أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Explainable Knowledge Distillation for Bone Fracture Classification Using X-ray Images

Samaher S. Alsharif, Durar A. Alqahtani, Somayah M. Alqahtani, Layan F. Alsulaiman, Yazan G. Rabaiah, Yasser J. Alqanatish, Dr. Salman Khan

samaher.alsharif@kaust.edu.sa, dorarali369@gmail.com, somayah09876@gmail.com, Layan.Alsulaiman@kaust.edu.sa, znrabaya@gmail.com, y.qanatish@outlook.com, salmank255@gmail.com

Abstract

Bone fracture classification from X-ray images is an important application of deep learning for supporting clinical diagnosis. However, high-capacity models are often computationally expensive and difficult to interpret, limiting their deployment in resource-constrained healthcare environments. This study investigates an Explainable Knowledge Distillation (EKD) framework that combines prediction-based knowledge distillation with class-discriminative attention transfer to develop a lightweight and interpretable fracture classifier. Using the FracAtlas dataset, EfficientNet-B6 was selected and optimized as the teacher model, while MobileNetV2 was selected as the lightweight student. The EKD student was evaluated under three training settings: No Augmentation, RandAugment, and AugMix, and compared with the corresponding MobileNetV2 baselines. EKD improved validation accuracy across all three settings, with AugMix achieving the highest validation accuracy of 0.8742. Qualitative comparisons of teacher and student attention maps showed partial spatial correspondence, although the student maps were generally broader and more diffuse. Overall, the results demonstrate the potential of explainability-guided knowledge distillation for improving lightweight fracture classification while encouraging similarity between teacher and student attention patterns.

Date: August 24, 2026
Code: GitHub Repository
Project Page: Project Page

Contents

1 Introduction **2**

2 Related Work **2**

3 Methodology **3**

3.1 Dataset and Preprocessing . . . 3

3.2 Model Selection and Training Setup . . . 4

3.3 Explainable Knowledge Distillation Framework . . . 5

3.4 Evaluation . . . 6

4 Results **6**

4.1 EKD Classification Performance . . . 6

4.2 Training Behavior . . . 7

4.3 Qualitative Explainability Analysis . . . 7

5 Discussion **8**

6 Limitations **9**

7 Conclusion **9**

7.1 Future Work . . . 9

Acknowledgments **9**

References **10**

1 Introduction

Bone fractures are commonly diagnosed using X-ray imaging due to its accessibility and efficiency, yet radiograph interpretation can be challenging, particularly for subtle fractures and in high-workload clinical environments [1, 2]. Deep learning has shown promising performance for automated fracture classification; however, high-capacity models can be computationally expensive and difficult to interpret.

Knowledge Distillation (KD) can reduce model complexity by transferring predictive knowledge from a high-capacity teacher to a lightweight student [3], while Explainable Artificial Intelligence (XAI) methods such as Grad-CAM provide visual explanations of model predictions [4]. However, existing fracture-classification studies primarily use explainability as a post-hoc analysis, while fracture-oriented KD methods do not explicitly transfer class-discriminative explanations. Although methods such as CAT-KD and Exp-KD incorporate class-attention transfer during distillation, they have mainly been evaluated on general computer-vision benchmarks [5, 6]. This motivates the investigation of explainability-guided knowledge distillation for bone fracture classification.

This work addresses this gap by combining predictive knowledge distillation with class-discriminative attention transfer to train a lightweight fracture classifier. The main contributions are:

- An Explainable Knowledge Distillation framework that jointly transfers predictive knowledge and class-discriminative spatial attention.
- Comparative evaluation and selection of teacher and student architectures across three training settings.
- A composite training objective combining supervised classification, prediction distillation, and attention alignment.
- An evaluation of the lightweight student using classification metrics and qualitative teacher–student Grad-CAM comparisons.

2 Related Work

Deep learning has achieved strong performance in automated bone fracture classification using CNNs, transfer learning, and attention mechanisms [1, 2, 7]. Explainable AI has also been incorporated into fracture-analysis systems, particularly through Grad-CAM, to visualize image regions influencing model predictions [8–10]. However, these explanation maps are primarily used for post-hoc interpretation rather than as supervisory signals during training.

Knowledge distillation has been applied in medical imaging to transfer knowledge from high-capacity models to lightweight students [11]. Fracture-specific approaches have also explored teacher–student learning, including Mixture-of-Experts-based distillation and task-specialized teachers [12, 13]. These methods demonstrate the potential of KD for efficient fracture classification, but do not explicitly distill class-discriminative explanation maps from teacher to student.

Explainability-guided distillation methods such as CAT-KD [5] and Exp-KD [6] transfer class-specific spatial information during distillation, but have primarily been evaluated on general computer-vision benchmarks. In medical imaging, teacher–student explanation maps have also been compared after conventional KD rather than transferred explicitly during training [14]. Table 1 summarizes the approaches most closely related to the proposed framework.

Table 1 Comparison with representative related work.

Work	Domain	Knowledge Transfer	Relation to This Work
Tamizifar et al. [8]	Fracture X-rays	None	Uses Grad-CAM for post-hoc fracture interpretation, not during training.
Sharma et al. [13]	Fracture X-rays	Predictions / features	Applies fracture-oriented KD but does not explicitly transfer class-discriminative explanation maps.
CAT-KD [5]	General vision	Class attention	Transfers class activation maps during KD, but is not evaluated on fracture radiographs.
Exp-KD [6]	General vision	Class activation maps	Transfers explainable spatial knowledge during KD, but is primarily evaluated on general vision benchmarks.
Proposed EKD	**Fracture X-rays**	**Predictions + class attention**	**Combines prediction distillation with class-discriminative attention transfer for lightweight fracture classification.**

Overall, prior fracture-classification studies largely treat explainability as a post-hoc tool, while fracture-oriented KD methods focus on predictive or feature knowledge. Conversely, explainability-guided distillation methods transfer class-discriminative attention but have mainly been studied outside the fracture domain. This gap motivates the proposed EKD framework, which combines prediction-based distillation and class-discriminative attention transfer for lightweight bone fracture classification.

3 Methodology

3.1 Dataset and Preprocessing

Experiments were conducted on the FracAtlas dataset [15], which contains 4,083 musculoskeletal X-ray images, including 3,366 non-fractured and 717 fractured images. Due to this class imbalance, class weighting was applied during training.

The dataset was divided into training, validation, and testing subsets using an 80%, 12%, and 8% split. The official FracAtlas partitions were retained for fractured images, while non-fractured images were randomly divided using the same ratios with a fixed seed of 42. No image overlap was found between the resulting subsets. Table 2 summarizes the final distribution.

Table 2 FracAtlas data distribution.

Split	Total	Non-Fracture	Fracture
Training	3266	2692	574
Validation	485	403	82
Testing	332	271	61

Images were resized to 224×224 using padding-based resizing to preserve aspect ratio and normalized to $[0, 1]$. Three training settings were evaluated: No Augmentation, RandAugment [16], and AugMix [17]. Augmentation was applied only to the training data, while validation and testing data remained unchanged. RandAugment used two random transformations with a magnitude of 0.15.

3.2 Model Selection and Training Setup

ResNet152V2 [18], DenseNet201 [19], and EfficientNet-B6 [20] were evaluated as candidate teacher models, while MobileNetV1 [21] and MobileNetV2 [22] were evaluated as lightweight student candidates. All models were initialized with ImageNet-pretrained weights and adapted for binary fracture classification. A common training configuration, summarized in Table 3, was used for comparison.

Table 3 Main baseline training configuration.

Setting	Value
Input size	224×224
Batch size	32
Maximum epochs	100
Optimizer	Adam
Initial learning rate	1×10^{-5}
Loss function	Categorical Cross-Entropy
Dropout rate	0.2
Fine-tuning	Last 10 trainable layers
ReduceLROnPlateau patience	7 epochs
Early stopping patience	10 epochs
Random seed	42

Baseline experiments were conducted using an NVIDIA T4 GPU, while the subsequent EKD experiments were executed on the KAUST IBEX cluster using an NVIDIA A100 GPU.

3.2.1 Teacher and Student Selection

The candidate models were evaluated under the three training settings. Table 4 summarizes the validation accuracy, F1-score, and AUC used for model selection.

EfficientNet-B6 showed the strongest and most consistent performance among the teacher candidates and was selected as the teacher. MobileNetV2 showed competitive validation performance across the evaluated settings and was selected as the lightweight student for the subsequent EKD experiments.

The selected EfficientNet-B6 teacher was further optimized using Optuna over 30 trials. The optimized configuration used a dropout rate of 0.4, the last 40 trainable layers, a head learning rate of 6.8056×10^{-4}, a fine-tuning learning rate of 3.2079×10^{-5}, and weight decay of 6.2290×10^{-6}. The optimized model was then frozen for EKD training.

Table 4 Validation performance of candidate teacher and student models.

Model	Setting	Accuracy	F1	AUC
ResNet152V2	No Aug.	0.7979	0.8161	0.8441
	RandAug.	0.7835	0.8036	0.8378
	AugMix	0.7918	0.8103	0.8330
DenseNet201	No Aug.	0.7711	0.7928	0.7733
	RandAug.	0.7608	0.7798	0.7469
	AugMix	0.7196	0.7437	0.6816
EfficientNet-B6	No Aug.	0.8412	0.8710	0.8497
	RandAug.	**0.8515**	**0.8858**	**0.8606**
	AugMix	0.8289	0.8595	0.8418
MobileNetV1	No Aug.	0.8000	0.8196	0.8691
	RandAug.	**0.8577**	**0.8658**	**0.8895**
	AugMix	0.8371	0.8498	0.8655
MobileNetV2	No Aug.	0.8454	0.8521	0.8621
	RandAug.	0.8433	0.8519	0.8682
	AugMix	0.8247	0.8371	0.8648

3.3 Explainable Knowledge Distillation Framework

The proposed EKD framework transfers predictive knowledge and class-discriminative spatial information from the frozen EfficientNet-B6 teacher to the MobileNetV2 student. As illustrated in Figure 1, the student is trained using ground-truth labels, temperature-scaled teacher predictions, and class-specific spatial attention.

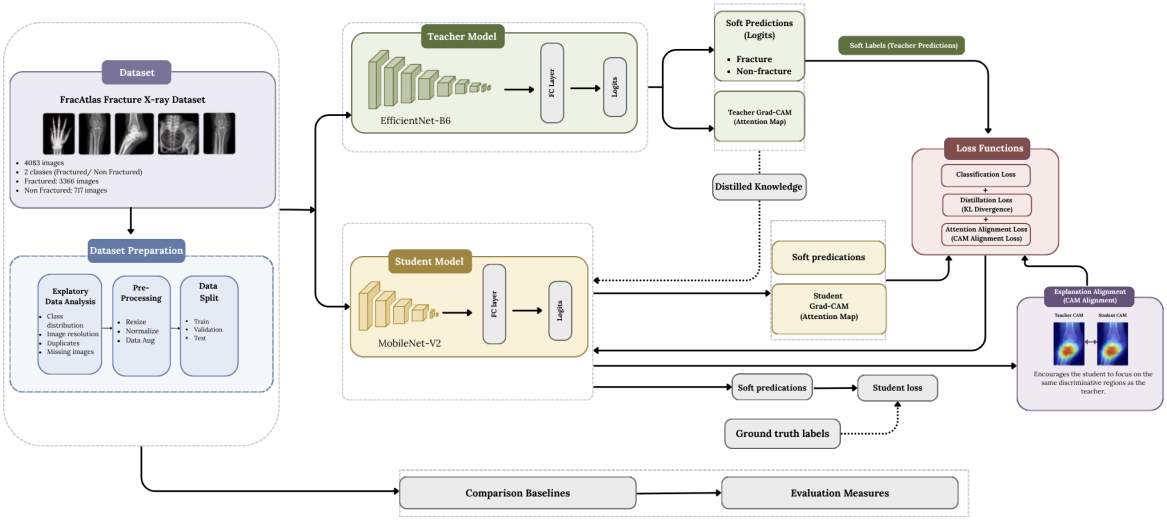


Figure 1 Proposed Explainable Knowledge Distillation framework.

3.3.1 Prediction and Attention Transfer

Weighted categorical cross-entropy is used for supervised classification to account for class imbalance. Predictive knowledge is transferred using temperature-scaled knowledge distillation

with $T = 4$ [3].

For spatial knowledge transfer, Grad-CAM maps are generated from the frozen teacher [4]. The student uses a differentiable CAM-compatible classification head to generate class-specific attention maps during the forward pass, following the principles of CAT-KD and Exp-KD [5, 6].

Teacher and student attention maps are adaptively average-pooled to a common 2×2 resolution and L_2-normalized. Their spatial discrepancy is minimized using the attention-alignment loss:

$$\mathcal{L}_{ATT} = \frac{1}{NCHW} \sum_{n,c,i,j} \left(\widetilde{G}^{S}_{n,c,i,j} - \widetilde{G}^{T}_{n,c,i,j} \right)^2, \tag{1}$$

where $\widetilde{G}^{T}$ and $\widetilde{G}^{S}$ denote the normalized teacher and student attention maps.

3.3.2 Overall EKD Objective

The complete EKD objective is:

$$\mathcal{L}_{EKD} = \mathcal{L}_{WCE} + \alpha \mathcal{L}_{KD} + \beta \mathcal{L}_{ATT}, \tag{2}$$

where $\mathcal{L}_{WCE}$ is the weighted supervised classification loss, $\mathcal{L}_{KD}$ is the temperature-scaled prediction distillation loss, and $\mathcal{L}_{ATT}$ is the attention-alignment loss. The implemented configuration uses $\alpha = 1.0$ and $\beta = 5.0$.

During training, the teacher remains frozen and only the student parameters are updated. The student therefore learns from the ground-truth labels, teacher predictions, and teacher spatial attention.

3.4 Evaluation

Classification performance was evaluated using accuracy, precision, recall, F1-score, AUC, and loss. Each EKD configuration was compared with its corresponding MobileNetV2 baseline under the same training setting. Explainability was evaluated qualitatively by comparing the attention maps produced by the teacher and EKD student.

4 Results

4.1 EKD Classification Performance

The proposed EKD framework was evaluated under No Augmentation, RandAugment, and AugMix. Each EKD-trained MobileNetV2 student was compared with its corresponding baseline MobileNetV2 model under the same training setting. Table 5 summarizes the validation results.

Table 5 Validation performance of MobileNetV2 before and after EKD.

Setting	Model	Accuracy	F1	AUC
No Aug.	Baseline	0.8454	0.8521	0.8621
	EKD	**0.8701**	**0.8710**	0.8615
RandAug.	Baseline	0.8433	0.8519	0.8682
	EKD	**0.8639**	**0.8646**	**0.8685**
AugMix	Baseline	0.8247	0.8371	0.8648
	EKD	**0.8742**	**0.8751**	**0.8664**

EKD improved validation accuracy and F1-score across all three training settings. The largest improvement was obtained with AugMix, where accuracy increased from 0.8247 to 0.8742 and F1-score from 0.8371 to 0.8751. AugMix also achieved the highest validation accuracy and F1-score among the three EKD configurations.

Changes in validation AUC were comparatively small. AUC decreased slightly from 0.8621 to 0.8615 under No Augmentation, while increasing from 0.8682 to 0.8685 with RandAugment and from 0.8648 to 0.8664 with AugMix.

4.2 Training Behavior

Figure 2 compares the training and validation accuracy and total EKD loss across the three training settings. No Augmentation showed a larger train–validation accuracy gap toward later epochs, while RandAugment maintained closer agreement between the two curves. Under AugMix, training accuracy remained lower while validation performance remained strong.

The total EKD loss decreased across all three configurations before approaching a stable region. These curves show differences in optimization behavior across the three training settings and complement the validation results reported above.

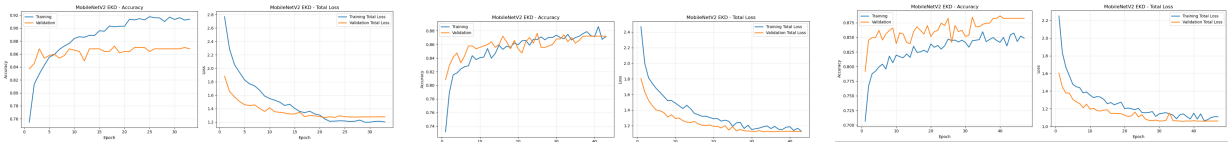


Figure 2 Training and validation accuracy and total EKD loss under No Augmentation, RandAugment, and AugMix, respectively.

4.3 Qualitative Explainability Analysis

Teacher and EKD-student attention maps were qualitatively compared to examine the spatial behavior transferred through the attention-alignment component. Representative Grad-CAM comparisons for the three training settings are shown in Figures 3–5.

Across the examined examples, the EKD student did not reproduce the teacher attention maps exactly. The student maps were generally broader and more diffuse, while partial spatial correspondence with the teacher was observed in several cases. These visualizations provide qualitative evidence of teacher–student attention correspondence rather than exact localization agreement.

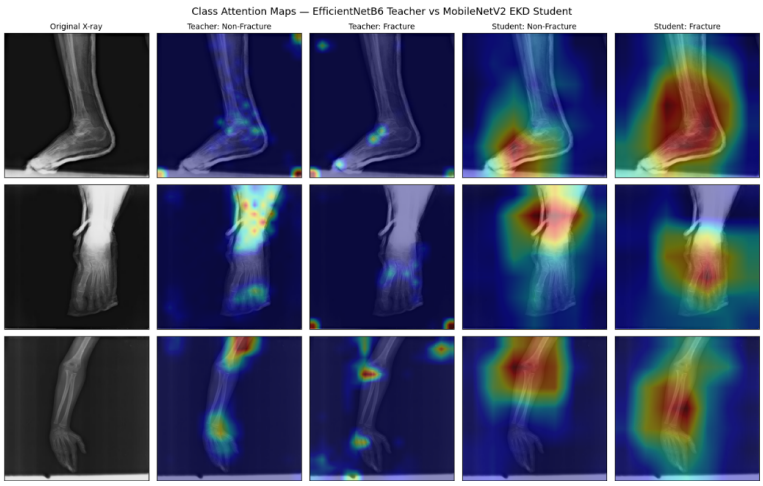


Figure 3 Representative teacher and EKD-student attention maps under No Augmentation.

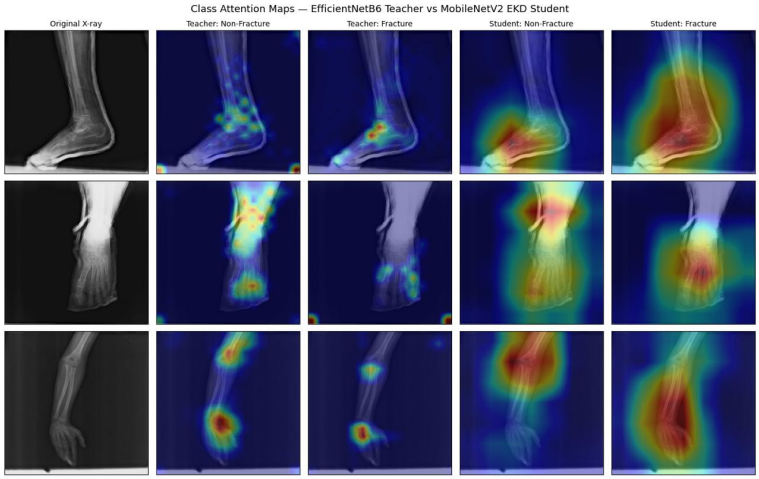


Figure 4 Representative teacher and EKD-student attention maps under RandAugment.

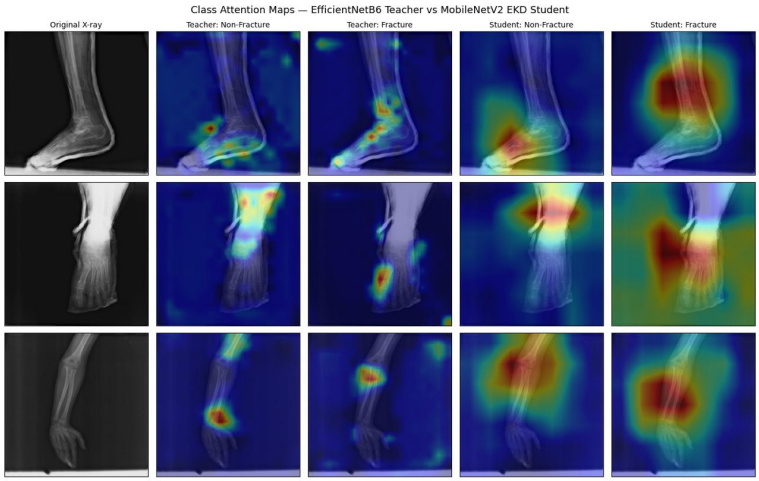


Figure 5 Representative teacher and EKD-student attention maps under AugMix.

5 Discussion

The results show that EKD consistently improved the MobileNetV2 student's validation accuracy and F1-score across all three training settings, with the largest improvement observed under AugMix. Validation AUC remained relatively stable across the three settings, while accuracy and F1-score improved consistently, indicating that the main benefit of EKD was reflected in the student's classification performance. The learning curves further indicate that augmentation influenced the optimization behavior of the distilled student.

The qualitative analysis showed partial spatial correspondence between teacher and student attention, although the student maps were generally broader and more diffuse. This suggests that attention alignment encouraged the student to preserve some of the teacher's class-discriminative spatial patterns without requiring exact map replication.

Overall, the findings support the potential of combining prediction distillation with attention transfer for lightweight fracture classification, while the qualitative nature of the attention analysis motivates further quantitative evaluation.

6 Limitations

The experiments were conducted on FracAtlas using a binary fracture/non-fracture classification setting; therefore, validation on additional datasets and classification tasks is needed to assess broader generalizability. Teacher–student attention correspondence was evaluated qualitatively, and the attention maps were pooled to a 2×2 resolution during training. In addition, EKD was evaluated using one teacher–student pair and fixed loss coefficients ($\alpha = 1.0$, $\beta = 5.0$), without an ablation study. Deployment-oriented metrics such as inference time, model size, memory usage, and FLOPs were outside the scope of the current evaluation.

7 Conclusion

This work investigated an Explainable Knowledge Distillation framework that combines prediction-based knowledge distillation with class-discriminative attention transfer for lightweight bone fracture classification. Using EfficientNet-B6 as the teacher and MobileNetV2 as the student, EKD improved validation accuracy and F1-score across No Augmentation, RandAugment, and AugMix, with AugMix achieving the strongest overall validation performance. Qualitative attention-map comparisons also showed partial teacher–student spatial correspondence. Overall, the results support explainability-guided distillation as a promising approach for improving lightweight fracture classification while encouraging teacher–student attention similarity.

7.1 Future Work

Future work should quantitatively evaluate teacher–student attention alignment, validate the framework on external datasets, and extend it to multi-class fracture classification. Future experiments should also evaluate additional teacher–student architectures and conduct ablation studies on the loss coefficients, while deployment-oriented evaluation should measure model size, inference time, memory usage, and computational cost.

Acknowledgments

We would like to express our sincere gratitude to KAUST Academy for providing this valuable opportunity and for its support throughout the project. We are especially grateful to Prof. Naeemullah Khan, Prof. Sultan Albarakati, and our project monitor, Dr. Salman Khan, for their guidance and support. We also extend our appreciation to the TAs for their continuous assistance and support throughout the project.

References

[1] Çağatay Berke Erdaş. "Automated fracture detection in the ulna and radius using deep learning on upper extremity radiographs". In: *Joint Diseases and Related Surgery* 34.3 (2023), pp. 598–604. DOI: 10.52312/jdrs.2023.1312.

[2] Shyam Gupta and Dhanisha Sharma. *Bone Fracture Classification using Transfer Learning*. 2024. DOI: 10.48550/arXiv.2406.15958. arXiv: 2406.15958 [eess.IV].

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network". In: *arXiv preprint arXiv:1503.02531* (2015).

[4] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[5] Ziyao Guo et al. "Class Attention Transfer Based Knowledge Distillation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 11868–11877. DOI: 10.1109/cvpr52729.2023.01142. URL: http://dx.doi.org/10.1109/CVPR52729.2023.01142.

[6] Tianli Sun et al. "Explainability-based knowledge distillation". In: *Pattern Recognition* 159 (Mar. 2025), p. 111095. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2024.111095. URL: http://dx.doi.org/10.1016/j.patcog.2024.111095.

[7] Sayeda Sanzida Ferdous Ruhi, Fokrun Nahar, and Adnan Ferdous Ashrafi. "A Novel Approach Towards the Classification of Bone Fracture from Musculoskeletal Radiography Images Using Attention Based Transfer Learning". In: *2024 27th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2024, pp. 517–522. DOI: 10.1109/ICCIT64611.2024.11022528.

[8] Ali Tamizifar, Shakiba Berenjkoub, and Mina Amiri. "Deep Learning Meets Explainable AI: A Robust Framework for X-Ray Fracture Detection". In: *2025 33rd International Conference on Electrical Engineering (ICEE)*. IEEE, May 2025, pp. 1–5. DOI: 10.1109/icee67339.2025.11213731. URL: http://dx.doi.org/10.1109/ICEE67339.2025.11213731.

[9] Enrique Queipo-de-Llano et al. "eXplainable Artificial Intelligence for Hip Fracture Recognition". In: *Applied Artificial Intelligence* 39.1 (May 2025). ISSN: 1087-6545. DOI: 10.1080/08839514.2025.2502568. URL: http://dx.doi.org/10.1080/08839514.2025.2502568.

[10] Thamizhazhagan Selvam and Hariprasad S. "XAI-based Advanced Computational Models for Predictive Analysis of Bone Fracture and Structural Repair Healing Prediction". In: *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE, 2025, pp. 388–393. DOI: 10.1109/icicv64824.2025.11085540. URL: http://dx.doi.org/10.1109/ICICV64824.2025.11085540.

[11] Saif Ur Rehman Khan et al. "KDLight: A Lightweight Knowledge Distillation Framework for Medical Image Classification". In: *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8. DOI: 10.1109/ijcnn64981.2025.11228615. URL: http://dx.doi.org/10.1109/IJCNN64981.2025.11228615.

[12] Jnanapriya B. et al. “Mixture of Experts-Based Knowledge Distillation for Robust Fracture Detection in X-Ray Images”. In: *2026 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*. IEEE, Mar. 2026, pp. 155–159. DOI: 10.1109/icmlas67792.2026.11483891. URL: http://dx.doi.org/10.1109/ICMLAS67792.2026.11483891.

[13] Sneha Sharma et al. “Task Specialized Bone Fracture Detection with Attention Models and Knowledge distillation”. In: *2026 7th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE, May 2026, pp. 1860–1866. DOI: 10.1109/icicv68925.2026.11554609. URL: http://dx.doi.org/10.1109/ICICV68925.2026.11554609.

[14] Aqib Nazir Mir and Danish Raza Rizvi. *Explainable Knowledge Distillation for Efficient Medical Image Classification*. 2025. DOI: 10.48550/ARXIV.2508.15251. URL: https://arxiv.org/abs/2508.15251.

[15] I. Abedeen et al. “FracAtlas: A Dataset for Fracture Classification, Localization and Segmentation of Musculoskeletal Radiographs”. In: *Scientific Data* 10 (2023), p. 521. DOI: 10.1038/s41597-023-02432-4.

[16] Ekin D. Cubuk et al. “RandAugment: Practical Automated Data Augmentation with a Reduced Search Space”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 3008–3017. DOI: 10.1109/CVPRW50498.2020.00359.

[17] Dan Hendrycks et al. “AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty”. In: *International Conference on Learning Representations (ICLR)*. 2020.

[18] Kaiming He et al. “Identity Mappings in Deep Residual Networks”. In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38. URL: https://doi.org/10.1007/978-3-319-46493-0_38.

[19] Gao Huang et al. “Densely Connected Convolutional Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–4708. DOI: 10.1109/CVPR.2017.243.

[20] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[21] Andrew G. Howard et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv: 1704.04861 [cs.CV]. URL: https://arxiv.org/abs/1704.04861.

[22] Mark Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.