Adversarial Evolving Neural Network for Longitudinal Knee Osteoarthritis Prediction

Kun Hu [Member, IEEE],

School of Computer Science, The University of Sydney, NSW 2006, Australia

Wenhua Wu,

School of Computer Science, The University of Sydney, NSW 2006, Australia

Wei Li,

School of Computer Science, The University of Sydney, NSW 2006, Australia

Milena Simic,

School of Health Sciences, The University of Sydney, NSW 2006, Australia

Albert Zomaya [Fellow, IEEE],

School of Computer Science, The University of Sydney, NSW 2006, Australia

Zhiyong Wang [Member, IEEE]

School of Computer Science, The University of Sydney, NSW 2006, Australia

Abstract

Knee osteoarthritis (KOA) as a disabling joint disease has doubled in prevalence since the mid-20th century. Early diagnosis for the longitudinal KOA grades has been increasingly important for effective monitoring and intervention. Although recent studies have achieved promising performance for baseline KOA grading, longitudinal KOA grading has been seldom studied and the KOA domain knowledge has not been well explored yet. In this paper, a novel deep learning architecture, namely adversarial evolving neural network (A-ENN), is proposed for longitudinal grading of KOA severity. As the disease progresses from mild to severe level, ENN involves the progression patterns for accurately characterizing the disease by comparing an input image it to the template images of different KL grades using convolution and deconvolution computations. In addition, an adversarial training scheme with a discriminator is developed to obtain the evolution traces. Thus, the evolution traces as fine-grained domain knowledge are further fused with the general convolutional image representations for longitudinal grading. Note that ENN can be applied to other learning tasks together with existing deep architectures, in which the responses characterize progressive representations. Comprehensive experiments on the Osteoarthritis Initiative (OAI) dataset were conducted to evaluate the proposed method. An overall accuracy was achieved as 62.7%, with the baseline, 12-month, 24-month, 36-month, and 48-month accuracy as 64.6%, 63.9%, 63.2%, 61.8% and 60.2%, respectively.

hukunsdu@hotmail.com .

Index Terms—

Knee osteoarthritis; deep learning; adversarial learning

I. INTRODUCTION

Knee osteoarthritis (KOA) as a disabling joint disease has doubled in prevalence since the mid-20th century [1]. The main radiographic characteristics of KOA are the formation of osteophytes, periarticular ossicles, narrowing of joint cartilage, small pseudo-cystic areas with sclerotic walls, and altered shape of the bone ends [2]. Currently, as there is no treatment to permanently cure KOA and the increasing trend in life expectancy and body mass index, early diagnosis and longitudinal KOA severity prediction of the disease progression have been increasingly important to help improve the life quality of patients [3]. For this purpose, the Kellgren and Lawrence (KL) grading system has been widely used to quantify the severity of the disease in clinical practice [2], in which radiology imaging techniques are utilised for KL grading. However, the grading requires well-trained experts for time-consuming and costly annotations, and it can be difficult for human experts to perform long-term KOA predictions using merely baseline images. Therefore, it is attractive to devise computer-aided automatic longitudinal KL grading methods for KOA using radiology images.

Existing studies mainly address the automatic KOA grading task of a given radiology image by treating it as an image classification task with multiple classes (i.e., KL grades). Early studies follow a conventional machine learning pipeline in which the image pre-processing and feature extraction steps are required for a classification model (e.g. [4]). To this end, the quality of the extracted features is critical to the grading performance. Owing to the great success of deep learning techniques in many visual applications, a data-driven approach can be devised in pursuit of more accurate grading algorithms. As many deep learning architectures have been proposed for general image classification problems, such as convolutiaon neural networks (CNNs) (e.g., VGG [5] and ResNet [6]) and transformer based networks (e.g., Visual Transformer (ViT) [7]), some of these architectures have been also adopted for KOA grading (e.g, [8]). Besides, there are also a few attempts to involve the domain knowledge of KOA into the development of deep learning based methods, such as the inclusion of the demographic features and the loss functions using the continuous grading property (e.g. [9], [10]). Although these methods have achieved impressive performance, the grading is only reported in regard to the baseline scan.

Many attempts have been recently undertaken for KOA progression analysis based on clinical assessment characteristic and medical imagining features (e.g., [11]–[14]). However, the prediction of the KOA conversion after a period of time is not accurate enough, compared to a longitudinal KOA grading scheme. Moreover, different from general image classification tasks, the intra-class variance of knee images can be higher than the inter-class variance in terms of the KOA grades which present fine-grained patterns. Therefore, it is reasonable to devise deep learning algorithms with adequate domain knowledge for accurately grading KOA in a longitudinal manner. In addition, although there have been

many studies recently on fine-grained learning (e.g. [15]–[19]), these studies are usually task-specified, which could not be applicable to longitudinal KOA grading.

Therefore, in this paper, a novel deep learning architecture, namely adversarial evolving neural network (A-ENN), with an adversarial training scheme is proposed for fine-grained longitudinal KOA grading using X-Ray images collected from clinical assessments. As the disease progresses from mild to severe, the longitudinal KOA grades of a patient can be investigated from an evolving perspective. By comparing an input X-Ray image to a set of template images of different KL grades, evolving traces, which indicate how the input image changes to/from templates of individual KL grades, can be helpful to formulate fine-grained KOA patterns. In detail, the proposed ENN introduces a set of convolution-deconvolution structures to transform an input image to the images of different KL grades with an adversarial training scheme. The feature maps during the forward propagation can be viewed as the trace of each transformation. Then, the representations of these traces and the input image can be fused to predict a longitudinal KL grade. Note that the proposed ENN can be applied to other learning tasks as well, in which the responses are of progressive and continuous properties. The proposed A-ENN method is evaluated on a widely used benchmark dataset - the Osteoarthritis Initiative (OAI) Dataset - for KOA grading and achieves an accuracy 62.6%.

In summary, the major contributions of this paper are three-fold:

- KOA patterns are formulated in a progressive manner by deriving the evolving traces with a set of template images of different KL grades.

- A novel fine-grained deep learning architecture ENN with an adversarial learning scheme is proposed to compute the evolving traces and predict the longitudinal KOA severity grades.

- Comprehensive experiments have been conducted on the OAI dataset to demonstrate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section II reviews the related work for KOA grading and progression prediction methods. Section III introduces the details of our proposed method. Section IV presents comprehensive experimental results and their analysis to evaluate the proposed method. Lastly, Section V concludes our study with discussions on future work.

II. RELATED WORK

In this section, the related studies are reviewed from two perspectives: (1) computer vision based KL grading methods and (2) longitudinal KOA progression prediction methods.

A. Computer Vision based KL Grading

Computer-aided KL grading methods identify the KL grade with the given information of a subject. When the information is of medical images such as X-ray and MRI scans, computer vision based methods can be devised for KL grading [20].

By following a conventional machine learning pipeline, early computer vision based KL grading algorithms mainly relied on hand-crafted features extracted from medical images. For example, features in regard to texture and edge were investigated for the classification from KL grade 0 to grade 3 [4]; regional features were used to distinguish images of KL grade 0 and grade 2 with Naive Bayes and random forest classifiers [21]. Note that obtaining KOA relevant features is a key step in these conventional algorithms and can be a key factor impacting the grading performance.

In recent years, various deep learning techniques have been introduced for KOA grading, by which hand-crafted features are no longer required. Many methods were based on X-ray images. In [22], pre-trained CNNs were used to extract deep representations as KOA features, and a support vector machine with a linear kernel was further adopted for the grading of KL grade 0 and grade 2. An end-to-end deep learning using CNNs for the same task treated the grading task as a combination of a regression and a classification task since the KOA grades present a continuous property [23]. More recently, a number of deep learning architectures such as VGG, ResNet and DenseNet were studied with an adjustable ordinal loss for KOA grading from grade 0 to grade 4 [8]. In [24], an X-ray based CNN modelling method was reported a performance close to that of a patients questionnaire data based modelling approach. An autoencoder based method to learn discriminative KOA representations was studied to improve classification performance [25]. Faster R-CNN was utilized to generate knee joint regional proposals and reduce the irrelevant information in X-ray images and CNNs can be further adopted for the localized classification to improve KOA grading performance. In addition, a focal loss was devised to address the imbalance issue of the grading classes [9]. Another study also was attempted for the localization of key regions with a YOLO algorithm [26]. An ordinal regression module was proposed for classification neural networks to perform ordinal regression for KOA grading [27]. A high-resolution network capturing the multi-scale features of knee X-rays was proposed to improve the grading [28]. Siamese neural networks were studied based on the similarities between input images [29], [30]. In [31], a semi-supervised learning approach was investigated to reduce the demands of large amounts of data for KOA severity assessment.

Besides X-Ray images, MRI images have been explored in a similar manner. For example, AlexNet based [32] and DenseNet based [33] CNN architectures were evaluated. In addition, a number of studies introduced demographics such as Body Mass Index (BMI), age and gender to improve the performance jointly with the medical images. For example, in [10], CNN representations and the demographic representations were jointly used for KL grading.

B. Longitudinal KOA Progression Prediction

One type of studies on KOA progression aims to discover the correlation between KOA progression and manually obtained clinical assessment outcomes. A group of individuals who were of high risks at the baseline visit were studied to predict the radiographic or pain progression (based on WOMAC - Western Ontario and McMaster Universities Osteoarthritis Index questionnaire) over 8 years [11]. In [12], a Least Absolute Shrinkage and Selection Operator (LASSO) regression model was proposed to predict patients into one

non-progressive class and three progressive classes by using various clinical assessments such as knee symptoms, medication usage, and knee alignment measured on x-ray images. Conventional machine learning methods were also utilized for KOA progression prediction by using features of X-ray image based assessments (e.g., semi-quantitative readings and JSW (joint space width)) [12], [13]. Clinical data including questionnaires and radiographic markers (e.g., joint space width and knee alignment) were studied to predict the KL grade for the next visit of a subject with an attention based long short term memory neural network [34]. By defining the progression based on the changes of both JSW and WOMAC pain score in 48 months, a multi-layer perceptron was studied using a wide range of clinical data, such as radiographic and symptomatic data, and history of the knee injury and surgery [35].

Medical images have also been analyzed automatically for KOA progression prediction. In [36], 344 knees from subjects with no sign of radiographic KOA at baseline visits were studied to predict whether a global radiographic KOA happens or not after 48 months, where a logistic regression model with trabecular bone texture features extracted from X-ray images was adopted. In [14], deep convolutional features of X-ray images together with questionnaire features were utilized to predict the probability of KOA progression as a multi-class classification task: no KOA progression, fast progression and slow progression. By using deep learning architecture DeepLabv3 [37] and U-Net [38] for a segmentation pre-processing of X-ray images, multiple JSW measurements can be estimated and further used with a gradient boosting machine for KOA progression prediction from KL 0–1 to 2–4 [39]. Deep learning methods including DenseNet and EfficientNet were adopted to predict the progression of the radiographic medial joint space loss [40] and pain [41] over 48 months for the KOA risk assessment using X-Ray images. MRI imaging also shows its effectiveness for KOA progression prediction [42], [43]. Cartilage damage index features based on informative regions of MRI images were studied for predicting the progression, which was defined by the changes of the KL, JSL and JSN grades in 24 months [44], [45].

Although these methods demonstrated encouraging results for KOA progression prediction, fine-grained medical imaging patterns have not been well explored. As deep learning techniques have achieved groundbreaking success in a wide variety of computer vision related tasks, it is desirable to explore a novel deep architecture to formulate fine-grained visual patterns of medical images for KOA progression prediction. To this end, this study aims to learn fine-grained representations of different KL grades using deep learning techniques in pursuit of effective longitudinal KOA grading for the first time.

III. Methodology

As illustrated in Fig. 2, the proposed A-ENN architecture consists of three key modules: 1) Evolving trace estimation which simulates the process how an input image evolves to the images of target KL grades with an adversarial training scheme; 2) Classifiers predicting the raw longitudinal KOA grades for the input image and its evolving traces; and 3) a fusion scheme obtaining the final grading probability map for longitudinal KL grading. In this section, after the formulation of longitudinal KOA KL grading is introduced, the details of these modules are explained.

A. KOA Images and Longitudinal KL Grading

The proposed method uses a baseline scan X-ray image as the input for the longitudinal KOA grading of a patient. Particularly, the image is denoted as $\mathbf{X} \in \mathbb{R}^{C \times W \times H}$, where C, W, and H are the number of channels, the width and the height of the image, respectively. A ground-truth KL grade at the time of the baseline scan can be scored by clinicians using $\mathbf{X}$, which indicates the status of the KOA severity. In detail, KL grade 0 represents definite absence of radiographic features of osteoarthritis; KL grade 1 indicates doubtful JSN and possible osteophytic lipping; KL grade 2 suggests possible JSN and definite osteophytes; KL grade 3 suggests definite JSN, moderate osteophytes, some sclerosis and possible bone-end deformity; KL grade 4 suggests marked JSN, large osteophytes, severe sclerosis and definite bone-end deformity [46].

Besides the baseline visit, follow-up visits of a patient after 12 months, 24 months, 36 months and 48 months provide additional X-ray scans to obtain the ground-truth KL grades in a longitudinal manner. Thus, for a patient, denote $\mathbf{Y} = [\mathbf{y}_0, ..., \mathbf{y}_t, ..., \mathbf{y}_T]^\top = (y_{t,k}) \in \mathbb{R}^{T \times K}$, where $\mathbf{y}_t$ denotes the t-th (e.g., $t = 0$ for the baseline and $t = 1$ for the 12-month follow-up) longitudinal grade. Note that $\mathbf{y}_t$ is a K-dimensional one-hot encoding vector and $K = 5$ represents the number of KL grading levels. The proposed method aims to estimate $\mathbf{Y}$ given only the baseline X-ray observation $\mathbf{X}$ without using the scans from the follow-up visits. Particularly, this estimation is denoted as $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_0, ..., \widehat{\mathbf{y}}_t, ..., \widehat{\mathbf{y}}_T]^\top = (\widehat{y}_{t,k}) \in \mathbb{R}^{T \times K}$.

B. Template-guided Evolving Trace Estimation

Without the future radiology scans, it can be challenging to predict the longitudinal grades of KOA. Intuitively, the disease's evolving traces of a given input image $\mathbf{X}$ to different severity levels can be helpful for KOA progression prediction by comparing the image with the radiology scans of different KL grades.

To achieve this goal, for each KOA severity level i ($i = 0, ..., K{-}1$), a number of images can be identified to construct a template image set $\mathbb{T}_i = \{\mathbf{T}_{i,j}\}$. In detail, $\mathbf{T}_{i,j}$ is the j-th selected template image of KL grade i. Note that there can be many different methods to construct the template sets. For example, to construct $\mathbb{T}_i$, the most representative images of KL grade i in the training set can be identified by clinicians, randomly selecting as a subset of images of KL grade i from the training set, or using all the images of KL grade i in the training set. Then the input image $\mathbf{X}$ is attempted to be evolved to the template images of each KOA severity level individually. Particularly, $K = 5$ traces can be obtained for $\mathbf{X}$.

As shown in Fig. 2, an evolving module is devised by utilising convolution and deconvolution filters for each KOA grade i. The feature maps computed by the filters of module i are treated as the evolving traces from the input image $\mathbf{X}$ to the templates of KOA grade i. In detail, a set of convolution layers are first adopted to formulate a latent representation of the input image $\mathbf{X}$. This latent representation is expected to include the implicit relations between $\mathbf{X}$ and target template images of grade i. Mathematically, denoting the computations of these convolution layers as $f_i^{e,1}$, the latent representation can be derived as:

$$\mathbf{H}_i = f_i^{e,\,1}(\mathbf{X}), \tag{1}$$

where $\mathbf{H}_i \in \mathbb{R}^{C^h \times W^h \times H^h}$, C^h indicates the number of channels, W^h and H^h represent the size of the hidden representations.

The latent representation is an intermediate result and its size is reduced compared to the original image (template) size. Hence, deconvolution layers, which are also known as transpose convolution layers, are further introduced to reconstruct a template image $\mathbf{T}_{i,j}$ of KL grade i. This is similar to the existing practices used in other computer vision tasks such as image super resolution [47]. Denoting the computations of the deconvolution layers as $f_i^{e,\,2}$, an estimation of $\mathbf{T}_{i,j}$ can be derived as:

$$\widehat{\mathbf{T}}_i = f_i^{e,\,2}(\mathbf{H}_i). \tag{2}$$

Note that the estimation $\widehat{\mathbf{T}}_i$ is expected to be close to any templates in $\mathbb{T}_i$, in regardless of j.

Note that the activation functions used in $f_i^{e,\,1}$ and $f_i^{e,\,2}$ are parametric rectified linear unit (PReLU) functions [48], which generalize the conventional rectified linear unit:

$$f(z_l) = \begin{cases} z_l, & if \quad z_l > 0 \\ a_l z_l, & if \quad z_l \le 0 \end{cases}, \tag{3}$$

where z_l is the l-th channel of a feature map and a_l is the associated learn-able weights to avoid zero gradients in an adaptive manner.

To guide the training of these evolving modules, a mean square error (MSE) is introduced to measure the distance between an estimated template $\widehat{\mathbf{T}}_i$ and every ground-truth template $\mathbf{T}_{i,\,j} \in \mathbb{T}_i$ in a pixel-wise manner. In detail, an MSE based loss function can be defined as in Eq. (4) to be minimized:

$$\underset{f_i^e}{\mathrm{argmin}}\, L_e = \frac{1}{K} \sum_i \frac{1}{|\mathbb{T}_i|} \sum_j \left\| f_i^e(\mathbf{X}) - \mathbf{T}_{i,\,j} \right\|_2^2, \tag{4}$$

where $f_i^e = f_i^{e,\,1} \circ f_i^{e,\,2}$ is a function composition, $|\mathbb{T}_i|$ is the number of templates in set $\mathbb{T}_i$, and $\|\cdot\|_2$ represents an ℓ_2 norm.

With the template-guided evolving modules, the evolving traces of an input $\mathbf{X}$ can be obtained for each KL grade i with the convolution and deconvolution feature maps computed during the forward propagation through $f_i^{e,\,1}$ and $f_i^{e,\,2}$. Particularly, in this study, the final output feature maps derived from $f_i^{e,\,2}$ are adopted as the evolving traces (i.e., $\widehat{\mathbf{T}}_i$, $i = 0, \ldots, K - 1$) for longitudinal KOA severity prediction. Although all the feature maps including the intermediate ones can be also used in the subsequent modules, using only the final ones

helps reduce the model complexity and is compatible with many existing pre-trained image classification backbones in terms of the input channels.

C. KOA Grading with Evolving Traces

A general pipeline for deep learning based KOA grading takes an input image with a classification neural network to obtain a grading score. To further leverage the success of the existing deep architectures, a simple yet flexible mechanism is proposed to integrate the evolving traces for longitudinal grading. In detail, define f_i^g, where $i = 0, 1, \ldots, K-1$, and f^g as the computations of particular neural network classifiers. KOA longitudinal prediction can be obtained from each evolving trace $\widehat{\mathbf{T}}_i$ and the original input image $\mathbf{X}$ as:

$$\widehat{\mathbf{Y}}^{\widehat{\mathbf{T}}_i} = \left(\hat{y}_{t,k}^{\widehat{\mathbf{T}}_i} \right) = f_i^g \left(\widehat{\mathbf{T}}_i \right), \tag{5}$$

$$\widehat{\mathbf{Y}}^{\mathbf{X}} = \left(\hat{y}_{t,k}^{\mathbf{X}} \right) = f^g(\mathbf{X}), \tag{6}$$

where $i = 0, \ldots, K-1$, and $\hat{y}_{t,k}^{\widehat{\mathbf{T}}_i}$ and $\hat{y}_{t,k}^{\mathbf{X}}$ are the output estimations of $y_{t,k}$ associated with their inputs. Note that f_i^g and f^g can be any existing deep architecture devised for general image classification tasks with a softmax activation function for final outputs (e.g., VGG-19, ResNet-50 and ViT). This provides the flexibility to integrate any deep learning architectures with the evolving traces, depending on the tasks in addition to longitudinal KOA grading.

D. Fusion of the Evolving Traces

To predict longitudinal KOA severity by referring to the results from each evolving trace and the original image, a two-stage fusion strategy is devised: 1) pooling from all estimations $\hat{y}_{t,k}^{\widehat{\mathbf{T}}_i}$, $i = 0, \ldots, K-1$, based on the evolving hints, and 2) fusing the pooled evolving estimation with the general estimation $\hat{y}_{t,k}^{\mathbf{X}}$.

Intuitively, at least one severity level of the templates and the associated evolving trace are assumed to be dominant in the fine-grained hints for the longitudinal grading at t, whilst other evolving traces may contain fewer longitudinal hints for the current time. Therefore, a maximum pooling can be conducted on the probability maps obtained from each severity level of the templates to select the most important elements as:

$$\hat{y}_{t,k}^{\widehat{\mathbf{T}}} = \max_i \hat{y}_{t,k}^{\widehat{\mathbf{T}}_i}, \tag{7}$$

which is the longitudinal estimation in line with the evolving traces. Note that in practice Eq. (7) is compatible with different numbers of evolving traces rather than a fixed settings. This provides the flexibility for other similar tasks, which have the progression property with a varying number of grading stages.

Then, a linear combination of $\hat{y}_{t,k}^{\hat{\mathbf{T}}}$ and $\hat{y}_{t,k}^{\mathbf{X}}$, can be adopted for the fusion purpose to derive the final estimation $\hat{\mathbf{Y}}$:

$$\hat{y}_{t,k} = \sum_{s \in \left\{\mathbf{T}, \hat{\mathbf{X}}\right\}} w_s \hat{y}_{t,k}^s, \tag{8}$$

where w_s is the weights to combine the predictions and it can be a hyper-parameter selected during the validation stage. Note that the overall fusion strategy can be viewed as late fusion as it works at the prediction stage of each classifier. The proposed method also works with other fusion strategies such as early fusion.

The prediction of longitudinal KOA grades is a multi-class classification task. To optimize the the parameters in the proposed methods, a cross-entropy loss function can be introduced:

$$\underset{f_i^g, f^g}{\operatorname{argmin}} L_c = -\frac{1}{TK} \sum_{t,k} y_{t,k} \log(\hat{y}_{t,k}). \tag{9}$$

In regard to the above discussions, a basic evolving neural network (ENN) architecture for longitudinal KOA grading is formulated. In line with the evolving traces $\hat{\mathbf{T}}^i$, $i = 0, \ldots, K{-}1$ and the input radiology image $\mathbf{X}$, fine-grained longitudinal KOA severity prediction $\hat{\mathbf{Y}}$ can be obtained.

E. Discriminator for Evolving

Although the template images in $\mathbb{T}_i$ can be representative for their corresponding KL grade i, an exact pixel-level evolution as guided in Eq. (4) between an input image and a template image can be extremely strict. For example, the geometry of the bones between two subjects can be very different and it can be difficult to compare the images of the two subjects directly. To this end, the pixel-level guidance could result in inaccurate evolving hints and negatively affect the longitudinal KL grading performance. Indeed, by treating $\mathbb{T}_i$ as the population subject to a particular probability distribution in regard to the corresponding KL grade i, we can formulate the evolving procedure as the transition from an original image following a source probability distribution to a destination distribution. This formulation is consistent with the design of the widely used generative neural networks which usually involve an adversarial training scheme.

In order to achieve this goal, additional classifiers are introduced to differentiate whether two images belong to the same distribution or not. In this study, K CNNs with simple structures such as ResNet-18 are adopted for this purpose for each evolving module f_i^e, $i = 0, \ldots, K-1$. In detail, the computations are denoted as f_i^d for the module using template images of KL grade i. The response of a real template image $\mathbf{T}_{i,j} \in \mathbb{T}_i$ is viewed as a positive sample and the one of a generated evolution image $\hat{\mathbf{T}}_i$ is viewed as a negative sample. Therefore, f_i^d conducts a binary classification task and outputs a probability with a

sigmoid activation function that an input is a real template image. A binary cross entropy loss function is introduced to optimize these discriminators as illustrated in Eq. (10):

$$\underset{f_i^d}{\operatorname{argmin}} L_d = -\frac{1}{K} \sum_i \frac{1}{|\mathbb{T}_i|} \sum_j \log\left(f_i^d\left(f_i^e(\mathbf{T}_{i,j})\right)\right)$$

$$-\frac{1}{K} \sum_i \log\left(1 - f_i^d\left(f_i^e(\mathbf{X})\right)\right). \tag{10}$$

Note that the parameters of the evolving modules f_i^e, $i = 0, ..., K-1$, are frozen during the optimization of L_d.

Next with these simple discriminator architectures, the proposed method is able to efficiently refine the parameters of the evolving modules f_i^e using an adversarial training scheme.

Algorithm 1: End-to-end adversarial ENN training

Input: A baseline X-ray scan image $\mathbf{X}$ and template sets $\mathbb{T}_i = \{\mathbf{T}_{i,j}\}$.

Output: Parameters of f_i^e, f_i^g, f^g and f_i^d.

1 **begin**
2   initialization;
3   **while** *not stopping criteria* **do**
      // Step 1 Evolving trace estimation
4   **for** $i = 0:K\text{-}1$ **do**
5   compute evolving traces $f_{p,i}^e(\mathbf{X})$;
6   **end**
7   minimize L_e in Eq. (4) and update the parameters of f_i^e;
      // Step 2 longitudinal prediction
8   estimate the longitudinal KOA grades $\hat{\mathbf{Y}}$;
9   minimize L_c in Eq. (9) and update the parameters of f_i^g $(i = 0, ..., K-1)$ and f^g;
      // Step 3 Adversarial training
10   minimize L_d in Eq. (10) and update the parameters of f_i^d $(i = 0, ..., K-1)$, while the parameters of f_i^e are frozen;
11   maximize L_a in Eq. (11) and update the parameters of f_i^e $(i = 0, ..., K-1)$, while the parameters of f_i^d are frozen;
12   **end**
13 **end**

F. Adversarial Training

To let the evolving trace generation modules f_i^e and the evolving discriminators f_i^d work together, an adversarial training scheme is introduced so that the training is not limited to the

two independent loss functions: L_e and L_d. It was originally studied in [49] and has achieved great success for many generation tasks (e.g. [50]).

During the adversarial training, the evolving generators continuously attempt to generate better evolution traces to their corresponding KL grades for a given image, whilst the evolving discriminators are trained to become a better detective for correctly judging whether an image follows the distribution of the real template images with a specified KL grade or not. The equilibrium of this game is achieved when the evolving modules are able to estimate the evolving traces accurately, and the evolving discriminators are left to always randomly guess at 50% confidence for the input images.

In detail, the adversarial training optimizes the loss function L_a as defined in Eq. (11), which further combines Eq. (4) and Eq. (10) to play a minimax game.

$$\underset{f_i^e}{\mathrm{argmax}}\, L_a = -\frac{1}{K} \sum_i \log\left(1 - f_i^d\left(f_i^e(\mathbf{X})\right)\right). \tag{11}$$

Maximizing L_a optimizes the parameters of f_i^e to confuse the discriminators. It helps to produce evolving traces of which the distributions can be close to the ones of the real template images. Note that the parameters of the discriminators f_i^d are frozen during the optimization of L_a. With the above discussions, now an end-to-end adversarial training scheme for KOA grading can be derived. Particularly, Algorithm 1 illustrates the key steps of the proposed adversarial training scheme for ENN.

IV. Experimental Results and Discussions

A. Dataset & Evaluation Metrics

Knee X-ray images from a widely used public dataset - the Osteoarthritis Initiative (OAI) [51] were used to evaluate our proposed method. Longitudinal assessments and measurements were conducted for 4,796 subjects with ages ranging from 45 to 79 for a better understanding of the prevention and treatment of KOA [52]. The outcomes include their clinical data, patient reported outcomes, biospecimen analyses, quantitative image analyses, radiographs and magnetic resonance images. Note that we followed the same the same pre-processing steps as in [8].

Specifically, the baseline cohort of the knee bilateral PA (Posterior-Anterior) fixed flexion X-ray images were adopted for the pre-processing pipeline: 1) resizing to obtain the same physical resolution of 0.14 mm/pixel; 2) center cropping with a window of 2560×2048 pixels; 3) knee joint patch detection and extraction. Note that only the images with available longitudinal KL grades on both knee joints from 0 to 48 months were kept. In total, 3,294 baseline scan X-ray images of knee joints were obtained. Table II lists the KL grade distributions along with the follow-ups. The figures indicate a significant proportion of the patients are on the trend that their KOA status progresses to severer stages. Note that there is a small proportion of the subjects, whose KL grades decreased during the longitudinal visits (e.g., the 36-month grade can be lower than the 24-month's). These images were further

resized to 224×224 and treated as the global scale images in this study. The dataset was partitioned as training, validation and testing sets with proportions of 70%, 10% and 20%, respectively. To evaluate the proposed method in a robust manner, the dataset was randomly split five times to conduct the experiments.

For the sake of convenience, a superscript (n) is introduced to indicate the n-th sample in a particular test split, which contains N samples in total. By comparing the ground truth $y_{t,k}^{(n)}$ and the model prediction $\hat{y}_{t,k}^{(n)}$, an accuracy metric can be obtained, which is the fraction of the correct predictions over all predictions:

$$\frac{1}{NT} \sum_{n,t} I\left(\underset{k}{\mathrm{argmax}}\, y_{t,k}^{(n)} = \underset{k}{\mathrm{argmax}}\, \hat{y}_{t,k}^{(n)} \right), \tag{12}$$

where I is an indicator function. Similarly, longitudinal accuracy metrics can be obtained to measure the accuracy at each temporal point independently. In addition to the categorical perspective, a KL grade is ordinal indicating the KOA severity. The estimation is expected to be close to the ground truth numerically. Thus, a mean absolute error (MAE) can be reported as:

$$\frac{1}{NT} \sum_{n,t} \frac{1}{K} \left| \underset{k}{\mathrm{argmax}}\, y_{t,k}^{(n)} - \underset{k}{\mathrm{argmax}}\, \hat{y}_{t,k}^{(n)} \right|, \tag{13}$$

where $\frac{1}{K}$ is a normalization factor.

B. Experimental Settings

In this study, two types of architectures in regard to the adversarial learning: ENN architecture and ENN architecture with the adversarial training scheme (A-ENN), were deployed to evaluate the performance of the proposed methods. Details of the two architectures are as follows:

- **ENN architectures** – these involve the template-guided evolving traces and KOA grading modules optimized by Eq. (4) and Eq. (1) only. The evolving module involves a convolution layer with 56 output channels and a kernel size of 5 for the latent trace representations, a convolution layer with 12 output channels and a kernel size of 1, and a deconvolution layer of a kernel size 9 to generate template images. Three different deep grading classifiers (i.e., backbone) were studied. 1) **VGG-19** achieved the best performance in [8] for baseline KOA grading. The last layer of VGG-19 was subsituted as a fully connected layer with an output dimension 5×5 in consistent with the KL grading scores and the longitudinal duration. 2) **ResNet-50** was adopted by altering its output layer similar to VGG-19. 3) **ViT** as one of the visual transformer models was also utilized, which have achieved the state-of-the-art performance in many computer vision tasks.

- **Adversarial ENN architectures** – these architectures are based on the ENN and additionally introduce ResNet-18 models as the discriminators. ResNet-18 is a simple architecture with less model complexity and computational costs. Note that an A-ENN contains 5 independent discriminators for the evolving traces using the templates of 5-level KL grades.

Pre-trained weights on ImageNet [53] were applied to initialize the VGG-19, ResNet-50, ResNet-18 and ViT networks. Overall, the ENN architectures were trained in an end-to-end manner using stochastic gradient descent optimizers [54] with a momentum 0.9, an initial learning rate of 5×10^{-4} and a weight decay of 5×10^{-4} every 5 epochs. The batch size was set to 1, 1 and 8 for the ENN of VGG-19, ResNet-50 and ViT backbones, respectively. The template sets were based on all training data in line with their KL grades. This could be helpful to formulate a comprehensive understanding of each evolution pipeline by considering adequate amount of data. Initially the evolving modules were independently trained for 10 epochs only with the loss function defined in Eq. (4). Then, the other losses were also estimated and the parameters of all modules were optimized. During the training, data augmentations were adopted to reduce the overfitting risk. In detail, the jitters of brightness, contrast, saturation and hue were set to 30%. The experiments were conducted using PyTorch 1.9.0 with an Nvidia GTX 3080 GPU.

C. Overall Performance

Table I lists the overall performance of the proposed methods. The A-ENN with VGG-19 Backbone achieves the best overall accuracy 62.7%. Compared to the ENN without the adversarial training scheme, of which the accuracy is 61.4%, it demonstrates the effectiveness of the adversarial training. Without the evolving traces as the fine-grained patterns for the KOA longitudinal progression hints, the conventional VGG-19 network achieves a less precise performance of accuracy 59.8%. As the experiments were evaluated five times based on different random splits of the dataset, the paired sample t-test was conducted to demonstrate the significance of the improvement of the (A-)ENN method compared to the backbone (baseline) method. *p*-values are listed in Table I, which suggest that the improvement is statistically significant. Standard deviation values are listed in Table I as well.

Similar trends can be found from the ResNet and ViT networks and the ENNs based on them. Note that the ViT based methods are not as good as the VGG-19 based ones, although it showed the state-of-the-art performance in general image classification tasks. The potential reasons could be that training transformer based neural networks for vision tasks usually requires a huge amount of data, whilst the size of the current OAI dataset is small. It indicates the necessity to devise proper architectures to formulate visual patterns from X-Ray images for KOA. The MAE metrics perform the same trend as the accuracy ones, except that the MAE of the VGG-19 based ENN increases compared to the VGG-19 itself. As MAE is not the objective to optimize the loss, this could be the reason leading to the outlier.

Fig. 3 visualizes the performances of longitudinal prediction accuracy at each temporal point. Overall, the ENNs with adversarial training achieve the best performance for the

predictions of the baseline and follow-ups. Note that downward trends can be noticed for all series (methods). For example, the accuracy values at the 36-month follow-ups decrease compared to those at the 24-month ones for all VGG-19 based methods. The trends indicate that predicting KOA grades can be more challenging for a follow-up after a longer time from the baseline visit.

Fig. 4 illustrates the evolution traces of an example image to different KL grades. It can be observed that each channel formulates its own feature map, which represents a particular perspective of the relation between the input image and the target KL grade in regard to the bone shape and texture.

D. Demographic based Performance

Tables III - V list the performance of the methods based on the demographic groups in terms of gender, age and BMI at the baseline visit. Overall, an increasing trend of the accuracy can be observed from a backbone model to its ENN models for each group.

For the two gender groups, the accuracy of the female group is consistently superior compared to that of the male group for all methods. In terms of different age groups, the KL grade prediction is more accurately identified for the elderly population. With the increasing BMI from normal to obesity, the KL grade predictions tend to be more accurate. This study mainly focuses on deep architectures for accurate visual patterns from medical images. Nonetheless, such inter-group variations suggest that these demographic information could be potential predictors for longitudinal KOA grading and a proper mechanism to consider them with radiology scans jointly is an encouraging future research direction.

E. Discriminator Analysis

A discriminator determines whether an input image belongs to the distribution of genuine template images or not. It is expected that it would be challenging to distinguish the representations from the generated evolving traces and the genuine template images. As all discriminators in this study conduct a binary classification, ROC curves and their associated AUC values can be used to analyze their effectiveness. Fig. 5 illustrates the AUC values of these discriminator ROC curves along their every 100 iteration steps for the A-ENN with a VGG-19 backbone. Along the iterations, all discriminators' AUC values vary around 0.5, which suggests that they are not able to tell the difference between the distributions. These findings confirm that the adversarial learning is helpful to reduce the gap between the generated evolving traces and the genuine template images, which increases the performance of the ENN in regard to the accuracy.

F. Feature Space Analysis

To further understand the improvements of the feature space obtained from A-ENNs and their associated backbone models, t-SNE (t-distributed stochastic neighbor embedding) [55] was conducted on the inputs of the last fully connected layers in these models. t-SNE creates a single map that reveals the structure of the feature space at different scales. Fig. 6 illustrates the comparisons of the t-SNE visualization between three groups of models: 1) ResNet-50 vs. A-ENN (ResNet-50 Backbone), 2) VGG-19 vs. A-ENN (VGG-19 Backbone)

and 3) ViT vs. A-ENN (ViT Backbone). For each method, in a 2-dimensional space, the first two major components of the t-SNE results obtained from the test set are visualized.

The colors of the data points represents the KL grade of each sample. It can be observed that for both comparison groups, A-ENNs present a more separable data distribution among the five categories, whilst the backbone models without evolving traces present dispersed data points over an extensive intersection area of the five categories. That is, the ENNs with an adversarial training scheme contribute to the robustness of the feature space, and thus the model's performance is superior to those of the backbone models.

G. Longitudinal Conversion Analysis

Longitudinal conversion is to predict how a patient could convert to more serious stage of the KOA disease in the future. By defining the case that the average longitudinal KL grades (i.e., from 12 months to 48 months) increase higher than 1 compared to the baseline assessment as a positive case, an evaluation based on a binary classification can be conducted for a change detection purpose. In detail, we have the score for the change as:

$$\hat{y}_c = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{k=0}^{K-1} k\hat{y}_{t,k} - \sum_{k=0}^{K-1} k\hat{y}_{0k}, \tag{14}$$

which can be viewed as a weighted score in line with the KL grades. As $\hat{y}_c$ presents its continuous property, a threshold can be identified in practice according to specific applications. To evaluate the diagnostic ability of $\hat{y}_c$ derived by different methods, the ROC curves and their associated AUC values are shown in Fig. 7 and Table VI, respectively. It can be observed that the A-ENN methods perform better than the backbone only methods. For example, the A-ENN with a VGG-19 backbone achieves an AUC 0.637, whilst VGG-19 achieves an AUC 0.561.

V. CONCLUSION

In this study, a novel deep learning architecture - evolving neural network with an adversarial training scheme is presented for a fine-grained learning task - longitudinal KOA grading. ENN is devised based on the fact that the disease presents progressive properties from mild to severe. It involves the evolving patterns of an input image by comparing it to the template images of different KL grades to formulate evolution traces as the fine-grained domain knowledge. Comprehensive experimental results demonstrated the effectiveness of the proposed method on a widely used public dataset - the Osteoarthritis Initiative (OAI) Dataset in terms of an accuracy 62.7%. Note that ENN can be introduced to any other tasks and deep architectures, in which the responses present progressive representations. For our future work, two major directions can be considered: 1) multi-task learning methods by investigating different progression objectives such as the perspectives from pain and JSN jointly for a robust medical imaging KOA representation; and 2) multi-modal learning methods by considering other commonly used modalities such as MRI scans and demographics for longitudinal modelling.

Acknowledgement

REFERENCES

[1]. Wallace IJ, Worthington S, Felson DT, Jurmain RD, Wren KT, Maijanen H, Woods RJ, and Lieberman DE, "Knee osteoarthritis has doubled in prevalence since the mid-20th century," Proceedings of the National Academy of Sciences, vol. 114, no. 35, pp. 9332–9336, 2017.

[2]. Kellgren JH and Lawrence J, "Radiological assessment of osteoarthrosis," Annals of the Rheumatic Diseases, vol. 16, no. 4, p. 494, 1957. [PubMed: 13498604]

[3]. Cross M, Smith E, Hoy D, Nolte S, Ackerman I, Fransen M, Bridgett L, Williams S, Guillemin F, Hill CL et al. , "The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study," Annals of the rheumatic diseases, vol. 73, no. 7, pp. 1323–1330, 2014. [PubMed: 24553908]

[4]. Shamir L, Ling SM, Scott W, Hochberg M, Ferrucci L, and Goldberg IG, "Early detection of radiographic knee osteoarthritis using computer-aided analysis," Osteoarthritis and Cartilage, vol. 17, no. 10, pp. 1307–1312, 2009. [PubMed: 19426848]

[5]. Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[6]. He K, Zhang X, Ren S, and Sun J, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[7]. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al. , "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[8]. Chen P, Gao L, Shi X, Allen K, and Yang L, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," Computerized Medical Imaging and Graphics, vol. 75, pp. 84–92, 2019. [PubMed: 31238184]

[9]. Liu B, Luo J, and Huang H, "Toward automatic quantification of knee osteoarthritis severity using improved faster r-cnn," International Journal of Computer Assisted Radiology and Surgery, vol. 15, no. 3, pp. 457–466, 2020. [PubMed: 31938993]

[10]. Norman B, Pedoia V, Noworolski A, Link TM, and Majumdar S, "Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs," Journal of Digital Imaging, vol. 32, no. 3, pp. 471–477, 2019. [PubMed: 30306418]

[11]. Halilaj E, Le Y, Hicks J, Hastie T, and Delp S, "Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative," Osteoarthritis and Cartilage, vol. 26, no. 12, pp. 1643–1650, 2018. [PubMed: 30130590]

[12]. Widera P, Welsing PM, Ladel C, Loughlin J, Lafeber FP, Petit Dop F, Larkin J, Weinans H, Mobasheri A, and Bacardit J, "Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data," Scientific Reports, vol. 10, no. 1, pp. 8427–8427, 2020. [PubMed: 32439879]

[13]. Ntakolia C, Kokkotis C, Moustakidis S, and Tsaopoulos D, "Prediction of joint space narrowing progression in knee osteoarthritis patients," Diagnostics (Basel), vol. 11, no. 2, pp. 285–, 2021. [PubMed: 33670414]

[14]. Tiulpin A, Klein S, Bierma-Zeinstra SM, Thevenot J, Rahtu E, van Meurs J, Oei EH, and Saarakkala S, "Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data," Scientific Reports, vol. 9, no. 1, pp. 1–11, 2019. [PubMed: 30626917]

[15]. Hu K, Wang Z, Ehgoetz Martens K, and Lewis S, "Vision-based freezing of gait detection with anatomic patch based representation," in Asian Conference on Computer Vision Springer, 2018, pp. 564–576.

[16]. Hu K, Wang Z, Wang W, Martens KAE, Wang L, Tan T, Lewis SJ, and Feng DD, "Graph sequence recurrent neural network for vision-based freezing of gait detection," IEEE Transactions on Image Processing, vol. 29, pp. 1890–1901, 2019.

[17]. Zheng H, Fu J, Zha Z-J, and Luo J, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5012–5021.

[18]. Mei S, Ma M, Wan S, Hou J, Wang Z, and Feng DD, "Patch based video summarization with block sparse representation," IEEE Transactions on Multimedia, vol. 23, pp. 732–747, 2020.

[19]. Hu K, Wang Z, Martens KAE, Hagenbuchner M, Bennamoun M, Tsoi AC, and Lewis SJ, "Graph fusion network-based multimodal learning for freezing of gait detection," IEEE Transactions on Neural Networks and Learning Systems, 2021.

[20]. Kokkotis C, Moustakidis S, Papageorgiou E, Giakas G, and Tsaopoulos D, "Machine learning in knee osteoarthritis: A review," Osteoarthritis and Cartilage Open, vol. 2, no. 3, p. 100069, 2020. [PubMed: 36474688]

[21]. Brahim A, Jennane R, Riad R, Janvier T, Khedher L, Toumi H, and Lespessailles E, "A decision support tool for early detection of knee osteoarthritis using x-ray imaging and machine learning: Data from the osteoarthritis initiative," Computerized Medical Imaging and Graphics, vol. 73, pp. 11–18, 2019. [PubMed: 30784984]

[22]. Antony J, McGuinness K, O'Connor NE, and Moran K, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in International Conference on Pattern Recognition. IEEE, 2016, pp. 1195–1200.

[23]. Antony J, McGuinness K, Moran K, and O'Connor NE, "Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks," in International Conference on Machine Learning and Data Mining in Pattern Recognition. Springer, 2017, pp. 376–390.

[24]. Abedin J, Antony J, McGuinness K, Moran K, OConnor NE, Rebholz-Schuhmann D, and Newell J, "Predicting knee osteoarthritis severity: comparative modeling based on patients data and plain x-ray images," Scientific reports, vol. 9, no. 1, pp. 1–11, 2019. [PubMed: 30626917]

[25]. Nasser Y, Jennane R, Chetouani A, Lespessailles E, and El Hassouni M, "Discriminative regularized auto-encoder for early detection of knee osteoarthritis: Data from the osteoarthritis initiative," IEEE transactions on medical imaging, vol. 39, no. 9, pp. 2976–2984, 2020. [PubMed: 32286962]

[26]. Dalia Y, Bharath A, Mayya V, and Kamath SS, "Deepoa: Clinical decision support system for early detection and severity grading of knee osteoarthritis," in International Conference on Computer, Communication and Signal Processing. IEEE, 2021, pp. 250–255.

[27]. Yong CW, Teo K, Murphy BP, Hum YC, Tee YK, Xia K, and Lai KW, "Knee osteoarthritis severity classification with ordinal regression module," Multimedia Tools and Applications, pp. 1–13, 2021.

[28]. Jain RK, Sharma PK, Gaj S, Sur A, and Ghosh P, "Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network," arXiv preprint arXiv:2106.14292, 2021.

[29]. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, and Saarakkala S, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," Scientific reports, vol. 8, no. 1, pp. 1–10, 2018. [PubMed: 29311619]

[30]. Li MD, Chang K, Bearce B, Chang CY, Huang AJ, Campbell JP, Brown JM, Singh P, Hoebel KV, Erdo mu D et al. , "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging," NPJ digital medicine, vol. 3, no. 1, pp. 1–9, 2020. [PubMed: 31934645]

[31]. Nguyen HH, Saarakkala S, Blaschko MB, and Tiulpin A, "Semixup: in-and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain

radiographs," IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 4346–4356, 2020. [PubMed: 32804644]

[32]. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, Bereket M, Patel BN, Yeom KW, Shpanskaya K et al. , "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet," PLoS medicine, vol. 15, no. 11, p. e1002699, 2018. [PubMed: 30481176]

[33]. Pedoia V, Lee J, Norman B, Link TM, and Majumdar S, "Diagnosing osteoarthritis from t2 maps using deep learning: an analysis of the entire osteoarthritis initiative baseline cohort," Osteoarthritis and Cartilage, vol. 27, no. 7, pp. 1002–1010, 2019. [PubMed: 30905742]

[34]. Wang Y, You L, Chyr J, Lan L, Zhao W, Zhou Y, Xu H, Noble P, and Zhou X, "Causal discovery in radiographic markers of knee osteoarthritis and prediction for knee osteoarthritis severity with attention–long short-term memory," Frontiers in Public Health, p. 845, 2020.

[35]. Chan L, Li H, Chan P, and Wen C, "A machine learning-based approach to decipher multi-etiology of knee osteoarthritis onset and deterioration," Osteoarthritis and Cartilage Open, vol. 3, no. 1, p. 100135, 2021. [PubMed: 36475069]

[36]. Janvier T, Jennane R, Toumi H, and Lespessailles E, "Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the osteoarthritis initiative," Osteoarthritis and cartilage, vol. 25, no. 12, pp. 2047–2054, 2017. [PubMed: 28935435]

[37]. Chen L-C, Papandreou G, Schroff F, and Adam H, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[38]. Ronneberger O, Fischer P, and Brox T, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[39]. Cheung JC-W, Tam AY-C, Chan L-C, Chan P-K, and Wen C, "Superiority of multiple-joint space width over minimum-joint space width approach in the machine learning for radiographic severity and knee osteoarthritis progression," Biology, vol. 10, no. 11, p. 1107, 2021. [PubMed: 34827100]

[40]. Guan B, Liu F, Haj-Mirzaian A, Demehri S, Samsonov A, Neogi T, Guermazi A, and Kijowski R, "Deep learning risk assessment models for predicting progression of radiographic medial joint space loss over a 48-month follow-up period," Osteoarthritis and cartilage, vol. 28, no. 4, pp. 428–437, 2020. [PubMed: 32035934]

[41]. Guan B, Liu F, Mizaian AH, Demehri S, Samsonov A, Guermazi A, and Kijowski R, "Deep learning approach to predict pain progression in knee osteoarthritis," Skeletal Radiology, vol. 51, no. 2, pp. 363–373, 2022. [PubMed: 33835240]

[42]. Pedoia V, Haefeli J, Morioka K, Teng H-L, Nardo L, Souza RB, Ferguson AR, and Majumdar S, "Mri and biomechanics multidimensional data analysis reveals r2-r1ρ as an early predictor of cartilage lesion progression in knee osteoarthritis," Journal of Magnetic Resonance Imaging, vol. 47, no. 1, pp. 78–90, 2018. [PubMed: 28471543]

[43]. Ashinsky BG, Bouhrara M, Coletta CE, Lehallier B, Urish KL, Lin P-C, Goldberg IG, and Spencer RG, "Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative," Journal of Orthopaedic Research, vol. 35, no. 10, pp. 2243–2250, 2017. [PubMed: 28084653]

[44]. Du Y, Shan J, and Zhang M, "Knee osteoarthritis prediction on mr images using cartilage damage index and machine learning methods," in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017, pp. 671–677.

[45]. Du Y, Almajalid R, Shan J, and Zhang M, "A novel method to predict knee osteoarthritis progression on mri using machine learning methods," IEEE Transactions on Nanobioscience, vol. 17, no. 3, pp. 228–236, 2018. [PubMed: 29994316]

[46]. Schiphof D, Boers M, and Bierma-Zeinstra SM, "Differences in descriptions of kellgren and lawrence grades of knee osteoarthritis," Annals of the Rheumatic Diseases, vol. 67, no. 7, pp. 1034–1036, 2008. [PubMed: 18198197]

[47]. Dong C, Loy CC, and Tang X, "Accelerating the super-resolution convolutional neural network," in European Conference on Computer Vision. Springer, 2016, pp. 391–407.

[48]. He K, Zhang X, Ren S, and Sun J, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[49]. Radford A, Metz L, and Chintala S, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[50]. Wang G, Kang W, Wu Q, Wang Z, and Gao J, "Generative adversarial network (gan) based data augmentation for palmprint recognition," in 2018 Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2018, pp. 1–7.

[51]. Nevitt M, Felson D, and Lester G, "The osteoarthritis initiative: Protocol for the cohort study," 2006.

[52]. Eckstein F, Wirth W, and Nevitt MC, "Recent advances in osteoarthritis imagingthe osteoarthritis initiative," Nature Reviews Rheumatology, vol. 8, no. 10, pp. 622–630, 2012. [PubMed: 22782003]

[53]. Deng J, Dong W, Socher R, Li L-J, Li K, and Fei-Fei L, "Imagenet: A large-scale hierarchical image database," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 248–255.

[54]. Robbins H and Monro S, "A stochastic approximation method," The Annals of Mathematical Statistics, pp. 400–407, 1951.

[55]. Van der Maaten L and Hinton G, "Visualizing data using t-sne." Journal of machine learning research, vol. 9, no. 11, 2008.
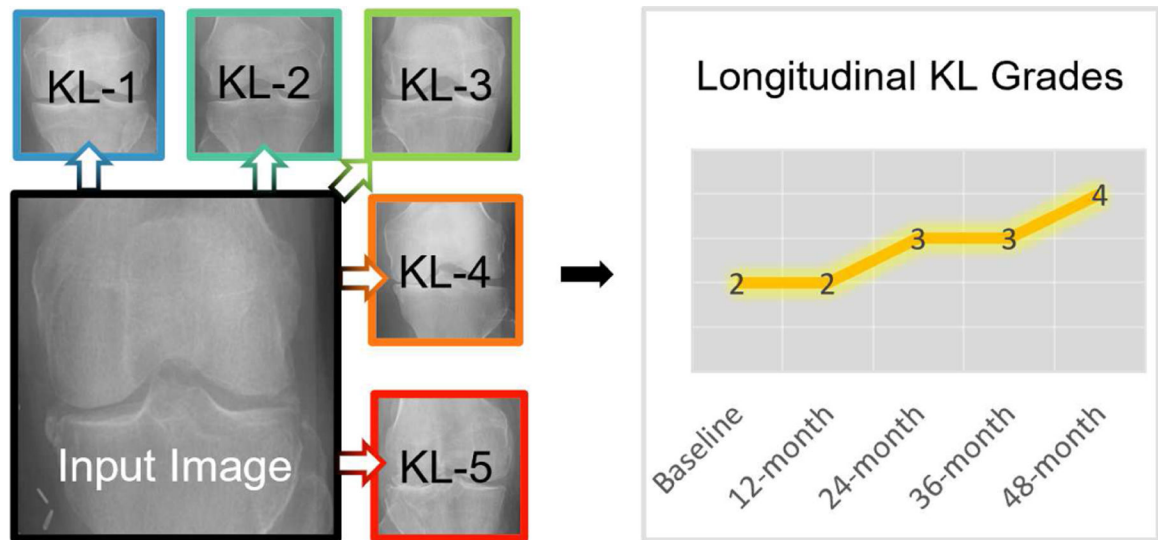
Fig. 1:
Illustration of longitudinal KOA severity prediction. Comparing and evolving an input image to the template images of different KL grades, fine-grained KOA patterns are obtained to predict the longitudinal grades of the disease from the baseline scan to up to 48 months.
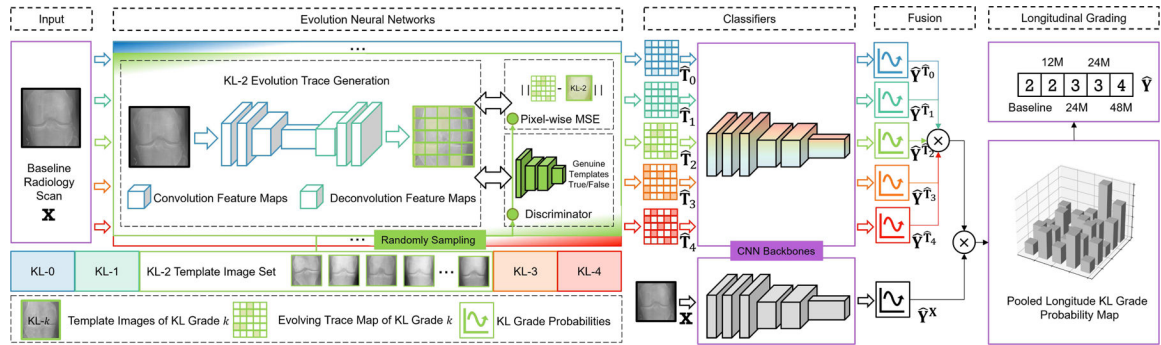
Fig. 2:

Illustration of the proposed ENN architecture with an adversarial learning scheme for longitudinal KOA severity prediction. It consists of three major components: 1) Evolving trace estimation with a discriminator for adversarial training; 2) Classification of raw longitudinal KOA grading probabilities from multiple evolving traces and the original input image; and 3) Fusion of the raw probabilities for producing the final KL longitudinal grade probability map.
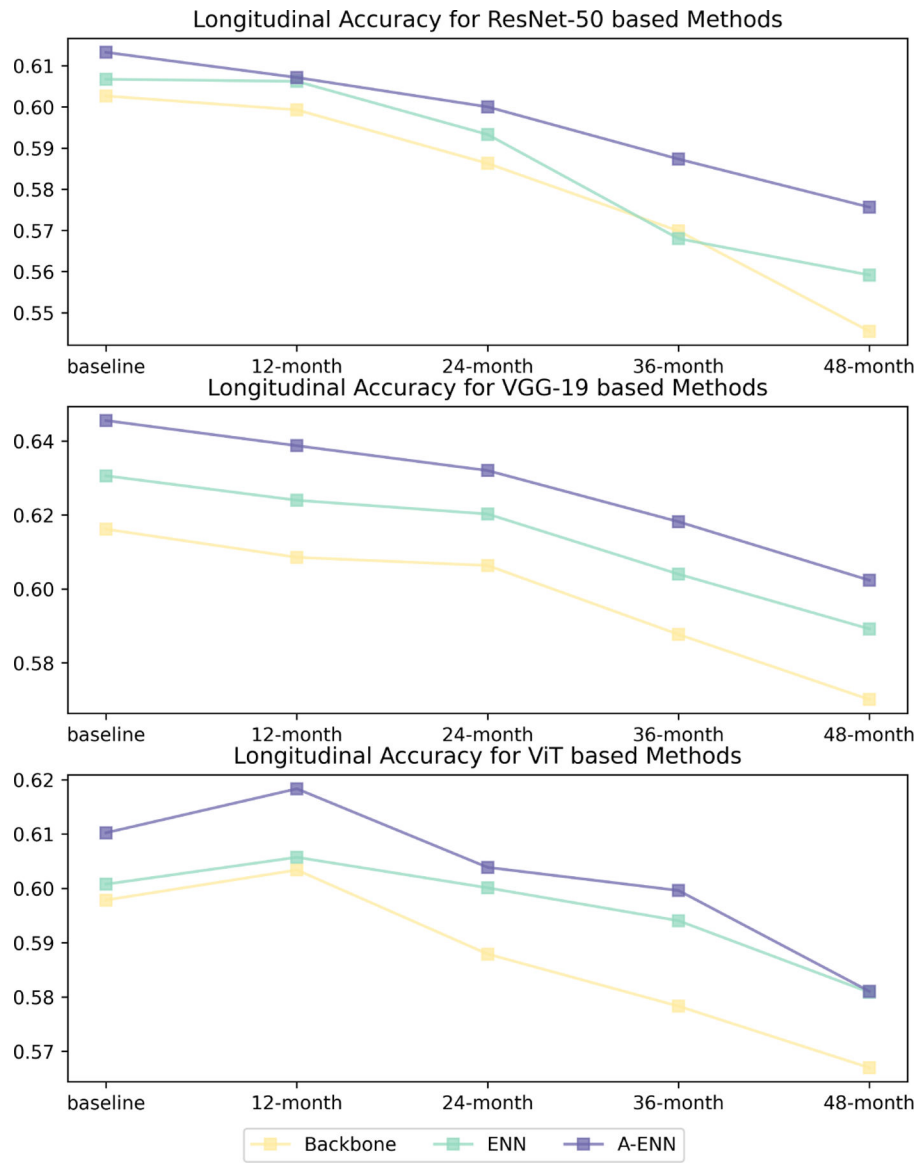
Fig. 3:
Illustration of the longitudinal accuracy over time of different modelling methods.
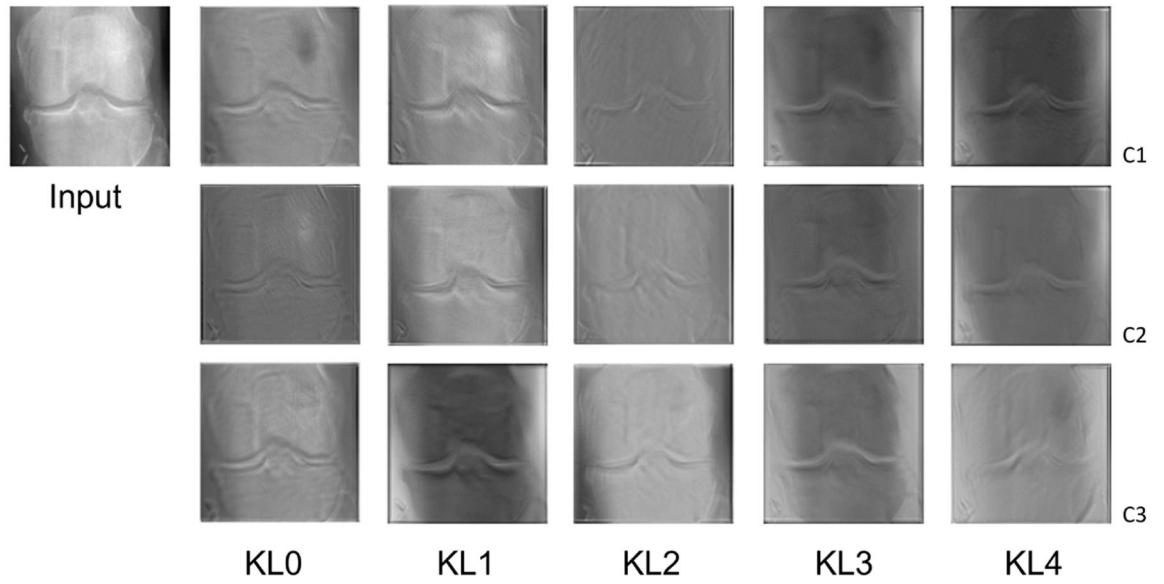
Fig. 4:
Illustration of the evolution traces of the three channels C1-C3 for each KL grade. For a particular channel, different proportions are enhanced based on the pixel brightness when comparing the traces, which suggests that specific patterns are focused in line with their target KL grade.
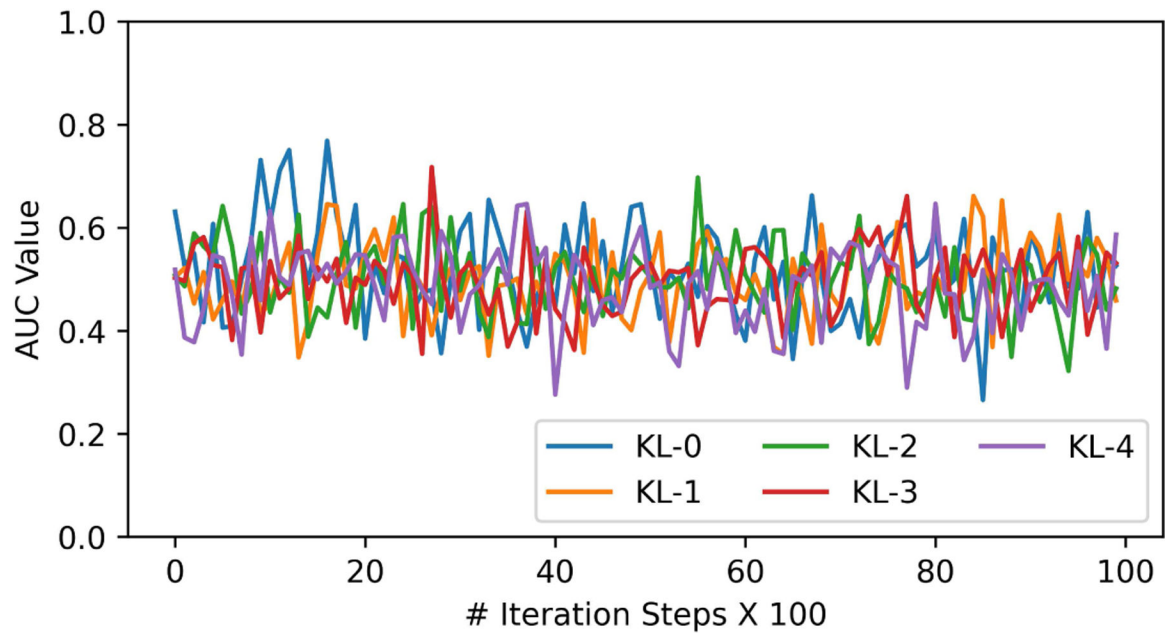
Fig. 5:
Illustration of the AUC values of the discriminators for different KL grades along the iteration steps.
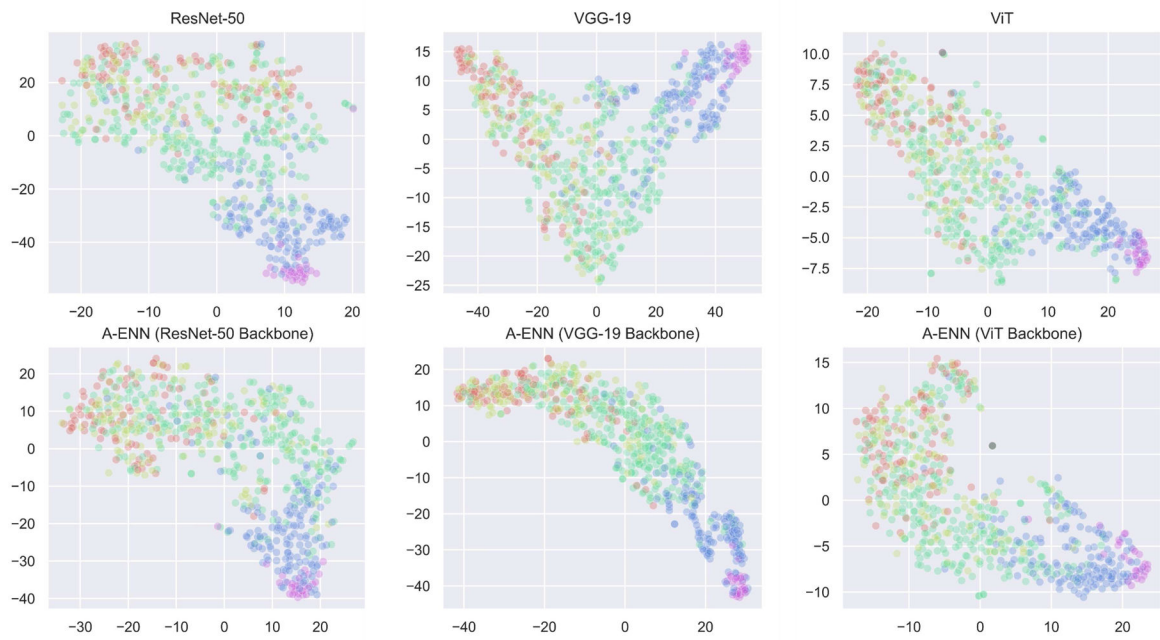
Fig. 6:

Visualization of the feature distributions by t-SNE for the comparisons between A-ENNs and their backbone only counterparts.
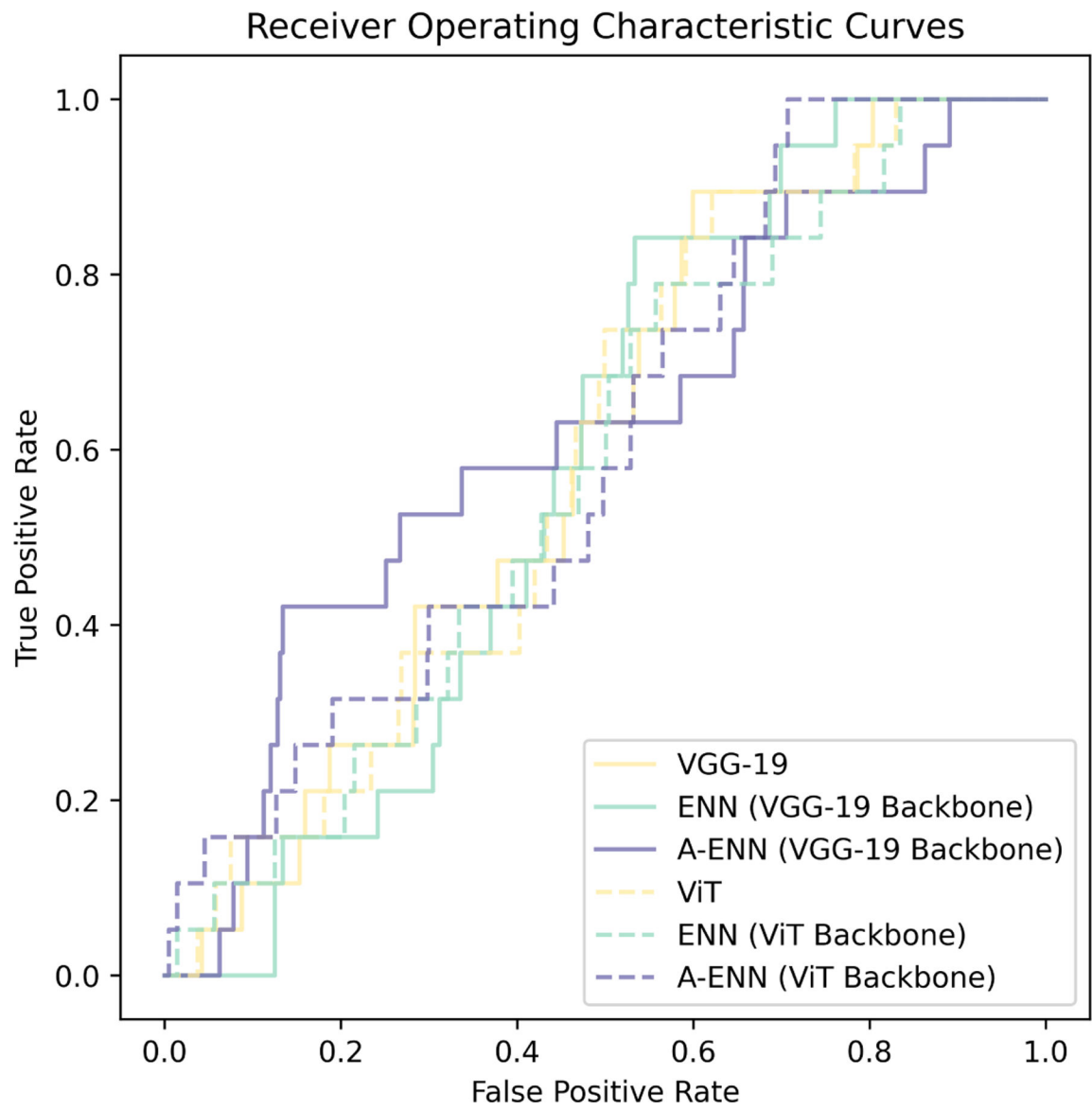
Fig. 7:
Illustration of the ROC curves for identifying KL grade change during the longitudinal progression.

TABLE I:

Overall performance of longitudinal KOA grading

| Method | | Accuracy | p | Baseline | Longitudinal Accuracy | | | | | MAE |
					12-Month	24-Month	36-Month	48-Month		
ResNet-50	Backbone	0.581±0.003	-	0.603±0.007	0.599±0.005	0.586±0.007	0.57±0.007	0.545±0.015		0.134±0.003
	ENN	0.587±0.005	0.061	0.607±0.011	0.606±0.011	0.593±0.008	0.568±0.009	0.559±0.016		0.131±0.005
	A-ENN	0.597±0.008	0.029	0.613±0.015	0.607±0.004	0.6±0.008	0.588±0.011	0.575±0.015		0.13±0.005
VGG19	Backbone	0.598±0.019	-	0.616±0.021	0.609±0.026	0.606±0.016	0.588±0.017	0.57±0.021		0.121±0.005
	ENN	0.614±0.009	0.090	0.631±0.015	0.624±0.01	0.62±0.011	0.604±0.009	0.589±0.01		0.124±0.006
	A-ENN	**0.627±0.007**	0.059	**0.646±0.009**	**0.639±0.009**	**0.632±0.01**	**0.618±0.008**	**0.602±0.012**		**0.117±0.004**
ViT	Backbone	0.587±0.015	-	0.598±0.014	0.603±0.016	0.588±0.021	0.578±0.019	0.567±0.022		0.133±0.005
	ENN	0.596±0.013	0.087	0.601±0.017	0.606±0.015	0.600±0.017	0.594±0.010	0.581±0.013		0.130±0.003
	A-ENN	0.603±0.011	0.016	0.610±0.020	0.618±0.014	0.604±0.015	0.600±0.013	0.581±0.010		0.129±0.005

TABLE II:

Statistics of KL grades in regard to the baseline and follow-up visits

KL	Baseline	12-Month	24-Month	36-Month	48-Month
0	502	440	422	394	379
1	560	482	456	406	375
2	1,451	1,486	1,480	1,485	1,465
3	658	725	748	783	786
4	123	161	188	226	289

TABLE III:

Accuracy by gender groups

	ResNet	ENN (ResNet)	A-ENN (ResNet)
Gender Gender	0.568±0.011	0.576±0.009	**0.591±0.023**
(Male) (Female)	0.591±0.006	0.595±0.013	**0.602±0.019**
	VGG19	**ENN (VGG19)**	**A-ENN (VGG19)**
Gender (Male)	0.589±0.032	0.606±0.02	**0.618±0.019**
Gender (Female)	0.605±0.025	0.619±0.021	**0.635±0.015**
	ViT	**ENN (ViT)**	**A-ENN (ViT)**
Gender (Male)	0.573±0.009	0.586±0.013	**0.595±0.010**
Gender (Female)	0.598±0.020	0.605±0.020	**0.608±0.013**

TABLE IV:

Accuracy by age groups

	ResNet	**ENN (ResNet)**	**A-ENN (ResNet)**
Age ($<$ 50)	0.524±0.043	**0.534±0.024**	0.515±0.039
Age (50–59)	0.597±0.017	0.595±0.017	**0.610±0.009**
Age (60 – 69)	0.584±0.014	0.578±0.022	**0.597±0.005**
Age (70)	0.573±0.019	**0.606±0.008**	0.605±0.026
	VGG19	**ENN (VGG19)**	**A-ENN (VGG19)**
Age ($<$ 50)	0.537±0.046	0.525±0.053	**0.580±0.051**
Age (50–59)	0.587±0.033	0.621±0.016	**0.627±0.031**
Age (60 – 69)	0.615±0.024	0.612±0.019	**0.627±0.022**
Age (70)	0.607±0.023	0.634±0.025	**0.640±0.032**
	ViT	**A-ENN (ViT)**	**ENN (ViT)**
Age ($<$ 50)	0.548±0.048	0.570±0.048	**0.579±0.032**
Age (50–59)	0.588±0.019	**0.600±0.012**	0.599±0.026
Age (60 – 69)	0.583±0.021	0.595±0.026	**0.597±0.026**
Age (70)	0.603±0.034	0.599±0.023	**0.620±0.030**

TABLE V:

Accuracy by BMI groups

	ResNet	ENN (ResNet)	A-ENN (ResNet)
BMI (Normal)	0.578±0.032	0.581±0.036	**0.597±0.031**
BMI (Pre-obesity)	0.564±0.016	0.573±0.023	**0.586±0.009**
BMI (Obesity)	0.599±0.017	0.604±0.019	**0.608±0.026**
	VGG19	ENN (VGG19)	A-ENN (VGG19)
BMI (Normal)	0.566±0.022	0.622±0.026	**0.622±0.02**
BMI (Pre-obesity)	0.596±0.017	0.594±0.021	**0.613±0.027**
BMI (Obesity)	0.616±0.034	0.632±0.032	**0.646±0.019**
	ViT	A-ENN (ViT)	ENN (ViT)
BMI (Normal)	0.578±0.04	0.585±0.038	**0.585±0.019**
BMI (Pre-obesity)	0.572±0.013	0.575±0.011	**0.592±0.026**
BMI (Obesity)	0.606±0.03	**0.625±0.039**	0.623±0.034

TABLE VI:

Longitudinal conversion analysis

Methods	AUC
ResNet-50	0.587
ENN (ResNet-50 Backbone)	0.578
A-ENN (ResNet-50 Backbone)	**0.588**
VGG-19	0.561
ENN (VGG-19 Backbone)	0.626
A-ENN (VGG-19 Backbone)	**0.637**
ViT	0.546
ENN (ViT Backbone)	0.596
A-ENN (ViT Backbone)	**0.630**