Curriculum learning for improved femur fracture classification: scheduling data with prior knowledge and uncertainty

Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A. González Ballester, Gemma Piella

***Abstract*— An adequate classification of proximal femur fractures from X-ray images is crucial for the treatment choice and the patients' clinical outcome. We rely on the commonly used AO system, which describes a hierarchical knowledge tree classifying the images into types and subtypes according to the fracture's location and complexity. In this paper, we propose a method for the automatic classification of proximal femur fractures into 3 and 7 AO classes based on a Convolutional Neural Network (CNN). As it is known, CNNs need large and representative datasets with reliable labels, which are hard to collect for the application at hand. In this paper, we design a curriculum learning (CL) approach that improves over the basic CNNs performance under such conditions. Our novel formulation reunites three curriculum strategies: individually weighting training samples, reordering the training set, and sampling subsets of data. The core of these strategies is a scoring function ranking the training samples. We define two novel scoring functions: one from domain-specific prior knowledge and an original self-paced uncertainty score. We perform experiments on a clinical dataset of proximal femur radiographs. The curriculum improves proximal femur fracture classification up to the performance of experienced trauma surgeons. The best curriculum method reorders the training set based on prior knowledge resulting into a classification improvement of 15%. Using the publicly available MNIST dataset, we further discuss and demonstrate the benefits of our unified CL formulation for three controlled and challenging digit recognition scenarios: with limited amounts of data, under class-imbalance, and in the presence of label noise. The code of our work is available at: https://github.com/ameliajimenez/curriculum-learning-prior-uncertainty.**

***Index Terms*— curriculum learning, self-paced learning, data scheduler, bone fracture, x-ray, multi-class classification, limited data, class-imbalance, noisy labels**

I. Introduction

PROXIMAL femur fractures are a significant cause of morbidity and mortality, giving rise to a notable socioeconomic impact [1], [2]. Elderly population in the western world are especially affected. The incidence of femur fractures increases exponentially from an age of 65 and is almost doubled every five years.

Surgery is the most common and preferred treatment for proximal femur fractures [4]. The exact classification of the fracture is crucial for deciding the surgical procedure and choosing the surgical implant if needed. The Arbeitsgemeinschaft für Osteosynthesefragen (AO-Foundation) has established a hierarchical classification system for fractures of all bones based on radiographs. For proximal femur fractures, the AO classification has been beneficial, in terms of reproducibility, when compared against other systems such as the Jensen classification [5]. The AO standard follows a hierarchy according to the location and configuration of the fracture lines, see Fig. 1. Fractures of type-A are located in the trochanteric region, and fractures of type-B are those affecting the area around the femur neck. Each type of fracture is further divided into 3 subclasses depending on the morphology and number of fragments of the fracture.

The ability to adequately classify fractures according to the AO standard based on radiographs is acquired through daily clinical routine in the trauma surgery department. Several years are needed until experienced trauma surgeons are significantly differentiated from residents. Inter-reader agreement varies from 66% among residents to 71% among experienced trauma surgeons [6]. To reach a precise classification, medical students and young trauma surgeons rely on a second opinion to choose the adequate treatment option for the patient. Our work aims to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713673 and by the Spanish Ministry of Economy [MDM-2015-0502]. A. Jiménez-Sánchez has received financial support through the "la Caixa" Foundation (ID Q5850017D), fellowship code: LCF/BQ/IN17/11620013. D. Mateus has received funding from Nantes Métropole and the European Regional Development, Pays de la Loire, under the Connect Talent scheme. Authors thank Nvidia for the donation of a GPU.

A. JS., G. P., and M. A. G. B. are with BCN MedTech, Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain. (e-mails: amelia.jimenez@upf.edu, gemma.piella@upf.edu, ma.gonzalez@upf.edu). M. A. G. B. is also with ICREA, Barcelona, Spain.

D. M. is with Ecole Centrale de Nantes, LS2N, UMR CNRS 6004, 44321, Nantes, France (e-mail: diana.mateus@ec-nantes.fr).

S. K., C. K., P. B. are with Department of Trauma Surgery, Klinikum rechts der Isar, Technische Universität München, 81675, Munich, Germany (e-mail: sonja.kirchhoff@me.com, dr.kirchhoff@me.com, peter.biberthaler@mri.tum.de). S. K. is also with Institute of Clinical Radiology, LMU München, Munich, Germany.

N. N. is with Computer Aided Medical Procedures, Technische Universität München, 85748, Munich, Germany (e-mail: nassir.navab@tum.de). N. N. is also with Johns Hopkins University, 21218, Baltimore, USA.

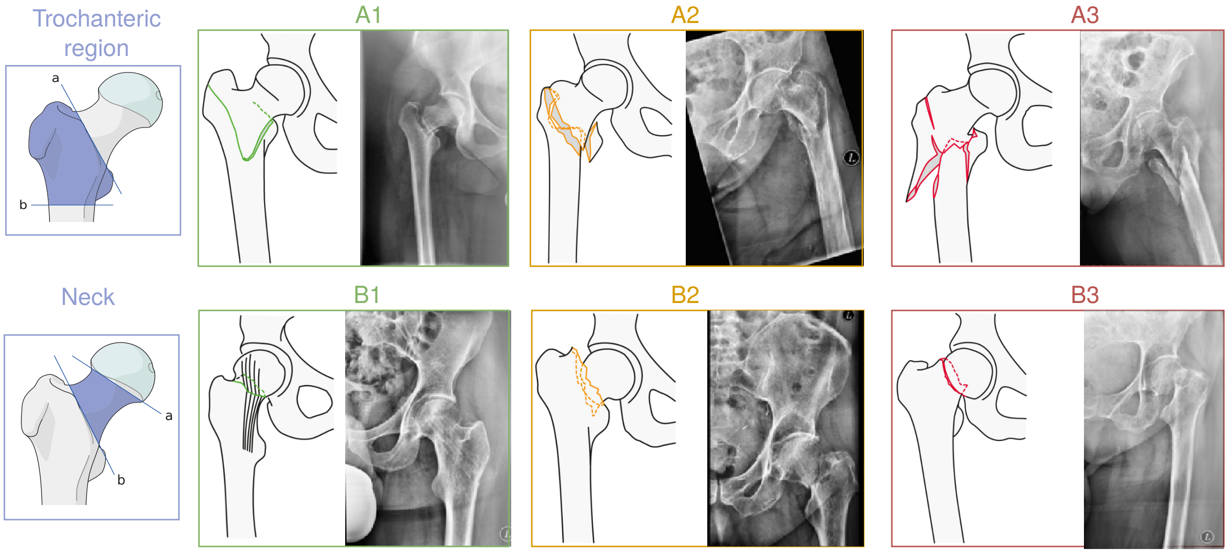


Fig. 1: Examples of proximal femur fractures and their fine-grained AO classification, adapted from [3].

provide support as a computer-aided diagnosis (CAD) system capable of classifying radiographs.

Convolutional neural networks (CNNs) are nowadays the model of predilection for CAD. They have been rapidly integrated in numerous medical applications [7]–[11] due to their strong capacity to learn, directly from data, meaningful and hierarchical image representations. However, their feature extraction ability heavily depends not only on the optimization scheme but also on the training dataset. To be properly trained, CNNs need a large dataset representative of the population of interest [12].

In general, in medical image analysis tasks, acquiring reliable and clinically relevant annotated data remains a key challenge. Apart from the intra- or inter-expert disagreement, typically, manual annotations call for the time and effort of clinical experts. In addition, medical datasets usually suffer from class imbalance due to difficulties in collecting cases and the incidence of rare diseases. Finally, medical image data needs also dealing with proprietary and/or privacy concerns. As a result, these datasets generally exhibit three main challenging characteristics: (i) limited amounts of data, (ii) class-imbalance, and (iii) uncertain annotations.

The most common approaches to alleviate these challenges have been transfer learning [7], [10], [13], [14], data augmentation [15] and semi-supervised learning [14], [16]. More recently, the attention has been shifted towards bootstrapping or weighting strategies [17], sample mining [18], active learning [19], and curriculum learning [20]–[23].

The underlying intuition of strategies such as reordering, sampling or weighting, is that they can significantly impact the optimization of CNNs during training. Towards this objective, we reunite and formulate the above curriculum learning (CL) strategies to improve the performance of fine-grained proximal femur fractures classification, by dealing with the lack of large annotated datasets, class imbalance, and annotation uncertainty. Inspired by the concept of curriculum in human learning, CL presents the training samples to the algorithm in a meaningful order (often by difficulty from "easy" to "hard") and has been shown to avoid bad local minima and lead to an improved generalization [24].

Lately, training CNNs with ordered sequences has been shown to improve medical image segmentation by gradually increasing the context around the areas of interest [25]–[27]. To the best of our knowledge, only few works have explored sample reordering for CAD with CNNS, for instance by extracting prior knowledge from radiology reports [20] or medical guidelines [23].

The ordering can be either fixed (*e.g.* set heuristically by a "teacher" or domain-specific knowledge) or, in the absence of a-priori knowledge, a self-paced order [28] derived from the algorithm's performance (*e.g.* the loss). Our unified CL formulation encompasses both approaches. We address the lack of prior knowledge to design an ad-hoc curriculum, by providing a ranking criterion based on uncertainty modelling. By using uncertainty to define our ranking, the classifier favors samples that it has not yet properly learnt, thus guiding it to explore "unseen" parts of the input space. We present three manners to actually implement the curriculum data sequencing. The first one is based on reordering the training set. The second uses a sampling strategy, *i.e.* selecting increasingly growing subsets. The last one employs a weighting scheme to give different importance to the training samples.

To show the impact of our proposed method, we perform two types of experiments. First, on the challenging problem of multi-class classification of proximal femur fractures. This multi-class problem is inherently imbalanced, as the frequency of the classes reflects their incidence. Moreover, the adequate classification takes several years of daily clinical routine in the trauma surgery department, limiting the collection of annotations and leading to potentially noisy labels. Thus, to deepen the understanding of the method and to verify its effectiveness under these challenging data conditions, we design a series of experiments on the MNIST dataset, controlling the amount of data, class-imbalance, and label noise.

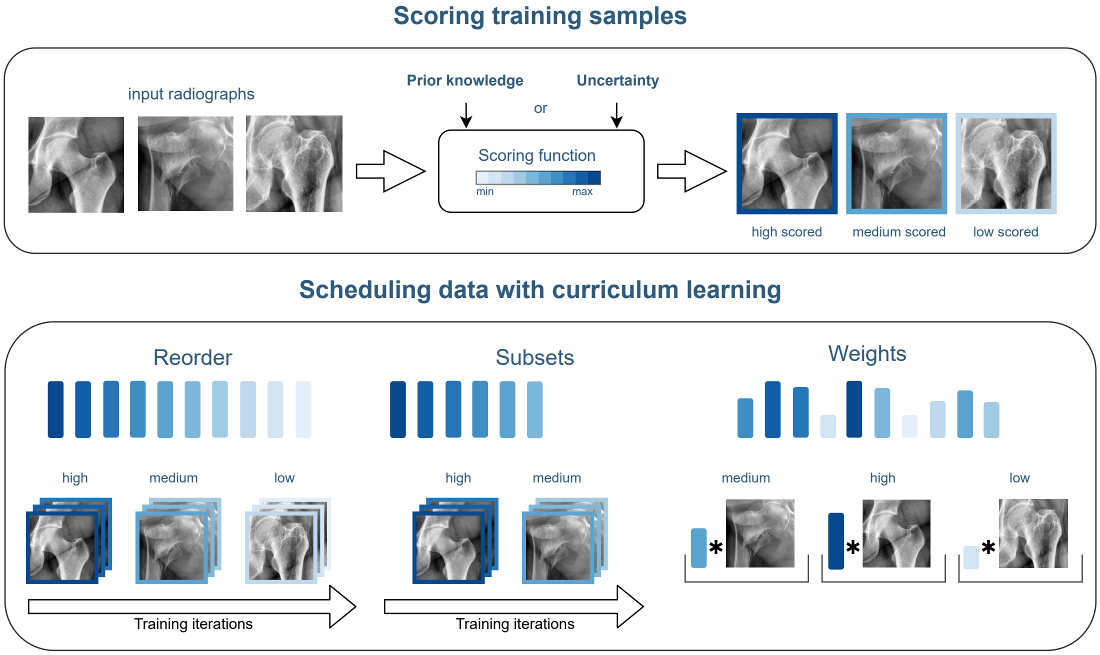


Fig. 2: Training a CNN with CL. Top: training samples are scored according to prior knowledge or uncertainty. Bottom: three CL strategies are presented to automatically schedule the order and pace of the training samples for multi-class classification.

Contributions: In this work, we propose three CL strategies to automatically schedule the order and pace of the training samples for an improved multi-class classification. Our contributions are:

- We identify common curriculum learning elements among different data scheduling strategies, and present them within a unified formulation.
- We propose two types of novel ranking functions guiding the prioritization of the training data.
- We leverage domain-specific clinical knowledge to define the first scoring function.
- In absence of domain knowledge, we propose to estimate the ranking of the training samples by dynamically quantifying the uncertainty of the model predictions.
- We validate our strategies on a clinical dataset for the multi-class classification of proximal femur fractures.
- With a controlled experimental setting, we confirm that our method is useful in reducing the classification error under limited amounts of data, imbalance in the class distribution, and unreliable annotations. We give recommendations about the best approaches for each scenario.

This paper is structured as follows. Section II covers CL related works that are relevant for the design of data schedulers. In Section III, the details of our proposed formulation are presented. Section IV describes the specifications of the experimental validation. Section V shows the classification performance. Section VI discusses our findings, recommendations and future work. Finally, Section VII summarizes our conclusions.

II. Related work

Recently, CL, self-paced learning (SPL), active learning (AL) and selection strategies have been studied to improve CNN-based image classification performance. These methods rely on ranking the training samples according to some criterion. In the following, we highlight some works that employ the two criteria related to our method: (i) domain-specific prior knowledge and (ii) data and model uncertainty.

Prior knowledge is leveraged in [20]–[23] to design a curriculum for classification. Yang *et al.* [21] exploited SPL to handle class-imbalance, by combining the number of samples in each class and the difficulty of the samples, which is derived from the loss. Tang *et al.* [20] proposed to feed the images in order of difficulty based on severity-levels mined from radiology reports to improve the localization and classification of thoracic diseases. Jiménez-Sánchez *et al.* [23] exploited the knowledge of the inconsistencies in the annotations of multiple experts and medical decision trees, to design a medical-based deep curriculum that boosted the classification of proximal femur fractures. Trying to mimic the training of radiologists, Maicas *et al.* [22] proposed to pretrain a CNN model with increasingly difficult tasks, before training for breast screening. The pretraining tasks were selected using teacher-student CL. In this work, we schedule our training data based on a scoring function that ranks the samples according to domain-specific prior knowledge or uncertainty. Different from previous works [20], [22], [23], which only considered reordering the training set, here we investigate two further curriculum strategies, namely, subset sampling and weighting. Furthermore, solely Yang *et al.* [21] targeted one of the mentioned data challenges: class-imbalance, whereas we investigate as well noisy labels and limited amounts of training data.

The second criterion that we consider for defining a curriculum is uncertainty. The estimation of uncertainty provides a way of systematically defining the difficulty of the samples.

Xue *et al.* [18] proposed online sample mining based on uncertainty to handle noisy labels in skin lesion classification. In their work, uncertainty is approximated through the classification loss. However, the most common methods for estimating classification uncertainty, in the context of deep learning, rely on Bayesian estimation theory, namely using Monte-Carlo (MC) dropout [29]. Uncertainty is probably the most frequent criterion in AL selection strategies. Recently, Wu *et al.* [30] combined uncertainty together with image noise into their AL scheme to alleviate medical image annotation efforts. Uncertainty and label correlation are integrated in the sampling process to determine the most informative examples for annotation. AL pays attention to examples near the decision surface to infer their labels. Similarly, we aim to gradually move the classification decision border by adding examples of increasing ranking scores. We prioritize in our second scoring function the most representative samples, letting uncertainty guide their order, pace or weight. Although uncertainty has been used as sampling criterion for AL, we employ this information, for the first time, to rank and define our curriculum.

We validate our proposed curriculum strategies for the classification of proximal femur fractures. Whereas most of the previous work on femur fractures focuses on the binary fracture detection task [31]–[33], we target the more challenging multi-class classification according to the AO standard [23], [34], [35].

Approaches to boost fracture classification accuracy comprise prior localization, transfer learning or medical knowledge. The localization of a region of interest before the classification of the full image has been studied either in a weakly-supervised [33], [34] or in a supervised [35] way. Knowledge transfer has been investigated across image domains, *i.e.* using ImageNet dataset for pretraining [31], [36], and across tasks, *i.e.* training first on body part detection (easier task) and then focusing on the hip fracture detection [32]. Medical knowledge has been proposed to train a hierarchical cascade of classifiers [37] or to schedule training data into a set of increasing difficulty [23]. Tanzi *et al.* [37] relied on a cascade of classifiers. However, this kind of strategies are prone to propagate errors in multi-class classification. Furthermore, our CL approach does not rely on a complicated multistage scheme. We do not introduce any further complexity to the CNN.

In our previous work [23], a series of heuristics, based on knowledge such as medical decision trees and inconsistencies in the annotations of multiple experts, were proposed as a scoring function to boost fracture classification performance. Here, we further propose two more strategies, and also provide an alternative mechanism to rank the samples, based on prediction uncertainty, in case prior knowledge is unavailable.

III. Method

Given a multi-class image classification task, where an image x_i needs to be assigned to a discrete class label $y_i \in \{1, \dots, T\}$, our training set is defined as $X = \{(x_1, y_1), \dots, (x_N, y_N)\}$. Assume a CNN model h with parameters θ is trained with stochastic gradient descent (SGD). During training, samples are typically randomly ordered. Our goal is to instead schedule the order and pace of the training data presented to the optimizer to better exploit the available data and annotations, and thereby improve the classification performance.

To learn the best CNN model h_{θ^*} from the input data, a common choice is to use empirical risk minimization:

$$\begin{aligned} \mathcal{L}(\theta) &= \tilde{\mathbb{E}}[L_\theta] = \frac{1}{N}\sum_{i=1}^{N} L_\theta(x_i, y_i) \\ \theta^* &= \arg\min_\theta \mathcal{L}(\theta) \end{aligned} \tag{1}$$

where $\tilde{\mathbb{E}}$ stands for the empirical expectation, L_θ is the loss function that measures the cost of predicting $h_\theta(x_i)$ when the correct label is y_i.

Optimization is conducted with SGD for a total of E epochs. Typically, the objective function L_θ is non-convex and is minimized in mini-batches of size B. Whereas convex learning is invariant to the order of sample presentation, CNNs are not. In the later case, the loss function usually presents a highly non-convex shape with many local minima, so the order of sample presentation affects learning, and thus, the final solution. It has been empirically shown that the variance in the direction of the gradient step defined by easier examples is significantly smaller than that defined by difficult ones, especially at the beginning of training [38], [39]. This suggests that favoring the easier examples may increase the likelihood to escape the attraction basin of an initial poor local minimum. Taking into account the mini-batches, we can rewrite Eq. (1) as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{j=1}^{N/B}\sum_{k=1}^{B} L_\theta(\hat{x}_{k,j}, \hat{y}_{k,j}), \tag{2}$$

where $\hat{x}_{k,j}$ is the k-th sample in the j-th batch, $\hat{x}_{k,j} = x_{k+(j-1)\cdot B}$, and $\hat{y}_{k,j}$ is the corresponding label.

We propose to modify Eq. (2) to schedule the training data. To do so, first, we formalize two types of scoring functions to assign a priority level to each data sample. The scoring is defined in Subsection III-A either according to domain-specific prior knowledge or to the samples' uncertainty measured with MC dropout. Then, in Subsection III-B, we introduce the different components required for reordering, pacing, and weighting the training data. Fig. 2 provides an schematic illustration of the process for training a CNN with CL. Finally, we cover the implementation details of the three variants of our unified CL formulation in Subsection III-C.

A. Scoring function definition

The key element of our approach is the definition of the scoring function s or, equivalently, the curriculum probabilities p, which corresponds to normalized score function values. The formal definition of the curriculum probabilities is presented in Subsection III-B. The curriculum allow us to sample the dataset and obtain a reordering function π that schedules the training samples. In this subsection, we present two alternative scoring functions. The first one is static and based on some

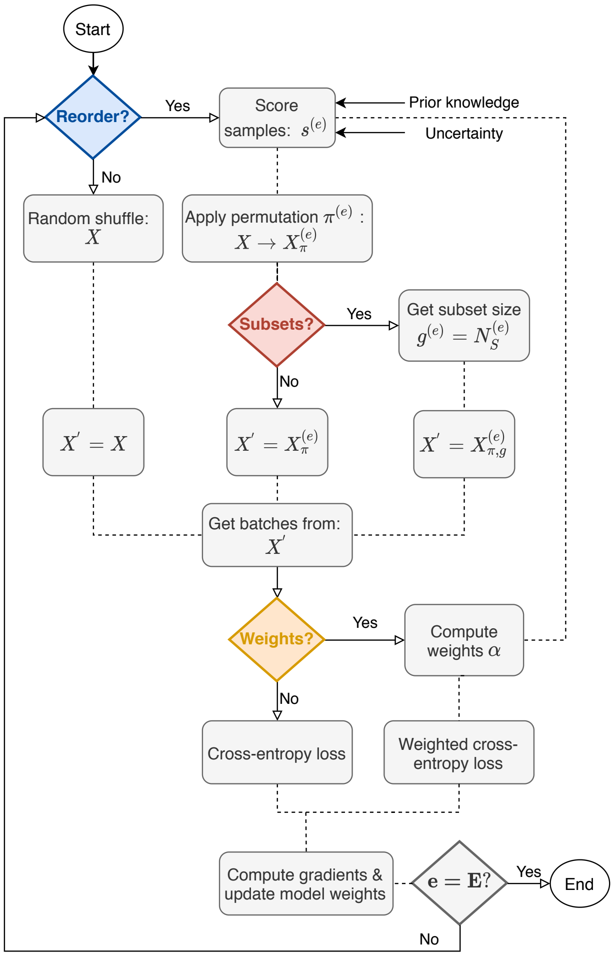

Start
Reorder?
Yes
Score samples: $s^{(e)}$
Prior knowledge
Uncertainty
No
Random shuffle: X
Apply permutation $\pi^{(e)}$: $X \rightarrow X_{\pi}^{(e)}$
Subsets?
Yes
Get subset size $g^{(e)} = N_S^{(e)}$
No
$X^{'} = X$
$X^{'} = X_{\pi}^{(e)}$
$X^{'} = X_{\pi,g}^{(e)}$
Get batches from: $X^{'}$
Weights?
Yes
Compute weights α
No
Cross-entropy loss
Weighted cross-entropy loss
Compute gradients & update model weights
$\mathbf{e} = \mathbf{E}$?
Yes
End
No

Fig. 3: Diagram illustrating the components of the proposed unified CL method reuniting the three scheduling strategies: reorder, subsets, and weights. Straight lines are employed after a Yes/No junction because the flow is split. Otherwise, dotted lines are employed when there is no split.

initial (domain) knowledge, as in classical CL [24]. The second one is dynamic and based on the estimation of uncertainty, inspired by SPL [28], [40].

1) Prior knowledge: In this scenario, the initial scoring $s^{(0)}$ and, thus, curriculum probabilities $p^{(0)}$, are specified based on domain prior knowledge. We assume in this variant that the scoring values are defined per class:

$$s^{(0)}(\cdot, y = t) = \omega_t, \tag{3}$$

where $t \in \{1, \ldots, T\}$ serves as index of the classes. ω_t is defined specifically for each task (or dataset). Once that the scoring values have been initialized, they can be kept fixed or decayed towards a uniform distribution [24]. In either case, as the curriculum probabilities are predetermined a priori in Eq. (3), we refer to this approach as static CL.

Prior knowledge can be obtained, for example, extracting keywords from medical reports [20], based on the frequency of samples [21], [23], employing medical classification standards or quantifying inconsistencies in the annotations [23]. Specifically for this work, we define the initial probabilities for the proximal femur fracture images based on the Cohen's kappa coefficient [41]. This statistic is used to measure the agreement of clinical experts on the classification between two readings. Basically, the kappa coefficient quantifies the ratio between the observed and chance agreement. To better understand and illustrate the potential of CL, we also analyze our method on MNIST dataset. In this case, we extract prior knowledge by ranking the per-class F_1-score performance after few epochs of training. The exact values used for our experiments are specified in Subsection IV-C.

2) Uncertainty estimation: In absence of domain knowledge, we propose to estimate the priority of the training samples by dynamically quantifying the uncertainty of the model predictions. Uncertainty provides a way of systematically ranking the training samples based on the model's agreement on the predictions, with the benefit of not requiring any prior knowledge. At each epoch e, we compute the uncertainty in predicting a sample x_k, and use such uncertainty as its scoring value s_k. See Subsection III-B for the definitions of x_k and s_k. The goal is to emphasize samples with high information gain at early stages of training, *i.e.* to rapidly reduce the error in highly-misleading samples.

To estimate the uncertainty of the model predictions, we employ MC dropout [29]. In this training regime, each epoch includes two stages [42]: uncertainty estimation and label prediction. In the uncertainty estimation stage, we perform L stochastic forward passes on the model under random dropout. The L estimators are used to measure the uncertainty of the output of the model. In the prediction stage, a single forward pass is performed. Then, the classification loss is used to measure the difference between the prediction and the label.

Let $\sigma \in \mathbb{R}^T$ be the (softmax) output of the CNN. This output represents the probability distribution of the predicted label over the set of the possible classes for sample x, *i.e.* $P(y = t \mid x, \theta) := \sigma_t$. We measure uncertainty as the entropy [43] of the output distribution, *i.e.* predictive entropy:

$$H(y|x, \theta) = -\sum_{t=1}^{T} P(y = t \mid x, \theta) \cdot \log P(y = t \mid x, \theta). \tag{4}$$

This measurement helps to discriminate points that are far from all training data, yet the model assigns high confident prediction (low predictive entropy). We aim to minimize the effect of these samples, with a small weight or bringing them at a later stage in training [**?**].

The output distribution $P(y = t|x, \theta)$ can be approximated using MC integration:

$$\tilde{P}(y = t \mid x, \theta) = \frac{1}{L} \sum_{l=1}^{L} P(y = t \mid x, \theta_l), \tag{5}$$

where $P(y = t \mid x, \theta_l)$ is the probability of input x to take class t with model parameters $\theta_l \sim q(\theta)$, with $q(\theta)$ being the (dropout) variational distribution. We set the scoring function to be the estimated predictive entropy, computed from the MC estimated output distribution $\tilde{\sigma}_t = \tilde{P}(y = t \mid x, \theta)$:

$$s = -\sum_{t=1}^{T} \tilde{\sigma}_t \cdot \log \tilde{\sigma}_t. \tag{6}$$

By assigning low scoring values to predictions with low predictive entropy, we decrease the priority of samples with low information gain. Note that in contrast with Eq. (3), here, the scoring elements s_i are defined independently for each sample, and updated after each epoch. Only few works measure uncertainty while learning the classification task [44]. To the best of our knowledge, our proposed dynamic uncertainty-driven curriculum strategy is novel for CAD.

B. Data scheduler

In the following, we define the scheduling elements required for reordering and pacing our training data: a scoring function s, curriculum probabilities p, a permutation function π, a pacing function g, and a weighting function α. The data scheduler takes as input the training set X, the scoring and pacing functions, s and g, respectively, and it outputs the reordered set/subset, partitioned in mini-batches. All components are updated at each epoch e.

- The *scoring function* $s : \mathcal{X} \to \mathbb{R}$ ranks the curriculum priority of each training pair. The curriculum priority can take various forms, such as difficulty or prediction disagreement. An example (x_i, y_i) has higher priority than example (x_j, y_j) if $s(x_i, y_i) > s(x_j, y_j)$. We define $s_i = s(x_i, y_i)$ and, in an abuse of notation, use s to denote both the scoring function and the vector $(s_1, \dots, s_N)$.
- The *curriculum probabilities* p are obtained by normalizing the score function values (while preserving the order and ensuring they add up to 1). For example, one can choose $p_i = s_i/||s||_1$, assuming $s_i \geq 0$. A pair (x_i, y_i) is more likely to be presented earlier to the optimizer than a pair (x_j, y_j) if $p_i > p_j$.
- The *reordering function* $\pi : [1, \dots, N] \to [1, \dots, N]$ is a permutation. It is determined by resampling without replacement X according to the curriculum probabilities p.
- The *pacing function* $g : \mathbb{N} \to \mathbb{N}$ controls the learning speed by presenting growing subsets of data. The batch size B is kept fixed. The non-decreasing mapping g determines the subset size $N_S \leq N$ at each training epoch e, *i.e.* $g(e) = N_S^{(e)}$.
- The *weighting function* $\alpha : \mathcal{X} \to \mathbb{R}$ favors the samples that have higher priority according to the curriculum probabilities. These per-sample weights are applied directly to the classification loss.

Taking into account the scheduling elements introduced, we can rewrite the optimization loss at epoch e as:

$$\mathcal{L}_\theta^{(e)} = \frac{1}{N_S^{(e)}} \sum_{j=1}^{N_S^{(e)}/B} \sum_{k=1}^{B} \hat{\alpha}_{k,j}^{(e)} \cdot L_\theta(\hat{x}_{k,j}^{(e)}, \hat{y}_{k,j}^{(e)}), \qquad (7)$$

where $\hat{x}_{k,j}^{(e)} = x_{\pi^{(e)}(k+(j-1)\cdot B)}$ corresponds to the k-th sample from the j-th batch at epoch e after reordering π. The same relation follows for its corresponding label and weight, $\hat{y}_{k,j}^{(e)}$ and $\hat{\alpha}_{k,j}^{(e)}$, respectively. We will drop superscript (e) when no confusion arises. Also, we simplify notation and use x_k (and y_k, α_k) to refer to a given (already reordered) sample (and label, weight). This equation encompasses the three main curriculum strategies from the literature: reordering, increasing subsets, and weighting.

C. Scheduling data with curriculum learning

In practice, any curriculum is implemented by assigning a predefined or estimated probability p_i to each training pair (x_i, y_i), as described in Subsection III-A. Fig. 3 visualizes the data flow in the different scheduling strategies, each of them being depicted by a diamond shape: reorder, subsets, and weights. The scoring function s and curriculum probabilities p are common to the three scheduling approaches, whereas the reordering function π is used in the reorder and subset strategies.

The first mechanism, *reorder*, presents the samples to the optimizer in a "smart" probabilistic order, instead of the typical random permutation. This strategy aims to deal with low-priority cases at a later stage of training [23], [24], [45]. At the beginning of every epoch e, the training set $X = \{(x_1, y_1), \dots, (x_N, y_N)\}$ is permuted to $X_\pi^{(e)} = \{(x_{\pi^{(e)}(1)}, y_{\pi^{(e)}(1)}), \dots, (x_{\pi^{(e)}(N)}, y_{\pi^{(e)}(N)})\}$ using the reordering function $\pi^{(e)}$. This mapping results from sampling the training set according to the curriculum probabilities $p^{(e)}$ at the current epoch e. Mini-batches are formed from $X_\pi^{(e)}$.

The second method, *subsets*, builds upon the reordered training set and selects gradually increasing subsets at every epoch. The purpose is to reduce the effect of outliers at the beginning of training [18], [40], [45]. Mini-batches are obtained from $X_{\pi,g}^{(e)} \subseteq X$, where $X_{\pi,g}^{(e)}$ are the first $N_S^{(e)}$ pairs of $X_\pi^{(e)}$. The subset size at every epoch $N_S^{(e)}$ is determined by the pacing function g. For simplicity, in our experiments we choose g to be a staircase function:

$$g(e) = N_S^{(e)} = \begin{cases} N_S^{(0)} + e \cdot \Delta & if \quad 1 \leq e < E_S \\ N & if \quad\quad e \geq E_S \end{cases} \qquad (8)$$

where $\Delta = (N - N_S^{(0)})/E_S$, $N_S^{(0)}$ is a predefined initial subset size, and E_S is the number of epochs before considering the whole training set.

A counter τ_i is introduced to track the selected pairs. Their scoring vector is decreased, thus favoring new pairs in the subsequent epoch. We choose to update the scoring vector using an exponential decay:

$$s_i^{(e)} = s_i^{(e-1)} \cdot \exp(-\tau_i^2/10) \quad e = 1, \dots, E. \qquad (9)$$

The third approach, *weights*, assigns scalar weights to training samples based on their curriculum probabilities [40]. We propose to weight the classification loss L_θ of each training sample in Eq. (7), in the form of a weighted cross-entropy loss. The role of the weights is to decrease the contribution to the classification loss of samples with low priority. We choose the weights $\hat{\alpha}_{k,j}$ to correspond to a per-batch normalization of the curriculum probabilities:

$$\hat{\alpha}_{k,j}^{(e)} = \frac{p_{k+(j-1)\cdot B}^{(e)}}{\max_m p_{m+(j-1)\cdot B}^{(e)}} = \frac{\hat{p}_{k,j}^{(e)}}{\max_m \hat{p}_{m,j}^{(e)}}. \qquad (10)$$

When the curriculum is driven by uncertainty, the resulting approach is similar to boosting [46]. In the boosting method, misclassified examples are given a higher weight than correctly classified ones. This is known as "re-weighting". Following the same principle, we use the uncertainty at every epoch, in our curriculum data scheduler, to update the values of the weights.

IV. Experimental validation

In order to validate the positive effect of data scheduling on the classification performance, we perform experiments on two types of image databases: (i) a real in-house dataset of a moderate size and naturally suffering from imbalance and noisy labels, and (ii) the MNIST dataset. The second one is used for additional analysis under controlled experiments to further illustrate the potential of CL.

A. Datasets

Proximal femur fractures: Our clinical dataset consists of anonymized X-rays of the hip and pelvis collected at the trauma surgery department of the Rechts der Isar Hospital in Munich. Images of 2500 × 2048 pixels were gathered from a group of 780 patients. Each patient study contained one or two radiographs. Most of the images were Anterior-Posterior (A-P), only 4% were side view. The collection of these radiographs was approved by the ethical committee of the Faculty of Medicine from the Technical University of Munich, under the number 409/15 S. The dataset consists of 327 type-A, 453 type-B fractures and 567 non-fracture cases. Class labels were assigned by clinical experts according to the AO classification standard [3]. Each type of fracture is further divided into 3 subclasses depending on the morphology and number of fragments of the fracture, see Fig. 1. Subtypes of the fracture classes are highly unbalanced, reflecting the incidence of the different fracture types. In particular, the number of images for the subclasses is as follows: type-A (114, 197, 16), and type-B (79, 241, 133). Clinicians also provided square bounding box annotations containing the head and neck of the femur. We leveraged these annotations, cropped and resized the image to 224 × 224 pixels. The dataset was split patient-wise into three parts with the ratio 70%:10%:20% to build respectively the training, validation and test sets. We evaluate the classification performance of the 3-class (type-A or type-B and non-fracture) and 7-class (fracture subtypes and non-fracture) classification tasks. The train, validation and test distributions were balanced between fracture type-A, type-B, and non-fracture cases. To achieve an equal proportion of subtype representation (of approximately 12%), data augmentation techniques were used. Specifically, techniques such as translation, scaling and rotation were combined.

MNIST: The MNIST handwritten digit database is publicly available[1]. It has a training set of 50000 examples and a validation and test sets of 10000 examples each. Classes are equally represented.

B. Experimental Setting

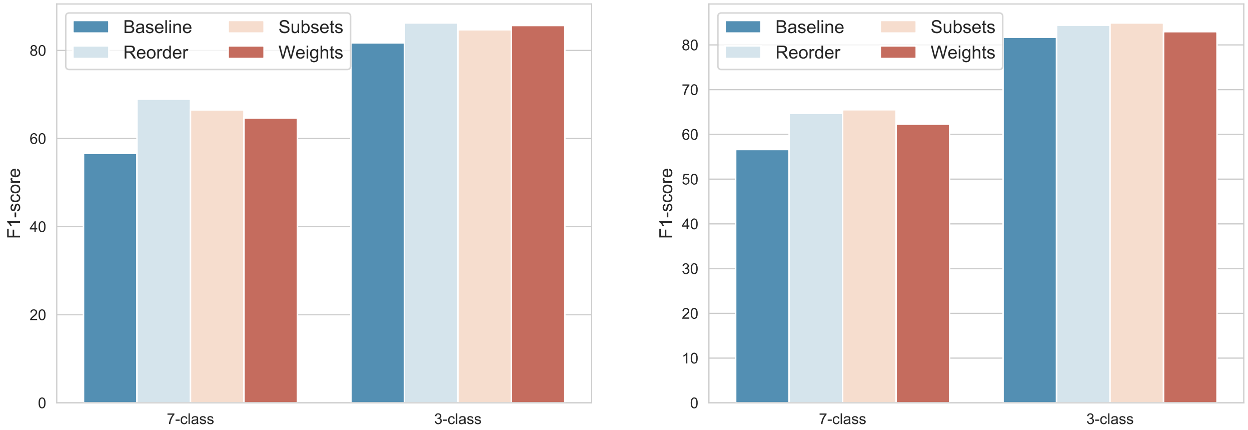
We perform a comparative evaluation of the classification task with five series of experiments. Our method is contrasted against its "anti-curriculum" approach, *i.e.* the curriculum probabilities are complemented so that training samples follow the reverse order, "random" criterion, *i.e.* the curriculum probabilities are assigned randomly, and the "baseline" model. The baseline model does not consider any data scheduling elements, and it is trained on randomly shuffled versions of the whole training set.

In the first series of experiments, we examine the performance of our method driven by prior knowledge. In the second series, we consider the use of uncertainty to overcome the lack of prior knowledge. Our clinical dataset inherently suffers from class-imbalance, unreliable annotations and a limited size. The 7-class discrimination task is challenging as reflected by i) the existing intra- and inter-expert agreement (66% among residents *vs.* 71% among experienced trauma surgeons); and ii) the long and shallow learning curve of young trauma surgery residents who acquire the classification skills during the daily routine. For the remaining experiments, we employ MNIST, as a controlled environment, to investigate such challenging scenarios. In the third series, we evaluate the classification performance when training with limited amounts of data. In the fourth series, we present the results that deal with class-imbalance. Finally, in our last series of experiments, we discuss and show the performance under the presence of label noise.

C. Implementation details

Architectures and optimization hyperparameters: We train our models 10 times for 30 epochs, with an early stopping criterion of no improvement in the validation set for 20 epochs. For the digit recognition task, we use an upgraded ConvPool-CNN-C [47] proposed by [48], illustrated in Fig. 6 of Suppl. Material. This architecture replaces pooling layers by convolutional layers with a stride of two. Besides, the small convolutional kernels greatly reduce the number of parameters of the network. It yielded competitive performance on several object recognition datasets (CIFAR-10, CIFAR-100, ImageNet). For the fracture classification, we deploy a ResNet-50 [49] pretrained on the ImageNet dataset, on account of the limited size of our dataset and the benefits of transfer learning [7], [14]. We limit our evaluation to those two CNNs, since Weinshall *et al.* [39] reported that CL lead to an improved generalization performance with both 'small' and 'large' architectures. For both architectures, we use a mini-batch size of 64, an initial learning rate of $1e{-}3$, and a dropout rate for the fully connected layer of 0.9 (0.7 for uncertainty estimation). Our ResNet-50 is trained with SGD and a momentum of 0.9. The learning rate is decayed by a factor of 10 every 10 epochs. ConvPool-CNN-C is trained with Adam. For the weighting strategy, since the batch size is directly related to the computation of the sample weights, we evaluated different batch sizes (16, 32, and 64). We found that the curriculum is robust, achieving the lowest standard deviation for $B = 64$ (see Table IX in Suppl. Material). For the

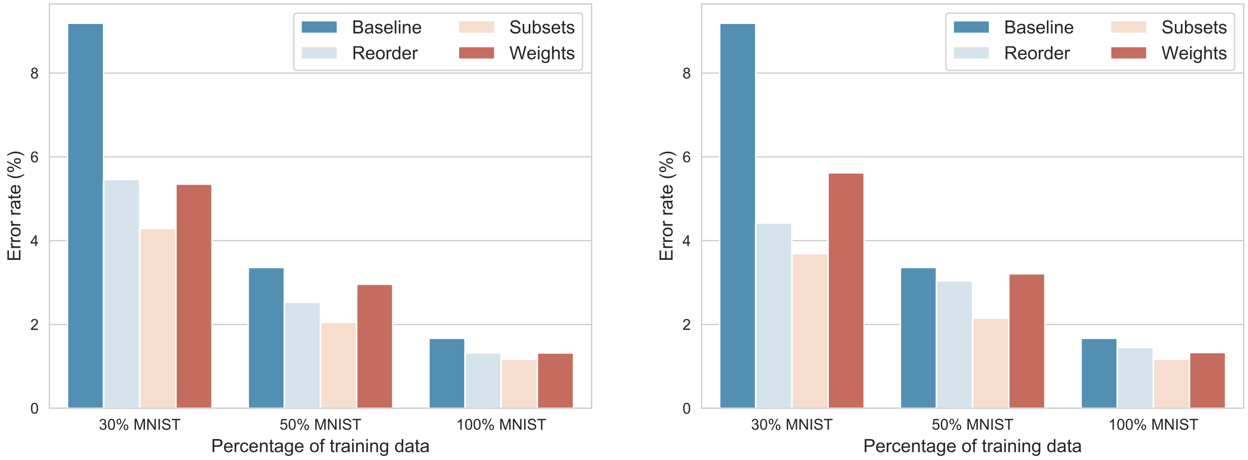
[1] http://yann.lecun.com/exdb/mnist/



(a) Prior knowledge-driven CL

(b) Uncertainty-driven CL.

Fig. 4: F1-score for multi-class fracture classification. The proposed curriculum method improves the classification performance in all variants.



(a) Prior knowledge-driven CL

(b) Uncertainty-driven CL.

Fig. 5: Classification performance for digit recognition. The proposed curriculum method reduces the baseline error rate in all variants under limited amounts of data.

subsets strategy, we choose as hyperparameters: the warm-up epochs $E_S = 10$ and the initial subset size $N_S^{(0)}$ to 25% of the training data size at each scenario. We evaluated several warm-up epochs $E_S = \{5, 10, 20\}$ and sizes for the initial subset $N_S^{(0)} = \{25\%, 40\%\}$. Results for the different configurations were comparable (see Tables VII-VIII in Suppl. Material).

Prior knowledge:

- Proximal femur fractures. In this setting, we leverage, as prior knowledge the intra-reader agreement from a committee of experts: a trauma surgery attendant with one year experience, a trauma surgery attending and a senior radiologist. The scoring values for the seven classes are the following:

$$\omega = (0.69, 0.56, 0.62, 0.60, 0.56, 0.38, 0.92). \qquad (11)$$

  These values correspond to the multi-read kappa agreement described in Results section [50].

- MNIST. In absence of domain-specific knowledge, a CNN is trained for 5 epochs. After observing the F_1-score of each of the classes, weights are assigned, by ranking the classes from easiest (highest F_1-score) to hardest (lowest F_1-score). Then, training is restarted from scratch using these particular weights. We specify the values for the experiments with limited amounts of data $\omega_{limited}$, under class-imbalance $\omega_{imbalance}$, and with noisy labels ω_{noise}:

$$\omega_{limited} = (7, 10, 5, 4, 9, 1, 8, 6, 2, 3) \qquad (12)$$

$$\omega_{imbalance} = (3, 10, 7, 8, 5, 6, 9, 4, 1, 2) \qquad (13)$$

$$\omega_{noise} = (8, 10, 9, 7, 5, 1, 2, 3, 4, 6). \qquad (14)$$

V. Results

A. Prior knowledge-driven CL

We evaluated the performance of the classifier with our data scheduler and verified that establishing a curriculum based on prior knowledge is a good and suitable option to improve classification performance. Results for proximal femur fracture are summarized in Table I-top, and for digit recognition in Table II-top. We found that the three variants helped to improve the performance of the two datasets. In contrast with the anti-CL approach, accuracy was increased with respect to the baseline.

For MNIST, we found that training starting with an easy subset, and gradually increasing the subset by adding more difficult samples was the best strategy for the three scenarios as shown in Fig. 5-a. A comparable improvement with respect to the baseline was found when the decay of Eq. (9) was introduced in reorder strategy and sampling with replacement was performed instead.

For fracture classification, the F_1-score for 7-class was improved up to 15% compared to the baseline (see Table I and Fig. 4). This score is comparable to state-of-the-art results [23] and experienced trauma surgeons [51]. Although related works report results for binary classification (fracture/no fracture), we are not aware of other teams doing fine-grained multi-class fracture classification. In this case, the best method was reordering the whole training set. We hypothesize that by re-ordering, we improve diversity by including the more challenging fine-grained fractures classification task than employing subsets of the data. Furthermore, as specified in Subsection IV-C, the CNN for fracture classification was pretrained, whereas for digit recognition the CNN was trained from scratch. From the results in Table I, we can say that our method is compatible with transfer learning.

B. Uncertainty-driven CL

Here, assuming lack of prior knowledge, we confirmed that uncertainty estimation can guide the data scheduling. Results are presented in Table I-bottom and Table II-bottom for fractures and MNIST, respectively. For the fine-grained 7-class proximal femur fractures classification, the F_1-score was improved up to 16% compared to the baseline. In this case, we found that weighting the samples was not as beneficial as reordering or sampling subsets. For digit classification, the error rate was reduced up to 30%, see Fig. 5-b. Anti-CL leading to a better performance than the baseline is a behaviour also reported in [24]. Furthermore, we found that this behaviour was sporadic and not statistically significant with respect to the baseline, whereas the CL approach was consistent and statistically significant.

Table III presents an analysis of the classification performance per class. We found a uniform improvement within each class when using our CL strategies with respect to the Baseline. Furthermore, not all classes are equally improved. For example, the easiest ‘Normal’ class is only slightly improved by a maximum of 6%, whereas more difficult classes like A1 or B2 are improved by 30% and 39%, respectively. The difference for A3 might not be significant due to the limited number of samples in the test set for this class.

C. Limited amounts of data

Table II shows the digit recognition performance when restricting the amount of training data to 30% and 50%. When employing our curriculum framework, the error rate for digit classification is reduced in all cases. We found that employing subsets in the first epochs based on uncertainty was the best strategy. Moreover, the effect of our curriculum approach was more evident on the more challenging scenario. The error rate was reduced by up to 59% training with only only 30% of the data. Interestingly, we found that when training with only 30% of data, the use of random subsets also reduced the error rate. This behaviour goes along with some findings about training with partial data [52].

D. Class-imbalance

We evaluated our proposed curriculum method in a controlled experiment under class-imbalance with the MNIST dataset. Specifically, the number of examples of two classes (digits 1 and 7) are limited to 30% of the available cases. Results in Table IV show that our approach can cope with class-imbalance and improved over the baseline result. Similar to the experiment with limited amount of data, the use of high-priority subsets, selected based on prior knowledge or uncertainty, was the best approach. The subsets approach reduced the error rate from 2.53% to 1.79%.

E. Noisy labels

Using MNIST and a controlled setting, we corrupted a randomly selected 30% of random training labels by assigning to them the subsequent label digit, *i.e.* zeroes become ones, ones become twos, *etc*. Table IV reports the mean error rate (%) when evaluating the digit classification. We found that all the variants of our unified CL framework were effective to deal with noisy labels and beat the baseline. In this case, prior knowledge was not as beneficial as the estimation of model prediction uncertainty. The best variant was using uncertainty to weight the classification loss, reducing the error rate by 43%. The fact that uncertainty performed better than prior knowledge was expected, since noise may affect individual samples and not entire classes. It is more reasonable to use a scoring function that independently affects the samples. Moreover, although reordering and subsets presented all the samples at convergence, weighting seemed to be the only strategy to remove or reduce the influence of the flawed labels.

VI. Discussion

In this work, we bring together several ideas from the literature and present them into a unified CL formulation. We experimentally demonstrate the effectiveness of ranking and scheduling training data for the challenging multi-class classification of proximal femur fractures. Most of the previous work [31]–[33] only target the fracture detection task, and Tanzi *et al.* [37] does not obtain the same level of granularity.

TABLE I: Fracture classification results over 10 runs: mean F_1-score. The highlighted indices in bold correspond to the best metric per curriculum method. The underlined values correspond to the best metric per scenario, *i.e.* 3-class (type-A or type-B and non-fracture) and 7-class (fracture subtypes and non-fracture) classification. Statistical significance with respect to baseline is marked with *.

Prior knowledge		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	56.62	34.56	**68.93***	58.90	50.89	**66.50***	58.26	55.20	**64.65***
3-class	81.71	60.46	**86.23***	80.82	75.64	**84.69***	80.66	75.33	**85.66***
Uncertainty		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	56.62	61.29	**64.70***	58.90	62.06	**65.51***	58.26	58.29	**62.29***
3-class	81.71	82.48	**84.38***	80.82	82.79	**84.90***	80.66	82.69	**82.96***

TABLE II: Digit classification results over 10 runs: mean error rate (%). The highlighted values in bold correspond to the best metric per curriculum method. The underlined values correspond to the best metric per scenario, *i.e.* percentage of data. Statistical significance with respect to baseline is marked with *.

Prior knowledge		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	9.19	9.28	**5.46***	5.60	13.17	**4.29***	8.01	5.78	**5.35***
50% MNIST	3.36	5.21	**2.53***	3.96	4.21	**2.05***	4.10	4.11	**2.96***
100% MNIST	1.67	2.53	**1.32***	1.96	1.78	**1.17***	1.98	1.79	**1.32***
Uncertainty		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	9.19	8.94	**4.42***	5.60	8.85	**3.69***	8.01	8.50	**5.62***
50% MNIST	3.36	3.23	**3.04**	3.96	4.21	**2.15***	4.10	4.77	**3.21**
100% MNIST	1.67	2.29	**1.45**	1.81	2.02	**1.17***	1.99	1.66	**1.33***

TABLE III: Relative per-class F_1-score with respect to the baseline. Up arrows indicate improvement, down arrows stand for degradation of performance.
The highlighted values in bold correspond to the best strategy per class.

Method	A1	A2	A3	B1	B2	B3	Normal	Avg.
Baseline	29.50	61.40	25.20	52.80	29.20	48.70	84.80	57.72
Reorder (Prior K.)	**↑30%**	**↑21%**	↑17%	↑22%	**↑39%**	**↑18%**	**↑6%**	**↑16%**
Subsets (Prior K.)	↑11%	↑15%	↑0%	**↑23%**	↑30%	**↑18%**	**↑6%**	↑13%
Weights (Prior K.)	↑22%	↑15%	↑23%	↑11%	↑17%	↑4%	↑5%	↑11%
Reorder (Uncertainty)	↑8%	↑16%	↑19%	↑17%	↑22%	↑12%	↑3%	↑11%
Subsets (Uncertainty)	↑19%	↑21%	**↑43%**	↑14%	↑25%	↑9%	↑2%	↑12%
Weights (Uncertainty)	↑11%	↑13%	↓46%	↑7%	↑23%	↑5%	↑2%	↑7%

TABLE IV: Comparison of curriculum strategies driven by prior knowledge and uncertainty, under class-imbalance and label noise for the MNIST dataset. Mean error rate (%). The highlighted values in bold correspond to the best strategy per scenario.

		Reorder		**Subsets**		**Weights**	
	Baseline	Prior K.	Uncertainty	Prior K.	Uncertainty	Prior K.	Uncertainty
Class-imbalance	2.53	2.08	2.05	**1.79**	2.08	2.31	2.22
Label Noise	9.46	8.76	8.42	8.28	7.24	8.49	**5.42**

Our CL schemes achieve state-of-the-art results on the 7-class classification task. Furthermore, we also show the benefits of our CL strategies in a controlled set-up with MNIST dataset, specifically, under demanding scenarios such as class-imbalance, limited amounts of data and noisy annotations.

Inspired by classical CL, we leveraged prior knowledge to define the data scheduling elements. In our formulation, this prior knowledge only requires defining a scalar value per class. In case of multiple experts annotating the dataset, this knowledge can be derived from their intra- or inter-expert variability, or by asking the experts about the perceived difficulty of each class. One limitation of this approach is that the use of prior knowledge at the class level may be less informative for the CNN than at the sample level. More sophisticated sample-level strategies could be used. For example, images with noise or low resolution could be assigned a lower confidence score. When prior knowledge is not available, we have shown that uncertainty can be used to guide the optimization. We used MC dropout to estimate uncertainty. This has the advantage of not requiring any change in the CNN architecture, but it is computationally demanding. Indeed, the training time is doubled in this variant. Instead, one could investigate the use of a Dirichtlet distribution to parametrize the output of the network. Then, the behavior of such predictor could be interpreted from an evidential reasoning perspective, such as in subjective logic [53], [54]. Future research directions for defining the scoring function could be based on other uncertainty measures such as quantifying out-of-distribution samples [55] or evidence theory [53]. We restricted our study to the predictive entropy of the model, which includes both aleatoric and epistemic uncertainty. We reckon that assessing separately each type of uncertainty could be advantageous for some applications. Moreover, if training time is not a concern, uncertainty does not only rank at the class but at the sample level. This scoring function is more appropriate for noisy annotations, since noise may affect individual samples and not entire classes. We restricted our experiments to individually applying each strategy, future work could try to combine them.

We evaluated three CL variants that consisted of reordering the whole training set, sampling subsets of data, or individually weighting training samples. Our CL schemes are compatible with any architecture and SGD training [39]. They only require domain-specific knowledge or the estimated uncertainty for the definition of the scoring function, hence the curriculum. The reordering and subsets performances are very similar but if the dataset is too complex for the amount of available data (fractures), it seems better to keep the entire training set. We found similar performance when the curriculum probabilities were decayed towards a uniform distribution [24] or maintained stable in our reorder and weights variants. Regarding the latter, we have proposed a simple and effective weighting scheme. In future work, we plan to explore other weighting strategies, *e.g.* the focal loss [56], which is well suited for class-imbalance scenarios, and the large margin loss [57], which has been shown beneficial under limited amounts of data and when noisy labels are present.

VII. Conclusions

In this work, we have designed three CL strategies for the multi-class classification of proximal femur fractures. We validated the benefits of our approach reaching a performance comparable to state-of-the-art and experienced trauma surgeons. We have identified common scheduling elements in the literature and unified their formulation in our approach. We have proposed two types of ranking functions to prioritize training data, leveraging: prior knowledge and uncertainty. The best strategy for the classification of proximal femur fractures employed reordering with prior knowledge. In controlled experiments with the MNIST dataset, we have shown that the proposed method is effective for datasets with class-imbalance, limited or noisy annotations. From our experiments, we can conclude that for datasets of limited size or under the presence of class-imbalance, the use of the subsets variant can lead to an improved classification performance. One can either exploit prior knowledge to achieve a better performance, or if the computational cost is not an issue, leverage uncertainty. In the case of unreliable labels, we found that the more advantageous approach is the combination of weights with uncertainty.

References

[1] D. J. Ryan, H. Yoshihara, D. Yoneoka, K. A. Egol, and J. D. Zuckerman, "Delay in hip fracture surgery: An analysis of patient-specific and hospital-specific risk factors," *Journal of Orthopaedic Trauma*, vol. 29, no. 8, pp. 343–348, aug 2015. [Online]. Available: https://doi.org/10.1097/bot.0000000000000313

[2] D. Giannoulis, G. M. Calori, and P. V. Giannoudis, "Thirty-day mortality after hip fractures: has anything changed?" *European Journal of Orthopaedic Surgery & Traumatology*, vol. 26, no. 4, pp. 365–370, mar 2016. [Online]. Available: https://doi.org/10.1007/s00590-016-1744-4

[3] E. Meinberg, J. Agel, C. Roberts, M. Karam, and J. Kellam, "Fracture and dislocation classification compendium—2018," *Journal of Orthopaedic Trauma*, vol. 32, pp. S1–S10, Jan. 2018.

[4] S. E. Sheehan, J. Y. Shyu, M. J. Weaver, A. D. Sodickson, and B. Khurana, "Proximal femoral fractures: What the orthopedic surgeon wants to know," *RadioGraphics*, vol. 35, no. 5, pp. 1563–1584, Sep. 2015. [Online]. Available: https://doi.org/10.1148/rg.2015140301

[5] M. Bhandari, P. J. Devereaux, T. A. Einhorn, L. Thabane, E. H. Schemitsch, K. J. Koval, F. Frihagen, R. W. Poolman, K. Tetsworth, E. Guerra-Farfan, K. Madden, S. Sprague, G. Guyatt, and H. Investigators, "Hip fracture evaluation with alternatives of total hip arthroplasty versus hemiarthroplasty (HEALTH): protocol for a multicentre randomised trial," *BMJ Open*, vol. 5, no. 2, pp. e006 263–e006 263, feb 2015. [Online]. Available: https://doi.org/10.1136/bmjopen-2014-006263

[6] J. D. Zuckerman, "Hip fracture," *New England journal of medicine*, vol. 334, no. 23, pp. 1519–1525, 1996.

[7] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[8] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017, pMID: 28301734. [Online]. Available: https://doi.org/10.1146/annurev-bioeng-071516-044442

[9] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "NiftyNet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 113 – 122, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260717311823

[10] I. Bonavita, X. Rafael-Palou, M. Ceresa, G. Piella, V. Ribas, and M. A. González Ballester, "Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline," *Computer Methods and Programs in Biomedicine*, vol. 185, p. 105172, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260719300975

[11] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[13] J. Zhou, Z. Li, W. Zhi, B. Liang, D. Moses, and L. Dawes, "Using convolutional neural networks and transfer learning for bone age classification," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2017, pp. 1–6.

[14] H. Shang, Z. Sun, W. Yang, X. Fu, H. Zheng, J. Chang, and J. Huang, "Leveraging other datasets for medical imaging classification: Evaluation of transfer, multi-task and semi-supervised learning," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 431–439.

[15] C. N. Vasconcelos and B. N. Vasconcelos, "Increasing deep learning melanoma classification by classical and expert knowledge based image transforms," *CoRR, abs/1702.07025*, vol. 1, 2017.

[16] H. Su, X. Shi, J. Cai, and L. Yang, "Local and global consistency regularized mean teacher for semi-supervised nuclei classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 559–567.

[17] A. G. Roy, S. Conjeti, D. Sheet, A. Katouzian, N. Navab, and C. Wachinger, "Error corrective boosting for learning fully convolutional networks with limited data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 231–239.

[18] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, "Robust learning at noisy labeled medical images: applied to skin lesion classification," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1280–1283.

[19] A. Smailagic, P. Costa, H. Y. Noh, D. Walawalkar, K. Khandelwal, A. Galdran, M. Mirshekari, J. Fagert, S. Xu, P. Zhang *et al.*, "Medal: Accurate and robust deep active learning for medical image analysis," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 481–488.

[20] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *Machine Learning in Medical Imaging*, Y. Shi, H.-I. Suk, and M. Liu, Eds. Cham: Springer International Publishing, 2018, pp. 249–258.

[21] J. Yang, X. Wu, J. Liang, X. Sun, M. Cheng, P. L. Rosin, and L. Wang, "Self-paced balance learning for clinical skin disease recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2019.

[22] G. Maicas, A. P. Bradley, J. C. Nascimento, I. Reid, and G. Carneiro, "Training medical image analysis systems like radiologists," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 546–554.

[23] A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. A. González Ballester, and G. Piella, "Medical-based deep curriculum learning for improved fracture classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 694–702.

[24] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th International Conference on Machine Learning*, ser. ICML 2009. New York, NY, USA: ACM, 2009, pp. 41–48.

[25] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Hetero-modal image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 469–477.

[26] A. Jesson, N. Guizard, S. H. Ghalehjegh, D. Goblot, F. Soudan, and N. Chapados, "CASED: Curriculum adaptive sampling for extreme data imbalance," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 639–646.

[27] H. Kervadec, J. Dolz, É. Granger, and I. Ben Ayed, "Curriculum semi-supervised segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 568–576.

[28] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1189–1197. [Online]. Available: http://papers.nips.cc/paper/3923-self-paced-learning-for-latent-variable-models.pdf

[29] 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. [Online]. Available: http://proceedings.mlr.press/v48/gal16.html

[30] J. Wu, S. Ruan, C. Lian, S. Mutic, M. A. Anastasio, and H. Li, "Active learning with noise modeling for medical image annotation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, April 2018, pp. 298–301.

[31] M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley, "Deep learning predicts hip fracture using confounding patient and healthcare variables," *npj Digital Medicine*, vol. 2, no. 1, Apr. 2019. [Online]. Available: https://doi.org/10.1038/s41746-019-0105-1

[32] C.-T. Cheng, T.-Y. Ho, T.-Y. Lee, C.-C. Chang, C.-C. Chou, C.-C. Chen, I.-F. Chung, and C.-H. Liao, "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs," *European Radiology*, vol. 29, no. 10, pp. 5469–5477, Oct 2019. [Online]. Available: https://doi.org/10.1007/s00330-019-06167-y

[33] Y. Wang, L. Lu, C.-T. Cheng, D. Jin, A. P. Harrison, J. Xiao, C.-H. Liao, and S. Miao, "Weakly supervised universal fracture detection in pelvic x-rays," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 459–467.

[34] A. Kazi, S. Albarqouni, A. J. Sanchez, S. Kirchhoff, P. Biberthaler, N. Navab, and D. Mateus, "Automatic classification of proximal femur fractures based on attention models," in *Machine Learning in Medical Imaging*, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds. Cham: Springer International Publishing, 2017, pp. 70–78.

[35] A. Jiménez-Sánchez, A. Kazi, S. Albarqouni, C. Kirchhoff, P. Biberthaler, N. Navab, S. Kirchhoff, and D. Mateus, "Precise proximal femur fracture classification for interactive training and surgical planning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 5, pp. 847–857, Apr. 2020. [Online]. Available: https://doi.org/10.1007/s11548-020-02150-x

[36] T. Urakawa, Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, and N. Endo, "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network," *Skeletal Radiology*, vol. 48, no. 2, pp. 239–244, Feb 2019. [Online]. Available: https://doi.org/10.1007/s00256-018-3016-3

[37] L. Tanzi, E. Vezzetti, R. Moreno, A. Aprato, A. Audisio, and A. Massè, "Hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach," *European Journal of Radiology*, vol. 133, p. 109373, 2020.

[38] D. Needell, N. Srebro, and R. Ward, "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," *Mathematical Programming*, vol. 155, no. 1-2, pp. 549–573, 2016.

[39] D. Weinshall, G. Cohen, and D. Amir, "Curriculum learning by transfer learning: Theory and experiments with deep networks," *arXiv preprint arXiv:1802.03796*, 2018.

[40] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan, "Dynamic curriculum

learning for imbalanced data classification," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[41] K. A. Hallgren, "Computing inter-rater reliability for observational data: an overview and tutorial," *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.

[42] Y. Li, L. Liu, and R. T. Tan, "Certainty-driven consistency loss for semi-supervised learning," *CoRR*, vol. abs/1901.05657, 2019. [Online]. Available: http://arxiv.org/abs/1901.05657

[43] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[44] F. C. Ghesu, B. Georgescu, E. Gibson, S. Guendel, M. K. Kalra, R. Singh, S. R. Digumarthy, S. Grbic, and D. Comaniciu, "Quantifying and leveraging classification uncertainty for chest radiograph assessment," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 676–684.

[45] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," *arXiv preprint arXiv:1904.03626*, 2019.

[46] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[47] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[48] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/pdf?id=BJ6oOfqge

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

[50] A. Jiménez-Sánchez, A. Kazi, S. Albarqouni, C. Kirchhoff, P. Biberthaler, N. Navab, D. Mateus, and S. Kirchhoff, "Towards an interactive and interpretable CAD system to support proximal femur fracture classification," *CoRR*, vol. abs/1902.01338v1, 2019. [Online]. Available: http://arxiv.org/abs/1902.01338v1

[51] D. van Embden, S. Rhemrev, S. Meylaerts, and G. Roukema, "The comparison of two classifications for trochanteric femur fractures: The AO/ASIF classification and the jensen classification," *Injury*, vol. 41, no. 4, pp. 377–381, apr 2010. [Online]. Available: https://doi.org/10.1016/j.injury.2009.10.007

[52] M. N. Mermer and M. F. Amasyali, "Training with growing sets: A simple alternative to curriculum learning and self paced learning," 2018. [Online]. Available: https://openreview.net/forum?id=SJ1fQYlCZ

[53] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 3179–3189. [Online]. Available: http://papers.nips.cc/paper/7580-evidential-deep-learning-to-quantify-classification-uncertainty.pdf

[54] A. Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, 1st ed. Springer, 2018.

[55] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2016. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl

[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[57] G. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Large margin deep networks for classification," in *Advances in Neural Information Processing Systems*, 2018, pp. 842–852.

Supplementary Material

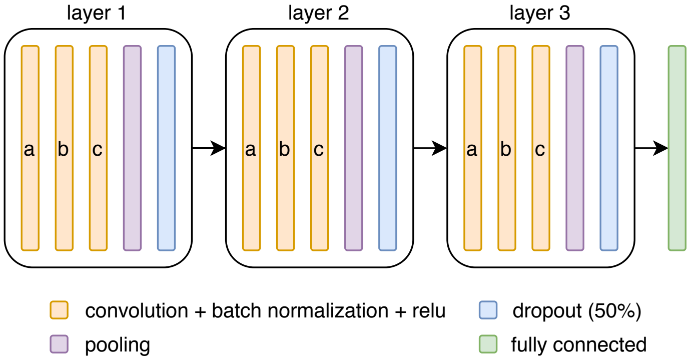


Fig. 6: Network architecture employed for the experiments with the MNIST dataset.

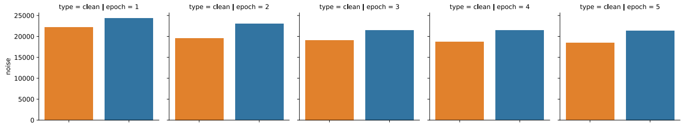


Fig. 7: Analysis of weights strategy under label corruption for MNIST dataset. Number of samples with a weight higher than the mean weight at that epoch. Random criterion and uncertainty are depicted in orange and blue, respectively.

TABLE V: Statistical significance analysis for proximal femur fracture experiments. T-test with respect to baseline. P-values below 0.05 are bold-faced.

Prior knowledge	**Reorder**		**Subsets**			**Weights**		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	**5.16E-04**	**3.03E-09**	1.33E-01	5.61E-02	**1.47E-06**	3.26E-01	6.34E-01	**1.48E-06**
3-class	**8.78E-04**	**5.43E-05**	6.02E-01	7.20E-02	**1.38E-02**	4.73E-01	**1.07E-02**	**1.57E-04**
Uncertainty	**Reorder**		**Subsets**			**Weights**		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	**7.10E-03**	**7.63E-05**	1.33E-01	**3.49E-03**	**2.27E-05**	3.26E-01	4.11E-01	**4.94E-05**
3-class	4.80E-01	**1.54E-02**	6.02E-01	4.37E-01	**2.98E-03**	4.73E-01	3.10E-01	**1.97E-02**

TABLE VI: Statistical significance analysis for MNIST experiments. T-test with respect to baseline. P-values are reported.

Prior knowledge	**Reorder**		**Subsets**			**Weights**		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	9.40E-01	**8.70E-04**	**1.55E-04**	1.62E-01	**3.52E-06**	3.29E-01	**1.08E-04**	**6.03E-05**
50%MNIST	**4.58E-04**	**8.37E-04**	7.19E-02	8.59E-02	**6.83E-05**	2.04E-01	5.69E-02	**7.23E-03**
100%MNIST	**3.75E-03**	**1.22E-02**	3.27E-01	3.88E-01	**5.98E-04**	**3.09E-02**	3.83E-01	**9.95E-03**
Uncertainty	**Reorder**		**Subsets**			**Weights**		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	8.38E-01	**7.16E-06**	1.55E-04	7.77E-01	**5.79E-07**	3.29E-01	5.01E-01	**3.11E-05**
50% MNIST	5.87E-01	3.77E-01	7.19E-02	1.15E-01	**1.86E-04**	2.04E-01	8.47E-02	6.43E-02
100% MNIST	**1.25E-02**	6.41E-02	3.27E-01	**3.44E-02**	**3.56E-03**	**3.09E-02**	7.11E-01	**2.49E-02**

TABLE VII: F_1-score for the 7-class fracture classification, mean (median) and standard deviation for the subsets strategy with different initial subset sizes $N_S^{(0)}$.

Prior knowledge-driven CL		**Uncertainty-driven CL**	
$N_S^{(0)} = 25\%$	$N_S^{(0)} = 40\%$	$N_S^{(0)} = 25\%$	$N_S^{(0)} = 40\%$
66.50 (66.02) ± 2.00	65.39 (65.76) ± 2.23	65.51 (66.32) ± 3.37	64.99 (65.63) ± 2.30

TABLE VIII: F_1-score for the 7-class fracture classification, mean (median) and standard deviation for the subsets strategy with different number of epochs E_S before considering the whole training set.

Prior knowledge-driven CL			**Uncertainty-driven CL**		
$E_S = 5$	$E_S = 10$	$E_S = 20$	$E_S = 5$	$E_S = 10$	$E_S = 20$
63.68 (63.42) ± 3.15	66.50 (66.02) ± 2.00	66.09 (64.04) ± 1.24	65.30 (65.78) ± 3.12	65.51 (66.32) ± 3.37	66.42 (66.68) ± 1.95

TABLE IX: F_1-score for the 7-class fracture classification, mean (median) and standard deviation for the weights strategy with different batch sizes.

Prior knowledge-driven CL			**Uncertainty-driven CL**		
$B = 16$	$B = 32$	$B = 64$	$B = 16$	$B = 32$	$B = 64$
65.35 (65.02) ± 2.97	65.69 (65.95) ± 2.11	64.65 (64.04) ± 1.56	64.66 (65.76) ± 2.27	66.92 (66.69) ± 2.18	62.60 (62.96) ± 1.63