Computer Methods and Programs in Biomedicine 242 (2023) 107807

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Stratifying knee osteoarthritis features through multitask deep hybrid learning: Data from the osteoarthritis initiative ☆

Yun Xin Teoh [a,b], Alice Othmani [b,*], Khin Wee Lai [a,*], Siew Li Goh [c,d], Juliana Usman [a]

[a] Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia
[b] LISSI, Université Paris-Est Créteil, Vitry sur Seine, 94400, France
[c] Sports Medicine Unit, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, 50603, Malaysia
[d] Centre for Epidemiology and Evidence-Based Practice, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, 50603, Malaysia

ARTICLE INFO

Keywords:
Deep hybrid learning
Computer-aided diagnosis
Joint-space narrowing
Knee osteoarthritis
Knee pain
Osteophytes

ABSTRACT

Background and objective: Knee osteoarthritis (OA) is a debilitating musculoskeletal disorder that causes functional disability. Automatic knee OA diagnosis has great potential of enabling timely and early intervention, that can potentially reverse the degenerative process of knee OA. Yet, it is a tedious task, concerning the heterogeneity of the disorder. Most of the proposed techniques demonstrated single OA diagnostic task widely based on Kellgren Lawrence (KL) standard, a composite score of only a few imaging features (i.e. osteophytes, joint space narrowing and subchondral bone changes). However, only one key disease pattern was tackled. The KL standard fails to represent disease pattern of individual OA features, particularly osteophytes, joint-space narrowing, and pain intensity that play a fundamental role in OA manifestation. In this study, we aim to develop a multitask model using convolutional neural network (CNN) feature extractors and machine learning classifiers to detect nine important OA features: KL grade, knee osteophytes (both knee, medial fibular: OSFM, medial tibial: OSTM, lateral fibular: OSFL, and lateral tibial: OSTL), joint-space narrowing (medial: JSM, and lateral: JSL), and patient-reported pain intensity from plain radiography.

Methods: We proposed a new feature extraction method by replacing fully-connected layer with global average pooling (GAP) layer. A comparative analysis was conducted to compare the efficacy of 16 different convolutional neural network (CNN) feature extractors and three machine learning classifiers.

Results: Experimental results revealed the potential of CNN feature extractors in conducting multitask diagnosis. Optimal model consisted of VGG16-GAP feature extractor and KNN classifier. This model not only outperformed the other tested models, it also outperformed the state-of-art methods with higher balanced accuracy, higher Cohen's kappa, higher F1, and lower mean squared error (MSE) in seven OA features prediction.

Conclusions: The proposed model demonstrates pain prediction on plain radiographs, as well as eight OA-related bony features. Future work should focus on exploring additional potential radiological manifestations of OA and their relation to therapeutic interventions.

1. Introduction

Osteoarthritis (OA) is the 11th most common leading cause of disability worldwide [1], and knee accounts for the major burden among the affected joints [2] (Fig. 1). The lifetime risk of developing symptomatic knee OA is 13.8% [3], where females and residents in rural area are the two high-risk populations [4]. Knee OA not only exists in elders as a result of aging cartilage tissues, it also affects young individuals as a result of post-trauma or joint overuse [5].

☆ In this work, we demonstrated several multitask models specifically for automatic knee osteoarthritis (OA) diagnosis using deep hybrid learning. 16 pretrained convolutional neural network (CNN) architectures from VGG, EfficientNet, ResNet, and DenseNet families were employed as feature extractors before passing a KNN classifier. A novel feature extraction method was proposed using global average pooling (GAP) layer as final layer of CNN. The findings of this study will pave the way for future development in the field, facilitating the development of more accurate diagnostic tools for medical image classification which could generate valuable insight into OA disease pathology.

* Corresponding authors.
E-mail addresses: alice.othmani@u-pec.fr (A. Othmani), lai.khinwee@um.edu.my (K.W. Lai).

https://doi.org/10.1016/j.cmpb.2023.107807
Received 18 January 2023; Received in revised form 2 August 2023; Accepted 8 September 2023
Available online 20 September 2023

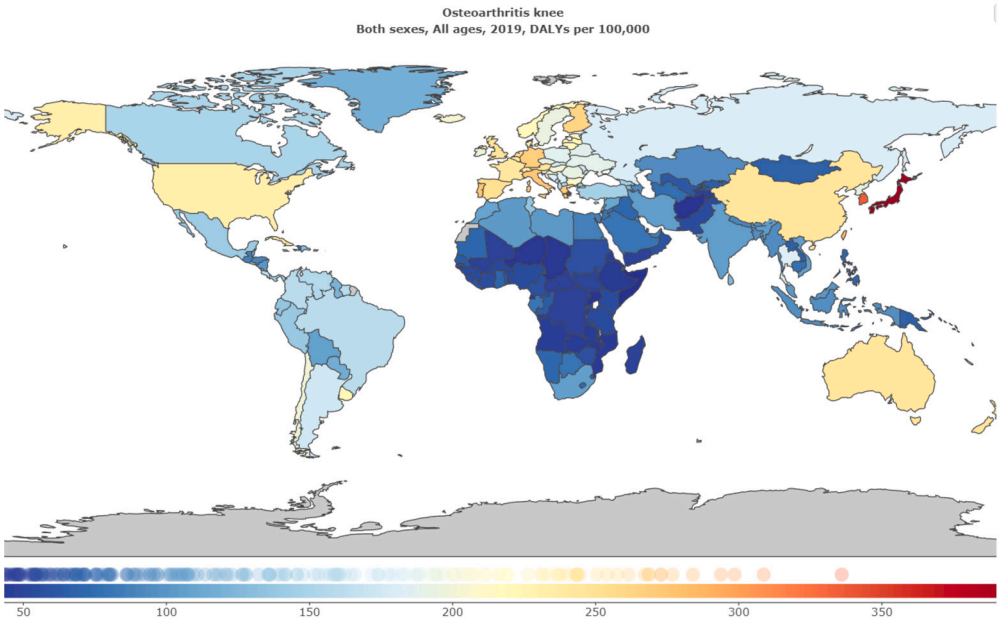

Fig. 1. This figure illustrates the global burden of knee OA as measured using disability-adjusted life years (DALYs).
Source: https://vizhub.healthdata.org/gbd-compare.

Knee OA manifests with cartilage degeneration [6], joint-space narrowing (JSN) [7], bone spurs (osteophytes), and development of bony deformities, that could be visually inspected through X-ray imaging modality. Patients with symptomatic knee OA will experience chronic knee pain, consistent joint stiffness and swelling, and poor functional capability [8]. At present, there is no cure for this disease, currently available interventions only work on providing a temporary pain relief. Yet, there is a scientific report showing that pre-OA is a reversible process [9], further recommending early OA diagnosis as a potential strategy to minimize clinical OA progression.

Early OA diagnosis possesses a significant challenge for medical experts and artificial neural networks. The main reason is the lack of radiographic knowledge indicating OA onset. At the early stage of OA, patients only have subtle change in joint structure. Medical experts usually confuse whether to grade it as normal or doubtful cases, and it is time-consuming for them to reach final decision. Computer-aided diagnosis methods based on deep learning offer a great opportunity to improve the aforementioned situation through automatic OA grading based on multi-feature learning.

Classification of OA symptoms is significant for knee OA diagnosis in terms of defining the severity of disease. Kellgren Lawrence (KL) rating [10] is the gold standard for knee OA diagnosis. It is a "top-down" classification system that classifies patient radiographs into five OA developmental stages. KL grade 0 indicates absence of OA radiologic features; grade 1 indicates doubtful OA; grade 2 indicates mild OA; grade 3 indicates moderate OA; whereas grade 4 indicates severe OA. Despite the wide-use of KL rating in clinical settings, KL grading scheme is prone to inter- and intra-reader variability [11]. Moreover, KL grade is a composite score, and did not focus on interpreting individual features and their anatomical sides (lateral or medial). To address the drawbacks of KL rating, Osteoarthritis Research Society International (OARSI) atlas [12] was proposed. OARSI enables grading of individual OA features, such as femoral osteophytes, tibial osteophytes, and joint space narrowing in a compartment-wise behavior.

Despite the aforementioned OA grading standards which are parts of interpretation from radiologists, another important OA indicator is patient-reported pain intensity. Challenge exists where the pain manifestation is not in parallel with KL rating. Despite the outcomes inconsistency, clinicians frequently use pain indicator to evaluate the health status of OA patients, as well as assessing the efficacy of the prescribed intervention in clinics. To the best of our knowledge, there is no deep learning model specifically for prediction of pain indicator. Existing projects only looked for correlation between pain and other risk factors [13,14], but they failed to reach consensus on developing a pain prediction model in knee OA.

In this paper, we proposed a multitask model based on convolutional neural network (CNN) feature extractors and machine learning classifiers to detect nine crucial OA features. Our study will be presented in the following structure. Section 2 described the motivation of the study. Related work was presented in Section 3. The proposed methodology was explained in Section 4. In Section 5, details about experimental outcomes were provided. The performance of deep knee OA feature learning of the several frameworks were evaluated. After presenting results and discussion, section 6 concluded the work.

2. Motivations

A few narrative reviews have summarized the emerging computerized knee OA diagnosis approaches [15–17]. To the best of our knowledge, there is no comparative analysis studying the performance of the most popular machine learning (ML) and deep learning (DL) models with the existing state-of-the-art hybrid ML-DL architectures.

The main contributions of this work are as follows:

- A comparative analysis of the performances of several popular deep neural networks for the stratification of knee osteoarthritis features from radiography images,
- Utilization of pretrained convolutional neural networks (CNNs) and global average pooling (GAP) as feature extractors for automatic knee OA diagnosis,
- Development of multitask deep hybrid learning models for multi-OA-feature diagnosis using machine learning classifiers,
- Development of deep hybrid learning models for knee pain classification from plain radiography.

3. Related work

3.1. Traditional knee OA feature learning

The hand-crafted individual knee OA imaging features have been studied comprehensively by researchers. First related work was initi-

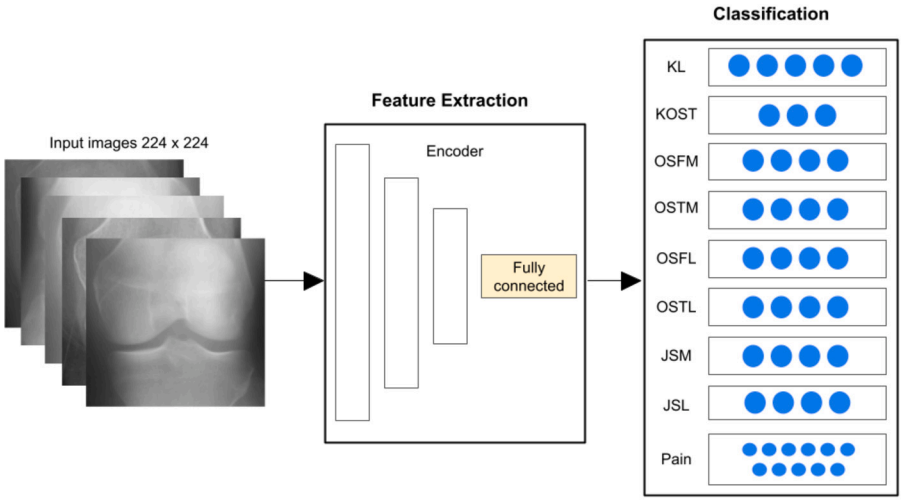

Fig. 2. Schematic representation of the workflow of the project.

ated by Oka et al. [18] for computation of knee OA features. In the later year, Thomson et al. [19] proposed a more robust setup methodology for evaluation of the existence of osteophytes and radiographic OA (KL $\geq$ 2) using shape and texture descriptors. The proposed approach obtained area under the receiver operating characteristic (AUROC) curve for detection of osteophytes as high as 0.85. Despite the good model performance, there were two critical limitations found in the study. First limitation was the usage of small test set size in the study. Second limitation was the doubtful clinical relevance of the classification model. The authors presented binary classification between osteophytes of OARSI grades 0 to 1 and 2 to 3, however, osteophytes could also develop in grade 1 patients. Saleem et al. [20] proposed an automatic joint width measuring approach using Canny edge detection algorithm and yielded highest accuracy at 0.9714 for OA discrimination. However, the proposed method required a lot of data preprocessing effort, which might be time-consuming.

3.2. Deep knee OA feature learning

In deep learning, imaging features are learned from a large pool of OA imaging data through a "bottom-up" hierarchical pattern. The learned features are adopted with a pre-defined classification algorithm for OA grade prediction. The limitations in earlier studies were addressed in the work by Antony [21], where the first CNN-based approach was proposed for simultaneous analysis of KL and OARSI grades. However, the study possessed a significant limitation in generalizing the prediction outcome from two different datasets, Multi-center Osteoarthritis study (MOST) and Osteoarthritis Initiative (OAI). Furthermore, the agreements between the method's predictions and the test set labels were shown to be lower than inter-rater agreements between the human observers for KL and OARSI grades. Tiulpin and Saarakkala [22] tackled those drawbacks by employing a cautious data preprocessing at metadata and image levels. They have produced latest state-of-the-art results in KL grading with Cohen's quadratic kappa of 0.83 and balanced accuracy of 66.71%, and ROC of 0.93 in radiographic OA detection. An excellent agreement on the test set was demonstrated by the authors using same datasets as in [21]. Other related works are contributed by the authors of [23–27] who mainly focused on proposal of new deep learning techniques based on KL severity grading scheme. Tiwari et al. [27] have launched an experiment of transfer learning on knee OA grade prediction using hospital dataset and DensetNet201 was reported as the optimal model. Mahum et al. [28] demonstrated a hybrid approach that combined both traditional feature descriptors and deep feature learning approaches. Combination of CNN, histogram of oriented gradient (HOG) feature descriptor, and k-nearest neighbor (KNN) classifier successively attained approximately 97% accuracy for classification of four KL-based OA grades. Although the reported accuracy was the highest in literature, but this study possessed one limitation as the amount of final classification outputs in this study was inconsistent with other studies. The authors have excluded non-OA condition (KL grade 0) as final classification output.

Most of the studies used standardized X-ray images with reduced visual disturbances through filtering or image processing. However, the use of such approach may also lead to the loss of important diagnostic information. In response to this concern, Olsson et al. [29] proposed a method that used unprocessed multi-view X-ray images with a higher degree of variations. They applied their approach to clinical data collected from a hospital in Sweden and obtained an overall AUC of more than 0.87 for all KL grades, except for KL grade 2, which yielded an AUC of 0.80. The authors observed that the middle KL classes, specifically KL 1, 2, and 3, posed the most challenges for classification. Moreover, they suggested the inclusion of patient symptoms and clinical signs, alongside radiographic findings, into a deep learning network for future work.

4. Methodology

The framework of proposed methodology (Fig. 2) was divided into three parts: data preprocessing, deep learning networks, and machine learning classifiers. Firstly, the images were normalized to generate a set of standardized 224 × 224 input images. Secondly, the preprocessed input images were fed into deep learning networks for feature extraction. Lastly, the extracted features were passed through machine learning classifiers for final classification outcome.

4.1. Data preprocessing

The main objective of data preprocessing is to transform the raw data into a format which is more suitable for a machine learning or deep learning model. In this study, we performed data preprocessing that could provide data normalization. To standardize the data, original 3D-array color images were resized into 224 × 224 and converted to grayscale 2D-array images. All images were then preprocessed and denoised using a low-pass filter, namely Gaussian blur, with a 3 × 3 Gaussian kernel (1).

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

where x and y are the number of pixels from the origin in the horizontal and vertical dimensions, respectively. σ denotes the standard deviation of the Gaussian distribution.

The image contrast after denoising was improved through histogram equalisation (2), (3).

$$p_x = \frac{n_x}{n}, \quad 0 \le x < L \tag{2}$$

where p_x is the normalized histogram for each possible intensity x; n is the number of pixels; and L is the total number of gray levels in the image, usually 256.

$$g_{i,j} = floor((L-1)\sum_{x=0}^{f_{i,j}} p_x) \tag{3}$$

where g is the histogram equalized image; $floor()$ means rounding down the output value to the nearest integer; and f is the input image.

The reason of applying these imaging processing techniques was to enhance image quality, so that the key imaging feature could be more noticeable.

4.2. Deep learning networks

We employed 16 CNNs from VGG, EfficientNet, ResNet and DenseNet families that were pretrained on the ImageNet dataset. In VGG, traditional feature extraction method harvested features after fully-connected (FC) layer. In this study, we proposed a new method by replacing the last two FC layers in VGG architecture with a global average pooling (GAP) layer. Thus, two feature extraction methods would be compared for VGG architecture.

4.2.1. VGG

VGG [30] is a classical CNN architectures, well known for its simplicity. Basically, VGG consists of a few convolutional layers with small receptive field (3x3 with a stride of 1), pooling layers and a fully connected layer. In this study, we used VGG16 and VGG19.

4.2.2. EfficientNet

EfficientNet [31] is a CNN architecture with a novel scaling method where all dimensions of depth, width, or resolution are uniformly scaled by a highly effective compound coefficient. The depth of EfficientNet is between 7 to 27 times deeper than the depth of VGG networks. The main building block in EfficientNet is a inverted linear bottleneck layer with depth-wise separable convolution (MBConv), which is originally known as MobileNetV2, a well-known cost-effective CNN architecture specifically designed for mobile device applications. Thus, the parameter size in EfficientNet is significantly smaller than that in VGG networks. EfficientNetB0 is the baseline network created by automated mobile neural architecture search (MNAS). The baseline network is extended into EfficientB1 to B7 through additional network scaling-up. In this study, EfficientNetB0, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, and EfficientNetB7 were used.

4.2.3. ResNet

ResNet [32] is a CNN architecture that utilizes residual connections, summing the output of a block of layers with its input before passing the input to the subsequent layer. ResNet learns a residual mapping instead of directly fitting a group of stacked layers into a desired underlying mapping. The element-wise feature summation minimizes the percentage of errors as the depth increases. The depth of ResNet is between 5 to 19 times deeper than the networks in VGG. ResNet50, ResNet101, and ResNet152 were used in this study.

4.2.4. DenseNet

DenseNet [33] is a CNN architecture that makes use of dense connections between layers through “Dense Blocks”. All layers are directly connected with each other. Each layer takes additional inputs from previous layer and adds input to the next layer, performing channel-wise concatenation to preserve most of the learned features from all layers. The depth of DenseNet is about 12 to 25 times deeper than VGG models. The testing DenseNet in this study included DenseNet121, DenseNet169, and DenseNet201.

4.2.5. Fully connected layer vs global pooling layer

Fully connected (FC) layers are linear transformation layers that map high-level features extracted by convolutional layers to all neurons in the following output layer [34]. They are a key component in the traditional feature extraction method used by VGG to produce full feature maps, and have been shown to be effective in image classification [35–37]. However, FC layers have the drawback of generating high-dimensional feature vectors, which can lead to overfitting and increased computational cost during training and inference [38].

To address the limitation, global pooling [39] was introduced as a replacement for FC layers. Global pooling condenses all feature maps directly into a single map by applying a simple mathematical operation on each respective feature map. There are two types of global pooling: global average pooling (GAP) and global max pooling (GMP). GAP works by calculating the average of all the values in each feature map, resulting in a single value for each feature map, which are then concatenated to create the final feature vector. Conversely, GMP works by considering the maximum value of each feature map, resulting in a single value for each feature map, which are then concatenated to create the final feature vector. The main advantage of global pooling lies in its ability to optimize the entire CNN network without the need for additional parameters, as pointed out by Guo et al. [38]. Previous research has revealed that GAP [40] is more effective than GMP [41] in extracting dominant OA features when working with knee X-ray images. According to the authors, GAP combines all features from low to high [42,41], whereas GMP only focuses on the highest features and could potentially miss out on crucial ones. Furthermore, GAP has been applied to an MRI-based model designed for deep knee OA feature learning [43] and was found to outperform a shallow model in discriminating OA knees. Since we are interested in utilizing both low and high-level features for model learning, GAP is a suitable choice for our study. Consequently, we integrated GAP into VGG as a technique to enhance model performance.

The current literature recommends using GAP instead of the fully connected (FC) layer for feature extraction in various domains [44,38, 45]. However, there is a lack of experiment to validate this approach in knee X-ray images. Therefore, our study aims to fill this research gap by validating the use of GAP for feature extraction in knee X-ray images and comparing its performance with the FC layer in VGG architecture.

4.3. Machine learning classifiers

Random forest (RF), logistic regression (LR), and K-nearest neighbor (KNN) classifiers were used as the top layer of each CNN model.

4.3.1. Random forest

Random forest (RF) [46] is a supervised learning algorithm that comprised of a collection of decision trees. Each decision tree in the RF ensemble is trained on a random subset of the training data, and a random subset of the features is selected for each split in the tree. In a typical classification task, the output of the RF algorithm is the class selected by most decision trees, or majority voting.

4.3.2. Logistic regression

Logistic regression (LR) [47] is a supervised learning algorithm that utilizes statistical approach and probability calculation to solve classification problems. The dependent variable (target) of this algorithm must be categorical.

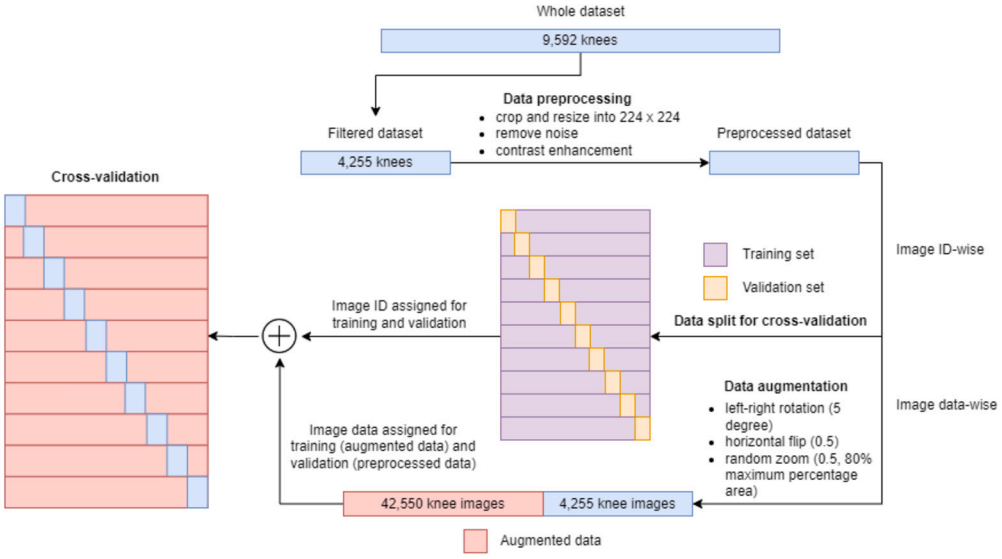

Fig. 3. The figure depicts the flow of data handling in this study.

4.3.3. K-nearest neighbor

K-nearest neighbor (KNN) [48] is a non-parametric supervised learning algorithm that assumes the similarity between new data point and available data points. The algorithm will put the new data point into a category with highest similarity.

5. Experiments and results

In this section, the presented results focused on reviewing experimental outcomes based on fully processed dataset, comparison between feature extraction methods, and comparison between machine learning classifiers.

5.1. Dataset

We utilized publicly available dataset from Osteoarthritis Initiative (OAI) (https://nda.nih.gov/oai) database. The radiographs in the database were annotated according to KL rating scheme and OARSI atlas by a team of recruited radiologists. Pain severity on each knee was measured using a numeric rating scale from 0 to 10 based on patient-perceived pain in past seven days. Our initial database comprised of 4,796 subjects and 9,592 individual knees. However, we conducted data filtering by removing distorted knee images and individual knees with missing data. As a result, we were able to include 4,255 individual knee images from 2,137 subjects, representing 42.3% male and 57.7% female. In terms of ethnicity, 76.5% were White or Caucasian, 21.1% were Black or African American, 1.0% were Asian, and 1.4% identified as other.

Our dataset had a complex multiclass nature, with each task involving a multiclass classification. The data was highly unbalanced in each task, as well as in overall multitask model and considered small for a DL model. To address these challenges, data augmentation was performed. We applied three data augmentation operations with following parameters: rotation with a maximum left and right rotation of 5 degrees, horizontal flipping with a probability of 0.5, and random zooming with a probability of 0.5 and a maximum percentage area of 80%. The factor of amplification was set at 10. After data augmentation, the amount of data was amplified from 4,255 to 42,550. Although the augmented data remained unbalanced with similar ratios as the original data, the increased amount of data for the minority class minimized the impact of the data imbalance issue.

We handled our data using both image ID-wise and image data-wise manipulations (Fig. 3). We partitioned the data into 10 folds based on image ID, and paired each fold with its corresponding images for model training and validation. Before the training phase, data augmentation was applied to all 10 fold sets, resulting in the creation of an original dataset (X) and an augmented version (X') in each fold. Throughout the training process, we used one fold for validation, ensuring that the augmented set (1X') of that fold was removed and excluded from the validation set. The validation set solely comprised the original data (1X) for unbiased evaluation. For model training, we combined the original (9X) and augmented (9X') sets from the remaining nine folds, constructing a training dataset of 9X + 9X' for each iteration. To prevent data leakage, we maintained strict separation between the training data, consisting of both the original and augmented data, and the validation data, containing only the original set. This careful approach ensured complete independence between the training and validation sets, eliminating any potential data leakage between them.

5.2. Implementation details

Our project was developed using Python 3.7.7 programming language in Jupyter Lab 3.4.4. TensorFlow version 2.4.0 and Keras version 2.2.4 were used for developing the deep learning model. NumPy version 1.19.5 was employed for numerical computations, data manipulation, and analytic processing. Scikit-learn version 1.0.2 was used for implementing the multioutput classification with machine learning algorithms. Lastly, the matplotlib version 3.4.3 was used for data visualization and plotting. All CNN networks (Table 1) were initialized with weights based on training outcome on ImageNet. The top layers of pretrained networks were excluded. A new dataset that consisted of feature abstract was generated from the pretrained networks and was used as the input of three testing classifiers (RF, LR, and KNN). The optimal configuration for each classifier was determined through grid search (Table 2).

The X-ray data were trained and validated through KL stratified 10-fold cross validation. In each iteration, the proposed model was trained using nine folds of data. Model performance was evaluated using the remaining one fold of data.

5.3. Evaluation metrics

We evaluated our models using four metrics: balanced accuracy (ACC) [49], F1 score, Cohen's kappa (K) [50], and mean squared error (MSE) [21].

Table 1
Size, depth, and number of features extracted from proposed CNN encoders. GAP: global average pooling, and FC: fully connected. Adopted from [51].

CNN	Size (MB)	Depth	No. of features	
			FC	GAP
VGG16	528	16	4096	512
VGG19	549	19	4096	512
EfficientNetB0	29	132	-	1280
EfficientNetB1	31	186	-	1280
EfficientNetB2	36	186	-	1408
EfficientNetB3	48	210	-	1536
EfficientNetB4	75	258	-	1792
EfficientNetB5	118	312	-	2048
EfficientNetB6	166	360	-	2304
EfficientNetB7	256	438	-	2560
ResNet50	98	107	-	2048
ResNet101	171	209	-	2048
ResNet152	232	311	-	2048
DenseNet121	33	242	-	1024
DenseNet169	57	338	-	1664
DenseNet201	80	402	-	1920

Table 2
Configuration of ML classifiers. lbfgs: limited-memory Broyden–Fletcher–Goldfarb–Shanno.

ML classifiers	Configuration
RF	Number of estimators = 100
LR	Solver = lbfgs
KNN	Number of neighbors = 5

ACC (4) is an improved version of the standard accuracy metric that is optimized to handle imbalanced datasets more effectively. It is calculated as the average of the recall obtained on each class.

$$\mathrm{ACC} = \frac{1}{n_{\text{classes}}} \sum_{i=1}^{n_{\text{classes}}} \frac{n_i}{TP_i + FN_i} \tag{4}$$

where n_{classes} is the number of classes, n_i is the number of samples belonging to class i, TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i.

F1 score (5) is a widely used metric that takes into account both precision and recall, making it particularly useful when dealing with imbalanced class distributions.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

K (6), (7), (8) measures the agreement between two raters, taking into account the possibility of agreement occurring by chance.

$$\mathrm{K} = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

$$p_o = \frac{\text{Number of agreements}}{\text{Total number of ratings}} \tag{7}$$

$$p_e = \frac{(\sum_{i=1}^{n} a_i b_i)/(\text{Total number of ratings})^2}{1/(\text{Total number of ratings})} \tag{8}$$

where n is the number of categories, a_i is the number of times the model classified an item into category i, and b_i is the number of times the true labels classified an item into category i.

In addition, we employed MSE (9) as a distance-based metric for assessing the accuracy of our automatic knee OA predictions. MSE is a measure of the average squared difference between predicted and actual values, and it provides more emphasis on large errors, allowing us to capture ordinal information of misclassification. This approach was endorsed by Antony [21], where the prediction of KL grades was formulated as a regression problem.

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9}$$

Table 3
Range of evaluation metrics for each ML classifiers with median written inside parentheses. ACC: balanced accuracy, K: Cohen's kappa, KNN: k-nearest neighbor, LR: logistic regression, MSE: mean squared error, RF: random forest.

ML	ACC	F1	K	MSE
RF	33.91-58.14 (51.63)	0.50-0.86 (0.70)	0.28-0.58 (0.51)	10.38-0.33 (0.68)
LR	29.95-70.73 (54.91)	0.24-0.77 (0.58)	0.21-0.56 (0.47)	16.91-0.68 (1.00)
KNN	**49.92-94.64 (82.66)**	**0.59-0.98 (0.89)**	**0.51-0.93 (0.84)**	**6.75-0.05 (0.24)**

Table 4
Range of balanced accuracy for each deep learning network and KNN combination. Their respective median was presented in parentheses. ACC: balanced accuracy, KNN: k-nearest neighbor.

ML	DL	ACC	Rank
KNN	VGG16-FC	82.25-90.28 (85.99)	5
	VGG19-FC	75.43-85.98 (80.39)	12
	VGG16-GAP	**89.75-94.64 (92.08)**	1
	VGG19-GAP	85.98-92.55 (88.42)	2
	EfficientNetB0-GAP	85.02-92.25 (88.38)	3
	EfficientNetB1-GAP	81.63-89.49 (85.53)	6
	EfficientNetB2-GAP	81.54-89.91 (85.23)	8
	EfficientNetB3-GAP	78.16-87.41 (83.20)	9
	EfficientNetB4-GAP	75.91-86.10 (81.01)	11
	EfficientNetB5-GAP	73.42-84.13 (78.57)	13
	EfficientNetB6-GAP	68.47-81.10 (74.64)	15
	EfficientNetB7-GAP	72.77-84.28 (78.27)	14
	ResNet50-GAP	84.08-91.51 (87.16)	4
	ResNet101-GAP	77.11-87.09 (82.25)	10
	ResNet152-GAP	81.17-89.31 (85.28)	7
	DenseNet121-GAP	49.92-70.85 (59.88)	18
	DenseNet169-GAP	50.97-71.69 (61.14)	17
	DenseNet201-GAP	62.19-78.41 (70.48)	16

where n is the number of samples, y_i is the true value of the target variable for the i-th sample, and $\hat{y}_i$ is the predicted value of the target variable for the i-th sample.

These metrics together provide a comprehensive evaluation of the models' performance.

5.4. Comparison between feature extraction methods

Our proposed GAP-based feature extraction produced eight times fewer features in VGG16 and VGG19 (Table 1) as compared to traditional FC-based feature extraction method, leading to decreased computational cost. Based on experimental outcomes, GAP significantly improved the performances of VGG16 and VGG19 in prediction of all OA features. This could be due to the elimination of useless features through GAP layer.

5.5. Comparison between machine learning classifiers

RF and LR algorithms demonstrated worse results with balanced accuracy below 60% for most of the predictions. KNN outperformed both RF and LR, with highest range of balanced accuracy, F1, Cohen's kappa and MSE (Table 3). Thus, only the results of KNN classifiers will be further discussed (Table 4). Among the combinations of feature extractors with KNN classifier, DenseNet121-GAP and KNN produced the worst classification outcomes in all predictions with median balanced accuracy of merely 59.88%, unlike other combinations which had achieved median balanced accuracy beyond 60%. Best performances were achieved by combination of VGG16-GAP and KNN.

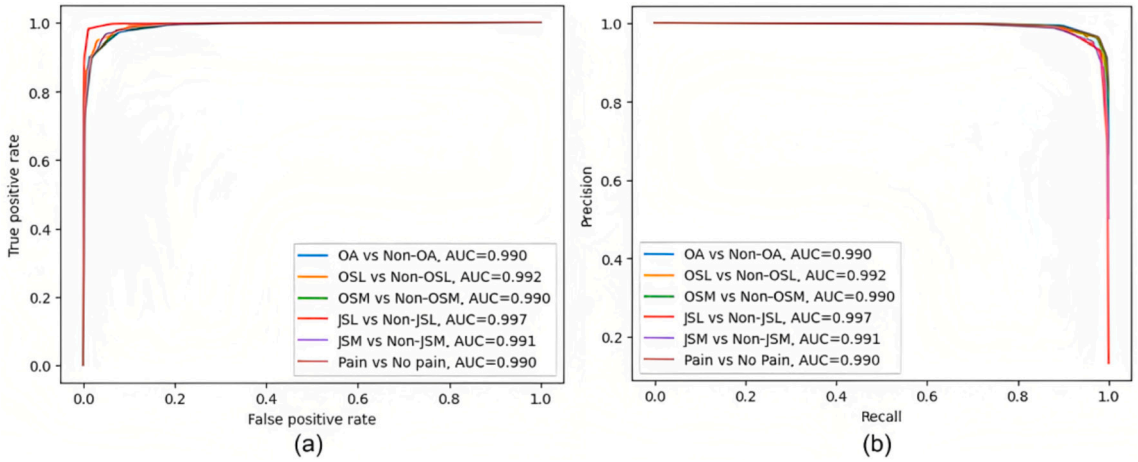

Fig. 4. The performance of proposed model in detecting the presence of radiographic OA (KL ≥ 2) on both sides of knee, osteophytes (grade ≥ 1) and joint-space narrowing (grade ≥ 1) on lateral and medial compartments, and knee pain in terms of (a) ROC and (b) precision-recall curves. OSL: presence of osteophytes on knee lateral compartment, OSM: presence of osteophytes on knee medial compartment, JSL: joint-space narrowing on knee lateral compartment, and JSM: joint-space narrowing on knee medial compartment.

Table 5
Knee OA severity prediction based on KL scheme.

Methods	ACC	F1	K	MSE
VGG16-FC + KNN	86.74	0.88	0.86	0.32
VGG19-FC + KNN	81.25	0.83	0.81	0.44
VGG16-GAP + KNN	***92.53**	***0.93**	***0.92**	***0.18**
VGG19-GAP + KNN	88.98	0.90	0.87	0.26
EfficientNetB0-GAP + KNN	89.67	0.91	0.90	0.24
EfficientNetB1-GAP + KNN	86.49	0.88	0.87	0.30
EfficientNetB2-GAP + KNN	86.49	0.88	0.87	0.30
EfficientNetB3-GAP + KNN	84.00	0.85	0.84	0.36
EfficientNetB4-GAP + KNN	82.23	0.84	0.83	0.39
EfficientNetB5-GAP + KNN	79.77	0.82	0.80	0.46
EfficientNetB6-GAP + KNN	76.35	0.79	0.76	0.55
EfficientNetB7-GAP + KNN	80.00	0.82	0.80	0.46
ResNet50-GAP + KNN	87.75	0.89	0.88	0.29
ResNet101-GAP + KNN	82.99	0.85	0.83	0.40
ResNet152-GAP + KNN	86.73	0.88	0.86	0.31
DenseNet121-GAP + KNN	60.37	0.66	0.59	0.89
DenseNet169-GAP + KNN	61.75	0.67	0.59	0.89
DenseNet201-GAP + KNN	71.51	0.75	0.71	0.64

Table 6
Pain severity prediction.

Methods	ACC	F1	K	MSE
VGG16-FC + KNN	82.25	0.86	0.84	2.30
VGG19-FC + KNN	75.43	0.81	0.77	3.30
VGG16-GAP + KNN	***89.75**	***0.92**	***0.91**	***1.30**
VGG19-GAP + KNN	88.60	0.89	0.87	1.92
EfficientNetB0-GAP + KNN	85.02	0.88	0.87	1.89
EfficientNetB1-GAP + KNN	81.63	0.86	0.84	2.35
EfficientNetB2-GAP + KNN	81.54	0.85	0.83	2.43
EfficientNetB3-GAP + KNN	78.16	0.83	0.80	2.89
EfficientNetB4-GAP + KNN	75.91	0.81	0.78	3.17
EfficientNetB5-GAP + KNN	73.42	0.79	0.74	3.62
EfficientNetB6-GAP + KNN	68.47	0.74	0.70	4.20
EfficientNetB7-GAP + KNN	72.77	0.78	0.74	3.65
ResNet50-GAP + KNN	84.08	0.87	0.86	2.05
ResNet101-GAP + KNN	77.11	0.82	0.79	3.02
ResNet152-GAP + KNN	81.17	0.85	0.83	2.46
DenseNet121-GAP + KNN	49.92	0.59	0.51	6.75
DenseNet169-GAP + KNN	50.97	0.60	0.52	6.62
DenseNet201 + KNN	62.19	0.70	0.64	5.02

5.6. Performances of knee OA feature learning

We performed an in-depth evaluation on variations of VGG, EfficientNet, ResNet, and DenseNet architectures using four evaluation metrics, namely balanced accuracy (ACC), F1 score, Cohen's kappa (K), and mean squared error (MSE). Primary attention would be paid on Cohen's kappa, which represented inter-rater reliability. To further confirm the robustness of the model, we also tested the binary classification for six main OA features (Fig. 4).

5.6.1. Kellgren Lawrence grade estimation

VGG16 and KNN model demonstrated the best performance with 92.53% ACC, 0.93 F1 score, 0.92 K, and 0.18 MSE, among all the models tested (Table 5). However, despite achieving a high ACC, the high MSE indicated a high rate of mispredictions between neighboring classes. This is particularly relevant to the difficulty in accurately assigning a patient's X-ray to the correct KL grade, which is a common misinterpretation error made by clinicians. We observed the optimal model predicted grade 3 with lowest true positive rate.

5.6.2. Pain severity estimation

VGG16 and KNN model achieved the highest performance, with 89.75% ACC, 0.92 F1 score, 0.91 K, and 1.30 MSE, outperforming the other tested models. (Table 6). The high MSE suggested that the model's prediction of pain severity had a significant amount of variability, indicating a lack of precision. One potential cause of this variability may be the high number of possible predicted pain severity classes, ranging from 0 to 10, which increases the risk of uncertainty. In addition, we found that the occurrence of grade 10 pain was very rare even in severe OA patient group (2.25%).

5.6.3. Knee osteophytes severity estimation

For osteophytes in both knees, best prediction outcomes were yielded by VGG16 and KNN model, with 94.64% ACC, 0.96 F1 score, 0.93 K, and 0.06 MSE. Knee osteophytes severity were also graded according to four specific locations: medial fibular (OSFM), medial tibial (OSTM), lateral fibular (OSFL), and lateral tibial (OSTL). K score of overall osteophyte features was 0.93, showing consistent inter-rater capability. Grading of knee osteophytes on medial tibial yielded highest ACC (92.57%), whereas grading of osteophytes on lateral tibial achieved least ACC (91.41%). MSE of medial fibular was the highest (0.15), possibly caused by local feature disruption.

5.6.4. Joint space narrowing severity estimation

VGG16 and KNN model produced the best performance, 90.16% and 92.97% ACC, 0.98 and 0.94 F1 score, 0.93 and 0.92 K score, 0.05 and 0.11 MSE for joint-space narrowing at lateral (JSL) and medial (JSM) compartments, respectively. The prediction of JSM was slightly better

Table 7
Comparison with state-of-the-art DL-based methods for knee OA feature severity grading task. ACC: accuracy, FCN: fully-connected layer, FE: feature extraction, FS: feature selection, K: Cohen's kappa, KNN: k-nearest neighbor, LBP: local binary pattern, MSE: mean squared error, and SVM: support vector machine.

Paper	FE	FS	Classifier	Dataset (No. of images)	Training approach	Targets	ACC	F1	K	MSE/loss
[21]	CNN	N/A	FCN	OAI (8,892) MOST (5,840)	Training (70%) and testing (30%) split	JSL	69.1	0.93	0.80	
						JSM	73.4	0.75	0.75	
						OSFM	45.8	0.61	0.48	
						OSTM	47.9	0.66	0.61	
						OSFL	44.3	0.67	0.47	
						OSTL	47.6	0.72	0.52	
						KL	63.6	0.60	0.69	
[22]	Ensemble method with SE-ResNet50 and SE-ResNext50-32x4d	N/A	FCN	OAI (19,704) MOST (11,743)	5-fold subject-wise stratified cross-validation	JSL	78.55	0.96	0.94	0.04
						JSM	80.66	0.82	0.90	0.20
						OSFM	72.02	0.81	0.84	0.41
						OSTM	65.49	0.77	0.83	0.26
						OSFL	68.85	0.81	0.79	0.33
						OSTL	63.58	0.83	0.84	0.22
						KL	66.68	0.65	0.82	0.68
[24]	Stacked ensemble CNN (6 base models)	N/A	SVM	OAI (37,996)	Training (60%), validation (20%), and testing (20%) split	KL	87.0	0.87		
[25]	Pretrained CNN	PCA	SVM	OAI (9,786)	Training (70%), validation (10%), and testing (20%) split	KL	74.57	0.83		
[26]	LBP, AlexNet and Dark-net-53	PCA	Fine KNN	OAI (3,795)	10-fold cross validation	KL	90.6	0.88		
[27]	Transfer learning using DenseNet201	N/A	FCN	Private hospital (2,068)	Training (70%), testing (10%), and validation (20%) split	KL	***92.87**	0.93		0.20
Proposed	VGG16	N/A	KNN	OAI (4,255)	10-fold KL stratified cross-validation	JSL	***90.16**	***0.98**	***0.93**	***0.05**
						JSM	***92.97**	***0.94**	***0.92**	***0.11**
						OSFM	***91.82**	***0.95**	***0.93**	***0.15**
						OSTM	***92.57**	***0.94**	***0.93**	***0.09**
						OSFL	***92.08**	***0.95**	***0.93**	***0.11**
						OSTL	***91.41**	***0.96**	***0.93**	***0.09**
						KL	92.53	***0.93**	***0.93**	***0.18**

* indicates best performance.

than the prediction of JSL, potentially due to the more balanced data distribution in JSM.

5.7. Comparison of proposed model with existing models

To demonstrate the proposed pipeline's effectiveness, Table 7 was presented for comparison of the results of the optimal hybrid model with the state-of-the-art methods trained and validated on standard OA datasets. Although the studies by Antony [21] and by Tiulpin and Saarakkala [22] were pioneering in the field, we have produced the new state-of-the-art results in radiographic OA detection based on all metrics, except the ACC of KL grade. Highest ACC for KL grade was achieved by Tiwari et al. [27] who used DenseNet201 as feature extractor and FC layer with softmax activation as classification layer. In contrast, we used a relatively small dataset to accomplish the task. However, concern of model generalization exists as we did not validate the model with second dataset. Furthermore, percentage of Asian in the dataset is very small, thereby the model may not effectively learning the OA patterns of Asian.

5.8. Discussion

In this study, we established a deep hybrid learning model to perform multitask classification of OA features, that involved quantification of radiographic features (bone spurs / osteophytes and joint space narrowing condition) and patient pain intensity. Among all nine tasks, overall osteophyte prediction achieved the highest ACC score, while JSL attained the highest F1 score and the lowest MSE. These findings suggested that the model's performance was particularly strong in these two tasks, indicating its effectiveness in quantifying radiographic features related to overall osteophytes and JSL in the context of OA classification. The K values for all tasks were around 0.91 to 0.93, indicating a good agreement rate between the model's predictions and the annotations provided by medical experts. Notably, pain severity prediction had slightly lower evaluation metrics compared to other tasks and thus warrants particular attention. Considering that clinical interpretation of pain manifestation can often be categorized into general groups such as mild, moderate, and severe [52], grouping similar pain levels for analysis could be a potential approach to address this limitation.

Producing accurate models for patient use is a realistic and important goal for successful automatic knee OA diagnosis in clinical settings. While our proposed model could provide a baseline reference for knee OA diagnosis on standard OAI data, it has certain limitations that need to be addressed to improve the accuracy of the models on real patient data. In order to contribute to future work, we would like to discuss these limitations and propose potential solutions to enhance the model's boundaries. Specifically, two contributions should be considered:

improvements in (1) data diversity and representation, and (2) model architecture and performance.

Our study was conducted on a standard dataset from OAI that only focuses United States (US) population, which represents the largest OA market [53,54]. However, the performance of the models on this dataset may not necessarily generalize to real global patient populations [55]. To address this limitation and improve the reliability of the models, we suggest exploring the use of more diverse and representative datasets that better reflect the complexities of real-world patient populations to tune the models. Ideally, we propose collaborating with health institutions from all regions to establish a more real-world represented dataset. Specifically, we suggest giving particular attention to fast-growing knee OA markets such as the Asia Pacific region [2] where the aging population is increasing rapidly [56].

Furthermore, we suggest the incorporation of external data sources and features that may be relevant to the OA prediction task, such as clinical notes, imaging data, and other patient-reported outcome measures. This will help to capture the unique characteristics of individual patients and improve the accuracy of the models in predicting outcomes. This effort aligns with the current trend in precision healthcare, where personalized approaches are becoming increasingly important in improving patient care.

Another crucial consideration is how to improve the accuracy of the model when applied to real patient data. Most models were trained on standardized data with reduced uncertainty factors, whereas real patient data can be subject to a wide range of uncertainties, and scarcity of labeled data. As such, it is crucial to develop models that are robust to these uncertainties and can effectively handle the variability of real patient data. One potential solution is to use weakly supervised learning (WSL) methods [57], which enable training ML models with reduced supervision, which is often the case in real patient data where only a subset of data is labeled (incomplete supervision), or the labels are coarse-grained (inexact supervision) or ambiguous (inaccurate supervision). Multiple Instance Learning (MIL) is a promising approach for addressing inexact supervision, especially in situations where only image-level labels are available, as opposed to more precise pixel-level labels. Another potential approach to improve the accuracy of a model on real patient data is to use knowledge distillation [58,59], which involves training a smaller, simpler model (the student) to mimic the output of a larger, more complex model (the teacher). The teacher model can be trained on standardized data, while the student model can be trained on real patient data. By distilling the knowledge learned by the teacher model into the student model, the student model can learn to effectively handle the variability of real patient data, while benefiting from the knowledge and generalization capabilities of the teacher model. To improve the generalization performance of the model on real patient data, we recommend a fine-tuning strategy that incorporates randomness into the formation of subsequent DL layers in the model architecture during training [60]. All suggested approaches can potentially improve accuracy, robustness, and generalization of models in real-world scenarios.

6. Conclusion and future works

In this paper, we studied the performance of 16 CNN architectures in extracting the imaging features from knee radiography. The framework based on VGG16 network outperforms the other CNNs, as well as the state-of-the-art methods. The experimental results showed that pain prediction model presented the highest error in terms of MSE metric. We can conclude that OA-associated pain pattern is more complicated than others. In future work, we plan to extend our existing model into a deep multiple instance learning classification model, with the aim of exploring additional potential OA features. We also plan to identify the key OA features that can facilitate the selection of subject-wise optimal therapeutic intervention.

CRediT authorship contribution statement

Yun Xin Teoh: Data curation, writing original draft preparation. **Alice Othmani:** Supervision of project, Software, produce final manuscript draft. **Khin Wee Lai:** Originate idea of study from engineering perspective. **Siew Li Goh:** Conceptualization of this study from clinical perspective. **Juliana Usman:** Originate idea of study from biomechanics perspective.

Declaration of competing interest

We confirm that this work is original, has not been published elsewhere, and is not currently being considered for publication elsewhere. The consent of all authors of this paper has been obtained for submitting the paper and all authors declare no conflict of interest.

Acknowledgement

This work was supported by the Ministry of Higher Education, Malaysia under Fundamental Research Grant Scheme (FRGS) Grant No. FRGS/1/2022/SKK01/UM/02/1 and the Malaysia-France University Centre (MFUC).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2023.107807.

References

[1] C. Palazzo, C. Nguyen, M.-M. Lefevre-Colau, F. Rannou, S. Poiraudeau, Risk factors and burden of osteoarthritis, Ann. Phys. Rehabil. Med. 59 (2016) 134–138.
[2] H. Long, Q. Liu, H. Yin, K. Wang, N. Diao, Y. Zhang, J. Lin, A. Guo, Prevalence trends of site-specific osteoarthritis from 1990 to 2019: findings from the global burden of disease study 2019, Arthritis Rheumatol. 74 (2022) 1172–1183.
[3] E. Losina, A.M. Weinstein, W.M. Reichmann, S.A. Burbine, D.H. Solomon, M.E. Daigle, B.N. Rome, S.P. Chen, D.J. Hunter, L.G. Suter, J.M. Jordan, J.N. Katz, Lifetime risk and age at diagnosis of symptomatic knee osteoarthritis in the US, Arthritis Care Res. 65 (2013) 703–711.
[4] D. Li, S. Li, Q. Chen, X. Xie, 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, Front. Med. 7 (2020).
[5] J.B. Driban, M.S. Harkey, S.-H. Liu, M. Salzler, T.E. McAlindon, Osteoarthritis and aging: young adults with osteoarthritis, Curr. Epidemiol. Rep. 7 (2020) 9–15.
[6] H. Madry, E. Kon, V. Condello, G.M. Peretti, M. Steinwachs, R. Seil, M. Berruto, L. Engebretsen, G. Filardo, P. Angele, Early osteoarthritis of the knee, Knee Surg. Sports Traumatol. Arthrosc. 24 (2016) 1753–1762.
[7] C.M. Parsons, A. Judge, R. Meyer, O. Bruyère, F. Petit-Dop, R. Chapurlat, J.-Y. Reginster, C. Cooper, H. Inskip, Determining individual trajectories of joint space loss: improved statistical methods for monitoring knee osteoarthritis disease progression, Osteoarthr. Cartil. 29 (2021) 59–67.
[8] D.K. White, Y. Zhang, J. Niu, J.J. Keysor, M.C. Nevitt, C.E. Lewis, J.C. Torner, T. Neogi, Do worsening knee radiographs mean greater chances of severe functional limitation?, Arthritis Care Res. 62 (2010) 1433–1439.
[9] C.R. Chu, A.A. Williams, C.H. Coyle, M.E. Bowers, Early diagnosis to enable early treatment of pre-osteoarthritis, Arthritis Res. Ther. 14 (2012) 212.
[10] M.D. Kohn, A.A. Sassoon, N.D. Fernando, Classifications in brief: Kellgren-Lawrence classification of osteoarthritis, Clin. Orthop. Relat. Res. 474 (2016) 1886–1893.
[11] K. Klara, J.E. Collins, E. Gurary, S.A. Elman, D.S. Stenquist, E. Losina, J.N. Katz, Radiographic assessment of severe knee osteoarthritis: role of training and experience, J. Rheumatol. 43 (2016) 1421–1426.
[12] R.D. Altman, G.E. Gold, Atlas of individual radiographic features in osteoarthritis, revised, Osteoarthr. Cartil. 15 (2007).
[13] A.M. Alenazi, M.M. Alshehri, S. Alothman, B.A. Alqahtani, J. Rucker, N. Sharma, N.A. Segal, S.M. Bindawas, P.M. Kluding, The association of diabetes with knee pain severity and distribution in people with knee osteoarthritis using data from the osteoarthritis initiative, Sci. Rep. 10 (2020).
[14] K.N. Kunze, S.J. Jang, T. Li, D.A. Mayman, J.M. Vigdorchik, S.A. Jerabek, A.T. Fragomen, P.K. Sculco, Radiographic findings involved in knee osteoarthritis progression are associated with pain symptom frequency and baseline disease severity: a population-level analysis using deep learning, Knee Surg. Sports Traumatol. Arthrosc. (2022).
[15] P.S.Q. Yeoh, K.W. Lai, S.L. Goh, K. Hasikin, Y.C. Hum, Y.K. Tee, S. Dhanalakshmi, Emergence of deep learning in knee osteoarthritis diagnosis, Comput. Intell. Neurosci. 2021 (2021) 4931437.

[16] Y.X. Teoh, J. Lai, K.W. Usman, S.L. Goh, H. Mohafez, K. Hasikin, P. Qian, Y. Jiang, Y. Zhang, S. Dhanalakshmi, Discovering knee osteoarthritis imaging features for diagnosis and prognosis: review of manual imaging grading and machine learning approaches, J. Healthc. Eng. 11 (2022) 4138666.
[17] M. Binvignat, V. Pedoia, A.J. Butte, K. Louati, D. Klatzmann, F. Berenbaum, E. Mariotti-Ferrandiz, J. Sellam, Use of machine learning in osteoarthritis research: a systematic literature review, in: International Workshop on Machine Learning in Medical Imaging, 2022.
[18] H. Oka, S. Muraki, T. Akune, K. Nakamura, H. Kawaguchi, N. Yoshimura, Normal and threshold values of radiographic parameters for knee osteoarthritis using a computer-assisted measuring system (koacad): the road study, J. Orthop. Sci. 15 (2010) 781–789.
[19] J. Thomson, T. O'Neill, D. Felson, T. Cootes, Detecting osteophytes in radiographs of the knee to diagnose osteoarthritis, in: International Workshop on Machine Learning in Medical Imaging, 2016.
[20] M. Saleem, M.S. Farid, S. Saleem, M.H. Khan, X-ray image analysis for automated knee osteoarthritis detection, Signal Image Video Process. 14 (2020) 1079–1087.
[21] A.J. Antony, Automatic quantification of radiographic knee osteoarthritis severity and associated diagnostic features using deep convolutional neural networks, Ph.D. thesis, Dublin City University, 2018.
[22] A. Tiulpin, S. Saarakkala, Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks, Diagnostics 10 (2020) 932.
[23] P. Chen, L. Gao, X. Shi, K. Allen, Y. Lin, Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss, Comput. Med. Imaging Graph. 75 (2019) 84–92.
[24] B.M. Muhammad, M. Yeasin, Interpretable and parameter optimized ensemble model for knee osteoarthritis assessment using radiographs, Sci. Rep. 11 (2021).
[25] S.M. Ahmed, R.J. Mstafa, Identifying severity grading of knee osteoarthritis from x-ray images using an efficient mixture of deep learning and machine learning models, Diagnostics 12 (2022) 2939.
[26] U. Yunus, J. Amin, M. Sharif, M. Yasmin, S. Kadry, S. Krishnamoorthy, Recognition of knee osteoarthritis (koa) using yolov2 and classification based on convolutional neural network, Life 12 (2022) 1126.
[27] A. Tiwari, M. Poduval, V. Bagaria, Evaluation of artificial intelligence models for osteoarthritis of the knee using deep learning algorithms for orthopedic radiographs, World J. Orthop. 13 (2022) 603–614.
[28] R. Mahum, S.U. Rehman, T. Meraj, H.T. Rauf, A. Irtaza, A.M. El-Sherbeeny, M.A. El-Meligy, A novel hybrid approach based on deep cnn features to detect knee osteoarthritis, Sensors 20 (2021) 6189.
[29] S. Olsson, E. Akbarian, A. Lind, A.S. Razavian, M. Gordon, Automating classification of osteoarthritis according to Kellgren-Lawrence in the knee using deep learning in an unfiltered adult population, BMC Musculoskelet. Disord. 22 (2021) 1–8.
[30] S. Karen, Z. Andrew, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, in: Conference Track Proceedings, 2015, pp. 1–14, http://arxiv.org/abs/1409.1556.
[31] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, PMLR, vol. 97, 2019, pp. 6105–6114, https://proceedings.mlr.press/v97/tan19a.html.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[33] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.
[34] R. Yamashita, M. Nishio, R.K. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging 9 (2018) 611–629.
[35] S.S. Basha, S.R. Dubey, V. Pulabaigari, S. Mukherjee, Impact of fully connected layers on performance of convolutional neural networks for image classification, Neurocomputing 378 (2020) 112–119.
[36] B. Subrahmanyeswara Rao, Accurate leukocoria predictor based on deep vgg-net cnn technique, IET Image Process. 14 (2020) 2241–2248.
[37] X. Qu, H. Lu, W. Tang, S. Wang, D. Zheng, Y. Hou, J. Jiang, A vgg attention vision transformer network for benign and malignant classification of breast ultrasound images, Med. Phys. 49 (2022) 5787–5798.
[38] Y. Guo, Y. Xia, J. Wang, H. Yu, R.-C. Chen, Real-time facial affective computing on mobile devices, Sensors 20 (2020) 870.
[39] X. Liu, L. Wang, J. Zhang, J. Yin, H. Liu, Global and local structure preservation for feature selection, IEEE Trans. Neural Netw. Learn. Syst. 25 (2013) 1083–1095.
[40] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint, arXiv:1312.4400, 2013.
[41] Z. Wang, A. Chetouani, D. Hans, E. Lespessailles, R. Jennane, Siamese-gap network for early detection of knee osteoarthritis, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE, 2022, pp. 1–4.
[42] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, S. Saarakkala, Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach, Sci. Rep. 8 (2018) 1–10.
[43] V. Pedoia, J. Lee, B. Norman, T.M. Link, S. Majumdar, Diagnosing osteoarthritis from t2 maps using deep learning: an analysis of the entire osteoarthritis initiative baseline cohort, Osteoarthr. Cartil. 27 (2019) 1002–1010.
[44] S. Zhang, S. Zhang, C. Zhang, X. Wang, Y. Shi, Cucumber leaf disease identification with global pooling dilated convolutional neural network, Comput. Electron. Agric. 162 (2019) 422–430.
[45] N. Zhou, R. Liang, W. Shi, A lightweight convolutional neural network for real-time facial expression detection, IEEE Access 9 (2020) 5573–5584.
[46] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[47] J. Tolles, W.J. Meurer, Logistic regression relating patient characteristics to outcomes, JAMA 316 (2016) 533.
[48] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, Knn model-based approach in classification, in: On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, 2003, pp. 986–996.
[49] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 3121–3124.
[50] M.J. Warrens, Category kappas for agreement between fuzzy classifications, Neurocomputing 194 (2016) 385–388.
[51] F. Chollet, et al., Keras, https://keras.io, 2015.
[52] D.S. Tsze, G. Hirschfeld, P.S. Dayan, Clinical interpretation of self-reported pain scores in children with acute pain, J. Pediatr. 240 (2022) 192–198.
[53] M.G. Cisternas, L. Murphy, J.J. Sacks, D.H. Solomon, D.J. Pasta, C.G. Helmick, Alternative methods for defining osteoarthritis and the impact on estimating prevalence in a US population-based survey, Arthritis Care Res. 68 (2016) 574–580.
[54] L.B. Murphy, M.G. Cisternas, D.J. Pasta, C.G. Helmick, E.H. Yelin, Medical expenditures and earnings losses among US adults with arthritis in 2013, Arthritis Care Res. 70 (2018) 869–876.
[55] E. Niinimäki, J. Paloneva, I. Pölönen, A. Heinonen, S. Äyrämö, Validation of knee kl-classifying deep neural network with Finnish patient data, in: Computational Sciences and Artificial Intelligence in Industry, Springer, 2022, pp. 177–188.
[56] R.G. Steinmetz, J.J. Guth, M.J. Matava, M.V. Smith, R.H. Brophy, Global variation in studies of articular cartilage procedures of the knee: a systematic review, Cartilage 13 (2022) 19476035221098169.
[57] Y.-F. Li, L.-Z. Guo, Z.-H. Zhou, Towards safe weakly supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 334–346.
[58] J. Zheng, C. Lu, C. Hao, D. Chen, D. Guo, Improving the generalization ability of deep neural networks for cross-domain visual recognition, IEEE Trans. Cogn. Dev. Syst. 13 (2020) 607–620.
[59] L. Schoneveld, A. Othmani, H. Abdelkawy, Leveraging recent advances in deep learning for audio-visual emotion recognition, Pattern Recognit. Lett. 146 (2021) 1–7.
[60] B. Swiderski, S. Osowski, G. Gwardys, J. Kurek, M. Slowinska, I. Lugowska, Random cnn structure: tool to increase generalization ability in deep learning, EURASIP J. Image Video Process. 2022 (2022) 3.

Yun Xin Teoh received the B.Eng. degree (Hons.) in biomedical engineering (prosthetics and orthotics) from Universiti Malaya, Malaysia. She continues her Ph.D. degree in Universiti Malaya, Malaysia. She is currently under Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Université Paris-Est Créteil, France, for mobility research stay. Her research interests include medical image analysis, artificial intelligent solution for clinical problems, and rehabilitation engineering.

Alice Othmani has been an Associate Professor at the Université Paris-Est Créteil, since 2017. Her research works concern developing computer vision and artificial intelligence solutions for healthcare, emotional intelligence, and psychiatry. She has been working in several international institutions, such as the Ecole Normale Supérieure de Paris, Collège de France, and Agency for Science, Technology and Research (A*STAR), Singapore.

Khin Wee Lai received the Ph.D. degree from the Technische Universitat Ilmenau, Germany, and Universiti Teknologi Malaysia, through the DAAD Ph.D. Sandwich Programme. He is currently the Programme Head of the M.E. degree (biomedical) with the Faculty of Engineering, Universiti Malaya. His research interests include computer vision, machine learning, medical image processing, and healthcare analytics.

Siew Li Goh is a clinician and medical lecturer in Sports Medicine in Universiti Malaya. She obtained her PhD from University of Nottingham after working a network meta-analysis in knee and hip osteoarthritis. She has an interest in evidence-based medicine and biomechanics of lower limb. She is leading the biomechanics interest group, under the newly formed Sports and Exercise Medicine Research and Education Group (SEMREG), to advance the research agenda of the Sports Medicine.

Dr. Juliana Usman is a Senior Lecturer from the Department of Biomedical Engineering, Universiti Malaya. She is a member of the Centre for Applied Biomechanics, UM and the appointed secretary for the Malaysian Society of Biomechanics. She is a certified Chartered Engineer (CEng) from the Engineering Council of UK and a member of the Institute of Engineering and Technology (UK) and the Board of Engineers (Malaysia). Her research areas include sports biomechanics in terms of injury prevention and performance enhancement and motion analysis.