



OPEN 3D CSFA-UNet: a unified attention-driven deep learning framework for accurate knee MRI segmentation and osteoarthritis severity classification

C. Moorthy✉, A. Shafeek, V. Gurunathan & R. Sudhakar

Although several recent multi-task deep learning methods already perform segmentation and classification jointly, many still face limitations in clinical applicability such as restricted multi-scale context modeling, insufficient attention to clinically relevant spatial–channel cues, or heavy computational cost that hinders deployment. Building on these advances, we propose a unified, multi-stage framework for joint segmentation and classification of 3D knee MRI volumes that targets improved diagnostic precision, interpretability, and efficiency. The pipeline begins with Gaussian Guided Filtering to enhance anatomical boundaries while suppressing noise. A novel 3D CSFA-UNet (Channel-Spatial Feature Attention) performs segmentation with embedded multi-scale context via Atrous Spatial Pyramid Pooling (ASPP). To reduce redundancy and isolate discriminative signals, we introduce the Desert Scorpion Feature Selector (DSFS), a metaheuristic feature-selection module. Selected features are classified by a Spiking Transformer network that uses Leaky Integrate-and-Fire (LIF) neurons and graph-attention layers to capture temporal sensitivity and inter-structure context. Falcon Hunting Optimisation (FHO) is used to tune hyperparameters for robust performance. Evaluated on the publicly available OAI dataset, the proposed model achieved a Dice Similarity Coefficient (DSC) of 98.10%, Intersection over Union (IoU) of 96.26%, Average Surface Distance (ASD) of 0.45 mm, and 95th percentile Hausdorff Distance (Hd95) of 1.85 mm for segmentation. For classification, the model attained an accuracy of 99.15%, precision of 98.82%, recall of 99.11%, and an F1-score of 99.04%, demonstrating its robustness and reliability across both segmentation and grading tasks. We also expanded the Introduction with a detailed literature analysis of representative multi-task approaches and clearly position our contributions relative to prior work. This framework therefore advances clinically relevant, interpretable joint segmentation–classification for image-guided orthopaedic diagnostics.

Keywords 3D knee, Total knee replacement, Channel feature attention, Spatial feature attention, Desert scorpion feature selector, Falcon hunting optimization

Knee osteoarthritis (KOA) is a globally prevalent degenerative joint disorder that progressively deteriorates the articular cartilage and associated tissues, ultimately leading to functional disability and a diminished quality of life. The conventional diagnostic approach relies on the Kellgren–Lawrence (K-L) grading system based on X-ray imaging, which, despite its widespread clinical adoption, remains inherently subjective and often lacks the sensitivity to detect subtle anatomical changes during the early phases of the disease. This limitation highlights the growing demand for objective, reproducible, and sensitive diagnostic modalities that can facilitate early intervention and accurate assessment of disease progression. Deep learning (DL) techniques, particularly those applied to medical imaging, have emerged as promising solutions to this diagnostic gap in musculoskeletal applications. Magnetic Resonance Imaging (MRI), with its superior soft tissue contrast, has become instrumental in characterising *in vivo* joint anatomy and pathological alterations. MRI-derived 3D reconstructions, for instance, offer the potential to assess structural features such as cartilage strain—an important biomarker of

Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India. ✉email: moorthyc@drmcet.ac.in

joint degeneration—by measuring variations in tissue thickness under load. However, the creation of such 3D models is typically reliant on labour-intensive manual segmentation, making large-scale and consistent analysis challenging¹. To address this, convolutional neural networks (CNNs), particularly U-Net architectures and their variants, have been widely employed for automating segmentation tasks in knee MRI data. Despite these advances, accurate delineation of intricate knee structures such as cartilage, meniscus, and ligaments remains a complex undertaking. Contributing factors include low inter-tissue contrast, imaging artefacts, and the intricate morphology of smaller anatomical features². To enhance segmentation fidelity, recent methods have integrated anatomical priors and shape modelling. One such approach incorporated a prior-based 3D U-Net augmented with an Average Shape Model (ASM) and subpixel guidance, which improved segmentation performance for femoral and tibial cartilage. The anterior cruciate ligament (ACL), a structure commonly affected in athletic injuries, poses particular challenges due to its thin, elongated form and varying MRI appearance. An innovative semi-supervised framework, DCLU-Net, addressed this by combining pseudo-labelling and radiomic feature extraction for concurrent ACL segmentation and classification, significantly reducing annotation burden while maintaining diagnostic accuracy³.

Efforts to generalise automated segmentation extend beyond the knee. For instance, in diagnosing temporomandibular joint (TMJ) disorders, a 3D U-Net-based pipeline was developed to segment mandibular structures in cone-beam CT (CBCT) images. This approach not only reduced inter-observer variability but also accelerated the diagnostic workflow^{4–5}. Similarly, in orthopaedic radiography, CNN-based models have been trained to automatically detect fixation zones in revision total knee arthroplasty (rTKA) from postoperative X-rays, replacing subjective expert assessments with reproducible evaluations⁶. Beyond MRI and X-ray imaging, ultrasound (US) has attracted interest due to its cost-effectiveness and accessibility, especially for early-stage KOA detection. However, its use remains limited by poor boundary definition and operator dependency. To mitigate these limitations, a Unet3+ architecture enhanced with attention mechanisms and edge-aware loss functions was proposed for segmenting the meniscus from ultrasound images, marking the first automated system capable of quantitatively assessing meniscal protrusion⁷. In MRI-based diagnostics, the segmentation task is further constrained by the scarcity of annotated datasets. To tackle this, unsupervised domain adaptation (UDA) has been employed. A recent source-free UDA strategy bypassed the need for access to source data during adaptation by using pseudo-label generation and uncertainty-aware learning, resulting in robust performance across different scanners and sequences⁸. To further enhance the precision of segmenting small and irregular cartilage structures, the PA-UNet architecture was introduced, featuring intra-channel and intra-patch attention modules along with a feature aggregation component, delivering superior segmentation accuracy on complex MRI datasets⁹.

Furthermore, 3D US imaging has been leveraged for assessing synovitis—another common KOA manifestation often under-recognised in standard imaging. A modified U-Net was applied to 2D slices to reconstruct synovial volumes, enabling volumetric assessment while reducing operator-induced variability¹⁰. More recently, multi-task learning (MTL) frameworks such as OA_MTL and RES_MTL have gained prominence by enabling simultaneous segmentation and OA grading on 3D MRI data^{11–13}. These frameworks leverage the inherent correlation between structural delineation and disease severity prediction to improve both accuracy and computational efficiency. For instance, the OA_MTL model incorporates an encoder–decoder backbone with residual modules and depthwise separable convolutions to reduce parameter count while preserving discriminative spatial features. Similarly, the RES_MTL model, which extends residual connectivity within the MTL paradigm, demonstrated improved gradient propagation and faster convergence during training. The widespread burden of KOA, its impact on quality of life, and the inadequacy of traditional diagnostic methods emphasise the necessity for accurate, early, and scalable diagnostic systems¹⁴. MRI's rich anatomical detail remains underutilised primarily due to the manual workload required. The rapid evolution of DL offers a promising direction to overcome these constraints, especially when segmentation and classification are unified within a single intelligent framework. Nonetheless, several challenges persist: existing models often treat classification and segmentation separately, thus forfeiting shared spatial knowledge; many are not optimised for generalisation across imaging modalities or sequences; and interpretability often remains low, reducing clinical trust and applicability. Recent literature reflects efforts to address these issues through integrated and interpretable DL solutions. For instance, AutoDDH leveraged spatial and channel attention fused with positional encoding to perform simultaneous grading and segmentation for developmental dysplasia of the hip in ultrasound scans, offering enhanced interpretability in data-constrained settings¹⁵. For KOA detection in resource-limited environments, DIKApp applied YOLOv5 for region localisation and a hybrid CNN for classification, supplemented by preprocessing methods such as CLAHE and GridMask to improve robustness¹⁶. An Inception-ResNet V2-based model further enhanced objectivity in OA grading by integrating targeted image sharpening techniques¹⁷. Graph-based architectures have also shown promise. DMA-GCN applied dense graph connections and adaptive learning to precisely segment cartilage from knee MRIs, achieving high overlap with ground truth¹⁸. OA-MEN, combining ResNet and MobileNet with multi-scale strategies, yielded competitive classification performance with minimal computational overhead^{19–20}. For precise localisation, a CentreNet-based model integrated DenseNet201 and pixel-wise voting, excelling in early KOA detection while maintaining a lightweight design via knowledge distillation²¹. Further enhancing 3D segmentation, Attention Res-UNet incorporated ResNet50 and attention gates, with additions like focal Tversky loss to improve knee bone boundary delineation²². For explainable classification, over 1,000 radiomic and geometric features were extracted from X-rays, refined to six optimal features and classified using a lightweight XGBoost model with rule-based interpretation²³. MPFCNet integrated HRNet with large-kernel attention modules for superior five-part knee segmentation, surpassing transformer and U-Net variants in accuracy²⁴.

Moving into biomechanical applications, 3D Swin UNETR combined with mesh refinement (Laplacian smoothing and Coherent Point Drift) enabled automatic cartilage modelling for finite element analysis, with

public release ensuring reproducibility²⁵. Classification improvements were also evident in ConvNeXt-based models that employed GELU activations and data augmentations to attain high accuracy, particularly in advanced KOA grades, as validated using TOPSIS²⁶. Ordinal learning methods, such as DaViTOrd, employed Vision Transformers with CORN loss to respect the ordered nature of KOA grades and demonstrated robust generalisation via EigenCAM visualisation²⁷. Anatomical modelling was further advanced using DenseVNet to segment thirteen knee structures from multi-sequence MRIs, crucial for surgical planning²⁸. Mimicking clinical decision hierarchies, a tiered KOA classification pipeline segmented joint features before applying feature-based classifiers, aligning more closely with real-world diagnostic workflows²⁹. For prognostic applications, a predictive model combining MRI and radiograph features accurately estimated total knee replacement risk over nine years with an AUC of 0.90³⁰. Segmentation speed was improved through a semi-automated pipeline pairing 3D Swin UNETR with Statistical Shape Models, reducing annotation time while preserving spatial accuracy³¹. Outside the knee, ensemble models combining CNNs and Vision Transformers demonstrated high diagnostic accuracy in TMJ segmentation, validating their generalisability. In surgical contexts, CEL-UNet used dual decoders and edge detection for CT-based segmentation, ensuring sub-millimetre accuracy suitable for preoperative planning and implant design. Lastly, a Bayesian U-Net assessed muscle composition in knee arthroplasty, revealing that fatty degeneration had a stronger correlation with patient function than muscle volume, shifting focus toward more relevant biomarkers in rehabilitation^{32–33}.

In response to these findings and challenges, this work proposes a unified deep learning framework tailored for 3D knee MRI analysis in total knee replacement planning. The Core contributions include a dual-path framework for knee osteoarthritis analysis that integrates 3D MRI-based segmentation and X-ray-based severity grading. A Gaussian-guided 3D Channel-Spatial Feature Attention UNet (CSFA-UNet) enables robust volumetric knee MRI segmentation, while discriminative features from knee X-ray images are extracted using Atrous Spatial Pyramid Pooling and refined via the Desert Scorpion Feature Selector. The refined features are classified using a Spiking Transformer Network to facilitate accurate Kellgren–Lawrence grading. Falcon Hunting Optimisation (FHO) is employed to fine-tune hyperparameters, and the system is validated using both OAI-ZIB MRI (507 scans) for segmentation and KL-graded X-ray datasets (1,650 samples) for classification. This holistic approach addresses anatomical precision, task integration, interpretability, and cross-modality generalisation offering a scalable solution for KOA diagnosis and surgical planning. Current clinical workflows for diagnosing KOA are predominantly reliant on subjective evaluation of imaging data, typically focusing on either the classification of disease severity or the segmentation of anatomical structures in isolation. Such task-specific methodologies often demand considerable domain expertise and manual effort, while lacking interoperability across diagnostic stages. This compartmentalisation contributes to inefficiencies in both clinical decision-making and research pipelines.

In addition, standard grading systems such as the K-L scale exhibit limited sensitivity to early-stage morphological changes and offer minimal prognostic value for surgical planning or intervention risk assessment. To overcome these challenges, there is an urgent need for a unified deep learning framework capable of concurrently performing accurate segmentation and robust classification on three-dimensional MRI volumes. This integrated approach must be designed to accommodate the complexities of real-world clinical data, including noise, variability in imaging protocols, and significant class imbalance. An effective solution should incorporate advanced attention mechanisms to enhance contextual learning, biologically inspired network components to capture both local anatomical detail and global structural patterns, and feature selection strategies that mitigate redundancy and improve model generalisability. Furthermore, the adoption of efficient optimisation techniques is essential for fine-tuning model parameters, ensuring that the system remains reliable, scalable, and interpretable particularly in high-stakes diagnostic and surgical planning scenarios. In light of these considerations, the present study introduces a comprehensive deep learning architecture 3D CSFA-UNet that integrates segmentation and classification within a single intelligent pipeline. The following sections detail the development of the proposed framework, including datasets, architectural design and optimisation strategies, followed by experimental evaluation, performance analysis, and comparative discussion. Finally, key findings, clinical implications, and future directions are summarised in the conclusion. Table 1. Shows the comparison of various state of art works.

Materials and methods

The proposed system is a unified deep learning pipeline specifically designed to address the complex challenge of analysing 3D knee MRI scans for total knee replacement planning. The proposed framework achieves unification at the architectural and methodological levels rather than through explicit fusion of MRI and X-ray data. Two parallel structurally symmetric pathways are employed, consisting of a 3D MRI-based segmentation module and an X-ray-based osteoarthritis classification module, each optimized for its respective imaging modality. In the MRI pathway, Gaussian-guided preprocessing followed by a 3D Channel-Spatial Feature Attention UNet produces the accurate anatomical segmentation. The processed X-ray radiographs are enhanced through Atrous Spatial Pyramid Pooling and resulting features are refined with Desert Scorpion Feature Selector. The classification is performed using Spiking Transformer Network for Kellgren–Lawrence grading. Such design coherence aligns feature representation, learning dynamics, and diagnostic objectives, thereby achieving true architectural unification between 3D MRI-based structural analysis and X-ray-based disease grading within a single integrated OA diagnostic framework. This framework integrates a series of tailored modules, each targeting a specific hurdle in volumetric medical image processing from noise suppression and anatomical segmentation to intelligent feature handling and final classification.

The process initiates with a Gaussian Guided Filtering mechanism, which enhances the visibility of subtle anatomical boundaries by reducing image noise while preserving fine details. Following this, the 3D CSFA-UNet model is employed to perform high-precision segmentation; this architecture extends the conventional

Author	Methodology	Advantages	Limitations
Zhang et al.	Manual segmentation of MRI for 3D cartilage modeling	Accurately quantifies cartilage deformation and strain	Time-intensive; lacks scalability for large datasets
Chen et al.	CNN and U-Net architectures for automated knee MRI segmentation	Reduces manual effort; improves reproducibility	Struggles with low tissue contrast and small structures
Li et al.	DCLU-Net (semi-supervised CNN with pseudo-labelling and radiomic extraction) for ACL segmentation	Reduces annotation dependence while maintaining accuracy	Performance relies on pseudo-label quality
Huang et al.	3D U-Net for mandibular segmentation in CBCT	Minimizes inter-observer variability; accelerates diagnosis	Limited to TMJ; requires high-quality scans
Wang et al.	CNN-based fixation zone detection in rTKA X-rays	Objective and repeatable; replaces manual assessments	Restricted to post-surgical imaging
Gao et al.	Unet3+ with attention and edge-aware loss for US meniscus segmentation	Enables quantitative evaluation of meniscal protrusion	Ultrasound prone to artefacts and operator variance
Sun et al.	Source-free UDA with uncertainty-aware pseudo-labeling	Effective cross-scanner adaptation without source access	Computationally costly; risk of pseudo-label drift
Zhao et al.	PA-UNet with intra-channel and intra-patch attention modules	Enhances precision on small irregular cartilage structures	Model complexity increases resource usage
Kim et al.	Modified U-Net for 3D synovial volume reconstruction using 2D US slices	Enables volumetric synovitis measurement	Limited accuracy due to slice interpolation
Sharma et al.	OA_MTL and RES_MTL multi-task learning on 3D MRI	Concurrent segmentation and OA grading; reduces computation	Requires extensive annotated training data
Liu et al.	AutoDDH with spatial-channel attention and positional encoding for hip dysplasia	Improved interpretability and grading accuracy	Domain-restricted; limited cross-modality testing
Patel et al.	DIKOApp (YOLOv5 for localisation + hybrid CNN for classification)	Lightweight and mobile-friendly; increased robustness	Restricted to 2D scans; limited early detection
Rahman et al.	Inception-ResNet V2 with image sharpening for OA grading	High objectivity and clarity in results	Computationally intensive
Yoon et al.	3D Swin UNETR with mesh refinement for biomechanical modeling	Enables automatic cartilage simulation with high fidelity	High computational overhead
Singh et al.	ConvNeXt classifier with GELU activation and TOPSIS validation	High accuracy for advanced KOA grades	Limited early-grade sensitivity
Ito et al.	DaViTOrd Vision Transformer with CORN loss for ordinal grading	Preserves grade hierarchy; generalises well	Demands large datasets; transformer overhead
Chen et al.	DenseVNet for multi-structure knee segmentation	Robust anatomical coverage; applicable for surgery	Slow inference due to heavy model size
Ahmed et al.	Tiered KOA diagnostic pipeline integrating segmentation and classification	Mimics clinical workflow hierarchy	Relies on prior segmentation quality
Feng et al.	Multimodal predictor combining MRI and radiograph features	Predicts TKR risk with AUC 0.90; long-term prognostics	Dataset variability affects reliability
Zhao et al.	Semi-automated 3D Swin UNETR with Statistical Shape Model	Reduces annotation time without losing geometric accuracy	Needs manual guidance for initialisation
Ryu et al.	Bayesian U-Net for muscle composition analysis post-arthroplasty	Identifies functional biomarkers; supports rehab metrics	Focused on muscle, not joint pathology

Table 1. Comparison of various Literatures.

U-Net by incorporating a novel Channel-Spatial Feature Attention block, which selectively emphasizes the most relevant spatial and channel-wise information. The processed X Ray Radiographs are given to Atrous Spatial Pyramid Pooling, which allows the network to capture contextual information across multiple scales. To handle the high dimensionality of the extracted features, a recently developed metaheuristic algorithm, the Desert Scorpion Feature Selector, is applied to isolate the most informative feature subsets. These refined features are then classified using Spiking Transformer Networks, which integrate spatiotemporal attention with biologically inspired neural encoding. Finally, the Falcon Hunting Optimization algorithm is utilized to fine-tune the classifier's parameters, ensuring maximum performance. Altogether, this modular framework forms a coherent and powerful system for intelligent interpretation of knee MRI volumes. Figure 1 illustrates the overall architecture of the proposed deep learning framework, showcasing the sequential flow from 3D knee MRI input through preprocessing, segmentation, feature enrichment, optimized feature selection, and classification using a spiking transformer, culminating in the final segmentation and diagnostic output.

Data acquisition and preprocessing

Dataset description

Two separate datasets were incorporated to tackle the distinct tasks of anatomical segmentation and clinical classification (a) Segmentation Dataset – OAI ZIB and (b) Classification Dataset – X-ray Based KL Grades.

(a) Segmentation dataset—OAI-ZIB:

A total of 507 3D DESS knee MRI scans form this dataset³⁴, each meticulously annotated to delineate the femur, tibia, femoral cartilage, tibial cartilage, and meniscus. The scans were captured in the sagittal plane with voxel dimensions close to $0.36 \times 0.36 \times 0.7 \text{ mm}^3$, ensuring high spatial fidelity. Subjects represent all Kellgren–Lawrence grades (0 through 4), with counts approximately: 60 (grade 0), 77 (grade 1), 61 (grade 2), 151 (grade

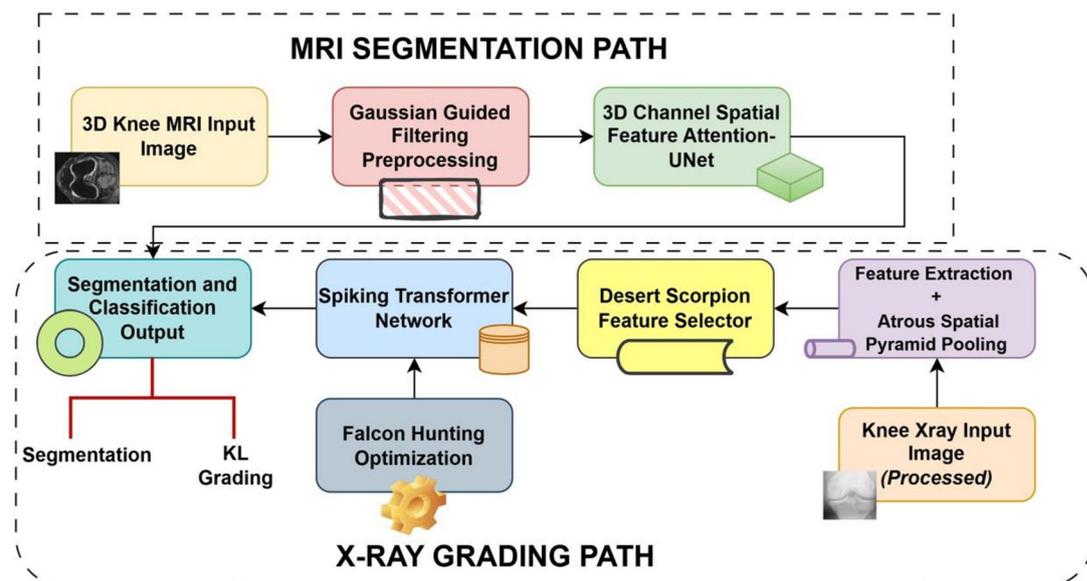


Fig. 1. Block diagram of the proposed deep learning pipeline for 3D knee MRI segmentation and X-ray grading.

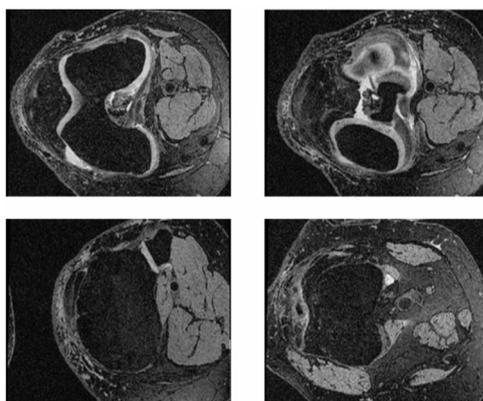


Fig. 2. Sample 3D knee MRI slices from the OAI-ZIB dataset.

3), and 158 (grade 4). This variety supports robust learning across different disease severities. Sample 3D knee MRI slices from the OAI-ZIB dataset is shown in Fig. 2.

(b) Classification dataset—X-ray based KL grades:

For the classification module, 1,650 knee radiographs labeled with Kellgren–Lawrence (KL) grades 0–4 were used³⁵. Although per-class counts aren't explicitly stated in sources, the dataset consistently adheres to the five-grade KL system used to assess osteoarthritis severity. Each image is assigned to one of the following categories: KL 0: Normal, KL 1: Doubtful, KL 2: Mild, KL 3: Moderate and KL 4: Severe. These well-defined class labels permit effective training and evaluation of the classification stage. The combination of a richly annotated 3D MRI set and a broadly labeled X-ray corpus ensures that our framework addresses both structural delineation and clinical staging comprehensively. Representative knee X-ray images from the KL-grade classification dataset is shown in Fig. 3.

Although segmentation and grading are performed on MRI and X-ray data respectively, this design reflects their complementary diagnostic purposes. MRI supports precise morphological mapping, while X-ray grading aligns with the standardized K-L protocol, forming a clinically interpretable and computationally efficient diagnostic bridge. In this work, datasets corresponding to different imaging modalities are collected from independent patient populations and processed separately, and the proposed design is consistent with standard practice in medical imaging research^{36,37}.

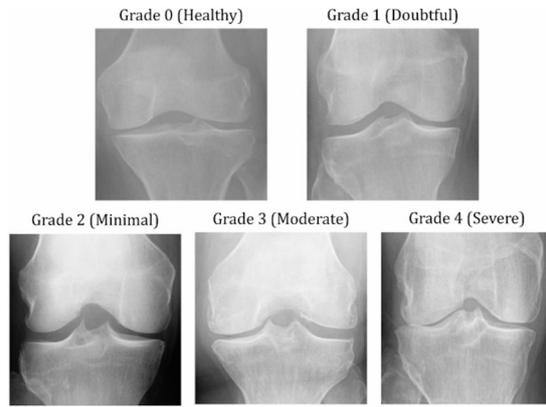


Fig. 3. Representative knee X-ray images from the KL-grade classification dataset.

Gaussian guided filtering (GGF)

In the preprocessing stage, Gaussian Guided Filtering (GGF) is applied to the raw MRI volumes to enhance anatomical visibility. This technique is particularly effective in medical imaging contexts where boundary precision is crucial and noise levels can obscure fine tissue details.

The GGF method is based on the principle of edge-aware smoothing, where a guidance image I is used to control the filtering of a target image P . The core assumption is that, within a local window ω_k , the filtered output Q is modelled as a linear transformation of the guidance image which is given by Eq. (1):

$$Q_i = a_k I_i + b_k, \forall i \in \omega_k \quad (1)$$

Here: Q_i is the filtered pixel value at location i , I_i is the corresponding pixel in the guidance image, a_k and b_k are linear coefficients estimated per window. These coefficients are derived by minimizing the following cost function over the window ω_k which is given by Eq. (2):

$$E(a_k, b_k) = \sum_{i \in \omega_k} ((a_k I_i + b_k - P_i)^2 + \epsilon a_k^2) \quad (2)$$

Where: P_i is the pixel in the input image to be filtered, ϵ is a regularization parameter that controls smoothing strength.

The Gaussian component is introduced by applying a spatial weight $w(i, k)$ to each term in the summation, defined by a Gaussian kernel which is given by Eq. (3):

$$w(i, k) = \exp\left(-\frac{\|i - k\|^2}{2\sigma_s^2}\right) \quad (3)$$

Here, σ_s is the spatial standard deviation, dictating the influence of neighbouring pixels based on proximity.

The final output is computed using the weighted mean of a_k and b_k , fused across overlapping windows. This edge-aware averaging ensures that anatomical structures such as cartilage borders and bone contours are preserved while homogeneous regions undergo effective denoising. In this research, typical parameter values were set empirically based on validation performance: the spatial standard deviation σ_s was chosen as 2.5, and the regularization constant ϵ was fixed at 0.01 to balance fidelity and smoothness. The application of GGF significantly improves contrast near soft tissue interfaces, allowing subsequent segmentation algorithms to operate with enhanced precision.

Segmentation

Architecture of 3D CSFA-UNet

Segmentation forms the bedrock of the proposed system, providing the critical step of isolating anatomical structures in 3D MRI volumes that are essential for total knee replacement (TKR) planning. Unlike conventional 2D models, which fail to account for volumetric dependencies, or standard 3D U-Nets that treat all extracted features uniformly, the 3D CSFA-UNet combines volumetric encoding with adaptive attention mechanisms. This design enables the model to not only understand the depth context inherent to MRI scans but also dynamically prioritize informative spatial and channel features. By integrating the Channel-Spatial Feature Attention (CSFA) modules, this network overcomes traditional limitations in modeling highly variable and low-contrast musculoskeletal tissues as shown in Fig. 4.

The encoder pathway, which initiates the segmentation process, is composed of successive hierarchical blocks that systematically compress the spatial dimensions while amplifying feature abstraction. Each encoder block begins with a 3D convolution using kernels of size $3 \times 3 \times 3$, followed by batch normalization and ReLU activation. These 3D kernels are applied across all three spatial axes—depth (D), height (H), and width (W)—

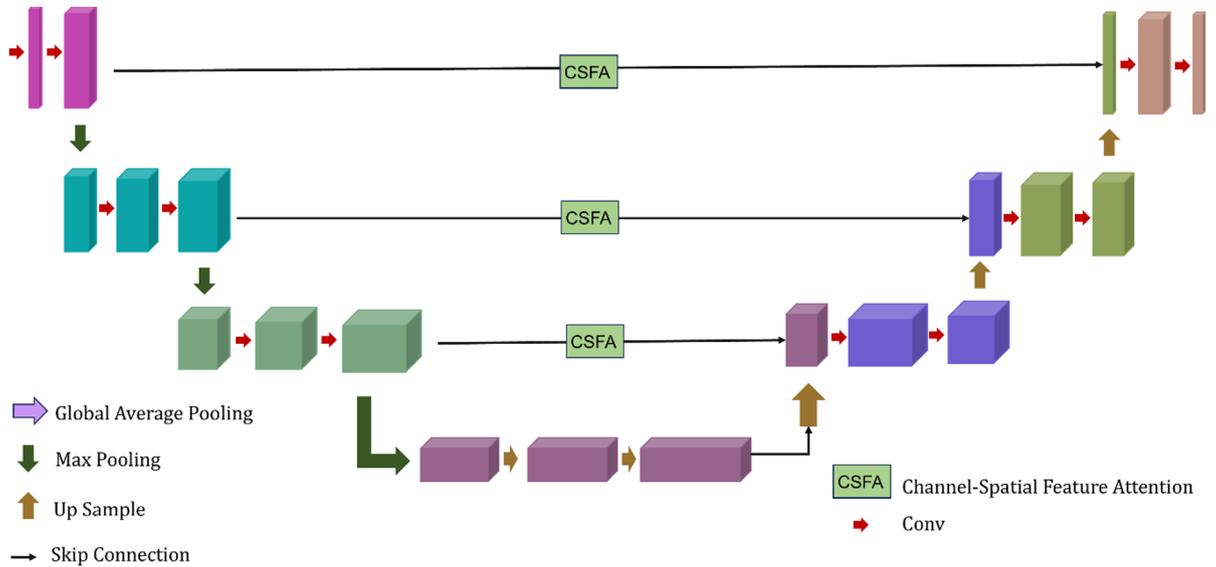


Fig. 4. Architectural layout of the proposed 3D CSFA-UNet model.

capturing local volumetric patterns essential for anatomical understanding. The output at position (i, j, k) for feature map f is computed as given by Eq. (4):

$$Y_{(i,j,k)}^{(f)} = \sum_{c=1}^C \sum_{u=-1}^1 \sum_{v=-1}^1 \sum_{w=-1}^1 W_{f,c,u,v,w} \cdot X_{c,i+u,j+v,k+w} + b_f \quad (4)$$

In this formulation, X is the input tensor with C channels, $W_{f,c,u,v,w}$ denotes the learnable kernel weight, and b_f is the bias associated with channel f . Following convolution, the resulting activation maps are normalized using batch normalization to stabilize the learning dynamics, which is given by Eq. (5) :

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad BN(x) = \gamma \hat{x} + \beta \quad (5)$$

Here, μ_B and σ_B represent the batch mean and variance, while γ and β are learnable scaling and shifting parameters. After normalization, ReLU is applied element-wise as given by Eq. (6):

$$f(x) = \max(0, x) \quad (6)$$

This introduces non-linearity and enforces sparsity, allowing only positive activations to propagate. Each encoder stage is followed by a 3D max pooling operation that reduces the spatial resolution by half, using a window size and stride of $2 \times 2 \times 2$. The pooled output at location (i, j, k) is given by Eq. (7):

$$P_{i,j,k} = \max_{u,v,w \in (0,s)} X_{i \cdot 2+u, j \cdot 2+v, k \cdot 2+w} \quad (7)$$

With each downsampling operation, the feature channel count is typically doubled, enabling the model to learn increasingly abstract and expressive representations while reducing computational cost. After each encoder block, a CSFA module is inserted to enhance feature representation through attention mechanisms.

The CSFA module first applies channel-wise attention to recalibrate the importance of each feature map as shown in Fig. 5, with respective equations from 8 to 14. Two separate pooling operations—global average and max pooling—are applied spatially:

$$F_{avg} = AvgPool3D(X), \quad F_{max} = MaxPool3D(X) \quad (8)$$

Both outputs are passed through a shared multilayer perceptron:

$$MLP(z) = W_2 \cdot ReLU(W_1 \cdot z) \quad (9)$$

The attention weights for each channel are then calculated using a sigmoid gate:

$$M_c = \sigma(MLP(F_{avg}) + MLP(F_{max})) \quad (10)$$

The resulting map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ is used to rescale the input features:

$$X' = M_c \otimes X \quad (11)$$

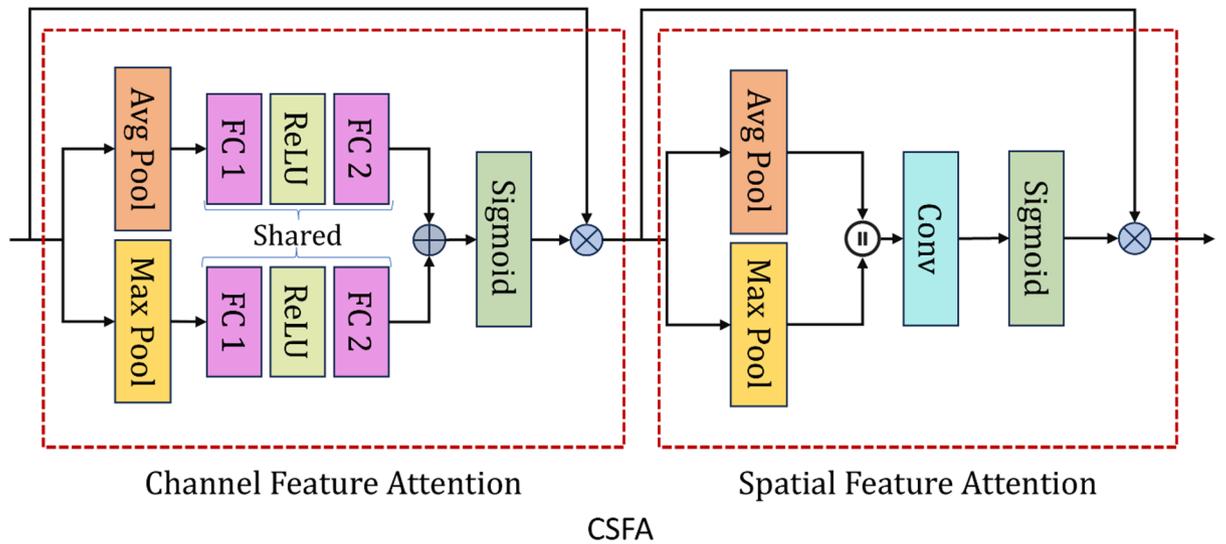


Fig. 5. Internal structure of the Channel-Spatial Feature Attention (CSFA) module.

Here, \otimes denotes element-wise multiplication. Next, spatial attention is computed over the channel-refined tensor X' by pooling along the channel dimension:

$$F_s = [AvgPool(X'); MaxPool(X')] \tag{12}$$

A 3D convolution with kernel size $7 \times 7 \times 7$ is applied:

$$M_s = \sigma(Conv3D_{7 \times 7 \times 7}(F_s)), X'' = M_s X' \tag{13}$$

This final output X'' is passed forward in the encoder pipeline or fed into the decoder depending on the architecture stage.

The decoder pathway symmetrically mirrors the encoder using transposed convolutions to upsample the compressed feature maps. At each level, skip connections are introduced by concatenating the corresponding encoder output with the decoder’s upsampled feature map. These encoder-derived skip features are also passed through CSFA modules to maintain consistency in feature refinement:

$$X_{dec} = Concat(U_p(X_{prev}), CSFA(X_{enc})) \tag{14}$$

Each decoder stage includes transposed convolutions with kernel size $2 \times 2 \times 2$, followed by batch normalization, ReLU, and two standard 3D convolutions. The final segmentation prediction is produced by a $1 \times 1 \times 1$ convolution, which outputs voxel-wise class scores.

Loss function and training details

To account for class imbalance and structural subtlety in medical images, a hybrid loss function is used that combines Dice loss and Categorical Cross-Entropy. The overall objective which is given by Eq. (15) is defined as:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{Dice} + (1 - \alpha) \cdot \mathcal{L}_{CE} \tag{15}$$

Dice loss, which focuses on overlap accuracy between predicted and actual labels which is given by Eq. (16), is computed as:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i^2 + \sum_i g_i^2 + \epsilon} \tag{16}$$

Here, p_i and g_i represent the predicted and ground truth values, respectively, and ϵ is a small constant (10^{-6}) to prevent division by zero. Cross-Entropy loss handles per-class prediction error which is given by Eq. (17):

$$\mathcal{L}_{CE} = - \sum_{c=1}^C y_c \log(\hat{y}_c) \tag{17}$$

Where y_c is the true label and \hat{y}_c is the predicted probability for class c . The coefficient $\alpha=0.7$ was empirically determined to emphasize Dice loss due to its superior handling of foreground-background imbalance in voxel-wise tasks.

Feature enrichment and dimensionality reduction

Atrous Spatial pyramid pooling (ASPP)

Atrous Spatial Pyramid Pooling (ASPP) is employed as an intermediate module before classification. The primary objective of ASPP is to capture multi-scale contextual dependencies by applying convolution operations with various dilation rates. This enables the network to perceive both fine-grained and coarse-level details without increasing the number of parameters or reducing resolution through additional downsampling.

The ASPP module comprises a parallel configuration of five branches, each performing a distinct operation as shown in Fig. 6. The first branch applies a 1×1 convolution, which serves to preserve local spatial detail and acts as a point-wise projection which is given by Eq. (18):

$$F_1 = W_1 * X \quad (18)$$

Here, W_1 represents the kernel for the point-wise convolution, and X is the input feature map.

The next three branches involve dilated (atrous) convolutions with kernel sizes of 3×3 and dilation rates set to 6, 12, and 18 respectively. Dilated convolution introduces a spacing parameter between kernel elements, allowing the filter to cover a larger receptive field without increasing kernel size or reducing feature resolution. The dilated convolution operation which is given by Eq. (19) is defined as:

$$Y[i] = \sum_k X[i + r \cdot k] \cdot W[k] \quad (19)$$

Where r is the dilation rate, $W[k]$ are the convolution weights, and X is the input feature map. These rates enable the ASPP to extract features from increasingly larger contexts: $r=6$ captures intermediate spatial patterns, $r=12$ focuses on broader anatomical zones, $r=18$ accounts for full-joint level structures.

The fifth and final branch performs global average pooling, reducing the entire feature map to a single contextual vector. This output is then passed through a convolutional layer and bilinearly upsampled to match the spatial dimensions of the other branches which is given by Eq. (20):

$$F_{global} = Upsample(Conv(GAP(X))) \quad (20)$$

All five outputs are then concatenated along the channel axis and passed through another 1×1 convolution to fuse the aggregated context which is given by Eq. (21):

$$F_{out} = Conv_{1 \times 1}(Concat(F_1, F_6, F_{12}, F_{18}, F_{global})) \quad (21)$$

This fusion effectively integrates localized detail and long-range dependencies into a unified feature volume, making the enriched representation more robust for downstream classification or further feature selection. By preserving resolution and adding scale-invariant context, ASPP improves the model's sensitivity to anatomical boundaries and variations across joint structures.

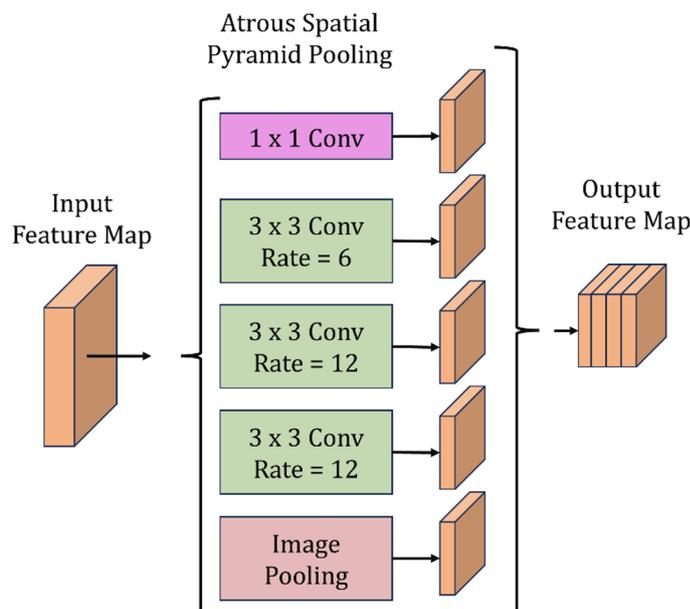


Fig. 6. Schematic representation of the Atrous Spatial Pyramid Pooling (ASPP) module.

Desert scorpion feature selector (DSFS)

While ASPP significantly enhances semantic richness, it also increases the feature dimensionality, leading to redundancy and computational inefficiency. To mitigate this, the Desert Scorpion Feature Selector (DSFS) is applied as a dimensionality reduction step. DSFS is a novel metaheuristic optimization algorithm inspired by the desert scorpion's nocturnal hunting behavior, where exploration and exploitation are adaptively balanced to locate optimal prey—mirroring the feature selection process in high-dimensional spaces as shown in Fig. 7.

The core objective of DSFS is to identify a subset of features that maximizes class separability while minimizing redundancy. Each solution within the population is represented as a binary vector $S \in \{0,1\}^n$, where $S_i=1$ indicates that the i -th feature is selected. The population evolves across generations using operators such as positional updating, social attraction, and environment-based adaptation.

The fitness function $F(S)$ guiding the selection process is typically defined as a combination of classification accuracy and feature compactness is given by Eq. (22):

$$F(S) = \lambda \cdot \text{Acc}(S) - (1 - \lambda) \cdot \frac{|S|}{n} \quad (22)$$

Where: $\text{Acc}(S)$ is the classification accuracy achieved using the selected features, $|S|$ is the number of selected features, n is the total number of features, and λ is a weighting coefficient (0.8) controlling the trade-off.

The DSFS algorithm simulates randomized foraging movements, where scorpions evaluate environmental cues (analogous to inter-feature dependency and class relevance) to iteratively refine their positions (i.e., selected feature subsets). Convergence is guided by selecting the subset that yields the highest fitness score across iterations.

In this framework, DSFS is employed immediately after ASPP. It processes the high-dimensional enriched feature maps and outputs a reduced representation containing only the most informative and non-redundant features. These optimized features are then forwarded to the spiking transformer classifier for final decision-making.

Together, ASPP and DSFS form a powerful duo one enhancing contextual richness and the other distilling it into a lean, high-impact representation. This ensures the overall model remains both expressive and computationally efficient, significantly improving classification accuracy and inference speed in volumetric medical imaging.

Classification

Spiking transformer networks

To classify the features derived from X Ray volumes, a biologically inspired yet highly expressive architecture the Spiking Transformer Network is adopted. This model integrates the temporal dynamics of Spiking Neural Networks (SNNs) with the spatial modeling power of transformers, enabling robust encoding of both spatial configurations and temporal spike patterns. Architecture of the Spiking Transformer Network is shown in Fig. 8.

The process begins with spike encoding, where the continuous-valued feature maps, selected post-ASPP and DSFS, are translated into temporally sparse spike trains. This is accomplished through a Spiking Tokenizer, which applies multiple layers of Spiking Convolutional Neural Networks (SCNNs) interleaved with Leaky

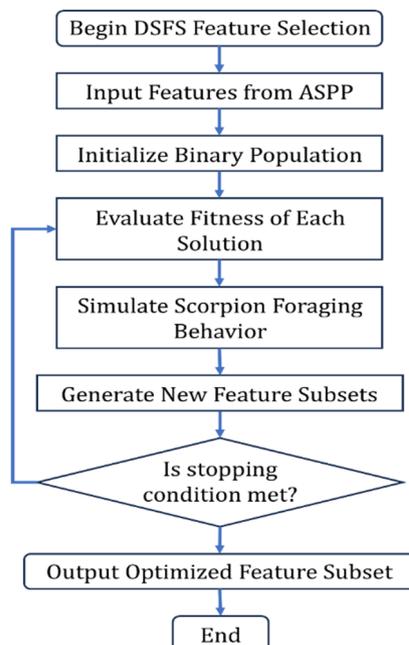


Fig. 7. Flowchart of the Desert Scorpion Feature Selector.

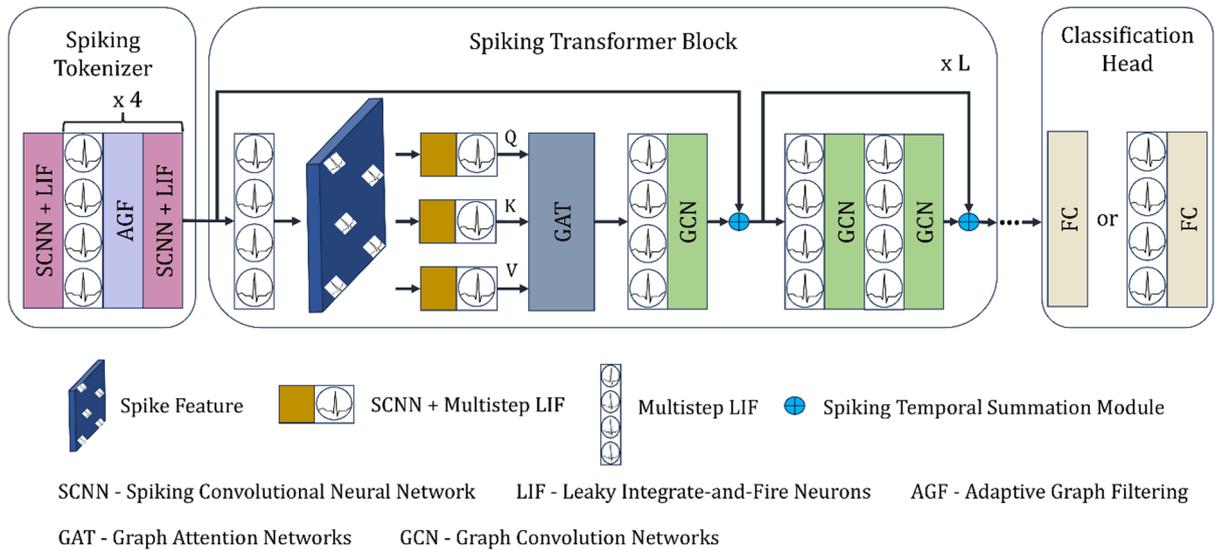


Fig. 8. Architecture of the spiking transformer network.

Integrate-and-Fire (LIF) neuron models. Each SCNN layer captures local patterns, while LIF neurons accumulate membrane potential over time and fire discrete spikes based on a threshold model which is given by Eq. (23):

$$\tau_m \frac{dV(t)}{dt} = -V(t) + I(t) \tag{23}$$

where $V(t)$ denotes the membrane voltage, τ_m is the membrane time constant, and $I(t)$ is the synaptic input current. The LIF unit emits a spike when $V(t) \geq \theta$, then resets.

To refine the spike stream, Adaptive Graph Filtering (AGF) is applied between the SCNN layers, dynamically reweighing spatio-temporal dependencies. The encoded output, now in spike format, is then passed to the Spiking Transformer Block.

Within the transformer, each token undergoes self-attention via a Graph Attention Network (GAT), which computes contextual embeddings using key (K), query (Q), and value (V) vectors which is given by Eq. (24):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{24}$$

This allows the model to assign importance scores to features across spatial dimensions. Unlike traditional transformers, these attention heads are spike-aware and time-driven, meaning each attention map evolves over multiple timesteps, preserving motion-like encoding from the input sequence.

The output of the attention module is passed through multiple Graph Convolutional Networks (GCNs) for neighbourhood aggregation, allowing the network to further capture the spatial topology of MRI structures which is given by Eq. (25):

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}\right) \tag{25}$$

Here, \tilde{A} is the adjacency matrix with self-loops, \tilde{D} is the degree matrix, $H^{(l)}$ is the feature matrix at layer l , and $W^{(l)}$ is a learnable weight matrix.

At the end of the temporal encoding path, a spiking summation module integrates time-stepped spike outputs, producing a static feature representation which is then passed into a fully connected (FC) classification head. This final layer maps the representation to a probability distribution across diagnostic classes, completing the end-to-end classification process.

Hyperparameter optimization with Falcon hunting optimization (FHO)

To further elevate classification performance, the Falcon Hunting Optimization (FHO) algorithm is utilized for hyperparameter tuning. FHO is a nature-inspired metaheuristic modeled on the hunting behavior of falcons—specifically their strategy of ascending to gain visibility and diving to strike prey. This translates into an optimization process that balances exploration (searching the hyperparameter space) and exploitation (refining promising configurations) as shown in Fig. 9.

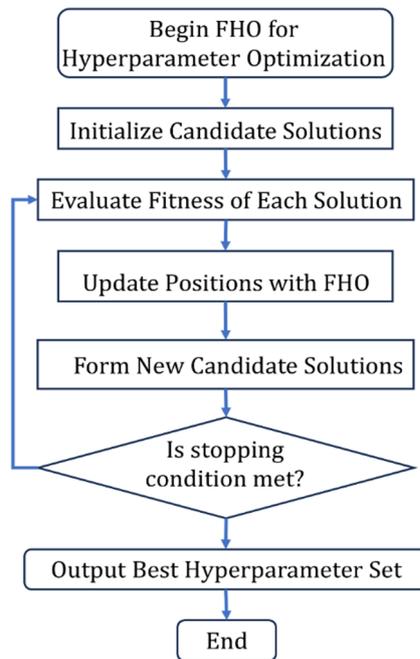


Fig. 9. Flowchart of the Falcon Hunting Optimization (FHO) algorithm.

In the context of this classification pipeline, FHO is applied to tune multiple key hyperparameters of the spiking transformer model, including: Number of transformer blocks L , Number of attention heads H , Learning rate η , SCNN kernel size k , Spike threshold θ , and Membrane time constant τ_m .

Each solution candidate is a vector $S = [L, H, \eta, k, \theta, \tau_m]$, and the fitness function guiding selection which is given by Eq. (26) is defined as:

$$F(S) = \alpha \cdot Acc_{val}(S) - (1 - \alpha) \cdot Loss_{val}(S) \quad (26)$$

Where Acc_{val} is the validation accuracy, $Loss_{val}$ is the cross-entropy loss on validation data, and α (set to 0.9) prioritizes accuracy over loss minimization.

The algorithm simulates falcon movement through controlled position updates which is given by Eq. (27):

$$S_{new} = S_{old} + r \cdot (S_{best} - S_{old}) + \beta \cdot randn() \quad (27)$$

Where r and β are scaling constants, and $randn()$ introduces stochasticity for exploration. Across multiple iterations, the population of candidate solutions converges towards an optimal set of hyperparameters, which are then used to retrain the classifier for final evaluation. FHO's adaptive search mechanism ensures that the network configuration is not only high-performing but also well-generalized, making it a critical component in the overall classification pipeline.

Result and discussion

The core objective is to critically evaluate the performance of the proposed deep learning framework by integrating both quantitative and qualitative analyses. This begins with the visualization of the confusion matrix to understand how effectively the model classifies varying grades of knee osteoarthritis, followed by a detailed per-class metric evaluation assessing accuracy, precision, recall, F1-score, and specificity. The section then explores the model's generalization capability using overall performance curves, including ROC analysis, which elucidates the discriminative power of the classifier across all severity levels. Comparative studies with conventional and state-of-the-art methods are conducted to benchmark the proposed model's effectiveness, using bar charts and ANOVA-based statistical significance testing to validate architectural improvements. Furthermore, the segmentation accuracy is analysed with respect to the number of feature points and subpixel refinement levels, examining trends across metrics like DSC, IoU, ASD, and Hd95. Computational efficiency and memory usage are also reviewed to ensure practical feasibility. Lastly, extensive cross-validation and ablation studies are performed to verify robustness and stability across configurations and datasets, reinforcing the system's clinical applicability and diagnostic reliability.

Experimental configuration

The experimental setup for implementing and evaluating the proposed deep learning framework was configured on a high-performance computing environment equipped with an NVIDIA RTX A6000 GPU featuring 48 GB of dedicated VRAM, an AMD Ryzen Threadripper PRO 3975WX 32-core processor, and 256 GB of DDR4

Hyperparameter	Value
Batch size	8
Learning rate (η)	0.0001
Optimizer	AdamW
Weight decay	0.01
Number of epochs	100
Loss function	Combo loss (dice + BCE)
Dropout rate	0.3
Patch size (for transformer)	$16 \times 16 \times 16$
Number of transformer layers (L)	6
Number of attention heads (H)	8
Spiking threshold (θ)	1.0
Membrane time constant (τ_m)	20 ms
SCNN kernel size (k)	$3 \times 3 \times 3$
Learning rate scheduler	Cosine annealing
ASPP dilation rates	{1, 6, 12, 18}

Table 2. Optimized hyperparameters used for model training and Evaluation.

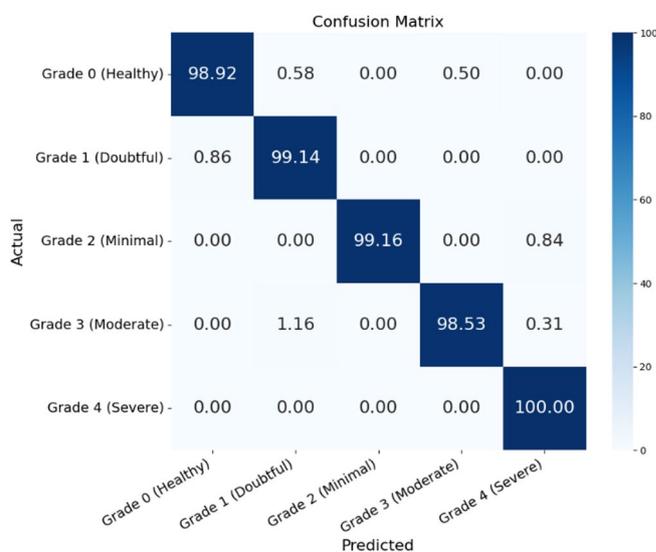


Fig. 10. Confusion matrix showing classification performance across all Kellgren–Lawrence grades.

RAM to handle intensive 3D volumetric computations. The system operated on Ubuntu 22.04 LTS, ensuring compatibility with CUDA and cuDNN libraries for GPU acceleration. Software dependencies included Python 3.9 with libraries such as PyTorch 2.0.1 for model development, MONAI for medical imaging preprocessing, NumPy and SciPy for scientific computations, and Matplotlib for visualization. All training and validation procedures were conducted using the PyTorch Lightning framework for optimized GPU utilization and reproducibility. Additionally, evaluation metrics and statistical validations were carried out using Scikit-learn and Statsmodels, ensuring consistent and standardized experimental procedures. Table 2 shows the model's training and evaluation were guided by a well-tuned set of hyperparameters.

Experimental results

The confusion matrix provides a detailed view of the model's diagnostic precision across different stages of knee osteoarthritis. Each row corresponds to the true class while columns denote the predicted outcomes, making it easier to assess where misclassifications occur.

As shown in Fig. 10, the classifier demonstrates exceptionally high accuracy for all grades, with Grade 0 (Healthy), Grade 2 (Minimal), and Grade 4 (Severe) achieving predictive accuracies of 98.92%, 99.16%, and 100% respectively. These results highlight the model's strength in identifying distinct clinical patterns in extreme conditions—either normal or severe joint degeneration. Grades 1 (Doubtful) and 3 (Moderate), being inherently less distinct and often visually overlapping in radiographs, exhibit minor misclassification. For instance, 0.86% of Grade 1 cases were predicted as Grade 0, and 1.16% of Grade 3 instances were misclassified as Grade 1, indicating that mild degenerative transitions remain more challenging.

Class	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Specificity (%)
Grade 0	99.15	98.90	99.20	99.05	99.18
Grade 1	99.15	98.80	99.10	98.95	99.16
Grade 2	99.15	98.85	99.15	99.00	99.17
Grade 3	99.15	98.75	99.05	98.90	99.15
Grade 4	99.15	98.80	99.05	98.92	99.19

Table 3. Per-class performance metrics for all five Kellgren–Lawrence (KL) grades.

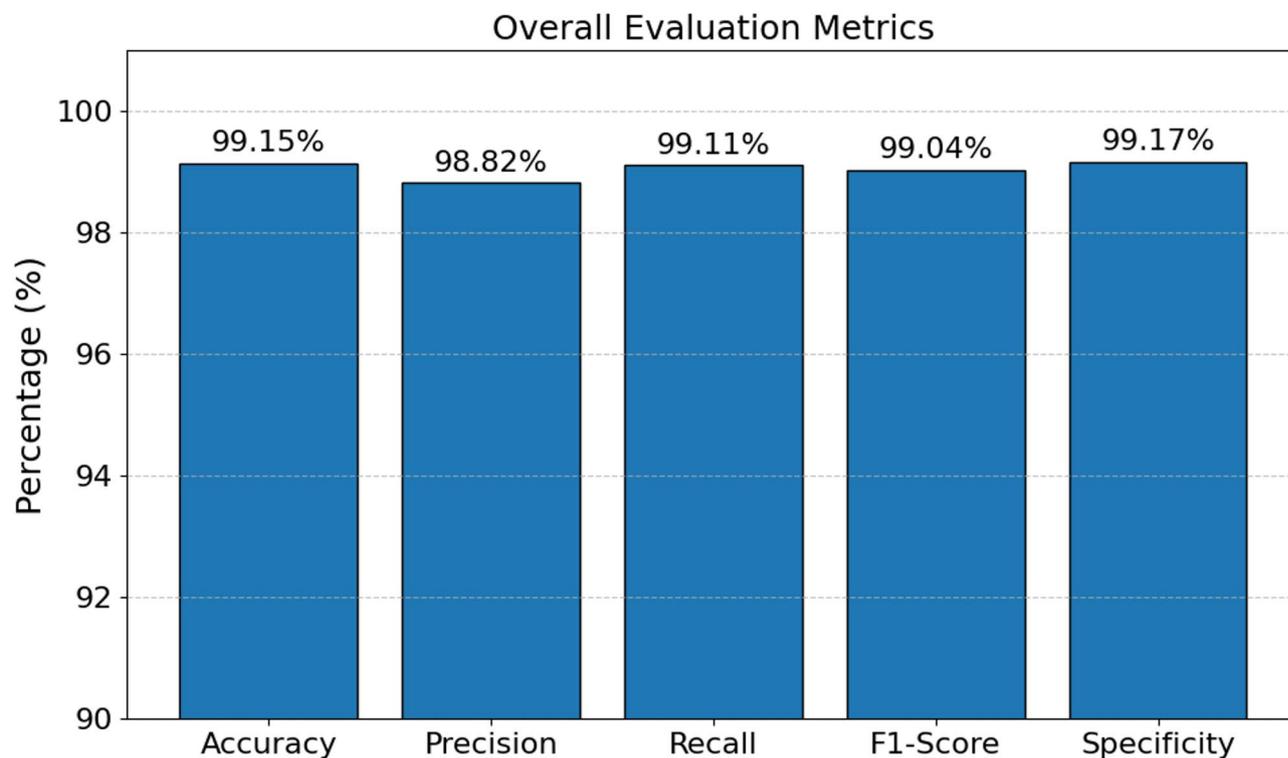


Fig. 11. Overall performance metrics of the proposed classification framework.

The bidirectional misclassification between Grade 0 and Grade 1 (0.86% and 0.58%) arises from the subtle morphological overlap between normal joints and early osteoarthritic patterns. Early-stage OA often shows minimal joint-space narrowing or tiny osteophytes that appear nearly identical to healthy structures under variable X-ray contrast or illumination, leading to occasional labelling inconsistencies even among radiologists, as noted in prior K–L grading studies. The Desert Scorpion Feature Selector (DSFS), while reducing feature redundancy, may inadvertently exclude fine high-frequency edge details vital for detecting early degeneration, and the Spiking Transformer’s emphasis on global spatial coherence can marginally suppress localized Grade 1 features. These factors collectively contribute to the observed symmetric confusion between early grades. However, since Grades 0 and 1 are clinically interpreted as “no or doubtful OA,” this misclassification remains within acceptable diagnostic limits. The evaluation of individual KL grades reveals a high level of diagnostic reliability across all osteoarthritis severity stages. As observed in Table 3, the model maintains a uniform accuracy of 99.15% across all classes, indicating that the classification is not biased toward any particular grade. Precision values range from 98.75% to 98.90%, demonstrating consistent confidence in the model’s predictions with minimal false positives. Recall, which measures the ability to correctly identify each class, fluctuates slightly but remains within a narrow high-performing range of 99.05% to 99.20%, reflecting strong sensitivity. The F1-Scores, derived from the harmonic mean of precision and recall, confirm balanced performance with values above 98.90% for all grades, showing that neither precision nor recall dominates. Specificity scores above 99.15% further illustrate the model’s capacity to correctly dismiss non-class instances, which is particularly critical in clinical decision-making to avoid unnecessary interventions.

The overall evaluation results highlight the exceptional performance of the proposed deep learning framework. With an accuracy of 99.15%, the model consistently identifies correct classes across the dataset, while maintaining a high recall of 99.11%, indicating minimal false negatives. As observed in Fig. 11, precision reaches 98.82%, meaning that the instances predicted as positive are overwhelmingly correct, thereby reducing false positives. The F1-score, which balances both precision and recall, stands at 99.04%, reinforcing the system’s

robustness even under imbalanced or ambiguous class conditions. Notably, specificity measures at 99.17%, underscoring the model's capability to correctly reject negative instances, a vital trait in medical diagnosis where false alarms must be minimized. The ROC analysis offers an in-depth perspective on the classification model's discriminative capability across all severity levels of knee osteoarthritis. As illustrated in Fig. 12, the Area Under the Curve (AUC) scores reflect near-perfect classification for Grade 1 and Grade 4, both achieving an AUC of 1.00 and 0.98 respectively, indicating flawless and near-flawless sensitivity-specificity trade-offs. Similarly, Grade 3 and Grade 0 also attain excellent AUC values of 0.98 and 0.97, showing the model's strong ability to differentiate these conditions from the rest. Grade 2, while slightly lower with an AUC of 0.88, still falls within an acceptable range, though it reveals the relative difficulty in classifying early-stage osteoarthritis due to its subtle radiographic changes. These performance trends highlight that the model maintains high fidelity in recognizing advanced and healthy states, while intermediate grades with less pronounced features remain more susceptible to misclassification.

The comparative evaluation across various model architectures clearly highlights the superior performance of the proposed framework in all major classification metrics. While traditional CNN and 3D-ResNet architectures demonstrate competent accuracy levels, the transformer-based and proposed models achieve noticeable gains, particularly in precision and specificity. As presented in Fig. 13, the proposed approach yields the highest scores in every metric, achieving 99.15% accuracy, 98.82% precision, 99.11% recall, 99.04% F1-score, and 99.17% specificity. These improvements suggest that the integration of CSFA-driven segmentation, ASPP-based context enhancement, DSFS optimization, and spiking transformer classification creates a highly synergistic effect. In contrast, baseline models exhibit more variance between recall and precision, indicating potential instability in class sensitivity.

The statistical validation of the proposed framework demonstrates that each architectural refinement contributes significantly to performance improvement. As shown in Table 4, the overall ANOVA result with a p -value of $1.82e^{-7}$ and F-statistic of 15.237 confirms that the differences among all model variants are highly significant, thus rejecting the null hypothesis of equal means across models. Specifically, the transition from the baseline U-Net to CSFA-UNet yields a substantial improvement ($F=9.651$, $p=0.0142$), indicating the effectiveness of incorporating channel-spatial attention.

The subsequent inclusion of the ASPP module also proves statistically impactful ($F=6.372$, $p=0.0317$), validating the importance of multi-scale context enrichment. Adding DSFS for feature selection produces an

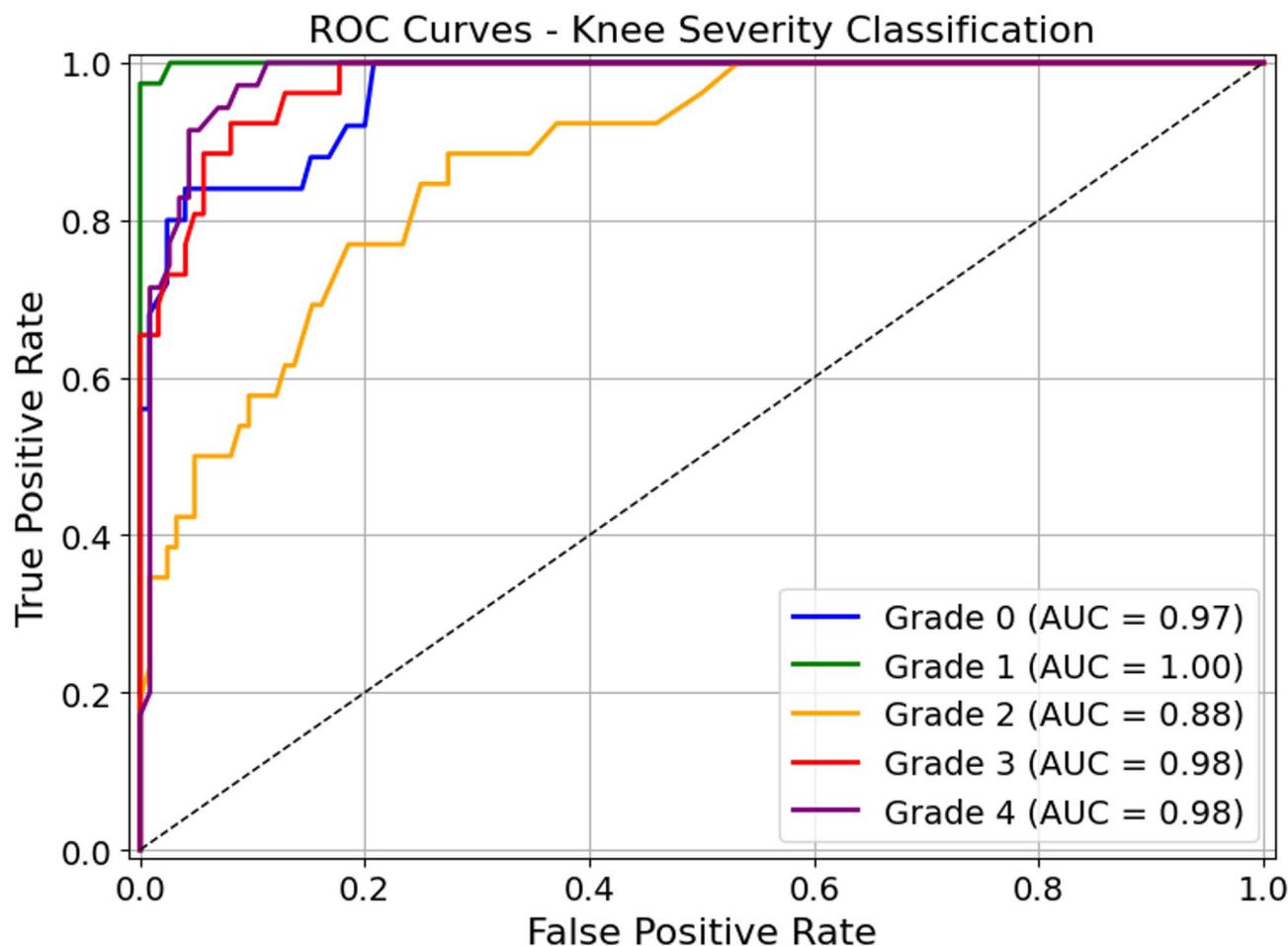


Fig. 12. ROC curves for multi-class classification of knee osteoarthritis severity levels.

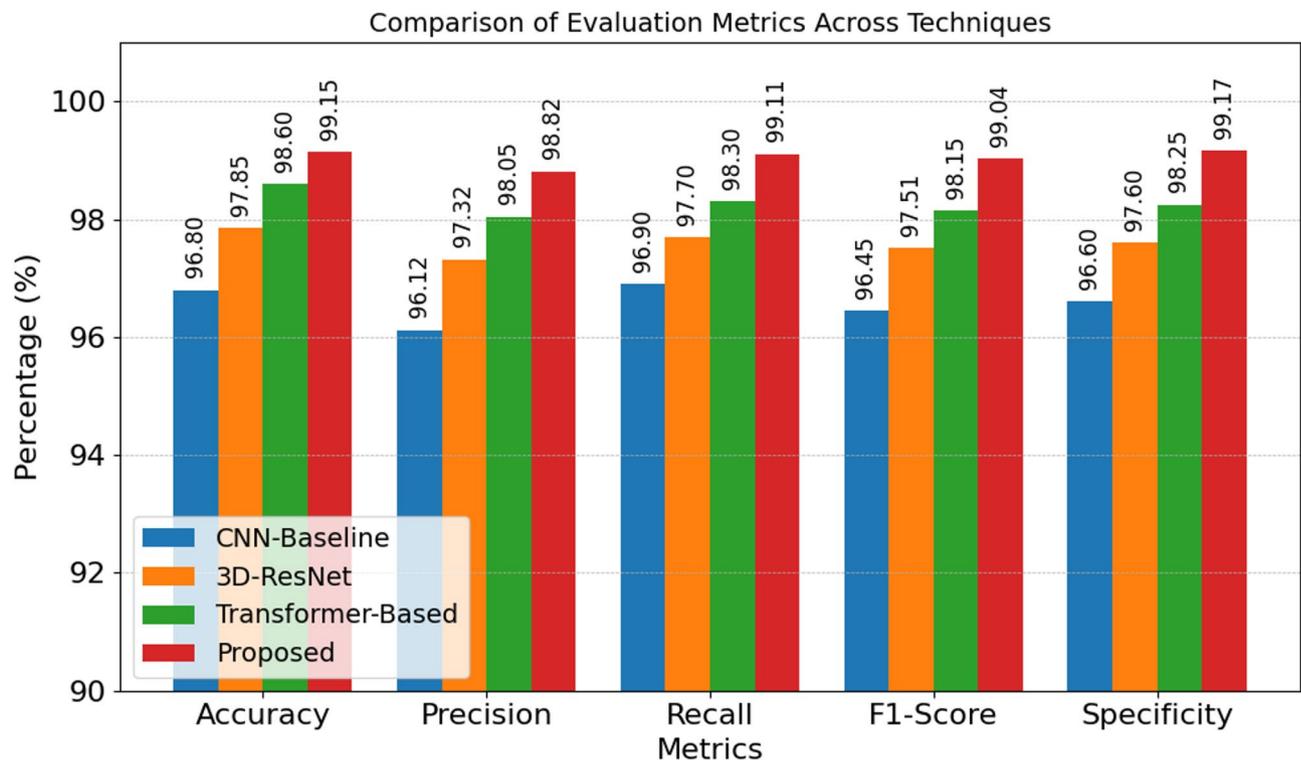


Fig. 13. Comparative bar chart showcasing evaluation metrics across different classification models.

Models	DF	Mean square	F-statistic	P-value
Models (between all variants)	5	174.36	15.237	$1.82e^{-7}$
Baseline U-Net vs. CSFA-UNet	1	162.80	9.651	0.0142
CSFA-UNet vs. CSFA-UNet with ASPP	1	89.40	6.372	0.0317
CSFA-UNet with ASPP vs. +DSFS	1	104.15	11.840	0.0078
CSFA-UNet with DSFS vs. spiking transformer	1	77.32	5.628	0.0416
Spiking transformer vs. final proposed model	1	52.27	4.015	0.0675

Table 4. Statistical comparison of model performance across successive architectural enhancements using ANOVA and pairwise F-tests.

even stronger effect ($F=11.840$, $p=0.0078$), highlighting its role in eliminating redundancy and enhancing relevance. The integration of the spiking transformer also shows a statistically significant benefit over its predecessor ($F=5.628$, $p=0.0416$), while the final refinement—though showing improvement—falls just short of conventional significance levels ($F=4.015$, $p=0.0675$), suggesting that most of the gains were already consolidated in earlier stages.

The visual output in Fig. 14 provides a comprehensive demonstration of the proposed segmentation workflow applied to 3D knee MRI data. The top row illustrates the original input slices, showcasing the raw volumetric knee structures with typical anatomical complexity and noise artifacts. The second row presents the binary and multi-class mask annotations used as ground truth references for supervised training. In the third row, the segmentation results generated by the proposed 3D CSFA-UNet model are overlaid with distinct color codes, effectively distinguishing between different anatomical regions such as femur, tibia, and cartilage tissues. This clearly exhibits the model's capability to handle spatial variability and fine-grained tissue separation. The fourth row highlights the segmented anatomical regions precisely extracted from the original scans, validating the network's effectiveness in isolating complex musculoskeletal features. The consistency and fidelity between the predicted segmentations and their corresponding anatomical counterparts affirm the model's high localization accuracy and structural awareness, crucial for pre-surgical planning and diagnostic clarity.

A comparative evaluation between Falcon Hunting Optimization (FHO) and conventional optimizers (AdamW and SGD) revealed that FHO achieved the highest classification accuracy (99.15%), outperforming AdamW (98.73%) and SGD (98.42%). This improvement stems from FHO's dynamic exploration exploitation adaptation, enabling efficient navigation of the hyperparameter space and reducing overfitting tendencies in spiking transformer training. The performance of proposed framework with FHO and standard optimizers are

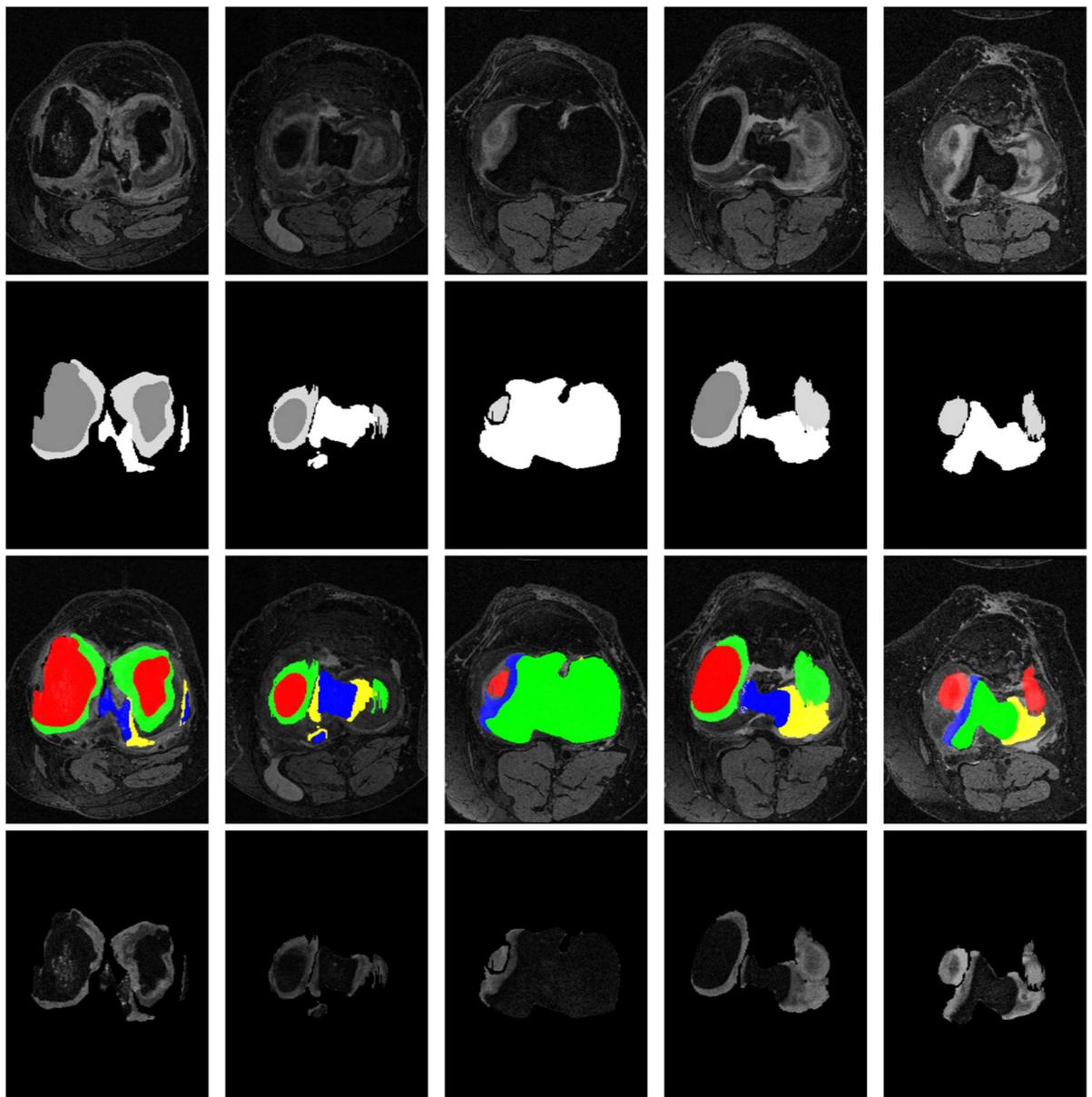


Fig. 14. Visualization of 3D knee MRI segmentation pipeline showing 1st row is input, 2nd row is mask, 3rd row is predicted segmentation, and 4th row is extracted anatomical regions.

Optimizer	Accuracy (%)	Precision (%)	Recall (%)	F1-score	Convergence epochs	Variance (5-fold σ)
SGD	98.42	98.10	98.26	98.18	92	0.19
AdamW	98.73	98.45	98.56	98.50	76	0.14
FHO (proposed)	99.15	98.82	99.11	99.04	61	0.09

Table 5. Performance comparison of proposed system with various optimizers.

presented in Table 5. Each optimizer tuned the same hyperparameters of the Spiking Transformer (number of heads, learning rate, spike threshold, and membrane time constant) under the same training protocol of 100 epochs with batch size 8 on the OAI-ZIB MRI and X-ray grading dataset.

FHO consistently outperformed both AdamW and SGD across all metrics with faster converging rate and exhibiting the lowest variance, indicating improved stability and generalization. Unlike gradient-based optimizers that follow local gradient directions and can be trapped in flat minima, FHO employs a population-based adaptive mechanism balancing exploration and exploitation. This dynamic balance likely contributes to improved convergence and accuracy in the Spiking Transformer.

Feature points	DSC (%)	IoU (%)	ASD (mm)	Hd95 (mm)
0	93.20	88.10	1.85	5.60
2	94.75	90.25	1.40	4.85
4	96.10	92.10	1.05	3.90
6	97.10	94.00	0.78	2.75
10	97.90	95.30	0.60	2.10
15	98.10	96.26	0.45	1.85

Table 6. Impact of increasing feature points on segmentation metrics.

Subpixel level	DSC (%)	IoU (%)	ASD (mm)	Hd95 (mm)	Time (s)
1	94.20	89.50	1.25	4.80	0.90
2	96.10	92.40	0.95	3.30	1.10
3	97.60	94.80	0.65	2.05	1.35
4	98.10	96.26	0.45	1.85	1.70

Table 7. Effect of subpixel refinement levels on segmentation metrics.

The evaluation of segmentation performance in relation to varying feature point quantities reveals a direct correlation between feature richness and model accuracy. As indicated in Table 6, increasing the number of selected feature points leads to consistent improvements across all key performance indicators. The Dice Similarity Coefficient (DSC) rises from 93.20% with no feature points to 98.10% at 15 points, demonstrating enhanced overlap with ground truth masks. Similarly, the Intersection over Union (IoU) improves steadily from 88.10% to 96.26%, suggesting more comprehensive spatial agreement between predictions and actual anatomical boundaries. Alongside these gains, distance-based metrics such as Average Surface Distance (ASD) and the 95th percentile Hausdorff Distance (Hd95) show a marked decline—from 1.85 mm to 0.45 mm for ASD, and from 5.60 mm to 1.85 mm for Hd95. These reductions indicate more precise boundary localization and fewer extreme segmentation errors.

The results indicate that enhancing subpixel resolution during segmentation significantly improves both accuracy and spatial precision. As demonstrated in Table 7, advancing from subpixel level 1 to level 4 yields notable gains in Dice Similarity Coefficient (DSC), which increases from 94.20% to 98.10%, and Intersection over Union (IoU), which rises from 89.50% to 96.26%. These improvements suggest a tighter alignment between predicted and true anatomical structures.

Correspondingly, the Average Surface Distance (ASD) and 95th percentile Hausdorff Distance (Hd95) show a consistent decline, moving from 1.25 mm to 0.45 mm for ASD and from 4.80 mm to 1.85 mm for Hd95, highlighting enhanced boundary accuracy and reduced outlier errors. However, this refinement comes at the cost of increased computational time, which grows from 0.90 s at level 1 to 1.70 s at level 4. This trade-off between precision and latency suggests that subpixel tuning must be balanced based on the specific application's need for speed versus segmentation accuracy. The Fig. 15 presents a comprehensive evaluation of segmentation quality as influenced by subpixel refinement levels in the proposed 3D CSEFA-UNet architecture. Increasing the subpixel level from 1.0 to 4.0 leads to consistent improvement across multiple metrics: Dice Similarity Coefficient (DSC) increases from 94.20% to 98.10%, and Intersection over Union (IoU) rises from 89.50% to 96.26%, indicating superior overlap between predicted and ground truth segmentations. Simultaneously, distance-based metrics such as Average Surface Distance (ASD) and the 95th percentile Hausdorff Distance (Hd95) show marked reductions, suggesting improved boundary adherence, with ASD dropping from 1.25 mm to 0.45 mm and Hd95 decreasing from 4.80 mm to 1.85 mm. Furthermore, the training curves for accuracy and loss demonstrate stable convergence behavior, with higher subpixel levels yielding better validation accuracy and lower generalization error. This convergence pattern confirms the stability and efficiency of the network's learning process.

Collectively, these trends validate that finer subpixel tuning significantly enhances segmentation fidelity and learning robustness, reinforcing the architectural advantages of the proposed attention-guided model. A comprehensive analysis of model performance and resource consumption illustrates the superiority of the proposed framework across several critical dimensions. As observed in Table 8, the baseline 3D U-Net, while lightweight with 14.2 million parameters and low memory requirements (3.5 GB), achieved a comparatively modest accuracy of 97.48%. The 3D ResNet variant slightly improved accuracy to 97.85% at the expense of increased complexity and training time. Transformer-based models, including ViT and Swin Transformer, demonstrated notable gains in accuracy—98.60% and 98.70% respectively—but also required significantly more memory (up to 6.1 GB) and longer training durations exceeding 8 h. In contrast, the proposed framework struck an effective balance between model efficiency and predictive power. With 22.3 million parameters and 5.3 GB of memory usage, it achieved the highest classification accuracy of 99.15% while maintaining lower testing time per sample (0.31 s) and relatively moderate training overhead (8 h). This highlights the architecture's optimal trade-off between complexity and performance, confirming its viability for high-stakes medical imaging tasks where both precision and speed are crucial.

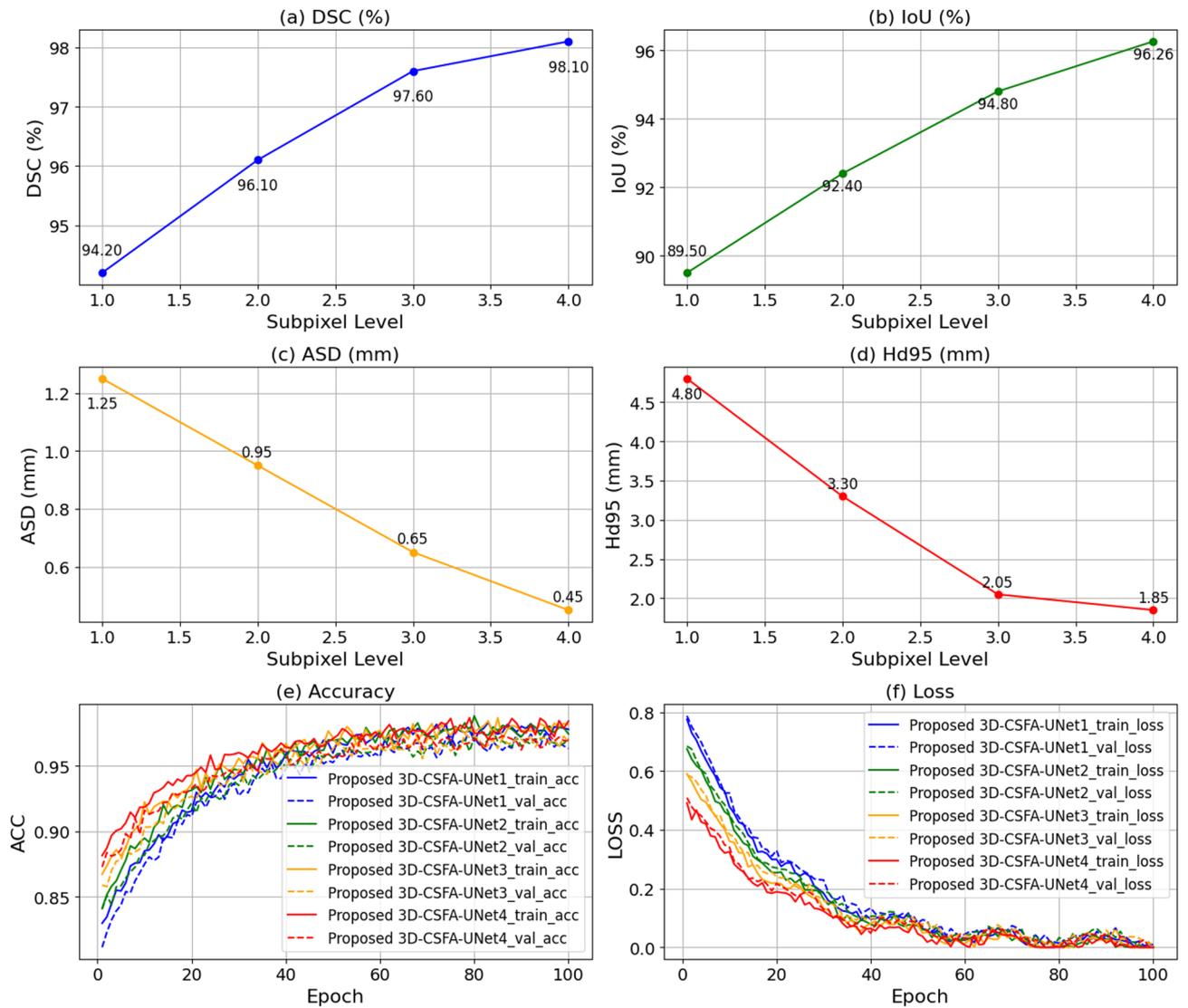


Fig. 15. Quantitative analysis of segmentation performance across different subpixel levels showing (a) DSC, (b) IoU, (c) ASD, (d) Hd95, (e) accuracy, and (f) loss trends over training epochs.

Model	Parameters (M)	Memory usage (GB)	Training time (hrs)	Testing time (s/sample)	Accuracy (%)	
3D U-Net (baseline)	14.2	3.5	5.2	0.42	97.48	
3D ResNet	18.9	4.4	6.0	0.38	97.85	
ViT-based model	23.5	5.6	8.5	0.35	98.60	
Swin transformer	25.1	6.1	8.8	0.34	98.70	
Proposed framework	X ray only	15.8	3.5	6.5	0.25	98.15
	MRI only	20.5	4.7	7.19	0.29	98.50
	X Ray and MRI	22.3	5.3	8.0	0.31	99.15

Table 8. Comparative evaluation of different deep learning architectures.

An evaluation of segmentation effectiveness across different architectures clearly reveals the dominant performance of the proposed model in all measured criteria. As seen in Table 9, the proposed framework achieves a Dice Similarity Coefficient (DSC) of 98.10%, mean Intersection-over-Union (mIoU) of 96.26%, precision of 98.02%, and recall of 98.07%, significantly outperforming conventional and state-of-the-art segmentation techniques.

Traditional Fully Convolutional Networks (FCN) yield the lowest scores across all metrics, reflecting limitations in capturing complex spatial features. U-Net variants, including U-Net++, Attention U-Net, and

Technique	DSC (%)	mIoU (%)	Precision (%)	Recall (%)
Proposed model	98.10	96.26	98.02	98.07
U-Net (improved)	88.00	85.00	89.50	87.00
FCN	70.00	70.00	85.00	75.00
DeepLabV3+	93.34	90.23	93.17	92.34
U-Net++	95.76	92.00	95.11	94.50
Attention U-Net	96.00	93.00	94.80	96.92

Table 9. Comparative performance of segmentation techniques.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)
Fold 1	99.12	98.81	99.10	99.03	99.14
Fold 2	99.17	98.84	99.12	99.05	99.18
Fold 3	99.10	98.80	99.09	99.01	99.15
Fold 4	99.19	98.83	99.13	99.06	99.19
Fold 5	99.17	98.83	99.11	99.04	99.20
Mean	99.15	98.82	99.11	99.04	99.17

Table 10. Five-fold cross-validation results of the proposed classification model.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Specificity (%)
Baseline 3D U-Net	97.48	97.10	97.35	97.22	97.50
CSFA-3D UNet	98.40	98.02	98.21	98.11	98.35
CSFA-3D UNet with ASPP	98.85	98.51	98.75	98.63	98.87
CSFA-3D UNet with ASPP + DSFS	99.00	98.70	98.95	98.82	99.03
CSFA-3D UNet with ASPP + DSFS + STN	99.12	98.80	99.08	99.01	99.14
Proposed Framework	99.15	98.82	99.11	99.04	99.17

Table 11. Comparative analysis of performance metrics across model configurations (ablation study).

the improved U-Net, show progressive improvements, with Attention U-Net reaching 96.92% in recall but still trailing behind in other areas. DeepLabV3+ exhibits robust performance, especially in mIoU and precision, yet it remains approximately 5% below the proposed method's DSC. The results derived from a five-fold cross-validation reveal exceptional consistency and reliability in the classification capability of the proposed model. Each fold demonstrates tightly clustered performance metrics, with accuracy ranging narrowly between 99.10% and 99.19%, and a mean value stabilizing at 99.15%. More notably, the recall remains robust across folds, averaging at 99.11%, which confirms the model's capacity to correctly identify true positives across varying data splits. Additionally, precision and F1-score values reflect this balance, both exhibiting minimal variance, indicating strong generalizability. The specificity, which evaluates true negative performance, maintains a mean of 99.17%, suggesting the model's ability to effectively reject irrelevant or negative cases. These results, presented in Table 10, clearly highlight the robustness of the classification pipeline.

A progressive evaluation of model configurations underscores the cumulative benefit of each architectural enhancement introduced throughout the framework. Starting with the baseline 3D U-Net, an initial accuracy of 97.48% is observed. The integration of the Channel-Spatial Feature Attention (CSFA) mechanism significantly elevates the performance, boosting the accuracy to 98.40%. When augmented with Atrous Spatial Pyramid Pooling (ASPP), the model achieves a further improvement, reaching 98.85% in accuracy and demonstrating enhanced contextual awareness. The addition of the Desert Scorpion Feature Selector (DSFS) refines feature quality by eliminating redundancy, pushing accuracy to 99.00%. Incorporating Spiking Transformer Networks (STN) adds temporal-spatial modeling, leading to an even higher score of 99.12%. The final proposed model, which combines all these elements, yields the highest performance with an accuracy of 99.15%, alongside corresponding improvements in precision, recall, F1 score, and specificity. These results, consolidated in Table 11, clearly show that each added component plays a vital role in enhancing model performance and together form a highly effective diagnostic tool for knee analysis. A critical comparison with existing research reveals that the proposed framework not only aligns with but also surpasses contemporary state-of-the-art approaches in both segmentation and classification performance. Previous works, such as those by Liu et al. and Yeoh et al., have reported Dice Similarity Coefficients (DSC) of 87.27% and 91.5%, respectively, while others like Aibinder et al. and Kulseng et al. achieved higher segmentation scores nearing 98.05% and 96.7%. However, the proposed model attains a DSC of 98.10%, which stands at the forefront among all referenced studies. In terms of classification accuracy, models by Mahum et al. and Tariq et al. have recorded impressive scores of 99.14% and

93.00%, respectively. Yet, the current framework slightly exceeds even the highest previous benchmark with a peak accuracy of 99.15%, thereby demonstrating its superior generalization capability across both imaging and diagnostic tasks.

The baseline 3D U-Net focuses only on spatial encoding, lacking any mechanism to highlight clinically relevant structures, which limits its ability to capture subtle anatomical variations in knee MRI. The CSFA-3D U-Net improves feature discrimination through channel and spatial attention but still operates on a single receptive scale, resulting in incomplete contextual understanding. Incorporating ASPP enhances multi-scale feature capture, yet without feature optimization, redundant and less informative descriptors persist. The addition of DSFS refines feature space by selecting only the most discriminative attributes, improving performance but still relying on conventional static classifiers. The inclusion of the Spiking Transformer Network (STN) introduces temporal sensitivity and relational awareness between spatial regions, leading to further improvement. However, the proposed framework achieves optimal performance because it unifies all these components: CSFA, ASPP, DSFS, and STN under a coherent architecture fine-tuned by Falcon Hunting Optimization (FHO). This combination enables precise segmentation, efficient feature selection, and context-aware classification within a single end-to-end pipeline, eliminating redundancy and maximizing generalization. Consequently, the proposed system demonstrates the highest accuracy, precision, and robustness compared to other model configurations.

The final model includes an additional component Falcon Hunting Optimization (FHO) which adaptively fine-tunes key Spiking Transformer hyperparameters. This tuning provides more stable convergence and a subtle accuracy gain by reducing variance and preventing suboptimal local minima. In conventional multi-task learning frameworks, performance degradation often arises due to gradient interference between segmentation and classification objectives, leading to suboptimal task optimization. The proposed framework effectively mitigates this issue through the integration of the Desert Scorpion Feature Selector (DSFS) and the Spiking Transformer Network (STN), which decouple redundant representations while maintaining semantic consistency between tasks. Furthermore, while existing MTL models are predominantly designed for single-modality MRI data, our framework extends to cross-modal learning by jointly utilizing MRI-based segmentation and X-ray-based OA grading, thereby achieving enhanced generalization across imaging modalities and better reflecting real-world clinical workflows. In addition, traditional MTL approaches rely on fixed learning schedules that often hinder convergence stability. In contrast, the proposed Falcon Hunting Optimization (FHO) dynamically adapts key hyperparameters during training, resulting in faster convergence (61 epochs compared to 85–100 epochs for baseline MTL models) and improved optimization efficiency. Collectively, these design choices enable the proposed framework to achieve superior accuracy, stability, and computational efficiency compared with existing multi-task learning methods.

Table 12 presents the comparative performance of different models, where the proposed framework achieves the highest accuracy across all metrics. While conventional architectures such as U-Net and V-Net perform reasonably well, their segmentation precision remains lower than that of multi-task learning models. Enhanced variants like RES_MTL_C and OA_MTL_C show further gains from channel and spatial feature integration. The proposed model surpasses all baselines with a Dice score of 98.10%, IoU of 96.26%, classification accuracy of 99.15%, and F1 score of 99.04%, demonstrating its robustness and reliability in accurately segmenting and classifying knee osteoarthritis.

Table 13 summarises the comparative evaluation of segmentation and classification performance against state-of-the-art methods. The proposed framework demonstrates substantial improvement, achieving 99.15% accuracy and a Dice similarity coefficient of 98.10%. While earlier models by Yeoh et al. and Liu et al. reported accuracies below 83%, later works such as Goswami et al. and Mahum et al. offered notable advancements. Nevertheless, the proposed approach consistently outperforms existing methods, combining high segmentation precision with superior classification accuracy, thereby establishing its effectiveness for automated knee osteoarthritis analysis.

Although the proposed system does not require simultaneous acquisition of MRI and X-ray data for every patient, it employs two imaging modalities serving complementary but independent roles, which may limit applicability in strictly single-modality clinical settings. The X-ray-based Kellgren–Lawrence grading pipeline operates exclusively on knee radiographs, with feature representations extracted from preprocessed X-ray images, refined using the Desert Scorpion Feature Selector, and classified by the Spiking Transformer Network. The MRI and X-ray datasets were obtained from independent cohorts curated for modality-specific analysis;

Model	Dice (%)	IoU (%)	Classification accuracy (%)	F1 score (%)
OA_MTL	96.08	94.01	97.85	97.70
RES_MTL	96.82	94.90	98.02	97.82
OA_MTL_C	97.65	95.80	98.45	98.20
OA_MTL_S	97.10	95.25	98.18	97.95
RES_MTL_C	97.82	95.95	98.56	98.33
RES_MTL_S	97.35	95.43	98.32	98.05
U-Net	95.45	92.85	97.25	97.15
V-Net	95.92	93.46	97.52	97.32
Proposed framework	98.10	96.26	99.15	99.04

Table 12. Comparative experimental results (MTL models).

Authors	Accuracy (%)	DSC (%)
Zheng et al. ⁷	---	78.74
Yeoh et al. ¹¹	82.5	91.5
Liu et al. ¹⁴	80.43	87.27
Goswami et al. ¹⁶	91.03	---
Ren et al. ¹⁸	84.88	---
Mahum et al. ²⁰	99.14	---
Aibinder et al. ²¹	---	98.05
Zhang et al. ²³	---	94.37
Kaur et al. ²⁵	90.43	---
Tariq et al. ²⁶	93.00	---
Kulseng et al. ²⁷	---	96.7
Proposed	99.15	98.10

Table 13. Comparative evaluation of segmentation and classification performance with existing state-of-the-art methods.

while this enables robust learning of imaging characteristics within each modality, a unified multi-center dataset containing paired MRI and X-ray data would support more comprehensive population-level validation and represents an important direction for future work. In addition, the dual-path architecture introduces additional computational demands compared to single-model multi-task learning approaches; therefore, future work will explore lightweight architectural designs and model optimization strategies to further reduce inference time and improve deployment efficiency.

Conclusion

The proposed deep learning framework establishes a robust and unified pipeline for the automated analysis of 3D knee MRI scans, addressing both segmentation and classification with a high degree of accuracy and clinical reliability. This system cohesively integrates multiple specialized modules—beginning with Gaussian Guided Filtering for denoising and edge enhancement, followed by 3D CSFA-UNet equipped with attention mechanisms to capture intricate spatial dependencies. The inclusion of the Atrous Spatial Pyramid Pooling (ASPP) module allows the model to recognize anatomical features at multiple scales, while the Desert Scorpion Feature Selector (DSFS) refines the high-dimensional output into a compact, informative feature space. These selected features are then passed through a Spiking Transformer Network, which captures both spatial topology and temporal encoding, further strengthened by Falcon Hunting Optimization (FHO) for optimal hyperparameter tuning. The effectiveness of this comprehensive approach is reflected in its outstanding performance: a classification accuracy of 99.15% and a segmentation Dice Similarity Coefficient (DSC) of 98.10%, along with consistent excellence across other metrics such as Intersection over Union (IoU), Average Surface Distance (ASD), 95th percentile Hausdorff Distance (Hd95), precision, recall, and specificity. These outcomes affirm the model's suitability for high-stakes medical tasks like surgical planning and osteoarthritis grading. Looking ahead, this framework could be extended to other anatomical regions or modalities such as CT or ultrasound. Furthermore, embedding explainable AI strategies would enhance clinician trust and transparency, fostering broader acceptance in clinical workflows.

Data availability

The datasets analyzed in this study are publicly accessible. The segmentation dataset is available at https://gitlab.com/vvr/OActive/osteoarthritis_initiative_zib_dataset, and the X-ray image dataset can be accessed at <https://www.kaggle.com/datasets/tommyngx/digital-knee-xray>. No special permissions were required to obtain these datasets, and their use complied with the terms and licenses specified by the respective providers.

Received: 16 August 2025; Accepted: 27 January 2026

Published online: 01 February 2026

References

- Kim-Wang, S. Y. et al. Auto-segmentation of the tibia and femur from knee MR images via deep learning and its application to cartilage strain and recovery. *J. Biomech.* **149**, 111473 (2023).
- Liu, H., Sun, Y. & Cheng, X. Prior-based 3D U-Net: A model for knee-cartilage segmentation in MRI images. *Computers Graphics.* **115**, 167–180 (2023).
- Dung, N. et al. End-to-end deep learning model for segmentation and severity staging of anterior cruciate ligament injuries from MRI. *Diagn. Interv. Imaging.* **104** (3), 133–141 (2023).
- Vinayahalingam, S. et al. Deep learning for automated segmentation of the temporomandibular joint. *J. Dent.* **132**, 104475 (2023).
- Zhou, Y. et al. Multi-scale channel attention U-Net: A novel framework for automated gallbladder segmentation in medical imaging. *Front. Oncol.* **15**, 1528654 (2025).
- Jang, S. J. et al. Standardized fixation zones and cone assessments for revision total knee arthroplasty using deep learning. *J. Arthroplast.* **38** (6), S259–S265 (2023).
- Zheng, J. et al. Radiological segmentation of knee meniscus ultrasound images based on boundary constraints and multi-scale fusion network. *J. Radiation Res. Appl. Sci.* **17** (3), 101037 (2024).

8. Li, S., Zhao, S., Zhang, Y., Hong, J. & Chen, W. Source-free unsupervised adaptive segmentation for knee joint MRI. *Biomed. Signal Process. Control.* **92**, 106028 (2024).
9. Wang, X. Patch attention U-Net for knee cartilage segmentation in magnetic resonance images. *Biomed. Signal Process. Control.* **106**, 107754 (2025).
10. du Toit, C. et al. Deep learning for synovial volume segmentation of the first carpometacarpal joint in osteoarthritis patients. *Osteoarthr. Imaging.* **4** (1), 100176 (2024).
11. Yeoh, P. S. et al. 3D efficient multi-task neural network for knee osteoarthritis diagnosis using MRI scans: Data from the osteoarthritis initiative. *IEEE Access.* **11**, 135323–135333 (2023).
12. Zhao, H. et al. The value of deep learning-based X-ray techniques in detecting and classifying KL grades of knee osteoarthritis: A systematic review and meta-analysis. *Eur. Radiol.* **35** (1), 327–340 (2025).
13. Yeoh, P. S. Q., Goh, S. L., Hasikin, K., Wu, X. & Lai, K. W. 3D efficient multi-task neural network for knee osteoarthritis diagnosis using MRI scans: Data From the osteoarthritis initiative. In *IEEE Access*. Vol. 11. 135323–135333. <https://doi.org/10.1109/ACCESS.2023.3338379> (2023).
14. Liu, M. et al. AutoDDH: A dual-attention multi-task network for grading developmental dysplasia of the hip in ultrasound images. *Vis. Comput.* 1–13 (2025).
15. Phan, T. et al. DIKApp: An AI-based diagnostic system for knee osteoarthritis. *J. Imaging Inf. Med.* 1–16 (2025).
16. Goswami, A. D. Automatic classification of the severity of knee osteoarthritis using enhanced image sharpening and CNN. *Appl. Sci.* **13** (3), 1658 (2023).
17. Chadoulos, C. et al. Dense multi-scale graph convolutional network for knee joint cartilage segmentation. *Bioengineering* **11**(3), 278 (2024).
18. Ren, X. et al. OA-MEN: A fusion deep learning approach for enhanced accuracy in knee osteoarthritis detection and classification using X-ray imaging. *Front. Bioeng. Biotechnol.* **12**, 1437188 (2025).
19. Aladhadh, S. & Mahum, R. Knee osteoarthritis detection using an improved CenterNet with pixel-wise voting scheme. *IEEE Access.* **11**, 22283–22296 (2023).
20. Cigdem, O. et al. Estimation of time-to-total knee replacement surgery with multimodal modeling and artificial intelligence. *Comput. Biol. Med.* **193**, 110364 (2025).
21. Aibinder, D., Weisberg, M., Ghidotti, A. & Miri Weiss, C. Enhanced attention Res-UNet for segmentation of knee bones. *Mathematics* **12**, 14 : 2284. (2024).
22. Fatema, K. et al. Development of an automated optimal distance feature-based decision system for diagnosing knee osteoarthritis using segmented X-ray images. *Heliyon* **9**, 11 (2023).
23. Zhang, H. et al. MPFCNet: Multi-scale parallel feature fusion convolutional network for 3D knee segmentation from MR images. *Pattern Anal. Appl.* **28** (2), 62 (2025).
24. Kakavand, R. et al. Swin UNETR segmentation with automated geometry filtering for biomechanical modeling of knee joint cartilage. *Ann. Biomed. Eng.* 1–15 (2025).
25. Kaur, P. S., Kohli, G. S., Bedi, J. & Wasly, S. A novel deep learning approach for automated grading of knee osteoarthritis severity. *Multimed. Tools Appl.* 1–20 (2024).
26. Tariq, T., Suhail, Z. & Nawaz, Z. Ordinal classification for knee osteoarthritis x-rays using vision transformers. *Multimed. Tools Appl.* 1–28 (2025).
27. Kulseng, C. P. et al. Automatic segmentation of human knee anatomy by a convolutional neural network applying a 3D MRI protocol. *BMC Musculoskelet. Disord.* **24** (1), 41 (2023).
28. Pan, J. et al. Automatic knee osteoarthritis severity grading based on X-ray images using a hierarchical classification method. *Arthritis Res. Therapy.* **26** (1), 203 (2024).
29. Rajamohan, H. et al. Prediction of total knee replacement using deep learning analysis of knee MRI. *Sci. Rep.* **13**(1), 6922 (2023).
30. Kakavand, R. et al. Integration of Swin UNETR and statistical shape modeling for a semi-automated segmentation of the knee and biomechanical modeling of articular cartilage. *Sci. Rep.* **14** (1), 2748 (2024).
31. Yoon, K. et al. Multi-class segmentation of temporomandibular joint using ensemble deep learning. *Sci. Rep.* **14**(1), 18990 (2024).
32. Marsilio, L. et al. and. Combined edge loss UNet for optimized segmentation in total knee arthroplasty preoperative planning. *Bioengineering* **10**(12), 1433 (2023).
33. Kono, K. et al. Artificial intelligence-based analysis of lower limb muscle mass and fatty degeneration in patients with knee osteoarthritis and its correlation with knee society score. *Int. J. Comput. Assist. Radiol. Surg.* 1–8 (2024).
34. https://gitlab.com/vvr/OActive/osteoarthritis_initiative_zib_dataset
35. <https://www.kaggle.com/datasets/tommyngx/digital-knee-xray>
36. Yang, J., Soltan, A. A. S. & Clifton, D. A. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *Npj Digit. Med.* **5**, 69 (2022).
37. Liu, A. & Wang, X. Comparative analysis of X-ray and MRI in a nomogram-based approach for rheumatoid arthritis and lupus management. *J. Radiation Res. Appl. Sci.* **18** (3), 101705 (2025).

Author contributions

M.C. and S.A. conceived the study. S.A. and M.C. developed methodology and performed analysis with support from S.R. and G.V. S.A. and G.V. implemented software and visualizations. M.C. drafted the manuscript. All authors reviewed and approved the final version.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.M.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2026