Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 13, No2 pp262–278 April -June,2025

journals.uoz.edu.krd

p-ISSN: 2663-628X
e-ISSN: 2663-6298

KNEE OSTEOARTHRITIS STAGE CLASSIFICATION BASED ON HYBRID FUSION DEEP LEARNING FRAMEWORK

Delveen Luqman Abd Alnabi [1, *], Shereen Sh. Ahmed[2], Nisreen Luqman Abd Alnabi[3]

[1] College of Administration and Economics, University of Duhok, Duhok, Kurdistan Region, Iraq.
[2] College of Science, University of Zakho, Duhok, Kurdistan Region, Iraq.
[3] Technical College of Administration, Duhok Polytechnique University, Duhok, Kurdistan Region, Iraq.

*Corresponding author email: delveen.luqman@uod.ac

Received: 22 Des 2024 Accepted:24 Mar 2025 Published:28 Apr 2025 https://doi.org/10.25271/sjuoz.2025.13.2.1450

ABSTRACT:

Knee osteoarthritis severity detection is one of the most challenging applications in computer vision due to the similarity between X-ray images of the adjacent stages. Handling huge number of X-ray images and the ability to detect the correct disease stage is based on advanced artificial intelligence technologies, like machine learning and deep learning. This study presents a novel deep learning-based fusion framework designed for detecting the severity of knee osteoarthritis and classifying its stages. The study utilizes two X-ray image datasets containing three challenges: imbalanced data, low contrast, and low data size. Data augmentation, adaptive histogram equalization, and limited oversampling techniques were used to solve these problems. Five deep learning architectures were utilized as base models (EfficientNetB0, EfficientNetV2B0, XceptionNet, ResNetRS101, and RegNetY032), followed by average pooling and dense layers. The feature-level, decision-level, score-level, and meta-based fusion technologies were also performed on the outputs of the best three trained models to minimize the individual models' errors. The study registered 70% and 90.61% classification accuracies using both datasets. The study also found that the best models are the score-level and meta-based fusion models in all scenarios.

KEYWORDS: Deep learning, Deep learning fusion, Disease classification, Feature-level fusion, Knee osteoarthritis

1. INTRODUCTION:

Knee osteoarthritis is one of the common diseases of knee joints that leads to significantly disabling human mobility at different levels (Han *et al.*, 2020; Zhu *et al.*, 2024; Pires *et al.*, 2024; Giannopapas *et al.*, 2024). In this disease, the slippery cartilage tissues covering the bone joints and providing good mobility cause humans to lose properties and wear out the tissues (Yildirim & Mutlu, 2024). This disease usually happens to elderly people and can affect the knee, the hand, the spine, etc. (Ilmi *et al.*, 2024; Komalasari & Motik, 2024). The pain and limited mobility persist the entire day during this disease and are caused by any hard activity or even the long period of inactivity of the patient (Watso & Vondrasek, 2024). The treatment of knee osteoarthritis requires multiple procedures and approaches to reduce the disease's symptoms and reduce its severity (Yildirim & Mutlu, 2024).

Knee osteoarthritis contains many stages according to the visual X-ray notice and disease symptoms: normal, doubtful, mild, moderate, and severe (Nasser *et al.*, 2023; Jahan *et al.*, 2024). In the first stage, the disease does not exist. In the 'doubtful' stage, the disease occurrence is uncertain and cannot exist. In the third stage, mild', the disease exists but is at the first stage with light symptoms and small joint space narrowing. In the fourth stage, 'moderate', the X-ray images of the patient start to show obvious osteophytes and reduction in the area of the knee joint. In the final stage, 'the severe' one, there will be a significant osteophyte, a huge joint narrowing, and severe sclerosis (Rehman & Gruhn, 2024).

Artificial intelligence (AI), machine learning (ML) and deep learning (DL) capabilities have been recently utilized for the aim of knee osteoarthritis stage classification in order to help physicians in their clinical investigations (Rani *et al.*,2024; Raza *et al.*, 2024; Zhao *et al.*, 2024). Many studies were introduced in this field for either binary classification (Ahmed & Omran, 2024) or severity stage detection (Bose *et al.*, 2024; Nurmirinta *et al.*, 2024; Rani *et al.*, 2024).
The Osteoarthritis Initiative (OAI) dataset (Chen, 2018) is a well-known X-ray image of knee Osteoarthritis stages. Many studies utilized this dataset. (Du *et al.*, 2018) utilized

[*] Corresponding author

the Kellgren-Lawrence (KL), and the Artificial Neural Networks (ANN) and obtained an area under curve AUC of 0.822 to 0.903. (Chen *et al.*, 2019) trained the VGG-19 model in a transfer learning way using the OAI dataset and achieved an accuracy of 69.7%. In another study (Ahmed & Mstafa, 2022), they also proposed transfer learning along with the principal component analysis (PCA) and support vector machines (SVM) classifier to achieve an accuracy of 62%, and an AUC score of 0.854. The same dataset was recently utilized in a study by (Apon *et al.*, 2024). The researchers utilized many DL architectures, including InceptionV3, and Vision transformers (ViT) models, and registered an accuracy of 66.14% and an AUC score of 0.835. The OsteoHRNet DL model proposed by (Jain *et al.*, 2024) was trained on a knee X-ray dataset. The attention mechanism was also utilized to improve performance. The obtained accuracy of their study was 71.74%. However, their architecture added overhead due to the extra computational time. Transfer learning of many DL architectures (DenseNet169, VGG, InceptionV3, ResNet50, Inception-ResNetV2, and Xception) was utilized by (El-Ghany *et al.*, 2023). They Found that the DenseNet169 was the best model with an accuracy of 95.93% for only binary classification (without severity detection). (M & Goswami, 2023) utilized an X-ray knee image dataset consisting of 1656 images and five severity stages. They obtained an accuracy of 72% using a ResNet-like CNN model. In a study by (Ahmed & Imran, 2024), they utilized pre-trained CNN models (VGG, ResNet, GradCAM) using a divide-and-conquer approach to step into the binary classification problem instead of severity classification mission. They utilized the Knee Osteoarthritis Severity Grading Dataset consisting of 8260 images. Although they achieved an accuracy of 99.13%, it dropped to 67% for multi-class severity detection. In another study (Jain *et al.*, 2024), they utilized the high-resolution net (HRNet) and the attention mechanism to extract the best multi-scale features of knee x-ray images. They achieved an accuracy of 71.74% on the OAI dataset.

In the conclusion of the previous studies, we found that most of these studies utilized only one dataset. Some of them applied the binary classification or only three-stage-based severity classification. Some studies developed complex architectures, while others were stuck in low performance due to using weak individual models. In this study, a novel hybrid DL framework is utilized to improve the performance of the individual models by minimizing individual errors. The study will also utilize two datasets with various challenges, including data imbalance, low dataset size, and low gray contrast.

The main contributions of the current study are:

1. This study utilizes the capabilities of the fusion techniques (score-level, decision-level, feature-level, and meta-based fusion) in improving the performance of the individual DL models in the field of knee osteoarthritis stage classification.
2. The study utilizes two datasets with different challenges: one with a data imbalance problem and the other with a small data size.
3. The study overcomes the challenges by introducing preprocessing using the CLAHE algorithm, SMOTE for data balance, data augmentation operations for improving the dataset size, and image preprocessing using adaptive histogram equalization.
4. The study proposes a new DL-based individual framework consisting of two parts: the feature extraction part, in which five different robust DL architectures are utilized, while in the second classification part, a proposed classification framework is utilized to handle non-linearity and satisfy the problem conditions.
5. The study replaces the binary classification (disease occurrence detection) with the stage classification for a better and more reliable disease staging system.

2. MATERIALS AND METHODS

Knee X-ray Datasets:

In this research, two X-ray open-source datasets are utilized (Nouman, 2024; Chen, 2018) with five different severity stages of the knee joint for both. The first one consists of 1650 X-ray images of knee joints without pre-split of training or test sets. The Second one is the "Osteoarthritis Initiative (OAI) dataset" (Chen, 2018), which includes five different stages of osteoarthritis severity of knee joint and 9786 X-ray images distributed on training, validation, and test sets. Figure 1 shows the distribution of samples among the five stages of both datasets, where the five stages are: 'Normal: no disease', 'Doubtful: it may contain the disease and may not', 'Mild: the first obvious stage of the disease', 'Moderate: the disease in the middle stage', and 'severe: the disease is in an advanced stage'.

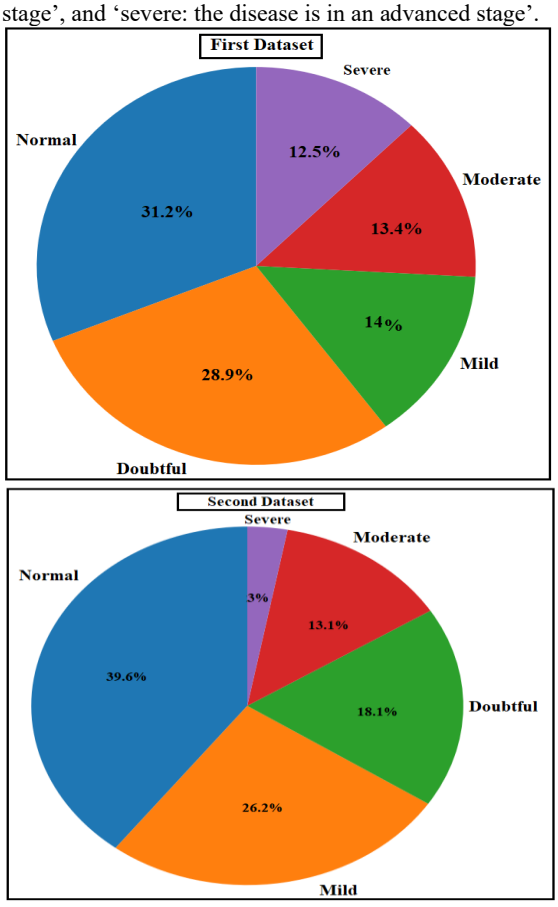


Figure 1: Distribution of the five classes of severity within the utilized dataset

The second dataset suffers from the imbalance problem in which there is a category 'Severe' has a small number of samples (only 3% of the entire dataset), leading to problems during the training due to the problem that DL models will bias to the dominant classes and learn their parameters to memorize their samples. To address this issue, a data balance is suggested in order to balance the training classes and prevent any possible biases. Figure 2 includes samples of both datasets.

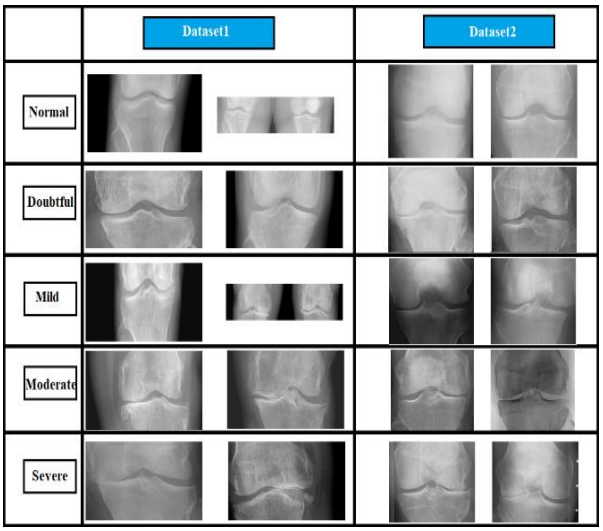


Figure 2: Some representative samples of the utilized datasets

3. The Proposed Methodology

In this study, deep learning, transfer learning, fine-tuning, and fusion techniques are all utilized to achieve the best performance of the knee osteoarthritis detection and staging classification system. Figure 3 shows the proposed methodology by which the dataset is first pre-processed and prepared using many operations, including splitting, image enhancement, data augmentation, and data balance. Figure 3-b shows an example of data preprocessing and augmentation.

In the preprocessing operations, the images are resized into a specific size (224*224 for the first dataset and 128*128 for the second one since the number of images is higher and the current resources are limited to the COLAB environment resources). Then, the dataset is split to train 80% and test 20% (only the first dataset since the second one is already split). The contrast-limited adaptive histogram equalization (CLAHE) algorithm is then applied to the images in order to dynamically improve the contrast of the images (Hu *et al.*, 2024). The data augmentation operations are then applied to the training set of the first and second datasets, while the validation and test sets are preserved without changes. The proposed data augmentation operations are random rotation in the range 0^0 to 25^0, horizontal flipping to get the flipped view of the sample, random width, and height shifting with a range of 0-0.2, zooming operation in the range (0-0.2), and shearing in the range (0-0.2). The data augmentation steps help to improve the training by increasing the training size and adding some noise and variation to the data leading to more powerful robust training (Chlap *et al.*, 2021; Shorten & Khoshgoftaar, 2019). The data balance operation is applied only to the training set of the second dataset. The study proposes using the Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al.*, 2002; Pradipta *et al.*, 2021) to increase the number of the minor categories' samples. The proposed technique in our study is based on increasing the minor classes' samples to a specific number (and not matching the number of the majority class) in order to minimize the effect of generating outlier samples. In our workflow, we first use SMOTE to balance the training set of the second dataset by increasing the number of samples in the minority classes to a predetermined level. Once the training set is balanced, the data generator then applies augmentation on the fly during training. The data balance is applied as follows: First, the image is flattened into a 1D array. After that, the flattened images are standardized using 'StandardScaler' to centers data (mean=0) which is important for SMOTE algorithm which utilizes distance metrics. In the third step, the SMOTE is applied to these flattened and standardized images using two neighbors (k=2). After that, the images are re-constructed again into the original 2D size.

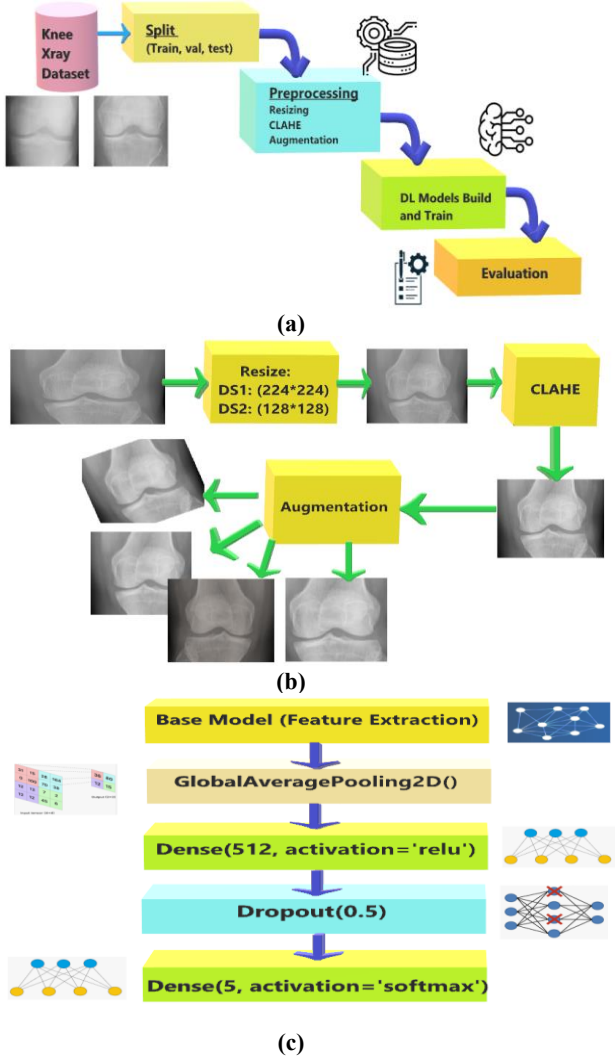


Figure 3: The proposed methodology: (a) The general steps, (b) the data preprocessing steps on a training sample, (c) The proposed architecture for all DL models.

After that, the deep learning architectures are loaded as pre-trained models. The inputs and outputs of the proposed models are modified to fit our data and classification labels (5 stages). So, the input layer is removed and replaced by the desired size (224*224 for the first dataset and 128*128 for the second one). The output layer is also modified to contain only 5 neurons (since we have 5 stages to be classified to define the knee osteoarthritis disease stage) and a 'Softmax' function is utilized as the activation function. The general architecture of each DL model is presented in Figure 3-C, where first the extracted feature vector of the DL model is obtained (base model), then a global average pooling layer is utilized to flatten the final feature matrix of the base model. After that, a dense layer of 512 neurons is utilized to minimize the dimensions, then a dropout layer is utilized to improve the non-linearity of the models. The final part is the classification layer with 5 neurons. The models are trained using the augmented and balanced training set, while the test set (without augmentation or balance) is utilized in the evaluation step. The next step performs the model fusion techniques (feature-level fusion, score-level fusion, decision-level fusion, and meta-fusion) on the best three trained DL models to reduce the individual models' errors and improve the performance. The proposed later-based fusion techniques (score-level, decision-level, and meta-fusion) are illustrated in Algorithms 1, 2, and 3.

Algorithm 1: Knee osteoarthritis stage classification-based ***score-level*** fusion framework (KOSCSLF)

Input: Test set (TS), trained DL models (model),

Output: Final classification (prediction).

Steps:

1- Compute prediction scores of the individual models using Equation 1.

$$prop_i = model_i.predict\ (TS) \quad (1)$$

where i=1,2,…,N. N: number of fused models.

2- Average the scores making one fused score as illustrated in Equation 2.

$$average\_scores = \sum_{i=1}^{N} w_i * prop_i \quad (2)$$

3- Compute the highest average probabilities to derive the final fused score:

$$Final\_\ prediction = \max\ (Fused\_score) \quad (3)$$

4- Output the Final prediction.

Algorithm 2: Knee osteoarthritis stage classification-based ***decision-level*** fusion framework (KOSCDLF)

Input: Test set (TS), trained DL models ($model_i$),
Output: Final classification (prediction).

Steps:

1- Compute prediction decisions of the individual models using Equation 4.

$$Pred_i = \max(model_i.predict(TS)) \quad (4)$$

2- Stack the decisions of the individual models in one matrix (all_predictions).

3- Compute the mode of the stacked predictions as in Equation 5:

$$Final\_prediction = mode(all\_predictions) \quad (5)$$

4- Output the Final_prediction.

Algorithm 3: Knee osteoarthritis stage classification based ***meta fusion*** framework (KOSCMF)

Input: Test set (TS), trained DL models ($model_i$),
Output: Final classification (prediction).

Steps:

1- Compute prediction decisions of the individual models.

2- Stack the decisions of the individual models in one matrix (all_predictions).

3- Choose a meta-classification model (Logistic regression) and train it in the outputs of the second step for 1000 epochs.

4- Compute the prediction of the meta-model using the same test set.

$$meta\text{-}pred_i = meta\text{-}model.predict(TS) \quad (6)$$

5- Output the meta-predictions as the Final_score.

The proposed transfer learning of the DL models:

As mentioned earlier, the transfer learning capabilities will be utilized in this study using five DL models, including EfficientNetB0, EfficientNetV2B0, ResNetRS101, XceptionNet, and RegNetY032 models. The selection of these models is based on three factors: high performance (RegNetY032 and ResNetRS101), lightweight architectures (EfficientNetB0), and the recent new and state-of-the-art DL architectures (RegNetY032, XceptionNet, EfficientNetV2B0).

Transfer learning is a well-known technique in which the original knowledge of the DL model (its original trained parameters on a specific dataset or application) is transferred to a new domain (application) with a new dataset and a new mission. In the new mission, the main architecture is maintained. The feature extraction part is also maintained, while the classification part is retrained to produce the prediction (classification) according to the new mission (El Gannour *et al.*, 2024; Lu *et al.*, 2015; Niu *et al.*, 2020). The comparison between the proposed DL models is shown in Table 1.

Table 1: DL models comparison.

Model	Total parameters according to our task	Size (MB)	Number of layers	Input size	Layers details	Main characteristics
EfficientNetB0 (Tan & Le, 2019)	4049571	15.93	237	224*224	7 MB convolutional layers. Block scales with depth and width. 1 Global average pooling (GAP) layer. 1 Fully-connected (FC) layer with 1000 neurons.	Includes Compound scaling and MobileNet-like inverted bottleneck modules.
EfficientNetV2B0 (Tan & Le, 2021)	5919312	23.149	218	224*224	6 Fused-MB Conv Blocks stages. 1 GAP, 1 FC (varying outputs)	Enhanced scaling and efficient training with Fused-MBConv layers
ResNetRS 101 (Wightman *et al.*, 2021)	61675296	236.33	101	224*224	101 Conv layers Bottleneck layers (conv layers 3*3 with residual connection shortcut), 1 GAP, 1 FC	Enhanced ResNet with selective kernel besides squeeze-and-excite layers
XceptionNet (Chollet, 2017)	20861480	79.8	71	299*299	36 Conv layers (distributed among 14 groups) Depthwise separable Conv layers 1 GAP, 1 FC	Contain the Depthwise separable convolutions and an Inception-inspired module
RegNetY032 (Radosavovic *et al.*, 2020)	17989498	69.49	32	224*224	Bottleneck layers with SE features 1 GAP 1 FC	Organized by neural architecture search. It focuses on width and depth balancing. Utilize Squeeze-and-Excitation (SE) blocks to improve performance.

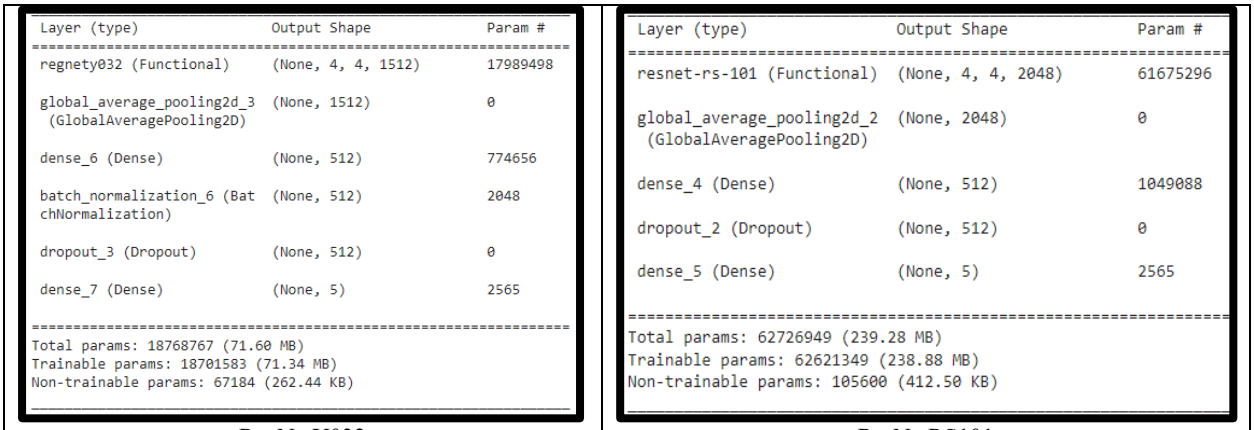
The classification part which contains two dense layers results in 658,437, 1,051,653, 1,051,653, 777,221, and 658,437 trainable parameters for EfficientNetB0, XceptionNet, ResNetRS101, RegNetY032, and EfficientNetV2B0, respectively. Figure 4 shows the architecture of the proposed DL models with the proposed classification part.

```
Layer (type)                 Output Shape              Param #
=================================================================
efficientnetb0 (Functional   (None, 7, 7, 1280)        4049571
)

global_average_pooling2d (   (None, 1280)              0
GlobalAveragePooling2D)

dense (Dense)                (None, 512)               655872

dropout (Dropout)            (None, 512)               0

dense_1 (Dense)              (None, 5)                 2565

=================================================================
Total params: 4708008 (17.96 MB)
Trainable params: 4665985 (17.80 MB)
Non-trainable params: 42023 (164.16 KB)
_________________________________________________________________
```

EfficientNet models

```
Layer (type)                 Output Shape              Param #
=================================================================
xception (Functional)        (None, 7, 7, 2048)        20861480

global_average_pooling2d_1   (None, 2048)              0
 (GlobalAveragePooling2D)

dense_2 (Dense)              (None, 512)               1049088

dropout_1 (Dropout)          (None, 512)               0

dense_3 (Dense)              (None, 5)                 2565

=================================================================
Total params: 21913133 (83.59 MB)
Trainable params: 21858605 (83.38 MB)
Non-trainable params: 54528 (213.00 KB)
_________________________________________________________________
```

Xception



Layer (type)	Output Shape	Param #
regnety032 (Functional)	(None, 4, 4, 1512)	17989498
global_average_pooling2d_3 (GlobalAveragePooling2D)	(None, 1512)	0
dense_6 (Dense)	(None, 512)	774656
batch_normalization_6 (BatchNormalization)	(None, 512)	2048
dropout_3 (Dropout)	(None, 512)	0
dense_7 (Dense)	(None, 5)	2565

Total params: 18768767 (71.60 MB)
Trainable params: 18701583 (71.34 MB)
Non-trainable params: 67184 (262.44 KB)

RegNetY032

Layer (type)	Output Shape	Param #
resnet-rs-101 (Functional)	(None, 4, 4, 2048)	61675296
global_average_pooling2d_2 (GlobalAveragePooling2D)	(None, 2048)	0
dense_4 (Dense)	(None, 512)	1049088
dropout_2 (Dropout)	(None, 512)	0
dense_5 (Dense)	(None, 5)	2565

Total params: 62726949 (239.28 MB)
Trainable params: 62621349 (238.88 MB)
Non-trainable params: 105600 (412.50 KB)

ResNetRS101

Figure 4: The proposed DL architectures with their backbones and classification parts**.**

The feature-level fusion framework:

The feature-level fusion is based on combining the feature vectors of two or more deep learning architectures by one vector and then utilizing a single classification part. Figure 5 illustrates the proposed feature-level fusion architecture for both datasets. The architecture includes a combination of multiple feature vectors and then utilizing a concatenation layer to merge the multiple features of many DL models, then a dense layer of 512 neurons (with 'Relu' activation function), a dropout layer to add non-linearity by a rate of 50%, and finally a classification (dense) layer with 5 neurons and a 'Softmax' activation function.

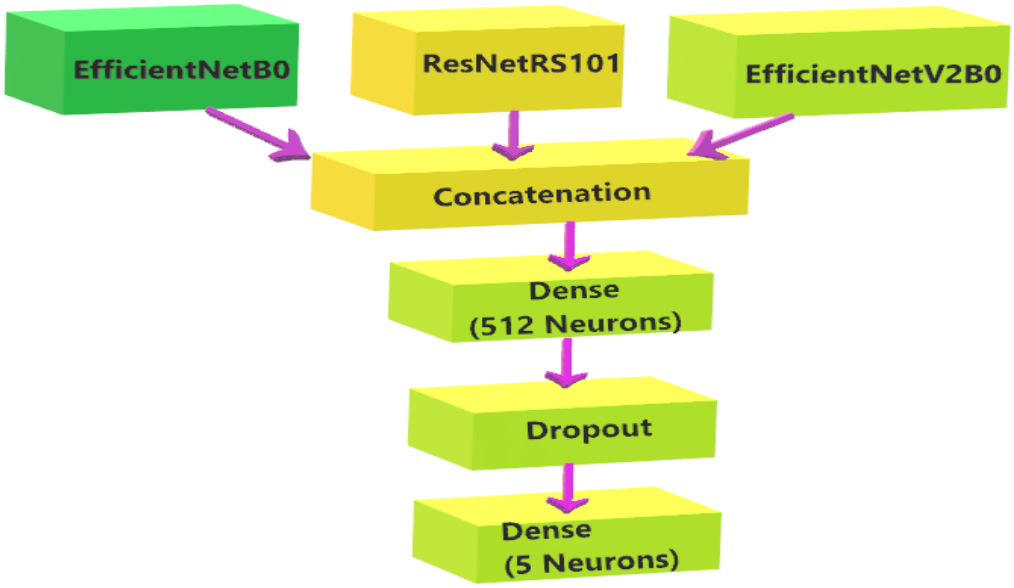


Figure 5: The proposed feature-level fusion architecture.

Performance Evaluation Metrics:

After training the individual DL models, and generating the fusion-based architectures, all these models need an evaluation step in which the models are assessed using many classification metrics, including accuracy, precision, recall, F1-score, the area under the curve (AUC), confusion matrix and training time (Hoang *et al.*, 2024; Khozama & Mayya, 2022; Szabó *et al.*, 2024).

Precision can be explained as the percentage of true positive TP samples out of all positive ones (TP and false positives (FP)) (Szabó *et al.*, 2024), meaning that if a sample with a category 'Mild' is predicted as 'Mild' so this is called one TP sample, while if a 'Doubtful' sample is accepted as a 'Mild' one, this is called FP sample. Recall, on the other hand, concentrates on minimizing the false negatives (FN) since its formula is TP/(TP+FN) (Szabó *et al.*, 2024), so if a 'Mild' sample is recognized as a 'Moderate' one, it will be considered as an FN, and this will reduce the recall

percentage. Accuracy is the general metric in which all true positives and true negatives (TN) are summed and divided by the total number of test samples (TP+TN)/(TP+TN+FP+FN) (Szabó *et al.*, 2024). AUC score is computed as the area under the ROC curve (Szabó *et al.*, 2024), representing the relationship between true and false positive rates. The confusion matrix is also one of the most commonly utilized performance assessment methods since it allows us to see the FP, FN, TP, and TN of the individual classes, and this can help to know exactly the performance degradation causes (at any specific category). All these metrics will be calculated for all models and all experiments.

Ethical approval and consent:

All authors gave verbally informed consent for their participation. The study's design and procedures were reviewed and approved by the Ethics and Scientific Committee of the College of Medicine at the University of Zakho with the reference number (FEB2024/UOZE446).

4. RESULTS AND DISCUSSION

In the experimental part, the experiments are performed on both datasets. Five individual DL models are trained and then fused using different fusion methodologies. Table 2 includes the training parameters utilized to train the DL models.

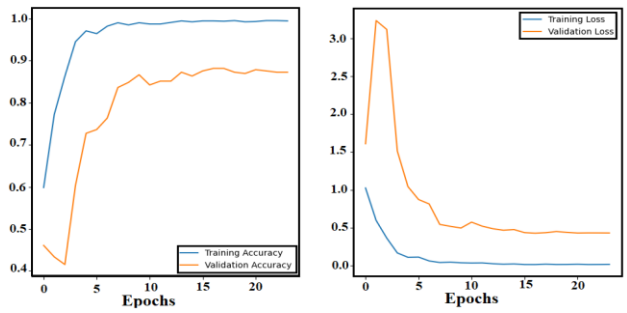
Table 2. Training parameters for both datasets.

Training parameter	Value or option
Initial Learning rate	0.001
Optimizer	Adam
Loss Function	Sparse categorical cross-entropy
Number of epochs	30
Early stopping	Monitor validation accuracy for 7 epochs; if no enhancement, stop the training.
Input Image size	For the first dataset, the utilized input size is 224*224 For the second dataset, the utilized input size is 128*128
Batch size	32

Results of the first dataset:

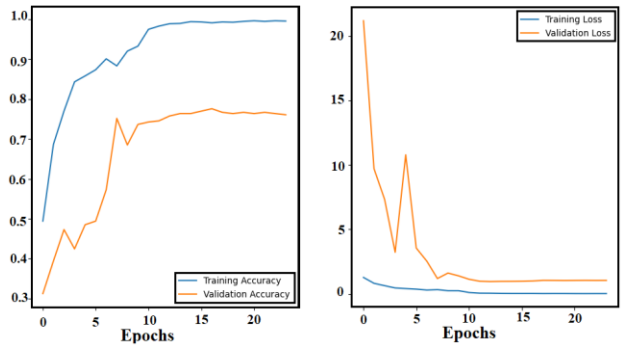
Figure 6 includes the model accuracy of the trained individual DL models (Orange color for validation curves and blue color for training curves). Figure 6 shows that the model with the best training and validation accuracy is EfficientNetB0, while other models include unstable training. The difference in the number of epochs among models is due to the utilization of the early stop condition to avoid potential overfitting or redundant training.

(a) (b)



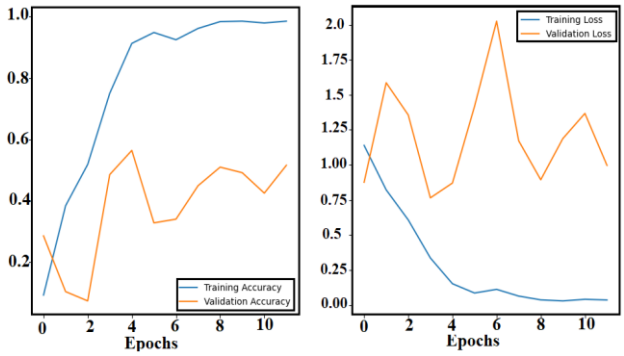
EfficicentNetB0

(a) (b)



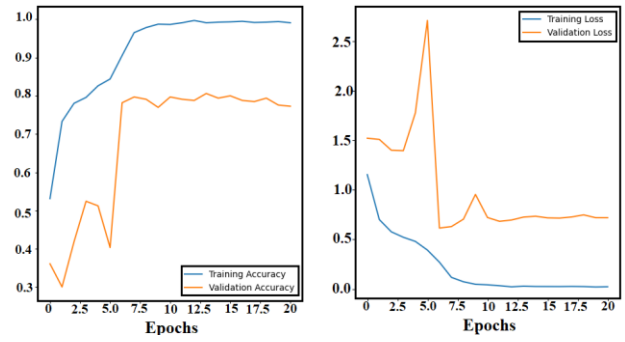
XceptionNet

(a) (b)



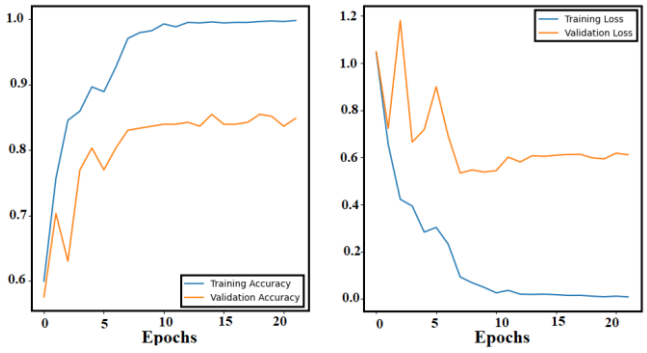
ResNetRS101

(a) (b)



RegNetY032

(a) (b)



EfficientNetV2B0

Figure 6: Performance evaluation metrics: (a) Training and validation accuracy, (b) Training and validation loss.

Table 3 includes the detailed performance metrics of the trained DL models using the first dataset. The performance metrics include precision, recall, and F1-score of the five classes of knee osteoarthritis. Besides this, the macro average and weighted average of the individual scores are also computed to get an overall assessment of the individual models. Table 3 proves that the best model with the best precision (89.86%), the best recall (86.98%), and the best F1-score (88.14%) is the EfficientNetB0. Table 3 also proves that the best performance of the five categories corresponds to the 'Moderate' category with 97.56%, 90.91%, and 94.14% for precision, recall, and F1-score. However, the worst performance of the categories is registered for the 'Doubtful' and the 'Mild' categories with 85.58% and 78.57% for F1-score, respectively. Figure 7 also includes snapshots from the results obtained from program output

Table 3: Individual model performance metrics were trained using the first dataset.

	EfficientNetB0			**Xception**			**ResNetRS101**			**RegNetY032**			**EfficientNetV2B0**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
N*	0.9293	0.8932	0.9109	0.8557	0.8058	0.8300	0.9348	0.8350	0.8821	0.8378	0.9029	0.8692	0.9670	0.8544	0.9072
D	0.7946	0.9271	0.8558	0.6881	0.7812	0.7317	0.6860	0.8646	0.7650	0.7451	0.7917	0.7677	0.7664	0.8542	0.8079
Mi	0.8684	0.7174	0.7857	0.6250	0.6522	0.6383	0.5439	0.6739	0.6019	0.7250	0.6304	0.6744	0.7273	0.6957	0.7111
Mo	0.9756	0.9091	0.9412	0.9459	0.7955	0.8642	0.9474	0.8182	0.8780	0.9189	0.7727	0.8395	0.9756	0.9091	0.9412
S	0.9250	0.9024	0.9136	0.8462	0.8049	0.8250	1.0000	0.5366	0.6984	0.8500	0.8293	0.8395	0.8511	0.9756	0.9091
MA	**0.8986**	**0.8698**	**0.8814**	0.7922	0.7679	0.7778	0.8224	0.7456	0.7651	0.8154	0.7854	0.7981	0.8575	0.8578	0.8553
WA	0.8873	0.8818	0.8818	0.7856	0.7758	0.7786	0.7847	0.7817	0.7856	0.8075	0.8061	0.8049	0.8620	0.8545	0.8557

***N: Normal, D: Doubtful, Mi: Mild, Mo: Moderate, S: Severe, MA: Macro average, WA: Weighted average, P: precision, R: Recall, F: F1-score.**

```
11/11 [==============================] - 3s 60ms/step
              precision    recall  f1-score   support

     0Normal     0.9293    0.8932    0.9109       103
  1Doubtful      0.7946    0.9271    0.8558        96
      2Mild      0.8684    0.7174    0.7857        46
  3Moderate      0.9756    0.9091    0.9412        44
     4Severe     0.9250    0.9024    0.9136        41

    accuracy                         0.8818       330
   macro avg     0.8986    0.8698    0.8814       330
weighted avg     0.8873    0.8818    0.8818       330
```

EfficientNetB0

```
11/11 [==============================] - 2s 152ms/step
              precision    recall  f1-score   support

     0Normal     0.8557    0.8058    0.8300       103
  1Doubtful      0.6881    0.7812    0.7317        96
      2Mild      0.6250    0.6522    0.6383        46
  3Moderate      0.9459    0.7955    0.8642        44
     4Severe     0.8462    0.8049    0.8250        41

    accuracy                         0.7758       330
   macro avg     0.7922    0.7679    0.7778       330
weighted avg     0.7856    0.7758    0.7786       330
```

ExceptionNet

```
11/11 [==============================] - 2s 190ms/step
              precision    recall  f1-score   support

     0Normal     0.9348    0.8350    0.8821       103
  1Doubtful      0.6860    0.8646    0.7650        96
      2Mild      0.5439    0.6739    0.6019        46
  3Moderate      0.9474    0.8182    0.8780        44
    4Severe      1.0000    0.5366    0.6984        41

    accuracy                         0.7818       330
   macro avg     0.8224    0.7456    0.7651       330
weighted avg     0.8177    0.7818    0.7856       330
```

ResNetRS101

```
11/11 [==============================] - 1s 123ms/step
              precision    recall  f1-score   support

     0Normal     0.8378    0.9029    0.8692       103
  1Doubtful      0.7451    0.7917    0.7677        96
      2Mild      0.7250    0.6304    0.6744        46
  3Moderate      0.9189    0.7727    0.8395        44
    4Severe      0.8500    0.8293    0.8395        41

    accuracy                         0.8061       330
   macro avg     0.8154    0.7854    0.7981       330
weighted avg     0.8075    0.8061    0.8049       330
```

RegNetY032

```
11/11 [==============================] - 3s 51ms/step
              precision    recall  f1-score   support

     0Normal     0.9670    0.8544    0.9072       103
  1Doubtful      0.7664    0.8542    0.8079        96
      2Mild      0.7273    0.6957    0.7111        46
  3Moderate      0.9756    0.9091    0.9412        44
    4Severe      0.8511    0.9756    0.9091        41

    accuracy                         0.8545       330
   macro avg     0.8575    0.8578    0.8553       330
weighted avg     0.8620    0.8545    0.8557       330
```

EfficientNetV2B0

Figure 7: Snapshots from the results obtained from program output.

The fusion step is performed on the best three models. The accuracy, AUC score, and training time of all ***fusion*** models using the first dataset are illustrated in Table 4 (where there is no training time for score-level and decision-level fusion models since we applied the fusion after the training operation). Table 4 illustrates that the best scores correspond to the score-level fusion scenario in which the precision, recall, and F1-score register 91.23%, 90.47%, and 90.66%, respectively. Table 4 also proves that the score-level fusion model outperforms all individual models' performance.

Table 4: Fusion models performance metrics trained using the first dataset.

	Feature-level fusion			**Score-level fusion**			**Decision-level fusion**		
	P	R	F	P	R	F	P	R	F
N*	0.8889	0.8544	0.8713	0.9789	0.9029	0.9394	0.9286	0.8835	0.9055
D	0.8119	0.8542	0.8325	0.8241	0.9271	0.8725	0.7818	0.8958	0.8350
Mi	0.8095	0.7391	0.7727	0.8500	0.7391	0.7907	0.7500	0.7174	0.7333
Mo	0.9773	0.9773	0.9773	0.9767	0.9545	0.9655	0.9756	0.9091	0.9412
S	0.9091	0.9756	0.9412	0.9318	1.0000	0.9647	0.9730	0.8780	0.9231
MA	0.8793	0.8801	0.8790	**0.9123**	**0.9047**	**0.9066**	0.8818	0.8568	0.8676
WA	0.8697	0.8697	0.8691	0.9098	0.9061	0.9058	0.8728	0.8667	0.8679

Table 5 includes the accuracy comparison between different models and the fusion models using the first dataset. Table 5 proves the same conclusion of previous tables and figures by which the score-level fusion achieves the best accuracy score of 90.61%.

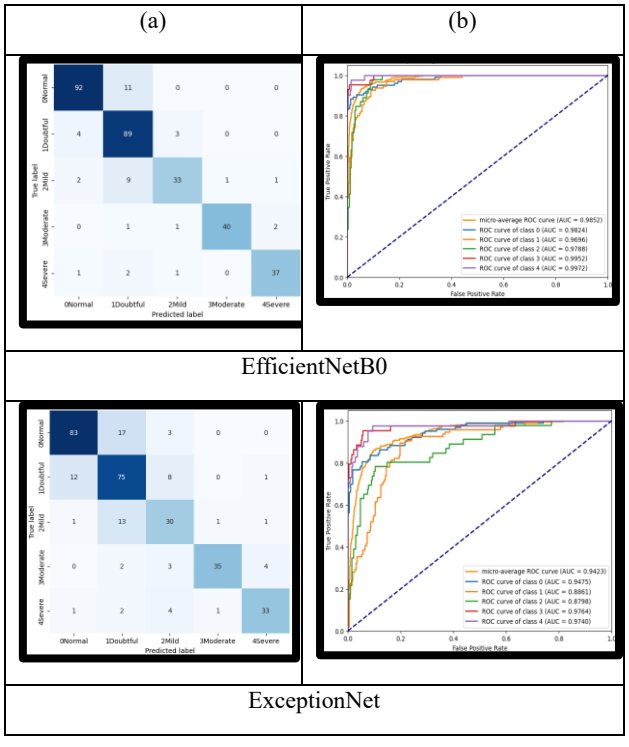
Table 5. Fusion models accuracy scores using the first dataset.

	Efficient NetB0	**Exception Net**	**ResNet RS101**	**RegNet Y032**	**Efficient NetV2B0**	**Feature-level fusion**	**Score -level fusion**	**Decision-level fusion**
Accuracy (%)	88.18	77.58	78.18	80.61	85.45	86.97	**90.61**	86.67

Discussion of the results of the first dataset:

The best individual model in terms of accuracy is the EfficientNetB0, with an accuracy of 88.18%. The second-best one is the EfficientNetV2B0 with an 85.45% score, and the third-best accuracy corresponds to the RegNetY032 with an 80.61% score. However, the score-level fusion improves the best individual accuracy by 5.16%.

The confusion matrix and ROC plots of the individual DL models trained using the first dataset are illustrated in Figure 8. Figure 8 proves that the EfficientNetB0 model performs best with the least number of false positive and false negative errors. Moreover, the EfficientNetB0 model has the best AUC score of 0.9824, while the worst performance corresponds to the ResNetRS101 with only 0.9408. Besides this, the category with the largest number of errors (according to the confusion matrixes) is the 'Mild' category, with 13 false negatives among 46 total samples and 5 false positives among 38 total samples (these calculations correspond to the best model EfficientNetB0). On the other hand, the category with the lowest error percentage is the 'Moderate' category, with 4 false negative errors out of 44 samples, and only one false positive error out of 41 samples. The fact that the errors are mostly located in the first three categories is caused by the similarity in X-ray images within the first three types of knee osteoarthritis disease (i.e. during the first three stages of knee osteoarthritis, the disease is either not exist (normal case), not obvious (Doubtful), or is at the first appearance in images (mild)).



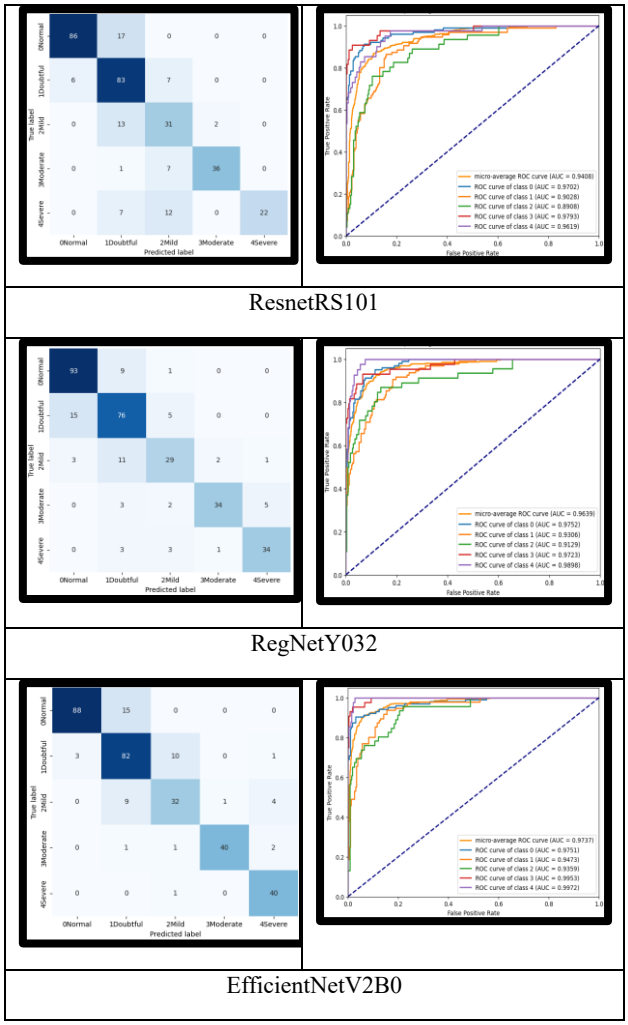


Figure 8. Individual models confusion matrixes and ROC plot: (a) Confusion matrix, (b) ROC plot.

It can be noticed that the score-level fusion model achieves the best performance with a 0.9859 AUC score and 31 false negative and positive errors out of 330 test samples. Figure 9 also shows that the errors of the 'Doubtful' and 'Mild' categories (which were the categories with the highest number of errors in individual models) are minimized significantly. The 'Doubtful' category registers 7 false negative errors out of 96 samples compared to 7 false negative errors, and 19 false positive errors compared to 23 false positive errors of the best individual model. The 'Mild' category includes 12 false negative and 6 false positive errors compared to 13 false negative and 6 false positive errors of the best individual model. The most enhanced category is the 'Mild' category which has 41 true positives compared to only 37 true positives of the best individual model.

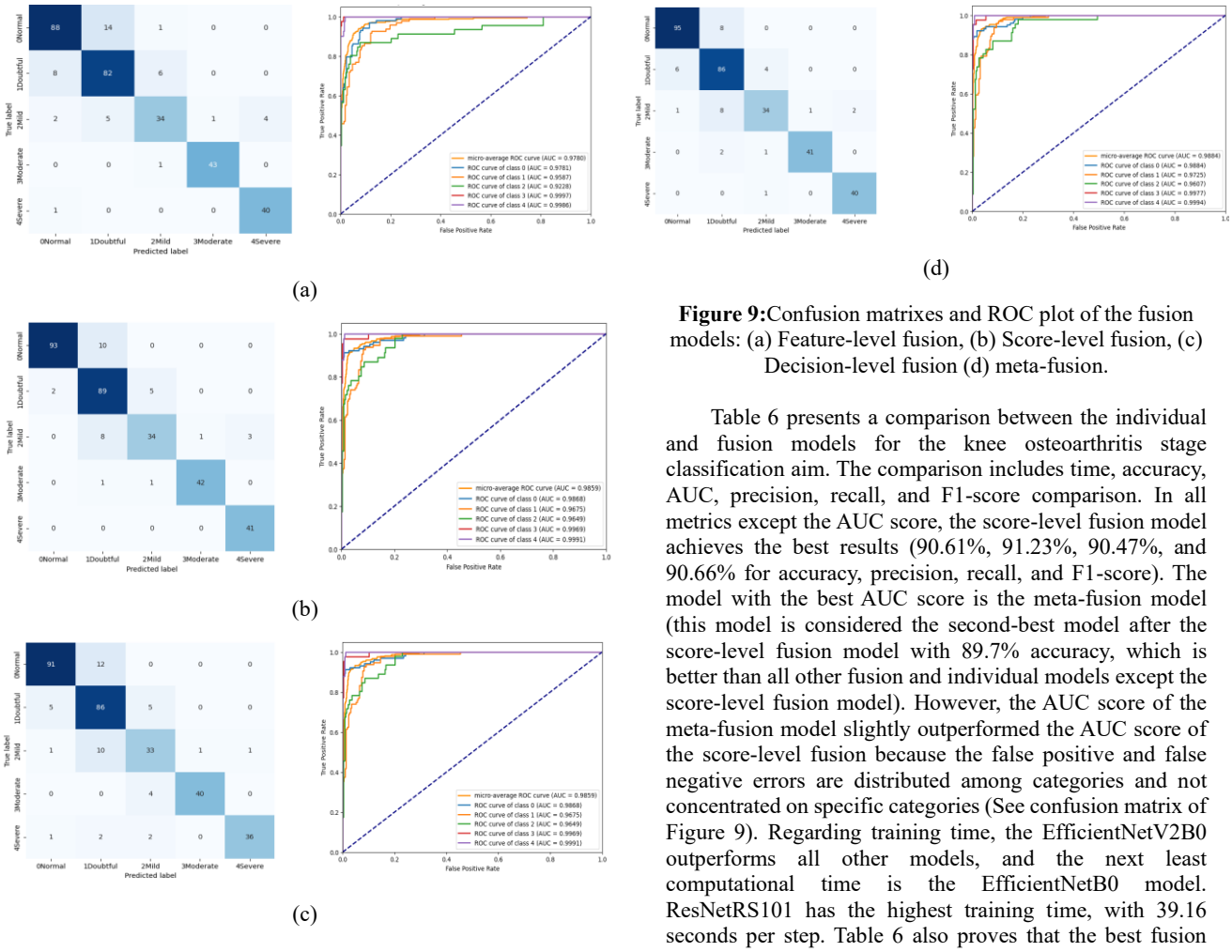


Figure 9: Confusion matrixes and ROC plot of the fusion models: (a) Feature-level fusion, (b) Score-level fusion, (c) Decision-level fusion (d) meta-fusion.

Table 6 presents a comparison between the individual and fusion models for the knee osteoarthritis stage classification aim. The comparison includes time, accuracy, AUC, precision, recall, and F1-score comparison. In all metrics except the AUC score, the score-level fusion model achieves the best results (90.61%, 91.23%, 90.47%, and 90.66% for accuracy, precision, recall, and F1-score). The model with the best AUC score is the meta-fusion model (this model is considered the second-best model after the score-level fusion model with 89.7% accuracy, which is better than all other fusion and individual models except the score-level fusion model). However, the AUC score of the meta-fusion model slightly outperformed the AUC score of the score-level fusion because the false positive and false negative errors are distributed among categories and not concentrated on specific categories (See confusion matrix of Figure 9). Regarding training time, the EfficientNetV2B0 outperforms all other models, and the next least computational time is the EfficientNetB0 model. ResNetRS101 has the highest training time, with 39.16 seconds per step. Table 6 also proves that the best fusion model (score-level fusion model) outperforms the best individual model (EfficientNetB0) by 2.43%, 1.37%, 3.49%, and 2.52% for accuracy, precision, recall, and F1-score, respectively.

Table 6: Comparison between individual and fusion models in terms of accuracy, AUC, and training time using the first dataset.

	EB0	XN	RRS101	RNY032	EV2	FF	SF	DF	MF
Accuracy	0.8818	0.7758	0.7818	0.8061	0.8545	0.8697	**0.9061**	0.8667	0.897
AUC	0.9852	0.9423	0.9408	0.9639	0.9751	0.9780	0.9858	0.9859	**0.988**
Precision	0.8986	0.7922	0.8224	0.8154	0.8575	0.8793	**0.9123**	0.8818	0.907
Recall	0.8698	0.7679	0.7456	0.7854	0.8578	0.8801	**0.9047**	0.8568	0.893
F1-score	0.8814	0.7778	0.7651	0.7981	0.8553	0.8790	**0.9066**	0.8676	0.899
TT (S/Step)	13.75	24.85	39.16	23.8	10.95	23.46	-	-	-

EB0: EfficientNetB0, XN: XceptionNet, RRS101: ResNetRS101, RNY032: RegNetY032, EV2: EfficientNetB0V2, FF: Feature-level fusion model, SF: Score-level fusion model, DF: Decision-level fusion model.

Figure 10 contains a visual experiment of testing the trained and fusion models using some test samples of the test set. In these test samples, the fusion model (score-level fusion) correctly classified all test samples to the correct stage

(severe, moderate, normal, doubtful, or mild), which proves the ability of the fusion model to define not only the occurrence of the disease but also its precise stage. On the other hand, some of the individual models misclassified the test samples. The ResNetRS101 model misclassifies samples 1, 3, 5, and 6. The EfficientNetB0 model misclassifies only sample number 3. EfficientV2B0 misclassifies samples 3 and 6.

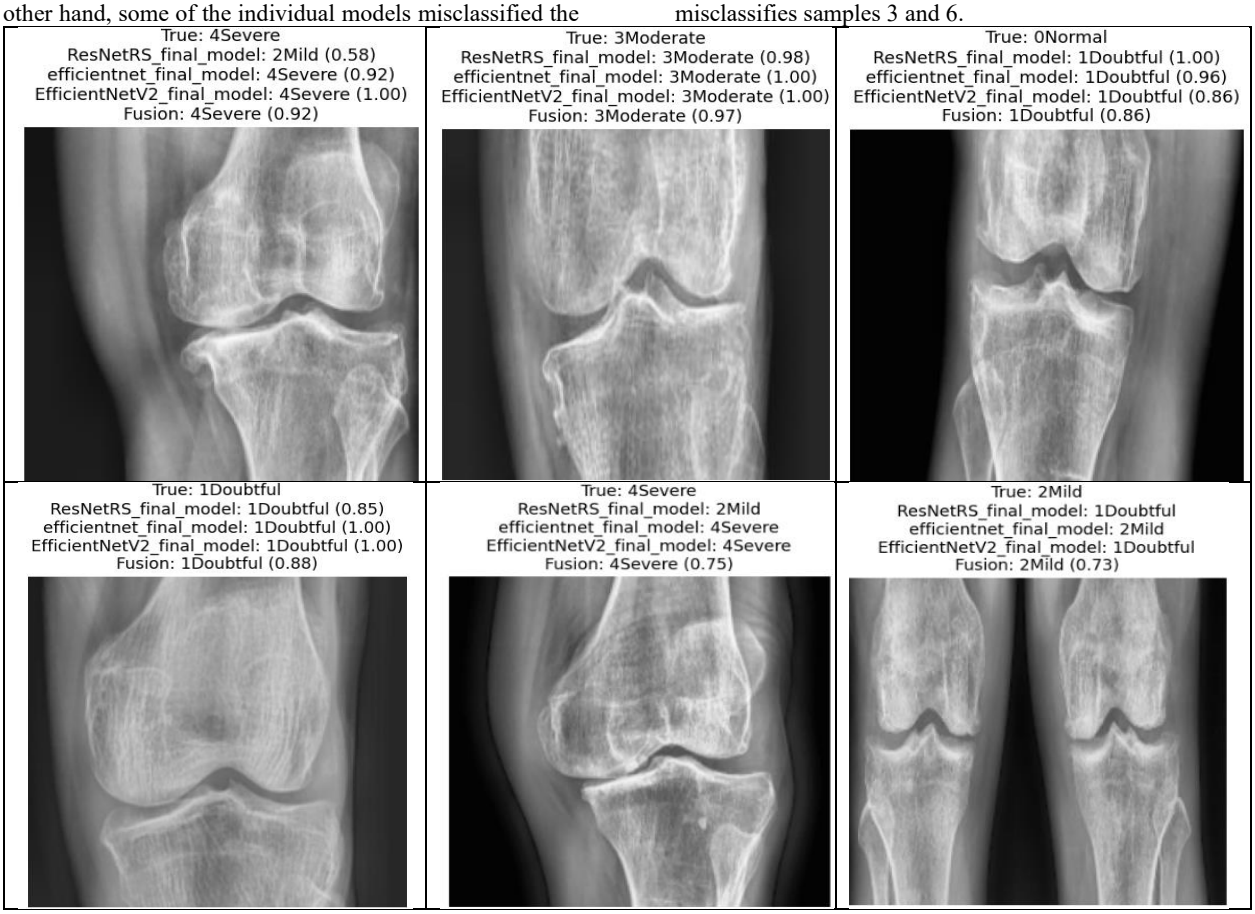


Figure 10: Visual test of some individual models and the score-level fusion model using some test samples of the first dataset.

Figure 11 proves that the utilization of the CLAHE preprocessing step improves the performance of the fusion model compared to the original case without the utilization of such a preprocessing step.

Figure 11 also shows that the utilization of the CLAHE preprocessing step enhanced the ability of the model to recognize the true positive and true negative samples since precision, recall, F1-score, and AUC scores are all improved by such modification.

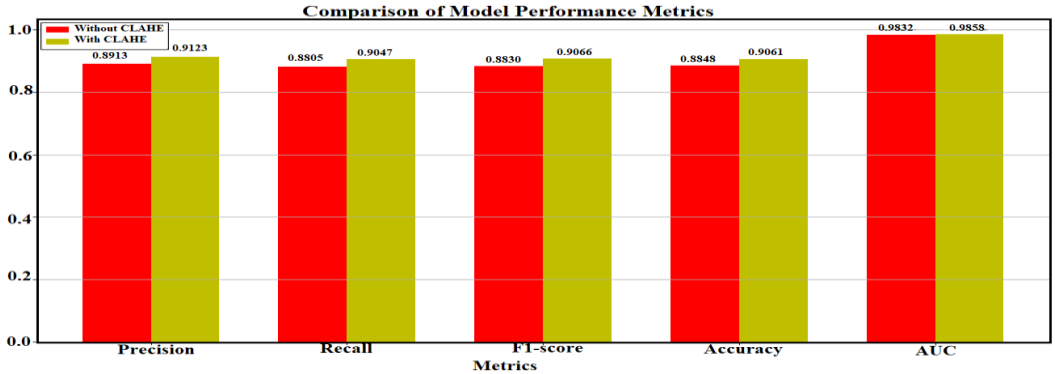


Figure 11: Effect of using CLAHE on the performance of the score-level Fusion Model.

Generalization study (Results on the second dataset):

In this part, another different dataset is utilized with a new challenge (unbalanced dataset) in which there are categories with a small number of samples, while others contain a large number of samples, leading to the unbalanced dataset. Table 7 includes details of the experiments applied

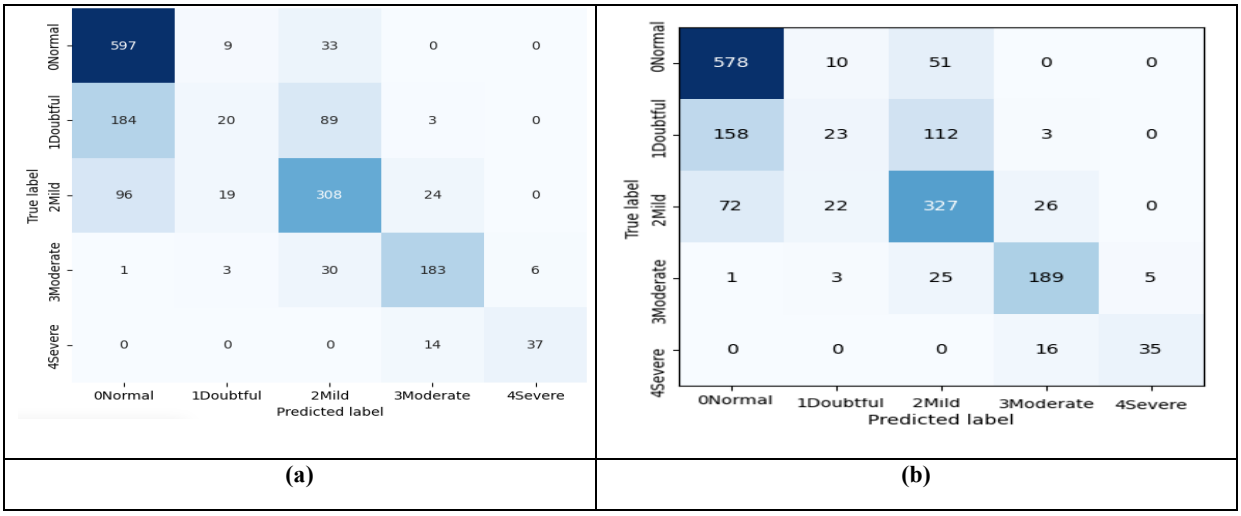
to the second dataset, including all performance metrics (precision, recall,
F1-score, AUC, accuracy, and training time). Table 7 proves that the best two models with the highest scores are the score-level fusion and the meta-fusion models with accuracies of 69.14% and 70%, respectively. In terms of the training time, again, the EfficientNetB0 model has the least training time

Table 7: Comparison between individual and fusion models in terms of accuracy, AUC, and training time using the second dataset.

*	EB0	XN	RRS101	RNY032	EV2	FF	SF	DF	MF
Accuracy	0.6594	0.6727	0.6763	0.6407	0.6310	0.4463	0.6914	0.6842	**0.70**
AUC	0.9122	0.9220	0.9240	0.9079	0.8988	0.7760	0.9280	0.9280	**0.923**
Precision	0.6614	0.6469	0.6481	0.6444	0.5952	0.4246	0.6838	0.6868	**0.69**
Recall	0.6382	0.6326	0.6195	0.5684	0.5668	0.3276	0.6474	0.6300	**0.65**
F1-score	0.6389	0.6197	0.6095	0.5635	0.5583	0.2894	0.6375	0.6265	**0.64**
TT (S/Step)	39.93	46.68	78.15	53	38.47	39.36	-	-	-

Discussion of the second dataset results:

Figure 12 shows the confusion matrixes of the best models. The meta-based fusion enhanced the number of true positives of the 'Doubtful' category from 20 (in score-level fusion model) to 23 TPs. It also improves the true positives of the third category, 'Mild,' by 19 samples and the true positives of the 'Moderate' class by 6 samples. While the 'Normal' category true positives are reduced from 597 to 578, and the 'Severe' category is reduced by 2 samples. However, the fusion models reduced the individual models' errors and improved their performance significantly. The meta-fusion model, for example, improves the performance of the best individual model (EfficientNetB0) by 4%, 2.8%, 1.18%, and 1.1% for accuracy, precision, recall, and F1-score, respectively.



(a) **(b)**

Figure 12: Confusion matrixes of the fusion models of the second dataset: (a) Score-level fusion, (b) meta-fusion.

Figure 13 includes three examples of predicting three test samples using the individual and fusion models. As seen in Figure 12, the fusion model successes to classify all three samples. On the other hand, individual models failed to predict them all correctly.

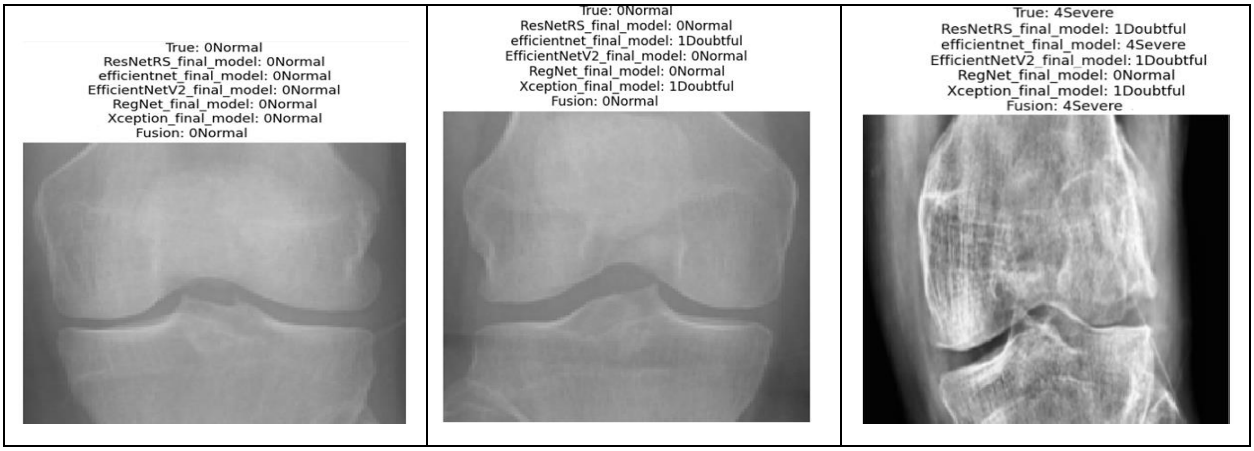


Figure 13: Visual test of some individuals and the score-level fusion model using some test samples of the second dataset

Comparison with related work:

In this section, the comparison with related work will be concentrated on the studies that utilize the same dataset to unify the comparison. Table 8 includes this detailed comparison taking into account the methodologies, the outcomes and the main consideration and limitations.

Table 8: Comparison with related work

Study	Methodology	Dataset	Results & Notes
(Du *et al.*, 2018)	Kellgren-Lawrence (KL), (ANN)	Osteoarthritis Initiative (OAI) dataset	AUC=0.822 to 0.903
(Chen *et al.*, 2019)	VGG-19	Osteoarthritis Initiative (OAI) dataset	Accuracy= 69.7%
(Ahmed & Mstafa, 2022)	CNN, Transfer learning, PCA, and SVM	Osteoarthritis Initiative (OAI) dataset	AUC=0.854 Precision=63.2% Recall=58.6% F1-score=59.6% Accuracy=62%
(Mohammed *et al.*, 2023)	VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121	Osteoarthritis Initiative (OAI) dataset (five classes, three classes and two classes)	For five classes: Accuracy=69%.
(Bhateja *et al.*, 2024)	CNN-based ensemble method	Osteoarthritis Initiative (OAI) dataset	Accuracy=68%
(Apon *et al.*, 2024)	VGG-19, Inception-V3, Da-VIT, GCViT and MaxViT	Osteoarthritis Initiative (OAI) dataset	Accuracy=66.14% AUC=0.835
Current research	Hybrid fusion DL models (Score-level, feature-level, and meta-fusion), CLAHE, SMOTE Balance	Osteoarthritis Initiative (OAI) dataset	Accuracy=70% Precision=69% Recall=65% F1-score=64 AUC=0.923
		Kaggle OA dataset (1650) images	Accuracy=90.61% Precision=91.23% Recall=90.47% F1-score=90.66% AUC=0.9858

Table 8 explores the studies that worked on the same dataset and proves that the current study outperforms all of them due to the following causes:1. In our study, the fusion of the best DL model is performed by reducing the individual errors and improving the performance, 2. In our study, the CLAHE preprocessing operation is applied to improve the contrast of the images, 3. The data balance applied to the

training set of the second dataset helps to prevent biasing to the dominant classes as much as possible.

It must be denoted that some pieces of research like a study by (Messaoudene & Harrar, 2024) utilized the same dataset but used only Grade0 and Grade1 of all datasets (only two stages) and achieved an accuracy of 94.59% which is normal since the main problem if the OAI dataset is the multi-level (grades) that makes the mission harder to be processed by best DL models. Another thesis master by (Rifat, 2024) utilized semi-supervised learning and deep learning on the same dataset but used only 'healthy', 'moderate', and 'severe' categories, which facilitated the training process and registered an accuracy of 82%.

CONCLUSION

In this study, a novel fusion framework of five robust deep learning models (EfficientNetB0, EfficientNetV2B0, ResNetRS101, RegNetY032, XceptionNet) was developed for the aim of knee osteoarthritis severity detection and stages classification. Five different severity stages were considered: normal, doubtful, mild, moderate, and severe. Two different open-access X-ray image datasets were utilized; the first one contains low data size, while the second one is imbalanced, and both datasets suffer from low contrast. To address such problems, the study suggested using specific data augmentation operations, CLAHE enhancement, and SMOTE-limited techniques. The feature-level fusion, score-level fusion, decision-level fusion, and meta-based fusion of the best-trained DL models were also developed. As a result, many training and evaluation scenarios were derived. The results showed that the best two models were score-level fusion and meta-based fusion. The results also revealed that the severity classification accuracy among the five classes was 70% and 90.61% on both datasets, respectively. The categories with the best detection accuracy of the first dataset were the 'severe' and 'moderate' categories with 96.47% and 96.55% F1-score; while for the second dataset, the 'normal' and 'moderate' categories were the best ones with 80% and 83% F2-score, respectively. The main issue is that the second and third severity stages were very similar, and this limited the performance of the DL individual models. Although the proposed fusion framework enhanced the performance, it only partially solved this issue. Future studies must work on the issue of similarity between these two stages by developing hybrid symptoms and image-based detection models. The study was also compared to the related work, and the comparison illustrated that the study outperforms all studies except those that developed binary classification or three-stage severity detection models.

ACKNOWLEDGEMENTS :

We would like to express our gratitude to the University of Duhok, the University of Zakho, and the Duhok Polytechnique University for their time and consideration in supporting our academic endeavors.

Statements and Declarations:

Ethical approval :

All authors gave verbal informed consent for their participation. The study's design and procedures were reviewed and approved by the Research Ethics Committee of the College of Medicine, UOZ, in compliance with ethical standards (Code UOZE446; 2024).

Conflict of Interest: The authors declared that no potential conflict of interest.

Author Contributions: All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Consent to Participate: All authors have consented to submit this article to this journal.

Consent to Publish: All authors have consented to publish this article in this journal.

Funding: The study did not receive specific funding from public, commercial, or any non-profit organizations.

Concept and Design: Delveen Luqman Abd Alnabi, Shereen Sh. Ahmed, Nisreen Luqman Abd Alnabi.

Acquisition, Analysis, or Interpretation of Data: Delveen Luqman Abd Alnabi, Shereen Sh. Ahmed, Nisreen Luqman Abd Alnabi.

Drafting of the Manuscript: Delveen Luqman Abd Alnabi.

REFERENCES

Ahmed, R., & Imran, A. S. (2024). Knee Osteoarthritis Analysis Using Deep Learning and XAI on X-rays. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3400987

Ahmed, S. M., & Mstafa, R. J. (2022). Identifying Severity Grading of Knee Osteoarthritis from X-ray Images Using an Efficient Mixture of Deep Learning and Machine Learning Models. *Diagnostics*, *12*(12), 2939. https://doi.org/10.3390/diagnostics12122939

Apon, T. S., Fahim-Ul-Islam, M., Rafin, N. I., Akter, J., & Alam, M. G. R. (2024). *Transforming precision: A comparative analysis of vision transformers, CNNs, and traditional ML for knee osteoarthritis severity diagnosis*. In 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) (pp. 31–36). IEEE. https://doi.org/10.1109/ICEEICT62016.2024.10534528

Bhateja, V., Dubey, Y., Maurya, N., Yadav, V. K., Shrivastava, S., Azar, A. T., Haider, Z., Amin, S. U., & Khan, Z. I. (2024). Ensemble CNN model for computer-aided knee osteoarthritis diagnosis. *International Journal of Service Science, Management, Engineering, and Technology, 15*(1), 1–17. https://doi.org/10.4018/IJSSMET.349913

Bose, A. S. C., Srinivasan, C., & Joy, S. I. (2024). Optimized feature selection for enhanced accuracy in knee osteoarthritis detection and severity classification with machine learning. *Biomedical Signal Processing and Control*, 97, 106670. https://doi.org/10.1016/j.bspc.2024.106670

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial*

Intelligence Research, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, P. (2018). Knee osteoarthritis severity grading dataset [Data set]. Mendeley Data. https://doi.org/10.17632/56rmx5bjcr.1

Chen, P., Gao, L., Shi, X., Allen, K., & Yang, L. (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, *75*, 84–92. https://doi.org/10.1016/j.compmedimag.2019.06.002

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563. https://doi.org/10.1111/1754-9485.13261

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp.*1800-1807). IEEE https://doi.org/10.1109/CVPR.2017.195.

Du, Y., Shan, J., Almajalid, R., & Zhang, M. (2018). Knee osteoarthritis severity level classification using whole knee cartilage damage index and ANN. In *Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies* (pp.19-21). ACM https://doi.org/10.1145/3278576.3278585

El Gannour, O., Hamida, S., Lamalem, Y., Mahjoubi, M. A., Cherradi, B., & Raihani, A. (2024). Improving skin diseases prediction through data balancing via classes weighting and transfer learning. *Bulletin of Electrical Engineering and Informatics*, *13*(1), 628–637. https://doi.org/10.11591/eei.v13i1.5999

El-Ghany, S. A., Elmogy, M., & El-Aziz, A. A. A. (2023). A fully automatic fine tuned deep learning model for knee osteoarthritis detection and progression analysis. *Egyptian Informatics Journal*, *24*(2), 229–240. https://doi.org/10.1016/j.eij.2023.03.005

Giannopapas, V., Smyrni, V., Kitsos, D. K., Stasi, S., Chasiotis, A. K., Moschovos, C., Papagiannopoulou, G., Stavrogianni, K., Kosmidou, M., Bakalidou, D., Tzartos, J. S., Tsivgoulis, G., & Giannopoulos, S. (2024). Osteoarthritis in people with multiple sclerosis: A systematic review and meta-analysis. *Journal of Clinical Medicine*, *13*(17), 5015. https://doi.org/10.3390/jcm13175015

Han, X., Cui, J., Xie, K., Jiang, X., He, Z., Du, J., Chu, L., Qu, X., Ai, S., Sun, Q., Wang, L., Wu, H., Zhang, W., Yu, Z., & Yan, M. (2020). Association between knee alignment, osteoarthritis disease severity, and subchondral trabecular bone microarchitecture in patients with knee osteoarthritis: A cross-sectional study. *Arthritis Research & Therapy*, *22*(1), 203. https://doi.org/10.1186/s13075-020-02274-0

Hoang, Q. T., Pham, X. H., Trinh, X. T., Le, A. V., Bui, M. V., & Bui, T. T. (2024). An efficient CNN-based method for intracranial hemorrhage segmentation from computerized tomography imaging. *Journal of Imaging*, *10*(4), 77. https://doi.org/10.3390/jimaging10040077

Hu, C., Li, H., Ma, T., Zeng, C., & Ji, X. (2024). An improved image enhancement algorithm: Radial contrast-limited adaptive histogram equalization. *Multimedia Tools and Applications*, *83*(36), 83695–83707. https://doi.org/10.1007/s11042-024-18922-5

Ilmi, B., Parellangi, P., Rizani, A., Hammad, H., Mallongi, A., & Palutturi, S. (2023). The effect of elderly hadrah gymnastics on muscle strength and scope of motion of lower extremity joints in elderly with osteoarthritis (Martapura River Region, South Kalimantan). *Pharmacognosy Journal*, *15*(6), 1126–1131. https://doi.org/10.5530/pj.2023.15.205

Jahan, M., Hasan, Md. Z., Jahan Samia, I., Fatema, K., Rony, Md. A. H., Shamsul Arefin, M., & Moustafa, A. (2024). KOA-CCTNet: An enhanced knee osteoarthritis grade assessment framework using modified compact convolutional transformer model. *IEEE Access*, *12*, 107719–107741. https://doi.org/10.1109/ACCESS.2024.3435572

Jain, R. K., Sharma, P. K., Gaj, S., Sur, A., & Ghosh, P. (2024). Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network. *Multimedia Tools and Applications*, 83(3), 6925-6942. https://doi.org/10.1007/s11042-023-15484-w

Khozama, S., & Mayya, A. M. (2022). A new range-based breast cancer prediction model msing the bayes' theorem and ensemble learning. *Information Technology and Control*, *51*(4), 757–770. https://doi.org/10.5755/j01.itc.51.4.31347

Komalasari, D. R., & Motik, A. F. (2024). Reliability test of the Timed Up and Go test in elderly people with knee osteoarthritis. *FISIO MU: Physiotherapy Evidences,* *5*(2), 129–140. https://doi.org/10.23917/fisiomu.v5i2.4227

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., & Zhang, G. (2015). Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, *80*, 14–23. https://doi.org/10.1016/j.knosys.2015.01.010

Goswami, M. K., & Das, A. (2023). Automatic classification of the severity of knee osteoarthritis using enhanced image sharpening and CNN. *Applied Sciences,* *13*(3), 1658. https://doi.org/10.3390/app13031658

Messaoudene, K., & Harrar, K. (2024). Computerized diagnosis of knee osteoarthritis from x-ray images using combined texture features: Data from the osteoarthritis initiative. *International Journal of Imaging Systems and Technology*, *34*(2). https://doi.org/10.1002/ima.23063

Mohammed, A. S., Hasanaath, A. A., Latif, G., & Bashar, A. (2023). Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed X-ray images. *Diagnostics*, *13*(8), 1380. https://doi.org/10.3390/diagnostics13081380

Nasser, Y., El Hassouni, M., Hans, D., & Jennane, R. (2023). A discriminative shape-texture convolutional neural network for early diagnosis of knee osteoarthritis from X-ray images. *Physical and Engineering Sciences in Medicine*, *46*(2), 827–837. https://doi.org/10.1007/s13246-023-01256-1

Niu, S., Liu, Y., Wang, J., & Song, H. (2020). A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, *1*(2), 151–166. https://doi.org/10.1109/TAI.2021.3054609

Nouman, H. (2024). Annotated dataset for knee arthritis detection [Data set]. Kaggle. https://www.kaggle.com/datasets/hafiznouman786/annotated-dataset-for-knee-arthritis-detection

Nurmirinta, T. A., Turunen, M. J., Korhonen, R. K., Tohka, J., Liukkonen, M. K., & Mononen, M. E. (2024). Two-stage classification of future knee osteoarthritis severity after 8 years using MRI: data from the osteoarthritis initiative. *Annals of Biomedical Engineering*, 52(12), 3172-3183. https://doi.org/10.1007/s10439-024-03578-x

Pires, D. P. D. C., Monte, F. A. Do, Monteiro, L. F., Soares, F. R. D. C., & Faria, J. L. R. De. (2024). Updates in the treatment of knee osteoarthritis. *Revista Brasileira de Ortopedia*, *59*(3), 337–348. https://doi.org/10.1055/s-0044-1786351

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021). Smote for handling imbalanced data problem: A review. *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–8. https://doi.org/10.1109/ICIC54025.2021.9632912

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (n.d.). *Designing network design spaces*. arXiv. https://arxiv.org/abs/2003.13678

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

Raza, A., Phan, T. L., Li, H. C., Hieu, N. V., Nghia, T. T., & Ching, C. T. S. (2024). A comparative study of machine learning classifiers for enhancing knee osteoarthritis diagnosis. *Information*, 15(4), 183. https://doi.org/10.3390/info15040183.

Rehman, S. U., & Gruhn, V. (2024). A sequential VGG16+CNN-based automated approach with adaptive input for efficient detection of knee osteoarthritis stages. *IEEE Access*, *12*, 62407–62415. https://doi.org/10.1109/ACCESS.2024.3395062

Rifat, R. H. (2024). A semi-supervised federated learning approach leveraging pseudo-labeling for knee osteoarthritis severity detection. *http://hdl.handle.net/10361/24036*

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Szabó, S., Holb, I. J., Abriha-Molnár, V. É., Szatmári, G., Singh, S. K., & Abriha, D. (2024). Classification assessment tool: A program to measure the uncertainty of classification models in terms of class-level metrics. *Applied Soft Computing, 155*, 111468. https://doi.org/10.1016/j.asoc.2024.111468

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *In Proceedings of the International Conference on Machine Learning* (pp. 6105-6114). https://doi.org/10.48550/arXiv.1905.11946

Tan, M., & Le, Q. (2021). *EfficientNetV2: Smaller models and faster training*. Proceedings of the 38th International Conference on Machine Learning (ICML 2021), PMLR 139, 10096–10106. https://proceedings.mlr.press/v139/tan21a.html

Watso, J. C., & Vondrasek, J. D. (2024). Risks of exercise in older adults. In *Exercise for aging adults* (pp. 29–45). Springer International Publishing. https://doi.org/10.1007/978-3-031-52928-3_3

Wightman, R., Touvron, H., & Jégou, H. (2021). ResNet strikes back: An improved training procedure in timm. *arXiv*. https://arxiv.org/abs/2103.00388

Yildirim, M., & Mutlu, H. B. (2024). Automatic detection of knee osteoarthritis grading using artificial intelligence–based methods. *International Journal of Imaging Systems and Technology, 34*(2), e23057. https://doi.org/10.1002/ima.23057

Zhao, H., Ou, L., Zhang, Z., Zhang, L., Liu, K., & Kuang, J. (2025). The value of deep learning-based X-ray techniques in detecting and classifying KL grades of knee osteoarthritis: a systematic review and meta-analysis. *European Radiology*, 35(1), 327-340. https://doi.org/10.1007/s00330-024-10928-9

Zhu, S., Qu, W., & He, C. (2024). Evaluation and management of knee osteoarthritis. *Journal of Evidence-Based Medicine, 17*(3), 675–687. https://doi.org/10.1111/jebm.12345