Computers and Electrical Engineering 78 (2019) 472–481

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

# Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks ☆

Janick Rohrbach [a], Tobias Reinhard [b], Beate Sick [a,c,*], Oliver Dürr [d,*]

[a] Zurich University of Applied Sciences, IDP, 8400 Winterthur, Switzerland
[b] Seantis GmbH, 6004 Luzern, Switzerland
[c] Zurich University, EBPI, 8006 Zürich, Switzerland
[d] Konstanz University of Applied Sciences, IOS, 78462 Konstanz, Germany

ARTICLE INFO

*Article history:*
Received 30 November 2018
Revised 14 May 2019
Accepted 5 August 2019
Available online 13 August 2019

*Keywords:*
Convolutional neural network
Deep learning
Machine learning
Classification
Rheumatoid arthritis
X-ray images
Medical image analysis

ABSTRACT

Rheumatoid arthritis is an autoimmune disease that causes chronic inflammation of synovial joints, often resulting in irreversible structural damage. The activity of the disease is evaluated by clinical examinations, laboratory tests, and patient self-assessment. The long-term course of the disease is assessed with radiographs of hands and feet. The evaluation of the X-ray images performed by trained medical staff requires several minutes per patient. We demonstrate that deep convolutional neural networks can be leveraged for a fully automated, fast, and reproducible scoring of X-ray images of patients with rheumatoid arthritis. A comparison of the predictions of different human experts and our deep learning system shows that there is no significant difference in the performance of human experts and our deep learning model.



## 1. Introduction

Rheumatoid arthritis is a common autoimmune disease affecting 0.5–1% of the worldwide population [1]. Caused by a malfunctioning immune system, the disease attacks healthy tissue instead of bacteria and viruses, provoking inflammation and resulting in warm, painful, swollen, and stiff joints. Specific causes of rheumatoid arthritis remain unclear. The disease is to date incurable; merely the symptoms can be treated. Irreversible structural damage to the bones in the joint can occur if the inflammation continues for an extended time. The severity of joint damage can be assessed by trained medical staff analyzing X-ray images of the hands and feet using the Ratingen scoring method [2]. Radiographic scoring methods play an important role in the detection and evaluation of the progression and treatment of rheumatoid arthritis. However, the process of manually evaluating radiographic damage in hands and feet requires several minutes of review per patient because many joints have to be evaluated individually. Additionally, it is often difficult to detect the subtle radiographic changes. Therefore, an automated system that makes bone erosion scoring faster and more consistent is desired in clinical practice and medical research.

The analysis of images with classical machine learning methods is rather difficult due to the lack of well-defined features. Hand-crafted features, often designed by domain experts, can be used as inputs for machine learning models such as Support

☆ This paper is for CAEE special section SI-mip. Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Li He.

* Corresponding authors.
*E-mail addresses:* beate.sick@zhaw.ch (B. Sick), oliver.duerr@htwg-konstanz.de (O. Dürr).

https://doi.org/10.1016/j.compeleceng.2019.08.003
0045-7906/

Vector Machine or Random Forest. Deep convolutional neural networks (CNN) greatly simplify this task, since it does not rely on hand-crafted features. The neural network will find appropriate features by itself during the training process. Therefore, deep CNN have replaced most classical machine learning methods in the field of computer vision.

In this work, we employ recent advances in computer vision to automate radiographic scoring. A similar approach to ours is shown by Murakami et al. in [3]. They focus on the extraction of regions of interest and subsequently use binary classification to predict whether bone erosion is present in those regions. In contrast to their work, our work classifies joints into six different classes. We distinguish between eroded and not eroded joints, as well as different stages of erosion. We also tackle the problem of highly imbalanced data with our weighted loss function. The extraction of regions of interest is not part of our work, since our dataset was already pre-processed to include only the joints.

The rest of this paper is arranged as follows. Section 2 reports on related works in automated radiograph scoring, medical image analysis with CNNs, and dealing with imbalanced data in deep learning. In Section 3 we describe the dataset, our different models and the achieved performances. In Section 4 we summarize the results and provides evidence that there is no disagreement between our model and domain experts. Finally, we conclude this paper in Section 5.

## 2. Related Works

Existing work on the automatic radiographic scoring of patients with rheumatoid arthritis mainly focuses on joint space measurement [4,5]. For example, Snekhalatha and Anburajan [4] use classical image recognition techniques such as the watershed algorithm to measure the joint space. More recent work exists for the degenerative, non-inflammatory osteoarthritis than for autoimmune disorders such as rheumatoid arthritis. Tiulpin et al. [6] use a deep siamese CNN for radiographic scoring of the severity of osteoarthritis, yielding an accuracy of 66.71%. Xue et al. [7] detect osteoarthritis with a CNN applied to X-ray images of hip joints with 92.8% accuracy.

Deep neural networks, such as CNNs, have gained significant attention in the medical image analysis field recently [8,9]. Goceri and Goceri present a multitude of application areas such as analysis of brain and breast images, or segmentation of organs [10]. Litjens et al. provide an extensive summary of over 300 different papers that apply deep learning for medical image analysis [8]. They show different applications such as mammographic mass classification, segmentation of lesions in the brain, leak detection in airway tree segmentation, diabetic retinopathy classification, prostate segmentation, breast cancer metastases detection in lymph nodes, skin lesion classification, and bone suppression in x-rays.

Several works have successfully applied deep learning to analyze biomedical images [8–15]. CNNs have been applied to MRI images of the brain to identify multiple sclerosis with a very high accuracy of 98.2% [11]. They report that the use of dropout and PReLU instead of ReLU increases the performance. In [12] CNNs have been used to analyze histopathological images of glioblastoma. Similar to our work, they classify their images into four different disease stages.

In our task we deal with imbalanced class sizes, a challenge for all classification models. In contrast to classical machine learning, only limited research on handling unbalanced data is available in deep learning. However, recently this problem has been investigated by Dong et al. [16]. They describe existing approaches of under-sampling the majority classes, over-sampling the minority classes as well as cost-sensitive learning, that applies a higher penalty to minority class samples. A new class rectification loss specifically designed for training on imbalanced datasets is also proposed in this paper. In a similar spirit, we use a weighting scheme based on the reciprocal class sizes to define a weighted version of the standard cross-entropy function, enhancing the importance of rare classes.

Pratt et al. [13] successfully apply CNNs to color fundus images to classify the severity of diabetic retinopathy. Similar to our work, their dataset consists of imbalanced, ordinal classes and they use standard cross-entropy loss function.

## 3. Materials and Methods

### 3.1. Dataset

Our data consists of X-ray images of the hands and feet of patients with rheumatoid arthritis. It was collected over 15 years by the Swiss Clinical Quality Management in Rheumatic Diseases Foundation (SCQM) as part of a Swiss register for rheumatoid arthritis. The registry has been described in detail elsewhere [17]. The collection of patient data for the SCQM register was approved by a national review board and, in accordance with the Declaration of Helsinki, all individuals included in this study have signed an informed consent form before enrolment. Fig. 1 shows a typical X-ray image of a hand with the five proximal interphalangeal (PIP) joints and five carpometacarpal (MCP) joints designated with blue bounding boxes. These joints in the hand are often affected in rheumatoid arthritis patients. The wrist joint can also be affected, but we limited our analysis to the ten finger joints.

Images of single joints were obtained from X-ray images of different resolution and quality. The dataset was restricted to joints of the left hand to simplify joint extraction. All joint images were scaled to a size of $150 \times 150$ pixels with an added black border if the extracted image resulted in a different aspect ratio. The resulting dataset consisting of 102,265 cropped grayscale images was used for the deep learning-based classification.

For the assessment of bone erosion, we use the Ratingen score [2], which is well established in the field. The different disease stages of this score are described in Table 1. The scoring was performed by medical raters trained explicitly for the task. These scores, along with the corresponding X-ray images, are used to train our model. We treat the six scores

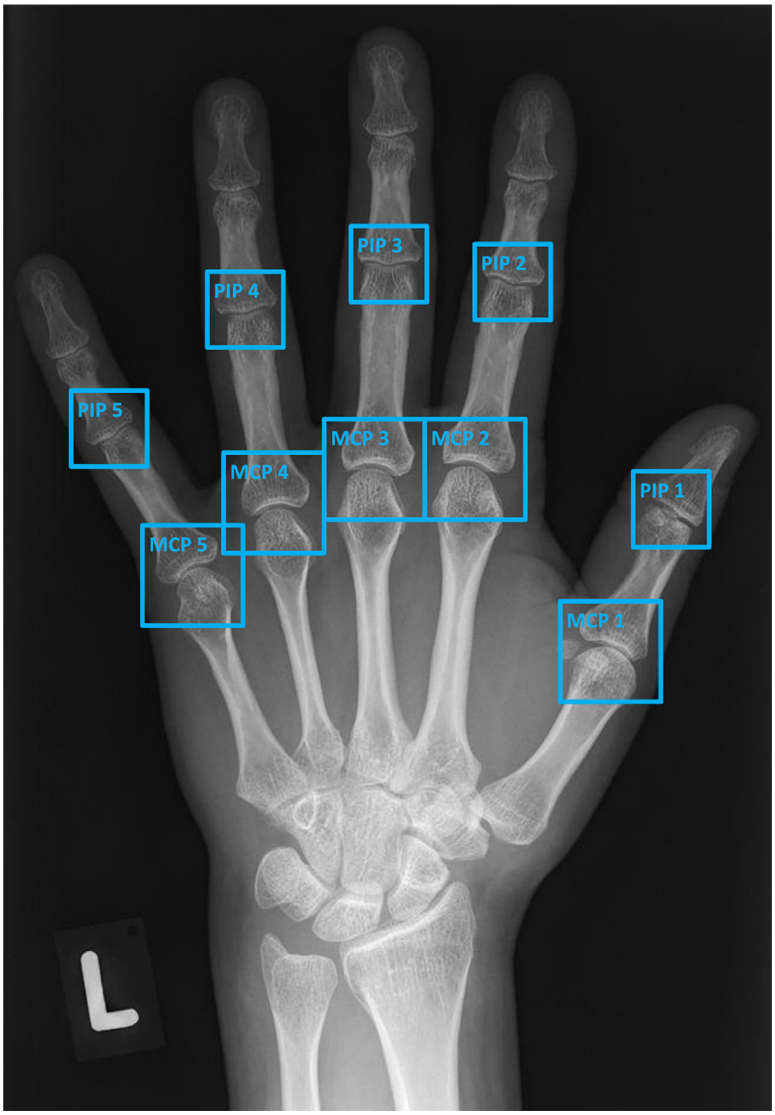


**Fig. 1.** Proximal interphalangeal (PIP) joints and carpometacarpal (MCP) joints of the left hand. The model was trained on cropped images of these ten joints.

**Table 1**
Disease stages according to the Ratingen score [2] used as class labels.

| Stage | Description | Proportion of the data | Number of images |
|---|---|---|---|
| 0 | 0% is eroded (normal joint) | 67.5% | 69,067 |
| 1 | Less than 20% of the joint surface is eroded | 27.2% | 27,853 |
| 2 | 21%–40% of the joint surface is eroded | 2.3% | 2316 |
| 3 | 41%–60% of the joint surface is eroded | 1.0% | 1041 |
| 4 | 61%–80% of the joint surface is eroded | 0.6% | 597 |
| 5 | More than 80% of the joint surface is eroded | 1.4% | 1391 |

in a nominal fashion (ignoring the order) and use them as class labels. In our dataset, these class scores are extremely imbalanced with most being 0, indicating a normal joint surface. Fig. 2 provides a representative image of a joint for each score class.

The data was split into a training set (70%), a validation set (20%), and a test set (10%). The splits were performed such that all images of a single patient are included in the same set. Otherwise, the splitting is random. To calculate inter-rater reliability, a re-scoring set of 308 images was chosen from the test set that includes approximately the same number of images per class. An independent expert then scored these images again.

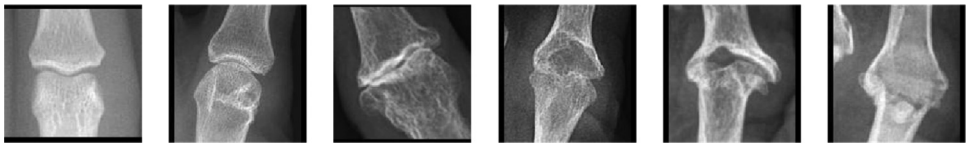

**Fig. 2.** Representative images for the six classes with Ratingen scores of 0 to 5 from left to right.

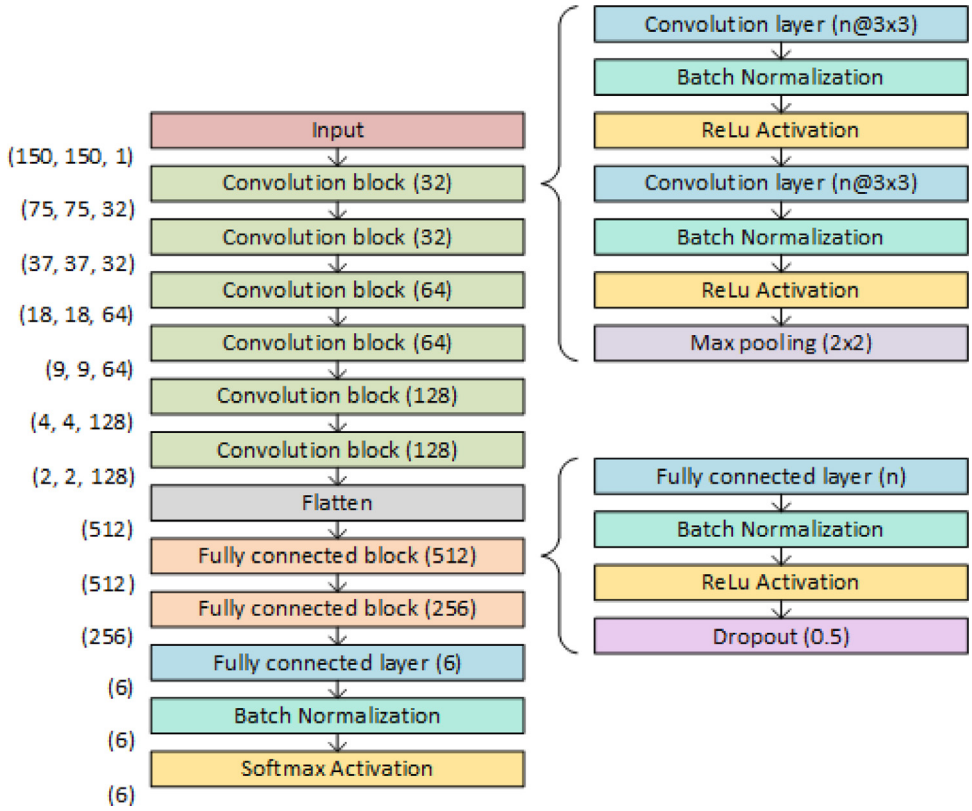


**Fig. 3.** Architecture of the neural network. All convolution blocks are identical with only the number of convolution filters differing. The two fully connected blocks are also identical except for the number of neurons in the fully connected layer.

### 3.2. Deep Learning Model

We use a deep learning model inspired by VGG16 [18] with six blocks of two convolutional layers with filter size 3×3 followed by a max pooling layer. The number of filters per convolutional layer increases with every second block from 32 to 64 to 128. The size of the feature maps decreases due to the max pooling layers and is only 2×2 after the last convolutional layer. Every convolutional layer uses batch normalization before the ReLU activation function. The output of the last convolution block is flattened and used as input to two dense layers which also contain batch normalization, ReLU activation, and dropout. Mathematical formulas and explanations for the success of these methods can, for example, be found in [19].

The last layer of the model is a softmax layer which predicts a probability for each of the six Ratingen scores (0–5) per image. The joint in the image is assigned to the class with the highest predicted probability. The architecture of the CNN is shown in Fig. 3.

The model includes 1,072,818 trainable parameters fitted via the Adam optimizer [18], which is a variant of stochastic gradient descent. To avoid overfitting we used data augmentation during training. For data augmentation the images were rotated randomly by up to ±30°, stretched in width and height by up to 10%, sheared by up to 10%, flipped horizontally and vertically, and zoomed in or out by up to 10%. The only pre-processing step was to normalize the pixel values to the range [0,1]. The model was trained for 250 epochs, each containing an augmented version of all images used as training data. The batch size is 128 and the initial learning rate is $10^{-3}$. This batch size of 128 was chosen because it is large enough to allow for efficient parallelized training but not so large as to decrease the generalization performance [20]. The training curves of our model are shown in Fig. 4.

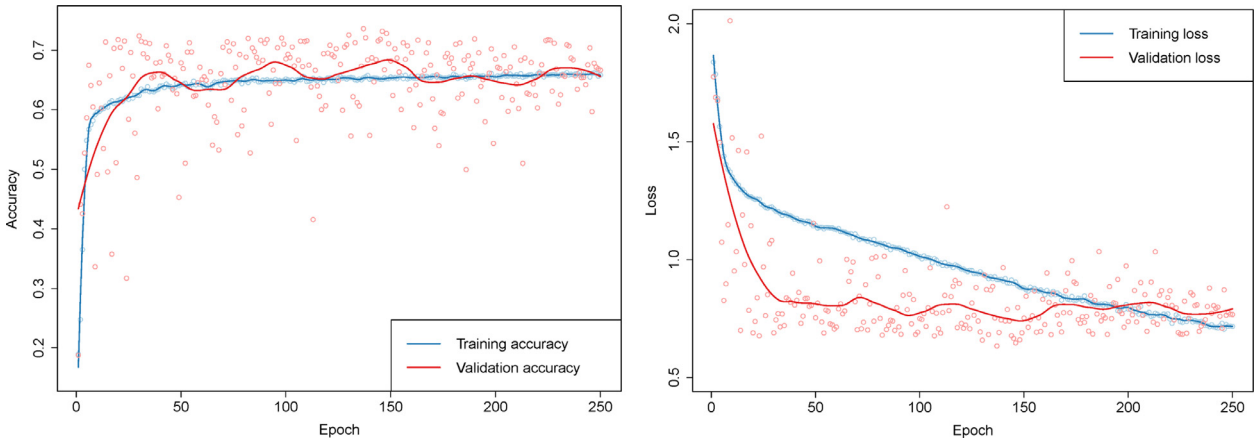


**Fig. 4.** Model accuracy and loss on training and validation set during training.

*3.2.1. Transfer-learning*

For transfer-learning we use a VGG16 model that was pre-trained on the ImageNet dataset [18]. Transfer learning on models trained with the ImageNet dataset has been successfully applied to images in other medical domains [6]. For transfer learning, we first fixed all weights in the convolutional part of the pre-trained VGG16. We replaced the three fully connected layers of the pre-trained model with two fully connected layers of size 512 and 256. Both fully connected layers use batch normalization before a ReLU activation [18]. A layer with six neurons and softmax activation is used as the output layer for our six-class classification problem. Because the pre-trained model requires an image with three channels as input, we replicated our grayscale images three times. First, we froze all convolutional layers and only trained the newly added dense part with the Adam optimizer and a learning rate of $10^{-4}$ for 25 epochs. We then unfroze also the convolutional layers and trained the whole network for an additional 10 epochs with a reduced learning rate of $10^{-5}$. In this step the convolutional filters are fine-tuned, such that they can adapt to the new domain. However, the chosen learning rate is much smaller than the standard Adam learning rate of $10^{-3}$. The small learning rate only changes the weights slightly every step, which ensures that the previously learned weights are not distorted too much or too fast.

### *3.3. Performance metrics and loss functions*

We applied a variety of commonly used metrics to evaluate the performance of the trained models. Since there is no gold standard and human expert raters assign different scores to a given joint, we additionally assessed the performance of our deep learning model in comparison to human experts by metrics used in interrater analysis.

*3.3.1. Metrics for evaluating the model performance*

To assess the prediction performance of a classification model, the observed agreement between predicted and true class label is quantified. In machine learning the best-known measure for agreement is the **global accuracy** $p_{obs}$, which is simply defined as the proportion of correctly classified images in the entire dataset of N images. Having K classes, $p_{obs}$ can be determined from the entries of the confusion matrix $n_{ij}$ by:

$$p_{obs} = \sum_{k=1}^{K} n_{kk}/N \text{ with } N = \sum_{k=1}^{K}\sum_{k'=1}^{K} n_{kk'}$$

where $K$ is the number of classes and $n_{ij}$ is the number of images belonging to class $i$ which are predicted as class $j$.

This metric is widely used in the deep learning context but might provide a misleadingly optimistic impression in imbalanced settings. This is because the classification rate in overrepresented classes dominates the global accuracy. For example, in our case, a classifier classifying all images as class 0 would still yield an accuracy of $p_{obs} = 0.675$, since the 67.5% of images belonging to class 0 are all correctly predicted. Therefore, we also use the **balanced accuracy**, which is the average of the classification rates determined separately for each class. All classes have an equal impact on this metric that is independent of the class size. Finally, since our outcome is an ordinal factor and a misclassification by $\pm 1$ class is acceptable from a medical viewpoint, we also determine the proportion of predictions not more than one class off and call it **±1 balanced accuracy**.

*3.3.2. Loss functions*

All three accuracy metrics discussed above are not differentiable and thus cannot directly be used as a loss function for the optimization procedure. The standard loss function for classification in deep learning is the **cross-entropy** [18]:

$$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} I(y_i = k)\ log(p_k)\cdot w_k$$

where $y_i$ denotes the (true) class of the training example $i$, $p_k$ the probability the model assigns to the $k$-th class, and $I$ the indicator function. For the standard cross-entropy, we set $w_k = 1$. To account for the pronounced imbalances in the Ratingen scores (see Table 1) in our data, we additionally use a **weighted cross-entropy**, where weights $w_k$ for class $k$ are given by:

$$w_k = N/\left(K\cdot\sum_{i=1}^{N} I(y_i = k)\right)$$

This weighting is also used in the software package scikit-learn [21] for the computation of balanced class weights. When computed on our training set, we get the following class weights: {0: 0.25, 1: 0.61, 2: 7.14, 3: 16.67, 4: 30.80, 5: 12.65}. Since the weights are very different (factor 123.2 between class 0 and class 4), we additionally tried logarithmic weights for the loss function. However, this gave worse results with respect to our performance measures. In addition, we also tried an approach where the underrepresented classes were oversampled to achieve a balanced training dataset. However, in prior experiments we observed that for the balanced accuracy the oversampling approach did not yield satisfactory results [22].

*3.3.3. Measures for assessing the agreement of different scorers*

In biostatistics the most prominent measure for agreement is Cohen's kappa [23], which is usually used to assess the agreement between two scorers in a diagnostic test. Cohen's kappa $\kappa$ quantifies the agreement beyond chance. It is given by the scaled difference of the observed agreement $p_{obs}$ (observed accuracy) and the agreement expected by chance $p_{chance}$, i.e., the accuracy derived from a confusion matrix where the elements are given by the product of the row and column marginals. It is common to formulate $\kappa$ based on the observed disagreement $q_{obs} = 1 - p_{obs}$ and the disagreement expected by chance $q_{chance} = 1 - p_{chance}$:

$$\kappa = \frac{p_{obs} - p_{chance}}{1 - p_{chance}} = 1 - \frac{q_{obs}}{q_{chance}}$$

Because the (dis-)agreement by chance depends on the estimated class frequencies, Cohen's kappa also depends on the imbalance of the classes. To additionally take into account the ordering of the classes we use Cohen's **quadratic kappa**. This measure assigns a penalty to each mismatch (corresponding to off-diagonal elements $n_{ij}$ in the confusion matrix) which quadratically increases with the distance between the assigned class levels [23].

$$\kappa_{squared} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left((i-j)^2\cdot n_{obs_{ij}}\right)}{\sum_{i=1}^{k}\sum_{j=1}^{k}\left((i-j)^2\cdot n_{chance_{ij}}\right)}$$

Possible values of $\kappa_{squared}$ are in the range [-1, 1] where -1 denotes complete disagreement and 1 represents complete agreement. A value of 0 indicates no systematic dependencies between the raters.

For more than two scorers the average kappa can be reported, which can be shown to be identical to the **intraclass correlation** ICC [24]. The ICC is commonly used to measure agreement between several scorers when the scores are continuous. The ICC quantifies to which degree the scorers are interchangeable. In the case of two scorers, this is given by the Pearson correlation of the scores after randomly permuting the scores assigned to individual subjects by the two scorers.

## 4. Results and discussion

### *4.1. Effect of the loss function and evaluation of transfer learning*

First, we investigate the performance of the deep learning model trained with the standard and the weighted categorical cross-entropy loss. Fig. 5 displays the resulting confusion matrices for the validation set.

Using instead of the normal categorical cross-entropy the weighted version of the categorical cross-entropy (see Fig. 5b), we observe less predictions to the majority class 1 and more predictions to the underrepresented classes 2-5. The effect of the different loss functions on the performance metrics is summarized in Table 2. The Wald confidence interval is shown in parentheses for a 95% confidence level.

For the task of scoring eroded joints, it is important to distinguish the different Ratingen scores. We therefore disregard global accuracy and focus on balanced accuracy. Considering this metric, the model trained with a weighted categorical cross-entropy loss function achieves the best results. Further, the $\pm 1$ balanced accuracy is also the highest when trained with the weighted loss function. Therefore, we use the weighted categorical cross-entropy in the final model.

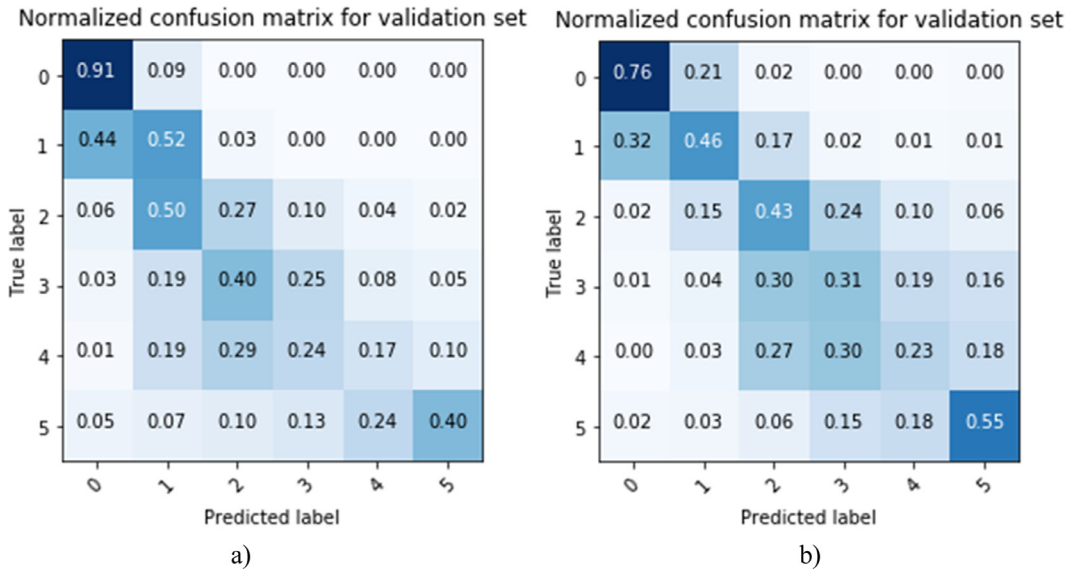


**Fig. 5.** Normalized confusion matrices for the validation set of the model trained with (a) categorical cross-entropy, (b) weighted categorical cross-entropy.

**Table 2**
Comparison of the models trained with different loss functions. The metrics along with 95% confidence intervals are calculated on the validation set.

| Columns: Loss<br>Rows: Metrics | Loss 1: Categorical cross-entropy | Loss 2: Weighted categorical cross-entropy |
|---|---|---|
| Global accuracy | **77.5%** (76.9%, 78.1%) | 66.3% (65.6%, 66.9%) |
| Balanced accuracy | 42.0% (40.0%, 44.0%) | **45.7%** (43.6%, 47.8%) |
| ±1 balanced accuracy | 79.2% (77.1%, 81.4%) | **82.9%** (80.9%, 84.9%) |

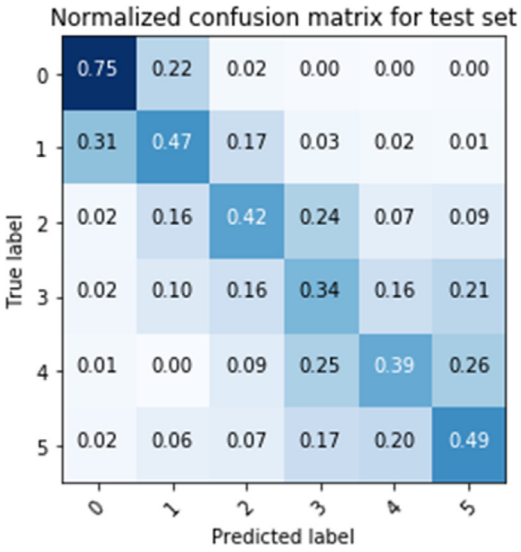


**Fig. 6.** The normalized confusion matrix of our model's predictions for the test set.

We also experimented with transfer learning which in our application achieved a comparable performance but did not lead to a significant improvement. This indicates that, due to the large dataset, we could train a model from scratch allowing it to learn domain tailored features without overfitting.

We now evaluate the performance of our trained network on the test set which was not used so far. The results are shown with the confusion matrix in Fig. 6 which is in good agreement with the confusion matrix of the validation set

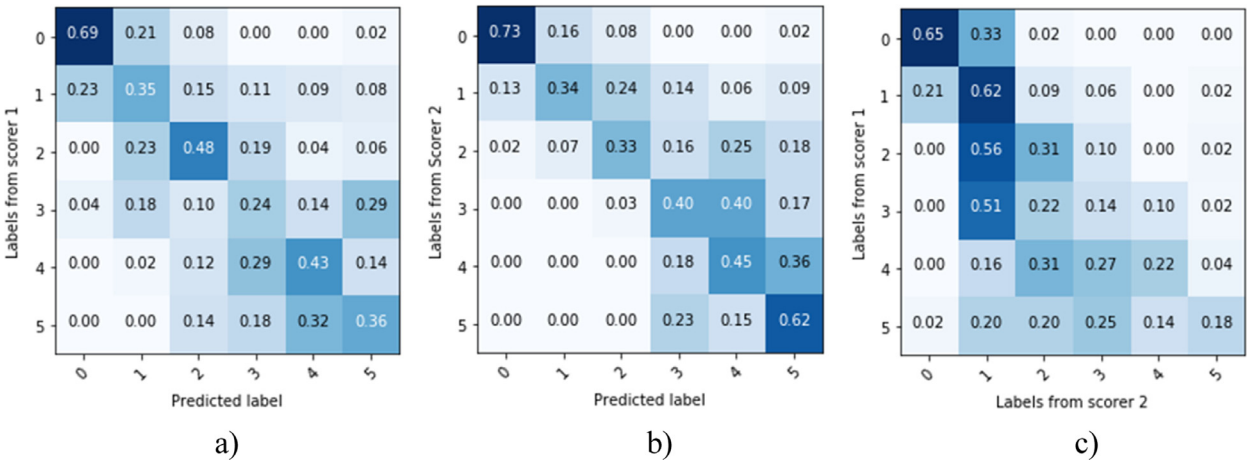


**Fig. 7.** Normalized confusion matrices comparing our model to scorer 1 (a) and scorer 2 (b) as well as a confusion matrix comparing the two scorers (c) for the re-scoring set.

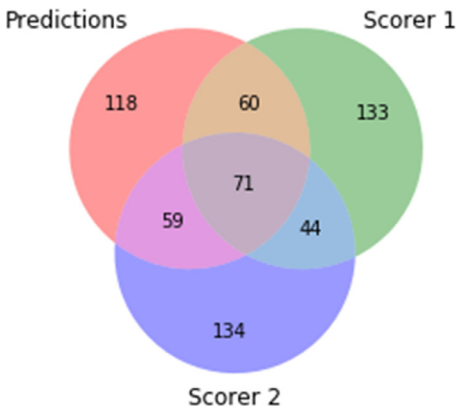


**Fig. 8.** Venn diagram of score overlap between the predictions of our model and the two human scorers.

shown in Fig. 5b. For the test set we achieve a global accuracy of 65.8%, a balanced accuracy of 47.5% and a ±1 balanced accuracy of 83.0%, all comparable with the validation set accuracies displayed in Table 2.

### *4.2. Comparison with human scorers*

To validate the CNN predictions against the human-assigned scores, we analyzed the agreement between the CNN's predictions and the scores given by a human scorer and compared this to the agreement between different human scorers. Scorer 1 denotes the group of experts who labeled the original data set. Scorer 2 is an independent scorer who evaluated images of the individual joints in a rescoring set that were also scored by our deep learning model. The global accuracy of the CNN predictions on the re-scoring set, taking the scores of rater 1 as gold standard, is 42.5% and the balanced accuracy is 42.6%. Notice, that now there is almost no difference between the two accuracies, because in the re-scoring set, all classes include roughly the same amount of observations. The ±1 balanced accuracy is 75.9%.

The global accuracy of the CNN predictions when taking the scores of Scorer 2 as the gold standard was 42.9% with a balanced accuracy of 47.9%. The ±1 balanced accuracy was 79.2%. Interestingly, the global accuracy of the scores of Scorer 2 when taking the scores of Scorer 1 as the gold standard was only 37.7% with a balanced accuracy of 35.6% and a ±1 balanced accuracy of 70.0%. This result indicates that the performance of the CNN prediction is within a comparable range as the performance of the scorer. In Fig. 7 the agreement of the different raters is shown in confusion matrices.

The frequency of agreement is visualized in the Venn diagram of Fig. 8, with the predictions from the CNN model having more overlap with the scores of the two human scorers than between assignments by both human scorers.

To quantify the inter-rater reliability, the quadratic weighted Cohen's Kappa for the three combinations of our model and the two scorers were calculated (results provided in Table 3). The 95% confidence intervals reveal that plausible values of agreement between all pairs of raters are in the range of 0.5 and 0.7, suggesting a fair agreement. There is no relevant

**Table 3**
Cohen's quadratic Kappa with a 95% confidence interval for the three combinations of our model and the two human scorers.

| Compared scores | Cohen's quadratic kappa | 95% Confidence interval |
|---|---|---|
| Model vs. Scorer 1 | 0.675 | (0.608, 0.742) |
| Model vs. Scorer 2 | 0.580 | (0.505, 0.655) |
| Scorer 1 vs. Scorer 2 | 0.562 | (0.487, 0.638) |

difference between the inter-rater reliability between our model versus Scorer 1 or Scorer 2, nor between the two scorers. This result suggests that our model can predict the Ratingen scores as well as a trained expert.

To perform a classical inter-rater reliability analysis for three raters, we calculate the ICC. An ICC of 0 indicates no agreement between raters, whereas an ICC of 1 means complete agreement. The ICC for our re-scoring set is 0.601 with a 95% confidence interval of (0.543, 0.656), indicating, according to Cicchetti [25], a "fair" agreement between all three raters. This provides evidence that the three raters (two human scorers and the DL model) do not disagree with each other. The null hypothesis of ICC=0 is rejected with a p-value $= 2.26 \times 10^{-72}$.

## 5. Conclusions

Automatically determining joint scores in rheumatic arthritis is a challenging problem: the scores are highly imbalanced and are ordinal data. Further, due to the limited agreement between different human scorers, there is no real ground truth.

Due to the high imbalance, the global accuracy of 77.5% is dominated by the overrepresented classes 0 (healthy joints), which accounts for 67.5% of all joints. We take imbalance into account by weighting the loss function and by considering a balanced accuracy, assigning the same importance to all 6 classes. This results in a balanced accuracy of 47.5%. While this result seems to be poor on first sight, we should acknowledge the ordinal nature of the scores. By allowing the classifier to be off by $\pm 1$ score, we take the ordinal nature into account, resulting in a $\pm 1$ balanced accuracy of 83%. Finally, we need to take into account the large label noise. Comparing scorers assigned by different human experts, we find a quite small agreement, quantified by a Cohen's kappa of 56.6%. It is difficult to train and validate a model on such inconsistent data because it is not clear which given label, if any, is correct. Still, the inter-rater reliability scores between the model and the 2 human experts are 0.675 and 0.580 and thus higher than between the human scorers. This clearly indicates that our model is (at least) as good as a human expert.

The main advantages of our deep learning approach compared to human scorers are the acceleration of the scoring process by several orders of magnitudes, i.e. milliseconds instead of minutes, and the reproducibility of the results (i.e., same images always get the same scores). We are also convinced that our model will further improve in the future when an extended data set is available for training.

## Declaration of Competing Interest

The SCQM Foundation is financially supported by pharmaceutical industries and donors. A list of financial supporters can be found on www.scqm.ch/sponsors. A list of rheumatology offices and hospitals that are contributing to the SCQM registries can be found on www.scqm.ch/institutions.

Tobias Reinhard works for Seantis GmbH.

## Acknowledgments

We thank the SCQM Foundation and all participating patients and rheumatologistics for providing the data for this study. Further, we would like to thank Christoph Molnar for early contributions to the work, Fabian Reinhard for the coordination, SCQM (Caroline Ensslin) for the rescoring, Christa Reinhard for the manual image labeling and Almut Scherrer for the feedback. We thank Kelly Reeve for carefully proofreading the manuscript.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compeleceng.2019.08.003.

## References

[1] Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet 2016;388:10055 (2023-2038). doi:10.1016/S0140-6736(16)30173-8.
[2] Rau, R. and Wassenberg, S. Scoringmethoden bei der rheumatoiden Arthritis. In: Bildgebende Verfahren in der Rheumatologie. Ed. by Deutsche Gesellschaft für Rheumatologie. Steinkopff, 2007. Chap. 2, pp. 27–46. https://doi.org/10.1007/978-3-7985-1721-9_2.
[3] Murakami S, Hatano K, Tan J, Kim H, Aoki T. Automatic identification of bone erosions in rheumatoid arthritis from hand radiographs based on deep convolutional neural network. Multimed Tools Appl 2018;77(9):10921–37.

[4] Snekhalatha U, Anburajan M. Computer-based measurements of joint space analysis using metacarpal morphometry in hand radiograph for evaluation of rheumatoid arthritis. Int J Rheum Dis 2017;20(9):1120–31.
[5] Langs G, Peloschek P, Bischof H, Kainberger F. Automatic quantification of joint space narrowing and erosions in rheumatoid arthritis. IEEE Trans Med Imaging 2009;28(1):151–64.
[6] Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 2018;8:1727. https://doi.org/10.1038/s41598-018-20132-7.
[7] Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One 2017;12(6):e0178992. https://doi.org/10.1371/journal.pone.0178992.
[8] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak GAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
[9] Dürr O, Sick B. Single-cell phenotype classification using deep convolutional neural networks. J Biomol Screen 2016;21(9):998–1003.
[10] Goceri E, Goceri N. Deep learning in medical image analysis: recent advances and future trends. In: 11th Int. Conf. on Computer Graphics, Vis., Computer Vision and Image Processing (CGVCVIP 2017); 20-23 July 2017. p. 305–11.
[11] Zhang Yu-Dong, Pan Chichun, Sun Junding, Tang Chaosheng. Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU. J Comput Sci 2018;28. doi:10.1016/j.jocs.2018.07.003.
[12] Yonekura A, Kawanaka H, Prasath VS, Aronow BJ, Takase H. Automatic disease stage classification of glioblastoma multiforme histopathological images using deep convolutional neural network. Biomed Eng Lett 2018;8(3):321–7.
[13] Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia Comput Sci 2016;90:200–5.
[14] Wang S, Muhammad K, Hong J, Kumar A, Zhang Y-D. Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization. Neural Comput Appl 2018. doi:10.1007/s00521-018-3924-0.
[15] Dürr O, Murina E, Siegismund D, Tolkachev V, Steigele S, Sick B. Know when you don't know: a robust deep learning approach in the presence of unknown phenotypes. Assay Drug Dev Technol 2018;16:343–9. doi:10.1089/adt.2018.859.
[16] Dong Q, Gong S, Zhu X. Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 2018.
[17] Uitz E, Fransen J, Langenegger T, Stucki G. Clinical quality management in rheumatoid arthritis: putting theory into practice. Swiss Clinical Quality Management in Rheumatoid Arthritis. Rheumatology (Oxford) 2000;39(5):542–9.
[18] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.
[19] Goceri E. Formulas behind deep learning success. In: Int. Conf. on Applied Analysis and Mathematical Modeling (ICAAMM2018); 2018. p. 156. pgJune 20-24.
[20] Goceri E, Gooya A. On the importance of batch size for deep learning. Int.Conf. on Mathematics (ICOMATH2018), an Istanbul meeting for world mathematicians, minisymposium on approximation theory & minisymposium on math education, 3-6 July Istanbul, Turkey; 2018.
[21] Pedregosa F, Varoquaux G, Gramfort A, Michel V. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[22] Rohrbach, J., Dürr, O., Sick, B. Project thesis: arthritis net, automated bone erosion scoring with deep convolutional neural networks. https://github.com/janickrohrbach/arthritis-net/blob/master/doc/project.pdf.
[23] Cohen J. Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213–20. PMID 19673146 https://doi.org/10.1037/h0026256 .
[24] Koch GG. Intraclass correlation coefficient. In: Kotz S, Johnson NL, editors. Encyclopedia of statistical sciences. New York: John Wiley & Sons; 1982. p. 213–17. 4.
[25] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284–90. doi:10.1037/1040-3590.6.4.284.

**Janick Rohrbach** graduated from Zurich University of Applied Sciences with a BSc in Engineering and Management. Currently, he is working as a research assistant at the institute of Data Analysis and Process Design and pursuing a MSc in Engineering specializing in Data Science. His research interests include machine learning and deep learning in various applications.

**Tobias Reinhard** received his Ph.D. in Computer Science from the University of Zurich in Switzerland. He is a co-founder of Seantis GmbH, where he develops software to bridge the gap between clinical practice and medical research.

**Beate Sick** is a professor for applied statistics at Zurich University of Applied Science and co-affiliated at University Zurich. After her PhD in experimental physics at ETHZ, she turned to biostatistics and was responsible for the bioinformatics at the DNA Array facility of UNIL and EPFL. Currently, her main research focus is on leveraging deep learning approaches for medical research.

**Oliver Dürr** is a professor of data science at the Konstanz University of Applied Sciences. After his PhD in theoretical physics, he worked for 10 years in a bioinformatics company developing and applying machine learning and statistical methods to all kind of -omics data. He is now working mainly on deep learning.