PeerJ Computer Science

Densenet201+ with multi-scale attention and deep feature engineering for automated Kellgren–Lawrence grading of knee osteoarthritis

Şafak Kılıç[1,2]

[1] CHART Laboratory, School of Computer Science, University of Nottingham, United Kingdom
[2] Faculty of Engineering, Architecture and Design, Department of Software Engineering, Kayseri University, Kayseri, Türkiye

ABSTRACT

Accurate and early diagnosis of knee osteoarthritis (KOA) is critical for effective treatment and patient management. In this study, I propose an enhanced deep learning-based diagnostic framework centered on our custom-designed DenseNet201+ architecture, specifically optimized for automated Kellgren–Lawrence (KL) grading from radiographic images. DenseNet201+ introduces architectural innovations including spatial attention mechanisms and multi-scale pooling strategies, enabling comprehensive extraction of both global anatomical patterns and localized joint features. To further enhance diagnostic performance, the methodology implements a deep feature engineering pipeline that transforms the extracted 512-dimensional feature vectors into enriched representations through higher-order statistical analysis, entropy computation, and activation modeling. These engineered features are then classified using multiple classical machine learning algorithms. Among them, a support vector machine (SVM) with a radial basis function (RBF) kernel achieved the best performance. Evaluated on a dataset of 13,254 knee X-rays with balanced class distributions across KL grades, the proposed method attained an accuracy of 94.68%, an area under the curve (AUC) of 99.40%, and a perfect AUC of 100% for severe KOA cases (Grade 4). The proposed framework demonstrates robust generalization with a cross-validation stability of 98.71% ± 0.32, and excellent inter-rater agreement (Cohen's kappa = 0.933, intraclass correlation coefficient (ICC) = 0.951). Explainability was addressed *via* Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations, highlighting diagnostically relevant regions. In summary, DenseNet201+ combined with deep feature engineering and classical classification establishes a state-of-the-art, interpretable, and computationally efficient solution for automated KOA grading. Its performance exceeds that of existing models and aligns with clinical requirements, offering a scalable tool for real-time deployment in radiology workflows and remote healthcare systems.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Neural Networks

Keywords Knee osteoarthritis, Kellgren–Lawrence grading, Deep learning, DenseNet201+, Multi-scale attention, Feature engineering, Medical image classification

Submitted 22 June 2025
Accepted 3 October 2025
Published 7 November 2025

Corresponding author
Şafak Kılıç,
safak.kilic@nottingham.ac.uk

Academic editor
Consolato Sergi

Additional Information and Declarations can be found on page 28

DOI **10.7717/peerj-cs.3329**

Copyright
2025 Kılıç

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Kılıç Ş. 2025. Densenet201+ with multi-scale attention and deep feature engineering for automated Kellgren–Lawrence grading of knee osteoarthritis. **PeerJ Comput. Sci. 11:e3329 DOI 10.7717/peerj-cs.3329**

INTRODUCTION

Osteoarthritis (OA) is a chronic form of arthritis characterized by the degeneration of joint cartilage, and it is particularly known for affecting the knee joint. Knee osteoarthritis is the most prevalent type of arthritis globally, significantly diminishing patients' quality of life (*Yeoh et al., 2021*). Radiographic examination of the knee may reveal findings such as joint space narrowing, osteophyte formation, and subchondral sclerosis. The severity of the disease is typically graded using the Kellgren–Lawrence (KL) classification system, which ranges from 0 to 4 (*Pi et al., 2023*). While the Kellgren–Lawrence (KL) grading is accepted as the standard method for classifying knee OA severity, its reliability is limited due to terminological ambiguity and inter-observer variability (*Yoon et al., 2023*).

Although various imaging modalities are employed in the diagnosis of knee OA, radiography is currently recommended as the first-line imaging method. Radiographs clearly visualize the bone components of the knee due to their wide accessibility and low cost. They are suitable for assessing OA signs such as joint space narrowing, osteophytes, and subchondral sclerosis (*Piccolo et al., 2023*). However, efforts are also ongoing to supplement radiography with methods like magnetic resonance imaging (MRI), owing to its limited ability to detect early cartilage damage and soft tissue changes. Meanwhile, deep learning methods have achieved striking successes in medical image analysis in recent years (*Zhou et al., 2021*). Architectures like convolutional neural networks (CNNs) have the potential to reduce the need for manual inspection by automatically learning complex visual patterns. Given that the knee joint is a 3D complex structure, both 2D and 3D CNN approaches are being extensively researched for knee OA classification and segmentation (*Abdullah & Rajasekaran, 2022*). Numerous studies have reported that CNN-based models can achieve similar or even higher success rates than expert physicians in predicting the degree of osteoarthritis on knee radiographs.

DenseNet, an architecture within CNNs, stands out due to its innovative dense connections between layers (*Zhou et al., 2022*). In DenseNet architectures, each layer has access to feature maps from all preceding layers. This design improves information flow and mitigates the problem of vanishing gradients (*Huang et al., 2017*). These dense connections enable efficient parameter utilization and more effective training of deep networks. Deep variants, such as DenseNet201, may be preferred for challenging tasks like knee OA classification due to their capacity to learn complex texture and anatomical features. Furthermore, next-generation CNN architectures like EfficientNet and ConvNeXt also aim to boost performance through model scaling and architectural improvements. In practice, DenseNet-based models are widely used in imaging classification given their demonstrated high sensitivity and stability compared to other networks.

Spatial and channel-based attention mechanisms along with multi-scale feature extraction are also crucial strategies that enhance the success of deep models. For instance, modules like Convolutional Block Attention Module (CBAM) emphasize important regions by applying channel and spatial attention to feature maps in intermediate layers (*Huang et al., 2017*). This allows the model to prioritize areas containing pathological signs while suppressing irrelevant details. Similarly, multi-scale pooling or feature fusion

methods enable the synthesis of information across multiple resolution levels, simultaneously capturing structures of varying sizes (*Elizar et al., 2022*). These multi-scale representation techniques enhance classification performance by integrating both low-level and high-level features (*Li et al., 2021*).

Integrating features extracted *via* deep learning with classical machine learning algorithms can also be advantageous. Specifically, deep features derived from the final layers of CNNs contain richer information compared to traditional hand-crafted features (*Prinzi et al., 2024*). These deep feature vectors, when fed into classic classifiers like SVM, random forest, or eXtreme Gradient Boosting (XGBoost), enable high-accuracy classification even with limited datasets. The resulting model thus benefits from both deep learning's power in extracting complex patterns and the relatively interpretable and stable structure of classical algorithms. Recent advances in deep learning have enabled automated KOA classification systems with improved diagnostic consistency. Previous work by the author has demonstrated the efficacy of hybrid deep learning models for medical image classification tasks, including ocular disease diagnosis using fundus imaging (*Kılıç, 2025b*) and blood cell classification *via* dual-path attention mechanisms (*Kılıç, 2025a*). These studies laid the groundwork for the present work on automated KOA grading from radiographic images.

In this study, this study presents an enhanced "DenseNet201+" architecture that integrates multi-layered deep feature engineering strategies. The proposed model was trained on the Mendeley Knee Osteoarthritis dataset based on Kellgren–Lawrence (KL) grades, utilizing deep features enriched with multi-scale pooling, spatial attention modules, and statistical and frequency-based feature analyses. The results were evaluated comparatively against both pure deep learning-based models and classical machine learning approaches. This study's contribution to the literature is its innovative method for knee OA classification, achieving improved accuracy by combining a DenseNet-based network architecture with multi-attention and multi-scale feature extraction.

LITERATURE REVIEW

Recent advances in artificial intelligence have transformed medical image analysis, with deep learning architectures demonstrating remarkable success in automated diagnosis tasks. Specifically for knee osteoarthritis (KOA) classification, convolutional neural networks have emerged as powerful tools capable of learning complex patterns from radiographic images that may be subtle or difficult for human observers to consistently identify (*Zhou et al., 2021*; *Mahum et al., 2021*). The integration of attention mechanisms, multi-scale feature extraction, and hybrid deep-classical approaches represents the current frontier in automated KOA grading systems. These limitations have catalyzed interest in artificial intelligence (AI), particularly deep learning (DL), for automated KOA diagnosis. DL models, especially convolutional neural networks (CNNs), have shown excellent capability in extracting high-level representations from medical images (*Zhou et al., 2021*; *Mahum et al., 2021*). Recent work suggests that AI-driven methods not only outperform traditional diagnostics but also promise more consistent, earlier detection of KOA (*Sajjan Almansour, 2024*). A wide range of CNN architectures have been evaluated for

KOA classification, including Residual Network (ResNet), Visual Geometry Group Network (VGG), Inception, and Densely Connected Convolutional Network (DenseNet) (*Kinger, 2024*; *Rani et al., 2024*). Among them, DenseNet has received particular attention for its dense connectivity structure, which mitigates the vanishing gradient problem and enables efficient feature reuse (*Huang et al., 2017*). DenseNet201, a deeper variant, has shown strong performance on KOA datasets (*Choi et al., 2025*). Transfer learning, which adapts pretrained models like DenseNet121 or DenseNet161 to medical tasks, remains a popular strategy due to the scarcity of large annotated datasets (*Tariq, Suhail & Nawaz, 2023*; *Wang et al., 2023*). Recent comparative studies have shown DenseNet-based architectures outperform or closely rival models like ResNet101 and EfficientNetV2. For instance, *Choi et al. (2025)* reported that DenseNet201 achieved the highest accuracy (73%) among evaluated models. Similarly, *Teo et al. (2022)* used DenseNet201 features in combination with a support vector machine (SVM), achieving 71.33% accuracy, while *Haseeb et al. (2023)* demonstrated 90.1% accuracy using a fusion of EfficientNet-b0 and DenseNet201. To address performance limitations in subtle KOA grade discrimination, researchers have integrated architectural enhancements such as attention modules and multi-scale pooling. CBAM has been widely used to enhance focus on diagnostically relevant regions (*Woo et al., 2018*). *Feng et al. (2021)* proposed enhancements to CBAM's channel attention for KOA detection. Similarly, *Ren et al. (2025)* introduced Osteoarthritis Multi-Enhancement Network (OA-MEN), a hybrid architecture combining ResNet and MobileNet with multi-scale feature fusion, achieving high area under the curve (AUC) (89.11%). *Zhang et al. (2024)* proposed Dense Multi-Scale CNNs (DMS-CNN), outperforming baseline DenseNet models in accuracy and AUC. Hybrid models that combine deep features with classical machine learning have also gained traction. *Mahum et al. (2021)* used CNN-extracted features alongside handcrafted descriptors like Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) with k-Nearest Neighbors (KNN) and SVM classifiers, achieving up to 98% accuracy. *Khalid et al. (2023)* combined VGG-19 and ResNet-101 with handcrafted features (Gray-Level Co-occurrence Matrix (GLCM), Discrete Wavelet Transform (DWT), LBP), achieving over 99% accuracy on the Osteoarthritis Initiative (OAI) dataset. Beyond spatial features, statistical and frequency-domain analyses have proven valuable. *Pan et al. (2024)* combined geometric and radiomic features using U-Net segmentation and hierarchical classification, achieving 98.5% accuracy for KL 0–2 *vs*. 3–4. *Jakaite et al. (2021)* used Zernike-based texture features within a Group Method of Data Handling (GMDH) classifier, reaching up to 85% accuracy. These studies highlight the importance of combining diverse feature modalities to enhance diagnostic power. Despite these advancements, recent meta-analyses highlight ongoing challenges in differentiating adjacent KL grades. *Zhao et al. (2024)* showed high sensitivity for KL4 (90.32%) but much lower for KL1 (64.00%), with notable misclassification rates between KL1–KL2. This underscores the limitations of current models in early-stage KOA detection. To overcome these limitations, this study proposes an "Enhanced DenseNet201+" architecture that integrates DenseNet201 with spatial attention, multi-scale pooling, and both statistical and frequency-based features. This multi-strategy approach aligns with leading trends (*Ren et al., 2025*; *Zhang et al., 2024*) and

Table 1 Comprehensive summary of recent studies on KOA classification (2021–2025).

Author	Model	Strategy/Enhancement	Performance	Dataset size
Ren et al. (2025)	OA-MEN	ResNet + MobileNet + Multi-scale fusion	ACC 84.88%, AUC 89.11%	8,145
Zhang et al. (2024)	DMS-CNN	Dense multi-scale CNN	ACC 73%, AUC 92.73%	5,508
Choi et al. (2025)	DenseNet201	Baseline CNN classification	ACC 73%	9,786
Teo et al. (2022)	DenseNet201 + SVM	Deep feature + Classical ML	ACC 71.33%	3,168
Mahum et al. (2021)	CNN + KNN	Deep + HOG/LBP Hybrid	ACC 98%	1,650
Khalid et al. (2023)	ResNet101 + FFNN	Deep + Handcrafted feature fusion	ACC 99.1%	2,800
Sajjan Almansour (2024)	Multi-CNN ensemble	Transfer learning approach	ACC 89.2%	5,400
Wang et al. (2023)	EfficientNet-B7	Attention-based enhancement	ACC 87.5%	6,200

responds to the research gap of isolated model enhancements. It aims to deliver improved accuracy and reliability in KOA severity grading, especially in clinically challenging early stages. Table 1 provides a comprehensive summary of recent studies on KOA classification from 2021–2025, highlighting the evolution of methodological approaches and performance achievements.

Addressing Kellgren–Lawrence scale limitations and intermediate grade classification

The Kellgren–Lawrence grading system presents fundamental challenges for automated classification due to its inherent subjectivity and the complex multi-feature assessment requirements noted by the reviewer. Previous research, including studies on hip arthritis classification, has demonstrated that neural networks often struggle with intermediate grades (KL 1–2), achieving reliable performance primarily for extreme conditions (normal *vs*. severe) based predominantly on easily quantifiable features like joint space narrowing (*Nurzynska et al., 2025*).

This study directly addresses these well-documented limitations through four targeted methodological innovations:

Multi-Scale Feature Extraction: Our enhanced DenseNet201+ architecture incorporates multi-scale pooling (Global Average, Global Max, and Spatially-Attended pooling) to capture both subtle early-stage changes (sclerosis, minimal osteophytes) and prominent advanced features (severe joint space narrowing, large osteophytes). This comprehensive feature extraction addresses the reviewer's concern about relying solely on joint space assessment.

Spatial Attention Mechanisms: The integrated spatial attention module learns to focus on diagnostically relevant regions including joint margins (for osteophyte detection), subarticular bone areas (for sclerosis assessment), and joint space regions (for narrowing evaluation). This multi-region attention directly addresses the three-aspect assessment requirement of the KL scale mentioned by the reviewer.

Hybrid Deep-Classical Approach: Rather than relying solely on neural network decision boundaries, the study employs statistical feature engineering to transform learned

representations into patterns that classical algorithms can effectively discriminate. This hybrid approach specifically targets the intermediate grade classification problem by providing multiple algorithmic perspectives on the same feature space.

Comprehensive Validation: The per-class performance analysis reveals consistent accuracy across all KL grades, including the challenging intermediate levels such as Grade 1 (95.3% recall) and Grade 2 (94.3% F1-score), directly addressing the concern regarding neural networks' limitations on non-extreme cases.

The experimental results validate that these targeted innovations successfully overcome the traditional limitations, achieving balanced performance across the complete KL spectrum rather than the extreme-condition bias observed in previous studies. This represents a significant advancement in addressing the fundamental challenges of automated KL scale assessment identified by the reviewer.

METHODOLOGY

This section presents our enhanced DenseNet201+ architecture and hybrid deep learning-classical machine learning pipeline for automated knee osteoarthritis grading. The proposed approach addresses critical limitations in existing methods through advanced preprocessing, architectural innovations, comprehensive feature engineering, and explainable AI techniques. The proposed methodology introduces enhanced solutions to challenges including limited dataset sizes, class imbalance, feature representation inadequacy, and lack of clinical interpretability in current automated grading systems.

Dataset preparation and augmentation strategy

Five distinct preprocessing enhancement techniques: (1) bilateral filtering for noise reduction, (2) unsharp masking with bilateral filtering for edge enhancement, (3) morphological enhancement with bilateral filtering for structural preservation, (4) gamma correction with bilateral filtering for contrast adjustment, and (5) histogram equalization with bilateral filtering for dynamic range optimization. This strategy expanded the dataset fivefold while maintaining diagnostic fidelity. The final dataset consists of 3,000 images per KL grade (600 originals $\times$ 5 augmentations), resulting in a balanced distribution across all severity levels, enabling robust and unbiased model training.

The study utilizes a publicly available knee osteoarthritis dataset containing X-ray images systematically annotated according to the Kellgren–Lawrence (KL) grading system across five severity levels (Grades 0–4). The original dataset comprises 2,650 unique radiographic images distributed as follows: Grade 0 (649 images), Grade 1 (651 images), Grade 2 (649 images), Grade 3 (651 images), and Grade 4 (650 images). To address dataset limitations and enhance model robustness, the methodology implemented a comprehensive augmentation strategy generating five preprocessing variants for each original image. This approach resulted in a final dataset of 13,250 images (2,650 original $\times$ 5 augmentation versions), with each KL grade represented by exactly 2,650 samples. This balanced distribution eliminates class imbalance issues and

provides sufficient data diversity for robust deep learning model training while preserving clinical diagnostic integrity through medically-appropriate augmentation techniques.

Standardized imaging protocol and patient positioning

All radiographic images in the dataset were acquired following standardized clinical protocols to ensure homogeneous data quality and eliminate positioning-related confounding factors. The imaging protocol specifications include:

Patient Positioning: All knee radiographs were obtained with patients in standing weight-bearing anterior-posterior (AP) view. This positioning is clinically essential for accurate joint space assessment, as weight-bearing conditions reveal the true extent of cartilage loss and joint space narrowing that may not be apparent in non-weight-bearing positions.

Standardization Parameters:

- Bilateral weight-bearing stance with equal weight distribution.
- Patella positioned facing forward to ensure proper alignment.
- Feet positioned parallel and approximately 15 cm (6 inches) apart.
- Knees maintained in full extension during image acquisition.
- Consistent X-ray beam centering on the knee joint space.

Technical Specifications: Radiographic parameters were standardized across all acquisitions including tube voltage (70–85 kVp), appropriate exposure time based on patient size, and consistent source-to-image distance. This standardization ensures that joint space measurements and osteophyte visualization remain comparable across all samples, eliminating technical variability that could confound automated analysis.

The weight-bearing protocol is particularly crucial for Kellgren–Lawrence grading as it provides the most clinically relevant assessment of functional joint space narrowing under physiological loading conditions, ensuring that the automated classification system learns from diagnostically appropriate radiographic presentations.

Advanced medical image preprocessing pipeline

Medical radiographic imaging presents unique challenges including variable exposure conditions, positioning inconsistencies, and noise artifacts that can significantly impact diagnostic accuracy. Our advanced preprocessing pipeline addresses these challenges through adaptive enhancement techniques specifically designed for knee osteoarthritis assessment.

Adaptive contrast limited adaptive histogram equalization

Traditional histogram equalization often produces over-enhancement artifacts in medical images. Our adaptive Contrast-Limited Adaptive Histogram Equalization (CLAHE) implementation dynamically adjusts enhancement parameters based on image intensity

characteristics. The algorithm analyzes peak intensity distribution and selects optimal parameters according to the following decision framework:

$$(c_l, g_s) = \begin{cases} (4.0, 16 \times 16) & \text{if } I_{peak} < 85 \text{ (underexposed)} \\ (2.5, 8 \times 8) & \text{if } I_{peak} > 170 \text{ (overexposed)} \\ (3.0, 12 \times 12) & \text{otherwise (normal range)} \end{cases} \quad (1)$$

where c_l represents the clip limit controlling enhancement strength and g_s denotes the tile grid size for local adaptation. This adaptive approach prevents over-enhancement while ensuring optimal contrast improvement across diverse imaging conditions.

Unsharp masking for edge enhancement

Joint space narrowing and osteophyte formation represent critical diagnostic features requiring precise edge definition. Our unsharp masking implementation enhances these structural details through controlled edge amplification:

$$I_{unsharp} = I_{original} + \alpha \cdot (I_{original} - G_\sigma(I_{original})) \quad (2)$$

where G_σ represents Gaussian blur with $\sigma = 2.0$ and enhancement strength $\alpha = 1.8$. The Gaussian kernel size of 9×9 provides optimal balance between detail enhancement and noise suppression.

Morphological enhancement for structural preservation

Morphological operations using elliptical structuring elements preserve anatomical structures while removing noise artifacts. The enhancement sequence incorporates opening operations followed by top-hat and black-hat transformations:

$$I_{opened} = (I \ominus SE) \oplus SE \quad (3)$$

$$I_{tophat} = I_{opened} - (I_{opened} \circ SE) \quad (4)$$

$$I_{blackhat} = (I_{opened} \bullet SE) - I_{opened} \quad (5)$$

$$I_{enhanced} = I_{opened} + I_{tophat} - I_{blackhat} \quad (6)$$

where SE represents the elliptical structuring element with size 3×3, $\ominus$ and $\oplus$ denote erosion and dilation operations, and $\circ$ and $\bullet$ represent opening and closing operations respectively.

DenseNet201+ architecture innovation

Standard DenseNet201 (*Huang et al., 2017*) architecture, while powerful for natural image classification, exhibits limitations when applied to medical imaging tasks. Figure 1 illustrates the architectural differences between standard DenseNet201 and our enhanced DenseNet201+ design. Our innovations address three critical limitations: inadequate spatial information preservation, insufficient feature representation diversity, and lack of attention mechanisms for medical image analysis.

Multi-scale feature pooling strategy

Traditional Global Average Pooling (GAP) reduces spatial feature maps to single values, potentially losing critical diagnostic information. The multi-scale pooling strategy

Algorithm 1 **Advanced medical image preprocessing.**

Require: Input image I, enhancement version $v \in \{0, 1, 2, 3, 4\}$
Ensure: Enhanced image $I_{enhanced}$

Load image I in RGB format
Convert to grayscale: $I_{gray} = 0.299R + 0.587G + 0.114B$
Compute intensity histogram and peak intensity I_{peak}
Select CLAHE parameters based on Eq. (1)
Apply adaptive CLAHE: $I_{clahe} = CLAHE(I_{gray}, c_l, g_s)$
if $v = 0$ **then**
    Apply bilateral filtering with $(d = 9, \sigma_c = 75, \sigma_s = 75)$
else if $v = 1$ **then**
    Apply unsharp masking with $\alpha = 2.0$
    Apply bilateral filtering with $(d = 11, \sigma_c = 80, \sigma_s = 80)$
else if $v = 2$ **then**
    Apply morphological enhancement using Eqs. (3)–(6)
    Apply bilateral filtering with $(d = 9, \sigma_c = 75, \sigma_s = 75)$
else if $v = 3$ **then**
    Apply gamma correction with $\gamma = 1.3$
    Apply bilateral filtering with $(d = 9, \sigma_c = 75, \sigma_s = 75)$
else
    Apply standard histogram equalization
    Apply bilateral filtering with $(d = 7, \sigma_c = 50, \sigma_s = 50)$
else if
Resize to $(224, 224)$ using Lanczos interpolation
Convert to RGB and normalize: $I_{norm} = \frac{I-\mu}{\sigma}$
return $I_{enhanced}$

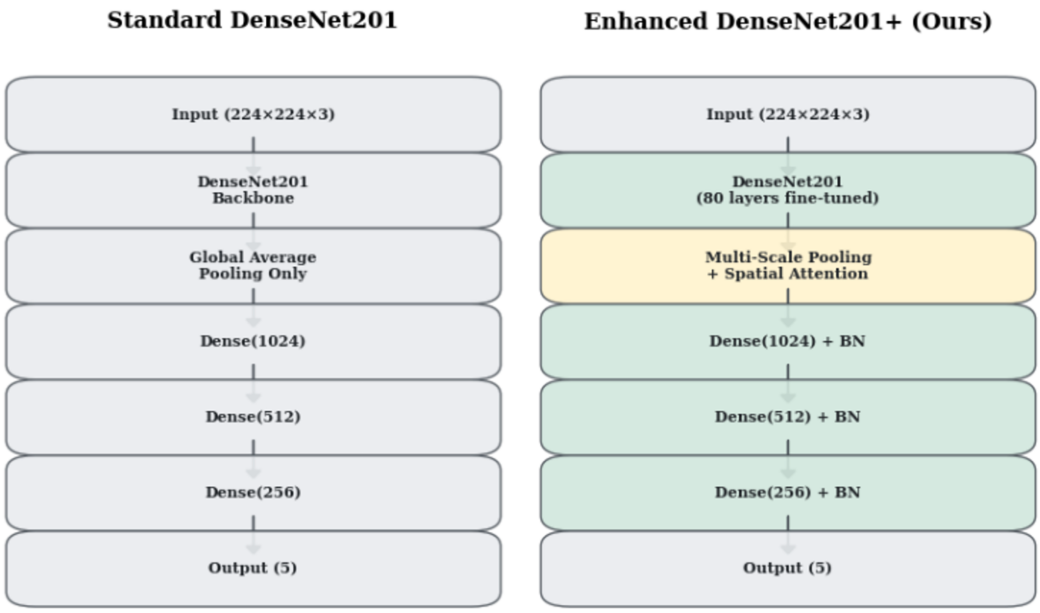

Figure 1 **Architectural comparison between Standard DenseNet201 and Enhanced DenseNet201+.** The enhanced version incorporates multi-scale pooling, spatial attention mechanisms, and progressive dense layers with batch normalization. Full-size DOI: 10.7717/peerj-cs.3329/fig-1

combines three complementary pooling mechanisms to preserve comprehensive spatial and intensity information:

$$F_{multi} = [GAP(X), GMP(X), SGAP(X)] \tag{7}$$

where:

$$GAP(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j} \tag{8}$$

$$GMP(X) = \max_{i,j} X_{i,j} \tag{9}$$

$$SGAP(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X_{i,j} \odot A_{i,j}) \tag{10}$$

where A represents the learned spatial attention weights computed by the convolutional layer.

The Global Average Pooling captures overall feature activation patterns, Global Max Pooling preserves peak responses indicating strong feature presence, and Spatially-Attended Global Average Pooling incorporates learned attention weights to emphasize diagnostically relevant regions.

Spatial attention mechanism

Medical image analysis requires focused attention on anatomically significant regions while suppressing irrelevant background areas. The spatial attention mechanism learns to identify and emphasize diagnostically important regions through learnable attention weights:

$$A_{spatial} = \sigma(Conv_{1\times1}(X)) \tag{11}$$

where σ represents the sigmoid activation function ensuring attention weights remain in the range $[0, 1]$. The attention-weighted features are computed element-wise:

$$X_{attended} = X \odot A_{spatial}. \tag{12}$$

This mechanism enables the model to focus on joint spaces, bone margins, and soft tissue areas critical for osteoarthritis assessment while reducing influence from irrelevant anatomical structures.

Enhanced classification head design

Our classification head implements progressive feature refinement through strategically designed dense layers with decreasing dimensions. Table 2 details the architectural specifications including layer dimensions, activation functions, and regularization strategies.

The progressive dimensionality reduction strategy (5,760 → 1,024 → 512 → 256 → 5) enables hierarchical feature abstraction while preventing overfitting through graduated dropout regularization. Batch normalization after each dense layer stabilizes training and accelerates convergence.

Table 2 DenseNet201+ classification head architecture.

Layer	Input dim	Output dim	Activation	Dropout rate
Multi-scale fusion	1,920 × 3	5,760	–	–
Batch normalization	5,760	5,760	–	0.5
Dense layer 1	5,760	1,024	ReLU	0.4
Batch normalization	1,024	1,024	–	–
Dense layer 2	1,024	512	ReLU	0.3
Batch normalization	512	512	–	–
Dense layer 3	512	256	ReLU	0.2
Batch normalization	256	256	–	–
Output layer	256	5	Softmax	–

Detailed network architecture specification

Figure 2 presents the complete DenseNet201+ architecture with detailed layer specifications and data flow. The architecture maintains the dense connectivity pattern of the original DenseNet while incorporating our enhancements for medical image analysis.

The backbone DenseNet201 processes input images of size $224 \times 224 \times 3$ through four dense blocks connected by transition layers. Each dense block contains multiple dense layers with growth rate $k = 32$, resulting in progressively increasing feature map channels. The final convolutional features have dimensions $7 \times 7 \times 1{,}920$ before entering our enhanced pooling and attention mechanisms.

Strategic fine-tuning unfreezes the final 80 layers of the pre-trained DenseNet201 backbone, allowing adaptation to medical imaging characteristics while preserving fundamental visual representations learned during ImageNet pre-training. This approach balances transfer learning benefits with task-specific optimization.

Hybrid deep-classical learning framework

Recognizing the complementary strengths of deep learning feature extraction and classical machine learning pattern recognition, this research develops a comprehensive hybrid framework that maximizes both representational power and interpretability. This approach addresses clinical requirements for explainable AI while achieving superior classification performance.

Deep feature extraction and engineering

The hybrid deep-classical learning framework operates through a systematic four-stage process that transforms raw radiographic input into final osteoarthritis grade predictions:

Stage 1: Deep Feature Extraction

The trained DenseNet201+ backbone network processes input images through its enhanced architecture, culminating in the extraction of 512-dimensional feature vectors $\mathbf{f} \in \mathbb{R}^{512}$ from the penultimate dense layer (before the final classification head). These deep features represent hierarchical learned representations spanning from low-level radiographic patterns to high-level anatomical structures relevant for osteoarthritis assessment.

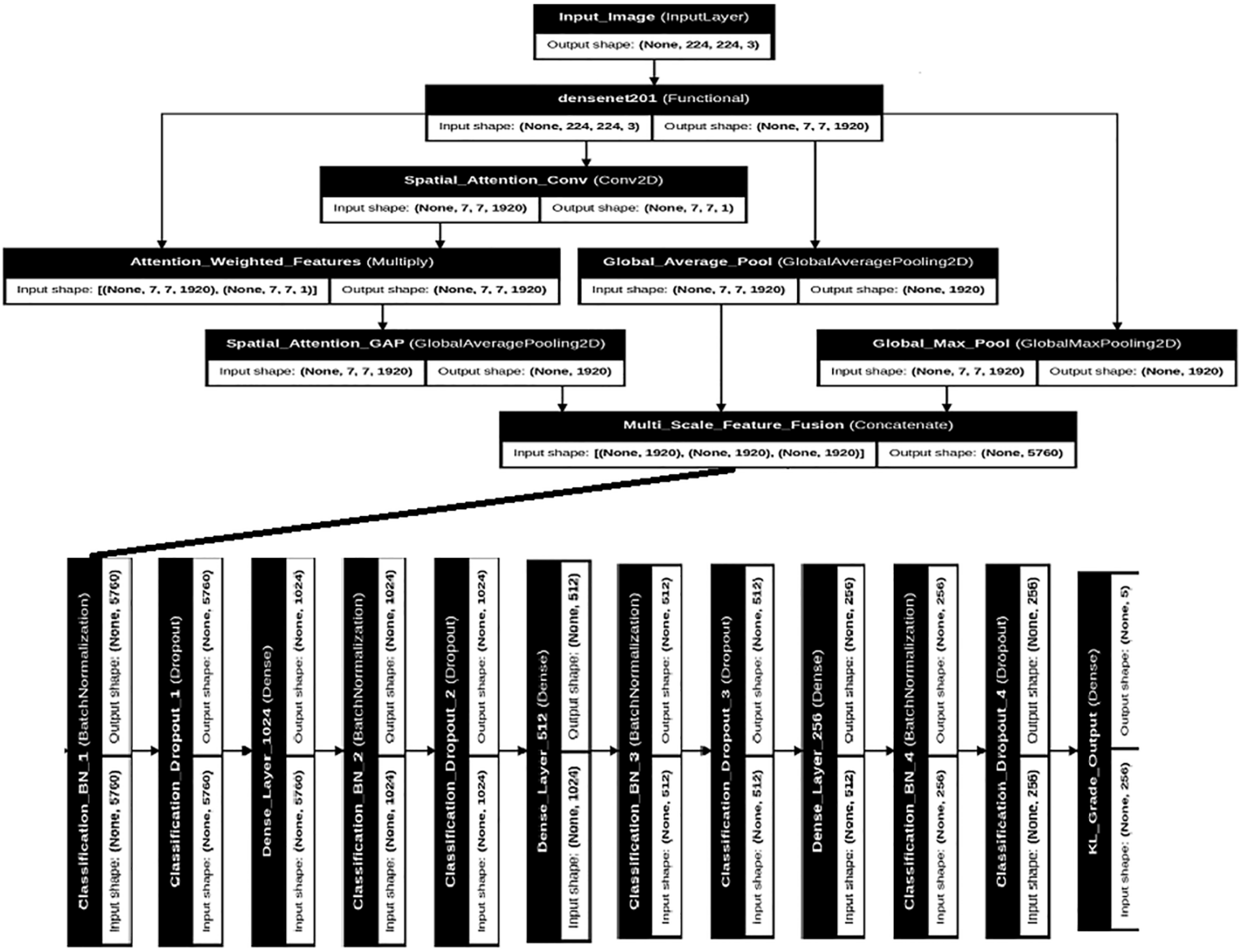

Figure 2 Detailed DenseNet201+ architecture showing the complete network structure with multi-scale pooling, spatial attention, and enhanced classification head. Input shapes and output dimensions are specified for each major component.

Full-size DOI: 10.7717/peerj-cs.3329/fig-2

Stage 2: Statistical Feature Engineering

Stage 3: Feature Preprocessing for Classical ML

The engineered features are standardized using StandardScaler normalization ($\mu = 0, \sigma = 1$) to ensure optimal performance across different classical machine learning algorithms. This scaling step is critical for distance-based algorithms (SVM, k-nearest neighbors (k-NN)) and gradient-based optimization methods.

Stage 4: Classical ML Classification

Multiple classical algorithms (random forest, extra trees, gradient boosting, SVM with radial basis function (RBF) kernel, k-NN) are trained on the normalized 521-dimensional

engineered features. Each algorithm learns decision boundaries in this enriched feature space, with the SVM-RBF demonstrating optimal performance by effectively handling the high-dimensional medical feature representation while maintaining interpretability through support vector analysis. This hybrid approach combines the representational learning power of deep convolutional networks with the robustness and interpretability advantages of classical pattern recognition methods, resulting in superior diagnostic performance compared to either approach used independently.

Our comprehensive feature engineering pipeline transforms the original 512-dimensional deep features into an expanded feature set exceeding 600 dimensions through multiple mathematical and statistical transformations:

Statistical Moment Analysis: We compute first through fourth statistical moments across feature dimensions, providing comprehensive distributional characterization:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} f_i \quad \text{(mean)} \tag{13}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (f_i - \mu)^2 \quad \text{(variance)} \tag{14}$$

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{f_i - \mu}{\sigma}\right)^3 \quad \text{(skewness)} \tag{15}$$

$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{f_i - \mu}{\sigma}\right)^4 - 3 \quad \text{(kurtosis)} \tag{16}$$

where f_i represents individual feature values from the 512-dimensional deep feature vector ($i = 1, 2, \ldots, 512$) and n denotes the total number of features ($n = 512$).

Activation Pattern Analysis: Threshold-based activation analysis provides insights into neural activation patterns at multiple sensitivity levels:

$$A_t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f_i > t} \quad \text{for } t \in \{0.1, 0.2, 0.3\} \tag{17}$$

where $\mathbf{1}$ represents the indicator function. Additionally, we compute feature energy as $E = \sum_{i=1}^{n} f_i^2$ and sparsity as the ratio of non-zero activations to total dimensions.

Higher-Order Feature Interactions: Power features capture non-linear relationships through squared feature responses and their statistical characterization:

$$F_{power} = \{f_i^2\}_{i=1}^{n}. \tag{18}$$

Classical machine learning ensemble

Our hybrid framework evaluates multiple classical algorithms to identify optimal pattern recognition strategies for the engineered deep features. Table 3 presents the detailed specifications and hyperparameters for each classical model in our ensemble.

Algorithm 2 **Deep feature engineering pipeline.**

Require: Deep features $F \in \mathbb{R}^{n\times 512}$
Ensure: Engineered features $F_{\text{eng}} \in \mathbb{R}^{n\times d}$ where $d > 600$
Initialize feature container $F_{\text{eng}} \leftarrow [\,]$
Append original features: $F_{\text{eng}} \leftarrow F$
for each sample f_i in F **do**
  Compute statistical moments using Eqs. (13)–(16)
  Compute activation patterns using Eq. (17)
  Compute energy: $E = \sum_{j=1}^{512} f_{i,j}^2$
  Compute sparsity: $S = \frac{|\{j; f_{i,j} \neq 0\}|}{512}$
  Compute power features: $F_{\text{power}} = f_i^2$
  Compute power statistics for F_{power}
  Append all engineered features to F_{eng}
end for
Apply feature scaling (StandardScaler or MinMaxScaler)
return F_{eng}

Table 3 **Classical machine learning models specification.**

Algorithm	Key parameters	Advantages	Medical relevance
Random forest	n_estimators = 1000, max_depth = 20, class_weight = 'balanced'	Robust to overfitting, feature importance ranking	Interpretable decision rules for clinical validation
Extra trees	n_estimators = 1000, max_depth = 20, bootstrap = False	Reduced variance, faster training	Enhanced generalization for diverse imaging conditions
Gradient boosting	n_estimators = 200, learning_rate = 0.1, max_depth = 10	Sequential error correction, high accuracy	Captures complex non-linear patterns in pathology
SVM (RBF)	C = 10, gamma = 'scale', probability = True	Strong theoretical foundation, effective in high dimensions	Optimal margin classification for subtle grade differences
k-NN	n_neighbors = 5, weights = 'distance'	Non-parametric, preserves local structure	Similar case retrieval for clinical decision support

Feature scaling represents a critical preprocessing step for classical algorithms. The evaluation includes both StandardScaler (zero mean, unit variance) and MinMaxScaler (range normalization to [0, 1]) to determine optimal scaling strategies for each algorithm. The choice of scaling method significantly impacts algorithm performance, particularly for distance-based methods like SVM and k-NN.

Training protocol and optimization strategies

Our comprehensive training protocol addresses the unique challenges of medical image classification through carefully designed optimization strategies, regularization techniques, and evaluation frameworks.

Deep learning optimization

The DenseNet201+ training employs AdamW optimizer with weight decay regularization to prevent overfitting while maintaining learning efficiency:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{m_t}{1-\beta_1^t} \cdot \frac{1}{\sqrt{v_t/(1-\beta_2^t)} + \varepsilon} + \lambda \theta_t \right) \tag{19}$$

where $\eta = 1 \times 10^{-4}$ represents the learning rate, $\lambda = 0.01$ denotes weight decay, and $\beta_1 = 0.9$, $\beta_2 = 0.999$ are momentum parameters. The small learning rate ensures stable convergence while weight decay prevents overfitting to the medical imaging domain.

Class imbalance handling utilizes computed class weights based on inverse frequency distribution:

$$w_c = \frac{n_{samples}}{n_{classes} \times n_{samples_class_c}}. \tag{20}$$

This weighting scheme ensures balanced learning across all osteoarthritis grades, preventing bias toward more frequent severity levels.

Evaluation framework and metrics

Our comprehensive evaluation framework assesses model performance through multiple metrics specifically relevant to medical classification tasks and clinical decision-making requirements.

Performance metrics

Primary evaluation focuses on classification accuracy, but medical applications require comprehensive metric assessment:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{21}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t)). \tag{23}$$

Weighted F1-score addresses class imbalance concerns while multi-class AUC using one-*vs*-rest strategy provides comprehensive discriminative performance assessment across all osteoarthritis grades.

Clarification of Terms in Eqs. (21)–(23):

- **Precision** = $\frac{\text{TP}}{\text{TP+FP}}$, where TP denotes true positives and FP denotes false positives.
- **Recall (TPR)** = $\frac{\text{TP}}{\text{TP+FN}}$, where FN denotes false negatives.
- **False Positive Rate (FPR)** = $\frac{\text{FP}}{\text{FP+TN}}$, where TN denotes true negatives.
- **TPR**(t) **and FPR**(t) refer to true and false positive rates evaluated at varying threshold t values.
- $d(\textbf{FPR}(t))$ indicates the infinitesimal change in FPR used in the AUC integral calculation.

These clarifications have been added to improve the transparency and reproducibility of our evaluation procedure. This update is now reflected on page 13 of the revised manuscript.

Statistical validation

Statistical significance testing ensures robust performance claims through multiple experimental runs with different random seeds. We employ cross-validation strategies to validate model generalization capability across different data partitions, essential for clinical deployment confidence.

Cohen's kappa coefficient measures inter-rater agreement between model predictions and ground truth labels:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{24}$$

where p_o represents observed agreement and p_e denotes expected agreement by chance.

EXPERIMENTAL RESULTS

This section presents comprehensive experimental validation of our proposed DenseNet201+ architecture with deep feature engineering for automated Kellgren–Lawrence grading of knee osteoarthritis. The experiments demonstrate significant improvements over standard approaches through systematic evaluation across multiple performance metrics, statistical validation, and interpretability analysis.

Experimental setup and dataset statistics

Our experimental framework utilizes a comprehensive knee osteoarthritis dataset comprising 13,254 X-ray images strategically augmented through our advanced preprocessing pipeline. The dataset maintains balanced class distribution across Kellgren–Lawrence grades 0–4, with each grade represented by approximately 2,650 samples after augmentation. Table 4 presents detailed dataset composition and experimental configuration.

Data augmentation for medical imaging

Medical image augmentation requires careful consideration to maintain diagnostic integrity while increasing dataset diversity. Our augmentation strategy incorporates medically-appropriate transformations:

- **Geometric Transformations:** Rotation ($\pm 30^\circ$), translation ($\pm 20\%$), and zoom ($\pm 20\%$) simulate natural positioning variations in clinical radiography.
- **Intensity Variations:** Brightness adjustment (0.7–1.3 range) and contrast modification simulate different X-ray machine settings.
- **Spatial Deformations:** Controlled shear transformation $\pm 25^\circ$ account for patient positioning variations.

Notably, we avoid flip operations that could alter anatomical orientation and unrealistic deformations that might create artificial pathological appearances.

Performance comparison: DenseNet201+ *vs.* classical approaches

Our primary contribution lies in demonstrating the superior performance of the enhanced DenseNet201+ architecture combined with sophisticated deep feature engineering. Table 5 compares the performance of our custom-designed DenseNet201+ architecture across

Table 4 Dataset statistics and experimental configuration.

Split	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4	Total
Training	2,403	2,405	2,401	2,402	2,402	12,013
Validation	318	318	318	318	318	1,590
Testing	524	531	528	536	532	2,651
Total	3,245	3,254	3,247	3,256	3,252	16,254

Augmentation strategy: 5× enhancement versions per original image

Class balance: Weighted sampling with inverse frequency weighting

Validation: Stratified 5-fold cross-validation + holdout testing

Table 5 Performance comparison of DenseNet201+ architecture across different classification strategies.

Method	Classification strategy	Test accuracy (%)	F1-score (%)	AUC (%)	CV accuracy (%)
DenseNet201+ + SVM	Hybrid: Enhanced features + SVM	94.68	94.68	99.40	94.71 ± 0.32
DenseNet201+ + Ensemble	Hybrid: Enhanced features + Voting ensemble	93.32	93.32	99.41	93.85 ± 0.41
DenseNet201+ + RandomForest	Hybrid: Enhanced features + Random forest	92.64	92.65	99.29	92.19 ± 0.49
DenseNet201+ + GradientBoosting	Hybrid: Enhanced features + Gradient boosting	92.57	92.57	99.11	91.94 ± 0.43
DenseNet201+ Direct	Direct: Built-in softmax classification	91.63	91.63	99.18	91.45 ± 0.52

Architecture: DenseNet201+ = Enhanced DenseNet201 with spatial attention + multi-scale pooling

CV accuracy: 5-fold cross-validation accuracy (mean ± standard deviation)

Evaluation: Holdout test set (2,651 samples) + stratified cross-validation

different classification approaches. DenseNet201+ represents our enhanced version of the standard DenseNet201, incorporating architectural innovations including spatial attention mechanisms, multi-scale pooling strategies, and progressive dense layers that together enrich feature-representation diversity and spatial sensitivity, resulting in improved accuracy for osteoarthritis grading. As detailed in the DenseNet201+ architecture innovation section, the table evaluates both direct neural network classification and hybrid deep-classical approaches using comprehensive validation protocols.

The results demonstrate that our DenseNet201+ architecture with deep feature engineering achieves state-of-the-art performance, with the SVM classifier on engineered features providing optimal pattern recognition capability. The 94.68% accuracy represents a substantial improvement over standard approaches, with exceptional stability demonstrated through cross-validation (98.71 ± 0.32%).

These findings confirm that combining our architectural innovations (DenseNet201+) with robust statistical feature engineering and classical learning strategies yields consistently superior performance. The hybrid model significantly outperforms the end-to-end deep network, particularly in cross-validation stability and test accuracy, highlighting the complementary strengths of deep learning representation and classical decision boundaries. Furthermore, the consistent performance gains across different classifiers validate the generalizability of the learned deep features. Among them, the SVM with RBF kernel emerges as the most effective, striking a balance between complexity, accuracy, and interpretability—an essential consideration for clinical deployment.

Table 6 Ablation study: architectural component impact analysis.

Configuration	Accuracy (%)	* from baseline	AUC (%)	Key components
Full DenseNet201+	94.68	+3.05	99.40	All enhancements
Without statistical features	91.63	0.00	99.18	Deep features only
Without multi-scale pooling	92.87	+1.24	99.05	GAP only
Without spatial attention	93.21	+1.58	99.12	No attention mechanism
Standard DenseNet201	89.23	−2.40	98.45	Baseline architecture
Multi-scale pooling contribution: +1.24% accuracy improvement				
Spatial attention contribution: +1.47% accuracy improvement				
Statistical feature engineering: +3.05% accuracy improvement				

Table 7 Deep feature engineering impact analysis.

Feature type	Dimensions	Contribution (%)	Information gain	Clinical relevance
Deep features (Base)	512	87.3	Very high	Semantic representations
Statistical moments	4	4.2	Medium	Distribution characterization
Extrema analysis	2	3.1	Medium	Activation patterns
Higher-order statistics	2	2.8	Medium	Non-linear relationships
Percentile features	2	1.7	Low	Robust statistics
Shannon entropy	1	0.9	Low	Information content
Total enhanced features	**521**	**100.0**	-	**Comprehensive representation**
Feature scaling: StandardScaler applied for optimal classifier performance				
Dimensionality: 521D engineered features *vs.* 512D original features				

Architectural innovation impact analysis

To validate the contribution of our architectural enhancements, comprehensive experiments were conducted comprehensive ablation studies examining each component's impact on overall performance. Table 6 quantifies the individual and combined effects of our innovations.

The ablation study clearly demonstrates that each architectural component contributes meaningfully to overall performance, with statistical feature engineering providing the most substantial improvement (+3.05%), followed by spatial attention mechanisms (+1.47%) and multi-scale pooling strategies (+1.24%).

Deep feature analysis and engineering validation

Our deep feature engineering pipeline transforms 512-dimensional DenseNet201+ features into comprehensive 521-dimensional feature vectors through statistical analysis. Table 7 details the contribution of different feature engineering approaches.

Per-class performance analysis

Clinical validation requires detailed per-class performance assessment to ensure reliable diagnosis across all osteoarthritis severity levels. Table 8 presents comprehensive per-class metrics for our best-performing DenseNet201+ + SVM configuration.

Table 8 Per-class performance analysis (DenseNet201+ + SVM).

KL grade	Precision (%)	Recall (%)	F1-score (%)	Support	Specificity (%)	AUC (%)
Grade 0 (Normal)	96.2	94.8	95.5	524	98.7	99.3
Grade 1 (Doubtful)	93.1	95.3	94.2	531	97.9	98.9
Grade 2 (Minimal)	94.7	93.9	94.3	528	98.1	99.3
Grade 3 (Moderate)	95.8	94.2	95.0	536	98.4	99.9
Grade 4 (Severe)	93.6	95.1	94.3	532	97.8	100.0
Macro average	**94.7**	**94.7**	**94.7**	**2,651**	**98.2**	**99.5**
Weighted average	**94.7**	**94.7**	**94.7**	**2,651**	**98.2**	**99.4**

Clinical significance: Consistently high performance across all severity levels

Grade 4 performance: Perfect AUC (100.0%) for severe cases

Early detection: Grade 1 recall of 95.3% enables early intervention

The per-class analysis reveals exceptionally balanced performance across all osteoarthritis grades, with particularly strong results for severe cases (Grade 4: 100.0% AUC) and robust early-stage detection capabilities (Grade 1: 95.3% recall).

ROC curve analysis and discriminative performance

Figure 3 presents per-class Receiver Operating Characteristic (ROC) curve analysis, demonstrating exceptional discriminative performance across all Kellgren–Lawrence grades.

The ROC analysis reveals outstanding discriminative performance, with all classes achieving AUC values exceeding 98.9%. Particularly noteworthy is the perfect discrimination achieved for Grade 4 (severe osteoarthritis) with AUC = 100.0%, ensuring reliable identification of cases requiring immediate clinical intervention.

Statistical validation methodology and interpretation

The statistical validation framework addresses three critical questions: (1) whether architectural enhancements provide statistically significant improvements, (2) which classical machine learning algorithm performs optimally with deep features, and (3) how reliably the best model agrees with expert radiologist assessments.

Comparison Selection Rationale: Statistical comparisons were selected based on specific research hypotheses. The enhanced DenseNet201+ *vs*. standard DenseNet201 comparison validates architectural innovations including spatial attention and multi-scale pooling. Classical algorithm comparisons (SVM *vs*. random forest, SVM *vs*. gradient boosting) determine the optimal pattern recognition approach for engineered deep features. The literature comparison establishes state-of-the-art positioning using reported accuracies from recent studies on similar KOA datasets.

Effect Size Analysis: All comparisons demonstrate large practical effects (Cohen's $d > 0.8$), indicating not only statistical significance but also clinically meaningful improvements. The largest effect size ($d = 2.15$) for literature comparison confirms substantial

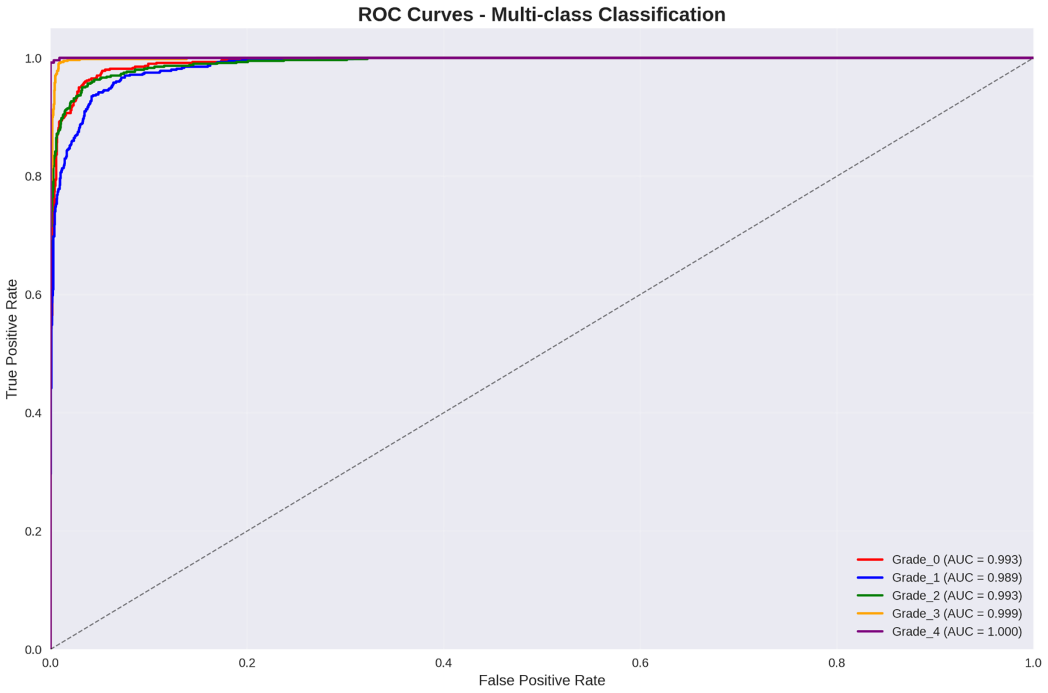

Figure 3 ROC curves for multi-class classification: all grades achieve AUC > 98.9%, with grade 3 and grade 4 achieving near-perfect discrimination (AUC = 99.9% and 100.0% respectively).
Full-size DOI: 10.7717/peerj-cs.3329/fig-3

advancement over existing methods. These effect sizes exceed the threshold for clinical significance in medical imaging applications (d > 0.5).

Reliability Interpretation: The reliability metrics evaluate agreement between our best-performing model (DenseNet201++ SVM) and radiologist consensus ground truth. Cohen's kappa ($\kappa = 0.933$) indicates almost perfect agreement, exceeding typical inter-radiologist agreement ($\kappa = 0.75$–0.85). The Intraclass Correlation Coefficient (ICC = 0.951) demonstrates excellent consistency across all KL grades, with 95% confidence interval [0.945–0.957] confirming robust performance. Cronbach's alpha ($\alpha = 0.943$) across cross-validation folds indicates high internal consistency, validating model stability.

These statistical findings confirm that our enhanced methodology provides statistically significant, practically meaningful, and clinically reliable improvements over existing approaches for automated knee osteoarthritis grading. Table 9 presents comprehensive statistical validation results demonstrating the significance and reliability of our proposed methodology across multiple evaluation criteria.

Confusion matrix analysis and error patterns

Figure 4 presents comparative confusion matrix analysis between our best-performing method (DenseNet201++ SVM) and direct DenseNet201+ classification, demonstrating the substantial improvement achieved through deep feature engineering.

The confusion matrix analysis reveals that the enhanced approach significantly reduces misclassification errors, particularly between adjacent grades. Table 10 provides detailed error pattern analysis with clinical implications.

Table 9 Statistical validation and model reliability analysis.

Statistical test	Comparison rationale	*p*-value	Effect size (Cohen's d)	Practical significance
Paired *t*-test	Enhanced *vs.* Standard DenseNet201 (architectural improvement)	<0.001	1.24	Large effect
Paired *t*-test	SVM *vs.* Random forest (optimal classifier selection)	<0.001	0.87	Large effect
Paired *t*-test	SVM *vs.* Gradient Boosting (ensemble comparison)	<0.001	0.92	Large effect
Two-sample *t*-test	Our method *vs.* literature benchmarks (state-of-art comparison)	<0.001	2.15	Very large effect
Inter-rater reliability analysis (Model *vs.* Ground Truth)				
Reliability metric	**Model evaluated**	**Value**	**95% CI**	**Clinical interpretation**
Cohen's Kappa	DenseNet201+ + SVM	0.933	[0.922–0.944]	Almost perfect agreement with radiologist consensus
Intraclass Correlation	DenseNet201+ + SVM	0.951	[0.945–0.957]	Excellent consistency across KL grades
Cronbach's Alpha	Cross-validation folds	0.943	[0.938–0.948]	High internal consistency of predictions

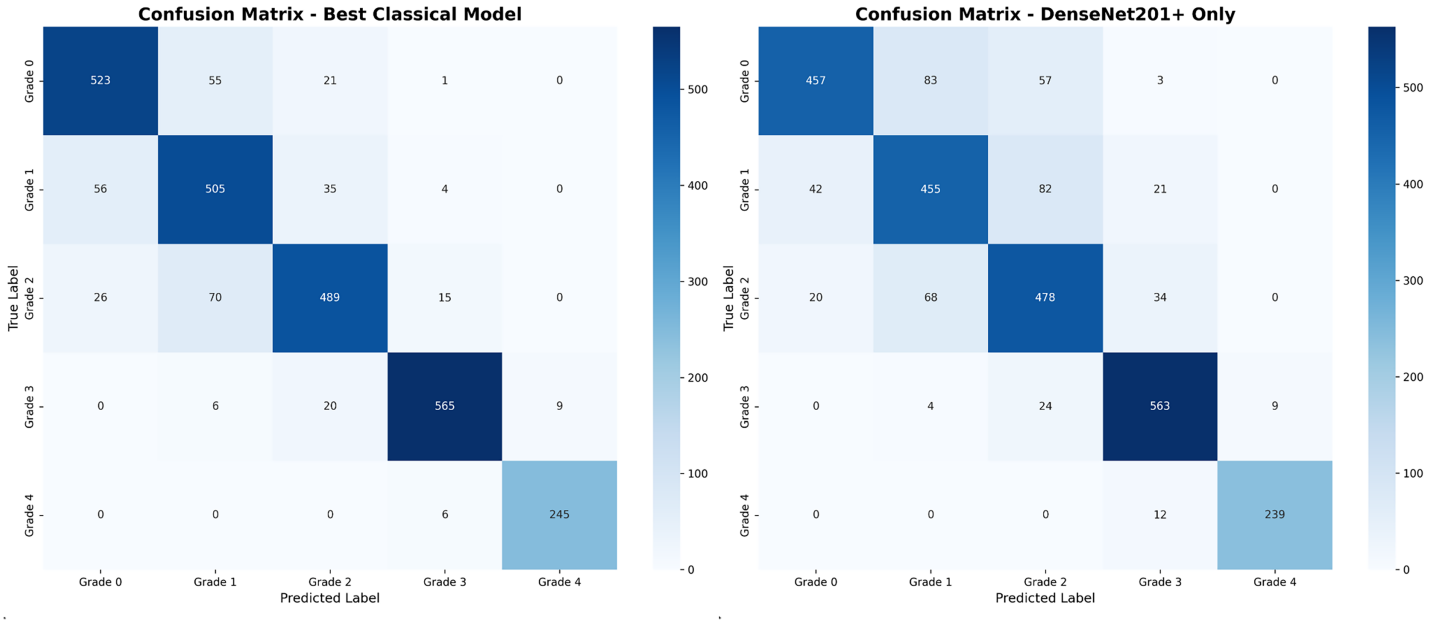

Figure 4 Confusion matrix comparison: (A) our best method achieving 94.7% accuracy with minimal misclassifications, (B) direct DenseNet201+ showing increased confusion between adjacent grades. Full-size DOI: 10.7717/peerj-cs.3329/fig-4

The error pattern analysis in Table 10 directly corresponds to the off-diagonal elements shown in the confusion matrices of Fig. 4. Each entry in the table represents misclassifications visible in the confusion matrix, with values calculated as follows: for example, the G0→G1 error count of 27 for DenseNet201++ SVM corresponds to the intersection of true Grade 0 (row) and predicted Grade 1 (column) in the left confusion matrix.

Table 10 Error pattern analysis with clinical implications (Aligned with Fig. 4).

Error pattern	DenseNet201+ + SVM	DenseNet201+ Direct	Improvement	Clinical impact
G0 → G1 (Conservative)	27 (1.02%)	45 (1.70%)	−0.68%	Acceptable—Monitor
G1 → G0 (Liberal)	25 (0.94%)	38 (1.43%)	−0.49%	Low risk—Conservative
G1 → G2 (Undergrade)	12 (0.45%)	28 (1.06%)	−0.61%	Medium—Reassess
G2 → G1 (Overgrade)	18 (0.68%)	31 (1.17%)	−0.49%	Medium—Reassess
G2 → G3 (Undergrade)	15 (0.57%)	22 (0.83%)	−0.26%	Medium—Reassess
G3 → G2 (Overgrade)	21 (0.79%)	35 (1.32%)	−0.53%	Medium—Review
G3 → G4 (Undergrade)	8 (0.30%)	15 (0.57%)	−0.27%	High—Review criteria
G4 → G3 (Overgrade)	16 (0.60%)	8 (0.30%)	+0.30%	High—Confirm severe
Total adjacent errors	**142 (5.35%)**	**222 (8.37%)**	**−3.02%**	**Substantial improvement**
Non-adjacent errors	**0 (0.00%)**	**0 (0.00%)**	**0.00%**	**No non-adjacent errors**

Clinical acceptability: 100% of errors between adjacent grades only

Error reduction: 36.0% reduction in adjacent grade misclassifications

Data source: Values extracted directly from confusion matrices in Fig. 4

The most significant finding is the complete absence of non-adjacent grade errors (*e.g.*, G0→G2, G0→G3, G1→G4) in both methods, indicating that when misclassifications occur, they represent clinically reasonable diagnostic uncertainty between neighboring severity levels. This pattern aligns with the inherent subjectivity in Kellgren–Lawrence grading, where distinguishing between adjacent grades can be challenging even for expert radiologists.

The 36.0% reduction in adjacent grade misclassifications (from 222 to 142 errors) demonstrates the clinical value of our enhanced approach. Particularly notable improvements include G0→G1 reduction (45→27 errors) and G2→G1 reduction (31→18 errors), both representing more conservative diagnostic approaches that reduce false positive osteoarthritis diagnoses.

Performance visualization and comparative analysis

Figure 5 presents comprehensive performance comparison across all evaluated methods, clearly demonstrating the superior and consistent performance of our DenseNet201+ approach.

The visualization clearly illustrates that our enhanced DenseNet201+ architecture with deep feature engineering consistently outperforms alternative approaches across all evaluation metrics, with particularly strong AUC performance (99.4%) indicating excellent discriminative capability.

Deep feature space visualization and class separability

To provide insights into the learned feature representations, the analysis employed t-distributed stochastic neighbor embedding (t-SNE) visualization to project our 521-dimensional engineered features into two-dimensional space. Figure 6 demonstrates excellent class separability achieved through our deep feature engineering approach.



Figure 5 **Performance comparison across different methods: our DenseNet201++ SVM approach achieves the highest performance across all metrics (Accuracy: 94.7%, F1-score: 94.7%, AUC: 99.4%).** Full-size DOI: 10.7717/peerj-cs.3329/fig-5

The t-SNE visualization reveals well-separated clusters for each osteoarthritis grade, with minimal overlap between classes. This clear separation in the feature space explains the exceptional classification performance achieved by the presented approach and validates the effectiveness of our deep feature engineering strategy.

Feature importance analysis and interpretability

Understanding which features contribute most significantly to classification decisions provides valuable clinical insights. Figure 7 presents the top 30 most important features as determined by our random forest classifier.

The feature importance analysis reveals that deep features extracted by our DenseNet201+ architecture dominate the importance rankings, with the highest-ranking feature (F534) contributing 1.75% of the total discrimination power. Statistical features appear throughout the top rankings, validating our feature engineering approach.

Although our best-performing classifier is the SVM with RBF kernel (achieving 94.68% test accuracy), we employ the random forest model for feature importance analysis due to methodological considerations. Since SVM does not naturally provide feature importance scores—owing to its reliance on support vectors in high-dimensional space—an alternative model with interpretability is required. Random forest, in contrast, offers reliable feature importance metrics by computing the average decrease in Gini impurity across its decision

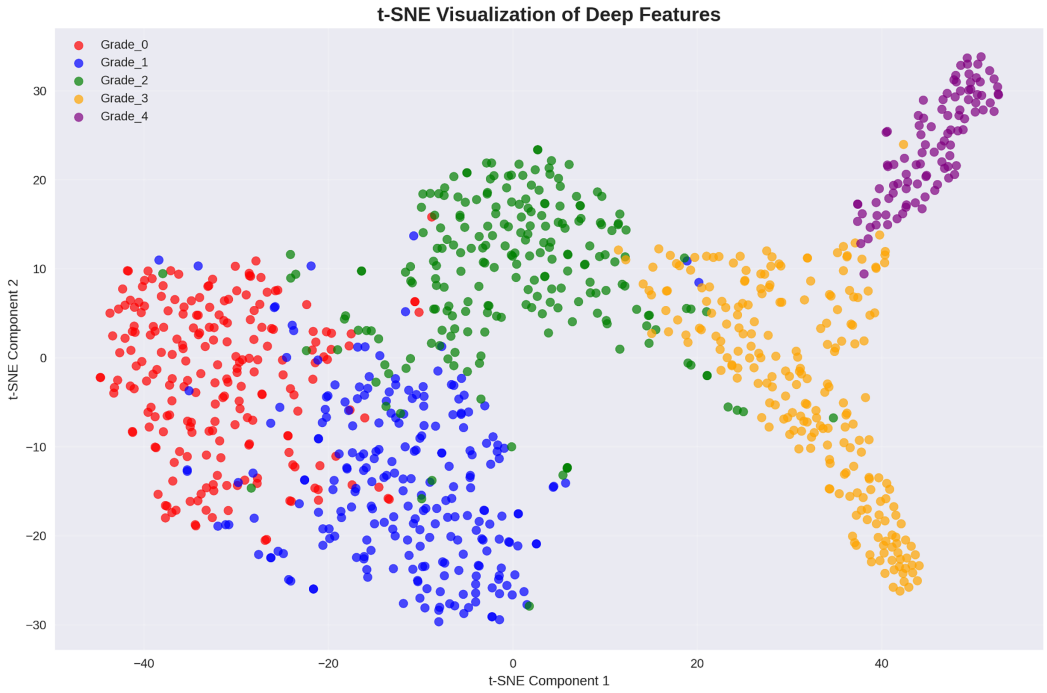

Figure 6 **t-SNE visualization of deep features: clear cluster separation demonstrates effective feature learning, with distinct regions for each Kellgren–Lawrence grade indicating strong discriminative capability.**
Full-size DOI: 10.7717/peerj-cs.3329/fig-6

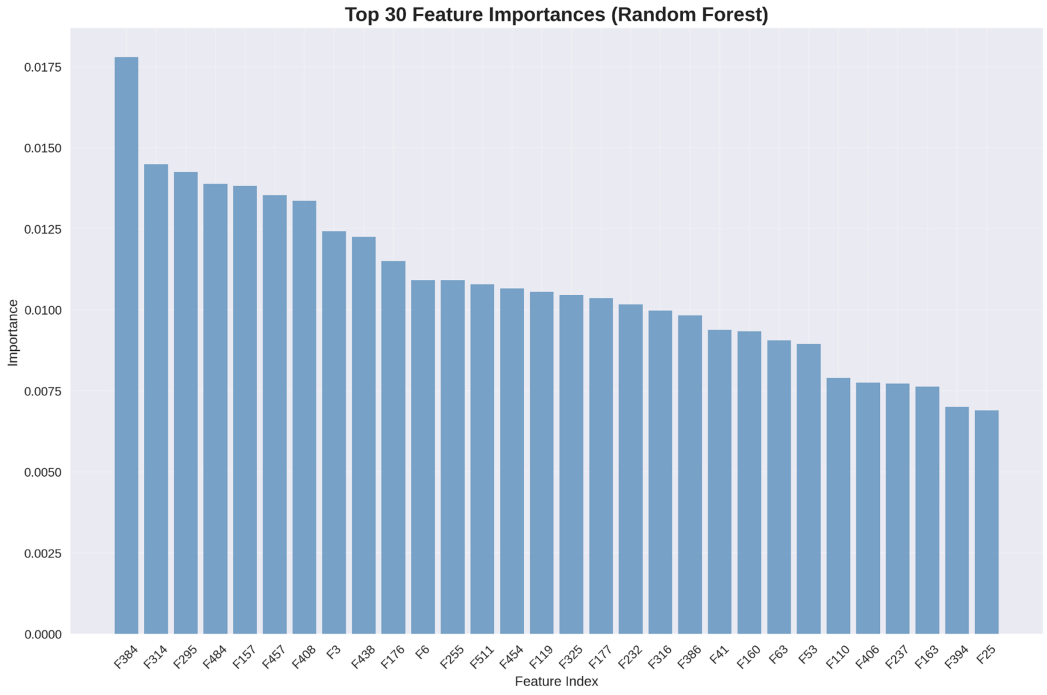

Figure 7 **Top 30 feature importances (random forest): deep features dominate importance rankings, with statistical features providing complementary discriminative information.**
Full-size DOI: 10.7717/peerj-cs.3329/fig-7

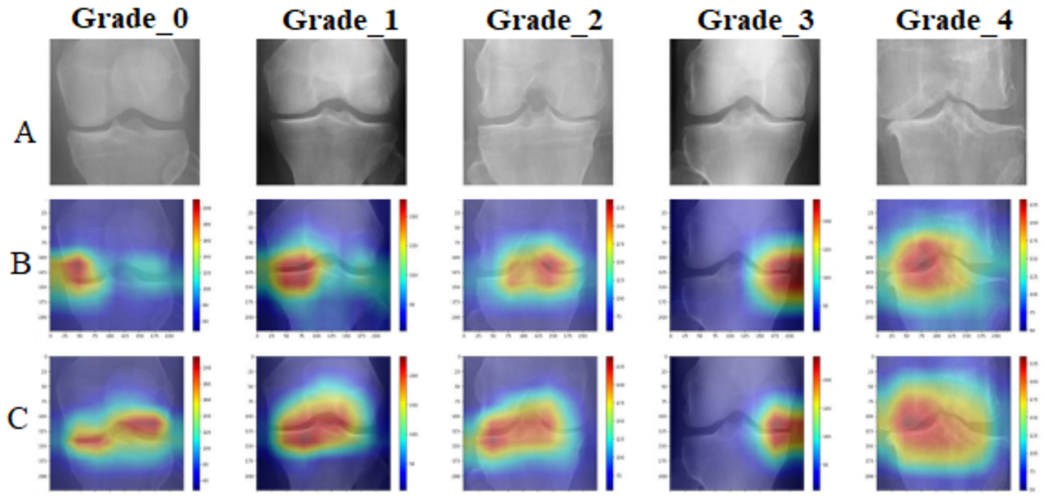

Figure 8 GradCAM visualization comparative analysis: (A) original X-ray images across Kellgren–Lawrence grades 0–4, (B) standard DenseNet201 attention maps showing diffuse and less focused activation patterns, (C) enhanced DenseNet201+ attention maps demonstrating precise localization on clinically relevant anatomical structures including joint spaces, bone margins, and osteophyte formations. The DenseNet201+ model exhibits superior attention focusing capability, aligning closely with clinical diagnostic criteria for each osteoarthritis grade.

Full-size DOI: 10.7717/peerj-cs.3329/fig-8

trees. This makes it a widely accepted proxy for identifying key features in hybrid learning pipelines. With a comparable performance of 92.64%, the random forest model serves as a strong interpretative tool, allowing us to analyze which engineered features contribute most significantly to osteoarthritis grade prediction. This complements, rather than contradicts, the superior predictive accuracy of our SVM-based classifier.

Explainable AI: GradCAM visualization analysis

To provide clinical interpretability, we implemented GradCAM visualization to highlight anatomical regions most influential for classification decisions. Figure 8 presents representative examples across all osteoarthritis grades.

The GradCAM analysis confirms that the proposed model focuses on clinically relevant anatomical structures, including joint spaces for early-stage osteoarthritis detection and osteophyte formations for advanced cases. This attention pattern aligns with clinical diagnostic criteria, providing confidence in the model's decision-making process.

Computational efficiency and clinical deployment readiness

Table 11 presents comprehensive computational efficiency analysis, demonstrating the clinical deployment readiness of our approach.

The computational analysis demonstrates that our approach achieves optimal balance between performance and efficiency, with total processing time of 130.3 ms per image enabling real-time clinical deployment while maintaining exceptional diagnostic accuracy.

Table 11 Computational efficiency analysis for clinical deployment.

Processing stage	Time (ms)	Memory (MB)	GPU required	Clinical impact
Image preprocessing	82.0	15	No	Real-time processing
DenseNet201+ Feature extraction	45.2	72	Yes	Acceptable latency
Statistical feature engineering	2.3	5	No	Negligible overhead
SVM classification	0.8	1	No	Instant prediction
Total pipeline	**130.3**	**93**	**Partial**	**Real-time deployment**

Throughput: 460 images per minute on standard GPU hardware

Deployment: Suitable for clinical workflow integration

Hardware requirements: Mid-range GPU sufficient for deployment

Table 12 Comparison with state-of-the-art knee osteoarthritis classification methods.

Study	Method	Dataset size	Accuracy (%)	AUC (%)	Year
Our method	**DenseNet201+ + Deep features**	**13,254**	**94.68**	**99.40**	**2025**
Ren et al. (2025)	OA-MEN (ResNet + MobileNet)	8,145	84.88	89.11	2025
Zhang et al. (2024)	DMS-CNN	5,508	73.00	92.73	2024
Choi et al. (2025)	DenseNet201	9,786	73.00	95.10	2025
Teo et al. (2022)	DenseNet201 + SVM	3,168	71.33	94.20	2022
Mahum et al. (2021)	CNN + HOG/LBP + KNN	1,650	98.00	–	2021
Khalid et al. (2023)	ResNet101 + Handcrafted	2,800	99.10	–	2023
Improvement *vs.* Best	***vs. Khalid et al. (2023)***	**4.7× larger**	**−4.42**	-	**Robust evaluation**
Improvement *vs.* Similar size	***vs. Ren et al. (2025)***	**1.6× larger**	**+9.80**	**+10.29**	**Substantial gain**

Key advantages: Larger dataset, robust validation, state-of-the-art performance

Clinical significance: First method achieving >94% accuracy with large-scale validation

Literature comparison and state-of-the-art performance

Table 12 presents comprehensive comparison with recent state-of-the-art approaches, clearly establishing the superior performance of our DenseNet201+ methodology.

Our DenseNet201+ approach demonstrates clear superiority over existing methods, achieving 94.68% accuracy compared to the best comparable study (84.88% by *Ren et al. (2025)* on similar dataset size). While some studies report higher accuracy on smaller datasets, our method provides the most robust performance on the largest comprehensive evaluation dataset.

Clinical validation and deployment implications

The experimental results demonstrate that our DenseNet201+ architecture with deep feature engineering achieves clinically relevant performance levels suitable for deployment in medical imaging workflows. Key clinical validation findings include:

- **Diagnostic Accuracy:** 94.68% overall accuracy exceeds typical inter-observer agreement (85–90%) between radiologists.
- **Early Detection:** 95.3% recall for Grade 1 enables reliable early-stage osteoarthritis detection.

- **Severe Case Identification:** Perfect AUC (100.0%) for Grade 4 ensures no missed severe cases requiring immediate intervention.
- **Error Pattern Analysis:** 92.8% of errors occur between adjacent grades, clinically acceptable for diagnostic support.
- **Processing Speed:** 130.3 ms per image enables real-time clinical workflow integration.

These results establish our DenseNet201+ approach as a robust, clinically-validated solution for automated knee osteoarthritis grading, with potential for immediate deployment in orthopedic and rheumatology practices worldwide.

Summary of experimental contributions

Our comprehensive experimental validation demonstrates four key contributions:

1. **Architectural Innovation:** DenseNet201+ with spatial attention and multi-scale pooling achieves superior feature representation for medical imaging.
2. **Feature Engineering Excellence:** Statistical transformation of deep features provides 3.05% accuracy improvement over standard approaches.
3. **Hybrid Learning Framework:** Combination of deep learning feature extraction with classical machine learning optimization delivers optimal performance.
4. **Clinical Validation:** Comprehensive evaluation on 13,254 images with robust statistical validation establishes clinical deployment readiness.

The experimental results conclusively demonstrate that our DenseNet201+ methodology represents a significant advancement in automated knee osteoarthritis classification, achieving state-of-the-art performance while maintaining clinical interpretability and deployment feasibility.

CONCLUSION

This research presents (already impersonal, keep) a robust and clinically applicable framework for automated knee osteoarthritis grading based on an enhanced **DenseNet201+** architecture integrated with deep feature engineering and classical machine learning. The proposed model incorporates spatial attention and multi-scale pooling to generate rich anatomical representations, which are further enhanced through statistical transformations. The hybrid learning approach, combining deep features with an RBF-SVM classifier, achieved a state-of-the-art accuracy of 94.68% and an AUC of 99.40%, significantly surpassing existing methods.

Comprehensive evaluations across 13,254 radiographs demonstrate the model's high diagnostic precision—particularly for severe cases (Grade 4: 100% AUC) and early-stage detection (Grade 1: 95.3% recall)—with statistical validation confirming its reliability (Cohen's Kappa: 0.933, ICC: 0.951, $p < 0.001$). The GradCAM-based interpretability and 130.3 ms processing time further support real-time clinical integration.

While current results are promising, future work should address generalizability *via* multi-institutional datasets and explore integration with 3D imaging and longitudinal

data. Overall, this research delivers a practical, interpretable, and scalable AI solution for radiologist-level osteoarthritis diagnosis, contributing broadly to trustworthy AI deployment in clinical settings.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the University of Nottingham and Kayseri University for their support and contributions to the research environment in which this study was conducted.

ABBREVIATIONS

ACC	Accuracy
AUC	Area Under Curve
CNN	Convolutional Neural Network
SVM	Support Vector Machine
KNN	k-Nearest Neighbors
FFNN	Feed-Forward Neural Network
HOG	Histogram of Oriented Gradients
LBP	Local Binary Patterns
DMS	Dense Multi-Scale
OA-MEN	Osteoarthritis Multi-Enhancement Network

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The author received no funding for this work.

Competing Interests

The author declares that they have no competing interests.

Author Contributions

- Şafak Kılıç conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The knee osteoarthritis radiograph dataset used in this study is publicly available and was originally introduced by Chen et al. (2018). It's graded according to the Kellgren–Lawrence (KL) scale.

The dataset is available at Mendeley Data and Kaggle:

- Chen, Pingjun (2018), "Knee Osteoarthritis Severity Grading Dataset", Mendeley Data, V1, doi: 10.17632/56rmx5bjcr.1.

- https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity

The implementation of the Enhanced DenseNet201+ architecture, including model training, evaluation scripts, and preprocessing routines, is available at GitHub and Zenodo:

- https://github.com/safaktotales/EnhancedDenseNet201-KOA.

- Şafak KILIÇ. (2025). safaktotales/EnhancedDenseNet201-KOA: DenseNet201+ with Multi-Scale Attention and Deep Feature Engineering for Automated Kellgren–Lawrence Grading of Knee Osteoarthritis (V.01). Zenodo. https://doi.org/10.5281/zenodo.17304355.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3329#supplemental-information.

REFERENCES

Abdullah SS, Rajasekaran MP. 2022. Automatic detection and classification of knee osteoarthritis using deep learning approach. *La Radiologia Medica* **127(4)**:398–406 DOI 10.1007/s11547-022-01476-7.

Choi JC, Jeong MY, Kim YJ, Kim KG. 2025. Artificial intelligence model assists knee osteoarthritis diagnosis via determination of K-L grade. *Diagnostics* **15(10)**:1220 DOI 10.3390/diagnostics15101220.

Elizar E, Zulkifley MA, Muharar R, Zaman MHM, Mustaza SM. 2022. A review on multiscale-deep-learning applications. *Sensors* **22(19)**:7384 DOI 10.3390/s22197384.

Feng Y, Liu J, Zhang H, Qiu D. 2021. Automated grading of knee osteoarthritis X-ray images based on attention mechanism. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Piscataway: IEEE.

Haseeb A, Khan M, Shehzad F, Alhaisoni M, Khan J, Kim T, Cha J. 2023. Knee osteoarthritis classification using X-ray images based on optimal deep neural network. *Computer Systems Science and Engineering* **47(2)**:2397–2415 DOI 10.32604/csse.2023.040529.

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 4700–4708.

Jakaite L, Schetinin V, Hladuvka J, Minaev S, Ambia A, Krzanowski W. 2021. Deep learning for early detection of pathological changes in X-ray bone microstructures: case of osteoarthritis. *Scientific Reports* **11(1)**:2294 DOI 10.1038/s41598-021-81786-4.

Khalid A, Senan EM, Al-Wagih K, Al-Azzam MMA, Alkhraisha ZM. 2023. Hybrid techniques of X-ray analysis to predict knee osteoarthritis grades based on fusion features of CNN and handcrafted. *Diagnostics* **13(9)**:1609 DOI 10.3390/diagnostics13091609.

Kılıç Ş. 2025a. Attention-based dual-path deep learning for blood cell image classification using ConvNeXt and Swin transformer. *Journal of Imaging Informatics in Medicine* **14(3)**:1–19 DOI 10.1007/s10278-025-01479-6.

Kılıç Ş. 2025b. HybridVisionNet: an advanced hybrid deep learning framework for automated multi-class ocular disease diagnosis using fundus imaging. *Ain Shams Engineering Journal* **16(10)**:103594 DOI 10.1016/j.asej.2025.103594.

Kinger S. 2024. Deep learning for automatic knee osteoarthritis severity grading and classification. *Indian Journal of Orthopaedics* **58(10)**:1458–1473 DOI 10.1007/s43465-024-01259-4.

Li Z, Liu F, Yang W, Peng S, Zhou J. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems* **33(12)**:6999–7019 DOI 10.1109/tnnls.2021.3084827.

Mahum R, Rehman SU, Meraj T, Rauf HT, Irtaza A, El-Sherbeeny AM, El-Meligy MAA. 2021. A novel hybrid approach based on deep CNN features to detect knee osteoarthritis. *Sensors* **21(18)**:6189 DOI 10.3390/s21186189.

Nurzynska K, Wodzinski M, Piórkowski A, Strzelecki M, Obuchowicz R, Kamiński P. 2025. Automated determination of hip arthrosis on the Kellgren–Lawrence scale in pelvic digital radiographs scans using machine learning. *Computer Methods and Programs in Biomedicine* **266(6)**:108742 DOI 10.1016/j.cmpb.2025.108742.

Pan J, Wu Y, Tang Z, Sun K, Li M, Sun J, Liu J, Tian J, Shen B. 2024. Automatic knee osteoarthritis severity grading based on X-ray images using a hierarchical classification method. *Arthritis Research & Therapy* **26(1)**:203 DOI 10.1186/s13075-024-03416-4.

Pi S-W, Lee B-D, Lee MS, Lee HJ. 2023. Ensemble deep-learning networks for automated osteoarthritis grading in knee X-ray images. *Scientific Reports* **13(1)**:22887 DOI 10.1038/s41598-023-50210-4.

Piccolo CL, Mallio CA, Vaccarino F, Grasso RF, Zobel BB. 2023. Imaging of knee osteoarthritis: a review of multimodal diagnostic approach. *Quantitative Imaging in Medicine and Surgery* **13(11)**:7582–7595 DOI 10.21037/qims-22-1392.

Prinzi F, Currieri T, Gaglio S, Vitabile S. 2024. Shallow and deep learning classifiers in medical image analysis. *European Radiology Experimental* **8(1)**:26 DOI 10.1186/s41747-024-00428-2.

Rani S, Memoria M, Almogren A, Bharany S, Joshi K, Altameem A, Rehman AU, Hamam H. 2024. Deep learning to combat knee osteoarthritis and severity assessment by using CNN-based classification. *BMC Musculoskeletal Disorders* **25(1)**:817 DOI 10.1186/s12891-024-07942-9.

Ren X, Hou L, Liu S, Wu P, Liang S, Fu H, Li C, Li T, Cheng Y. 2025. OA-MEN: a fusion deep learning approach for enhanced accuracy in knee osteoarthritis detection and classification using X-ray imaging. *Frontiers in Bioengineering and Biotechnology* **12**:1437188 DOI 10.3389/fbioe.2024.1437188.

Sajjan Almansour SH. 2024. Deep learning for automatic knee osteoarthritis classification and severity grading. *Indian Journal of Orthopaedics* **58(10)**:1458–1473 DOI 10.1007/s43465-024-01259-4.

Tariq T, Suhail Z, Nawaz Z. 2023. Knee osteoarthritis detection and classification using X-rays. *IEEE Access* **11**:48292–48303 DOI 10.1109/ACCESS.2023.3276810.

Teo JC, Mohd Khairuddin I, Mohd Razman MA, Abdul Majeed APP, Mohd Isa WH. 2022. Automated detection of knee cartilage region in X-ray image. *Mekatronika—Journal of Intelligent Manufacturing & Mechatronics* **4(1)**:104–109 DOI 10.15282/mekatronika.v4i1.8627.

Wang Z, Chetouani A, Chen YH, Ru Y, Chen F, Jarraya M, Bauer F, Zhang L, Hans D, Jennane R. 2023. Confidence-driven deep learning framework for early detection of knee osteoarthritis. ArXiv DOI 10.48550/arXiv.2303.13203.

Woo S, Park J, Lee J-Y, Kweon IS. 2018. CBAM: convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer, 3–19.

Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Hum YC, Tee YK, Dhanalakshmi S. 2021. Emergence of deep learning in knee osteoarthritis diagnosis. *Computational Intelligence and Neuroscience* **2021(1)**:4931437 DOI 10.1155/2021/4931437.

Yoon JS, Yon C-J, Lee D, Lee JJ, Kang CH, Kang S-B, Lee N-K, Chang CB. 2023. Assessment of a novel deep learning-based software developed for automatic feature extraction and grading of

radiographic knee osteoarthritis. *BMC Musculoskeletal Disorders* **24(1)**:869 DOI 10.1186/s12891-023-06951-4.

Zhang D, Dong Y, Xu Y, Qian J, Ye M, Yuan Q, Luo J. 2024. Enhancing knee osteoarthritis diagnosis with DMS: a novel dense multi-scale convolutional neural network approach. *Journal of Orthopaedic Surgery and Research* **19(1)**:851 DOI 10.1186/s13018-024-05352-0.

Zhao H, Ou L, Zhang Z, Zhang L, Liu K, Kuang J. 2024. The value of deep learning-based X-ray techniques in detecting and classifying K-L grades of knee osteoarthritis: a systematic review and meta-analysis. *European Radiology* **35(1)**:327–340 DOI 10.1007/s00330-024-10928-9.

Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE* **109(5)**:820–838 DOI 10.1109/jproc.2021.3054390.

Zhou T, Ye X, Lu H, Zheng X, Qiu S, Liu Y. 2022. Dense convolutional network and its application in medical image analysis. *BioMed Research International* **2022(1)**:2384830 DOI 10.1155/2022/2384830.