Malaysian Journal of Computing, 9 (2): 1944-1954, 2024

eISSN: 2600-8238

OSTEOARTHRITIS GRADING: A SYNTHESIZED MAGNETIC RESONANCE IMAGES TECHNIQUE

Qiu Ruiyun [1], Siti Khatijah Nor Abdul Rahim[2*], Nursuriati Jamil [3], Raseeda Hamzah[4] and Fu Xiaoling[5]

[1,2,3,4] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia*

[5]*Department of Orthopedics, The Second Affiliated Hospital of Nanchang University,Jiangxi,China*

[1]qryun1228@qq.com, [2*]sitik781@uitm.edu.my, [3]lizajamil@computer.org, [4]raseeda@uitm.edu.my, [5]fxl982@sina.com

ABSTRACT

Osteoarthritis (OA) in the knee is a major cause of decreased activity and physical limitations among older people. Identifying and treating knee osteoarthritis in its early stages can help patients delay the progression of the condition. Currently, early detection of knee osteoarthritis involves the use of X-ray images and assessment using the Kellgren-Lawrence (KL) grading system. Doctors' evaluations can be subjective and may differ among different doctors. Similar to a computer systems analyst, the automatic knee OA grading and diagnosis can be a valuable tool for doctors, enabling them to streamline their workload and provide more efficient care. An innovative network named OA_GAN_ViT has been developed to autonomously detect knee OA. The network is a ViT architecture consisting of two branches: one branch utilizes the synthesized MR image derived from X-ray images for data processing before classification operations via the GAN network, while the other branch employs a histogram-equalized X-ray image. The OA_GAN_ViT network demonstrated superior performance in terms of accuracy and MAE compared to well-known neural networks such as ResNet, DenseNet, VGG, Inception, and ViT. It achieved an impressive accuracy of 79.2 and an MAE of 0.492, highlighting its effectiveness.

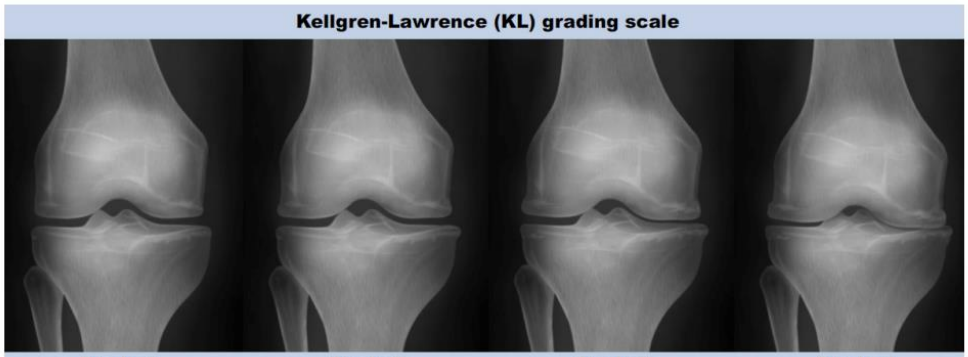
Keywords*: Deep Learning, Multimodal Synthesis, OA Grading, Pre-process.*

Received for review: 02-09-2024; Accepted: 27-09-2024; Published: 01-10-2024
DOI: 10.24191/mjoc.v9i2.27015

1. Introduction

Knee OA is a prevalent kind of arthritis that primarily affects aged people, leading to reduced physical activity and impairment. It is characterized by the deterioration of articular cartilage (Wesseling et al., 2009). The presence of OA significantly impairs the overall quality of life due to the considerable impact of pain and other associated symptoms. Regrettably, there is

now no medication available that can prevent the degenerative structural changes that occur throughout the course of knee osteoarthritis. Nevertheless, timely identification and intervention might assist elderly individuals in postponing the advancement of OA and enhancing their overall well-being. Imaging examinations, such as X-rays and magnetic resonance imaging (MRI), are crucial tools for assessing the course of OA. MRI provides a depiction of the knee joint's three-dimensional configuration. Nevertheless, MRI is solely accessible in expansive medical facilities, and the exorbitant expense of the examination renders it impractical for routine diagnosis of knee OA (Li et al., 2020). Contrarily, X-ray examination is known for its safety, affordability, and widespread use, and has consistently been regarded as the benchmark for knee osteoarthritis evaluation. The Kellgren-Lawrence (KL) grading system, established by the World Health Organization in 1961, is the predominant method for assessing the severity of knee OA (Kellgren & Lawrence, 1957). The KL method categorizes the severity of knee OA into five grades, denoted by grades 0 to 4, in ascending order. Figure 1 displays the KL grading system. Medical professionals typically examine scanned X-ray images of the knee joint and assign a KL grade to assess its condition. The accuracy of this diagnosis heavily relies on the doctor's expertise and attentiveness. Misclassification can result in incorrect treatment methods and prognosis. Therefore, the implementation of computer-aided technology to automatically grade the KL level would greatly enhance the diagnosis of OA (Saini et al., 2023).



CLASSIFICATION	Normal	Doubtful	Mild	Moderate	Severe
DESCRIPTION	No features of OA	Minute osteophyte: doubtful significance	Definite osteophyte: normal joint space	Moderate joint space reduction	Joint space greatly reduced: subchondral sclerosis

Figure 1. The KL grading system to assess the severity of knee OA (Antony et al., 2017)

Starting in 2009, significant progress was made in classifying the KL grade of knee osteoarthritis. Shamir et al. (2009) introduced a weighted closest neighbor approach that necessitates the human creation of characteristics, such as texture features, Chebyshev statistics, and Haralick texture features, *etc* (Shamir et al., 2009).

Recently, deep learning has led to the emergence of well-established convolutional neural network (CNN) classification models, including ResNet (He et al., 2016), VGG (Simonyan & Zisserman, 2014), Inception-V3 (Szegedy et al., 2015), and DenseNet (Huang et al., 2017). ResNet is easier to optimize and performs better in terms of generalization on recognition tasks. It was shown to make considerable improvements on the CIFAR-10 and COCO datasets, with a 28% relative improvement on the COCO object detection dataset. In the ILSVRC-2014 competition, the VGG model secured first place in the localization job and second place in the classification challenge on the ImageNet dataset. The VGG architecture of the model, which employs 16–19 layers, greatly increased recognition of images accuracy. Inception-V3 deep neural network achieved dermatologist-level accuracy in classifying skin cancer using a dataset of 129,450 clinical images. It matched the performance of 21 board-certified dermatologists in distinguishing malignant melanomas and carcinomas from benign

lesions, using biopsy-proven clinical images, in medical classification jobs, it is commonly employed. By tackling the vanishing gradient problem and improving feature reuse with fewer parameters, the DenseNet model showed notable improvements in object identification tasks on several datasets, including CIFAR-10, CIFAR-100, SVHN, and ImageNet. It is intended to promote feature reuse and enhance feature propagation. Antony and his colleagues employed deep learning techniques to classify KL in their study. They developed a novel CNN architecture and enhanced the weighted combination of cross-entropy loss and mean square error loss, resulting in a recognition rate of 63.6% (Antony et al., 2016, 2017, 2020). Gorriz et al. (2019) introduced a new CNN framework that can automatically measure the extent of knee arthritis using X-ray pictures. This framework achieved a recognition accuracy of 64.3%. Tiulpin et al. (2018) employed a deep convolutional network to assess the extent of knee osteoarthritis and obtained a success rate of 66.7%. Zhang et al. (2020) introduced a knee osteoarthritis Kellgren-Lawrence grade classification model using a CNN and an attention mechanism. The ResNet model was initially employed to extract knee joint characteristics from X-ray images. These characteristics were then coupled with the information retrieved by the convolutional attention module to automatically estimate KL grades.

The main contributions of the proposed method are as follows:

(1) An end-to-end X-ray images to MR image synthesis network is proposed.

(2) We propose an OA_GAN_ViT network, a GAN and ViT hybrid network, for automatic grading of knee OA severity, and demonstrate through extensive experiments that we can achieve 79.2% classification accuracy and the MAE of 0.492 with this approach after possible pre-processing.

2. Method

The OA_GAN_ViT network proposed consists of two parts, the first part is demonstrated in section 2.2. The step is to synthesize X-ray images to MRI images by using a GAN network. The second step is to grade the knee OA by using the synthesized MRI image as elaborated in section 2.3. The details of each step will be introduced in the sections below.

2.1 Dataset

The dataset used in this study is composed of two datasets.

One dataset is utilized for training, verification, and testing the source from the Osteoarthritis Initiative (OAI) (Peterfy et al., 2008). This program serves as a permanent repository for the clinical data, patient-reported outcomes, biospecimen studies, quantitative image analyses, radiographs (X-Rays), and MRI. The X-ray is often be called CR(Computed Radiography) in some studies. This repository contains longitudinal assessments and measures from 4,796 participants, encompassing data from over 431,000 clinical and imaging sessions, and a total of over 26,626,000 pictures.

The other dataset utilized for evaluation is the data collected from the orthopedics department of the second affiliated hospital of Nanchang University. The dataset includes the paired X-ray and MRI images of 200 patients.

2.2 X-ray Images Synthesize to MRI Images

2.2.1 Preprocessing

We screened out a total of 72 sets of images from the dataset. The X-ray images of these 72 patients were highly consistent with their MRI images, and all were scanned in Sag T1 FSE (iQMR) format. MRI images generally include data from 1 to 20 layers, and all images are DCM files. We used X-ray images and MR images of the 7th or 8th layer as training sets to train the generative network. Among them, the MR images of the 7th layer had the best effect.

The selected data set will be preprocessed as follows.

1. Input the DCM data and cut off the excess parts to make the width and height consistent. The original X-ray image size is generally about 1988*2300. The height is greater than the width, and the excess parts at the top and bottom are cut off. The original MR image size is generally about 576*574. When the height is greater than the width, the excess parts are cut from the top and bottom. When the height is less than the width, the excess parts are cut from the left and right.

Due to the grayscale range contained in the DCM file is larger than that of PNG data, and the grayscale details are more distinct, there is no need to convert the DCM data into PNG data. The reason for cropping to the same width and height is that when the image is stretched proportionally, the texture details at the joints will not be distorted.

2. Scale the data proportionally to 256*256.

3. Perform histogram equalization on the data. The reason for using histogram equalization is to make the joints clearer.

2.2.2 Synthesis Process

The training process of synthesize X-ray to MR is shown in Figure 2. This process is based on CycleGAN (Zhu et al., 2020). The details of the network are illustrated below.

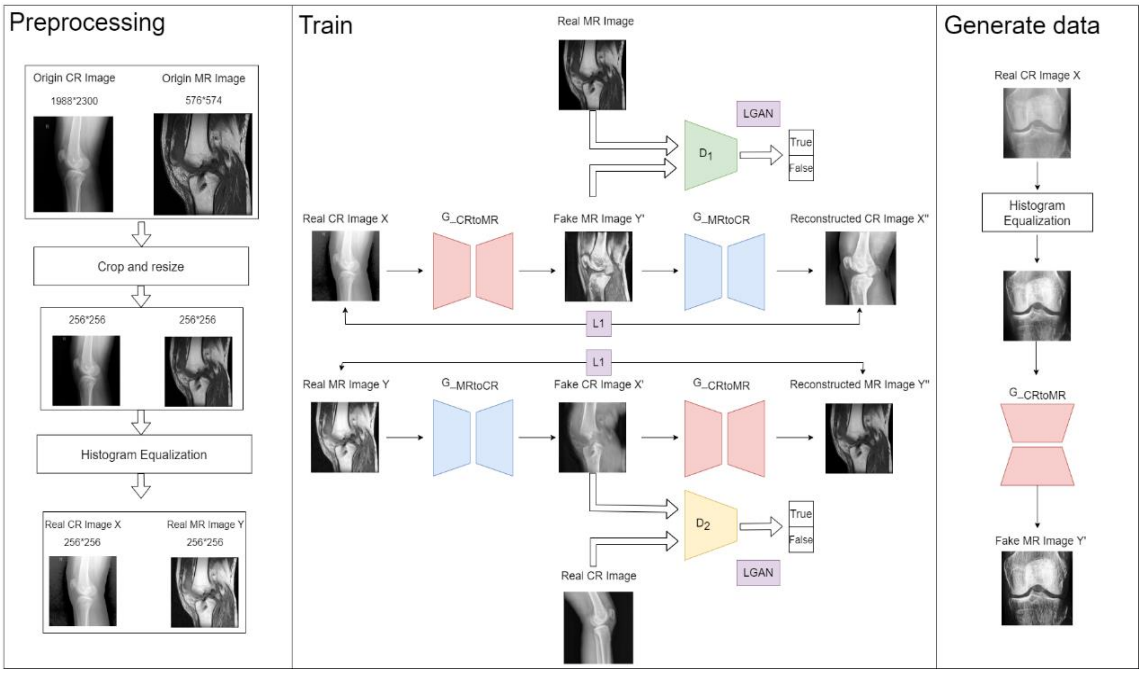


Figure 2: The Synthesis process from X-ray to MR

The model consists of two generators and two discriminators.

The generator uses a 4-layer U-Net network, inputs a 256*256 image, and outputs 256*256 [0,1] distributed data. The data is "reverse normalized", the [0,1] distributed tensor data is mapped to the interval [0,255] and converted to a grayscale image for output. The discriminator uses a 4-layer fully convolutional network, inputs a 256*256 image, and outputs 32*32 [0,1] distributed data. The cross-entropy calculation is performed on the data and the 32*32 all-0 or all-1 labels, which are used as adversarial losses to update the generator and discriminator.

The architecture of the generator and the discriminator are shown in Figure 3. The generator generates the MR image from the features selected from the X-ray image, then, the discriminator tells the synthesized image from the truth from the coordinate X-ray image.

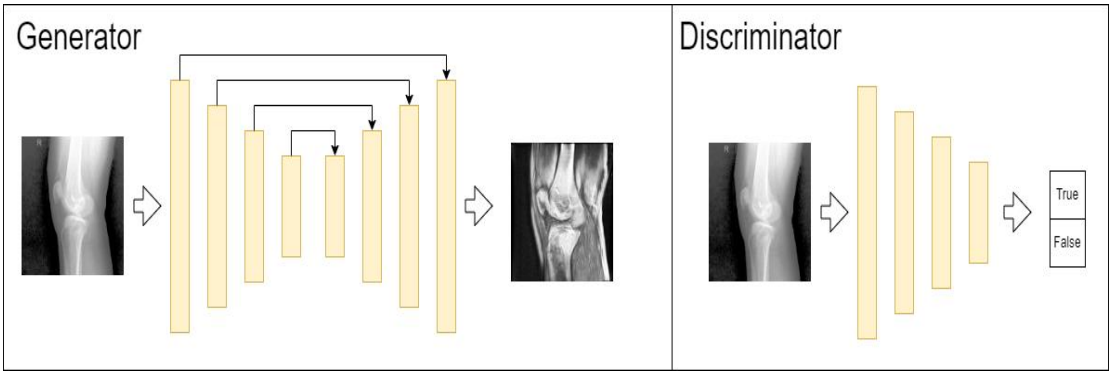


Figure 3 The Generator and the Discriminator

2.2.3 Loss Function

The loss functions are composed of cycle consistency loss and adversarial loss.
Cycle consistency loss is shown in Equation 1,

$$L_{cycle}(X,Y) = |Y - X| \quad (1)$$

Adversarial loss is demonstrated in Equation 2,

$$L_{GAN}(G,D,X) = \log D(X) + \log(1 - D(G(X))) \quad (2)$$

The loss function is shown in Equation 3,

$$L_{full} = L_{cycle}(X,X'') + L_{cycle}(Y,Y'') + L_{GAN}(G_{CRtoMR},D_1,X) + L_{GAN}(G_{MRtoCR},D_2,Y) \quad (3)$$

2.2.4 The Output of the Data for OA Grading

After training the generative network model, the trained generator $G_{\_MRtoCR}$ is used to enhance the kneeKL224 dataset, generating MR images for the 7335 training set X-ray images and the 815 test set X-ray images. Before the images are put into the generator, histogram averaging is also performed. Since the dataset is already 256*256 in size, no additional stretching is performed.

2.3 OA Grading Using the Synthesized MR Image

2.3.1 OA_GAN_ViT Network

A network named OA_GAN_ViT is proposed to perform the OA grading using the synthesized MR Images. The framework of the network is illustrated in Figure 4, and the details of the network are demonstrated below.

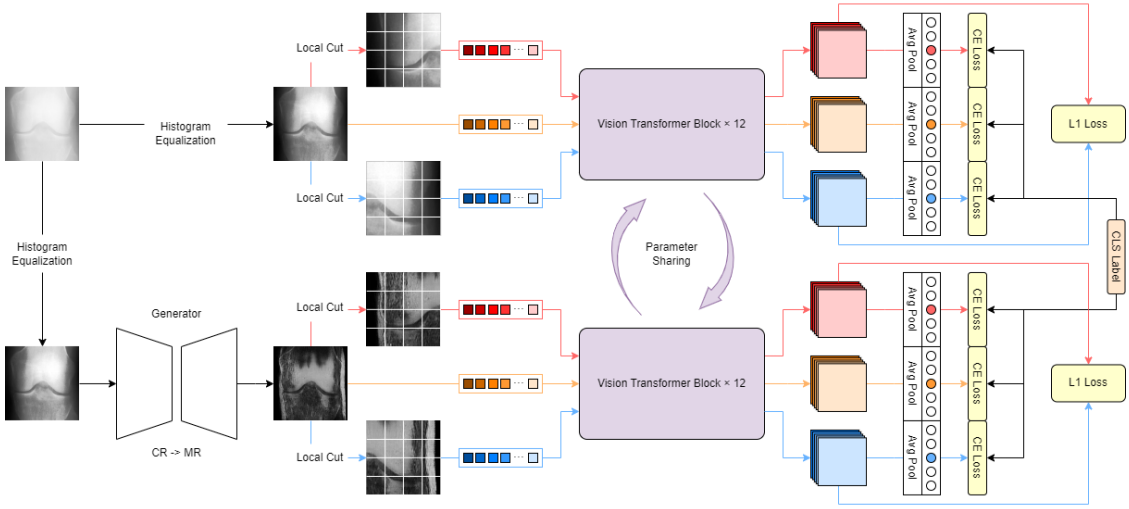


Figure 4 The Framework of OA_GAN_ViT Architecture

The input image of OA_GAN_ViT is composed of two branches. One branch is a X-ray image that has been histogram equalized. It must be histogram equalized because the original image is not clear, histogram equalization can make the image clearer and highlight the texture of the joints. The equalization method used is from OpenCV. The other branch is the MR image synthesized from the steps mentioned in section 2.2. Before inputting into the network, the image is first partially cut, the joint gap is taken, and then the joint is cut in half. Assuming that the size of the complete image is (224, 224), the left and right joint blocks correspond to (56:168, 0:112) and (56:168, 112:224) in the image respectively. Then the right joint gap is flipped, thus obtaining the complete original image, the left joint gap image, and the right joint gap flipped image.

The same operation is performed on these three images. First, Patch Embedding is performed to obtain their respective Patch Tokens, and then Class Tokens are spliced, and Position Embedding is performed. The process here uses the original ViT operation method. The Patch Embedding operation is to use a convolution with a kernel of 16 and a step length of 16 for cutting. The image size changes from (B, 3, 224, 224) -> (B, 768, 14, 14), and then flatten and transpose operations are performed. The size changes from (B, 768, 14, 14) -> (B, 196, 768). Class Token is a set of learnable parameters with a size of (B, 1, 768). After splicing with the image, the size is (B, 197, 768). Position Embedding is also a set of learnable parameters with a size of (B, 197, 768), which is directly added to the image size, so the final size is (B, 197, 768).

After that, they are put into the ViT network. The ViT network calls the pre-trained parameters. The three pictures are put into ViT respectively, which can be understood as ViT with shared parameters. There are 12 ViT blocks in total. Each ViT is LN+MHSA+LN+FFN, which is consistent with the original ViT. After this process, each outputs a feature map from the ViT network. The size of the feature map is (B, C, 14, 14). The prediction result is obtained by averaging the feature map and then squeezing it. The size is (B, C). Finally, the classification loss is calculated for each prediction result and the classification label, and the L1 loss is calculated for the feature map of the left joint gap map and the right joint gap flip map.

2.3.2 Loss Function

The loss function is shown as follows.

$$l_{cr\text{-}cls} = l_{cr\text{-}lcls} = l_{cr\text{-}rcls} = -\frac{1}{C}\sum_{C=1}^{C}\left[l_c \log\left(\frac{1}{1+e^{-s_c}}\right) + \left(1-l_c\right)\log\left(\frac{e^{-s_c}}{1+e^{-s_c}}\right)\right] \tag{4}$$

$$l_{mr\text{-}cls} = l_{mr\text{-}lcls} = l_{mr\text{-}rcls} = -\frac{1}{C}\sum_{C=1}^{C}\left[l_c \log\left(\frac{1}{1+e^{-s_c}}\right) + \left(1-l_c\right)\log\left(\frac{e^{-s_c}}{1+e^{-s_c}}\right)\right] \tag{5}$$

$$l_{ptp} = \left\| F_l - F_r \right\|_1 \tag{6}$$

$$L_{cr} = l_{cr\text{-}cls} + l_{cr\text{-}lcls} + l_{cr\text{-}rcls} + l_{cr\text{-}ptp} \tag{7}$$

$$L_{mr} = l_{mr\text{-}cls} + l_{mr\text{-}lcls} + l_{mr\text{-}rcls} + l_{mr\text{-}ptp} \tag{8}$$

$$L = L_{cr} + L_{mr} \tag{9}$$

Among them, the labels l_c and s_c show the results of the guess for class c. The number of groups is C, the feature map for the left joint gap map is F_l, and the feature map for the right joint gap map is F_r.

The classification loss for the whole image is shown by l_{cls}. l_{lcls} is the classification loss for the left joint gap map, l_{rcls} is the classification loss for the right joint gap map, and l_{ptp} is the equivalence loss used to calculate L1 loss. The equation with a cr prefix is the loss of the cr branch, and the one with mr prefix is the loss of the MR branch. The letter L stands for the loss of the complete structure.

3. Results and Discussions

3.1 Evaluation Metrics

This study primarily employs two assessment metrics, accuracy (ACC) and mean absolute error (MAE), to assess the outcomes of the network. The next section offers a comprehensive overview of these two metrics.

3.1.1 ACC

Accuracy is the proportion of correct forecasts compared to all other predictions. An ideal model should possess a high level of accuracy. Nevertheless, achieving a high or 100% accuracy on a machine learning model does not necessarily indicate a well-constructed model. Instead, it often suggests a problem, such as overfitting (Hui et al., 2023). Mathematically, accuracy can be defined in Equation 10.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{10}$$

where, *tp* represents True Positive, *tn* stands for True Negative, *fp* indicates False Positive, and *fn* refers to False Negative.

3.1.2 MAE

As accurac**y** is merely a criterion for evaluating the efficacy of a model, physicians are more concerned with the distance between the incorrect category and the correct category when assessing the KL grading task (Chen et al., 2019). For instance, in the case where the actual label is zero, the cost of a misjudgment of four is unquestionably considerably greater than the cost of a misjudgment of one; therefore, an alternative metric is required to quantify the cost of a misjudgment. The mean absolute error can describe this point very well, and its definition is shown in Equation 11.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|P_i - T_i\right| \tag{11}$$

Where P_i stands for the prediction, T_i represents the true value, and n refers to the total number of data points. It is evident that a decreased mean absolute error (MAE) corresponds to a reduced cost of model misjudgment. Lower MAE values indicate more accurate predictions, which is important in reducing the risk of large misclassifications that can significantly impact the treatment of patients with varying severity of OA**.** The MAE is particularly significant because it reflects the overall accuracy of the model by quantifying how far predictions deviate from true values.

3.2 Results

3.2.1 The Comparison of Different Kinds of Networks

Table 1 displays the experimental findings for several networks. CE stands for cross-entropy loss, Ordinal represents the sequence penalty weight loss introduced by Chen et al. (2019), and Liu denotes the approach introduced by Liu et al (2021).

Table 1 The Comparison of Different Kinds of Networks

Model		ACC	MAE
ResNet	ResNet-34-CE	66.4	0.585
	ResNet-34-Ordinal	65.8	0.550
	ResNet-34-Liu	64.3	0.513
	ResNet-50-CE	65.2	0.618
	ResNet-50-Ordinal	62.7	0.591
	ResNet-50-Liu	63.2	0.563
DenseNet	DenseNet-161-CE	66.6	0.519
	DenseNet-161-Ordinal	65.2	0.524
	DenseNet-161-Liu	66.9	0.515
VGG	VGG-16-Proposed	69.1	**0.428**
	VGG-19-Ordinal	68.1	0.450
Inception	Inception-V3-CE	66.5	0.563
	Inception-V3-Liu	65.8	0.506
ViT	Original dataset	72.9	0.574
	Augment dataset	68.3	0.489
Dual-ViT	Augment dataset	78.4	0.471
OA_GAN_ViT	ours	**79.2**	0.492

The suggested network does the best in terms of accuracy, but only the fourth best in terms of MAE, as shown in the table.

3.2.2 The Performance of Different Kinds of Pre-process

Table 2 shows the performance of ViT and the proposed dual-ViT network when utilizing different pre-processing methods.

Table 2 The Performance of Using Different Pre-processing Methods

Model	Dataset	ACC	MAE
ViT	origin_onlygap	69.6	0.574
	Origin_fullfilling	72.9	0.612
	Eqhist_fullfing	70.6	0.594
	Eqhist_onlygap	70.4	0.510
	cutout	66.1	0.524
	Eqhist_cutout_flip	68.3	0.489
Dual-ViT	origin_onlygap	71.4	0.562
	Origin_fullfilling	69.9	0.549
	Eqhist_fullfing	72.3	0.553
	Eqhist_onlygap	71.4	0.502
	cutout	66.4	0.510
	Eqhist_cutout_flip	78.4	0.471
OA_GAN_ViT	origin_onlygap	62.2	0.689
	Origin_fullfilling	63.6	0.512
	Eqhist_fullfing	**79.2**	0.492
	Eqhist_onlygap	70.1	0.509
	cutout	73.2	0.677
	Eqhist_cutout_flip	69.3	0.515

The table shows that the OA_GAN_ViT network works best when the data has only been histogram equalized when pre-processing. As elaborated in section 3.1.2, MAE play a significant part in evluate the knee OA grading network performance, but the highest accuracy still is the best performance instead of the lowest MAE, because the highest accuracy is considered the best performance because it directly reflects how well the model can correctly classify the knee OA severity into the correct KL grade. While MAE measures the average deviation between predicted and actual grades, accuracy emphasizes the model's ability to make exact predictions. In clinical settings, exact predictions are crucial for determining the

appropriate treatment, making accuracy a more significant metric in this context. So, OA_GAN_ViT works well with only histogram equalized data pre-processed.

4. Discussion

The OA is divided into five grades, and the key area for distinguishing them is at the joint position, so the proposed network is specially designed for feature extraction at the joint position. The network extracts features from the complete image in order to obtain the global features of the image, and extracts features from the joint gap in order to capture the key discriminant features at the joint gap. Since the joint gap is relatively symmetrical, feature extraction is performed on the left and right sides respectively to reduce the difference in perception between the two sides and balance the feature extraction capabilities on both sides of the gap. Through this design, the ability to extract key features at the joint gap is enhanced, which can improve the ability to distinguish OA grades.

5. Conclusions

We introduced a method for knee OA grading utilizing X-ray synthesized MR images and proposed OA_GAN_ViT, a new end-to-end network. The classification operations of this network are initially executed on synthesized MR images and histogram-equalized X-ray images via the OA_GAN_ViT network. The experimental results demonstrate that the proposed network outperforms other neural networks including ResNet, DenseNet, VGG, Inception, and ViT in terms of both accuracy and MAE metrics. This suggests that the network is indeed effective.

In comparison to alternative networks, there is still potential for enhancing the MAE of the proposed method. In subsequent investigations, we aim to enhance the approach and augment the precision of model recognition through the fortification of the adaptive strategy and the incorporation of model combination methods. The KL automatic osteoarthritis grading system provides a more accurate auxiliary diagnosis.

Acknowledgements

The authors would like to thank Universiti Teknologi MARA for the continuous support and encouragement in conducting this research project.

Author Contribution

The authors have been involved equally in every section when writing this paper.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Antony, J., McGuinness, K., Moran, K., & O'Connor, N. E. (2017). *Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks*. 376–390.

Antony, J., McGuinness, K., Moran, K., & O'Connor, N. E. (2020). Feature learning to automatically assess radiographic knee osteoarthritis severity. *Deep Learners and Deep Learner Descriptors for Medical Applications*, 9–93.

Antony, J., McGuinness, K., O'Connor, N. E., & Moran, K. (2016). Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1195–1200. https://doi.org/10.1109/ICPR.2016.7899799

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks*. 4700–4708.

Hui, S. H., Khai, W. K., XinYing, C., & Wai, P. W. (2023). Prediction of customer churn for ABC Multistate Bank using machine learning algorithms/Hui Shan Hon...[et al.]. *Malaysian Journal of Computing (MJoC)*, *8*(2), 1602–1619.

Kellgren, J. H., & Lawrence, J. S. (1957). Radiological Assessment of Osteo-Arthrosis. *Annals of the Rheumatic Diseases*, *16*(4), 494–502. https://doi.org/10.1136/ard.16.4.494

Li, D., Li, S., Chen, Q., & Xie, X. (2020). The Prevalence of Symptomatic Knee Osteoarthritis in Relation to Age, Sex, Area, Region, and Body Mass Index in China: A Systematic Review and Meta-Analysis. *Frontiers in Medicine*, *7*, 304. https://doi.org/10.3389/fmed.2020.00304

Liu, W., Luo, linkai, Peng, H., Zhang, Q., & Huang, W. (2021). Grading scoring of knee osteoarthritis based on adaptive ordinal penalty weighted deep neural networks. *Chinese Journal of Scientific Instrument*, *42*(7), 145–154. https://doi.org/10.19650/j.cnki.cjsi.J2107390

Peterfy, C. G., Schneider, E., & Nevitt, M. (2008). The osteoarthritis initiative: Report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and Cartilage*, *16*(12), 1433–1441. https://doi.org/10.1016/j.joca.2008.06.016

Saini, D., Khosla, A., Chand, T., Chouhan, D. K., & Prakash, M. (2023). Automated knee osteoarthritis severity classification using three-stage preprocessing method and VGG16 architecture. *International Journal of Imaging Systems and Technology*, ima.22845. https://doi.org/10.1002/ima.22845

Shamir, L., Ling, S. M., Scott, W. W., Bos, A., Orlov, N., Macura, T. J., Eckley, D. M., Ferrucci, L., & Goldberg, I. G. (2009). Knee X-Ray Image Analysis Method for Automated Detection of Osteoarthritis. *IEEE Transactions on Biomedical Engineering*, *56*(2), 407–415. https://doi.org/10.1109/TBME.2008.2006025

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). *Going deeper with convolutions*. 1–9.

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), 1727.

Wesseling, J., Dekker, J., van den Berg, W. B., Bierma-Zeinstra, S. M. A., Boers, M., Cats, H. A., Deckers, P., Gorter, K. J., Heuts, P. H. T. G., Hilberdink, W. K. H. A., Kloppenburg, M., Nelissen, R. G. H. H., Oosterveld, F. G. J., Oostveen, J. C. M., Roorda, L. D., Viergever, M. A., Wolde, S. ten, Lafeber, F. P. J. G., & Bijlsma, J. W. J. (2009). CHECK (Cohort Hip and Cohort Knee): Similarities and differences with the Osteoarthritis Initiative. *Annals of the Rheumatic Diseases*, *68*(9), 1413–1419. https://doi.org/10.1136/ard.2008.096164

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks* (arXiv:1703.10593). arXiv. http://arxiv.org/abs/1703.10593