QUANTIFYING UNCERTAINTY IN KNEE OSTEOARTHRITIS DIAGNOSIS

Mame Diarra FALL

QUANTIFYING UNCERTAINTY IN KNEE OSTEOARTHRITIS DIAGNOSIS

Mame Diarra Fall †

† Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS, 45100 Orléans, France.

ABSTRACT

Knee OsteoArthritis (OA) is one of the most common causes of physical disability in the world, causing a large personal and socio-economic burden. Visual assessment of OA still suffers from subjectivity. Deep learning (DL), and in particular convolutional neural networks (CNN), has recently led to remarkable improvements in knee OA detection. However, traditional deep learning-based knee OA classification algorithms lack the ability to quantify decision uncertainty. This is a key point in the medical field where, due to the high cost of labelling, we are faced with a lack of sufficient data to train a learning model. We propose here an alternative approach based on the the concept of *Evidential Deep Learning* (EDL). Unlike Bayesian neural networks which indirectly infer prediction uncertainty through uncertainties in the network weights, EDL approaches explicitly model this uncertainty using the theory of subjective logic. Experimental results on the Osteoarthritis (OAI) database demonstrate the potential of the proposed approach.

Index Terms— Knee OsteoArthritis, Deep-learning, Uncertainty, X-ray images, classification.

1. INTRODUCTION

Knee OsteoArthritis is a degenerative disease characterized by deterioration and damage of the articular cartilage, joint edges, and reactive hyperplasia of the subchondral bone [1]. Multiple factors, including age, weight, stress, trauma, etc. may contribute to its occurrence [2]. The disease is associated with stiffness, swelling, and pain. Knee OA is recognized as the main cause of reduced mobility in the elderly, and is now recognized as an independent risk factor for increased mortality. Since no treatment can prevent the degenerative structural changes responsible for the progression of knee osteoarthritis, early detection is essential so that timely behavioural therapies, such as weight loss, can be implemented to delay the onset and progression of knee OA [3].

The Kellgren and Lawrence (KL) grading system [4] defines knee OA severity in five grades, from 0 (normal) to 4 (severe), according to the existence and severity of symptoms such as osteophytes and joint space narrowing (see Fig. 1). This criterion is however semi-quantitative, and suffers from subjectivity and ambiguity, which makes early OA diagnosis very challenging.

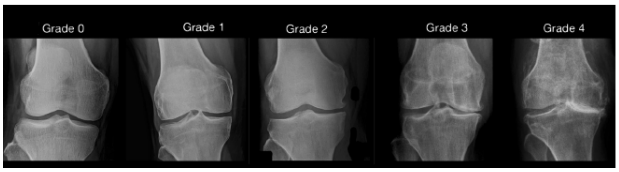


Fig. 1. Radiological grades of OA according to the KL scale

It is therefore necessary to develop automated methods to facilitate the knee OA diagnosis. Several deep learning (DL) have been proposed to this aim. Most of them are based on convolutional neural networks (CNN). The latter have become state-of-the-art technology for image classification tasks due to their ability to capture spatial information in images. One can refer to [5] for a brief review of current CNN-based methods for knee OA. Nevertheless, one major drawback of CNN models is their dependency on large amounts of labelled data. In the field of medical imaging the small quantity of valid medical datasets, due to the high cost of labelling, still poses a major challenge for training the learning model. In this context it is crucial to quantify the *uncertainty* associated with the model prediction.

In classical CNN for classification, the softmax function is used to predict class assignment probabilities. A softmax output is however often misinterpreted as an indication of model confidence. Nevertheless a model can be uncertain in its predictions even with a high softmax output for a particular class [6]. Bayesian neural networks can address this issue by putting a prior distribution on the neural network parameters and inferring their posterior distribution using approximations such as Variational Bayes [6]. The posterior predictive distribution is approximated using Monte-Carlo sampling methods. These models are however computationally demanding since each data point has to pass through the network at several times, and also require in addition considerable modifications to existing baselines.

Recently, a new class of models based on the concept of *evidential deep-learning* (EDL) has been proposed [7], [8]. They allow to compute uncertainty in a single forward pass by parameterizing distributions on distributions, and involve minimal modifications to the architecture of standard neural networks.

Is this paper, we develop an EDL-based method for knee OA diagnosis. The potential of the proposed approach is eval-

uated for binary and multi-class classification tasks. To the best of our knowledge, this is the first work based on deep neural networks that allows to quantify uncertainty in the diagnosis of knee OA.

The rest of the paper is organized as follows. In Section 2 the EDL theory is briefly reviewed before presenting our model. Experimental setups are described in Section 3. Results are presented in Section 4. We conclude in Section 5 and point some future research directions.

2. METHODS

2.1. Theory of Evidence and uncertainty

Let $\mathbf{x} \in \mathbb{R}^d$ be an input image (in vectorized form) and $f(\cdot|\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}^K$ a NN with parameters $\boldsymbol{\theta}$, where $K \geq 2$ stands for the number of classes. We denote $\mathbf{a} = f(\mathbf{x}|\boldsymbol{\theta}) \in \mathbb{R}^K$ the raw output vector of the NN corresponding to the input $\mathbf{x}$. The output is $\hat{\mathbf{y}} = \Phi(\mathbf{a})$. A schematic representation of such a network is depicted in Fig. 2. In classical NN for classification, Φ is given by the softmax function.

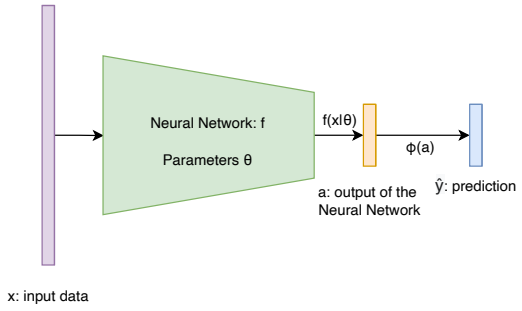


Fig. 2. Schematic representation of the NN.

The training dataset consists of $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, N_{\mathcal{D}}\}$ where $\mathbf{y}_i$ is the ground-truth class associated to $\mathbf{x}_i$, in the form of a one-hot encoded vector. The idea behind EDL is to train the network to give an opinion on a classification problem [7]. This originates from the Dempster-Shafer Theory of Evidence (DST), a generalization of the Bayesian theory to subjective probabilities. Using the DST theory, we assign a belief mass $b_k \geq 0$ to each class k and consider $u \geq 0$ an overall uncertainty mass such that

$$u + \sum_{k=1}^{K} b_k = 1.$$

A belief mass can be computed using the evidence. If we let $e_k \geq 0$ be the evidence derived for the k-th class, then

$$b_k = \frac{e_k}{S} \quad \text{and} \quad u = \frac{K}{S},$$

where $S = \sum_{k=1}^{K}(e_k + 1)$. The uncertainty is inversely proportional to the total evidence; and if there is no evidence, the belief for each class is zero and the uncertainty u is one.

Using subjective logic, we can formalize the DST as a Dirichlet distribution, quantifying belief masses and uncertainty [9].

2.2. Dirichlet Modeling

The Dirichlet distribution is a distribution over probability mass functions. A random vector $\mathbf{P} = (P_1, \ldots, P_K)$ is distributed according to the Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ if its probability density function is

$$f(\mathbf{p}|\boldsymbol{\alpha}) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \quad \text{where} \quad \beta(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(S_\alpha)}$$

is the K-dimensional multinomial beta function, $S_\alpha = \sum_{k=1}^{K} \alpha_k$ is the Dirichlet strength, with $\alpha_k > 0 \, \forall k$.

Since classical CNN using sotfmax only provide a point estimate of the class probabilities for a given sample $\mathbf{x}$, one solution is to use a Dirichlet distribution to model the probability distribution of these class probabilities. To this end, one should define a relationship between the NN output $\mathbf{a}$ and the Dirichlet parameters in the form $\boldsymbol{\alpha} = \Phi(\mathbf{a})$, where Φ must comply with the strictly positive constraint of $\boldsymbol{\alpha}$. We consider $\Phi(\mathbf{a}) = \mathrm{ReLU}(\mathbf{a}) + 1$ which satisfies both conditions. To any input data $\mathbf{x}_i$ we associate the random variable

$$\mathbf{P}_i = (P_{i1}, ..., P_{iK}) \sim \mathrm{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i), \tag{1}$$

where $\boldsymbol{\alpha}_i = \Phi(\mathbf{a}_i) = \mathrm{ReLU}(f(\mathbf{x}_i|\boldsymbol{\theta})) + 1$. The class prediction $\hat{\mathbf{y}}_i$ is given by the mean of the Dirichlet distribution $\mathrm{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i)$, that is,

$$\hat{\mathbf{y}}_i = \frac{\boldsymbol{\alpha}_i}{S_{\boldsymbol{\alpha}_i}} = \left(\frac{\alpha_{i1}}{S_{\boldsymbol{\alpha}_i}}, \ldots, \frac{\alpha_{iK}}{S_{\boldsymbol{\alpha}_i}} \right).$$

We used a similar idea in [10] to estimate proportions in a mixture for X-Ray diffraction and hyperspectral imaging. The problem considered in this paper is different, since it is a classification problem for which we want to estimate the uncertainty at the same time.

2.3. Loss function

To train the neural network, we must define a loss function using the Dirichlet model (1). As in [10], we consider minimizing the mean square error. That is for any input $\mathbf{x_i}$, minimize the expectation of the squares of the errors between $\mathbf{y_i}$ and $\mathbf{P}_i$,

$$\mathcal{L}_i^{SE}(\boldsymbol{\theta}) = \mathbb{E}(\| \mathbf{y}_i - \mathbf{P}_i \|^2) = \| \mathbf{y}_i - \hat{\mathbf{y}}_i \|^2 + \mathrm{Var}(\mathbf{P}_i), \tag{2}$$

where $\mathrm{Var}(\mathbf{P}_i) = \sum_{j=1}^{K} \mathrm{Var}(P_{ij})$. This loss aims to achieve the joint goal of minimizing the prediction error and the variance of the Dirichlet distribution output by the NN.

In addition, we also need to ensure that the NN has the expected behaviour under foreign data inputs, by reinforcing for example the flatness of the Dirichlet distribution. We consider a regularization term in the loss function, consisting of the KL divergence with respect to the uniform Dirichlet distribution [7]. The overall loss considered is then,

$$\mathcal{L}_i(\boldsymbol{\theta}) = \mathcal{L}_i^{SE}(\boldsymbol{\theta}) + \lambda_t \mathrm{KL}\left[\mathrm{Dir}(\mathbf{p}_i|\tilde{\boldsymbol{\alpha}}_i)||\mathrm{Dir}(\mathbf{p}_i|\mathbf{1})\right], \tag{3}$$

where $\mathrm{Dir}(\mathbf{p}_i|\mathbf{1})$ refers to the uniform Dirichlet distribution, $\tilde{\boldsymbol{\alpha}}_i = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i$ (with $\odot$ referring to the element-wise product) are the Dirichlet parameters after removing the non-misleading evidence from predicted parameters $\boldsymbol{\alpha}_i$ for sample i, $\lambda_t = \min(1, t/k)$ is an annealing coefficient, t is the index of the current training epoch and k is a fixed number (often set to $k = 10$).
The KL divergence in the loss (3) exhibits a closed-form,

$$\mathrm{KL}\left[\mathrm{Dir}(\mathbf{p}_i|\tilde{\boldsymbol{\alpha}}_i)||\mathrm{Dir}(\mathbf{p}_i|\mathbf{1})\right] = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\tilde{\alpha}_{ik}\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{ik})}\right) + \sum_{k=1}^{K}(\tilde{\alpha}_{ik}-1)\left[\psi(\tilde{\alpha}_{ik}) - \psi\left(\sum_{k=1}^{K}\tilde{\alpha}_{ik}\right)\right],$$

where $\psi(\cdot)$ is the digamma function.

3. EXPERIMENTAL SETUP

We evaluate our method on the publicly available Osteoarthritis Initiative (OAI) data set. This cohort recruited 4796 participants with age ranging from 45 to 79. As the OAI is a multi-center study, the physical resolution and dimension of the knee X-ray images collected from the baseline cohort are not homogeneous. Pre-processing is required to ensure that all images have the same size. We use the 4130 X-ray images with 8260 knee joints from [11]. The data distribution is represented in Fig. 3 below.

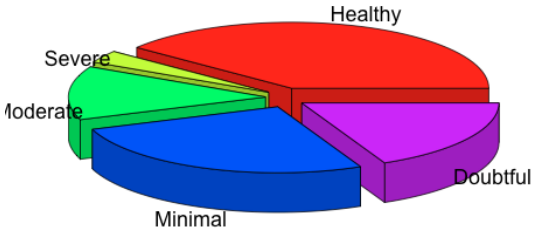


Fig. 3. Dataset distribution

3.1. Dataset preparation

The dataset of X-ray images of knee joints is not suitable in terms of clarity and localization as input for the DL models. It is therefore necessary to perform a data pre-processing step, during which the images are transformed to clearly capture the joint area where OA information is likely to exist. This involves first cropping the image to the desired region of the knee, so that all undesirable regions are excluded. To achieve this, we cropped the images by 60 pixels top and bottom, resulting in images of size $224 \times 104 \times 3$. The second step was to perform histogram equalization to improve contrast and ensure good visibility of the desired areas.

We randomly assigned images to the training, validation, and test sets with respectively the following percentages 64%, 16% and 20%. All data were normalized, and we also used a data augmentation scheme in the training phase using a horizontal random flip with a probability of 0.5.

3.2. Training phase

We conduct the experiments with the ResNet-101 architecture [12]. The standard softmax layer is replaced by a ReLU whose output is used as an evidence vector for the Dirichlet distribution. The network is trained from scratch for 200 epochs in experiments 1 (described below), and 500 in experiments 2 and 3. We choose Adam as the optimizer, with an initial learning rate of 10^{-3} and a weight decay of 10^{-4} to avoid overfitting. We also choose a learning rate schedule, decreasing the learning rate by 10% every 20 epochs.

3.3. Performance metrics

In our experiments, quantitative evaluation was performed using four different metrics:

1. Accuracy: percentage of predictions that match exactly the ground-truth.
2. Precision: fraction of true positives among predicted positives.
3. Recall: fraction of the total number of true positives retrieved.
4. F1-score: harmonic mean of precision and recall.

We now present the results on different test sets derived from the original data.

4. NUMERICAL RESULTS

In this section, the performance of the proposed method is assessed on binary and multi-class classification tasks.

Binary classification

First, the model has been trained to detect two classes. We consider two classification scenarios .

- **Experiment 1.** The aim of this experiment is to detect knee OA at an early stage. We therefore classify normal patients (KL0) versus those with mild OA (KL2). The test database includes 1086 images.

- **Experiment 2.** This aims at detecting the presence/absence of knee OA. We create a binary dataset by combining classes KL0 and KL1 to represent negative diagnosis of KOA (denoted as "Normal"); KL2, KL3, and KL4 are combined to represent positive diagnosis ("Abnormal"). There are 1656 images in the test database.

Multi-class classification

- **Experiment 3.** The aim of this third experiment is to determine the severity of KOA. We therefore remove classes KL0 and KL1 and consider classes KL2, KL3, and KL4. The number of images in the test database is 721.

Table 1 summarizes the results of the three classification experiments on the test set. Note that the table can be misleading for our approach, since totally uncertain predictions (i.e. $u = 1$) are also considered failures when computing overall accuracy. Despite this, the performance results are good, and quite similar for Experiments 1 and 2. This is promising, as experiment 1 is challenging due to the similarity between KL0 and KL2 images. This also shows the model's ability for early detection. The results of experiment 3 are slightly less good. Confusion matrices (not shown here) indicate large misclassifications of knee joints categorized as KL3 (moderate) and predicted as KL2 (minimal). These images show minimal variations in terms of joint space width and osteophytes formation, making them challenging to distinguish. Moreover, due to the very unbalanced data distribution (see Fig. 3), we have much less data for these classes, making the classification task even more difficult.

Experiment	Accuracy	F1score	Precision	Recall
1	0.72	0.71	0.72	0.70
2	0.73	0.73	0.73	0.73
3	0.66	0.64	0.62	0.66

Table 1. Model performance on the three experiments.

Uncertainty quantification

The strength of our approach lies in its ability to quantify uncertainty. We first examine the average uncertainty for right and wrong predictions. The results are presented in Table 2 for the three experiments. Our model associates higher uncertainty to erroneous predictions, in all the three data sets. Uncertainty is also higher for experiment 3, for which the classification results are less good.

Experiment	Right prediction	Wrong prediction
1	0.14	0.23
2	0.18	0.34
3	0.43	0.59

Table 2. Average uncertainty for right and wrong predictions.

It may be useful to look at the images and the results obtained when the classification results matches the groundtruth. Fig. 4 shows that the proposed method in this case produces high probability and low uncertainty, indicating a confident classification.

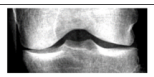
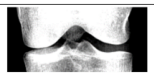
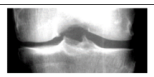
Experiment	Test image	Actual Class	Predicted Class	Probability	Uncertainty
1		Healthy	Healthy	0.93	0.12
2		Normal	Normal	0.90	0.19
3		Minimal	Minimal	0.84	0.24

Fig. 4. Results with correctly classified data.

Finally, we examine the results obtained in the case of misclassification. Fig. 5 displays some of these results. In the case of a severe misclassification in Experiment 3, for example with a KL4 (severe) grade predicted as KL2 (minimal), the model produces a high uncertainty ($u = 0.57$).

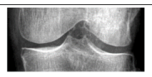
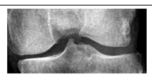
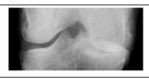
Experiment	Test image	Actual Class	Predicted Class	Probability	Uncertainty
1		Minimal	Healthy	0.53	0.24
2		Abnormal	Normal	0.69	0.30
3		Severe	Minimal	0.65	0.57

Fig. 5. Results with misclassified data

5. CONCLUSION AND PERSPECTIVES

In this paper, we have proposed a new classification method for knee OA diagnosis. These preliminary results are promising since classification task can be challenging particularly when dealing with complex data such as X-ray images used for the knee OA diagnosis. Being able to quantify uncertainty is a key point of the proposed method

There are, however, some limitations. The dataset was relatively small and highly unbalanced. Another limitation was the lack of clean classes due to the high similarity of X-ray images particularly in early stages of knee OA. Future work will be devoted to explore other kind of data augmentation [13], and combining texture and shape information [14].

Acknowledgement

This work was supported by the French National Agency of Research (ANR) through the MIMOSA project (ANR-20-CE45-0013) and the BACKUP project (ANR-23-CE40-0018-01).

6. REFERENCES

[1] R. Loeser, S. Goldring, S. C.R., and G. M.B., "Osteoarthritis: a disease of the joint as an organ," *Arthritis and Rheumatism*, vol. 64, no. 6, pp. 1697–1707, 2012.

[2] A. Litwic, M. H. Edwards, E. M. Dennison, and C. Cooper, "Epidemiology and burden of osteoarthritis," *British medical bulletin*, vol. 105, no. 1, p. 185–199, 2013.

[3] A. E. Wluka, C. B. Lombard, and F. M. Cicuttini, "Tackling obesity in knee osteoarthritis," *Nature Reviews Rheumatology*, vol. 9, no. 4, p. 225–235., 2013.

[4] J. Kellgren and J. Lawrence, "Radiological assessment of osteo-arthrosis." *Annals of the rheumatic diseases*, vol. 16, no. 4, pp. 494–502, 1957.

[5] Y. Nasser, R. Jennane, A. Chetouani, E. Lespessailles, and M. E. Hassouni, "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.

[6] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059.

[7] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3183–3193.

[8] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7047–7058.

[9] A. Jsang, *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[10] T. Simonnet, M. D. Fall, B. Galerne, F. Claret, and S. Grangeon, "Proportion inference using deep neural networks. applications to x-ray diffraction and hyperspectral imaging," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 1310–1314.

[11] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "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.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Z. Wang, A. Chetouani, and R. Jennane, "Key-exchange convolutional auto-encoder for data augmentation in early knee osteoarthritis classification," *arXiv preprint arXiv:2302.13336*, 2023.

[14] Y. Nasser, M. El Hassouni, D. Hans, and R. Jennane, "A discriminative shape-texture convolutional neural network for early diagnosis of knee osteoarthritis from x-ray images," *Physical and Engineering Sciences in Medicine*, pp. 1–11, 2023.