Medical-based Deep Curriculum Learning for Improved Fracture Classification

Amelia Jiménez-Sánchez[1], Diana Mateus[2], Sonja Kirchhoff[3,4], Chlodwig Kirchhoff[4], Peter Biberthaler[4], Nassir Navab[5], Miguel A. González Ballester[1,6], Gemma Piella[1]

[1]BCN MedTech, DTIC, Universitat Pompeu Fabra, Barcelona, Spain
[2]Ecole Centrale de Nantes, LS2N, UMR CNRS 6004, Nantes, France
[3]Institute of Clinical Radiology, LMU München, Munich, Germany
[4]Department of Trauma Surgery, Klinikum rechts der Isar, Technische Universität München, Munich, Germany
[5]Computer Aided Medical Procedures, Technische Universität München, Munich, Germany
[6]ICREA, Barcelona, Spain

INTRODUCTION

- In a typical **educational** system, **learning relies on a curriculum** that introduces concepts building upon previously acquired ones.
- Bengio *et. al.* [1] made the connection between cognitive science and machine learning, demonstrating a boost in classification performance by combining **Curriculum Learning (CL)** and convolutional neural networks (CNN).
- These techniques have been successful in applications such as image segmentation or computer-aided diagnosis. However, they remain agnostic of clinical standards and medical protocols.
- Our **contribution** is on the **integration of knowledge**, extracted **from medical guidelines**, directly from expert recommendations or from ambiguities in their annotations, **to ease the learning process of CNNs**.

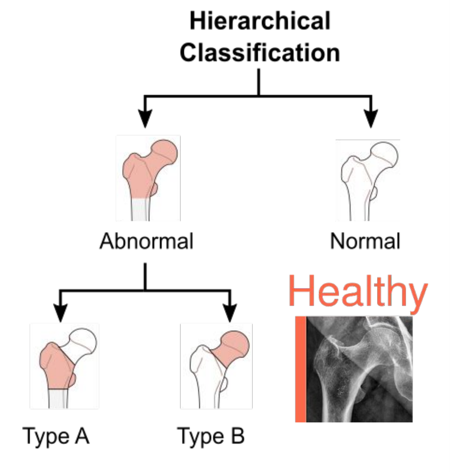

Figure 1. Hierarchical classification according to the AO standard.

DATASET

- The **proximal femur fractures** study was collected at Klinikum rechts der Isar, Munich (Germany). According to the **AO standard**, it consists of **327 type-A**, **453 type-B** and **567 non-fracture** cases. Subtypes of the fracture classes (A1-A3, B1-B3) are highly unbalanced, reflecting their incidence, as depicted in Fig. 3-(ii).

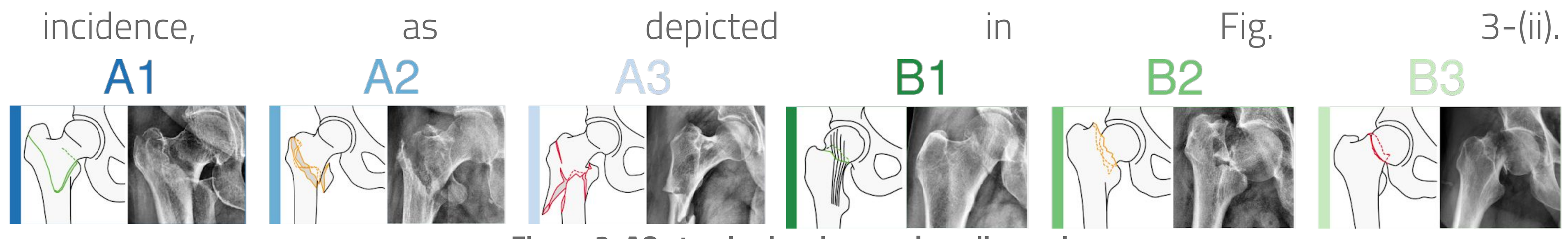

Figure 2. AO standard and example radiographs.

- **Early detection** and **classification** of such fractures are **essential** for guiding appropriate treatment and intervention. However, several years of training are needed, and **inter-reader agreement** ranges between **66-71%** [2] (trauma surgery residents and experienced surgeons).
- The dataset was split patient-wise into training (70%), validation (10%) and test (20%) sets.
- Offline data augmentation techniques such as translation, scaling and rotation were used.

MEDICAL-BASED DEEP CURRICULUM LEARNING

- We tackle multiclass classification problems where an **image** x_i needs to be assigned to a discrete **class label** y_i. The **training set** is denoted as $\{X, Y\}$.
- A **curriculum** $c \in C$ induces a **bias** of presenting samples to the optimizer. It reflects a notion of **"hardness"**, which in our work depends upon different forms of **prior medical knowledge**.
- Initially, each image x_i is assigned a **curriculum probability** $p_i^{(0)}$.
- At the beginning of every epoch e, $\{X, Y\}$ is permuted *to* $\{X, Y\}^c$ via a **reordering function** $f^{(e)}$. This mapping results from **sampling** $\{X, Y\}$ according to the probabilities at the current epoch $p^{(e)}$.
- **Probabilities** are **decayed** towards a uniform distribution based on:

$$q_i^{(e)} = p_i^{(e-1)} \cdot exp(-cn_i^2/10) \quad \forall\ e > 0 \qquad (1)$$

$$p_i^{(e)} = q_i^{(e)} / \sum_i q_i^{(e)} \qquad (2)$$

where cn_i is a counter that is incremented when the i-th sample is selected.

- **Initial curriculum probabilities** (see Fig. 3) are given by:

$$p^{(0)}(y_i=m) = w^c_m$$

where $m \in [1, 2, \ldots, M]$ *serves as index of the classes, and* w^c_m is defined according to the curriculum c:

(i) c = **uniform**: all classes are treated equally.

(ii) c = **frequency**: assigned a probability equal to their original incidence in the dataset.

(iii) c = **AO**: naively considered equally spaced, according to the difficulty ranking of the AO categories provided by an experienced radiologist.

(iv) c: **kappa**: given a probability proportional to the intra-reader agreement, measured with Cohen's kappa coefficient, found by a committee of experts.

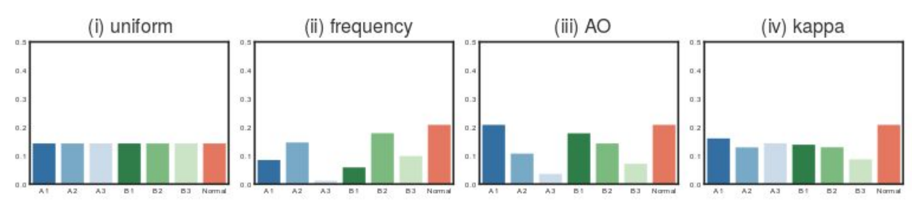

Figure 3 Initial probabilities $p^{(0)}$ **for our medical-based curriculums.**

Algorithm 1: CNN with medical curriculum data scheduler

```
input : X (X-ray images), Y (classification labels), c ∈ C (curriculum)
        B (mini-batch size) , E (expected training epochs)
for each epoch e do
    if first epoch then
        Define initial probabilities: p_i^(0) = w_m^c;
    else
        Update probabilities with Eqs. (1-2);
    end
    Get reordering function f^(e) by sampling {X, Y} according to p^(e)
    Permute training set f^(e) : {X, Y} ↦ {X, Y}^c ;
    for each training round do
        Get the next mini-batch from {X, Y}^c : {x_b, y_b}_{b=1}^B;
        Calculate cross-entropy loss L(y_b, ŷ_b);
        Compute gradients and update model weights;
    end
end
```

RESULTS

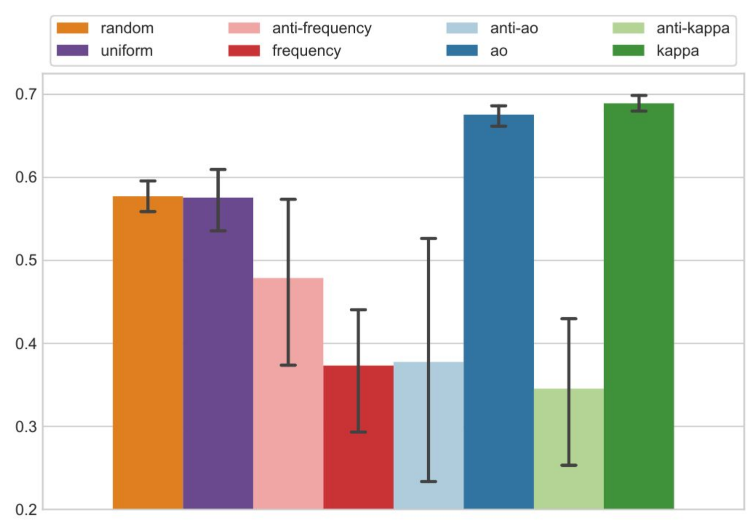

Figure 4. Mean F_1**-score and variance over 10 runs on the 7-way classification** of the curriculum strategies, together with their corresponding anti-curriculum, and compared against random and uniform-curriculum.

	7 classes			3 classes		
F_1-score	Mean	Median	SD	Mean	Median	SD
Random	0.5662	0.5731	0.0423	0.8063	0.8171	0.0337
Uniform	0.5757	0.5923	0.0590	0.8011	0.7971	0.0399
AO	0.6757	**0.6783**	0.0197	0.8651	**0.8657**	0.0172
Kappa	0.6893	**0.6900**	0.0150	0.8623	**0.8657**	0.0146
AO - 60%	0.6325	0.6188	0.0302	0.8457	0.8486	0.0191
Kappa - 60%	0.6352	0.6500	0.0398	0.8446	0.8457	0.0222

Table 1. Classification results over 10 runs. The highlighted indices in bold correspond to the best two models.

1. **Similar** performance found **randomly** shuffling the training data and learning with a **uniform-curriculum**.
2. The sequence of the samples has a significant effect in the classification, *i.e.* meaningful difference between curriculum and **anti-curriculum** strategies.
3. **AO-** and **kappa-curriculums** boost the median **F1-score** by approximately **15%** compared to the baselines.
4. Interestingly, our experiments suggest that, in the case of the **frequency**-curriculum, the **easy** scenario is the **class-imbalance** as in [3].
5. **3-class problem**: we achieve **state-of-the-art** results [4] and about **7%** better than random and uniform.
6. **60%** balanced **data**: AO- and kappa-curriculums perform **even better** than the **baselines** using **100% training data**.

CONCLUSIONS

- The integration of **medical knowledge** is useful for the **design** of **data schedulers** by means of **CL**.
- If observers agreement is not available, clinicians' **perception** of **difficulty** is a good estimate, as shown by our **AO**-curriculum.
- Our method can be used in **other applications** where medical decision trees are available, such as grading malignancy of tumors, as well as whenever intra- or inter-expert agreement is available.

FUTURE WORK

- Explore the combination of our medical curriculum data scheduler with **uncertainty** of the model, and investigate which samples play a more significant role in the decision **boundary**.

ACKNOWLEDGEMENTS

Obra Social "la Caixa"

This project has received funding from the European Unions 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.

REFERENCES

[1] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41-48). ACM.
[2] Embden, D.V., Rhemrev, S.J., Meylaerts, S.A., & Roukema, G.R. (2010). The comparison of two classifications for trochanteric femur fractures: the AO/ASIF classification and the Jensen classification. Injury, 41 4, 377-81 .
[3] Wang, Y., Gan, W., Wu, W., & Yan, J. (2019). Dynamic Curriculum Learning for Imbalanced Data Classification. arXiv, abs/1901.06783.
[4] Kazi, A., Albarqouni, S., Sanchez, A. J., Kirchhoff, S., Biberthaler, P., Navab, N., & Mateus, D. (2017, September). Automatic classification of proximal femur fractures based on attention models. In International Workshop on Machine Learning in Medical Imaging (pp. 70-78). Springer, Cham.