DEEP SEMI-SUPERVISED ACTIVE LEARNING FOR KNEE OSTEOARTHRITIS SEVERITY GRADING

Abu Mohammed Raisuddin[*] *Huy Hoang Nguyen*[*] *Aleksei Tiulpin*

Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Finland

[*]Equal Contributions

ABSTRACT

This paper tackles the problem of developing active learning (AL) methods in the context of knee osteoarthritis (OA) diagnosis from X-ray images. OA is known to be a huge burden for society, and its associated costs are constantly rising. Automatic diagnostic methods can potentially reduce these costs, and Deep Learning (DL) methodology may be its key enabler. To date, there have been numerous studies on knee OA severity grading using DL, and all but one of them assume a large annotated dataset available for model development. In contrast, our study shows one can develop a knee OA severity grading model using AL from as little as 50 samples randomly chosen from a pool of unlabeled data. The main insight of this work is that the performance of AL improves when the model developer leverages the consistency regularization technique, commonly applied in semi-supervised learning.

***Index Terms*—** Deep Active Learning, Knee Osteoarthritis, Epistemic Uncertainty, Monte-Carlo Dropout, Consistency Regularization

1. INTRODUCTION

Osteoarthritis (OA) is a joint disease affecting hundreds of millions of people worldwide [1]. This musculoskeletal disorder is degenerative. The severity of OA is commonly assessed using X-ray imaging according to the Kellgren-Lawrence (KL) system [2], which consists of 5 grades scaling from KL 0 (no OA) to KL 4 (severe OA) (see Figure 1).

The knee joint is the largest one in the human body, and it is one of the most affected joint by OA. A patient with knee OA (KOA) may experience knee pain at its early stages, and has poor life-quality due to disability at the final stage. Undergoing total knee replacement is the only available treatment option for patients with severe KOA. As a result, the burden of OA is large at a societal level. For instance, OA was in the top-2 of the most expensive health expenditures in the United States in 2013 [3]. One of the potential solutions to reduce the negative impacts of OA is its early detection where deep learning (DL) can be useful.

Recent literature has shown that DL-based methods have high potential for automatic KOA severity grading [4, 5, 6].

Fig. 1. Knee radiographs with KL grades

Nevertheless, these supervised learning (SL)-based methods were conducted in an idealized scenario, in which a large amount of labeled samples was available. Such a huge demand for annotations is expensive in practice, as it requires highly skilled radiologist(s) to generate the training dataset.

Recently, there was an attempt to achieve annotation-efficient training in KOA grading via semi-supervised learning (SSL), which needs a small amount of labeled samples along with low-cost unlabeled data [7]. The core idea was to use the consistency regularization (CR) approach to make the model invariant to perturbations in input and parameter spaces. However, the authors of [7] assumed that sets of labeled samples with *equally distributed KL grades* were given. We argue that such an assumption is *not realistic*, as no prior knowledge on labels in the unlabeled data pool is usually available.

In this paper, contrary to the prior art, we instead consider a realistic scenario, in which we do not have any prior knowledge about the data distribution. We aim to construct the set of labeled samples in an iterative manner while leveraging pre-selected unlabeled data during the iterative acquisition process. Here, we adapt a Deep Active Learning (DAL) approach and combine it with CR approach of SSL. In summary, our contributions are as follows:

1. We apply DAL methods for annotation-efficient training in automatic KOA severity grading. To the best of our knowledge, this is the first study to apply DAL for KOA severity prediction.
2. We utilize CR during DAL training processes.
3. We systematically study the effect of including SSL training in DAL methods on the performance of OA severity grading.

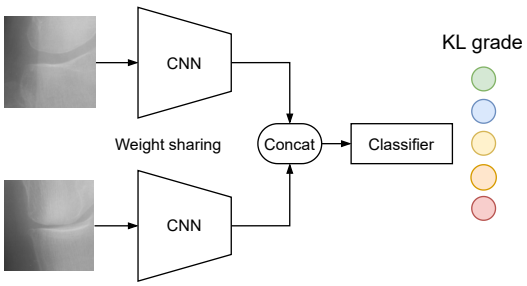


Fig. 2. Siamese network

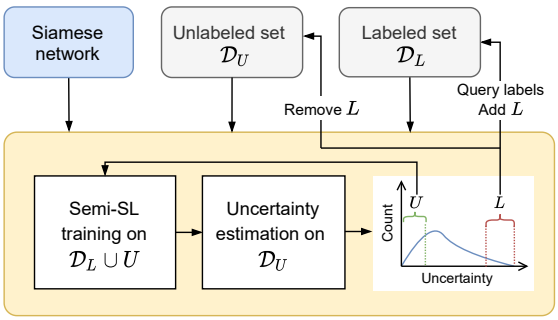


Fig. 3. Our workflow

2. METHODS

2.1. Semi-supervised learning

Consistency regularization (CR) is a central idea of a variety of SSL methods [7, 8]. Consider a data sample $\mathbf{x}$, two semantics-preserving augmentations T and T', and a neural network, denoted by the parametric function f_θ with the parameters θ. CR aims to enforce invariance of f_θ to minor input and parameter perturbations. Formally, the consistency regularizer can be written as

$$\|f_{\theta'}(T(\mathbf{x})) - f_{\theta''}(T'(\mathbf{x}))\|_2^2 \tag{1}$$

where θ' and θ'' denote the model parameters under dropout. We note that, in the context of this study, the term SSL and CR can be read interchangeably.

2.2. Uncertainty measures

When performing active learning, one needs to consider so-called acquisition functions, which help to obtain the points of interest for annotations. The natural choice for acquisition function is some form of uncertainty. There are two main types of uncertainty considered in machine learning – aleatoric and epistemic [9]. The former represents the expected amount of noise in the data, and the latter the lack of model knowledge.

For a deep neural network trained with training data $\mathcal{D}$, the prediction $\hat{y}$ of a new data point $\hat{\mathbf{x}}$ and model parameters θ, total uncertainty can be computed as $\texttt{Unc}_t = \mathcal{H}[\mathbb{E}_{p(\theta|\hat{\mathbf{x}},\mathcal{D})} p(\hat{y}|\theta,\hat{\mathbf{x}},\mathcal{D})]$, and data uncertainty as $\texttt{Unc}_d = \mathbb{E}_{p(\theta|\hat{\mathbf{x}},\mathcal{D})} \mathcal{H}\left[p(\hat{y}|\theta,\hat{\mathbf{x}},\mathcal{D})\right]$. In our notation we denote $\mathcal{H}$ an the entropy, and $p(\hat{y}|\theta,\hat{\mathbf{x}},\mathcal{D})$ – predictive posterior for $\hat{\mathbf{x}}$ given the model parameters θ and training dataset $\mathcal{D}$. Our method relies on computing posterior samples with the Monte-Carlo Dropout (MCD) approach [10].

Epistemic uncertainty is often represented by mutual information (MI) between the model prediction and the posterior distribution over the model parameters [9], but can be estimated using the total and aleatoric uncertainty, that is

$$\mathcal{MI}(\hat{y}; f|\hat{\mathbf{x}},\mathcal{D}) = \texttt{Unc}_t - \texttt{Unc}_d. \tag{2}$$

We finally note here that in practice it is popular to use *Least Confidence* (LC) [11, 12] and *Entropy* (Ent) of the output as total uncertainty measures. Here, it is assumed that $p(\theta|\hat{\mathbf{x}},\mathcal{D})$ is simply a point estimate obtained with a standard training technique.

2.3. Siamese network

For training DAL, we use a Siamese network from Nguyen *et al*. [7]. This architecture's backbone is a VGG-like neural network that processes a pair of knee patches in parallel as depicted in Figure 2. Once the representations of the patches are obtained, we concatenate and pass them through a classifier to predict the KL grade of the given knee.

2.4. Regularized active learning

Our pipeline is presented in Figure 3. First, we create a labeled set $\mathcal{D}_L$ by randomly sampling n_{init} data points from the full dataset $\mathcal{D}$, obtain an unlabeled set $\mathcal{D}_U = \mathcal{D} \setminus \mathcal{D}_L$, and initialize an empty set U for later use. We then run an AL algorithm that iteratively learns to identify (i) the least uncertain samples U with respect to the model to be added to the SSL training process, and (ii) the hardest samples L to be annotated for improving the model's performance on the main task.

At each iteration, we train the model using the SSL algorithm with the CR on the set $\mathcal{D}_L \cup U$. While we use a cross-entropy loss for predictions of $\mathcal{D}_L$'s samples, we apply the consistency regularizer on representations of samples from $\mathcal{D}_L \cup U$ using (1) as in [8]. Subsequently, we utilize the trained model to estimate the uncertainties of unlabeled samples in $\mathcal{D}_U$.

Since our model has the least knowledge on samples with the highest uncertainties, we query labels of the top-n_b and add them to $\mathcal{D}_L$ for the next iteration. Simultaneously, we generate a set U of n_{SS} samples with the lowest uncertainties as the unlabeled set of the SSL algorithm. We continue this cycle until a pre-defined budget N of labeled samples is reached.

Table 1. Train and test dataset details.

Split	Acquisition Site	# Knees	KL0	KL1	KL2	KL3	KL4
Train	A	1397	483	239	438	194	43
	B	1883	765	336	445	272	64
	C	2535	1065	450	632	307	81
	D	2235	841	432	572	323	67
Test	E	903	294	140	287	143	39

3. EXPERIMENTS

3.1. Data

We used plain radiographs from the Osteoarthritis Initiative (OAI) cohort (`https://nda.nih.gov/oai`). The dataset consists of 4796 patients who were examined at a *baseline* (first visit) and follow-up visits. In this work, we only used data from the baseline. Among the 5 acquisition sites of the cohort, we allocated data from sites A, B, C, and D for training (8050 samples), and the data from site E for independent evaluation (953 samples). The KL grade distribution for the whole baseline is shown in Table 1. We followed pre-processing steps in [4, 7] to extract knee patches from bilateral knee radiographs.

3.2. Implementation details

We implemented our codebase in PyTorch and ran all our experiments on NVidia V100 GPUs. All the models were trained for 500 epochs with a batch size of 32. Following [7], we used the Adam optimizer with a learning rate of $1e{-}3$, β values of $(0.9, 0.99)$, and without weight decay. During training, we dropped the learning rate by a factor of 10 at 300^{th} and 400^{th} epochs.

Regarding the architecture, we used the horizontal-vertical pooling, based on the reported empirical results in [7]. For the DAL setting, we set $n_{init} = 50$, $n_b = 50$, and $N = 800$ (approximately 10% of full baseline data). For SSL, we dynamically let n_{SS} be the same as $|X_L|$ in each iteration. Besides the random query as a baseline method, we conducted experiments on four other uncertainty estimation algorithms – namely entropy [13, 9], least confidence [11, 12], and mutual information from MCD (MI-MCD) [13, 9, 10] and entropy from MCD (ENT-MCD). In MI-MCD and ENT-MCD, we used 50 forward passes of MCD.

In the data preparation step, we split the data into a training set (80%) and a validation set (20%) with a stratification by patient. Once models were trained, we independently evaluated them on data from the site E for final reports.

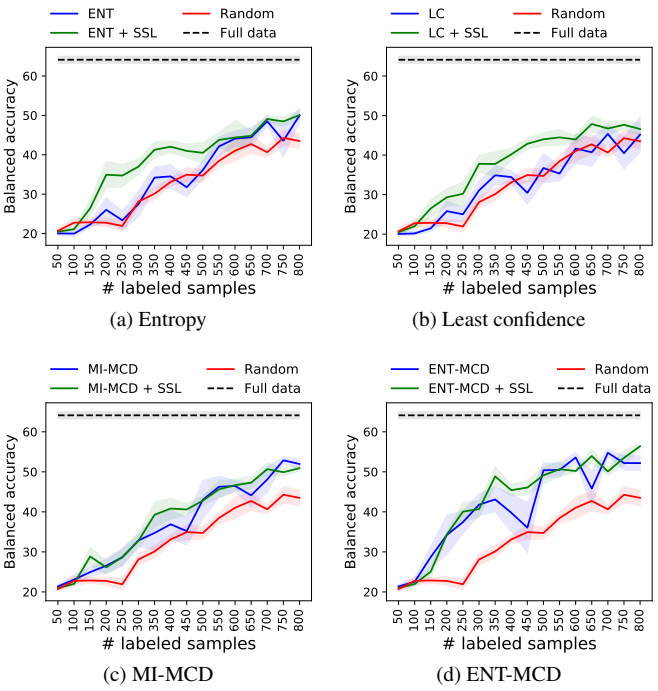
Since the data were imbalanced as shown in Table 1, we evaluated models and performed model selection based on balanced accuracy (BA). We trained each setting 5 times with different random seeds and reported average BAs and standard errors.

Table 2. Effect of semi-supervised learning (SSL) on balanced accuracy (BA) for different query methods. Model trained on the full dataset achieved $64.13_{\pm 0.88}$ BA. ENT-MCD indicates the entropy calculated from MCD samples. MI-MCD indicates mutual information between model posterior and data target distributions (epistemic uncertainty) computed from MCD [10] samples.

Query	SSL	50	250	500	800
ENT-MCD	✓	$20.96_{\pm 0.22}$	$40.06_{\pm 2.34}$	$49.14_{\pm 2.07}$	$56.39_{\pm 0.59}$
	✗	$21.37_{\pm 0.52}$	$37.43_{\pm 2.83}$	$50.39_{\pm 1.85}$	$52.18_{\pm 1.83}$
MI-MCD	✓	$20.96_{\pm 0.22}$	$28.68_{\pm 2.03}$	$42.81_{\pm 1.08}$	$50.91_{\pm 0.73}$
	✗	$21.37_{\pm 0.52}$	$28.66_{\pm 2.13}$	$43.02_{\pm 4.81}$	$51.97_{\pm 1.35}$
LC	✓	$20.49_{\pm 0.29}$	$30.19_{\pm 2.97}$	$43.99_{\pm 1.73}$	$46.59_{\pm 1.55}$
	✗	$20.04_{\pm 0.59}$	$25.01_{\pm 1.77}$	$36.73_{\pm 3.68}$	$45.17_{\pm 4.45}$
ENT	✓	$20.49_{\pm 0.29}$	$34.72_{\pm 2.76}$	$40.49_{\pm 2.66}$	$50.07_{\pm 1.08}$
	✗	$20.04_{\pm 0.59}$	$23.37_{\pm 2.38}$	$36.12_{\pm 2.26}$	$49.97_{\pm 1.77}$
Random	✗	$20.71_{\pm 0.67}$	$21.92_{\pm 1.39}$	$34.73_{\pm 1.54}$	$43.5_{\pm 1.93}$

3.3. Results

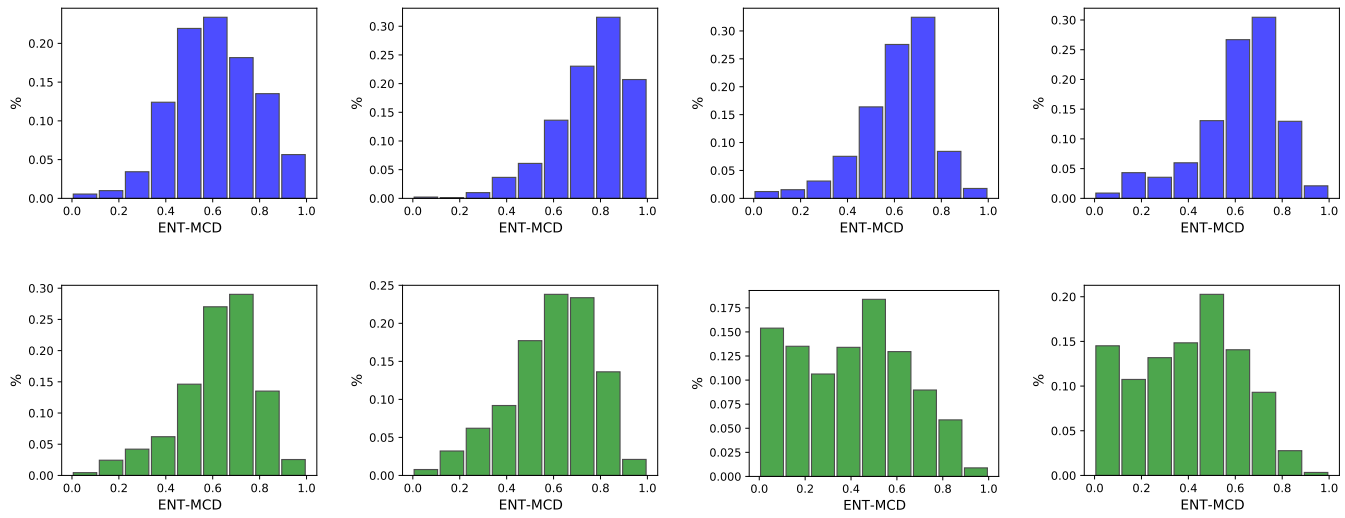
We present our detailed and graphical results in Table 2 and Figure 4, respectively. Here, we considered different



(a) Entropy (b) Least confidence

(c) MI-MCD (d) ENT-MCD

Fig. 4. Added value of CR for different active learning query strategies: (a) Entropy of network's probabilities, and (b) Least Confidence, (c) epistemic uncertainty (MI) from MCD, (d) total uncertainty (entropy) from MCD. All sub-figures show comparison between random sampling (red), naïve query sampling (blue) and query-based sampling with CR and black dashed line indicates the BA with full baseline.



(a) Iteration 1 (b) Iteration 4 (c) Iteration 8 (d) Iteration 15

Fig. 5. Effect of CR on ENT-MCD uncertainty distribution of the test set - without (top row) and with (bottom row) SSL training.

acquisition functions representing total uncertainty and epistemic uncertainty computed from MCD (denoted as MI-MCD). When training the methods with the CR, we found that ENT-MCD (total uncertainty) outperformed the others in most cases.

Naïve Entropy and LC based sampling techniques have shown improvements with CR. For MI-MCD and ENT-MCD based sampling techniques, the model outperforms random baseline with and without SSL, but SSL makes the performance curve more smooth and stable (e.g. there are some sudden big performance drops for without SSL cases which are resolved with SSL).

In other words, after enforcing the CR over $\mathcal{D}_L \cup U$ in each iteration, we observe that all the uncertainty-based methods were stabilized, which results in substantial improvements from the random query baseline once at least 300 labeled samples were added to the data pool.

We visualize the effect of CR on uncertainty distribution in Figure 5 over DAL iterations on the independent test set. Specifically, we show the ENT-MCD distributions from the 1^{st}, 4^{th}, 8^{th}, and 15^{th} iterations. As the DAL iterations progress, the total uncertainty distribution did not shift towards zero for models without SSL, i.e. the amount of low-uncertainty samples did not increase over time. However, for models with SSL, we observed the opposite.

4. DISCUSSION

We studied DAL for KOA severity prediction and empirically showed that CR from SSL stabilized and enhanced the performance of DAL and improved the quality of uncertainty measure. This facilitates the use of AL in practical applications of DL to knee OA grading.

Though our results show clear benefit of SSL for DAL, some limitations of this study should be discussed. First, we selected the initial data randomly. Although realistic, random selection does not guarantee good data distribution. In such a scenario, it is possible that the selected data will be highly biased, there could be zero sample for a certain KL grade. That may result in poorly-trained models and low-quality uncertainty estimates, negatively affecting the next iterations. This could ripple through all the iterations of DAL. We think future studies should focus on optimizing initial data subset selection. Second, DAL is expensive in terms of computational power even for 8050 data points. We believe, future integration of transfer learning by Self-Supervised learning with Semi-Supervised DAL can help the model learn faster. Finally, the use of MCD could potentially result in mode collapse which in turn will result in poor uncertainty estimation. Combating mode collapse could help gain even better performance with MI-MCD and ENT-MCD queries. We believe future studies should consider using deep ensemble [14] or deep ensemble of MCD [15] to fight this limitation.

To conclude, we have shown that SSL improves DAL in the context of knee OA grading. In general, our results indicate that DAL may be beneficial for cost-effective DL in medical image analysis or interpretation, and we call for more studies investigating such approaches.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by Osteoarthritis Initiative (`https://nda.nih.gov/oai`). A

new ethical approval was not required, as the ethical approval and informed consent of the patients were obtained by OAI and published under open access permission group.

6. ACKNOWLEDGEMENTS

This study is supported by the internal funds of the University of Oulu, Finland and the Finnish Center for Artificial Intelligence (FCAI). CSC – IT Center for Science is acknowledged for providing generous computational resources. Egor Panfilov and Terence McSweeney are acknowledged for proofreading. The authors declare no conflict of interest.

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

7. REFERENCES

[1] Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[2] Jonas H Kellgren and JS1006995 Lawrence, "Radiological assessment of osteo-arthrosis," *Annals of the rheumatic diseases*, vol. 16, no. 4, pp. 494, 1957.

[3] Celeste M Torio and Brian J Moore, *National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013: Statistical Brief #204*, Agency for Healthcare Research and Quality (US), Rockville (MD), 2006.

[4] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[5] Aleksei Tiulpin and Simo Saarakkala, "Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks," *Diagnostics*, vol. 10, no. 11, pp. 932, 2020.

[6] Kevin Leung, Bofei Zhang, Jimin Tan, Yiqiu Shen, Krzysztof J Geras, James S Babb, Kyunghyun Cho, Gregory Chang, and Cem M Deniz, "Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative," *Radiology*, vol. 296, no. 3, pp. 584–593, 2020.

[7] Huy Hoang Nguyen, Simo Saarakkala, Matthew B Blaschko, and Aleksei Tiulpin, "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.

[8] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[9] Yarin Gal, "Uncertainty in deep learning," 2016.

[10] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[11] Burr Settles, "Active learning literature survey," 2009.

[12] Dan Wang and Yi Shang, "A new active labeling method for deep learning," in *2014 International joint conference on neural networks (IJCNN)*. IEEE, 2014, pp. 112–119.

[13] Claude Elwood Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *arXiv preprint arXiv:1612.01474*, 2016.

[15] Remus Pop and Patric Fulop, "Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles," *arXiv preprint arXiv:1811.03897*, 2018.