An additional disadvantage of some traditional search strategies, like grid search and random search, is that these methods disregard previous data in the optimization process (Liashchynskyi and Liashchynskyi, 2019). Bayesian optimization is a statistical modeling-based optimization algorithm that creates a probabilistic mode using Bayes' theorem to predict how well different hyperparameters will perform (Raiaan et al., 2024). A block diagram of our transfer learning and model interpretability pipeline is shown in Fig. 3.

2.3. Model interpretability

Gradient-weighted Class Activation Mapping (Selvaraju et al., 2017) is a technique that provides visual explanations that give insight into regions within an image that drive model predictions. This technique is a class-discriminative approach and generalization of the Class Activation Mapping method (Zhou et al., 2016). GradCAM uses information from class-specific gradients flowing into the final convolutional layer of a CNN to produce coarse localization of the regions that drive a classification decision. The class discriminative localization maps produced by GradCAM not only provide a way to interpret CNN results, but can also give insights into model failures such as network bias.

The Local Interpretable Model-agnostic Explanations technique is a popular explainable AI method that can be applied to any ML classifier (Ribeiro et al., 2016). When applied to an image, LIME, segments the image and creates synthetic samples in the neighborhood of the input image by randomly removing some of these localized features. The black-box model that we want to explain, is then used to make predictions on the synthetic samples and weights are assigned according to their distance from the original image. An interpretable, frequently linear, model is then trained on the synthetic sample set to approximate the behavior of the original network. The interpretable model is used to generate a map that highlights the relative importance of the areas of an image that most influence the classification.

Table 2
Osteopenia classification performances using fine-tuned CNNs. Reported results are averaged over 10 experiment runs using the top performing training parameters obtained through Bayesian optimization (standard deviation in parenthesis).

Deep network	TPR %	TNR %	ACC %	AUC	F1 %
DenseNet201	87.49 *(2.19)*	96.61 *(0.32)*	**95.21** *(0.22)*	**0.9827** *(0.0013)*	84.90 *(0.85)*
InceptionV3	**90.85** *(2.24)*	95.55 *(0.28)*	94.83 *(0.18)*	0.9790 *(0.0024)*	84.40 *(0.76)*
InceptionResNetV2	86.42 *(1.76)*	95.80 *(0.33)*	94.36 *(0.30)*	0.9736 *(0.2552)*	82.52 *(0.94)*
ResNet101	90.32 *(1.39)*	95.79 *(0.37)*	94.94 *(0.20)*	0.9805 *(0.0019)*	84.63 *(0.50)*
Xception	90.26 *(2.64)*	95.96 *(0.67)*	95.08 *(0.46)*	0.9825 *(0.0013)*	**84.98** *(1.30)*
ViT	76.45 *(2.62)*	**97.04** *(0.42)*	93.87 *(0.20)*	0.9765 *(0.0023)*	79.34 *(0.95)*

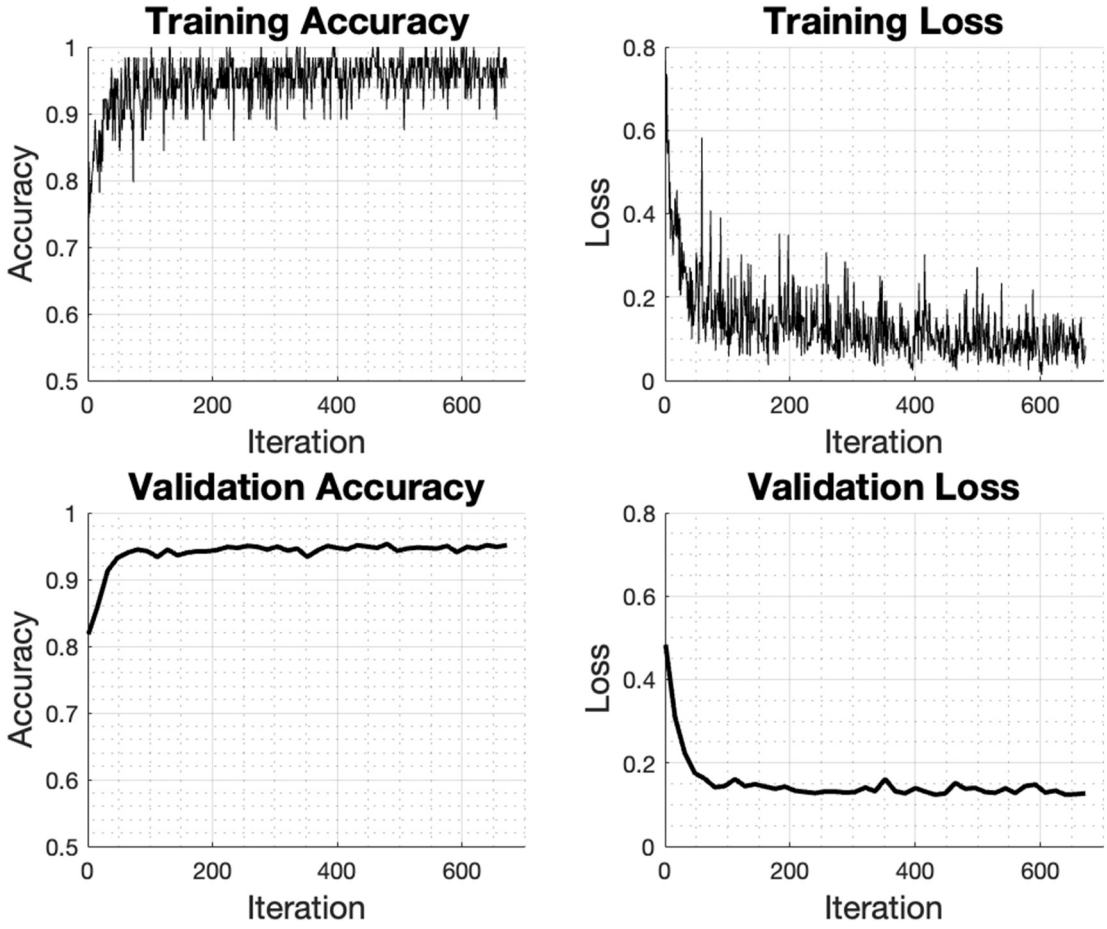

Fig. 4. Training accuracy and loss function graphs of a single experiment using optimized hyperparameters (batch size of 64 and 10 epochs). These graphs show that the DenseNet201 model learns effectively the application domain patterns with no overfitting.

Table 3
Osteopenia classification performances using DF-SRC, DF-SVM, and DF-RF classifiers. Decision fusion results are provided for each classifier group.

	TPR %	TNR %	ACC %	AUC	F1%
DenseNet201-SRC	66.30	82.10	68.72	0.8372	44.72
InceptionV3-SRC	**81.43**	69.85	71.64	0.8432	46.95
InceptionResNetV2-SRC	79.80	58.10	61.45	0.7628	38.95
ResNet101-SRC	24.76	97.92	86.65	0.8014	36.36
Xception-SRC	67.78	66.41	66.62	0.7312	38.48
Fusion (CNN-SRC decisions)	61.24	88.31	84.14	0.8531	54.36
ViT-SRC	43.00	**99.05**	**90.41**	**0.9006**	**58.02**
Fusion (CNN + ViT-SRC decisions)	55.05	92.88	87.05	0.8656	56.71
DenseNet201-SVM	72.64	97.09	93.32	0.9623	77.03
InceptionV3-SVM	70.68	96.80	92.77	0.9580	75.09
InceptionResNetV2-SVM	66.78	96.50	91.92	0.9464	71.80
ResNet101-SVM	70.68	95.07	91.32	0.9447	71.50
Xception-SVM	63.19	96.80	91.62	0.9416	69.91
Fusion (CNN-SVM decisions)	72.31	**97.63**	93.72	0.9757	78.03
ViT-SVM	**79.48**	96.80	**94.13**	0.9715	**80.66**
Fusion (CNN + ViT-SVM decisions)	74.82	97.45	93.98	**0.9774**	79.31
DenseNet201-RF	32.25	98.58	88.35	0.8580	46.05
InceptionV3-RF	32.90	97.86	87.85	0.8417	45.50
InceptionResNetV2-RF	21.82	98.40	86.60	0.8063	33.42
ResNet101-RF	25.41	98.16	86.95	0.8263	37.50
Xception-RF	20.52	98.69	86.65	0.8170	32.14
Fusion (CNN-RF decisions)	22.15	99.41	87.50	0.9107	35.32
ViT-RF	**42.35**	98.46	**89.81**	0.9021	**56.16**
Fusion (CNN + ViT-RF decisions)	25.73	**99.47**	88.10	**0.9250**	40.00

3. Results

3.1. Training and validation set generation

For our experiments, we selected a subset of the dataset containing the AP image view only, including longitudinal scans from all clinical visits for each patient. When available, both laterality images (left and right wrist) were incorporated, excluding cases with metal implants, a cast, and uncertain diagnosis. Our selected data subset captured morphological variability across the 0–19 year age spectrum. We specifically selected the AP view for our experiments to provide a more comprehensive view of the wrist. In this view, the radius and ulna bones are visible, thus offering more specific spatial information than the lateral view, where the bones overlap. We included images depicting visible fractures in both healthy and osteopenia cases to ensure algorithm robustness to realistic variations in patient data. Considering that for some patients more than one image is available, we applied a patient-based testing and training split ensuring that images from the same patient are not present in both the testing and training groups. We performed an approximately 70 %/30 % training and testing randomized split of the data for the experiments. In our test set, we have 1992 X-ray images representing 1613 subjects. The training set consists of 4725 X-ray images from 3764 subjects. The hold-out cross-validation approach was selected to ensure a direct comparison to the performance of the other ML classifiers evaluated in this work. Furthermore, the size of our dataset allows for a comprehensive representation of the underlying data distribution and helps mitigate the risk of overfitting. Table 1 lists the number of osteopenic and healthy X-ray images used in the training and testing sets.

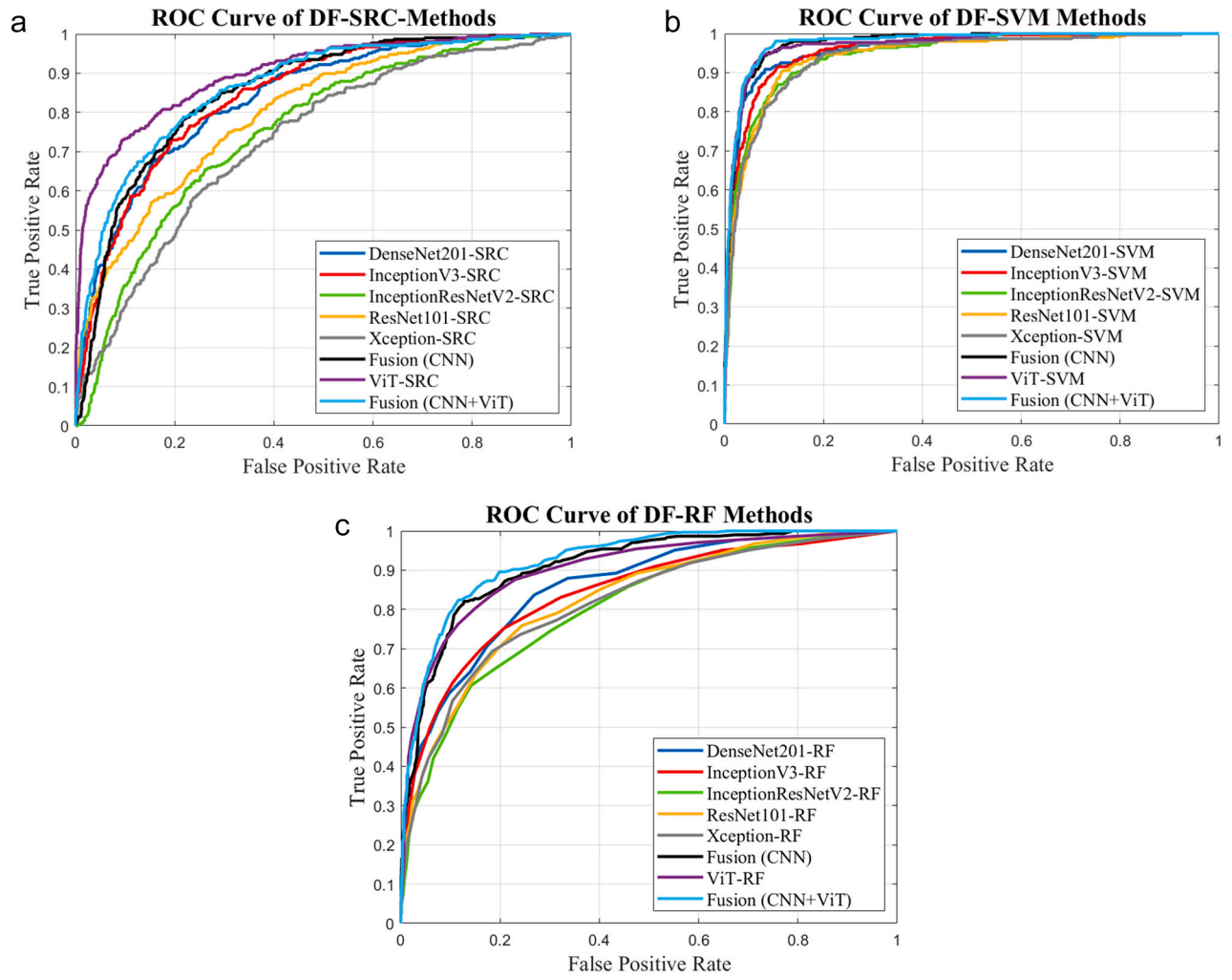

Fig. 5. ROC curves of the DF-SRC, DF-SVM, and DF-RF classifiers.

Table 4
Osteopenia classification performance on GRAZPEDWRI-DX dataset by other methods in the literature. Summary of osteopenia classification performance from methods used in this work. *Best performing model in each group of classifiers used in this work.

Method	TPR %	TNR %	ACC %	AUC	F1 %
Siamese-based network with CBAM (Wang et al., 2023)	–	–	88.09	–	82.89
VGG19 (confounding variable occlusion) (Mikulić et al., 2024)	96.27	–	91.71	–	93.74
DenseNet201 (confounding variable occlusion) (Mikulić et al., 2024)	97.93	–	90.91	–	93.28
ViT-SVM*	79.48	96.80	94.13	0.9715	80.66
ViT-SRC*	43.00	99.05	90.41	0.9006	58.02
ViT-RF*	42.35	98.46	89.81	0.9021	56.16
DenseNet201 (training from scratch)	84.36	95.79	94.03	0.9719	81.32
InceptionV3 (training from scratch)	91.21	94.90	94.33	0.9751	83.21
Xception (training from scratch)	84.36	95.61	93.88	0.9681	80.94
DenseNet201 with transfer learning*	87.49	96.61	95.21	0.9827	84.90

3.2. Evaluation metrics

In this study, we measure classifier performance using the true positive rate (TPR), true negative rate (TNR), F1-score, classification accuracy (ACC), and the area under the receiver operating characteristic curve (AUC) (Grandini et al., 2020). Classification accuracy reflects the proportion of correct predictions made by the classifier and is calculated using the following equation:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

(TP: true positives, TN: true negatives, FP: false positives, and FN: false negatives). The true positive rate indicates how well the CNN classifies osteopenia instances, where the true negative rate is an indication of healthy prediction.

$$TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP} \tag{2}$$

The F1-score is a measure of the weighted average of the CNN classifiers precision and recall.

$$F1 = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

The area under the curve (AUC) of the receiver operator characteristic (ROC) is evaluated as the integral of the ROC curve. The ROC curve is the graph of TPR values versus false positive rate (FPR) values.

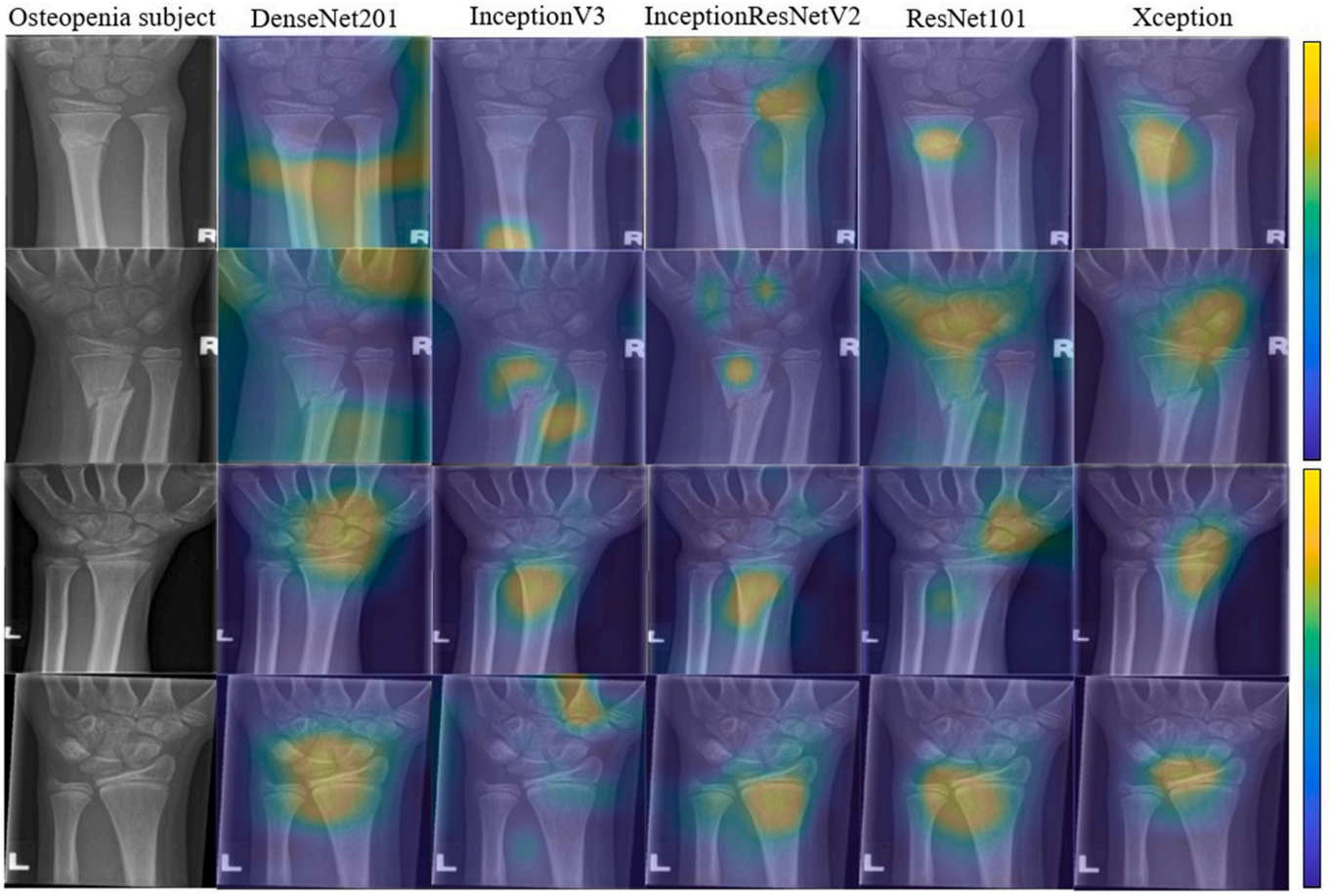

Fig. 6. Four examples of osteopenia test instances, two with a fracture (rows 1 and 2) and two without a fracture (rows 3 and 4). Corresponding GradCAM maps for each network are presented for each test instance.

3.3. Deep network classification of osteopenia using transfer learning

Each pre-trained deep network model underwent transfer learning as the last fully connected layer of each original network was replaced with an average pooling layer followed by two trainable fully connected layers. In Table 2, we report classification results averaged over 10 experiment runs along with standard deviations. All tested models produced very good classification rates. The fine-tuned DenseNet201 network outperforms the other models on average in TPR, ACC and AUC rates; reaching the highest averaged ACC of 95.21 %. To illustrate the training process of the top performing fine-tuned CNN, DenseNet201, we show graphs of the training accuracy, training loss, validation accuracy, and validation loss performance of a single experiment using optimized hyperparameters (batch size of 64 and 10 epochs) in Fig. 4. The accuracy and loss function graphs indicate that the DenseNet201 model learns the new patterns effectively without overfitting.

To investigate the effectiveness of a more recent deep network architecture, we performed transfer learning on a pre-trained vision transformer (ViT). Vision transformers are a relatively newer class of deep learning models that follow the transformer architecture. ViTs employ self-attention mechanisms, unlike CNNs, and have achieved excellent results in image classification tasks (Chen et al., 2021; Maurício et al., 2023). Similar to the CNNs used in this work, the network weights of the pre-trained ViT were initialized from training on the ImageNet dataset. However, we applied additional fine-tuning on the ViT by unfreezing the attention layer weights. We report the averaged results of 10 ViT classification experiments where 80 training epochs and a batch size of 64 was used to train the model.

3.4. Comparison to SRC of deep features and other deep feature-based classifiers

In this section, we evaluate the performance of other machine learning techniques to be used as baselines for comparisons. We evaluate sparse approximation classifiers, support vector machines (SVM) and random forests (RF) (Bishop, 2006). For each classifier, we applied the same training and testing features extracted from the pre-trained CNN models. In addition, to investigate the effectiveness of a more recent deep network architecture, we extracted deep features from a pre-trained vision transformer network and evaluated the SRC, SVM, and RF using the ViT deep features.

Our earlier Sparse Representation Classification of Deep Features method (DF-SRC) is a hybrid framework designed to classify deep features of the X-rays using sparse approximations. To extract deep features from the X-ray data, we utilized the five pre-trained deep network models aforementioned. Deep features were extracted from the network layer preceding the first fully connected layer, followed by average pooling to reduce the feature dimensionality before sparse coding and to promote translation invariance. This technique is described in detail in (Harris et al., 2023; Harris and Makrogiannis, 2022). We also explore the regularization technique of decision score fusion by averaging the log-likelihood scores from each DF-SRC method as explained in (Harris et al., 2024). In our DF-SRC experiments, we evaluated deep feature average pooling lengths of 512, 256, and 128. The average pooling of the extracted deep features reduces the feature dimensionality, thus ensuring an over-complete dictionary that is required to achieve sparse approximation solutions.

The SVM and RF classifiers were trained using 5-fold cross validation

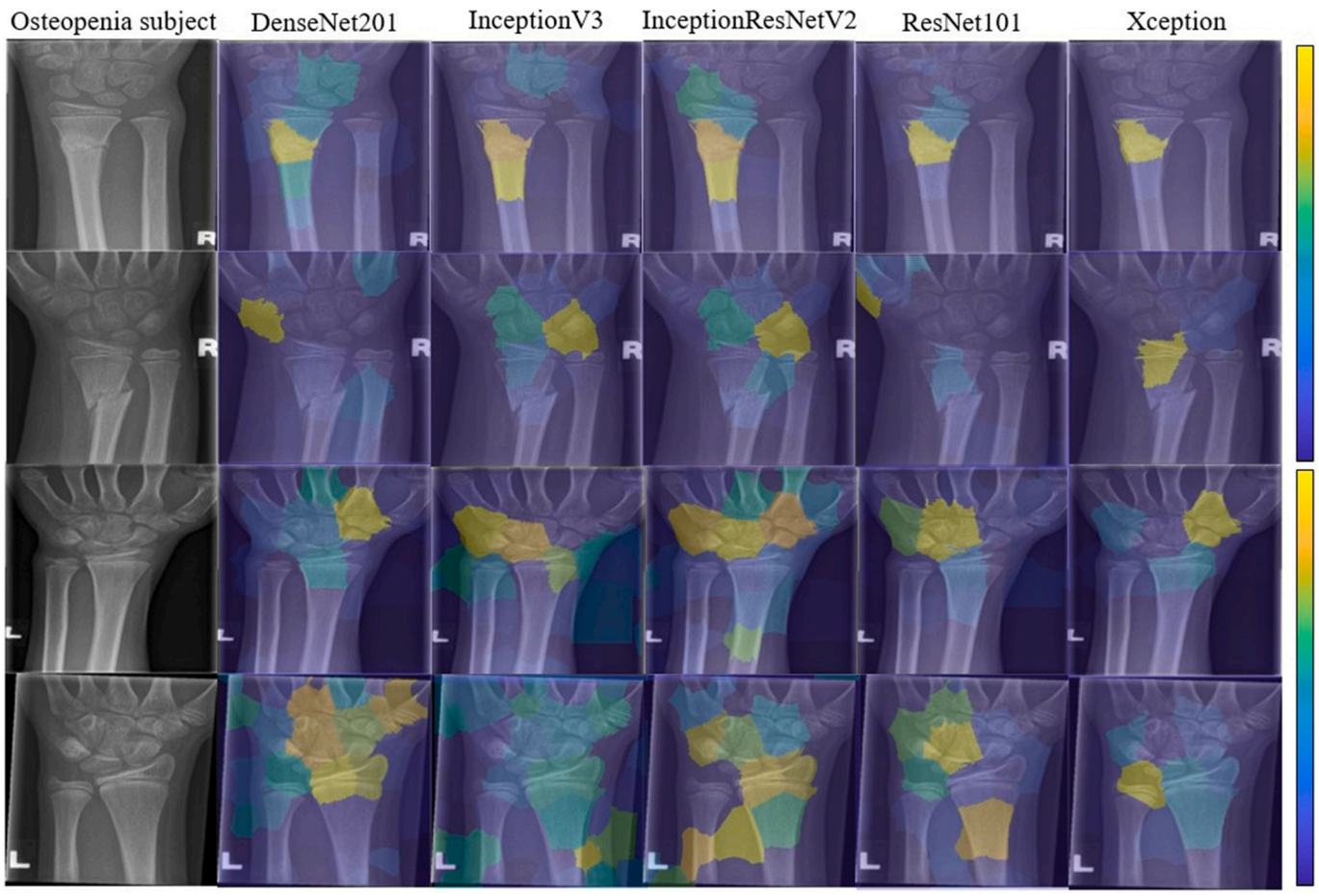

Fig. 7. Four examples of osteopenia test instances, two with a fracture (rows 1 and 2) and two without a fracture (rows 3 and 4). Corresponding LIME maps for each network are presented for each test instance.

on the deep feature training sets. The SVM classifiers were optimized using Bayesian optimization over 30 iterations to optimize hyperparameters of kernel function (i.e. linear, Gaussian, quadratic, or cubic), box constraint, and data standardization. The RF were optimized using bagging with 30 Decision Tree learners, and 4724 maximum splits. We applied majority voting decision fusion of all SVM and all RF trained classifiers to obtain ensemble classification results. Therefore, a final class prediction is made based on the most common prediction among all DF-SVM and DF-RF classifiers.

For CNN+ViT decision fusion experiments, we evaluated our fusion techniques combining the CNN based feature models (DenseNet201, InceptionV3, ResNet101, and Xception) with ViT-based feature models. The performance results for the DF-SRC, DF-SVM, and DF-RF classifiers and their ensembles are detailed in Table 3. Each classifier is identified by the deep features used as the model prefix, followed by the classifier type as the suffix. The decision score fusion of DF-SRC model predictions, where CNN features are used, generally outperforms most individual DF-SRC models across various metrics, particularly in TNR and ACC scores; and all individual DF-SRC models in AUC and F1 performance. This fusion approach also proves advantageous for SVM classifiers where CNN features are used achieving better TNR, ACC, AUC and F1 performance that individual CNN-SVM models. Notably, the highest classification accuracy and F1 among all deep feature classifiers, 94.13 %, is achieved with ViT-SVM, and is comparable to the performance of top deep network model classifiers. Fig. 5 displays ROC curves of the DF-SRC, DF-SVM, and DF-RF classifiers to offer deeper insight into sensitivity and specificity values across different thresholds.

3.5. Comparison to other state-of-the-art methods

We present a comparison of osteopenia classification results with two recent works, Wang et al. (Wang et al., 2023) and Mikulic et al. (Mikulić et al., 2024). Both studies utilized the GRAZPEWRI-DX dataset and applied CNN-based methods for osteopenia classification. Although the data splits, validation techniques, image pre-processing procedures, and overall methodologies employed in these works differ from those in our study, we include their results for the purpose of benchmarking our approach against other state-of-the-art models in this classification task. In addition, we compare the results of the top three performing CNNs with transfer learning (i.e. DenseNet201, InceptionV3 and Xception) to the same network models trained from scratch to perform another baseline method comparison. To train the networks from scratch, we set all layer biases to zero and randomly initialize the layer weights. The training process employed an initial learning rate of 10^{-3}, a learning rate drop factor of 0.9, 80 epochs, and a batch size of 64. It is important to highlight that in Mukilic et al., the DenseNet201 network achieved the highest F1 score among the methods compared here. However, in our approach, the DenseNet201 network with transfer learning achieved the highest accuracy among all the methods compared, surpassing the ACC of the methods in both (Wang et al., 2023; Mikulić et al., 2024). Additionally, our DenseNet201 model outperformed the F1 score reported in (Wang et al., 2023). Table 4 summarizes the osteopenia prediction performance on the GRAZPEWRI-DX dataset with results from our study alongside those of other works.

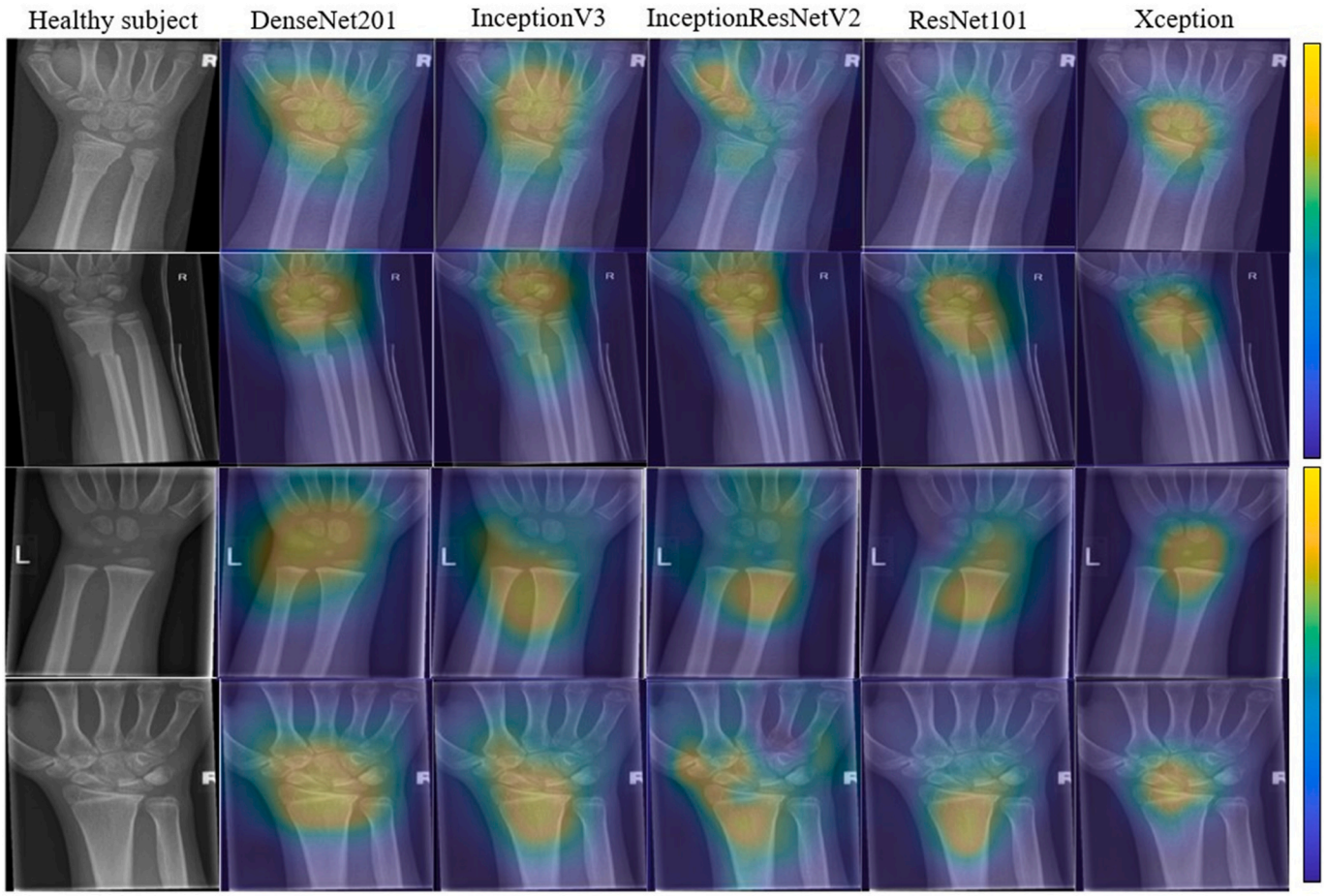

Fig. 8. Four examples of healthy test instances, two with a fracture (rows 1 and 2) and two without a fracture (rows 3 and 4). Corresponding GradCAM maps for each network are presented for each test instance.

3.6. Model interpretability and visual explanations

In this section, we present the results of GradCAM and LIME methods based on the five fine-tuned CNNs used in our classification experiments. We provide the visual explanation results produced by each XAI technique for four osteopenic and four healthy wrist X-rays from our test set. In Figs. 6 and 7, we show four examples of osteopenia test instances; two with a visible fracture and two without a fracture. The class activation heatmap produced using each network is overlayed on each X-ray image to provide a superimposed visualization. Regions of yellow and orange reveal areas within the image indicate regions where the highest gradient values are concentrated. There is a noticeable similarity in some of the visual explanations generated by both methods, for instance, those based on the fine-tuned Xception network and the osteopenia X-ray images. The GradCAM maps show a consistent trend that the models do not focus on non-imaging patterns such as the laterality label present in the image ('L' or 'R'). All networks accurately predicted osteopenia with the exception of the InceptionResNetV2 network in the osteopenia subject with fracture. All networks misclassified the osteopenia subject without fracture as healthy.

In Figs. 8 and 9, we show four examples of healthy test instances; two with a visible fracture and two without a fracture. The LIME results on healthy X-rays are particularly intriguing. In several instances, they highlight that regions within the bone are the most critical for the model's predictions. In contrast, the LIME results from the InceptionV3 model for healthy X-rays are less interpretable, as they often assign higher importance to background and non-bone regions. The yellow and orange regions of the heatmap in LIME visual explanations represent superpixels with the greatest weight on the class prediction. The superpixels themselves provide insight into regions within the input image that were used to create random permutations, or synthetic samples, to fit the learned local interpretable model.

3.7. Medical expert analysis of XAI results

Two medical experts (expert 1: L.A. and expert 2: C.K.) evaluated the effectiveness of GradCAM and LIME visual explanations in aligning with the diagnostic approach typically employed by radiologists when interpreting X-ray images for bone abnormalities. To facilitate this assessment, we provided a set of 40 X-ray images, including 20 pediatric images diagnosed with osteopenia and 20 normal bone X-rays. Each X-ray was paired with GradCAM and LIME outputs from the five CNNs used in this work. Additionally, to introduce variability in the subjects and assess the influence of visible fractures on model behavior, we included 8 healthy wrist X-rays with visible fractures and 5 osteopenic wrist X-rays without visible fractures. This diverse subset of test dataset used in this work allowed for a more comprehensive evaluation of the models' performance in detecting bone abnormalities.

Each heatmap was rated on a 6-point scale, ranging from "poor" (0) to "excellent" (5). A "poor" rating indicated that no discernible features within the heatmap aligned with the medical expert's approach for identifying bone abnormalities. In contrast, an "excellent" rating reflected a strong alignment with the expert's examination process, with clear identification of key bone features relevant to disease prediction. The agreement between the two medical expert ratings was correlated with the Cohen's kappa inter-rater agreement. The quadratic weighted

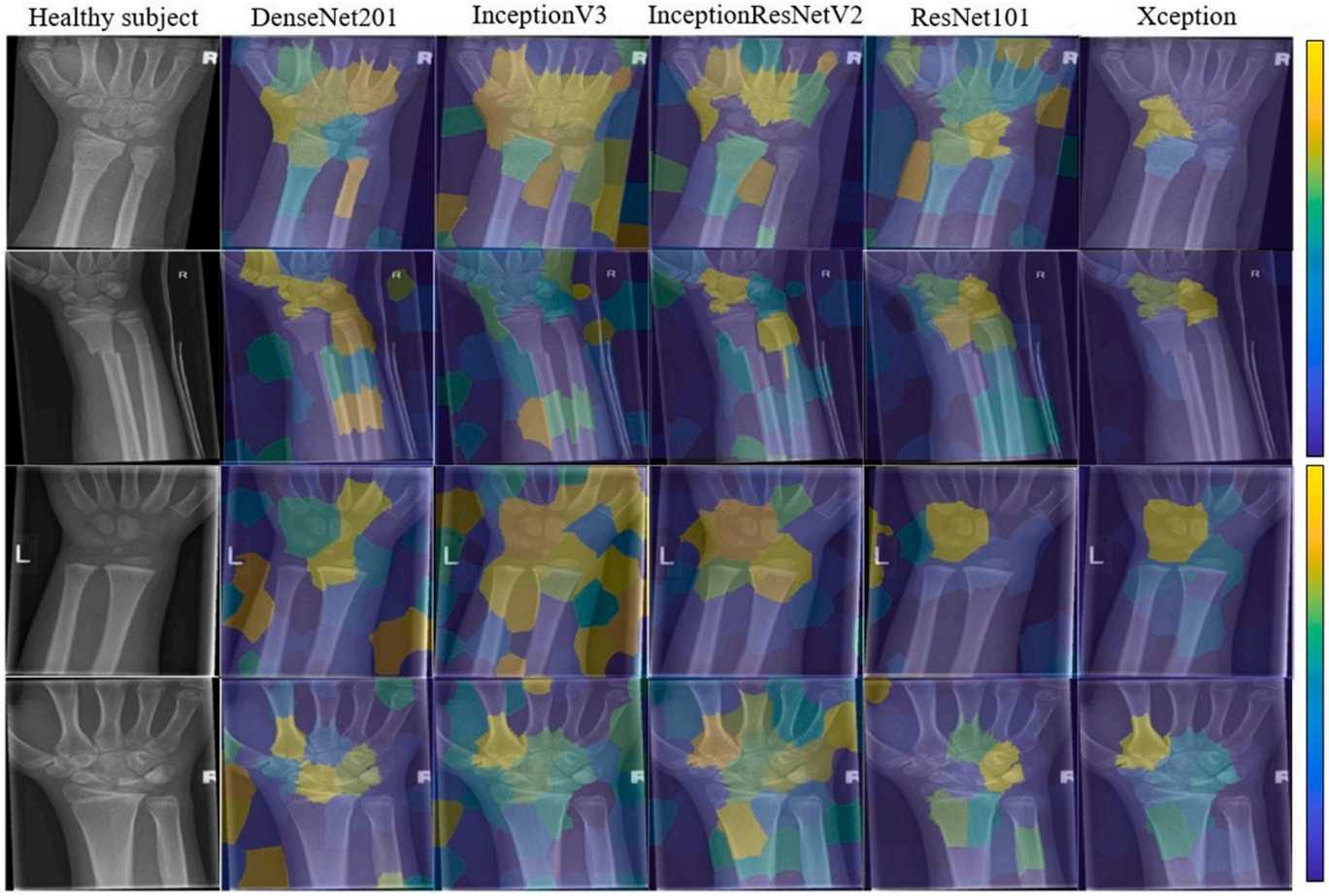

Fig. 9. Four examples of healthy test instances, two with a fracture (rows 1 and 2) and two without a fracture (rows 3 and 4). Corresponding LIME maps for each network are presented for each test instance.

Kappa statistic (κ) for each XAI method and deep network is provided in Table 5. The average scores for each XAI method and deep network are presented in Table 5. Violin plots illustrating the average XAI visual explanation scores across all raters are shown in Fig. 10.

4. Discussion

The results of this work highlight several promising aspects and strengths of this work. One being that all fine-tuned deep networks used in the study achieved an accuracy rate exceeding approximately 94 %, indicating strong performance in differentiating between osteopenic and healthy bones in pediatric wrist X-rays. Of note is that all deep network models with transfer learning achieved an F1 score exceeding approximately 80 %. In addition, our top performing fine-tuned model demonstrated a high sensitivity (or TPR) of 90.85 %. High sensitivity in ML models is especially important in this medical image classification application, as accurate identification of patients with osteopenia can ensure timely and proper treatment. Furthermore, the top specificity (or TNR) achieved by our fine-tuned models was 97.04 %, indicating its effectiveness in accurately identifying healthy bone features. Specificity in this application is also crucial considering the critical impact of false negatives in medical diagnoses. By applying transfer learning to pre-trained models such as DenseNet201, Xception, and others, this study leverages existing knowledge embedded in these models, thus improving efficiency and potentially the overall performance. Our experiments were conducted using a comprehensive dataset consisting of over 6700 images from pediatric patients, which adds robustness to the model's training and testing processes. This research is among the few machine learning applications focused on osteopenia diagnosis, highlighting the critical role of early and accurate detection in children, which can significantly enhance long-term musculoskeletal health and lower the risk of future complications. The developmental and physiological differences inherent in pediatric versus adult bone tissue, underscore the greater complexity of pediatric bone disease diagnosis.

The XAI visual explanation results help to interpret the local behavior of the models and provided insight into which features were influencing the decisions on a case-by-case basis. The heat maps produced by GradCAM and LIME methods show that the networks tend to use relevant bone feature information to make classification predictions. The analysis and ratings provided by medical experts on the XAI visual explanations corroborates this finding.

According to the medical experts' analysis, the regions highlighted in yellow and orange in the GradCAM and LIME heatmaps with high scores are often indicative of fracture sites and areas with low bone density in the distal forearm. The majority of medical expert comments regarding the GradCAM and LIME results for osteopenic X-rays noted a clear overlay of the heatmap with the fracture site or regions exhibiting the lowest bone density in the anatomical area. The XAI explanation using the Xception network were rated "excellent" in all 20 samples by Expert 1. Expert 2 yielded ratings that were consistent. The medical expert comments on the XAI outputs of healthy X-rays revealed that the key areas highlighted in the GradCAM heat maps mostly overlay regions where low bone density is present in the distal forearm in healthy children. This property was also noted in the majority of the LIME outputs of healthy X-rays. However, the medical experts noted that color diffusion in the activation heatmaps caused by localization inaccuracies,

Table 5
Mean (std. dev.) ratings of GradCAM and LIME maps for precision in highlighting key areas of X-ray images relevant for disease prediction in healthy and osteopenic subject by Expert 1 (a,b) and Expert 2 (c,d). Average ratings from both experts for (e) healthy and (f) osteopenic XAI results, and (g) the corresponding Cohen's κ coefficient.

	DenseNet201	InceptionV3	InceptionResNetV2	ResNet101	Xception	AVG
(a) Average ratings Expert 1 - Healthy						
GradCAM	4.65 *(1.14)*	4.40 *(1.14)*	3.70 *(1.30)*	4.80 *(0.62)*	4.85 *(0.67)*	4.48
LIME	3.90 *(1.65)*	2.10 *(1.68)*	3.00 *(1.84)*	2.95 *(1.95)*	4.65 *(0.81)*	3.32
(b) Average ratings from Expert 1 - Osteopenia						
GradCAM	4.15 *(1.63)*	3.75 *(3.75)*	4.10 *(1.97)*	4.85 *(0.36)*	5.00 *(0)*	4.37
LIME	4.90 *(0.45)*	4.55 *(1.10)*	4.65 *(0.93)*	3.20 *(2.21)*	5.00 *(0)*	4.46
(c) Average ratings from Expert 2 - Healthy						
GradCAM	4.60 *(1.14)*	4.40 *(1.14)*	3.70 *(1.30)*	4.85 *(0.49)*	4.85 *(0.67)*	4.48
LIME	3.95 *(1.57)*	2.05 *(1.70)*	3.00 *(1.84)*	2.90 *(1.92)*	4.65 *(0.81)*	3.31
(d) Average ratings from Expert 2 - Osteopenia						
GradCAM	4.10 *(1.62)*	3.75 *(1.97)*	4.10 *(1.65)*	4.85 *(0.37)*	4.95 *(0.22)*	4.35
LIME	4.90 *(0.45)*	4.45 *(1.10)*	4.60 *(1.14)*	3.15 *(2.18)*	4.95 *(0.22)*	4.41
(e) Average ratings from both Expert1 and Expert 2 - Healthy						
GradCAM	4.625	4.40	3.70	4.825	4.85	4.48
LIME	3.925	2.075	3.00	2.925	4.65	3.315
(f) Average ratings from both Expert1 and Expert 2 - Osteopenia						
GradCAM	4.125	3.75	4.10	4.85	4.975	4.36
LIME	4.90	4.5	4.625	3.175	4.975	4.435
(g) Cohen's kappa between Expert 1 and Expert 2						
κ - GradCAM	0.9815	1.00	1.00	0.8892	0.8851	0.9615
κ - LIME	0.6098	0.9890	0.7059	0.9939	0.9645	0.8526

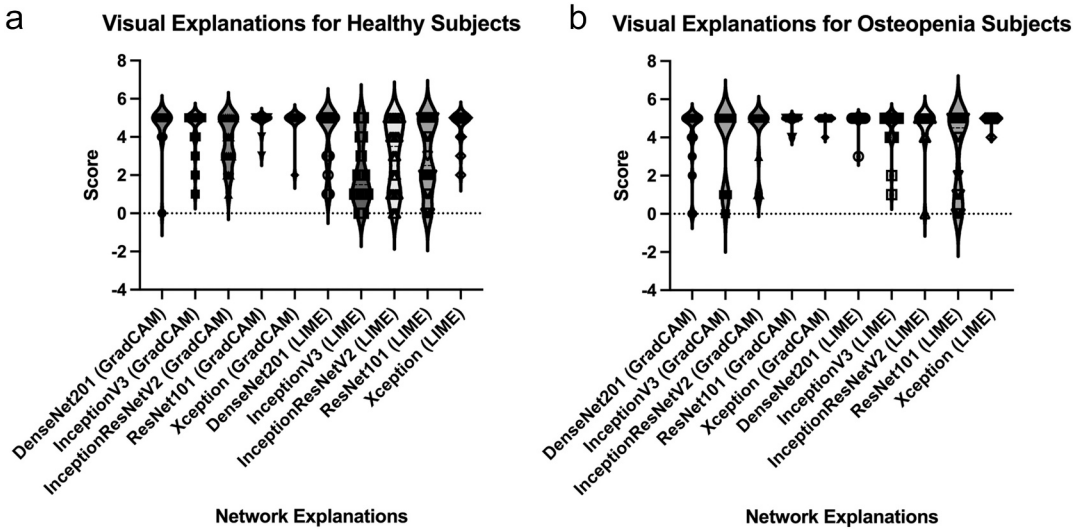

Fig. 10. Violin plots of XAI visual explanation average scoring for all raters grouped by network model for each class.

often seen in the LIME results, complicates the interpretation of the heatmap, implying reduced precision. The GradCAM heatmaps on the other hand show better localization of the abnormalities. The GradCAM and LIME results using the Xception network generally exhibited the highest average score of the 5 networks. Overall, the GradCAM results exhibited higher average scores than the LIME results for healthy X-rays, suggesting a stronger alignment with medical expert examinations. In general, the two experts yielded consistent ratings with a few exceptions in LIME maps, mainly for DenseNet201 and InceptionResNetV2. The few deviations in ratings of LIME may be caused by the reduced specificity of these maps.

While there are several strengths in the methods and performances presented in this work, there are specific areas that could benefit from further exploration and validation. One area would be the use of multiple X-ray views for network training; additional image views may help the networks learn more diverse imaging patterns. In future research, we plan to incorporate diverse clinical settings to ensure better representation of pediatric X-ray data. Another direction of further research would be to investigate the applicability of this technique in specific scenarios, such as detecting osteopenia in subjects without fractures.

5. Conclusion

In this work, we developed a deep learning framework for osteopenia vs. healthy bone classification and decision interpretation using X-ray imaging from pediatric wrist data. Our findings demonstrate the effectiveness of deep network models in distinguishing osteopenic bones from healthy ones in pediatric wrist X-rays. The high classification accuracy and the interpretability of deep network decisions reinforce the potential of integrating machine learning techniques into clinical workflows for early and accurate diagnosis of osteopenia. This approach may not only enhance diagnostic accuracy but can also contribute to the transparency and trustworthiness of AI-driven biomedical applications.

CRediT authorship contribution statement

Chelsea E. Harris: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Lingling Liu:** Writing – original draft, Visualization, Validation, Software, Investigation, Data curation. **Luiz Almeida:** Writing – review & editing, Validation, Investigation, Formal analysis. **Carolina Kassick:** Writing – review & editing, Validation, Formal analysis. **Sokratis Makrogiannis:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) (award #SC3GM113754), the Interdisciplinary Health Equity Research Center (IHER) NIH award #U54MD015959.

Data availability

Data will be made available on request.

References

Bartl, R., Frisch, B., 2009. Osteoporosis: Diagnosis, Prevention, Therapy. Springer Science & Business Media.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Chen, C., Fan, Q., Panda, R., 2021. Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366.

F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

E. M. Erzen, E. Bütün, M. A. Al-Antari, R. A. Saleh, D. Addo, Artificial intelligence computer-aided diagnosis to automatically predict the pediatric wrist trauma using medical x-ray images, in: 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS), IEEE, 2023, pp. 1–7.

Ferjani, H., Cherif, I., Nessib, D., Kaffel, D., Maatallah, K., Hamdi, W., 2024. Pediatric and adult osteoporosis: a contrasting mirror. Ann. Pediatr. Endocrinol. Metab. https://doi.org/10.6065/apem.2346114.057.

Fernandes, L., Fernandes, J.N.D., Calado, M., Pinto, J.R., Cerqueira, R., Cardoso, J.S., 2024. Intrinsic explainability for end-to-end object detection. IEEE Access 12, 2623–2634. https://doi.org/10.1109/ACCESS.2023.3347038.

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: an overview, arXiv preprint. arXiv:2008.05756.

Harris, C., Okorie, U., Makrogiannis, S., 2023. Spatially localized sparse approximations of deep features for breast mass characterization. Math. Biosci. Eng. 20 (9), 15859.

Harris, C.E., Makrogiannis, S., 2022. Sparse analysis of block-boosted deep features for osteoporosis classification. In: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), pp. 1–5. https://doi.org/10.1109/IVMSP54334.2022.9816199.

Harris, C.E., Okorie, U., Makrogiannis, S., 2024. Mammographic breast density classification by integration of deep dictionaries and multi-model sparse approximations, in. IEEE International Symposium on Biomedical Imaging (ISBI) 2024, 1–5. https://doi.org/10.1109/ISBI56570.2024.10635360.

K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

Kanis, J.A., Melton III, L.J., Christiansen, C., Johnston, C.C., Khaltaev, N., 1994. The diagnosis of osteoporosis. J. Bone Miner. Res. 9 (8), 1137–1141.

Karaguzel, G., Holick, M.F., 2010. Diagnosis and treatment of osteopenia. Rev. Endocr. Metab. Disord. 11 (4), 237–251.

A. Kumar, R. C. Joshi, M. K. Dutta, R. Burget, V. Myska, Osteo-net: a robust deep learning-based diagnosis of osteoporosis using x-ray images, in: 2022 45th International Conference on Telecommunications and Signal Processing (TSP), 2022, pp. 91–95. doi:https://doi.org/10.1109/TSP55681.2022.9851342.

P. Liashchynskyi, P. Liashchynskyi, Grid search, random search, genetic algorithm: a big comparison for NAS, in: arXiv preprint arXiv:1912.06059, 2019.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable AI: a review of machine learning interpretability methods. Entropy 23 (1). https://doi.org/10.3390/e23010018.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Makrogiannis, S., Zheng, K., 2021. AIM in osteoporosis. Springer International Publishing 1–17.

Maurício, J., Domingues, I., Bernardino, Jorge, 2023. Comparing vision transformers and convolutional neural networks for image classification: a literature review. Appl. Sci. 13.

Mikulić, M., Vičević, D., Nagy, E., Napravnik, M., Štajduhar, I., Tschauner, S., Hržić, F., Jul 2024. Balancing performance and interpretability in medical image analysis: case study of osteopenia. J. Imaging Inform. Med. https://doi.org/10.1007/s10278-024-01194-8.

Morid, M.A., Borjali, A., Del Fiol, G., 2021. A scoping review of transfer learning research on medical image analysis using imagenet. Comput. Biol. Med. 128, 104115. https://doi.org/10.1016/j.compbiomed.2020.104115.

Nagy, E., Janisch, M., Hržić, F., Sorantin, E., Tschauner, S., 2022. A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. Sci. Data 9 (1), 222.

Y. Nasser, M. El Hassouni, A. Brahim, H. Toumi, E. Lespessailles, R. Jennane, Diagnosis of osteoporosis disease from bone x-ray images with stacked sparse autoencoder and SVM classifier, in: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, 2017, pp. 1–5.

Pisani, P., Renna, M.D., Conversano, F., Casciaro, E., Muratore, M., Quarta, E., Di Paola, M., Casciaro, S., 2013. Screening and early diagnosis of osteoporosis through x-ray and ultrasound based techniques. World J. Radiol. 5 (11), 398.

Raiaan, M., Sakib, S., Fahad, N., Al Mamun, A., Rahman, M., Shatabda, S., Mukta, M., 2024. A systematic review of hyperparameter optimization techniques in convolutional neural networks. Decis. Anal. J. 11, 100470. https://doi.org/10.1016/j.dajour.2024.100470.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Y. Sato, N. Yamamoto, N. Inagaki, Y. Iesaki, T. Asamoto, T. Suzuki, S. Takahara, Deep learning for bone mineral density and T-score prediction from chest X-rays: A multicenter study, in: Biomedicines, vol. 10, 2022.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Shin, H., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R. M., 2016. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35 (5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162.

Smets, J., Shevroja, E., Hügle, T., Leslie, W.D., Hans, D., 2021. Machine learning solutions for osteoporosis—a review. J. Bone Miner. Res. 36 (5), 833–851.

J. Snoek, H. Larochelle, R. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, 2012, Vol. 25.

Su, R., Liu, T., Sun, C., Jin, Q., Jennane, R., Wei, L., 2020. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 385, 300–309. https://doi.org/10.1016/j.neucom.2019.12.083.

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

C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312.

Z. Wang, A. Chetouani, R. Jennane, A siamese-based network for the detection of osteopenia in paediatric digital x-rays of the wrist, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 2023, pp. 1–4.

Yousfi, L., Houam, L., Boukrouche, A., Lespessailles, E., Ros, F., Jennane, R., 2020. Texture analysis and genetic algorithms for osteoporosis diagnosis. Int. J. Pattern Recognit. Artif. Intell. 34 (05), 2057002.

B. Zhang, K. Yu, Z. Ning, K. Wang, Y. Dong, X. Liu, S. Liu, J. Wang, C. Zhu, Q. Yu, et.al. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: a multicenter retrospective cohort study, in: Bone, vol. 140, 2020.

K. Zheng, S. Makrogiannis, Bone texture characterization for osteoporosis diagnosis using digital radiography, in: 2016 38th Annual International Conference of the IEEE

Engineering in Medicine and Biology Society (EMBC), 2016, pp. 1034–1037. doi:https://doi.org/10.1109/EMBC.2016.7590879.
Zheng, K., Harris, C.E., Jennane, R., Makrogiannis, S., 2020. Integrative blockwise sparse analysis for tissue characterization and classification. Artif. Intell. Med. 107, 101885.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).