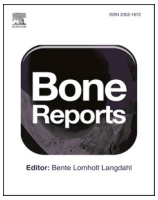
Bone Reports 22 (2024) 101801

Contents lists available at ScienceDirect

Bone Reports

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



Classification of AO/OTA 31A/B femur fractures in X-ray images using YOLOv8 and advanced data augmentation techniques

Giorgia Marullo [a,*], Luca Ulrich [a], Francesca Giada Antonaci [a], Andrea Audisio [b], Alessandro Aprato [c], Alessandro Massè [c], Enrico Vezzetti [a]

[a] Department of Management, Production, and Design, Politecnico di Torino, C.so Duca degli Abruzzi, 24, Torino 10129, Italy
[b] Pediatric Orthopaedics and Traumatology, Regina Margherita Children's Hospital, Torino 10126, Italy
[c] Department of Surgical Sciences, University of Turin, Torino 10124, Italy

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural networks
Femur fracture
Computer assisted diagnosis

ABSTRACT

Femur fractures are a significant worldwide public health concern that affects patients as well as their families because of their high frequency, morbidity, and mortality. When employing computer-aided diagnostic (CAD) technologies, promising results have been shown in the efficiency and accuracy of fracture classification, particularly with the growing use of Deep Learning (DL) approaches. Nevertheless, the complexity is further increased by the need to collect enough input data to train these algorithms and the challenge of interpreting the findings. By improving on the results of the most recent deep learning-based Arbeitsgemeinschaft für Osteosynthesefragen and Orthopaedic Trauma Association (AO/OTA) system classification of femur fractures, this study intends to support physicians in making correct and timely decisions regarding patient care. A state-of-the-art architecture, YOLOv8, was used and refined while paying close attention to the interpretability of the model. Furthermore, data augmentation techniques were involved during preprocessing, increasing the dataset samples through image processing alterations. The fine-tuned YOLOv8 model achieved remarkable results, with 0.9 accuracy, 0.85 precision, 0.85 recall, and 0.85 F1-score, computed by averaging the values among all the individual classes for each metric. This study shows the proposed architecture's effectiveness in enhancing the AO/OTA system's classification of femur fractures, assisting physicians in making prompt and accurate diagnoses.

1. Introduction

Deep Learning (DL), a powerful branch of Artificial Intelligence (AI), is rapidly gaining attention in the medical field as an optimization tool for all phases of clinical practice (Checcucci et al., 2023), from diagnosis and surgery (Cannavò et al., 2020) to pharmaceutical investigation (Rajula et al., 2020). AI along with Extended Reality has empowered the "Augmented Humanity" paradigm, which involves the study of techniques, technologies, and applications for improving human productivity or capabilities through interactive digital extension of individual abilities (Cannavò et al., 2020), such as senses, motor skills, or cognitive capacities (Raisamo et al., 2019). This concept is of remarkable importance in the medical field, where an efficient interaction between physicians and computers could optimize clinical processes and overcome bottlenecks in decision-making processes from the perspective of supporting humans instead of replacing them. Literature shows that combining AI and human judgment can increase diagnostic precision and effectiveness (Göndöcs and Dörfler, 2024) in several medical disciplines, among which orthopedics (Tariq et al., 2023; Wang et al., 2022), urology (Bulten et al., 2021), and dermatology (Hekler et al., 2019; Tschandl et al., 2020). In particular, promising outcomes have been observed in the accuracy and efficiency of fracture classification when using computer-assisted diagnosis (CAD) tools (Lindsey et al., 2018; Yang et al., 2020; Tanzi et al., 2020). These systems use various methods, including DL algorithms, to evaluate medical images and offer diagnostic judgments. This can be especially helpful in emergency scenarios where prompt and precise diagnosis is essential, as postponing surgical treatment or inaccuracies in surgical planning may result in higher morbidity and death rates (Ryan et al., 2015).

Due to the high occurrence (more than 10 million cases annually (Wu et al., 2021)), femur fractures continue to be a substantial global public health concern, which affects patients, their families, and healthcare systems (Dyer et al., 2016). This illness frequently implies a significant turning point in life, particularly for elderly individuals as it further

* Corresponding author.
E-mail address: giorgia.marullo@polito.it (G. Marullo).

https://doi.org/10.1016/j.bonr.2024.101801
Received 26 June 2024; Received in revised form 20 August 2024; Accepted 5 September 2024
Available online 16 September 2024

reduces already compromised capacity for self-sustenance. Furthermore, within a year, just 40–60 % of senior people recover their pre-fracture level of mobility to carry out everyday activities (Dyer et al., 2016). Considerable attention has been given to this issue due to the risk of severe disability (Johnell and Kanis, 2005; Bäcker et al., 2021), and its high morbidity and mortality rates (Sing et al., 2023), mainly because, as society ages, the yearly number is continuously rising (Sambrook and Cooper, 2006; Gullberg et al., 1997) and could double in the upcoming 20 to 30 years (Sing et al., 2023). Therefore, to enhance patient outcomes and reduce the risk of hip fractures in the future, interventions are required to prevent, close the treatment gap, and offer post-fracture care (Sing et al., 2023; Fischer et al., 2021).

The classification of femur fractures is the current study's main topic and is based on multiple systems to emphasize a distinct fracture component. Several attempts have been made to address this problem supported by DL algorithms. Some studies focused on a high-level classification, such as distinguishing between atypical femoral fractures (AFFs) from normal femur fractures (NFFs) (Zdolsek et al., 2021) or considering only high-level classes that merge multiple low-level categories. For instance, Mutasa et Al. worked on a CNN to classify the images as Garden (Garden, 1961) I/II fracture, Garden III/IV fracture, or no fracture for femoral neck fracture detection and classification (Mutasa et al., 2020), while Alzaid et al. considered only type A, B, C of Vancouver Classification System (D., 1995) for classifying and localizing Peri-prosthetic Femur Fracture (PFF) (Alzaid et al., 2022). Some other studies exploited custom classes, such as Krogue et al. (2020), which considered normal, displaced femoral neck fracture, nondisplaced femoral neck fracture, intertrochanteric fracture, previous open reduction and internal fixation, or previous arthroplasty. Recently, the Arbeitsgemeinschaft für Osteosynthesefragen (AO) and Orthopaedic Trauma Association (OTA) classification system (Meinberg et al., 2018) has been commonly adopted for femur fractures. Some methods were proposed, ranging from fully automatic tools that concentrated only on a limited set of classes, such as “A”, “B” and “not-fractured” (Jiménez-Sánchez et al., 2020), or “no fracture” and each type of A fractures (Lee et al., 2020), to other methods which merged the visual data ancillary information at training from radiology reports. In this way, the model learned from X-ray images and textual radiology reports simultaneously (Lee et al., 2020). In a previous study, remarkable results were obtained concerning femur fracture classification according to the AO/OTA classification system. The proposed system explored the Vision Transformer (ViT) architecture to address the categorization of seven classes: Unbroken, A1, A2, A3, B1, B2, B3 (Tanzi et al., 2022).

Due to the rapid advancement in the development of DL technology, new algorithms with increasingly high performance are continuously updated and introduced to support individuals in different application fields. Nevertheless, healthcare entails some issues that cannot be easily addressed. For instance, neural networks require a huge amount of data to be properly trained. In this sense, medical information could not be available due to a limited number of patients suffering from a specific pathology, or gathering sufficient input data could be challenging due to privacy-related issues (Lee et al., 2020).

Although deep learning technologies can produce incredibly dependable results, it can be challenging to comprehend their behaviors because they are frequently very opaque, if not completely undetectable. It could still be challenging for even highly qualified professionals to fully understand these so-called “black-box” models (Yang et al., 2022). One of the most common criticisms of this research area is the difficulty in explaining the results of these algorithms and their lack of interpretability (Rai, 2020), particularly in the medical industry where a high degree of reliability is necessary to allow physicians and patients to accept and trust the technology. To this purpose, the class of systems known as explainable AI (XAI) offers transparency into the decision-making, prediction, and action execution processes of AI systems, unboxing how they make their “black-box” decisions (Yang et al., 2022). In addition to outlining the method's advantages and disadvantages, XAI indicates the system's future behavior and justifies the decision-making process (Du et al., 2019). Various XAI methods have been explored in digital healthcare and medicine (Yang et al., 2022), including dimension reduction (Hao et al., 2018; Bernardini et al., 2019), feature importance (Suh et al., 2020; Singh et al., 2020), attention mechanism (Kaji et al., 2019; Kim et al., 2021; Rajpurkar et al., 2020; Porumb et al., 2020), knowledge distillation (Lee et al., 2019; Prentzas et al., 2019), and surrogate representations (Panigutti et al., 2020; Lauritsen et al., 2020).

A reliable and precise method for identifying the different types of femur fractures could be essential for clinical practice to ensure rapid diagnosis and improved outcomes following surgery (Qi et al., 2020). As a matter of fact, patients frequently suffer grave effects when a fracture in their radiograph is missed, including poor function recovery and therapy delays (Jiménez-Sánchez et al., 2020). Nonetheless, emergency physicians lacking subspecialized training in orthopedics frequently examine radiographs in emergencies out of necessity, and up to four out of five documented diagnostic errors are fracture misdiagnoses (Lindsey et al., 2018).

In a previous study, a novel DL-based CAD system was introduced, demonstrating the power of the ViT architecture for a more efficient and accurate prediction of femur fracture type according to the AO/OTA classification (Tanzi et al., 2022). The primary aim of this paper is to enhance the findings of state-of-the-art deep learning-based classification of femur fractures according to the AO/OTA system, thereby providing better support for healthcare professionals in accurately classifying femur fractures. For this purpose, findings obtained from the previous study based on the ViT Transformer for the classification of Unbroken, A1, A2, A3, B1, B2, and B3 classes (Tanzi et al., 2022) were further improved, by exploring the current advancement in DL research and changing the architecture of the model. A cutting-edge architecture, YOLOv8 (Jocher et al., 2023), was involved and fine-tuned, taking into careful consideration the model interpretability through the analysis of the attention maps related to the model. The YOLOv8 real-time object detector (Jocher et al., 2023) is the most recent version in the YOLO series and, to the best of our knowledge, is considered the state-of-the-art in terms of speed and accuracy across a broad range of applications, such as object detection (Reis et al., 2023; Talaat and ZainEldin, 2023), semantic segmentation (Yue et al., 2023; Pandey et al., 2023), and image classification (Al Mudawi et al., 2023; Quach et al., 2024). Moreover, to overcome the skepticism about the use of deep learning in the medical field due to their challenging interpretation and enigmatic prediction-making process, the Eigen-CAM model (Muhammad and Yeasin, 2020) was used to compute the principal components of the learned features from the convolutional layers and to visually display the regions that the network mostly concentrated on to provide the prediction.

This approach has been conceived to foster the comparisons between artificial intelligence and human judgment, especially for borderline cases. In addition, data augmentation techniques were applied at the preprocessing stage by applying image processing transformations to the input images.

2. Materials and methods

An overview of the proposed approach is shown in Fig. 1. The original database was initially divided into training, validation, and test sets. Employing data augmentation techniques, the training portion's images were processed to double the number of training images and, as a result, optimize the algorithm's training phase. Once the trained model was refined, it was assessed utilizing a set of test images that hadn't been seen before. Given the test set as input, the trained model can predict the type of fracture and matching attention map for each image.

The dataset, the architecture of the YOLOv8 DL model, and the training process and metrics are described in the following sections.

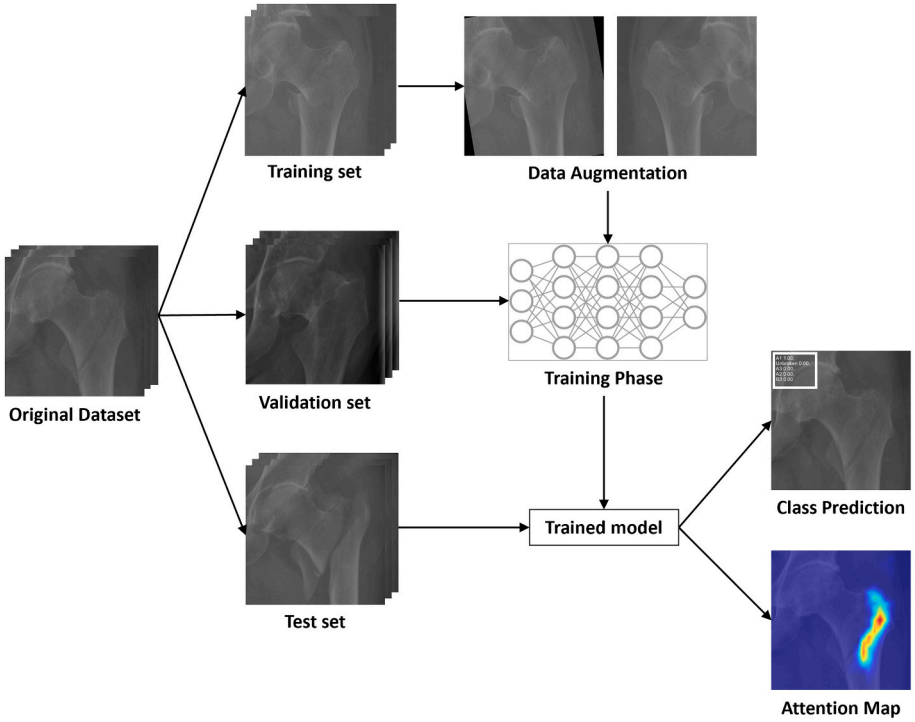


Fig. 1. Overview of the proposed system. The original database was divided into training, validation, and test sets. The training images were processed through data augmentation to optimize the training phase. The trained model, given the test set as input, can predict for each image the fracture type and the corresponding attention map.

2.1. Original dataset

The initial dataset was retrieved from the previous work (Tanzi et al., 2022). Patients' details are provided in Table 1. The dataset was composed of 4233 manually annotated images. To guarantee the ground truth, three orthopaedic surgeons carried out the labeling process. The images were divided into different fracture types: 1976 Unbroken femurs, 535 type A1, 473 type A2, 171 type A3, 637 type B1, 333 type B2, and 108 type B3. Because C-type femur fractures are so uncommon, they were not included in the current study's data. From this initial dataset, 85 % was used for training, and 15 % for validation and testing. Consequently, the dataset is composed of 3587 training images, 646 test images, and 289 validation images, which are a subset of the test ones.

2.2. Data augmentation

To further improve the system reliability and the challenge of acquiring labeled medical data, the number of images of the initial dataset has been increased through data augmentation (Xu et al., 2023; Liu et al., 2024; Beddiar et al., 2023; Islam et al., 2024; Ayan and Unver, 2018; Hussain et al., 2017). Appropriate changes were chosen with the medical team's assistance to increase the quantity and heterogeneity of the database without compromising the data quality (Kora Venu and Ravula, 2021). To prevent bias in the testing phase and enhance the robustness of the learning process, data augmentation was exclusively applied during the training phase. No data augmentation was employed during testing. Furthermore, a comparative analysis of several configurations of transformations was performed to evaluate the effective advantages.

Table 1
Dataset details.

Age (*yr*)	Median (IQR)	81 (73–86)
Sex	F	67.5 %
	M	32.5 %
Ethnicity	Caucasian	93 %
	African	4 %
	Hispanic	2 %
	Asian	1 %

The following transformations were selected: horizontal flipping, vertical flipping, rotation, color jitter in brightness and contrast, inversion, equalization, sharpness adjustment, auto contrast, and Gaussian blurring. Fig. 2 shows a visual example of each transformation concerning a sample picture. Firstly, a horizontal flip was applied, which mirrors the image along the vertical axis, and a vertical flip, which mirrors it along the horizontal axis. These flips assist the network in recognizing femurs in different orientations. Secondly, each image was rotated by 10 degrees, thus exposing the network to images that were slightly tilted in various directions. To diversify the dataset color jittering was employed, which randomly adjusts brightness and contrast. Depending on the chosen parameters, the contrast and brightness can be reduced by up to 50 % or increased by up to 50 % compared to their original value. This transformation simulates the variability of lighting conditions and color profiles according to real-world imaging scenarios. Gaussian blur was applied to each image, with a randomly chosen kernel size between 5 and 9 and a standard deviation value ranging from 0.1 to 5. This technique simulates various degrees of image blurring, enhancing the network's ability to recognize key features even when the images are not sharp. Furthermore, random inversion and sharpness adjustment were added, which reversed the colors and varied the sharpness to enhance or soften details. These transformations help to handle different levels of contrast and sharpness. Additionally, autocontrast and equalization techniques were employed. The former adjusts the contrast of the image automatically based on its histogram, the latter redistributes the brightness values to spread them out more evenly across the image. Each image in the training set was subjected to a single transformation at a time, resulting in nine new augmented images per original image. Subsequently, some images were drawn such that the classes were numerically twice as large as the initial training classes. The augmented images were randomly selected from the set of images

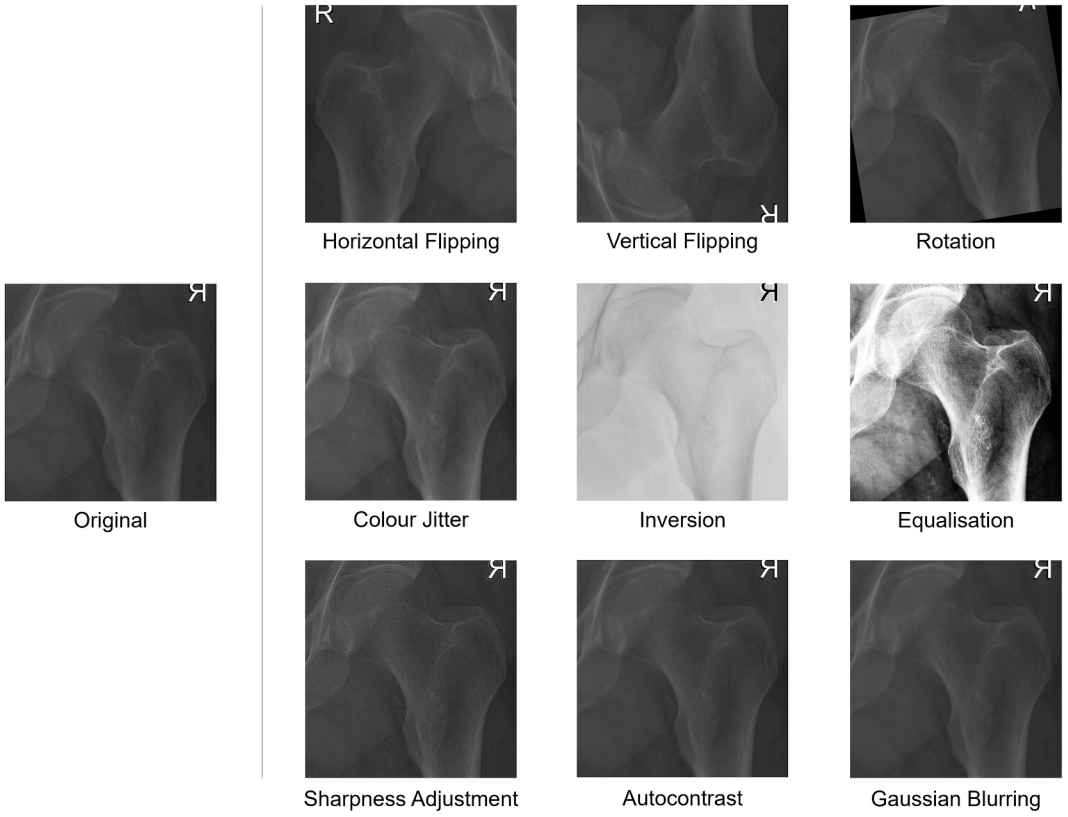


Fig. 2. Graphical representation of the selected transformation applied during the data augmentation process. The original unbroken femur image is shown on the left and the nine augmented images obtained from the applied transformations are arranged in the three rows on the right.

obtained through rotation, vertical and horizontal flipping and thus did not transform in contrast and brightness, and from the images obtained through the other transformations whose new luminosity and contrast was between the 25th and 75th percentile of the actual training images (Goceri, 2023). The best configuration of augmented transformations, which minimized false positives and negatives, particularly between configurations A1 and A2 and A1 and B3 was obtained by doubling the numerosity of samples for each class of the training dataset, creating a new training dataset of 7174 images.

2.3. *Deep learning model architecture*

With its release at the start of 2023, Ultralytics' YOLOv8 model established a new benchmark for numerous tasks, including posture estimation, object detection, instance segmentation, and classification. This powerful deep learning model consists of a convolutional neural network with three primary building blocks: head, neck, and backbone. A graphical representation of the YOLOv8 model is shown in Fig. 3. Significant features are extracted from the input by the backbone, further referred to as the feature extractor. The neck behaves as a connector, carrying out actions related to feature fusion and incorporating contextual information. The head is the final part of the network and is responsible for generating the outputs. To solve the image classification problem, YOLOv8 employs a variant of the EfficientNet architecture and several model configurations, all pre-trained on the ImageNet dataset. The five model configurations are known as nano, small, medium, large, and extra-large, arranged in ascending order according to the depth, width, and number of parameters of the architecture layers. A lighter, faster, but less accurate trained model could come from a configuration with fewer, simpler layers. On the other hand, a more complex model results in a higher accuracy, a heavier architecture, and slower performance. Because the model configurations range concerning speed, accuracy, and parameters, this architecture is more adaptable to the diverse applications' accuracy and computation needs. Moreover, a collection of augmentation methods is automatically integrated into the training process to add variability and improve the

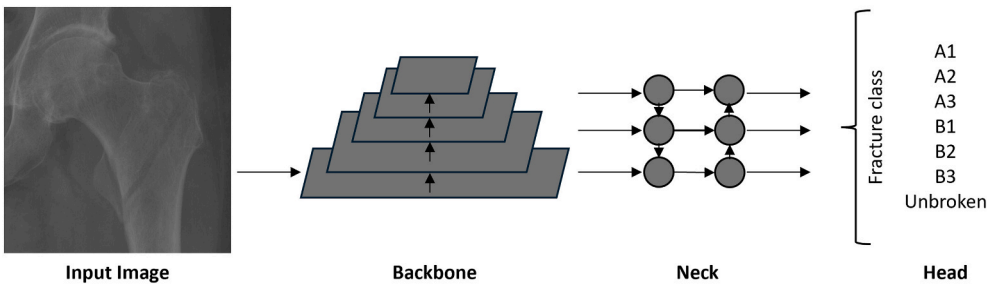


Fig. 3. Overview of the Yolov8 model architecture. The head, neck, and backbone are the three main structural components of the architecture. The backbone extracts important information from the input. The neck functions as a connector, integrating contextual data with feature fusion. The head, which is the last component of the network, is in charge of producing the fracture class as output.

capability of the model to generalize to new data. A benchmark concerning the different Yolov8 configurations was performed to assess the most suitable setup considering the non-augmented dataset and default training and data augmentation parameters, confirming that the YOLOv8-x configuration performed the best overall. This is consistent with the complexity of the medical issue, which necessitates a tailored network in terms of architecture and a greater number of parameters that must be optimized to identify the different features within the medical images and produce a reliable prediction.

Since YOLO was created for a generalist context, an additional effort has been required to match the medical situation. In particular, the original dataset was affected by a lack of images related to the B3 class, representative of a less frequent event. Despite at each epoch, the model sees a slightly different variation of the images through YOLO data augmentation, this operation does not increase the number of images. Consequently, the original dataset was increased through data augmentation to mitigate this issue and further enhance the other classes, as described in Section 2.1. It was also observed that YOLO's data augmentation sometimes produces a lot of artifacts, making it difficult to track an image back to the source, as visible from the sample training batch shown in Fig. 4. This led to the deactivation of certain default transformations. In particular, the erasing transform, which randomly erases a portion of the image, was deactivated to avoid that, if the portion containing the fracture were randomly deleted, the network would learn about incorrect portions of the image. Additionally, the mosaic augmentation, which entails piecing four images together, degrades performance if used for the full training program. Since it was empirically shown that turning the mosaic augmentation off for the final

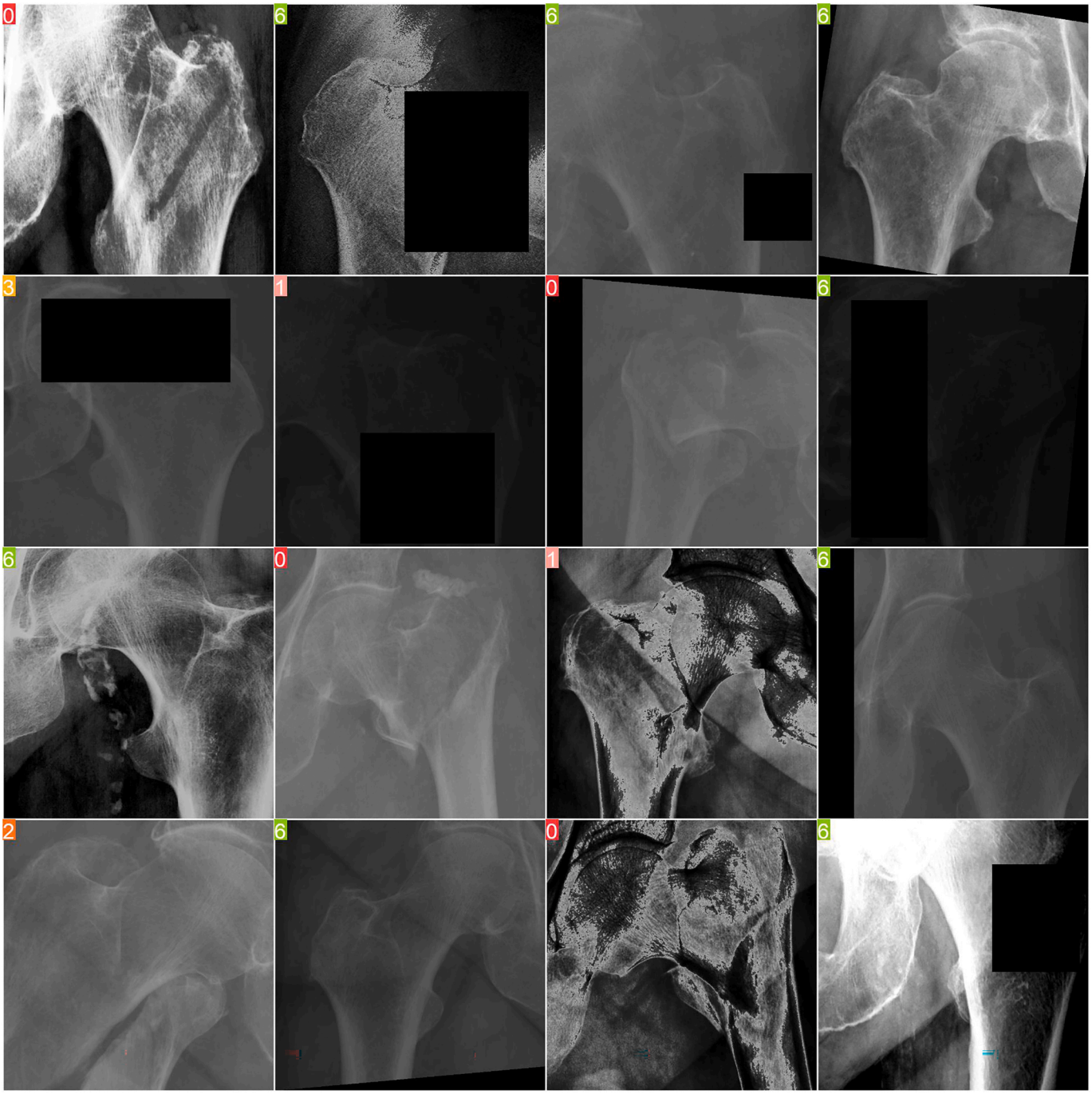


Fig. 4. Example of images belonging to a training batch, transformed through default YOLO data augmentation.

ten epochs is beneficial, this parameter was adjusted with the number of epochs.

2.4. Data interpretability

The healthcare domain requires the model to be as interpretable as possible to improve the prediction's reliability and effectively support physicians' assessment. This includes explaining to the user what the model learns from the data or why it performs badly in a particular situation. Eigen-CAM is a class discriminative method that has proven its robustness against classification errors made by fully connected layers in CNNs. It is one of the models that allows obtaining a class activation map to improve model interpretability. It does not rely on the back-propagation of gradients, class relevance score, maximum activation locations, or any other form of weighting features (Muhammad and Yeasin, 2020). To compute and show the key elements of the learned features/representations from the convolutional layers and to produce visual explanations of the output decision by identifying the region in the input space that generates the decision itself, this model was integrated into the YOLOv8 architecture. The comparison between the areas that the deep learning model prioritizes to get the prediction and the ones that the physician uses to identify the type of fracture allowed for major comprehension of how the algorithm works.

2.5. Training and metrics

The model ran on a Windows 11 system with an NVIDIA Quadro RTX 6000 GPU, using Python-3.11.8 and PyTorch 2.2.2 + cu118 library. The final classification model, which contained 183 layers, 56,150,807 parameters, and 56,150,807 gradients, was trained for 50 epochs and a batch size of 8, requiring approximately an hour and a half. However, in most cases about 20 epochs proved sufficient for the model to learn, as shown in Fig. 5.

Furthermore, all the input images were resized to 800 × 800 to fit the maximum computational capacity of the GPU. To further enhance generalization, label smoothing was additionally applied to ground truth labels, softening hard labels to a combination of the target labels and a uniform distribution across labels. The AdamW optimizer was adopted with an initial learning rate of 0.000714 and a momentum of 0.9. Besides, the cosine learning rate scheduler adjusted the learning rate following a cosine curve the learning rate over time for better convergence.

YOLOv8's emphasis on preserving the best possible balance between speed and accuracy makes it appropriate for real-time classification (Jocher et al., 2023). Indeed, computational power is mainly required for the training phase, while the inference could be easily implemented even in mid-range devices commonly used in hospitals.

For what concerns evaluation metrics, Cross Entropy loss was chosen to assess the convergence of the model during training, while accuracy, precision, recall, and F1 score were selected to evaluate the results to obtain a reliable result even if the dataset is not perfectly balanced.

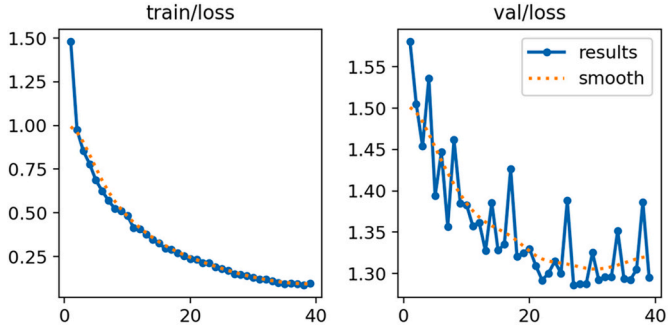


Fig. 5. Training and validation loss trend.

Specifically, the following formula was used to calculate the classification metrics. Given the values of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), accuracy, precision, and recall are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The f1-score is defined as the harmonic mean between precision and recall and it is computed as:

$$F1 = 2^{*}\frac{precision^{*}recall}{precision + recall}$$

Because precision and recall have an inverse relationship, the F1 score functions as a performance indicator by integrating both values through harmonic mean. To give a more comprehensive overview of the performance, the metrics for accuracy, precision, recall, and f1-score were calculated separately for each class and then macro-averaged.

3. Results

The optimized combination of parameters, in terms of both data augmentation and neural network architecture, led the trained model to obtain an overall accuracy of 0.9 and a general improvement concerning all the evaluation metrics compared to the standard YOLOv8 configuration and the state-of-the-art models based on the same classification systems and number of classes. In particular, compared to the earlier work based on the ViT model (Tanzi et al., 2022), which achieved an average accuracy of 0.83, 0.77 for precision and F1-score, and 0.76 for recall, the Yolov8-based model outperformed the previous model by almost 10 % across all metrics. Fig. 6 and Table 2 report the normalized confusion matrix and the corresponding quantitative results, in terms of accuracy, precision, recall, and F1-score, respectively. Evaluating metrics associated with false positives and false negatives is of utmost importance to assess the actual effectiveness of the proposed methodology both in terms of performance and interpretability.

As can be seen in Table 2, the most critical class is B3, as expected due to issues related to the number of samples and similarities in fracture physiological features, that remarkably increase the learning challenge.

Attention maps were originated jointly with predictions to demonstrate model reliability and improve its interpretability. Fig. 7 reports some examples of images with the prediction outputs on the top left corner and the corresponding attention maps, obtained through the Eigen-CAM model on the right.

The image shows that the accuracy in the prediction probability is always higher than 90 %. Furthermore, it is clear from the attention maps that the model focuses on specific parts of the bone to determine the type of fracture, indicating that the model learns the areas of the X-ray on which to focus to obtain a correct classification.

4. Discussion

The present study aimed to enhance the classification accuracy of femur fractures according to the AO/OTA system by utilizing the YOLOv8 architecture and incorporating advanced interpretability measures. The outcomes demonstrate significant improvements over our previous model (Tanzi et al., 2022), providing valuable insights for clinical applications and highlighting the potential of deep learning (DL) in medical image analysis, by assisting the doctor in making an accurate diagnosis in less time, limiting errors, minimizing the diagnosis phase

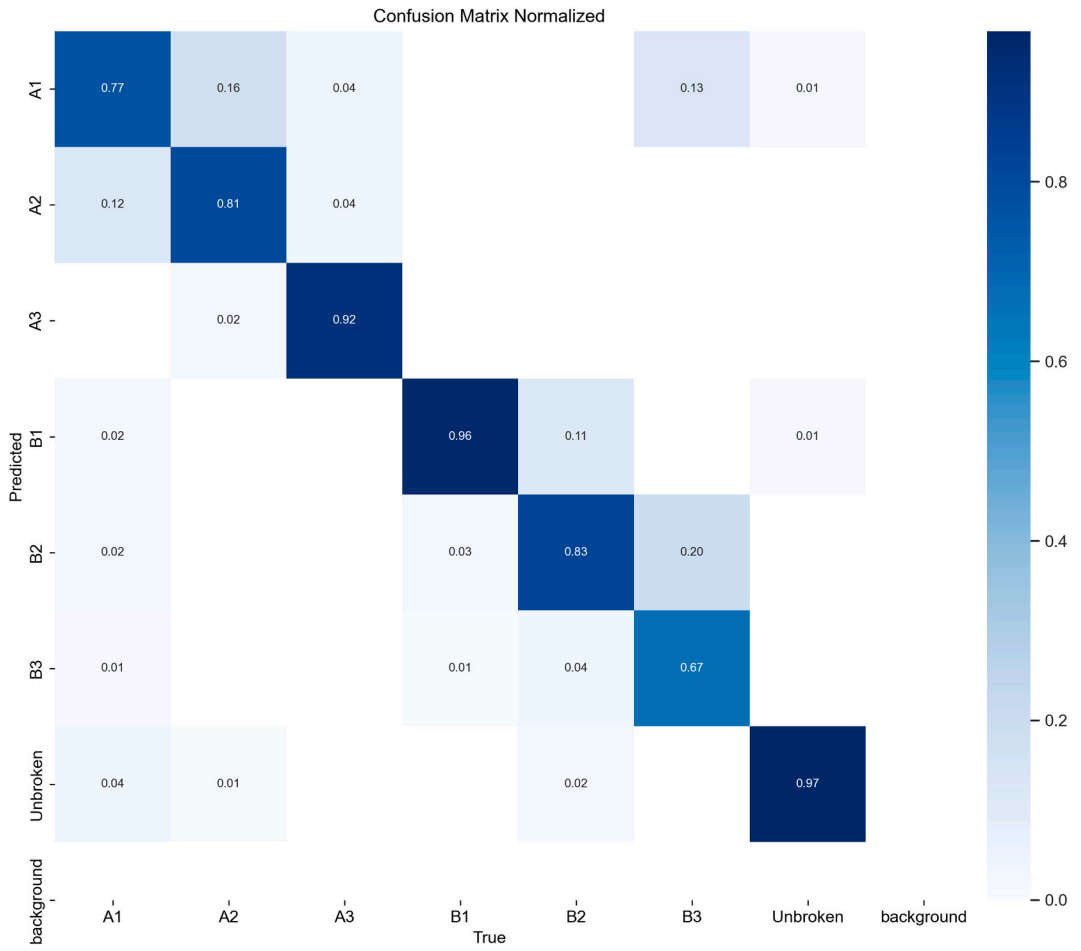


Fig. 6. Normalized confusion matrix related to fine-tuned YOLOv8 model.

Table 2
Values of accuracy, precision, recall, and F1-score concerning the final fine-tuned Yolov8 architecture.

Class	Accuracy	Precision	Recall	F1-score
A1	0.77	0.78	0.72	0.75
A2	0.81	0.81	0.88	0.84
A3	0.92	0.92	0.92	0.92
B1	0.96	0.96	0.88	0.91
B2	0.83	0.83	0.83	0.83
B3	0.67	0.67	0.71	0.69
Unbroken	0.97	0.97	0.98	0.97
Macro AVG	0.90	0.85	0.85	0.85

bottleneck, and taking immediate action to help the patient recover more quickly and effectively.

Despite the excellent results reported by algorithms in detecting hip fractures on X-ray images (Lee et al., 2020; Potter et al., 2024), there is still a lack of comprehensive studies focused on classifying fractures into subtypes. Accurate therapeutic management strongly depends on precise characterization of the fracture to plan the treatment properly and predict potential complications in advance. Our model, enhanced with data augmentation techniques and fine-tuning, achieved a macro average accuracy of 90 % classifying not only fractured and unbroken bones but discriminating among 7 classes. In particular, the model accurately classified A3, B1, and unbroken femurs, indicating its robustness in identifying these categories. However, challenges remain in accurately classifying B3 fractures, which may be attributed to the limited number of samples.

In the medical field, collecting a comprehensive dataset can be challenging due to the relative rarity of some specific fractures (Potter et al., 2024; Sundkvist et al., 2021), stringent privacy regulations (Dierks et al., 2021), and the labor-intensive process of manually annotating medical images (Grünberg et al., 2017). For instance, Lin et al. (Haj-Mirzaian et al., 2020) demonstrated that while traditional augmentation methods can improve model performance, they are insufficient to address specific fracture patterns' scarcity. Their study utilized Generative Adversarial Networks (GANs) to create synthetic images, significantly enhancing the performance of their hip fracture detection model. Similarly, a study by Mutasa et al. (2020) showed that using advanced augmentation techniques like digitally reconstructed radiographs (DRRs) and GANs can substantially improve the accuracy of femoral neck fracture classification applying Garden classification system, achieving an Area Under Curve (AUC) of 0.92 with DRRs and 0.87 with GANs.

However, while data augmentation can significantly enhance model performance, it has limitations. Traditional methods like flipping and rotating images are beneficial but insufficient to fully capture the complexity and variability of medical images. For instance, recent studies have highlighted that classic augmentation methods do not significantly increase the diversity needed for small datasets (Lin and Chung, 2019). GANs have shown promise in this area by creating new, realistic images that add substantial variability to the dataset, thereby improving model performance even further (Mutasa et al., 2020). Nonetheless, these advanced techniques also have limitations, as they depend on the quality and quantity of the original data. Therefore, while data augmentation improves generalization and performance, it cannot completely overcome the challenges posed by the limited availability of certain medical images. To further enhance the robustness and

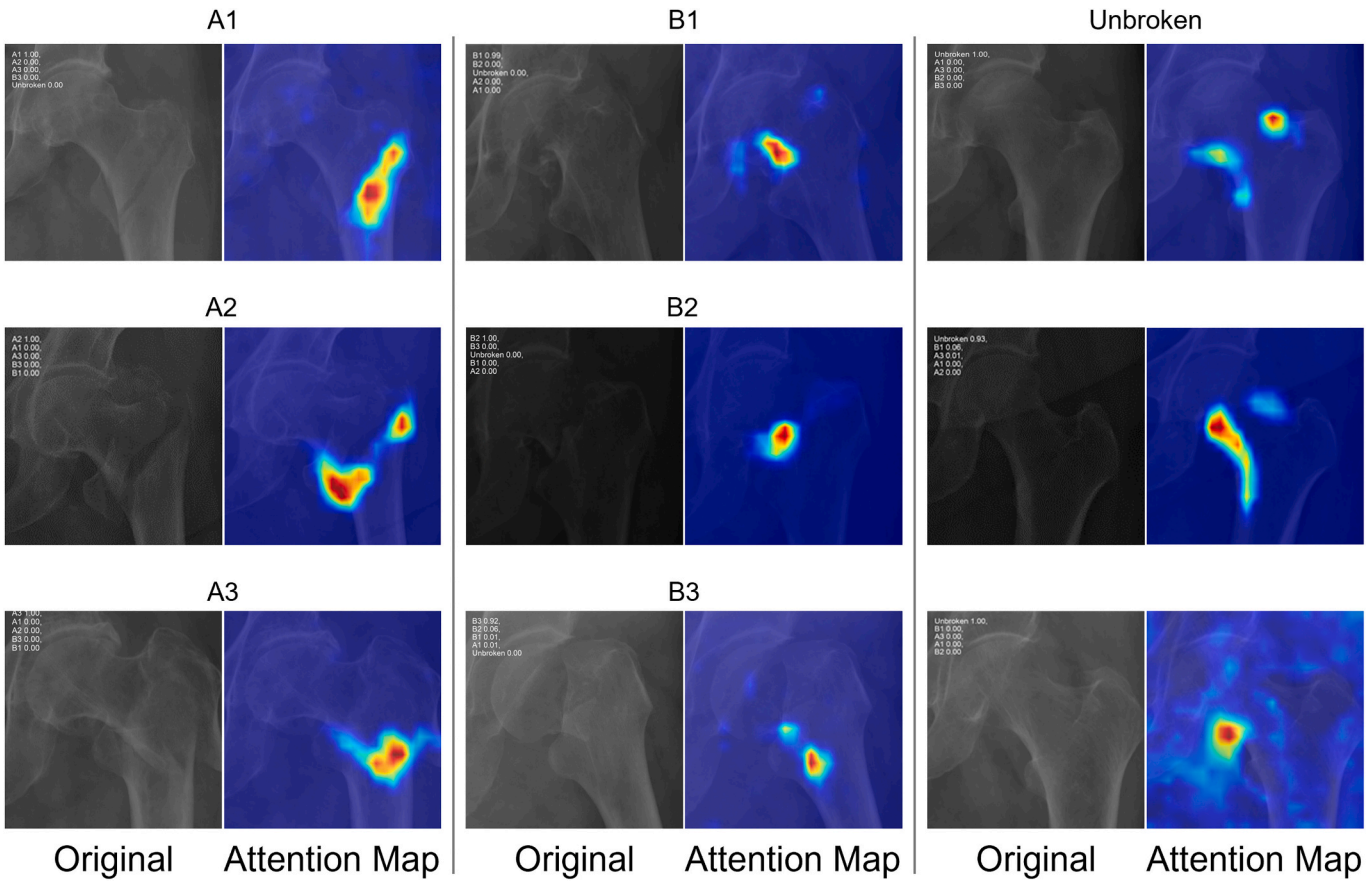


Fig. 7. Examples of original images with the corresponding prediction (on the top left corner) and attention map for each class.

reliability of CAD tools, combining data augmentation with other strategies, such as collaborative data sharing across institutions and the inclusion of synthetic data generated by advanced techniques, is essential (Prediger et al., 2024).

One of the key contributions of this study is the incorporation of the Eigen-CAM (Muhammad and Yeasin, 2020) model to enhance the interpretability of the YOLOv8 predictions. Explainable AI (XAI) methods are crucial in medical applications to ensure transparency and trust in the automated decision-making process (Lötsch et al., 2021; Antoniadi et al., 2021). The attention maps generated by Eigen-CAM provide visual insights into the regions of the X-ray images that the model focuses on to make its predictions. This not only helps in validating the model's decision-making process but also aids clinicians in understanding, trusting, or even out-performing the AI system. Indeed, Tanzi et al.'s previous study (Tanzi et al., 2022) showed that using the CAD system in conjunction with diagnostic expertise increased the diagnosis accuracy of 11 clinicians by a factor of 29 %, even surpassing the performance of the algorithm. Similarly, Deep learning models could significantly reduce fractures' misinterpretation rate (47.0 % (95 % CI, 37.4–53.9 %) in emergency medicine clinicians who lack specific training in orthopedics (Lindsey et al., 2018). Moreover, A deep learning-based AI model significantly improved the performance of inexperienced radiologists (0.094 (95 % CI, 0.020–0.168; $p = 0.012$)) and emergency physicians (0.069 (95 % CI, 0.002–0.136; $p = 0.043$)) in diagnosing pediatric skull fractures on plain radiographs (Choi et al., 2022). This suggests that combining technology and medical knowledge can improve clinical judgment and strengthen skills, above all in less experienced physicians. The attention maps suggest that the model prioritizes relevant anatomical features of the femur, aligning with clinical expectations and reinforcing the predictions' reliability. Despite the promising results, this study has some limitations. The dataset, although augmented, may still lack sufficient diversity to cover all possible variations of femur fractures. The performance discrepancies observed in the B3 classes suggest that augmentation techniques are not sufficient where a dataset class is strongly unbalanced and additional data is required to improve the model's robustness. Furthermore, the black-box nature of deep learning models poses challenges in fully understanding their decision-making processes net of Eigen-CAM employment. Future research should focus on integrating more advanced XAI techniques to further elucidate the model's internal workings and enhance its interpretability. Additionally, expanding the dataset to include more diverse and rare fracture types, such as C-type fractures and multiple fractures, could provide a more refined tool to assist physicians during diagnosis. Collaborative efforts with medical institutions to gather extensive and diverse datasets, along with continuous refinement of the model's architecture and training protocols, are essential steps toward achieving higher accuracy and reliability.

5. Conclusion

DL-based CAD systems have shown great promise in the detection of femur fractures, providing improved precision and effectiveness over conventional diagnostic techniques. These devices can accurately and quickly analyze medical images, spot fractures, and help radiologists by cutting down on diagnostic time and the possibility of human mistakes. This study demonstrates the potential of the YOLOv8 architecture in improving the classification of femur fractures according to the AO/OTA system, thereby supporting healthcare professionals in making accurate and timely diagnoses. To the best of our knowledge, the proposed methodology outperforms the current state-of-the-art solutions by discriminating among seven different fracture sub-classes. The

incorporation of interpretability measures further enhances the model's applicability in clinical settings by providing transparency and fostering trust in the AI system, combining physicians' assessment, and improving the doctor-patient relationship. Crucially, technology should be at the service of medical professionals, strengthening their skills and boosting their clinical judgment. Enhancing diagnostic confidence, enabling early intervention, and improving patient outcomes are all possible by combining DL algorithms with medical knowledge.

Funding

This study was carried out within Ministerial Decree no. 1062/2021 and received funding from the FSE REACT-EU-PON Ricerca e Innovazione 2014–2020. This manuscript only reflects the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them.

CRediT authorship contribution statement

Giorgia Marullo: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luca Ulrich:** Writing – review & editing, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Francesca Giada Antonaci:** Writing – original draft, Visualization, Methodology, Investigation, Data curation. **Andrea Audisio:** Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Alessandro Massè:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Enrico Vezzetti:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Al Mudawi, N., Qureshi, A., Abdelhaq, M., Alshahrani, A., Alazeb, A., Alonazi, M., Algarni, A., 2023. Vehicle detection and classification via YOLOv8 and deep belief network over aerial image sequences. Sustainability 15, 14597.

Alzaid, A., Wignall, A., Dogramadzi, S., Pandit, H., Xie, S.Q., 2022. Automatic detection and classification of peri-prosthetic femur fracture. Int. J. Comput. Assist. Radiol. Surg. 17, 649–660.

Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C., 2021. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Appl. Sci. 11, 5088.

Ayan, E., Unver, H., 2018. Data augmentation importance for classification of skin lesions via deep learning. In: 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018, pp. 1–4. https://doi.org/10.1109/EBBT.2018.8391469.

Bäcker, H.C., Wu, C.H., Maniglio, M., Wittekindt, S., Hardt, S., Perka, C., 2021. Epidemiology of proximal femoral fractures. Journal of Clinical Orthopaedics and Trauma 12, 161–165.

Beddiar, D., Oussalah, M., Muhammad, U., Seppänen, T., 2023. A deep learning based data augmentation method to improve COVID-19 detection from medical imaging. Knowl.-Based Syst. 280.

Bernardini, M., Romeo, L., Misericordia, P., Frontoni, E., 2019. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. IEEE J. Biomed. Health Inform. 24, 235–246.

Bulten, W., Balkenhol, M., Belinga, J., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farré, X., Geronatsiou, K., Molinié, V., et al., 2021. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod. Pathol. 34, 660–671.

Cannavò, A., D'Alessandro, A., Maglione, D., Marullo, G., Zhang, C., Lamberti, F., et al., 2020. Automatic generation of affective 3d virtual environments from 2d images. In: Proc. 15th International Conference on Computer Graphics Theory and Applications (GRAPP 2020). SCITEPRESS, pp. 113–124.

Checcucci, E., Piazzolla, P., Marullo, G., Innocente, C., Salerno, F., Ulrich, L., Moos, S., Quarà, A., Volpi, G., Amparore, D., et al., 2023. Development of bleeding artificial intelligence detector (blair) system for robotic radical prostatectomy. J. Clin. Med. 12, 7355.

Choi, J., Cho, Y., Ha, J., Lee, Y., Koh, S., Seo, J., Choi, Y., Cheon, J., Phi, J., Kim, I., et al., 2022. Deep learning-assisted diagnosis of pediatric skull fractures on plain radiographs. Korean J. Radiol. 23, 343.

D., C.P., 1995. Fractures of the femur after hip replacement. Instr. Course Lect. 44, 293–304.

Dierks, C., Kircher, P., Husemann, C., Kleinschmidt, J., Haase, M., 2021. Data Privacy in European Medical Research: A Contemporary Legal Opinion, Vol. 18. MWV Medizinisch Wissenschaftliche Verlagsgesellschaft.

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Commun. ACM 63, 68–77.

Dyer, S.M., Crotty, M., Fairhall, N., Magaziner, J., Beaupre, L.A., Cameron, I.D., Sherrington, C., F. F. N. F. R. R. S. I. Group, 2016. A critical review of the long-term disability outcomes following hip fracture. BMC Geriatr. 16, 1–18.

Fischer, H., Maleitzke, T., Eder, C., Ahmad, S., Stöckle, U., Braun, K., 2021. Management of proximal femur fractures in the elderly: current concepts and treatment options. Eur. J. Med. Res. 26, 1–15.

Garden, R.S., 1961. Low-angle fixation in fractures of the femoral neck. The Journal of Bone & Joint Surgery British 43, 647–663.

Goceri, E., 2023. Medical image data augmentation: techniques, comparisons and interpretations. Artif. Intell. Rev. 56, 12561–12605.

Göndöcs, D., Dörfler, V., 2024. AI in medical diagnosis: AI prediction & human judgment. Artif. Intell. Med. 149, 102769.

Grünberg, K., Jimenez-del Toro, O., Jakab, A., Langs, G., Salas Fernandez, T., Winterstein, M., Weber, M.-A., Krenn, M., 2017. Annotating medical image data. In: Cloud-Based Benchmarking of Medical Image Analysis, pp. 45–67.

Gullberg, B., Johnell, O., Kanis, J., 1997. World-wide projections for hip fracture. Osteoporos. Int. 7, 407–413.

Haj-Mirzaian, A., Eng, J., Khorasani, R., Raja, A.S., Levin, A.S., Smith, S.E., Johnson, P.T., Demehri, S., 2020. Use of advanced imaging for radiographically occult hip fracture in elderly patients: a systematic review and meta-analysis. Radiology 296, 521–531.

Hao, J., Kim, Y., Kim, T.-K., Kang, M., 2018. Pasnet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. BMC Bioinformatics 19, 1–13.

Hekler, A., Utikal, J., Enk, A., Hauschild, A., Weichenthal, M., Maron, R., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al., 2019. Superior skin cancer classification by the combination of human and artificial intelligence. Eur. J. Cancer 120, 114–121.

Hussain, Z., Gimenez, F., Yi, D., Rubin, D., 2017. Differential data augmentation techniques for medical imaging classification tasks, AMIA... Annual Symposium proceedings. In: AMIA Symposium 2017, pp. 979–984.

Islam, T., Hafiz, M., Jim, J., Kabir, M., Mridha, M., 2024. A systematic review of deep learning data augmentation in medical imaging: recent advances and future research directions. Healthcare Analytics 5.

Jiménez-Sánchez, A., Kazi, A., Albarqouni, S., Kirchhoff, C., Biberthaler, P., Navab, N., Kirchhoff, S., Mateus, D., 2020. Precise proximal femur fracture classification for interactive training and surgical planning. Int. J. Comput. Assist. Radiol. Surg. 15, 847–857.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics yolov8. URL: https://github.com/ultralytics/ultralytics.

Johnell, O., Kanis, J., 2005. Epidemiology of osteoporotic fractures. Osteoporos. Int. 16, S3–S7.

Kaji, D.A., Zech, J.R., Kim, J.S., Cho, S.K., Dangayach, N.S., Costa, A.B., Oermann, E.K., 2019. An attention based deep learning model of clinical events in the intensive care unit. PLoS One 14, e0211057.

Kim, J., Kim, H.J., Kim, C., Kim, W.H., 2021. Artificial intelligence in breast ultrasonography. Ultrasonography 40, 183.

Kora Venu, S., Ravula, S., 2021. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest X-ray images. In: Future Internet, 13. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 8.

Krogue, J.D., Cheng, K.V., Hwang, K.M., Toogood, P., Meinberg, E.G., Geiger, E.J., Zaid, M., McGill, K.C., Patel, R., Sohn, J.H., et al., 2020. Automatic hip fracture identification and functional subclassification with deep learning. Radiology: Artificial Intelligence 2, e190023.

Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M. J., Lange, J., Thiesson, B., 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. 11, 3852.

Lee, H., Kim, S.T., Ro, Y.M., 2019. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9. Springer, pp. 21–29.

Lee, C., Jang, J., Lee, S., Kim, Y.S., Jo, H.J., Kim, Y., 2020. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. Sci. Rep. 10, 13694.

Lin, Y.-J., Chung, I.-F., 2019. Medical data augmentation using generative adversarial networks: X-ray image generation for transfer learning of hip fracture detection. In: 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, pp. 1–5.

Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., et al., 2018. Deep neural network improves fracture detection by clinicians. Proc. Natl. Acad. Sci. 115, 11591–11596.

Liu, X., Ono, K., Bise, R., 2024. A data augmentation approach that ensures the reliability of foregrounds in medical image segmentation. Image Vis. Comput. 147.
Lötsch, J., Kringel, D., Ultsch, A., 2021. Explainable artificial intelligence (xai) in biomedicine: making AI decisions trustworthy for physicians and patients. BioMedInformatics 2, 1–17.
Meinberg, E.G., Agel, J., Roberts, C.S., Karam, M.D., Kellam, J.F., 2018. Fracture and dislocation classification compendium—2018. J. Orthop. Trauma 32, S1–S10.
Muhammad, M.B., Yeasin, M., 2020. Eigen-cam: class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–7.
Mutasa, S., Varada, S., Goel, A., Wong, T.T., Rasiej, M.J., 2020. Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. J. Digit. Imaging 33, 1209–1217.
Pandey, S., Chen, K., Dam, E., 2023. Comprehensive multimodal segmentation in medical imaging: combining yolov8 with sam and hq-sam models. In: Proceedings Of The IEEE/CVF International Conference On Computer Vision, pp. 2592–2598.
Panigutti, C., Perotti, A., Pedreschi, D., 2020. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 629–639.
Porumb, M., Stranges, S., Pescapè, A., Pecchia, L., 2020. Precision medicine and artificial intelligence: a pilot study on deep learning for hypoglycemic events detection based on ecg. Sci. Rep. 10, 170.
Potter, Í. Yıldız, Yeritsyan, D., Mahar, S., Kheir, N., Vaziri, A., Putman, M., Rodriguez, E. K., Wu, J., Nazarian, A., Vaziri, A., 2024. Proximal femur fracture detection on plain radiography via feature pyramid networks. Sci. Rep. 14, 12046.
Prediger, L., Jälkö, J., Honkela, A., Kaski, S., 2024. Collaborative learning from distributed data with differentially private synthetic data. BMC Medical Informatics and Decision Making 24 (1), 167.
Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A., Pattichis, C., 2019. Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, pp. 817–821.
Qi, Y., Zhao, J., Shi, Y., Zuo, G., Zhang, H., Long, Y., Wang, F., Wang, W., 2020. Ground truth annotated femoral X-ray image dataset and object detection based method for fracture types classification. IEEE Access 8, 189436–189444.
Quach, L., Quoc, K., Quynh, A., Ngoc, H., Nghe, N., 2024. Tomato health monitoring system: tomato classification, detection, and counting system based on YOLOv8 model with explainable MobileNet models using Grad-CAM++. IEEE Access 12, 9719–9737.
Rai, A., 2020. Explainable AI: from black box to glass box. J. Acad. Mark. Sci. 48, 137–141.
Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., Farooq, A., 2019. Human augmentation: past, present and future. International Journal of Human-Computer Studies 131, 131–143.
Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, C.P., Lungren, M.P., Ng, A.Y., Patel, B.N., 2020. Appendixnet: deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining. Sci. Rep. 10, 3958.
Rajula, H.S.R., Verlato, G., Manchia, M., Antonucci, N., Fanos, V., 2020. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina 56, 455.
Reis, D., Kupec, J., Hong, J., Daoudi, A., 2023. Real-time flying object detection with YOLOv8. ArXiv Preprint. https://doi.org/10.48550/arXiv.2305.09972. ArXiv: 2305.09972.
Ryan, D.J., Yoshihara, H., Yoneoka, D., Egol, K.A., Zuckerman, J.D., 2015. Delay in hip fracture surgery: an analysis of patient-specific and hospital-specific risk factors. J. Orthop. Trauma 29, 343–348.
Sambrook, P., Cooper, C., 2006. Osteoporosis. Lancet 367, 2010–2018.
Sing, C.-W., Lin, T.-C., Bartholomew, S., Bell, J.S., Bennett, C., Beyene, K., Bosco-Levy, P., Bradbury, B.D., Chan, A.H.Y., Chandran, M., et al., 2023. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. J. Bone Miner. Res. 38, 1064–1075.
Singh, A., Mohammed, A.R., Zelek, J., Lakshminarayanan, V., 2020. Interpretation of deep learning using attributions: application to ophthalmic diagnosis. In: Applications of Machine Learning 2020, 11511. SPIE, pp. 39–49.
Suh, J., Yoo, S., Park, J., Cho, S.Y., Cho, M.C., Son, H., Jeong, H., 2020. Development and validation of an explainable artificial intelligence-based decision-supporting tool for prostate biopsy. BJU Int. 126, 694–703.
Sundkvist, J., Brüggeman, A., Sayed-Noor, A., Möller, M., Wolf, O., Mukka, S., 2021. Epidemiology, classification, treatment, and mortality of adult femoral neck and basicervical fractures: an observational study of 40,049 fractures from the swedish fracture register. J. Orthop. Surg. Res. 16, 1–10.
Talaat, F., ZainEldin, H., 2023. An improved fire detection approach based on YOLO-v8 for smart cities. Neural Comput. & Applic. 35, 20939–20954.
Tanzi, L., Vezzetti, E., Moreno, R., Aprato, A., Audisio, A., Massè, A., 2020. Hierarchical fracture classification of proximal femur X-ray images using a multistage deep learning approach. Eur. J. Radiol. 133, 109373.
Tanzi, L., Audisio, A., Cirrincione, G., Aprato, A., Vezzetti, E., 2022. Vision transformer for femur fracture classification. Injury 53, 2625–2634.
Tariq, A., Gill, A.Y., Hussain, H.K., 2023. Evaluating the potential of artificial intelligence in orthopedic surgery for value-based healthcare. International Journal of Multidisciplinary Sciences and Arts 2, 27–35.
Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al., 2020. Human–computer collaboration for skin cancer recognition. Nat. Med. 26, 1229–1234.
Wang, Y., Li, R., Zheng, P., 2022. Progress in clinical application of artificial intelligence in orthopedics. Digital Medicine 8, 4.
Wu, A.-M., Bisignano, C., James, S.L., Abady, G.G., Abedi, A., Abu-Gharbieh, E., Alhassan, R.K., Alipour, V., Arabloo, J., Asaad, M., et al., 2021. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the global burden of disease study 2019. The Lancet Healthy Longevity 2, e580–e592.
Xu, M., Yoon, S., Fuentes, A., Park, D.S., 2023. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recogn. 137, 109347.
Yang, S., Yin, B., Cao, W., Feng, C., Fan, G., He, S., 2020. Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. Clin. Radiol. 75, 713–e17.
Yang, G., Ye, Q., Xia, J., 2022. Unbox the black-box for the medically explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Information Fusion 77, 29–52.
Yue, X., Qi, K., Na, X., Zhang, Y., Liu, Y., Liu, C., 2023. Improved YOLOv8-Seg network for instance segmentation of healthy and diseased tomato plants in the growth stage. Agriculture 13, 1643.
Zdolsek, G., Chen, Y., Bögl, H.-P., Wang, C., Woisetschläger, M., Schilcher, J., 2021. Deep neural networks with promising diagnostic accuracy for the classification of atypical femoral fractures. Acta Orthop. 92, 394–400.