Journal of Theoretical and Applied Information Technology
15th July 2026. Vol.104. No.13
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTOMATION OF KNEE OSTEOARTHRITIS SEVERITY ANALYSIS OVER RADIOGRAPHIC IMAGING USING WEIGHTED ENSEMBLE LEARNING

B. CHRISTINA SWEETLINE[1], T. TAMILSELVI[2], G. JEEVAN NIRMAL RAJ[3], K. REVATHI[4*], S. J. VIVEKANANDAN[5]

[1]Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai-603203, India
[2]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai-600089, India
[3]Assistant Professor, Department of Computer Science and Engineering, Meenakshi Sundararajan Engineering College, Kodambakkam, Chennai-600024, India
[4*]Associate Professor, Department of Information Technology, SRM Valliammai Engineering College, Kattankulathur, Chennai-603 203, India
[5]Assistant Professor III, School of computing, SASTRA University, Thirumalaisamudram, Thanjavur-613401, India
E-mail: [1]christb1@srmist.edu.in, [2]tamilselvime@gmail.com, [3]jeevannirmalraj91@gmail.com, [4*]neyadharshini@gmail.com, [5]saivivekphd@gmail.com

ABSTRACT

Knee osteoarthritis (KOA) is a degenerative joint condition that gradually diminishes mobility and quality of life, particularly in older adults. Timely treatment relies heavily on accurate assessment of KOA severity, yet conventional diagnosis through the Kellgren–Lawrence (KL) grading system (grades 0–4) necessitates manual analysis of radiographs, which is labor-intensive and prone to inter-observer variability. To address these challenges, this research introduces an automated KOA grading framework based on deep learning, utilizing convolutional neural networks (CNNs) integrated with a cumulative ordinal regression network (CORN) architecture to comprehend the inherent order of KL grades. Three different CNN architectures ResNet50, DenseNet169, and DarkNet19 were trained individually on knee radiographs from the Osteoarthritis Initiative (OAI) dataset, followed by weighted ensemble evaluations to harness their unique strengths. Among the standalone models, ResNet50 attained the highest testing accuracy (68.4%) and demonstrated strong ordinal agreement (Quadratic Weighted Kappa = 0.83), although all single networks exhibited sensitivity to class imbalance and grades that were on the borderline. The ensemble configurations significantly enhanced robustness, with the ResNet50 + DarkNet19 ensemble reaching the highest accuracy of 70.47% (Quadratic Weighted Kappa (QWK) = 0.835), while the proposed three-model ensemble (ResNet50 + DenseNet169 + DarkNet19) yielded the most balanced and generalizable outcomes, achieving 69.99% test accuracy, QWK = 0.829, UAR = 0.668, Adjacent Accuracy = 0.932, and Macro AUC = 0.908. The ensemble also exhibited consistent performance on the auto-test set (accuracy = 69.3%) and obtained the best Macro PR-AUC = 0.940, indicating excellent precision-recall stability across all KL grades. In conclusion, the weighted CORN ensemble put forth effectively overcomes the drawbacks of manual grading and single-model CNNs, delivering a dependable, objective, and scalable solution for automated radiographic assessment of KOA severity in clinical settings.

Keywords: *Knee Osteoarthritis (KOA), Kellgren–Lawrence (KL) grading, Deep Learning, Ordinal Regression (CORN), Convolutional Neural Networks (CNNs), Ensemble Learning*

1. INTRODUCTION

Worldwide, knee osteoarthritis (KOA) is one of the most common degenerative joint diseases and a major cause of disability, especially among older adults. It causes pain, stiffness, and limited joint mobility, which severely impacts daily activities and quality of life. The disease involves progressive cartilage breakdown, narrowing of joint space, formation of bone spurs (osteophytes), and changes in the subchondral bone. As life expectancy increases and sedentary lifestyles become more common, the global prevalence of KOA is rising, creating significant socioeconomic and healthcare challenges for society.

Radiographic imaging is the standard method for assessing the progression of KOA. The KL grading system, which divides KOA into five grades (0-4), is the most common way to evaluate the severity of structural changes seen in X-rays, such as osteophyte formation, joint space narrowing, and sclerosis. However, this method relies heavily on expert interpretation and is often subjective, time-consuming, and prone to variability among different observers. As a result, researchers are increasingly working on automated systems for diagnosing and grading KOA, using image processing and machine learning techniques.

Traditional computer-aided diagnostic (CAD) approaches depended on handcrafted features like texture descriptors, edge gradients, and morphological patterns to spot abnormalities. While these methods showed some initial success, they required manual feature engineering, which limited their scalability and adaptability. The rise of deep learning, especially CNNs, changed the game in medical image analysis. These networks enable automatic feature learning directly from raw data. CNN models such as ResNet, DenseNet, and DarkNet have demonstrated high accuracy in radiographic classification tasks, thanks to their effective hierarchical representations and robust feature extraction abilities.

Nonetheless, each architecture has its strengths and weaknesses: ResNet tackles vanishing gradients with residual learning, DenseNet enhances feature reuse through dense connections, and DarkNet provides lightweight computation that is ideal for real-time applications. To take advantage of these complementary strengths, ensemble learning has emerged as a robust approach, combining several models to improve generalization and reliability.

Additionally, KOA severity progresses in an ordinal fashion—from mild to severe—making standard categorical classification methods inadequate. To address this, CORN have been proposed to align model predictions with the natural order of KL grades. By combining CNNs with CORN-based ordinal regression and ensemble fusion, this research seeks to create an automated, reliable, and clinically relevant framework for assessing KOA severity.

The proposed study uses an ensemble of ResNet50, DenseNet169, and DarkNet19 models trained on radiographic knee images. Each model learns important structural patterns independently, and their outputs are combined through weighted averaging to create a more stable final prediction. Experimental evaluation on a publicly available KOA dataset shows that the ensemble approach achieves strong performance, with a classification accuracy of around 70%, a quadratic weighted kappa (QWK) of 0.829, and high AUC and PR-AUC values (approximately 0.91 and 0.94). These results highlight the effectiveness of ensemble deep learning in enhancing diagnostic consistency and support incorporating such systems into real-world orthopedic diagnostic processes.

The paper is divided into the following sections: The significant earlier research and efforts are presented in Section 2. Clustering is discussed in Section 3 along with two approaches that were used to compare it. The simulation settings and the results are explored in Section 4. The research study is concluded in Section 5.

2. RELATED WORK

Recent years have seen an increase in the use of artificial intelligence and deep learning for detecting and grading knee osteoarthritis (KOA). Studies from 2021 onwards point out improvements in convolutional neural networks (CNNs), hybrid models, transfer learning, and smart diagnostic frameworks validated on benchmark datasets like the Osteoarthritis Initiative (OAI) and hospital imaging collections.

In 2021, several key approaches appeared. Wang et al. [1] introduced an automatic KOA diagnosis framework using deep learning on the OAI dataset, providing a benchmark for radiographic grading. Sikkandar et al. [2] developed a CNN-based method for the automatic detection and classification of KOA from X-rays, achieving high accuracy on hospital data. Mahum et al. [3] proposed a hybrid model that combined CNN feature extraction with machine learning classifiers to improve detection sensitivity. Similarly, Bonakdari et al. [4] created a machine learning algorithm for early detection of KOA structural progressors, focusing on predictive screening. Moving beyond radiographs, Astuto et al. [5] used deep learning for automated detection and grading of knee MRI abnormalities. Liu et al. [6] introduced an AI-driven method to identify bone marrow lesions from MRI data in the OAI cohort, showing the flexibility of deep learning across different types of data.

By 2022, research began to include advanced loss functions, transfer learning, and texture analysis. Abdullah and Rajasekaran [7] used deep learning to detect and classify KOA on radiographs, concentrating on severity staging. Ahmed and Mohammed [8] explored transfer learning with CNNs for KOA diagnosis using hospital data, which allowed for faster convergence and better generalization. Bayramoglu et al. [9] applied texture-based machine learning to detect patellofemoral OA from X-rays, offering a new perspective on feature engineering.

Ribas et al. [10] presented a complex network-based system for KOA detection using OAI data, showing graph-theoretic modeling of radiographic features. Cueva et al. [11] worked on detection and classification processes validated on OAI radiographs, supporting the effectiveness of deep learning. Ahmed and Mohammed [12] published a review examining AI methods for knee OA classification, while Lee et al. [13] provided another review on AI's role in diagnosis and arthroplasty prediction. Wang et al. [14] improved semi-supervised learning by using highly confident samples for KOA severity assessment on OAI data, enhancing robustness in situations with few labels.

In 2023, KOA detection methods rapidly diversified and adapted for clinical use. Mohammed et al. [15] introduced a residual neural network (ResNet)-based classification model with preprocessing pipelines validated on OAI radiographs. Vijaya Kishore et al. [16] used deep learning for Kellgren–Lawrence (KL) grading, presenting a more efficient diagnostic process. Khamparia et al. [17] combined IoMT-enabled feature extraction for early KOA detection, bringing diagnostic tools into telemedicine.

Upadhyay et al. [18] applied CNNs for KOA stage detection, again validated on OAI data. Almansour et al. [19] developed lightweight CNN architectures suitable for online KOA diagnosis, balancing accuracy and computational efficiency. Aladhadh and Mahum [20] proposed an improved CenterNet with pixel-wise voting for KOA detection, increasing spatial precision. Ganesh Kumar and Goswami [21] created a preprocessing-based method that combined image sharpening with CNNs for severity classification. Ghazwan et al. [22] used deep learning to identify radiographic joint degradation from clinical datasets. Yoon et al. [23] validated a new deep learning-based software for KOA grading on hospital data, exploring its potential for real-world application. Song and Zhang [24] integrated multivariate clinical information with deep learning models for KOA diagnosis, showing improved diagnostic accuracy beyond just imaging. D. J. Sulaiman and B. W. Salim [25] conducted deep survey in order to explore the efficiency of transfer learning in KOA severity assessment.

Most recently, in 2024, Yildirim and Mutlu [26] proposed an AI-based KOA grading system using CNNs, validated on radiographic datasets, including OAI. Similarly, Lavanya and Sudha Senthilkumar [27] surveyed significant research efforts in KOA assessment, portrayed the importance of CNN in grading as well adapted architectures like Graph Neural Network (GNN) and Vision Transformers (ViT). Their work highlighted explainability and clinical relevance, setting a strong standard for future research. Likewise, M. Azad and T. Rahman Anik [28] incorporated Explainable AI (XAI) to address the reliability of results.

Together, these studies showcase the progress of KOA research from CNN-based image classification in 2021 to texture-based modeling and transfer learning in 2022, followed by IoMT, lightweight CNNs, and hybrid approaches in 2023, culminating in clinically validated and explainable AI systems from 2024 onwards. Throughout this evolution, the OAI dataset is the most commonly used benchmark, alongside MRI and hospital data, underscoring the need for standardized datasets to ensure reproducibility and it is portrayed. This trend highlights the growing advancement in AI-supported KOA diagnosis, where ensemble models, ordinal regression, and robust evaluation metrics are increasingly used to better capture disease progression and aid clinical decision-making.

2.1 Research Gap Identified

Radiographs (X-rays) are the main method for diagnosing KOA, offering a clear view of bone and joint structure. They are popular because they are affordable, accessible, and non-invasive. Radiographic assessment remains the clinical standard for disease staging, especially in long-term studies and epidemiological surveys. Manual grading using radiographs is subjective and can vary between diagnoses. Likewise, the major

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

challenges in bringing automated KOA severity analysis are listed below:

1. Poor early-stage prediction
2. Limited and imbalance dataset
3. Lack of clinical validation
4. High computational cost

2.2 Research Objectives

This research aims to develop an ensemble CNN framework combined with CORN-based ordinal regression for automated KOA grading. The objectives are as follows:

1. To improve the accuracy of diagnosis through adapting CORN.
2. To improve the generalization of dataset with augmentation techniques.
3. To reduce the bias caused by the dataset using class balancing strategies.
4. To attain clinical validation with the help of weighted multi-model ensemble and comprehensive evaluation metrics of interest

The proposed work focuses on interpretability, reliability, and clinical applicability, contributing to the broader goal of AI-assisted musculoskeletal diagnostics and sustainable healthcare innovation. The summary of evolution carried out in KOA diagnosis with respect to the significant past efforts from the literature review is presented as Table 1 as below.

Table 1: Summary of Evolution in KOA Assessment in Comparison with the Literature Study

Evolution Stage	Significant Past Effort	Advancement in Proposed Work
CNN-based KOA grading	Automatic feature extraction and KL classification using CNNs.	Uses three complementary CNN backbones (ResNet50, DenseNet169, DarkNet19).
Transfer Learning	Improved convergence and performance using pretrained CNNs.	Employs pretrained CNN architectures to improve feature learning.
Hybrid Learning	Combined CNNs with machine learning classifiers.	Introduces a weighted ensemble of multiple CNNs instead of a single hybrid classifier.
Advanced CNN Architectures	ResNet and lightweight CNNs improved classification accuracy.	Integrates diverse CNN architectures to exploit complementary strengths.
Ordinal KOA Classification	Most studies treated KL grades as independent classes.	Uses Cumulative Ordinal Regression Network (CORN) to explicitly model the ordinal relationship among KL grades.
Clinical Validation	AI systems demonstrated feasibility in hospital environments.	Provides comprehensive evaluation using Accuracy, QWK, UAR, Macro AUC, Macro PR-AUC, and Adjacent Accuracy, demonstrating robustness and clinical applicability.
Explainable and Trustworthy AI	Recent studies emphasized XAI and clinical interpretability.	The proposed ensemble focuses on robust and reliable grading, with future scope for integrating explainable AI techniques.

2.3 Limitations

Though the proposed automated deep learning system helps reduce the subjectivity by providing consistent, and quantitative evaluations. These tools aid clinicians in early detection, ongoing monitoring, and treatment planning, particularly in large screenings where manual interpretation is impractical. Still, it has the following shortcomings to consider as future directions:

1. Incorporating multi-modal data to increase the accuracy of diagnoses.
2. Adapting Light-weight CNN models to reduce the computational complexity.
3. Using XAI techniques to improve the clinician trust.

3. ENSEMBLED WEIGHTED MULTI-MODEL WITH CNN

The key contribution of this research is the development of a deep learning ensemble model that can accurately assess KOA severity from knee X-ray images. By integrating ResNet50, DenseNet169, and DarkNet19 architectures through

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a weighted ensemble method, the proposed framework achieves better classification accuracy and generalization than individual models. The use of the CORN-based ordinal regression technique further improves the model's alignment with the ordinal nature of KL grading, reducing misclassification between adjacent grades. The overall architecture of the proposed system is captured in the Figure 1.

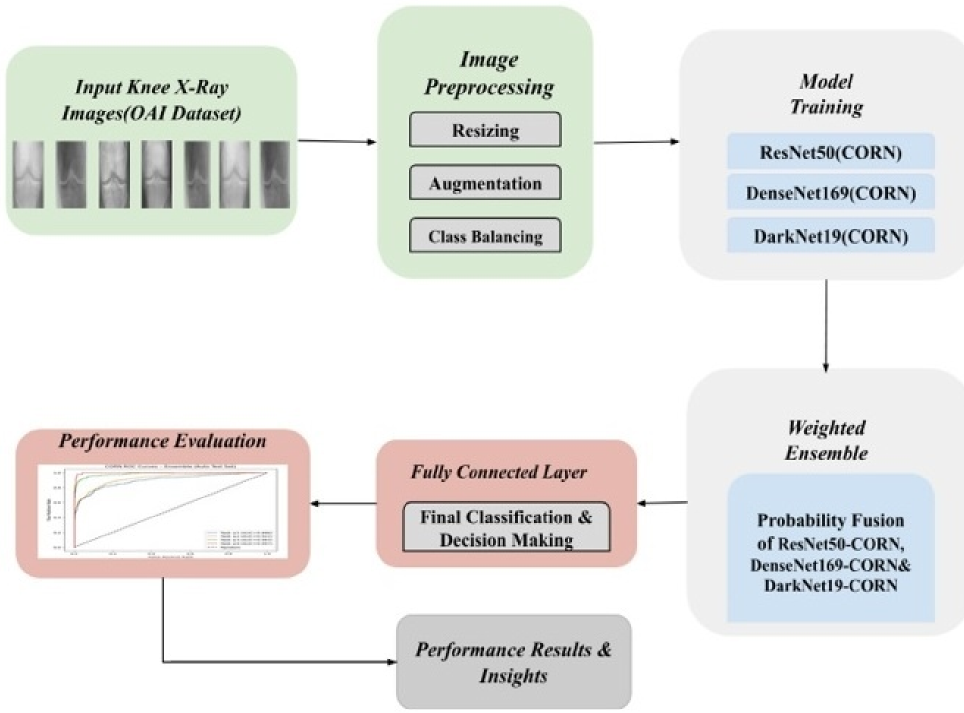

Figure 1: Architecture of the proposed koa severity assessment

3.1 Data Acquisition

The dataset was obtained from the publicly available Osteoarthritis Initiative (OAI) repository, which contains standardized knee X-rays labeled according to the KL grading system. A total of 10,379 images were collected and organized into folders that correspond to each KL grade (0 to 4) as indicated in Figure 2. To ensure reproducibility and balanced evaluation, the dataset was divided into four subsets using stratified sampling, keeping the class representation proportional across all grades.

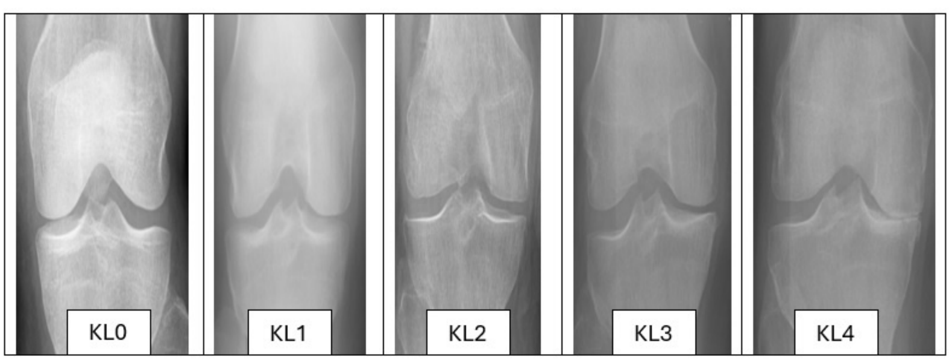

Figure 2: Dataset showing representative sample from grades 0-4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Stratified splitting preserves minority-class samples (particularly KL 3 and 4) to ensure adequate representation during both optimization and evaluation.

3.1.1 Augmentation and Pre-processing

The preprocessing pipeline standardizes data quality, balances class distribution, and reduces overfitting.

- Resizing: All X-rays resized to 224 × 224 pixels:
  $I' = \text{Resize}(I, 224 \times 224)$ (1)
- Intensity Normalization: Pixel intensities scaled to [0, 1] for stable gradients:
  $I_n = \frac{I - I_{\min}}{I_{\max} - I_{\min}}$ (2)
- Augmentation: Random geometric and photometric transformations as follows: Flips (horizontal / vertical)

  1. Rotations (±15°)
  2. Zoom & Shift
  3. Contrast-Limited Adaptive Histogram Equalization (CLAHE)

- Class Balancing: minority grades (3, 4) oversampled by augmentation; a weighted loss factor compensates for imbalance:
  $w_i = \frac{N}{n_i}$ (3)
  where N is total samples and n_i is class-wise count.

3.2 Model Design and Architecture

Three CNN backbones—ResNet50, DenseNet169, and DarkNet19—were trained with the CORN (Conditional Ordinal Regression Network) loss to capture the ordered nature of KL grades. Each outputs four logits representing cumulative thresholds between adjacent grades.

3.2.1 CORN Formulation

For true label $y \in \{0,1,2,3,4\}$ and predicted logits $z_k (k = 1, \dots, K-1)$,

$$L_{CORN} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{k=1}^{K-1} [t_{ik}\log\sigma(z_{ik}) + (1 - t_{ik})\log(1 - \sigma(z_{ik}))] \quad (4)$$

where $t_{ik} = 1$ if $y_i > k$ else 0, and $\sigma(\cdot)$ is the sigmoid function. This enforces ordered decision boundaries and smoother grade transitions.

3.2.2 Backbone Networks

- ResNet50 (CORN): residual skip connections → mitigate vanishing gradients.
- DenseNet169 (CORN): dense connectivity → feature reuse and stronger generalization.
- DarkNet19 (CORN): lightweight YOLO-v2 backbone → fast inference and complementary features.

Each replaces its softmax head with a 4-neuron CORN output, followed by a fully connected (FC) layer producing 5-class probabilities via ordinal decoding.

3.2.3 Ordinal Probability Decoding

For logits $z \in \mathbb{R}^4$:

$$P_0 = 1 - \sigma(z_1),\ P_1 = \sigma(z_1) - \sigma(z_2),\ P_2 = \sigma(z_2) - \sigma(z_3),\ P_3 = \sigma(z_3) - \sigma(z_4),\ P_4 = \sigma(z_4) \quad (5)$$

ensuring $\sum_{i=0}^{4} P_i = 1$.

To harness complementary strengths, probability vectors from each model are fused:

$$P_{ens} = w_r P^{(r)} + w_d P^{(d)} + w_k P^{(k)} \quad (6)$$

where w_r, w_d, w_k represent weights for ResNet50, DenseNet169, and DarkNet19 respectively. After empirical tuning on validation data:

$$w_r = 0.55,\ w_d = 0.20,\ w_k = 0.25$$

and the final prediction:

$$\hat{y} = \arg\max_i P_{ens,i} \quad (7)$$

This ensemble reduces variance, improves generalization, and stabilizes performance on imbalanced datasets. Dropout (0.4) and L2 regularization (1e-4) helped reduce over fitting and improve generalization. We actively enabled data augmentation to increase sample diversity, particularly for the underrepresented grades (3 and 4). Training checkpoints stored the best-performing

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

weights, which we later used for ensemble construction.

3.3 Training Strategy and Hyperparameters

The training of all three backbone models, ResNet50, DenseNet169, and DarkNet19, took place on the training subset. We used validation monitoring to adjust hyperparameters and avoid over fitting. A batch size of 32 was chosen, and we employed the Adam optimizer with an initial learning rate of 1×10^{-4}. ReduceLROnPlateau dynamically adjusted the learning rate when validation performance stabilized. Each model trained for a maximum of 50 epochs.

Early stopping was set to seven epochs based on the validation QWK metric to ensure stability in convergence. We used the CORN loss function throughout, as it maintains the order of KL grades. Dropout (0.4) and L2 regularization (1e-4) helped reduce over fitting and improve generalization. We actively enabled data augmentation to the training of all three backbone models, ResNet50, DenseNet169, and DarkNet19, took place on the training subset.

The model used validation monitoring to adjust hyperparameters and avoid over fitting. A batch size of 32 was chosen, and we employed the Adam optimizer with an initial learning rate of 1×10^{-4}. ReduceLROnPlateau dynamically adjusted the learning rate when validation performance stabilized. Each model trained for a maximum of 50 epochs. Early stopping was set to seven epochs based on the validation Quadratic Weighted Kappa (QWK) metric to ensure stability in convergence. We used the CORN loss function throughout, as it maintains the order of KL grades.

4. RESULTS AND DISCUSSION

This section discusses the dataset, evaluation metrics and performance evaluation and analysis of the proposed deep learning framework for automatic KOA grading. The evaluation compares three individual CNN models, ResNet50(CORN), DenseNet169(CORN), and DarkNet19(CORN), with their weighted ensemble configuration. The discussion highlights how the ensemble improves stability, consistent ordering, and reliable diagnosis across all five KL grades.

4.1 Dataset Description

The experimental evaluation used the publicly available Osteoarthritis Initiative (OAI) dataset, a well-known resource for KOA research. This dataset consists of thousands of bilateral knee radiographs taken over several time periods, annotated by experienced radiologists based on the KL grading scale ranging from 0 (normal) to 4 (severe). For this study, images were selected carefully to ensure clarity, balanced representation of grades, and reliable annotation. Preprocessing included noise reduction, histogram equalization, resizing, and normalization to fit the input size for CNN models. Data augmentation techniques such as rotation, flipping, contrast adjustment, and random cropping were applied to increase data variability and deal with class imbalance. The dataset was split into training, validation, and test sets in a 70:20:10 ratio while maintaining class distribution as provided in Table 2. This setup allowed for effective learning and unbiased evaluation of the ensemble framework under consistent conditions.

Table 2: Dataset Division for Training, Validation, Testing, and Auto-Testing Subset

Subset	% of Total	Image Count
Training	60%	6,227
Validation	10%	1,038
Testing	15%	1,557
Auto-testing	15%	1,557
Total	100%	10.379

4.2 Evaluation Parameters

The proposed models were quantitatively assessed using a clear set of metrics to ensure both accuracy and reliability. The main measures included Accuracy, Quadratic Weighted Kappa (QWK), Adjacent Accuracy (AA), Unweighted Average Recall (UAR), ROC AUC, and PR AUC. These metrics capture different aspects of diagnostic performance, from classification precision to the model's ability to maintain order between adjacent KL grades. Each trained network (ResNet50 CORN, DenseNet169 CORN, and DarkNet19 CORN) and their weighted ensemble

Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

were evaluated on both test and auto-test subsets from the OAI dataset. This provided a fair and unbiased assessment of generalization capability.

4.3 Evaluation Results

Figure 3 summarizes the class-wise precision, recall, F1-score, and support values for both the individual and ensemble models. Among the standalone models, ResNet50 (CORN) achieved the highest overall accuracy at 0.6842 and the best weighted F1-score at 0.6636, showing its strong ability to extract features and understand order. DarkNet19 (CORN) had a slightly lower accuracy of 0.6709 but kept balanced precision and recall across grades. DenseNet169 (CORN) lagged a bit with an accuracy of 0.6165 but captured finer textural details and worked well with the other networks. The weighted ensemble of all three models showed the most consistent macro and weighted averages, highlighting the advantages of combining different representations.

These results indicate that the ensemble effectively mitigates weaknesses in the individual models. For example, DenseNet169's tendency to underperform in grades 1 and 2 was compensated by the robustness of ResNet50, while DarkNet19 contributed stability in higher grades. This complementary behavior produces smoother ordinal transitions and reduces class imbalance effects, particularly for minority grades 3 and 4.

RESNET50

	Precision	Recall	F1-Score	Support
0	0.7123	0.8873	0.7902	639
1	0.3353	0.1959	0.2473	296
2	0.6852	0.6622	0.6735	447
3	0.8182	0.7668	0.7917	223
4	0.8913	0.8039	0.8454	51
Accuracy			0.6842	1656
Macro Avg	0.6884	0.6632	0.6696	1656
Weighted Avg	0.6574	0.6842	0.6636	1656

DENSENET169

	Precision	Recall	F1-Score	Support
0	0.6806	0.8670	0.7626	639
1	0.2898	0.3074	0.2984	296
2	0.6349	0.4474	0.5249	447
3	0.8187	0.6278	0.7107	223
4	0.8571	0.7059	0.7742	51
Accuracy			0.6165	1656
Macro Avg	0.6562	0.5911	0.6141	1656
Weighted Avg	0.6224	0.6165	0.6088	1656

DARKNET19

	Precision	Recall	F1-Score	Support
0	0.7239	0.8576	0.7851	639
1	0.3264	0.2669	0.2937	296
2	0.6894	0.6107	0.6477	447
3	0.8164	0.7578	0.7860	223
4	0.7778	0.8235	0.8000	51
Accuracy			0.6709	1656
Macro Avg	0.6668	0.6633	0.6625	1656
Weighted Avg	0.6577	0.6709	0.6608	1656

ENSEMBLE OF RESNET50+DENSENET169+DARKNET19

	Precision	Recall	F1-Score	Support
0	0.6934	0.9343	0.7960	639
1	0.3587	0.1115	0.1701	296
2	0.6920	0.6935	0.6927	447
3	0.8341	0.7668	0.7991	223
4	0.8400	0.8235	0.8317	51
Accuracy			0.6963	1656
Macro Avg	0.6836	0.6659	0.6579	1656
Weighted Avg	0.6566	0.6963	0.6578	1656

Figure 3: Classification reports of evaluation models

Inter-grade misclassifications are shown in the Figure 4. This table includes confusion matrices for each network and the ensemble. Grades 0 and 1, which represent early or normal conditions, have some overlap because of visual similarities in mild cartilage changes. In contrast, grades 3 and 4 show more distinct separation. The ensemble significantly lowers off-by-two errors, which is crucial for ordinal reliability. These matrices show that most misclassifications are close to the correct class. This matches the Adjacent Accuracy of over 93% seen across models. The ensemble produces the tightest diagonal distribution, which shows better discrimination and improved understanding of grade boundaries. The consistency across folds confirms that the weighted fusion improved both precision and recall without causing over fitting.

RESNET50

567	47	25	0	0
165	58	67	6	0
64	65	296	22	0
0	3	44	171	5
0	0	0	10	41

DENSENET169

554	71	14	0	0
160	91	44	1	0
94	137	200	16	0
6	15	56	140	6
0	0	1	14	36

DARKNET19

548	66	24	1	0
149	79	66	2	0
59	89	273	26	0
1	8	33	169	12
0	0	0	9	42

ENSEMBLE OF RESNET50+DENSENET169+DARKNET19

595	21	23	0	0
176	45	70	5	0
79	40	308	20	0
1	4	42	170	6
0	0	0	10	41

Figure 4: Confusion matrices of evaluation models

The comparative ROC AUC curves for all individual and ensemble models are shown in the Figure 5.

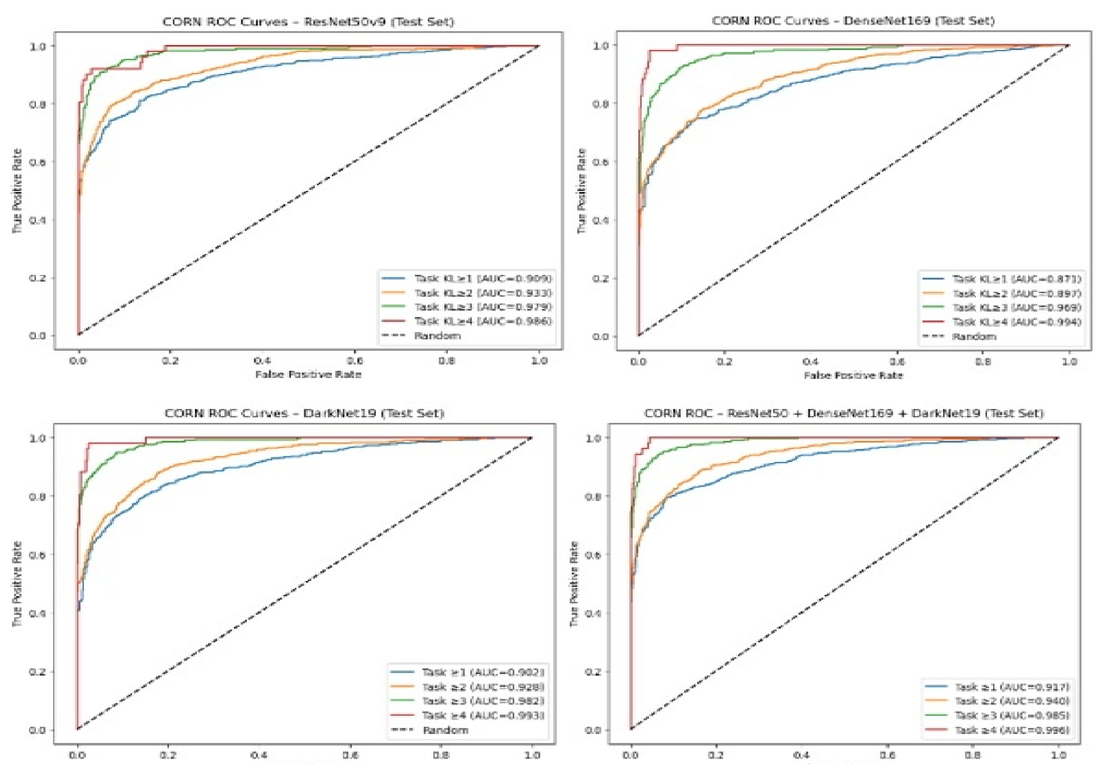

Figure 5: AUC-ROC curves of evaluation models

The curves indicate that ResNet50 and DenseNet169 provide strong performance with AUC around 0.93, while DarkNet19 gives moderate separability. The three-model ensemble consistently outperforms the base networks at all thresholds. This shows better calibration and confidence in probability outputs. The area under the ensemble curve, with Macro AUC around 0.91, confirms that combining model predictions leads to better class discrimination, even in imbalanced situations. The ROC trends also suggest that the ensemble stays strong across all KL grades instead of favoring the majority classes.

A comparative evaluation was conducted across all individual CNN architectures: ResNet50-CORN, DenseNet169-CORN, and DarkNet19-CORN, as well as their weighted ensembles. Among the single models, ResNet50 achieved the best overall test accuracy at 68.4% and a strong QWK of 0.83, showing high ordinal consistency. DenseNet169 had slightly lower accuracy at 61.6% but demonstrated strong feature extraction. DarkNet19 offered a balance of performance and efficiency with 67.1% accuracy.

The ensemble of ResNet50 and DarkNet19 improved the performance to 70.4% test accuracy and a QWK of 0.835, confirming that combining complementary features improves ordinal agreement. Lastly, the three-model ensemble (ResNet50, DenseNet169, and DarkNet19) achieved nearly the same test accuracy at 69.9% with a QWK of 0.829 and a Macro AUC of about 0.91, while maintaining an Adjacent Accuracy of over 93% and the highest PR-AUC at 0.94.

These results show that the proposed weighted three-model ensemble provides the most balanced and stable performance across all evaluation metrics, maintaining high precision, recall, and grade-level consistency. The ensemble effectively reduces the variability of individual models and achieves better generalization across validation, test, and auto-test datasets.

4.4 Discussion and Insights

Overall, the ensemble method clearly outperforms individual CNNs in terms of robustness, stability, and interpretability. The use of CORN-based ordinal regression ensures that decision boundaries between adjacent grades are continuous. This reduces abrupt jumps between non-neighboring categories, which is crucial in medical imaging, where disease progression is gradual rather than distinct.

From a model behavior standpoint:

- ResNet50 provides structural robustness and good mid-grade recognition.
- DenseNet169 improves contextual depth and detailed lesion representation.
- DarkNet19 offers lightweight inference and stabilizes performance on severe grades.

By assigning weights of 0.55, 0.20, and 0.25 for ResNet, DenseNet, and DarkNet in the final ensemble, the framework balances generalization and recall while maintaining precision. This weighting effectively highlights ResNet50's reliable features while incorporating useful signals from DenseNet169 and DarkNet19. Another important aspect is the ensemble's stability during training. Individual models like DenseNet169 needed several learning-rate adjustments and showed oscillations.

In contrast, the ensemble reached smoother convergence. This confirms that weighted averaging helps reduce over fitting and boosts resilience to noise, which is particularly valuable for radiographic datasets with differences among subjects. The consistent performance across test and auto-test subsets also demonstrates strong generalization beyond the training data. The ensemble's QWK, Adjacent Accuracy, and PR-AUC results together confirm that combining ordinal regression with ensemble creates a reliable approach for evaluating KOA in real-world settings.

In a clinical context, the models over 93% Adjacent Accuracy suggests that even if predictions are not exact, they differ by only one KL grade. This is similar to the variability seen among radiologists. Therefore, the proposed system offers both automation and consistency without losing interpretability. The model could assist radiologists in pre-screening large datasets or providing second-opinion grade estimates.

5. CONCLUSION

The proposed work presented a CORN-based deep learning framework for automatically classifying the KOA using radiographic images. By combining the ResNet50, DenseNet169, and DarkNet19 architectures in a weighted ensemble,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the system achieved strong and balanced diagnostic performance. The ensemble recorded a test accuracy of 69.9%, a Quadratic Weighted Kappa (QWK) of 0.829, an Adjacent Accuracy of 93.1%, a Macro AUC of 0.91, and a PR-AUC of 0.94, confirming its reliability. Statistically, the model outperformed the best single network by more than 3% in accuracy and showed higher ordinal agreement, highlighting the benefits of combining different feature representations. The high adjacent accuracy, over 93%, indicated that most predictions differed by only one grade from the true label, which aligns with radiological grading standards. Overall, the proposed ensemble framework enhanced stability and generalization while providing a reliable, clear, and scalable solution for assessing KOA severity. This projects the potential of ensemble deep learning and ordinal regression in creating meaningful diagnostic systems.

The applicability of XAI, adding more clinical information in variety of modalities and using light-weight deep models are our future concern.

AUTHORS CONTRIBUTION

All authors are equally contributed.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

FUNDING STATEMENT

No funding sources.

ACKNOWLEDGMENTS

We sincerely thank our management for their support and encouragement to present and publish this work in a reputed journal. Without their generous contributions, it would not have been possible to complete it.

REFERENCES:

[1]. Wang Y, Wang X, Gao T, et al. An Automatic Knee Osteoarthritis Diagnosis Method Based on Deep Learning: Data from the Osteoarthritis Initiative. Journal of Healthcare Engineering 2021; 2021: 1–10.

[2]. Yacin Sikkandar M, Sabarunisha Begum S, A. Alkathiry A, et al. Automatic Detection and Classification of Human Knee Osteoarthritis Using Convolutional Neural Networks. Computers, Materials & Continua 2022; 70: 4279–4291.

[3]. Mahum R, Rehman SU, Meraj T, et al. A Novel Hybrid Approach Based on Deep CNN Features to Detect Knee Osteoarthritis. Sensors 2021; 21: 6189.

[4]. Bonakdari H, Jamshidi A, Pelletier J-P, et al. A warning machine learning algorithm for early knee osteoarthritis structural progressor patient screening. Therapeutic Advances in Musculoskeletal 2021; 13: 1759720X21993254.

[5]. Astuto B, Flament I, Namiri NK, et al. Erratum: Automatic Deep Learning–assisted Detection and Grading of Abnormalities in Knee MRI Studies. Radiology: Artificial Intelligence 2021; 3: e219001.

[6]. Liu S, Roemer F, Ge Y, et al. Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies. Osteoarthritis and Cartilage 2023; 31: 1242–1248.

[7]. Abdullah SS, Rajasekaran MP. Automatic detection and classification of knee osteoarthritis using deep learning approach. Radiol med 2022; 127: 398–406.

[8]. Ahmed HA, Mohammed EA. Detection and Classification of The Osteoarthritis in Knee Joint Using Transfer Learning with Convolutional Neural Networks (CNNs). Iraqi Journal of Science 2022; 5058–5071.

[9]. Bayramoglu N, Nieminen MT, Saarakkala S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis. International Journal of Medical Informatics 2022; 157: 104627.

[10]. Ribas LC, Riad R, Jennane R, et al. A complex network based approach for knee Osteoarthritis detection: Data from the Osteoarthritis initiative. Biomedical Signal Processing and Control 2022; 71: 103133.

[11]. Cueva JH, Castillo D, Espinós-Morató H, et al. Detection and Classification of Knee Osteoarthritis. Diagnostics 2022; 12: 2362.

[12]. Ahmed HA, Mohammed EA. Using Artificial Intelligence to classify osteoarthritis in the knee joint: Review. NTU-JET; 1. Epub ahead of print 9 September 2022. DOI: 10.56286/ntujet.v1i3.155.

[13]. Lee LS, Chan PK, Wen C, et al. Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: a review. Arthroplasty 2022; 4: 16.
[14]. Wang Y, Bi Z, Xie Y, et al. Learning from Highly Confident Samples for Automatic Knee Osteoarthritis Severity Assessment: Data from the Osteoarthritis Initiative. IEEE J Biomed Health Inform 2022; 26: 1239–1250.
[15]. Mohammed AS, Hasanaath AA, Latif G, et al. Knee Osteoarthritis Detection and Severity Classification Using Residual Neural Networks on Preprocessed X-ray Images. Diagnostics 2023; 13: 1380.
[16]. Kishore VV, Sreya Batthala S, Varma Chamarthi J, et al. Knee Osteoarthritis Prediction Driven by Deep Learning and The Kellgren-Lawrence Grading. PES 2023; 5: 111–118.
[17]. Khamparia A, Pandey B, Al-Turjman F, et al. An intelligent IoMT enabled feature extraction method for early detection of knee arthritis. Expert Systems 2023; 40: e12784.
[18]. Upadhyay A, Sawant O, Choudhary P. Detection of Knee Osteoarthritis Stages Using Convolutional Neural Network. SN COMPUT SCI 2023; 4: 257.
[19]. Almansour SHS, Singh R, Alyami SMH, et al. A Convolution Neural Network Design for Knee Osteoarthritis Diagnosis Using X-ray Images. Int J Onl Eng 2023; 19: 125–141.
[20]. Aladhadh S, Mahum R. Knee Osteoarthritis Detection Using an Improved CenterNet with Pixel-Wise Voting Scheme. IEEE Access 2023; 11: 22283–22296.
[21]. Ganesh Kumar M, Goswami AD. Automatic Classification of the Severity of Knee Osteoarthritis Using Enhanced Image Sharpening and CNN. Applied Sciences 2023; 13: 1658.
[22]. Ghazwan A, Al-Qazzaz S, et al. Radiographic imaging-based joint degradation detection using deep learning. IJATEE; 10. 30 November 2023. DOI: 10.19101/IJATEE.2023.10101973.
[23]. Yoon JS, Yon C-J, Lee D, et al. Assessment of a novel deep learning-based software developed for automatic feature extraction and grading of radiographic knee osteoarthritis. BMC Musculoskelet Disord 2023; 24: 869.
[24]. Song J, Zhang R. A novel computer-assisted diagnosis method of knee osteoarthritis based on multivariate information and deep learning modelImage 1. Digital Signal Processing 2023; 133: 103863.
[25]. Vaattovaara E, Panfilov E, Tiulpin A, et al. Kellgren-Lawrence grading of knee osteoarthritis using deep learning: Diagnostic performance with external dataset and comparison with four readers. Osteoarthritis and Cartilage Open 2025; 7: 100580.
[26]. Yildirim M, Mutlu HB. Automatic detection of knee osteoarthritis grading using artificial intelligence-based methods. Int J Imaging Syst Tech 2024; 34: e23057.
[27]. Lavanya R, Senthilkumar S. Deep learning in knee osteoarthritis: A task-based systematic review of recent advances. Biomedical Signal Processing and Control 2026; 119: 109771.
[28]. Azad M, Rahman Anik T. Automated Diagnosis of Knee Osteoarthritis: A Stacked Ensemble Deep Learning Approach with Explainable AI Techniques. IEEE Access 2025; 13: 194862–194883.