Residual Network with Triple-Attention Mechanisms for Knee Osteoarthritis Severity Classification

Hongyu Tian[1], Wenming Cao[1], Qiyu Ran[2,*]

[1]Systems science, School of Mathematics and Statistics, Chongqing Jiaotong University, China
[2]Fuling Center for Disease Control and Prevention, Chongqing, China

Abstract: As the quality of life continues to improve, people are more concerned about all types of diseases. Knee osteoarthritis (KOA) is a type of arthritis that is characterized by limited movement, joint stiffness and pain. This degenerative disease leads to gradual wear and tear of the knee joint and in severe cases, disability. Conventional radiographic diagnosis remains challenging due to the subtle morphological changes in early-stage KOA that often resemble age-related physiological variations. Meanwhile, applying convolutional neural networks to the prediction of KOA has become an effective method. To address this diagnostic bottleneck, we introduce Triplet Attention (TA) in Residual Network (ResNet) for KOA recognition. The architecture innovatively integrates cross-dimensional attention modules within residual blocks, enabling simultaneous modeling of channel-wise dependencies, spatial correlations, and hierarchical feature relationships. Our experimental framework was rigorously evaluated on two medical imaging datasets, comprising 8,263 standardized knee radiographs with Kellgren-Lawrence grading annotations. The proposed architecture demonstrated statistically improvements over other existing convolutional networks (VGG-19, DenseNet-121, GoogleNet, etc.).

1 Introduction

Knee osteoarthritis (KOA) is a chronic degenerative disease caused by the deterioration of articular cartilage, primarily resulting from wear and tear of the knee joint cartilage that leads to direct bone contact and friction [1]. It is characterized by progressive cartilage degeneration, inflammation, narrowing of joint space, and osteophyte formation, with pain, stiffness, and functional impairment as hallmark symptoms. The pathogenesis is associated with multiple risk factors including genetic predisposition, obesity, aging, trauma, and occupational strain. Given global trends of population aging and rising obesity rates, KOA prevalence is projected to increase significantly, posing substantial threats to quality of life and functional capacity. As an irreversible and progressive condition with heterogeneous early manifestations and variable progression rates among individuals, KOA management remains clinically challenging [2]. Without timely intervention, it can lead to disability, highlighting the importance of early diagnosis and treatment to improve patient outcomes and slow disease progression.

Current diagnostic approaches for KOA predominantly rely on conventional imaging modalities such as X-ray, CT, and MRI, with X-ray being widely adopted due to cost-effectiveness [3]. Kellburg-Lawrence (KL) grading system is the standard scheme used to assess the severity of KOA. n 1961, the World Health Organization (WHO) adopted the Kellgren-Lawrence (KL) grading system for knee osteoarthritis severity classification, categorizing the condition into five distinct grades ranging from grade 0 (G0) to grade 4 (G4). Representative examples of KOA severity grades are illustrated in Figure 1.

However, these techniques exhibit inherent limitations, particularly in their dependence on clinician expertise, which introduces subjectivity, time-intensive analysis, and diagnostic inconsistency, especially in early-stage detection where subtle pathological changes are prone to misdiagnosis. Recent advancements in deep learning, particularly deep convolutional neural networks (CNNs), have demonstrated potential in KOA identification by automatically detecting nuanced structural alterations in X-ray images, thereby enhancing diagnostic objectivity and accuracy for early intervention [4]. Nevertheless, despite promising capabilities in early disease recognition and risk stratification, the generalizability of deep learning models and data diversity constraints may impede their performance across heterogeneous clinical environments.

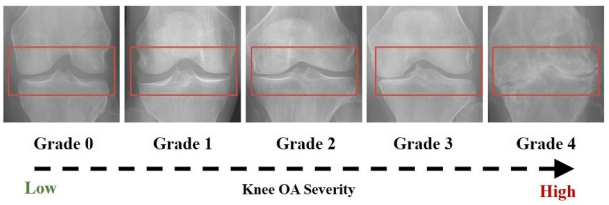

Figure 1. Progression of KOA disease: sample radiographs and corresponding KL levels.

* Corresponding author: Rqy2077@outlook.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

To address these challenges, this study proposes a novel ResNet architecture incorporating triplet attention (TA) mechanisms within residual modules. The triplet attention framework employs a three-branch structure that establishes interdependencies across multiple image dimensions through rotation operations and residual transformations. Compared to conventional convolutional architectures, this network enhances the learning of semantically meaningful image representations, thereby improving diagnostic accuracy for KOA detection in X-ray radiography.

2 Related work

2.1 CNNs applied in KOA detection

The advent of convolutional neural networks (CNNs) has driven remarkable progress in visual recognition systems, fundamentally transforming methodological frameworks for image classification tasks in computer vision. With the continuous refinement of network architectures and the development of novel training paradigms, substantial advancements in classification accuracy have been consistently demonstrated across diverse multimodal imaging datasets. This technological evolution has catalyzed significant breakthroughs in medical imaging analytics, particularly in developing computer-aided diagnostic solutions for early-stage KOA detection.

Yeoh et al. [5] developed a 3D convolutional neural network architecture for KOA detection. Their study demonstrated significant diagnostic efficacy of this 3D deep learning model in the early identification of degenerative knee joint lesions. Sarvamangala et al. [6] proposed an innovative Multi-scale Feature Fusion Network (MCBCNN), which integrated three pre-trained models (MobileNetV2, ResNet-50, and InceptionV3) for joint feature extraction. The enhanced ResNet-50 architecture achieved over 95% accuracy and an F1-score of 0.8 in the KOA classification task. Sivakumari et al. [7] developed an automatic classification system for joint space narrowing (JSN) using the AlexNet architecture. This system employed convolutional neural networks to encode features from X-ray images, implementing a five-level classification system ranging from grade 0 (no JSN) to grade 4 (significant osteophyte formation). Abdullah et al. [8] developed a multimodal diagnostic system based on transfer learning, which combined pre-trained models including ResNet-50 and AlexNet through feature transfer learning to achieve automatic quantitative assessment of joint space narrowing in knee X-ray images. Harish et al. [9] conducted a comparative analysis of traditional CNN and VGG-16 architectures for osteoarthritis diagnosis. Experimental results indicated that the VGG-16 transfer learning model achieved 93% accuracy, significantly outperforming the baseline CNN model's 67% classification efficacy. Alshamrani et al. [10] performed a systematic evaluation of three transfer learning models (sequential CNN, VGG-16, and ResNet-50) for knee X-ray image analysis, ultimately developing an intelligent osteoarthritis recognition framework through feature fusion. Sarhan et al. [11] developed a multimodal fusion framework utilizing transfer learning from convolutional neural networks. This framework achieved high diagnostic accuracy in X-ray image analysis through the implementation of data augmentation strategies and feature visualization techniques, offering visual decision support for clinical interpretation of KOA pathologies. Muzaffar et al. [12] proposed an innovative methodology integrating local phase quantization (LPQ) to extract microtexture features from articular cartilage. By embedding a convolutional block attention module within the VGG network architecture, their experimental results demonstrated that the fusion of differential image features with attention weighting mechanisms enhances the characterization of KOA radiographic images. Kadu et al. [13] introduced a bilinear interactive convolutional neural network to model spatial correlations and channel interdependencies among multiscale features in knee joint images, leveraging quantitative analysis of joint space narrowing for improved detection accuracy.

2.2 Attention applied in CNNs

The integration of attention mechanisms in deep CNNs has become an important strategy to improve the performance of large-scale visual tasks. This mechanism simulates the prioritized focus of information in human cognitive systems, strengthening key features and suppressing redundant information by dynamically allocating computing resources. As a pioneering work on channel attention mechanisms, the Squeeze-and-Excitation Network (SENet) [14] achieves adaptive re-calibration of feature channels by constructing nonlinear interactions between channels. The core of this method is to introduce learnable channel modulation weights and build a channel dependency model through global average pooling (GAP), which achieves significant performance improvements while maintaining low computational overhead. Subsequent studies further explore the potential of this framework through compound attention mechanisms. For example, the Convolutional Block Attention Module (CBAM) [15] integrates spatial attention components with channel attention, establishing a dual spatio-channel attention mechanism by leveraging both maximum pooling and average pooling for feature extraction. Experiments show that this synergistic effect yields performance gains surpassing those of single-channel attention. These methods collectively demonstrate that attention mechanisms can effectively enhance feature representation capabilities of deep networks through feature selection and reinforcement.

The Dual Attention Network (A^2-Nets) [16] introduces an enhanced relational operator that reformulates the computational framework of the classical Non-Local (NL) module. The original NL module [17], designed to establish feature correlations through a global interaction mechanism, features a modular architecture characterized by high architectural compatibility and low computational complexity. For feature aggregation, the Global Second-order Pooling Network (GSoP-Net) [18] em-

ploys second-order statistical modeling, leveraging covariance matrix operations to achieve deep feature integration. Its innovation lies in constructing cross-spatial higher-order feature correlations and enhancing network representational capacity via a parametric channel recalibration strategy. Xiao et al. [19] developed a multi-level attention framework incorporating reverse attention pathways, improving fine-grained classification performance through hierarchical feature refinement.

Among the aforementioned methods, some leverage the capability of CNNs to learn image features, while others incorporate channel and spatial attention mechanisms. However, none of these approaches consider cross-dimensional interactions within images, and they have largely overlooked the dependencies between different dimensions. In contrast, our method introduces a triplet attention module [20] into residual blocks of ResNet, aiming to capture cross-dimensional interactions and better learn subtle changes between images of varying KOA severity levels.

3 Methods

3.1 ResNet

After the emergence of AlexNet [21], the first CNN-based architecture, the deep learning community widely adopted the strategy of increasing network depth to enhance model performance. However, when network depth increases significantly, abnormal phenomena emerge in gradient propagation across layers: gradient magnitudes tend toward zero (vanishing gradients) or grow exponentially (exploding gradients). Both scenarios destabilize parameter updates during model training, which not only contradicts the theoretical expectation that increased model capacity should enhance representational power but may also exacerbate overfitting due to higher parameter space complexity.

To address these abnormal gradient propagation issues, Residual Networks (ResNet) [22] innovatively introduced residual connections. The core concept of this architecture lies in directly transmitting shallow-layer feature information to deeper modules through skip connections. Specifically, given a module input x and its nonlinearly transformed output $H(x)$, traditional CNN architectures directly propagate this transformed result to subsequent layers. In contrast, residual modules reformulate the learning objective as the difference between the transformed output and original input, constructing a residual function $F(x)=H(x)-x$. This design enables gradient signals to be directly backpropagated through skip connection paths, effectively mitigating gradient attenuation in deep networks. From an architectural implementation perspective, ResNet employs residual modules as fundamental building blocks, each containing two to three convolutional layers. By progressively increasing network depth, ResNet learns more complex and abstract feature representations, which prove critical for computer vision tasks.

In ResNet variants such as ResNet-18 and ResNet-34, which feature relatively shallow depths, each residual block consists of two convolutional layers with 3×3 kernels, forming a stack of convolutional layers followed by a shortcut connection. As network depth increases, to reduce computational costs and enhance model efficiency, deeper architectures like ResNet-50 and ResNet-101 replace these blocks with bottleneck blocks comprising three convolutional layers. These layers sequentially employ 1×1, 3×3, and 1×1 convolutions, where the 1×1 layers first reduce computational complexity by lowering dimensionality and then restore it, while the 3×3 layer acts as a bottleneck operating on reduced input/output dimensions. Figure 2 illustrates the distinct building blocks of residual learning: (a) the schematic diagram of residual learning; (b) the residual block in shallower networks; (c) the bottleneck block in deeper networks.

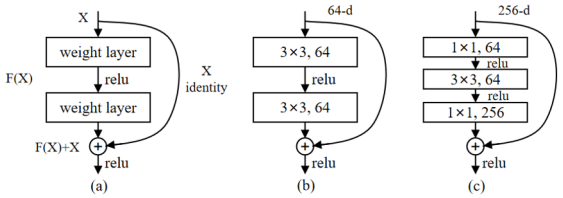

Figure 2. Residual learning: different building blocks.

3.2. Triplet Attention

In traditional attention computation methods, channel attention mechanisms typically employ global average pooling to compress spatial dimensions, reducing each channel to a single pixel. This approach compromises spatial detail features significantly. While existing improved methods introduce independent spatial attention modules as supplements, they fail to effectively establish interdependencies between channel and spatial dimensions.

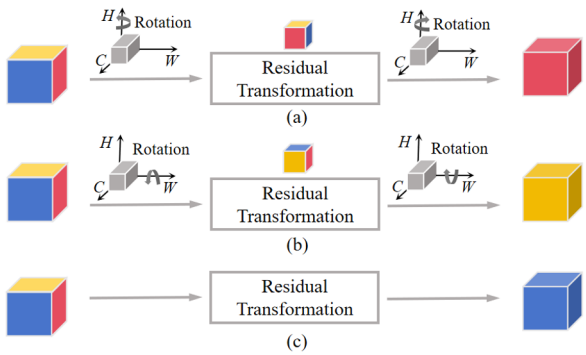

Figure 3. An illustration of triplet attention mechanism.

The triplet attention mechanism proposed in this study constructs a cross-dimensional dependency model by establishing synergistic relationships between channel dimension C and spatial dimensions H or W to achieve feature refinement. This attention mechanism comprises three parallel branches: the first two branches respectively build interactive mappings between channel dimension and vertical/horizontal spatial dimensions, while the last branch retains a spatial attention structure similar to CBAM. The final optimized feature representation is

generated through weighted average integration of tensors from all branches, maintaining identical dimensions with the input tensor. Figure 3 presents an abstract representation of the triplet attention mechanism, where three branches capture cross-dimension interactions. (a) corresponds to the first branch; (b) illustrates the second branch; (c) depicts the third branch.

The specific implementation workflow is described as follows. Given an input tensor $\chi \in \mathbb{R}^{C\times H\times W}$, we first pass it to each of the three branches in the proposed triplet attention module. In the first branch, we capture dependencies between the (C,H) dimensions of the tensor. Rotate the input χ counterclockwise by 90° along the H axis. The rotated tensor is denoted as $\hat{\chi}_1$, with a shape of $(W\times H\times C)$. Subsequently, pass it through a Z-Pool layer, which reduces the 0th-dimension of the tensor to 2 by concatenating average-pooled features and max-pooled features along that dimension. The mathematical expression is as follows:

$$Z-\mathrm{pool}(\chi)=\left[MaxPool_{0d}(\chi),AvgPool_{0d}(\chi)\right] \quad (1)$$

where $0d$ is 0th-dimension across which the max and average pooling operations take place.

This layer simplifies $\hat{\chi}_1$ into $\hat{\chi}_1^*$ which is a shape of $(2\times H\times C)$. Subsequently, $\hat{\chi}_1^*$ is passed through a standard convolution layer with a kernel size of $k\times k$, followed by a batch normalization layer, which yields an intermediate output of dimensions $(1\times H\times C)$. The resulting attention weights are then generated by passing the tensor through a sigmoid activation layer (σ). These attention weights generated are applied to $\hat{\chi}_1$, which is then rotated clockwise by 90° along the H axis to restore the original input shape of χ. The process by which the rotated $\hat{\chi}_1$ goes through a series of transformations to get a new $\hat{\chi}_1$ is called the residual transformation

In the second branch, we rotate the tensor χ counterclockwise by 90° along the W axis to obtain $\hat{\chi}_2$, capturing dependencies between the (C,W) dimensions of the tensor. Subsequent operations are analogous to those in the first branch. In the third branch, the input tensor χ is not rotated, preserving its original orientation to capture dependencies between the (H,W) dimensions of the tensor, with follow-up operations similarly mirroring the first branch.

The refined tensors with shape $(C\times H\times W)$ generated by each of three branches are aggregated via simple averaging. In summary, the overall process of the triplet attention mechanism, from an input tensor to an output tensor, can be formulated by the following equation:

$$y=\frac{1}{3}\left(\hat{\chi}_1\sigma(\varphi_1(\hat{\chi}_1^*))+\hat{\chi}_2\sigma(\varphi_2(\hat{\chi}_2^*))+\chi\sigma(\varphi_3(\hat{\chi}_3^*))\right) \quad (2)$$

where σ denotes sigmoid activation function. φ_1, φ_2 and φ_3 denote standard two-dimensional convolutional layers defined by kernel size k in three branches of triplet attention. Simplifying Eq. (2), y becomes:

$$y=\frac{1}{3}\left(\hat{\chi}_1\omega_1+\hat{\chi}_2\omega_2+\chi\omega_3\right)=\frac{1}{3}\left(\overline{y}_1+\overline{y}_2+y_3\right) \quad (3)$$

where ω_1, ω_2 and ω_3 are the three cross-dimensional attention weights computed in triplet attention. The $\overline{y}_1$ and $\overline{y}_2$ in Eq. (3) represents the 90° clockwise rotation to retain the original input shape of $(C\times H\times W)$.

3.3 Network architecture

The proposed network architecture employs ResNet101 as its backbone, which comprises an initial 7×7 convolutional layer, a max-pooling layer, four bottleneck blocks (with quantities of 3, 4, 23, and 3 respectively), a global average pooling layer, and a fully connected layer. The total depth of 101 layers is derived by summing the initial 7×7 convolutional layer, 99 layers from the bottleneck blocks (each containing three convolutional layers), and the final fully connected layer, the specific network architecture is shown in Figure 4.

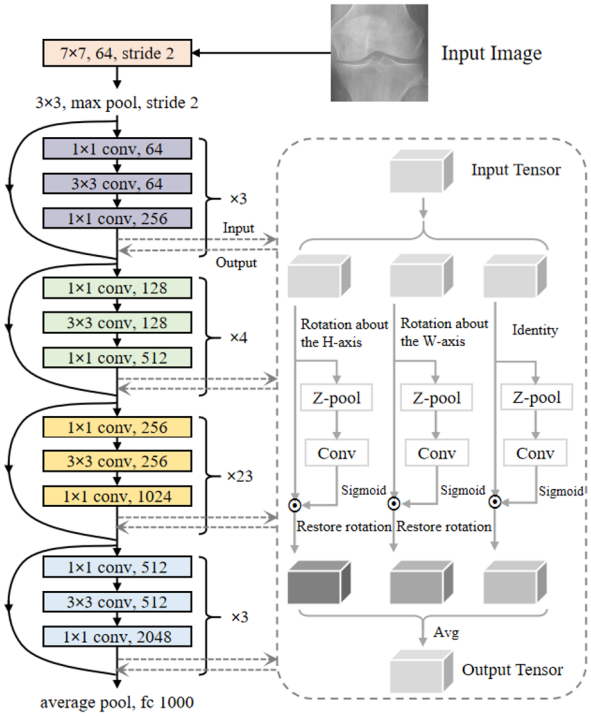

Figure 4. The architecture of proposed ResNet with triplet attention.

In the modified design, triplet attention mechanisms are integrated into the end of each bottleneck block — specifically inserted after the three convolutional layers but before the residual connection. Leveraging the triplet attention's three-branch structure, each bottleneck block enhances its capability to model interdependencies across multiple dimensions of input tensors, thereby capturing finer-grained features to improve classification accuracy for early-stage KOA symptoms.

4 Experiments and results

4.1 Datasets

We conducted image classification tasks using the proposed framework on two knee-related datasets: the Knee Osteoarthritis Dataset with Severity Grading and the Knee X-ray Osteoporosis Database, respectively.

Knee Osteoarthritis Dataset with Severity Grading: It contains five categories of knee osteoarthritis (KOA) X-ray data for KOA detection and Kellgren-Lawrence (KL) grading, with uniform image dimensions of 224×224 pixels. The dataset is partitioned into training (5,778 images), validation (826 images), and test sets (1,659 images) in a ratio of 7:1:2. The grading descriptions are as follows: Grade 0 (Healthy): Normal knee joint appearance; Grade 1 (Doubtful): Suspected joint space narrowing with possible osteophytes; Grade 2 (Mild): Definite osteophytes and possible joint space narrowing; Grade 3 (Moderate): Multiple osteophytes, marked joint space narrowing, and mild sclerosis; Grade 4 (Severe): Large osteophytes, severe joint space narrowing, and significant sclerosis. The dataset is publicly available at https://www.kaggle.com.

Knee X-ray Osteoporosis Database: The dataset is publicly available at https://data.mendeley.com. It comprises three categories of knee osteoporosis X-ray data: Normal, Osteopenia, and Osteoporosis, with a total of 239 images of non-uniform dimensions.

4.2 Implementation details

Prior to the experiments, both datasets were preprocessed. The first dataset underwent data augmentation, including brightness adjustment, random rotation and translation, and cropping. The second dataset was partitioned into training (2,004 images), validation (276 images), and test sets (590 images) in a ratio of 7:1:2. Preprocessing steps for the second dataset included proportionally cropping images into varying sizes, equalization, addition of Gaussian noise to both original and cropped images, and random rotation to augment the dataset size.

All experiments were conducted in an environment with PyTorch 1.11.0, Python 3.8, and CUDA 11.3, utilizing a 15-core Intel(R) Xeon(R) Platinum 8474C CPU and an RTX 4090D GPU. The Adam optimizer was employed for model training across all the experiments, with an initial learning rate of 1e-4 and a batch size of 32. Performance evaluation and comparison between our model and classical CNN models were performed using accuracy, precision, recall, F1-score, and AUC.

4.3 Results

Under the aforementioned data preprocessing and experimental setup, we conducted classification tasks on two distinct datasets: the Knee Osteoarthritis Dataset with Severity Grading (hereafter referred to as Dataset 1) and the Knee X-ray Osteoporosis Database (hereafter referred to as Dataset 2). The experimental framework encompassed multiple established architectures including AlexNet [21], VGG-19 [23], DenseNet-121 [24], GoogLeNet [25], Vision Transformer Large/16 (ViT-L/16) [26], along with the backbone network ResNet employed in this study. Additionally, comprehensive evaluations were performed using the proposed model developed in this work.

Table 1 presents the detailed predictive performance metrics of all the models on Dataset 1, including accuracy, AUC, and macro/micro-averaged precision, recall, and F1-scores. The results demonstrate that among classical image classification models, VGG exhibits superior performance in macro-precision.

Table 1. Comparison of evaluation results between the proposed model and the classical model on dataset 1.

Methods	ACC	Precision		Recall		F1 score		AUC
		micro	macro	micro	macro	micro	macro	
AlexNet	0.618	0.618	0.606	0.618	0.616	0.618	0.595	0.859
VGG19	0.626	0.626	**0.703**	0.626	0.627	0.626	0.588	0.858
DenseNet-121	0.650	0.650	0.653	0.650	0.644	0.650	**0.639**	0.875
GoogleNet	0.650	0.650	0.621	0.650	0.645	0.650	0.612	0.868
VIT-L/16	0.548	0.548	0.505	0.548	0.393	0.548	0.369	0.807
ResNet-101	0.650	0.650	0.643	0.650	0.625	0.650	0.618	**0.879**
ResNet-101+TA	**0.655**	**0.655**	0.653	**0.655**	**0.646**	**0.655**	0.637	**0.879**

Table 2. Comparison of evaluation results between the proposed model and the classical model on dataset 2.

Methods	ACC	Precision		Recall		F1 score		AUC
		micro	macro	micro	macro	micro	macro	
AlexNet	0.653	0.653	0.218	0.653	0.333	0.653	0.263	0.464
VGG19	0.651	0.651	0.217	0.651	0.333	0.651	0.263	0.558
DenseNet-121	0.617	0.617	0.394	0.617	0.354	0.617	0.339	0.638
GoogleNet	0.651	0.651	0.217	0.651	0.333	0.651	0.263	0.482
VIT-L/16	0.639	0.639	0.377	0.639	0.344	0.639	0.297	0.555
ResNet-101	0.644	0.644	0.328	0.644	0.335	0.644	0.276	**0.657**
ResNet-101+TA	**0.663**	**0.663**	**0.839**	**0.663**	**0.359**	**0.663**	**0.315**	**0.657**

Meanwhile, DenseNet, GoogLeNet, and ResNet generally outperform other models across most metrics, with DenseNet achieving the balanced performance among all baselines (including the backbone architectures utilized in this study). For the novel model architecture proposed in this study, the structural modifications have resulted in optimal comprehensive performance on Dataset1. Through architectural refinement, the originally slightly inferior ResNet (compared to DenseNet across multiple metrics) has been elevated to achieve state-of-the-art performance. This advancement not only demonstrates our framework's enhanced capability in learning discriminative features from radiographic images, but also substantiates the effectiveness of capturing cross-dimensional interactions for differentiating varying severity levels of KOA (Knee Osteoarthritis) symptoms.

Similarly, Table 2 summarizes the predictive performance metrics of all models on Dataset 2. Compared to Dataset 1, none of the existing models demonstrate a balanced capability to achieve strong performance across all metrics. In contrast, our proposed model fulfills this requirement by attaining optimal performance across all evaluation metrics. The experimental outcomes from both datasets validate that the architecture proposed in this study enhances early prediction of KOA. Particularly when datasets exhibit ambiguous performance across different baseline classification models, our framework provides a robust solution to assist clinicians in making more accurate diagnostic assessments.

5 Conclusion

In this paper, we propose a novel framework for detecting knee osteoarthritis (KOA) in radiographs. Traditional computer vision feature extraction methods typically rely on aggregating visual features, which fail to capture fine-grained details. While convolutional neural network (CNN) models can extract dominant image features, their capability to focus on and capture subtle variations remains limited, particularly given the high similarity of early-stage KOA radiographic manifestations. Our proposed method integrates rotational triple attention mechanisms within the ResNet backbone network, enabling the capture of interactive information across multiple dimensions of input image tensors. By incorporating attention layers at the terminal ends of bottleneck blocks, each feature extraction module achieves earlier refinement and feature enhancement. Comparative classification experiments with conventional model architectures were conducted on two relevant datasets. Experimental results demonstrate that the proposed architecture effectively improves the network's ability to discriminate between different KOA stages, while providing more decisive predictions in cases where conventional architectures yield ambiguous classification outcomes.

References

1. D. Saini, T. Chand, D.K. Chouhan, et al., Biocybern. Biomed. Eng. **41**, 419 (2021).
2. A. Brahim, R. Jennane, R. Riad, et al., Comput. Med. Imaging Graph. **73**, 11–18 (2019)
3. I. Malik, M. Yasmin, A. Iqbal, et al., Multimed. Tools Appl. **83**, 86923–86954 (2024)
4. K.A. Thomas, Ł. Kidziński, E. Halilaj, et al., Radiol. Artif. Intell. **2**, e190065 (2020)
5. P.S.Q. Yeoh, K.W. Lai, S.L. Goh, et al., Comput. Intell. Neurosci. **2021**, 4931437 (2021)
6. D.R. Sarvamangala, R.V. Kulkarni, Neural Process. Lett. **53**, 2985–3009 (2021)
7. T. Sivakumari, R. Vani, Proc. Int. Conf. Commun. Electron. Syst. (ICCES), 1405–1409 (2022)
8. S.S. Abdullah, M.P. Rajasekaran, Radiol. Med. **127**, 398–406 (2022)
9. H. Harish, A. Patrot, S. Bhavan, et al., Proc. Int. Conf. Recent Adv. Inf. Technol. Sustain. Dev. (ICRAIS), 14–19 (2023)
10. H.A. Alshamrani, M. Rashid, S.S. Alshamrani, et al., Healthcare **11**, 1206 (2023)
11. A.M. Sarhan, M. Gobara, S. Yasser, et al., Int. J. Comput. Intell. Syst. **17**, 241 (2024)
12. A.W. Muzaffar, F. Riaz, M. Tahir, IEEE Access (to be published) (2025)
13. R. Kadu, S. Pawar, J. Electron. Electromed. Eng. Med. Inform. **7**, 80–90 (2025)
14. J. Hu, L. Shen, G. Sun, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 7132–7141 (2018)
15. S. Woo, J. Park, J.Y. Lee, et al., Proc. Eur. Conf. Comput. Vis. (ECCV), 3–19 (2018)
16. Y. Chen, Y. Kalantidis, J. Li, et al., Adv. Neural Inf. Process. Syst. **31** (2018)
17. X. Wang, R. Girshick, A. Gupta, et al., Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 7794–7803 (2018)
18. Z. Gao, J. Xie, Q. Wang, et al., Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 3024–3033 (2019)
19. T. Xiao, Y. Xu, K. Yang, et al., Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 842–850 (2015)
20. D. Misra, T. Nalamada, A.U. Arasanipalai, et al., Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 3139–3148 (2021)
21. A. Krizhevsky, I. Sutskever, G.E. Hinton, Adv. Neural Inf. Process. Syst. **25** (2012)
22. K. He, X. Zhang, S. Ren, et al., Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 770–778 (2016)
23. K. Simonyan, A. Zisserman, arXiv preprint arXiv:1409.1556 (2014)
24. G. Huang, Z. Liu, L. Van Der Maaten, et al., Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 4700–4708 (2017)
25. C. Szegedy, W. Liu, Y. Jia, et al., Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1–9 (2015)
26. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., arXiv preprint arXiv:2010.11929 (2020)