Automated Grading of Modic Changes Using CNNs – Improving the Performance with Mixup

Dimitrios Damopoulos[1](✉), Daniel Haschtmann[2], Tamás F. Fekete[2], and Guoyan Zheng[1]

[1] Institute for Surgical Technology and Biomechanics,
University of Bern, Bern, Switzerland
dimitrios.damopoulos@istb.unibe.ch

[2] Schulthess Clinic, Spine Center, Zürich, Switzerland

Abstract. We propose a method for automated grading of the vertebral endplate regions according to the Modic changes scale based on the VGG16 network architecture. We evaluate four variations of the method in a standard 9-fold cross-validation study setup on a heterogeneous dataset of 92 cases. Due to the very weak representation of the Modic Type III in the dataset, we focus on the grading of Modic Type I and Modic Type II. Despite the relatively small size of our dataset, the pipeline demonstrated a performanc1e that is similar to or better than those achieved by the state-of-the-art methods. In particular, the most performant variant achieved an accuracy of 88.0% with an average-per-class accuracy of 77.3%. When the method is used as a binary detector for the presence or not of Modic changes, the achieved average-per-class accuracy is 92.3%. Our evaluation also suggests that the so-called mixup strategy is particularly useful for this type of classification task.

Keywords: Modic changes · Automated grading · Mixup · VGG

1 Introduction

The term *Modic changes* (MCs) refers to specific patterns of intensity variation in the signal of the T1 and T2 MR scans of the spine, occurring in the bone marrow region around the vertebral endplates. They were first mentioned in 1987 [1] and they are then named after the first author of [2, 3], where three types of such patterns were defined and their possible association with degenerative disk disease (DDD) was discussed.

Specifically, a *Type I Modic change* is defined as the presence of a bone morrow region which has a lower intensity than its surrounding tissue in a T1 scan and a higher intensity in a T2 scan, indicating a bone marrow oedema. In a *Type II Modic change*, the intensity of the region is higher than its surrounding tissue in both the T1 and T2 scans, indicating local fatty degeneration. Finally, in *Type III Modic change* the intensity is lower in both the T1 and T2 scans, representing sclerotic changes of the endplates that result in low signal in both sequences. For brevity, we will refer to these grades as *MC-I*, *MC-II* and *MC-III* respectively. Figure 1 shows a representative example for MC-I and MC-II from the dataset of our study.

G. Zheng et al. (Eds.): CSI 2018, LNCS 11397, pp. 41–52, 2019.
https://doi.org/10.1007/978-3-030-13736-6_4

MCs are considered to be clinically important, especially MC-I and MC-II. There has been evidence suggesting a correlation between the presence of these two MC types (especially MC-I) and low back pain (LBP) [2, 4–7], however the etiology of the MCs and how they are linked to either LBP or DDD remains an active research topic [4, 9]. Research related to MCs is complicated by the subjectability of the MC grading to inter-rater disagreement [10]. The variability in grading can be reduced substantially if the grading process is more strictly standardized and when the raters get more experienced in the task [4, 8, 10]. However, both of these conditions need time to be satisfied. Moreover, grading every case manually on a large dataset is a time-consuming process.

In this study, we propose a pipeline for the automated grading of the endplate regions around the intervertebral disks (IVDs) of the lumbar spine according the MC scale. As a component of a computer aided diagnosis system, we envision that such a method can be useful in a clinical setting by automatically pinpointing IVD regions in an MRI which might require further attention by the clinician, as possible sources of LBP. Furthermore, it can facilitate the conduction of large population studies related to MCs, as it can minimize the tedious task of annotating manually the large number of cases typically stored in a PACS system.

Despite the small size of the dataset that was used, we were able to achieve a MC-detection rate which is on par with the reported performances of human raters. A highlight of the presented work is the application of the mixup strategy [16], which we found to be effective for this particular type of classification task. In particular, we make the following contributions:

(a) We present a learning-based method for the automated grading of MCs, reporting an accuracy which is better than that of the other published work on this task [12]. The dataset that was available in the present study consisted of 92 cases, as opposed to 388 of [12].

(b) We present a successful application of the mixup strategy as introduced in [16] for data augmentation. Mixup appears to be well-suited for this problem, assisting us in coping with the inherent imbalance of this classification task.

1.1 Related Work

A method for the automated detection of MCs was first proposed in [11], which requires the manual segmentation of the IVDs, it consults only one T2 slice for the detection and it performs only binary classification (presence or not of a MC). The first complete system for automatic MC grading was proposed in [12], which also includes an automatic module for the localization of the vertebrae and their corners, making the whole pipeline fully automatic. Their proposed system achieved a 87.8% classification accuracy. The same authors proposed a multi-task CNN architecture in [14], yielding impressive performance in a variety of spine-related computer-aided diagnosis tasks. One of them was the detection of bone marrow defects of the upper and lower vertebrae of IVD regions, which they were able to detect with an accuracy of 91.0% and 90.3% respectively with their best performing models, approaching their intra-rater accuracy. These defects appear to be very similar to MCs, however they are not the same, since they are graded after consulting T2 slices only. The datasets used in the latter two works were rather extensive, consisting of the annotated scans of 388 patients in the case of [12] and of 2009 patients in the case of [14].

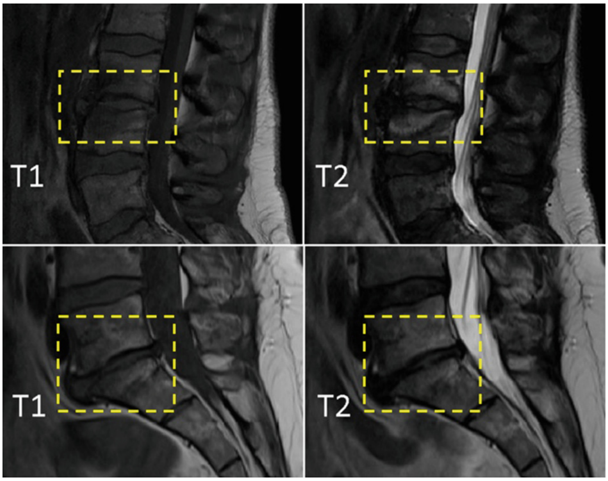

Fig. 1. Two characteristic cases of MCs from the dataset of the present study. The top T1 and T2 slices depict a MC-I case at level L3/L4 and the bottom a MC-II case at level L5/S1. The affected regions are highlighted with yellow rectangles. For the case of MC-I, the affected bone marrow region is visibly hypointense on T1 and hyperintense on T2, whereas the for the MC-II case the bone marrow is hyperintense on both modalities. (Color figure online)

2 Method

The aim of the present study is to automatically grade a given pair of T1 and T2 sequences of regions of interest that capture the upper and lower marrow regions around a particular IVD according to the MC grading system. We will refer to the two sequences of regions of interest as *IVD volumes* and to their individual slices are *IVD regions*. We restrict our attention to the following six spinal levels: T12/L1, L1/L2, L2/L3, L3/L4, L4/L5 and L5/S1. For the rest of the discussion, we will refer to these as *IVD levels*. Example IVD volumes are illustrated in Fig. 2.

As mentioned earlier, there are three MC types defined. However, in the dataset that was utilized in this study, the number of cases with MC-III was very limited. Due to this limitation and also because of the limited clinical significance of MC-III, it was decided to restrict the set of the grades to the first two types, i.e. MC I and MC II. For convenience, we will also refer to the absence of any MC on some IVD level as *MC-0* (background class). Thus, there are $M = 3$ classes in total, MC-0, MC-I and MC-II.

2.1 Isolation of the IVD Regions

The input to the pipeline is a sequence of T1-weighed and a sequence of T2-weighted sagittal MR scans of a lumbar spine. Additionally, it is assumed that a prior localization step has taken place that can provide estimations for the centers of the depicted IVDs, their orientations and their widths. We are interested only in the projection of these elements on the sagittal plane, therefore, for each IVD, we assume the availability of: (a) its 2D center on the sagittal plane and (b) a 2D vector, whose angular displacement represents its orientation and its length is equal to the width of the IVD.

The supplied information is utilized for the extraction of rectangular IVD regions. The extracted IVD regions are centered around their corresponding 2D IVD center, they are parallel to the identified orientation and their size is proportional to the identified width, with their aspect ratio set to 2:1. For the extraction of the IVD volume, this operation is carried out on five of the slices of the input T1 sequence and five of these slices of the input T2 sequence, symmetrically around their midsagittal slice. The same center, orientation and width are used for the extraction of all the 10 IVD regions of the two IVD volumes. In [12], a rigid registration step was also employed for the extraction of the IVD regions in order to account for the small possible movement of the patient between the acquisition of the two sequences. In the present work, no further attempt is made to register one of the two modalities to the other. Finally, the intensity of the extracted regions is rescaled linearly to the 0-255 range. The result of this region extraction stage is illustrated in Fig. 2.

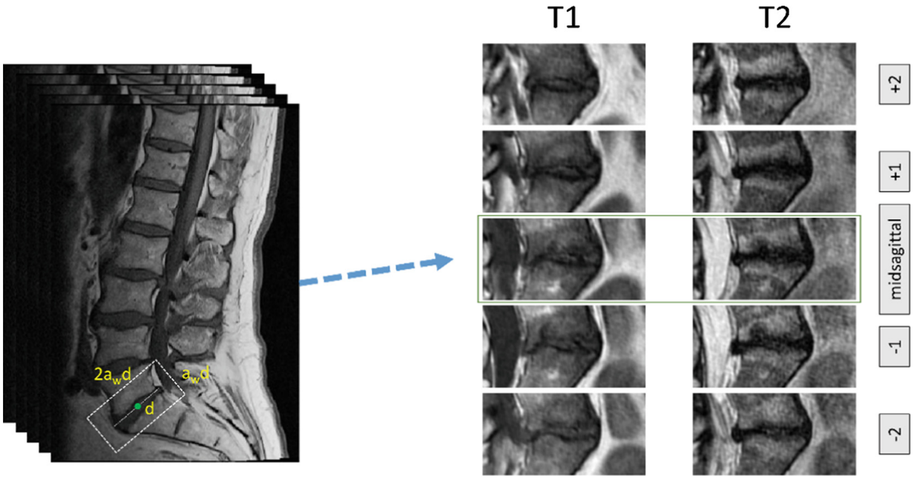

Fig. 2. Schematic illustration of the procedure for the extraction of the IVD volumes. The width of the IVD, its center and its orientation are given as input (left). Then, a region of aspect ratio 2:1 is extracted from five T1 and five T2 slices, symmetrically around the midsagittal slice (the two slices from the left side of the midsagittal, the midsagittal and the two slices from its right side). The result of this procedure is 10 aligned IVD regions, as shown on the right.

2.2 Network Architecture

A variant of VGG16 CNN of [15] is employed, with the following modifications to the vanilla architecture: (a) The size of the input layer is 112 × 224 × 6; (b) A dropout layer is added after the last convolutional layer and (c) there is only one fully-connected layer before the softmax classification. The six channels of the input volume consist of three IVD regions extracted from the T1 slices and the corresponding three IVD regions from the T2 slices, in that order. An illustration of the employed architecture is presented in Fig. 3.

The weighted cross entropy is used as a loss function in all the conducted experiments. Due to the imbalanced representation of the classes in the dataset, the contribution of every training sample to the loss function is weighted according to its class. These class weights are set to be inversely proportional to the frequency of the respective class in the training set.

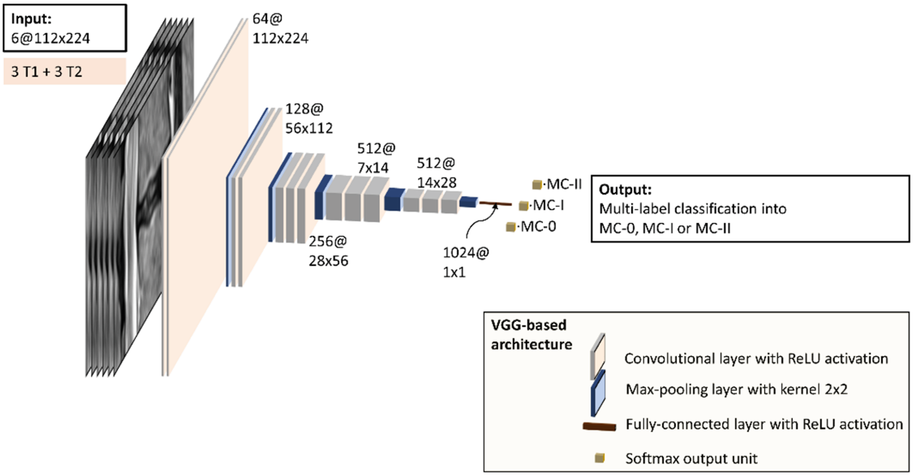

Fig. 3. Schematic illustration of the VGG16 architecture of [15] as employed in the single-stage variant of the present work. The input to the CNN consists of six channels of size 112 × 224, populated with the IVD regions that are extracted from the three T1 and and three T2 slices. The output is a softmax layer of three units, corresponding to the grades MC-0, MC-I, MC-II. Each layer of the architecture is represented with a rectangle, whose width is proportional to the number of feature maps in the layer. For the two-stage variant, the CNN architectures of the first and the second stage differ only on the final classification layer (two output units instead of three).

2.3 Mixup

An important challenge in this study is the lack of a satisfactory number of cases with a non-background label. Only 123 IVD regions (22.3% of the total number) in our dataset have a label which is not MC-0, with 84 of them labeled as MC-II. We attempt to partially address this problem by using the so-called *mixup* strategy, introduced in [16].

The mixup approach was motivated in [16] by the desire to reduce the oscillating predictive behavior of a trained classification model when it encounters samples that fall outside of its training set. The basic idea is the following: Given two training samples $S_1 = (\boldsymbol{x}_1, \boldsymbol{y}_1)$ and $S_2 = (\boldsymbol{x}_1, \boldsymbol{y}_2)$ where $\boldsymbol{y}_1, \boldsymbol{y}_2 \epsilon [0,1]^M$ are 1-hot vectors and M is the number of classes in the problem, a new training sample S_m is formed by a linear interpolation of S_1, S_2:

$$S_m = (\boldsymbol{x}_m, \boldsymbol{y}_m) = (\lambda \cdot \boldsymbol{x}_1 + (1-\lambda) \cdot \boldsymbol{x}_2, \lambda \cdot \boldsymbol{y}_1 + (1-\lambda) \cdot \boldsymbol{y}_2), \lambda \in [0,1] \tag{1}$$

Or more concisely:

$$S_m = \lambda \cdot S_1 + (1-\lambda) \cdot S_2, \lambda \in [0,1] \tag{2}$$

Where the weight λ is a random variable. When λ is very close to either 0 or 1, S_m will very similar to one of the original training samples, whereas values near 0.5 lead to maximum blending. Following [16], λ is drawn from a beta distribution, giving flexibility on specifying how aggressively new mixup samples are formed. An illustration of the mixup procedure is shown in Fig. 4.

Although after the application of mixup the target $\boldsymbol{y}_m$ is no longer an 1-hot vector, it can still be treated as a probabilistic distribution. Indeed, let $\boldsymbol{y}_1 = \left[y_{1,1}, \cdots y_{1,M}\right]$, $\boldsymbol{y}_2 = \left[y_{2,1}, \cdots y_{2,M}\right]$, $\boldsymbol{y}_m = \left[y_{m,1}, \cdots y_{m,M}\right]$ be the elements of the target vectors we interested in. The original targets $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are 1-hot vectors, so $\sum_{i=1}^{M} y_{1,i} = \sum_{i=1}^{M} y_{2,i} = 1$. Then:

$$\sum_{i=1}^{M} y_{m,i} = \sum_{i=1}^{M} \lambda \cdot y_{1,i} + (1-\lambda) \cdot y_{2,i}$$

$$= \lambda \cdot \sum_{i=1}^{M} y_{1,i} + (1-\lambda) \cdot \sum_{i=1}^{M} y_{2,i} = \lambda + (1-\lambda) = 1$$

Also, the elements $y_{m,i}$ are all positive since $\lambda, (1-\lambda) > 0$. Therefore, the target $\boldsymbol{y}_m$ can be treated as a discrete probability distribution over the M classes. This is important because it allows us to continue using the cross-entropy as a loss function for training the network, which is the standard choice for classification tasks.

In practice, mixup can be understood as a data augmentation method [16] and it can be implemented with minimal modifications to the standard training pipeline. In particular, for every training mini-batch, a random permutation of it is constructed and the λ values are sampled from the beta distribution. Then, the original mini-batch and its permuted version are multiplied with λ and $(1-\lambda)$ and they are added together to form the mixup mini-batch.

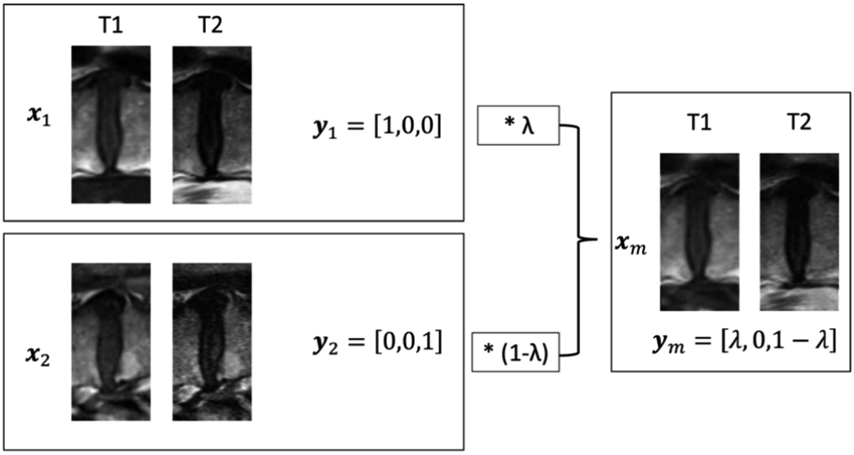

Fig. 4. Illustration of the generation of the mixup training samples for a 3-class classification scenario (MC-0, MC-I, MC-II). On the left, two training samples from the original mini-batch with labels MC-0 and MC-II (top to bottom). The two training samples are linearly interpolated with weights λ and $(1 - \lambda)$ to form a mixup training sample. In this example, $\lambda = 0.769$. The entire IVD volumes are interpolated (for convenience, only one pair of T1 and T2 slices are shown on the figure).

2.4 IVD Level Grading

A straightforward approach for predicting the label of a certain IVD level would be to pass the isolated T1 and T2 IVD volumes (of five slices each) to the network and use the prediction of the network as the prediction for the MC of that level. In practice however, we found out that the accuracy is improved if narrower volumes are used as input. The adopted strategy is as follows: firstly, from the original T1 and T2 IVD volumes of five slices, three sub-volumes are constructed with three consecutive IVD regions each. The constructed T1 and T2 sub-volumes are combined, forming three volumes of size 112 × 224 × 6. Then, these three combined volumes are passed in succession to the network in order to get one prediction for each of them. If all of these predictions are MC-0, the assigned grade for this particular IVD level is MC-0. Otherwise, it is the grade that corresponds to the most confident prediction.

Such an approach can be justified on the grounds that it mimics the process that a human rater is following when rating a given pair of T1 and T2 sequences: One examines one T1 slice and one T2 slice at a time, in order to assess whether MC intensity patterns are present or not. The two adjacent slices are also taken into account in order to decide whether any pinpointed pattern is consistent with the presence of a MC or it is an unrelated artifact. If it is decided that a MC is present, this is enough to assign a MC grade to whole IVD level, even if the detected pattern is not visible throughout the sagittal length of the endplate.

3 Experimental Design and Results

3.1 Dataset

The dataset of this study consists of a fully anonymized dataset of 92 pairs of T1 and T2 sequences. These sequences were acquired using a variety of protocols, including fat-suppressed T1 and T2, some of them employing the Dixon method. The inclusion criteria were the following: (a) The entire sacrolumbar region should be visible; (b) any deformities of the spinal curvature should be limited enough for a midsagittal slice to still be definable and (c) no implants should be present on the lumbar region of the spine. All of these sequences were sagittally acquired, with the total number of slices per sequence being in the 9–17 range.

The endplate regions of six spinal levels from T12/L1 to S1/L5 of every case were rated from two spine surgeons according to the MC grading system with every case being rated by exactly one rater. The acquisition of the ground truth was guided by the following criteria:

(a) Only the five slices closest to the midsagittal one were taken into consideration during grading;
(b) Only intensity changes of the bone marrow that extend from an endplate were graded as a MC;
(c) The MC pattern must be visible in at least two adjacent sagittal slices for an endplate region to be graded as MC-I or MC-II.

Similar criteria have been used in literature in order to standardize the annotation process [4, 8, 10]. The localization of the IVDs that is required for the extraction of the IVD regions was performed with the help of a manual, approximate segmentation of the IVDs. The orientation of each IVD was given by the first component of a principal component analysis (PCA) on the corresponding binary segmentation mask of the IVD and the IVD center by the centroid of the mask.

3.2 Evaluation

The dataset was split in 9 folds; 8 folds have the sequences of 10 patients and one fold of 12 patients. A standard 9-fold cross-validation study was then conducted and the achieved accuracy was compared with the annotations provided by the experts. Therefore, every case participated exactly once in the study as a member of a testing fold, permitting the computation of the evaluation metrics on the whole dataset.

The principal evaluation metric that was used is the *average-per-class accuracy* (APCA), which is especially suited for unbalanced classification tasks. In particular, let $ACC_{MC-0}, ACC_{MC-I}, ACC_{MC-II}$ by the achieved accuracies for the classes MC-0, MC-I and MC-II respectively. Then, the APCA will be:

$$ACC_{APC} = \frac{ACC_{MC-0} + ACC_{MC-I} + ACC_{MC-II}}{3} \tag{3}$$

We also record the accuracy of detecting whether certain degree of MC is present, i.e. the accuracy on the union of the MC-I and MC-II classes. We will denote this measurement with ACC_{MC}. The APCA for this binary classification task is:

$$ACC_{APC,bin} = \frac{ACC_{MC-0} + ACC_{MC}}{2} \tag{4}$$

Finally, the total, unweighted accuracy ACC is also reported, i.e. the rate of the correct automatic classifications.

3.3 Multiclass Classification vs. Two-Stage Classification

In addition to the classification pipeline as presented in the previous section, we evaluated an alternative scheme where the multiclass CNN classifier is replaced by two binary classifiers, assembled in a two-stage classification fashion. In particular, the first-stage binary classifier makes a prediction on whether the given IVD volume has a MC grade of MC-0 or not. If the first-stage classifier does not detect a MC-0 grade, the second-stage classifier further classifies the same IVD volume into MC-I or MC-II. Except for the final softmax layer, both of these binary classifier share exactly the same architecture as the multiclass classifier, including the size of the input IVD volume. Both pipelines were evaluated with and without the application of mixup strategy.

3.4 Hyperparameters

The values of the hyperparameters were set using two splits of a subset of 78 cases of the dataset into 70 training and 8 testing cases. The width of the extracted IVD regions was set to be 1.7 times the width of the IVD. The parameter alpha of the beta distribution of mixup was set to 0.1 for all experiments. Mixup was applied indiscriminately to all the MC classes, thus all the combinations of MC classes were possible during the creation of the mixup mini-batch. The dropout rate was set to 0.2. The mini-batch for the training of the multiclass classifier and of the first-stage classifier was formed from the IVD volumes of three cases of the training set. The size of the mini-batch of the second-stage classifier was set to six IVD volumes, drawn from all the IVD volumes of the training set with a non-MC-0 label. The weights of all the network were initialized from a VGG16 network pre-trained on ImageNet [17]. The networks were trained for 40 epochs when mixup strategy was not used and for 80 epochs when mixup strategy was used. The difference in the number of epochs was due to our observation that, when mixup is applied, more epochs are need for the training error to reach the same level (this observation agrees with [16]). When mixup was not used, the performance seemed to actually get worse when the network was trained to the same number of epochs (80), hence we decided to keep it much lower (40), in an attempt to make the comparison fair. Furthermore, as in [14], we noticed that the increased network capacity offered by the two additional fully-connected layers of the default VGG16 architecture hurt the performance, therefore we kept only one fully-connected layer before the softmax layer.

Similar to [14], extensive training-time data augmentation was applied: rotation of the IVD regions by ±7.5°, change of their scale by a factor of 0.8–1.2, displacement of their center by ±5 pixels in the coronal and axial dimensions, random flipping in the coronal direction with a probability of 0.5 and swapping of the order of the IVD regions in a volume with a probability of 0.5.

3.5 Results

Four variations of the proposed method were evaluated, corresponding to the four configuration combinations of using/not-using mixup and for multiclass/two-stage classification. The achieved evaluation scores are reported in Table 1. For the case of the two-stage classification, the ACC_{MC-0} and ACC_{MC} metrics depend only the performance of the first-stage classifier. On the other hand, the ACC_{MC-I} and ACC_{MC-II} metrics depend on both on the accuracy of the first-stage on detecting MCs and on the ability of the second-stage classifier to discriminate between MC-I and MC-II.

From this table, we can make some observations: firstly, the application of mixup resulted in an improvement in five out of the six recorded evaluation metrics, both in the multiclass and in the two-stage classification scheme. The metric that got worse in both scenarios was the accuracy on the MC-0 class (ACC_{MC0}). These results leave the impression that mixup improves the accuracy on the underrepresented classes, at the modest expense of the most common class.

The second observation is that the two-stage classification scheme seems to be performing better than the multiclass one. Even though ACC_{MC-II} got worse in the two-stage pipeline, leading to a worse ACC_{MC} too, the APCA is much higher (both the binary and the multiclass one), as well as the total accuracy.

Table 1. The achieved accuracies of the four variation of the pipeline (with and without mixup, multiclass vs. two-stage classification). All the shown values are percentages over the whole dataset. The number of IVD levels with labels MC-0, MC-I and MC-II is 429, 39, 84 respectively. The best values are highlighted with bold font.

	MC-0	MC-I	MC-II	MC-I + MC-II	APCA	APCA binary	Accuracy
Multiclass classifier							
No mixup	90.4	41.0	78.6	88.6	70.0	90.0	85.1
Mixup	88.8	59.0	**81.0**	**94.3**	76.2	91.6	85.5
Two-stage classification							
No mixup	**93.2**	51.3	71.2	87.8	71.6	90.5	86.8
Mixup	92.8	**64.1**	75.0	91.9	**77.3**	**92.3**	**88.0**

4 Discussion and Conclusion

This paper proposed a method for the automated detection of MC-I and MC-II in IVD regions. The four variants of the proposed method were evaluated in a standard 9-fold cross-validation setup with a heterogeneous dataset of 92 cases. The evaluation demonstrated the usefulness of the recently proposed mixup strategy for this type of classification task. Interestingly, a two-stage classification scheme achieved a generally better performance than a multiclass classification approach.

Despite the relatively small size of the dataset used, the proposed method seems to achieve a performance that is similar to or even better than those achieved by the state-of-the-art methods. In particular, the most performant variant achieved an accuracy of 88.0%, which compared favorably to the 87.8% accuracy of [12], the only other published method for the automatic MC multiclass grading. However, a direct comparison is difficult to make, since our dataset is different and smaller compared to the one used in [12] and we also opted to omit the MC-III grade from our study. On the other hand, when the proposed method is used as a binary detector for the presence or not of MCs, the achieved performance is very good, with an average-per-class accuracy of 92.3%, for the most performant variant.

Despite the success of the method in detecting the presence of MCs, their classification into MC-I or MC-II was proved to be a more challenging task for the present method: the average-per-class-accuracy of 77.3% leaves much to be desired. The accuracy on the MC-I class was particularly low (64.1%), likely related to the small number of cases with such a grading on our dataset (39 in total).

As part of future work, we plan to collect and annotate additional cases, since we feel that this is a limiting factor of the current study. A larger dataset could hopefully allow us to consider MC-III in our study. Even though MC-III is clinically less significant, the etiology of MCs is still not well understood and an automated system for the identification of all the recognized MC types would be beneficial to MC-related research. From a technical standpoint, it would be also interesting to see if mixup remains effective on a larger dataset.

References

1. De Roos, A., et al.: MR imaging of marrow changes adjacent to end plates in degenerative lumbar disk disease. Am. J. Roentgenol. **149**(3), 531–534 (1987)
2. Modic, M.T., et al.: Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology **166**(1), 193–199 (1988)
3. Modic, M.T., et al.: Imaging of degenerative disk disease. Radiology **168**(1), 177–186 (1988)
4. Wang, Y., Videman, T., Battié, M.C.: Modic changes: prevalence, distribution patterns, and association with age in white men. Spine J. **12**(5), 411–416 (2012)
5. Zhang, Y.-H., et al.: Modic changes: a systematic review of the literature. Eur. Spine J. **17** (10), 1289–1299 (2008)
6. Albert, H.B., et al.: Modic changes, possible causes and relation to low back pain. Med. Hypotheses **70**(2), 361–368 (2008)
7. Järvinen, J., et al.: Association between changes in lumbar Modic changes and low back symptoms over a two-year period. BMC Musculoskelet. Disord. **16**(1), 98 (2015)
8. Fayad, F., et al.: Reliability of a modified Modic classification of bone marrow changes in lumbar spine MRI. Jt. Bone Spine **76**(3), 286–289 (2009)
9. Crockett, M.T., et al.: Modic type 1 vertebral endplate changes: injury, inflammation, or infection? Am. J. Roentgenol. **209**(1), 167–170 (2017)

10. Wang, Y., et al.: Quantitative measures of Modic changes in lumbar spine magnetic resonance imaging: intra-and inter-rater reliability. Spine **36**(15), 1236–1243 (2011)
11. Vivas, E.L.A., et al.: Application of a semiautomatic classifier for Modic and disk hernia changes in magnetic resonance. Coluna/Columna **14**(1), 18–22 (2015)
12. Jamaludin, A., Kadir, T., Zisserman, A.: Automatic Modic changes classification in spinal MRI. In: Vrtovec, T., Yao, J., Glocker, B., Klinder, T., Frangi, A., Zheng, G., Li, S. (eds.) CSI 2015. LNCS, vol. 9402, pp. 14–26. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41827-8_2
13. Jensen, T.S., Sorensen, J.S., Kjaer, P.: Intra-and interobserver reproducibility of vertebral endplate signal (Modic) changes in the lumbar spine: the Nordic Modic consensus group classification. Acta Radiol. **48**(7), 748–754 (2007)
14. Jamaludin, A., Kadir, T., Zisserman, A.: SpineNet: automatically pinpointing classification evidence in spinal MRIs. In: Ourselin, S., Joskowicz, L., Sabuncu, Mert R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 166–175. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_20
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Zhang, H., et al.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
17. Deng, J., et al.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. IEEE (2009)