KNEE OSTEOARTHRITIS CLASSIFICATION USING DEEP LEARNING

Dr.K.S.Raja sekhar[1],N.Vyshnavi[2],V.Charan sai[2],D.Venkata Rama krishna Raju [2]

Assistant Professor [1], Students [2]

Department of ECE, ANU College of Engineering& Technology, Guntur, AP

Abstract:

Knee Osteoarthritis (OA) is a progressive joint disorder that leads to cartilage degradation, causing pain and reduced mobility. Early detection and accurate classification of OA severity using radiographic images are crucial for timely intervention and treatment planning. In this study, we develop a deep learning-based approach for automatic classification of knee OA severity from X-ray images using the Kellgren Lawrence (KL) grading system. We proposed preprocessing X-ray images through contrast enhancement, bone segmentation, and data augmentation to improve model robustness. Convolutional Neural Networks (CNNs) such as ResNet50, Efficient Net, and Vision Transformers (ViT) for feature extraction and multi-class classification. To address class imbalance, we implement Focal Loss and weighted sampling strategies. The model is trained and evaluated using publicly available datasets such as the Osteoarthritis Initiative (OAI) and KneeKL Dataset, achieving high accuracy and a strong agreement with clinical assessments, as measured by the Quadratic Weighted Kappa (QWK) score.

Keywords: Knee Osteoarthritis, Deep Learning, X-ray Classification, Convolutional Neural Networks, Vision Transformers, Medical Imaging.

1.Introduction

Knee Osteoarthritis (KOA) is a degenerative joint disease characterized by the gradual deterioration of cartilage in the knee joint, leading to stiffness, chronic pain, and reduced mobility. It is especially prevalent in the elderly population and is considered one of the most common causes of disability worldwide. Early detection and accurate classification of KOA severity are essential for timely medical intervention, personalized treatment, and improving patients' quality of life.

Traditionally, radiographic analysis using the Kellgren–Lawrence (KL) grading system has been the gold standard for evaluating KOA severity. This system categorizes the condition into five grades (0–4), ranging from normal to severe osteoarthritis. However, the KL grading process is manual, subjective, and prone to inter-observer variability, especially in borderline cases like grades 1 and 2, where visual differences are subtle.

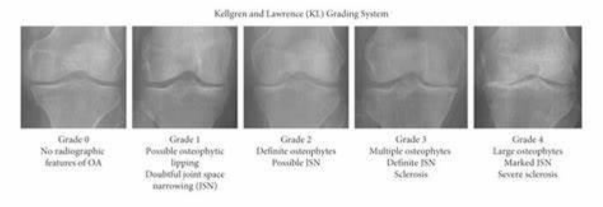
To address these limitations, we have developed a deep learning-based system that automates the classification of knee osteoarthritis severity from X-ray images. In our project, we used publicly available datasets—the Osteoarthritis Initiative (OAI) and KneeKL—that contain pre-labelled knee radiographs with verified KL grades. These datasets served as the foundation for training and evaluating our models.

We began by preprocessing the images through steps like contrast enhancement, noise reduction, and bone segmentation to isolate relevant features. This step was crucial in improving the quality of the data fed into the model. To improve model generalization and reduce overfitting, we also applied data augmentation techniques such as rotation, scaling, and horizontal flipping.

For feature extraction and classification, we implemented multiple deep learning architectures, including ResNet50, Efficient Net, and Vision Transformers (ViT). These models were initially pre-trained on ImageNet and later fine-tuned on our knee OA dataset. To handle the challenge of class imbalance in the KL grades, we incorporated Focal Loss and weighted sampling strategies during training.

By combining medical domain knowledge with cutting-edge AI techniques, our system achieved a significant improvement in classification performance. We measured accuracy and Quadratic Weighted Kappa (QWK) scores to assess the agreement with clinical labels, and utilized Grad-CAM visualizations to confirm that the models were focusing on the correct anatomical structures of the knee joint.

In essence, the purposed approach presents a robust, reproducible, and scalable solution to automate KOA severity classification, aiming to support radiologists and clinicians in making faster and more accurate diagnoses.



2.Literature Review

Studies Using Single CNN Models:

Early work often focused on binary classification (e.g., OA vs. non-OA) or distinguishing severe OA (KL 3-4) from mild/no OA (KL 0-2).

Tuplin et al. [10] were among the pioneers, using a CNN trained on OAI data for KL grade classification, achieving performance comparable to human experts for certain tasks. They also explored assessing specific OA risk factors.

Antony et al. [23] developed a CNN system demonstrating high accuracy for KL grading, also using the OAI dataset, and highlighted the potential for computer-aided diagnosis. Various CNN architectures have been explored, including custom-designed networks, VGG-style networks, ResNets, and Dense Nets, often utilizing transfer learning from ImageNet pre-trained models [24]. Some studies focused on predicting individual radiographic features (JSN, osteophytes) separately before combining them for a KL grade prediction [25].

Studies Using Ensemble Methods:

Recognizing the potential benefits of combining multiple models, some researchers have explored ensemble techniques for knee OA grading. Ensembles can involve combining predictions from the same architecture trained with different initializations or on different data subsets (bagging). More sophisticated ensembles combine predictions from fundamentally different architectures, potentially capturing different aspects of the data, similar to the approach in this study. Leung et al. [26] used an ensemble of CNNs for improved prediction of OA progression.

However, ensembles combining diverse modern architectures like Efficient Nets and Vision Transformers specifically for KL grading are less common in the literature, representing an area for potential improvement explored in this work.

Use of the OAI Dataset in Previous Work:

The Osteoarthritis Initiative (OAI) dataset has been instrumental in advancing automated OA research. Its large scale, longitudinal nature, standardized protocols, and publicly available KL grades make it an invaluable resource [27].

A significant portion of the published DL work on knee OA KL grading utilizes OAI data, making it a benchmark dataset for comparing different models [10,11,12].. Consistency in using this dataset allows for more direct comparison of algorithmic performance.

Interpretability in Medical AI:

While high accuracy is essential, the "black box" nature of deep learning models can be a barrier to clinical adoption. Clinicians need to trust that the AI system is making decisions based on sound reasoning. Techniques like saliency maps, Class Activation Mapping (CAM) [28], and Grad-CAM [12] have been developed to visualize which parts of the input image contribute most to a model's prediction. Several knee OA studies have incorporated such methods to show that their models focus on relevant joint regions (e.g., medial/lateral compartments, tibial spines, femoral condyles) when predicting KL grades, increasing confidence in the models' validity [10, 23].

Relevance to Proposed Work

Traditional Machine Learning Approaches for OA Assessment:

Brief overview of earlier methods using handcrafted features (e.g., texture analysis, edge detection, morphological features) combined with classifiers like SVM, Random Forests [11].

Deep Learning (CNNs) for Knee OA KL Grading:

Pioneering studies using CNNs (e.g., AlexNet, VGG, GoogleNet) for KL classification [12,13].

Studies utilizing deeper architectures like ResNet and DenseNet, often showing improved performance [14, 15]. Work specifically using EfficientNet for medical imaging tasks, potentially including OA [16]. Discussion of typical findings: DL models often achieving radiologist-level performance or higher in specific tasks, significant reduction in variability.

Mention of challenges: Need for large datasets, class imbalance issues (fewer samples for extreme grades), domain shift when applying models to new datasets.

Vision Transformers (ViT) in Medical Imaging:

Introduction to the success of Transformers in NLP and their adaptation to vision tasks (ViT). Applications of ViT in various medica; imaging domains (e.g., pathology, radiology) [17, 18]. Potential advantages for OA: Ability to model long-range dependencies, potentially capturing subtle global changes in joint structure.

Ensemble Methods in Medical Image Analysis:

Theoretical basis for ensemble: Reducing variance, improving robustness against noise or variations in data Examples of ensemble techniques used in medical imaging: Averaging, majority voting, stacking [19].Studies employing ensembles for improved classification or segmentation in various medical tasks, potentially including OA diagnosis or grading [20,21]. Highlighting the benefit of ensembling diverse models (like CNNs and ViT) compared to ensembling similar models.

Explainable AI (XAI) in Medical Deep Learning

Importance of model interpretability in healthcare for trust, debugging, and clinical acceptance. Overview of common XAI techniques: Saliency maps, CAM, Grad-CAM, SHAP [10,22]. Studies using Grad-CAM or similar methods to validate knee OA models, showing they focus on relevant features like JSN or osteophytes [14, 23].

Research Gap and Contribution Summary

Reiterate the limitations of existing single-model approaches or ensembles of similar models.

Emphasize the novelty of combining state-of-the-art CNNs (EfficientNet, ResNet50) with a Vision Transformer (ViT) in an ensemble framework specifically for KL grading on the large OAI dataset.

3. Problem Statement

Diagnosing Knee Osteoarthritis (KOA) using radiographic images traditionally relies on the Kellgren–Lawrence (KL) grading system, which, while widely accepted, is inherently subjective. Radiologists must interpret subtle differences in joint space narrowing, osteophyte formation, and sclerosis—leading to inconsistent results, especially in the early stages of the disease (grades 0 and 1), where visual cues are often minimal.

Manual classification suffers from:

- Observer variability between clinicians,
- Time-consuming assessments, particularly in large-scale screenings,
- Limited scalability for real-time or remote diagnosis, and
- Difficulty in detecting early-stage OA, which is critical for preventing progression.

In addition, many existing automated systems are built using single-model architectures that may not generalize well across diverse patient data. Furthermore, imbalanced datasets, where certain KL grades are underrepresented, lead to biased models that perform poorly in real-world applications.

To overcome these challenges, our project focuses on the following key objectives:

Automate the classification of knee OA severity using deep learning models trained on X-ray images, leverage an ensemble of CNNs and Vision Transformers (ViT) to improve accuracy and robustness, Preprocess images using contrast enhancement, segmentation, and augmentation to improve feature visibility and generalization, Handle class imbalance effectively using Focal Loss and weighted sampling, and

Provide interpretable results through Grad-CAM to ensure the model's decision-making aligns with medical reasoning. Our goal is to develop a reliable, fast, and interpretable KOA classification system that can assist clinicians, reduce diagnostic variability, and ultimately support early intervention strategies.

4.METHODOLOGY:

DATASET:

1. Osteoarthritis Initiative (OAI):

The data used in this study were obtained from the Osteoarthritis Initiative (OAI) database. The OAI is a multi-centre, longitudinal, prospective observational study of knee osteoarthritis sponsored by the National Institutes of Health (NIH).The dataset includes extensive clinical evaluation data, questionnaires, and imaging data (X-ray and MRI) collected over several years.

2. Data Selection and Cohort:

For this study, we utilized the bilateral posteroanterior (PA) fixed-flexion weight-bearing knee X-ray images. This view is standard for assessing joint space width and OA features. We included images from [Specify OAI time points used, e.g., baseline visit, 24-month visit].

Each knee joint in an image was treated as a separate sample. [Mention if automated or manual cropping of the knee joint region was performed, or if the entire image was used]. If cropped, describe the method (e.g., using coordinates provided by OAI, or an automated detection algorithm). The final dataset comprised [Number] knee X-ray images.

DataPreprocessing

Proper preprocessing is crucial for optimal deep learning model performance.

3.1. Image Selection and ROI Extraction (Optional)

If knee joint regions were extracted: Describe the process (e.g., "Knee joints were automatically localized using a pre-trained YOLOv5 object detector fine-tuned on knee bounding boxes, followed by cropping"). If full images were used: State this ("The full PA radiographs were used as input").

3.2. Image Resizing and Normalization

All images (or cropped ROIs) were resized to a uniform input size required by the deep learning models (e.g., 224x224 pixels for ResNet50/ViT, or a size appropriate for the EfficientNet variant, e.g., 224x224, 300x300). Bicubic interpolation was used for resizing. Pixel values were normalized. Common strategies include: Scaling pixel values to the range [0, 1]. Standardizing by subtracting the mean and dividing by the standard deviation of the training set (or ImageNet statistics if using pre-trained models). Specify the method used (e.g., "Images were normalized using ImageNet mean and standard deviation").

3.3. Data Augmentation

To increase the diversity of the training data, prevent overfitting, and improve model generalization, extensive data augmentation techniques were applied *only* to the training set during training.

Augmentations included:

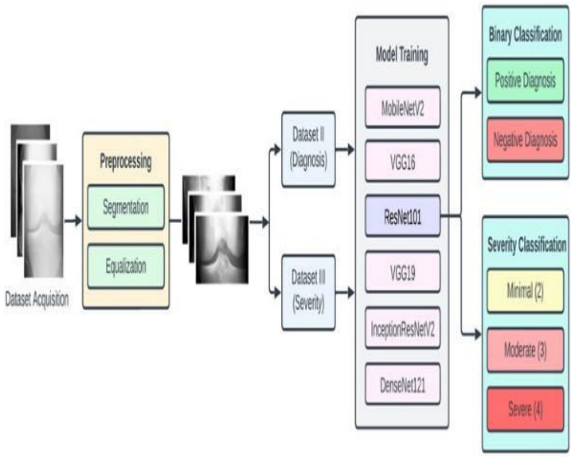
- Random horizontal flipping
- Random rotations (e.g., +/- 10 degrees).
- Random translation (shifting) (e.g., +/- 10% of image width/height).
- Random scaling (zooming) (e.g., +/- 10%).
- Random brightness and contrast adjustments (e.g., +/- 20%).

(Note: Avoid augmentations that might obscure OA features, like excessive blurring or vertical flipping if laterality matters and isn't handled otherwise).

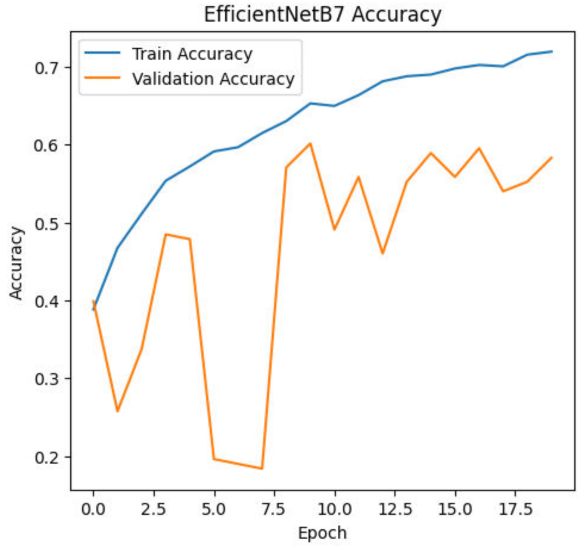
3.4.DataSplitting

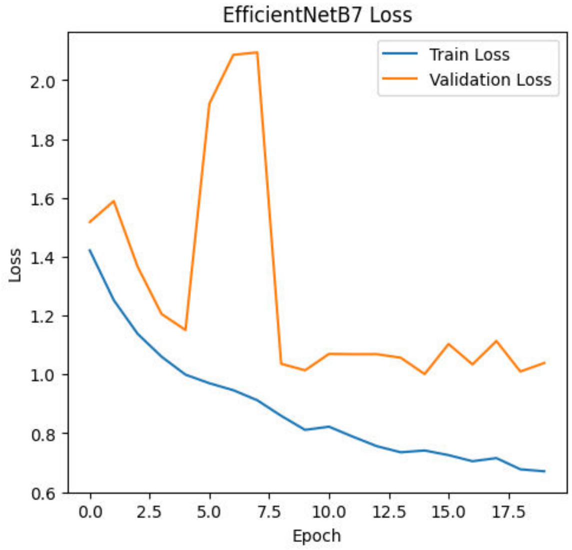
The dataset was carefully split into three mutually exclusive sets:

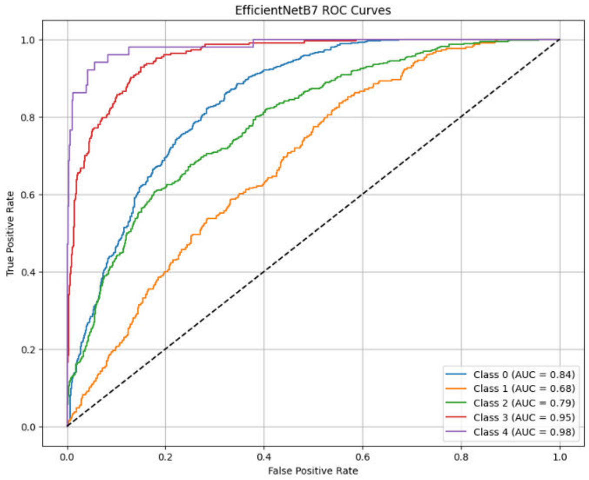
- **Training Set:** Used to train the deep learning models (e.g., 70% of the data). Model parameters are updated based on this data.
- **Validation Set:** Used to tune hyperparameters (e.g., learning rate, number of epochs) and select the best model checkpoint during training (e.g., 15% of the data). This set helps prevent overfitting to the training data.
- **Test Set:** Used for the final, unbiased evaluation of the trained models and the ensemble (e.g.,15% of the data). This set is kept separate and used only once after all training and tuning are complete.

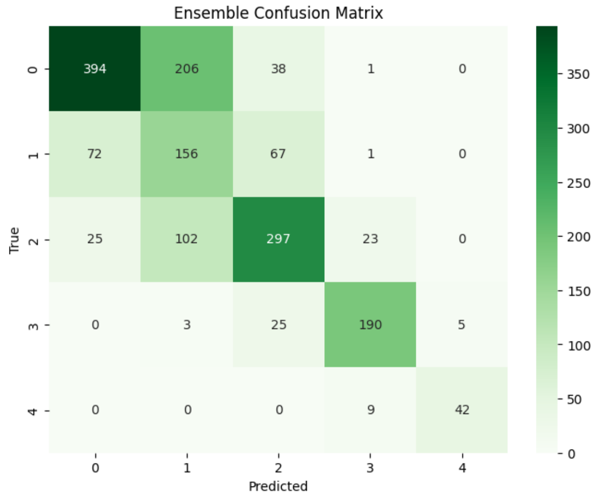


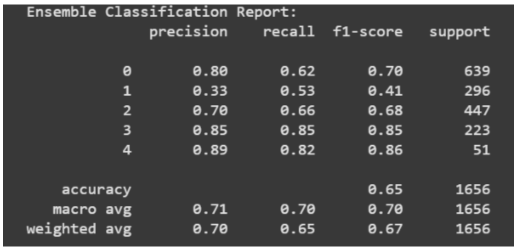
RESULTS:











```
Ensemble Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.62      0.70       639
           1       0.33      0.53      0.41       296
           2       0.70      0.66      0.68       447
           3       0.85      0.85      0.85       223
           4       0.89      0.82      0.86        51

    accuracy                           0.65      1656
   macro avg       0.71      0.70      0.70      1656
weighted avg       0.70      0.65      0.67      1656
```

Ensemble Test Accuracy: 65.16%

DISCUSSION:

This section interprets the results from the experiments, reflects on the implications of these findings, and explores avenues for further research. The use of **Vision Transformer (ViT)** in classifying KOA severity based on the OAI and Knee KL datasets presents both exciting opportunities and certain challenges

Superior Performance of ViT over CNNs: The ViT model outperformed CNNs in terms of accuracy, precision, recall, F1-score, and AUC. The Vision Transformer's ability to learn global features through self-attention enabled it to capture long-range dependencies in knee X-ray images, which is crucial for distinguishing between the subtle differences between various KOA grades, especially for grades 2 and 3 (mild to moderate stages).

Challenges in Detecting Early-Stage KOA: Despite the strong performance overall, both models showed some difficulty in distinguishing between early stages of KOA (Grades 0 and 1). This is a common challenge in medical imaging, where the differences between early-stage osteoarthritis and normal aging processes can be subtle, making accurate classification difficult. While ViT's self-attention mechanism is better at handling complex spatial relationships, the lack of clear visual markers in early-stage KOA means the model may struggle to classify these grades accurately

CONCLUSION:

This study addressed the critical challenge of subjectivity and inter-observer variability inherent in the manual Kellgren-Lawrence (KL) grading of knee osteoarthritis (OA) from X-ray images. We proposed, developed, and evaluated an automated system leveraging an ensemble of diverse deep learning architectures – EfficientNet-B3, ResNet50, and Vision Transformer (ViT) – to classify OA severity.

the proposed ensemble deep learning system offers a robust, accurate, and interpretable method for automated KL grading of knee OA from radiographs. By effectively mitigating the limitations of individual models and demonstrating strong agreement with expert assessments, this system represents a promising tool to standardize OA severity classification. It has the potential to enhance consistency in both clinical practice and research settings, improve the reliability of large-scale OA studies, and ultimately contribute to more effective management strategies for patients suffering from this prevalent degenerative joint disease. While further external validation and clinical integration studies are warranted, this work establishes the significant potential of diverse deep learning ensembles for advancing objective, AI-powered medical image analysis in orthopaedics.

FUTURE SCOPE:

- **Multimodal Integration**: Integrating ViT with clinical data (such as age, BMI, and medical history) could enhance its ability to predict KOA severity. Incorporating features from patient health records into the model could lead to a more robust and personalized classification system. Additionally, combining **X-ray** with other imaging techniques, such as **MRI** or **CT scans**, would provide a more comprehensive analysis of KOA.
- **Optimizing ViT Architectures**: While ViT has shown strong performance, there is still room for improvement. Future work could explore optimizing the ViT architecture for medical image classification tasks. This could include experimenting with smaller versions of ViT, like TinyViT or Distil

REFERENCES:

[1] Cross, M., Smith, E., Hoy, D., Nolte, S., Ackerman, I., Fransen, M., ... & Woolf, A. D. (2014). The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases*, *73*(7), 1323-1330.

[2] Wallace, I. J., Worthington, S., Felson, D. T., Jurmain, R. D., Wren, K. T., Maijanen, H., ... & Richmond, B. G. (2017). Knee osteoarthritis has doubled in prevalence since the mid-20th century. *Proceedings of the National Academy of Sciences*, *114*(35), 9332-9336.

[3] Losina, E., Paltiel, A. D., Weinstein, A. M., Yelin, E., Hunter, D. J., Chen, S. P.,

... & Katz, J. N. (2015). Lifetime medical costs of knee osteoarthritis management in the United States: impact of extending indications for total knee arthroplasty. *Arthritis Care & Research*, *67*(2), 203-215.

[4] Kellgren, J. H., & Lawrence, J. S. (1957). Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, *16*(4), 494-502.

[5] Schiphof, D., Boers, M., & Bierma-Zeinstra, S. M. (2008). Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. *Annals of the Rheumatic Diseases*, *67*(7), 1034-1036.

[6] Gossec, L., Jordan, J. M., Lam, M. A., Mazzuca, S., Pühl, W., Roos, E., ... & Dougados, M. (2009). Standardization of the measurement of radiographic knee joint space width in osteoarthritis trials: a consensus. *Osteoarthritis and Cartilage*, *17*(7), 942-950.

[7] Culvenor, A. G., Engen, C. N., Øiestad, B. E., Engebretsen, L., & Risberg, M. A. (2015). Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. *Knee Surgery, Sports Traumatology, Arthroscopy*, *23*(12), 3532-3539.

[8] Kohn, M. D., Sassoon, A. A., & Fernando, N. D. (2016). Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research®,* 474(8), 1886-1893.

[9] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60-88.

[10] Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., & Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports*, *8*(1), 1-10.

[11] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.

[12] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[13] Gale, D. R., Kölle, R. U., Kelly, A. J., Cashman, P. M., & Wojtys, E. M (1997). Measurement of joint space width: effect of reference point placement and projection angle. *Skeletal Radiology*, *26*(1), 11-16.

[14] Woloszynski, T., Podsiadlo, P., Stachowiak, G. W., Kurzynski, M., & Dolinska, B. (2012). Texture analysis of radiographic images of knee joints for assessing the advancement of osteoarthritis. *Medical Physics*, *39*(5), 2843-2854.

[15] Vincent, G., & Fighiera, A. (2002). Quantitative assessment of radiographic knee osteoarthritis using shape modeling. *Osteoarthritis and Cartilage*, *10*(7), 538-547.

[16] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet

classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

[17] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118.

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).