Combining Mixed-Format Labels for AI-Based Pathology Detection Pipeline in a Large-Scale Knee MRI Study

Micha Kornreich(✉), JinHyeong Park, Joschka Braun, Jayashri Pawar, James Browning, Richard Herzog, Benjamin Odry, and Li Zhang

Covera Health, New York, NY 10013, USA
micha.kornreich@coverahealth.com
https://www.coverahealth.com/

Abstract. Labeling for pathology detection is a laborious task, performed by highly trained and expensive experts. Datasets often have mixed formats, including a mix of pathology positional labels and categorical labels. Successfully combining mixed-format data from multiple institutions for model training and evaluation is critical for model generalization. Herein, we describe a novel machine-learning method to augment a categorical dataset with positional information. This is inspired by the emerging data-centric AI paradigm, which focuses on systematically changing data to improve performance, rather than changing the model. In order to improve on a baseline of reducing the positional labels to categorical data, we propose a generalizable two-stage method that directs model attention to regions where pathologies are highly likely to occur, exploiting all the mixed-format data. The proposed approach was evaluated using four different knee MRI pathology detection tasks, including anterior cruciate ligament (ACL) integrity and injury age (5082 cases), and medial compartment cartilage (MCC) high-grade defects and subchondral edema detection (4251 cases). For these tasks, we achieved specificities and sensitivities between 90–94% and 78–93%, respectively, which were comparable to the inter-reader agreement results. On all tasks, we report an increase in AUC score, and an average of 8% specificity and 4% sensitivity improvement, as compared to the baseline approach. Combining a UNet network with a morphological peak-finding algorithm, our method also provides defect localization, with average accuracies between 4.3–5.1 mm. In addition, we demonstrate that our model generalizes well on a publicly available ACL tear dataset of 717 cases, without re-training, achieving 90% specificity and 100% sensitivity. The proposed method can be used to optimize image classification tasks in other medical or non-medical domains, which often have a mixture of categorical and positional labels.

Keywords: MRI · Pathology detection · ACL · Cartilage · Data-centric

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16452-1_18.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13438, pp. 183–192, 2022.
https://doi.org/10.1007/978-3-031-16452-1_18

1 Introduction

One of the key challenges to the clinical deployment of artificial intelligence models in medical imaging is the failure of models to generalize across institutions, demographics, and imaging protocols [7]. Accordingly, it is important to train and evaluate models over a broad variety of data sources, and to be able to combine these sources efficiently in training.

However, different data sources which focus on the same pathology detection task could include different label data types, as well as different image types. This is especially true in knee MRI studies, where MRI orientations, protocols, and scanners can significantly differ between unassociated institutions. In addition, some data sources include positional labels such as point-landmarks or bounding boxes, while others only include labels in the form of text or categories.

Combining defect bounding box labels with categorical labels was recently explored in a large scale chest x-ray study, using multiple instance learning [8]. The results suggested that combining positional and categorical *defect* labels can improve network attention and performance over purely categorical training. Such attention focusing was previously accomplished in MRI studies by using *anatomical* landmarks (*i.e.*, that do not label a defect), including in recent knee ACL and cartilage studies [1,9–11]. However, models combining different defect label types were not addressed.

We present a two stage model to combine positional point-like defect landmark labels with categorical defect labels. The first stage was trained on positional labels to predict possible defect locations. A compact volume-of-interest (VOI) was cropped around each predicted defect location, to improve network attention. The second stage classified the pathologies in the VOIs using a convolutional network, and was trained on a combination of positional and categorical labels. This two-stage technique overcomes the difficulties object and point detection models face while training on class imbalanced sets [12] - a common scenario in the medical field.

Our method was evaluated on four knee MRI defect detection and localization tasks, including anterior cruciate ligament (ACL) integrity and injury age, as well as medial compartment cartilage (MCC) high-grade defect and subchondral osteoarthritis related edema underlying the cartilage defect. The classification performance was comparable to the inter-reader agreement levels in the radiologists' reviews, and superior to a purely categorical baseline.

For a recent general review of pathology detection in MRI for ACL and cartilage, we refer readers to [6].

The main contributions of this paper are:

- A method to efficiently combine categorical-label datasets with positional-label datasets during training.
- The proposed method can train and infer on studies that include one or multiple series.
- First models, to our knowledge, trained at this scale for ACL injury-age and Osteoarthritis associated subchondral edema underlying the high grade cartilage defect pathology detection.

– Leveraging the above-mentioned series and label type flexibility during training, we were able to use over 5,000 studies from over 25 institutions, as well as validate on a publicly available dataset.

2 Data

The dataset included 5676 ACL reviews collected from 5082 imaging studies, and 4759 MCC reviews, collected from 4251 studies. Studies were split between training (66%), validation (21%) and test sets (13%). The split was performed in two stages. First, we randomly sampled at a 70-30-10% ratio. Then, we randomly sampled positive cases from the training set, until each positive category in the test set had at least 100 cases. The data did not include multiple studies for any single patient.

Studies were collected at over 25 different institutions, and differed in scanner manufacturers, magnetic field strengths, and imaging protocols (Supplementary Fig. 1). The most common series types included fat-suppressed (FS) sagittal (Sag), coronal (Cor) and axial (Ax) orientations, using either T2-weighted (T2) or proton-density (PD) protocols (Supplementary Table 1). For pathology detection, we used either SagFS, SagPD, or both.

Ground Truth Labeling Process. Each study was reviewed by at least one of eight board-certified radiologists with an MSK fellowship. The review was performed using either a structured form (for categorical labels) or a custom viewer. Radiologists using the viewer also annotated the position of the defect. In both formats, the same ACL and MCC defect categories were used (see Supplementary Tables 2 and 3). ACL categories included *ACL defect* (normal, degeneration, partial tear, or complete tear) and *ACL injury age* (non-acute, or acute). For the MCC, structured report categories included *Cartilage defects* (normal or slight thinning, small high-grade defect, moderate high-Grade defect, or large high-grade defect) and *Edema underlying cartilage defects* (none or trace edema, or more than trace edema). The edema labeled in our dataset differs from the one labeled in previous studies [1], since it is limited to non-traumatic, osteoarthritis associated edema that is underlying a high-grade defect. This distinction is

Table 1. Available categorical and position-labeled data for different pathologies.

Labels	Class	ACL C. Tear	ACL acute	MCC edema	MCC grade
Categorical	0	2323	2161	1327	1122
	1	147	80	198	403
Positional	0	1794	1907	2260	1808
	1	818	705	466	918
Total		5082	4853	4251	4251

clinically important, since Osteoarthritis associated edema is a good predictor of structural deterioration in knee osteoarthritis [5].

Notably, an annotated review could include the same location label type (*e.g.*, a small high-grade defect) multiple times in the same series, one for each such observed defect on the cartilage surface.

Labels Used by Models. For model training and evaluation, we grouped label categories to create 4 tasks that can assist in surgical decision making. For ACL, we trained a model to differentiate *Complete tear* from *Not-complete tear*, and another to predict *Acute* vs. *Non-acute* states. In the MCC, one task was *High-grade defect* vs. *Not-high-grade defect*, and another was *Underlying edema* vs. *None or trace edema* (Table 1 and Supplementary Table 4).

Inter-reader agreement analysis was conducted on 1398 studies with multiple reviews. For training and testing, if two conflicting reviews for the same study existed, the position-annotated review was preferred over the categorical-only one.

3 Methods

Preprocessing Using Deep Reinforcement Learning. Images were automatically cropped around the ACL or MCC prior to pathology detection. Two anatomical landmarks, the Intercondylar Eminence and the Fibular Styloid, were detected using a deep reinforcement learning model [2], and a VOI was positioned with respect to the location of the landmarks. VOI dimensions were determined by clinical experts to include the anatomy of interest (ACL or MCC).

The ACL VOI was a $75 \times 75 \times 75$ mm^3 cube, centered 2.5 mm anteriorly and 2.5 mm medially from the Intercondylar Eminence. The MCC VOI dimensions were 80 mm (superior-inferior), 95 mm (Anterior-posterior) and 75 mm (left-right). The VOI was located 27.5 mm superior, 12.5 mm anterior, and 12.5 mm medial to the Intercondylar Eminence.

Cropped images were linearly interpolated in-plane to a 0.325 mm resolution. Images with out-of plane resolution below 2 mm were sub-sampled (but never interpolated out of plane) to approximately a 4 mm resolution. Images were then intensity-standardized by clipping the 1$^{\text{st}}$ and 99$^{\text{th}}$ percentile intensities, followed by volume normalization to 0 mean intensity and 1 standard deviation.

Baseline Convolutional Network. As a baseline to our proposed method, we trained a 3D ResNet50 for each of the four tasks using the same preprocessing steps described above. In order to include the position-annotated labels in baseline training, we used their categorical labels only. Following preprocessing, the VOI was in-plane padded to a square, and resized to 256×256 pixels. The number of slices was fixed to 24, by either padding or slicing. We also experimented with 128×128 and 320×320 pixel images, which both achieved slightly worse baseline results and were discarded.

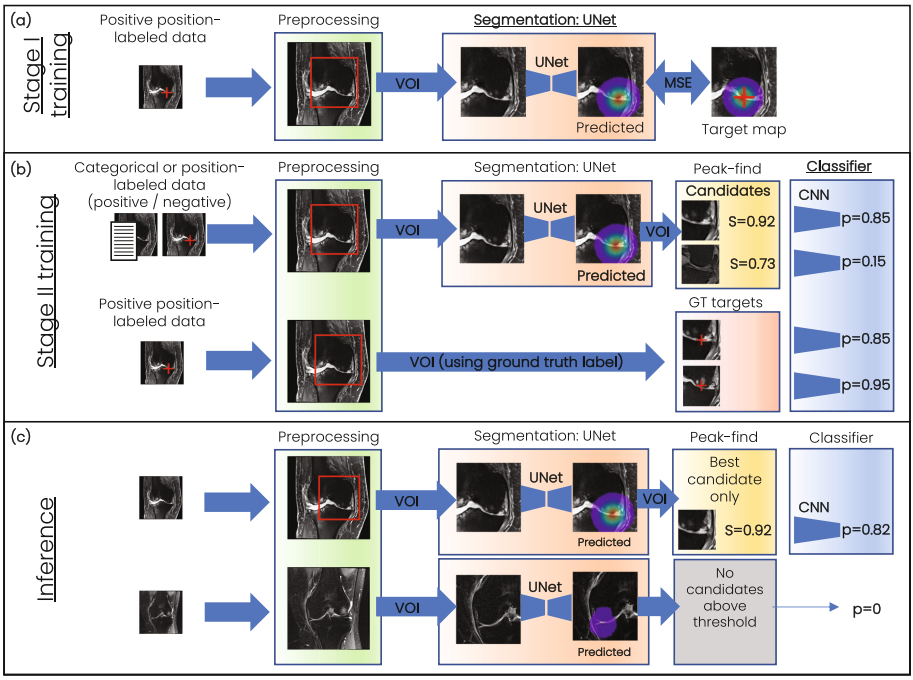

Fig. 1. Architecture. (a) The segmentation UNet is trained during Stage I, using MSE loss to the Target map. (b) The classifier is trained during Stage II, using cropped VOI that were centered either around a positional label created by the expert annotator, or around a candidate label predicted by the UNet and peak-finding algorithm. (c) During inference, positional labels are not used. If the peak-finding failed for the case, the model predicts Class 0 (negative).

A dropout modification to the convolutional network allowed us to train a single model on studies with either Sag FS, Cor FS, or both. The network had two parallel encoders, one for Sag and another for Cor images. The features from the encoders were concatenated and forwarded to a fully connected network. Whenever one of the series was missing, its corresponding feature vector was dropped-out, while the other feature vector was multiplied by two. This was performed both in training and inference.

3.1 Proposed Model Using Mixed-Format Labels

The proposed method utilizes both categorical and positionally-labeled data formats during training, which is performed in two stages, as explained below. Models were trained using PyTorch 1.7.1 and Albumentations 1.1.0 software on an AWS *p3.x2large* instance (16 GB v100 Tesla GPU), where 200 Stage I followed by 50 Stage II epochs took 24 h. GPU memory allowed batch sizes up to 10 and 36 in Stage I and II, respectively.

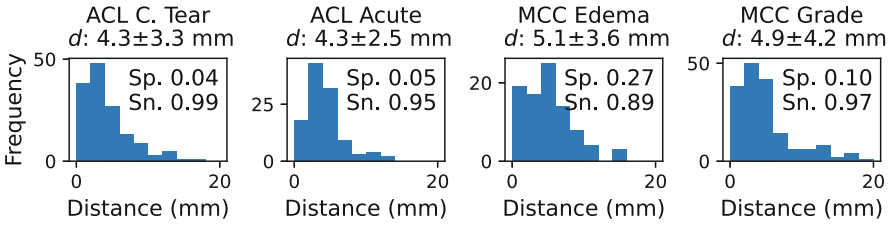

Fig. 2. Stage I results. Histograms for distances between the best candidate and the nearest defect, along average distance d±std, sensitivity (Sn.) and Specificity (Sp.) for each task. Distances are calculated for cases which were predicted positive by Stage II.

Stage I: Landmark. To locate potential defects, we used a Residual UNet model [13], where each volume could have none, one, or many target positional labels. During training, this stage used only pathology-positive studies, and only series (*i.e.*, volumes) with at least one positional annotation (Fig. 1a). Following [15], for each volume, a target map was created, where each label in location μ was replaced by an isotropic Gaussian sphere: $I_G = \frac{1}{\sigma\sqrt{2\pi}} e^{-\|x-\mu\|^2/2\sigma^2}$, with $\sigma = 10$ mm. Training was performed with an MSE loss function, ADAM optimizer (lr = 0.0001), 128 × 128 × 24 volume size, and a batch size of ten. A separate UNet was independently trained for each of the four tasks.

Coordinates of potential defect landmarks (candidates) were extracted from the UNet output, using a fast peak-finding algorithm, originally developed for particle-tracking [3]. During training, all candidates were forwarded to Stage II. During test and validation, only the "best" candidate was selected from each series for stage II (see Supplementary Fig. 2). Whenever Stage I found no defect candidates, the full model prediction was negative (class 0).

Stage II: Pathology Detection. Classification was performed using a 3D ResNet50 with an ADAM optimizer (lr = 0.0001). For each task, the model was trained and evaluated on $40 \times 40 \times 40\,\text{mm}^3$ defect-VOI cubes, which were resized to 128 × 128×12 pixels. Each cube was centered around a single candidate defect location that was predicted in stage I, or a location provided by our ground-truth annotations (Fig. 1b). Limiting the classifier to these compact VOIs was meant to improve network attention, and subsequent performance. Series for which stage I found no candidates were not included in stage II training. Training was performed by cross-entropy loss on each series, where the categorical ground truth for the study (rather than series) was used. To assess the performance of our complete model on unlabeled data, we only used locations predicted by the trained stage I model during inference (Fig. 1c).

Perturbations and Augmentation Strategy. In both stages, augmentations included blurring, uniform noise, gamma shift, horizontal flips (for Cor only) and reverse ordering of slices (for Sag only). In addition, during stage II training, the defect location was shifted uniformly in the $[-3.5, 3.5]$ mm range in all axes.

4 Results and Discussion

Baseline Convolutional Network Performance. The baseline convolutional model was evaluated on two datasets: 1) including only categorical labels (Labels = *Categ.* and Method = *ResNet* in Table 2) and, 2) a unified dataset which included categorical labels originating from both the categorical and the positional-annotated datasets (Labels = *Both* and Method = *ResNet* in Table 2). When training the baseline model using the unified dataset, all positional-annotated data was reduced to categorical format by taking the most severe label for each study, and removing the positional information.

4.1 Performance of Proposed Model Using Mixed-Format Labels

Stage I: Landmark. Stage I training was designed to achieve high sensitivity, since false positive studies would be filtered by stage II. Indeed, in three tasks we observed sensitivity exceeding 95% (Fig. 2). By training on positive samples only at stage I and using stage II for filtering, we circumvent difficulties encountered when training object detection models on mostly negative samples.

The localization accuracy of stage I confirms that the defects are captured by the $40 \times 40 \times 40\,\mathrm{mm}^3$ VOI cubes used by the following stage (see Fig. 2).

Table 2. Ablation study of different training methods and datasets, where each model was run 5 times using randomly initialized weights to produce average ± std. Inter-reader sensitivity and specificity appear in the last row.

Labels		Categ	Positional	Positional	Both	Both	Both	Inter-
method		ResNet	two-stage	two-stage	ResNet	two-stage	two-stage	reader
perturb				+			+	
ACL	Sp.	82.8 ± 1.3	**92.4** ± 0.8	90.2 ± 0.7	89.4 ± 0.8	92.2 ± 0.4	**92.2** ± 0.7	97
C. Tear	Sn.	47.8 ± 5.8	65.4 ± 1.9	89.4 ± 1.0	**91.0** ± 0.6	76.4 ± 1.2	**92.6** ± 0.5	84
	AUC	72.0 ± 0.6	90.8 ± 0.4	94.8 ± 0.4	**95.8** ± 0.3	91.8 ± 0.4	**97.0** ± 0.2	
ACL	Sp.	65.8 ± 1.7	89.8 ± 0.7	**93.4** ± 1.0	**94.0** ± 1.4	89.8 ± 1.0	92.2 ± 0.7	97
acute	Sn.	**74.8** ± 3.0	69.8 ± 1.5	73.0 ± 1.4	64.0 ± 0.9	72.4 ± 1.0	**78.4** ± 1.0	74
	AUC	75.6 ± 1.6	84.4 ± 0.5	87.6 ± 0.5	**88.0** ± 0.6	85.8 ± 0.4	**89.2** ± 0.4	
MCC	Sp.	73.6 ± 2.4	89.2 ± 0.7	**94.0** ± 0.9	82.8 ± 1.2	90.8 ± 0.7	**92.8** ± 0.7	92
edema	Sn.	64.4 ± 2.7	68.3 ± 0.8	74.2 ± 0.7	**80.8** ± 0.4	70.2 ± 0.7	**78.4** ± 0.5	69
	AUC	74.2 ± 1.2	86.4 ± 0.5	**88.4** ± 0.5	87.4 ± 0.5	87.0 ± 0.9	**90.2** ± 0.4	
MCC	Sp.	66.2 ± 1.5	**92.4** ± 0.8	**89.8** ± 0.7	87.6 ± 1.0	83.2 ± 1.2	89.0 ± 0.6	87
grade	Sn.	77.8 ± 1.6	75.2 ± 0.7	80.4 ± 1.0	77.4 ± 0.5	**85.4** ± 0.5	**88.2** ± 0.4	84
	AUC	79.6 ± 1.4	90.4 ± 0.5	**91.6** ± 0.5	90.4 ± 0.5	91.0 ± 0.6	**93.8** ± 0.4	

Stage II: Pathology Detection. Final classification results were obtained at stage II inference, using the cropped volumes. Each volume was centered around a best defect candidate predicted by stage I. The sensitivity, specificity, and AUC are detailed under *Two-stage* method in Table 2. For ablation research, four different models were trained. Two models only used positional-labels in training (Labels = *Positional* in Table 2). This was achieved by removing the positional information from the annotations, and using only the categorical information. The other two used both categorical and positional datasets, facilitated by our two-stage combined approach (Labels = *Both* in Table 2). In addition, in two of the four models we added a random perturbation shift to the best candidate location, sampled uniformly in the range [−3.5, 3.5] mm in each direction. Evaluation was performed on the same data set for all four models.

Our combined method achieved specificities and sensitivities between 89–94% and 78–93%, respectively, which were comparable to the inter-reader agreement results. These results were, on average, 8% specificity and 4% sensitivity over the baseline model. A McNemar's test [4] comparing our method with the baseline model (columns 7 and 5 in Table 2, respectively) yielded a 10^{-6} p-value. Since the dataset is not class balanced, we also computed the p-values for positive cases (sensitivity, p-value = 0.0001) and negative cases (specificity, p-value = 0.0008), indicating that the performance improvement was statistically significant.

Our best results are comparable to the inter-reader agreement between the board-certified MSK fellowship trained radiologists that labeled our ground-truth data. Notably, it is unusual for a model to outperform noisy ground-truth agreement rates in evaluation, unless the evaluation set is obtained from a higher-quality source. However, our test set had disproportionately more positional-annotated studies, which the radiologists established as more reliable (Supplementary Table S5). Therefore, we maintain that higher performance evaluation is possible, given the assumption that the test had higher quality labels.

Public Dataset Validation. The trained model utilized data from multiple institutions, using various protocols, with either Cor, Sag or both orientations. To further validate its generalizability, we evaluated performance, without any re-training, on a publicly available ACL dataset [14]. The dataset included 917 Sag PDFS 12-bit grayscale images. 717 studies (≈76%) were classified as non-injured, 182 (≈19%) partially injured, 45 (≈5%) completely ruptured. Since our model differentiates complete tear from not complete tears, we mapped their class labels from 1 to 0 and 2 to 1. Even though our network was trained on data that typically had both Sag and Cor series, it achieved 90% specificity and 100% sensitivity, which were comparable to results on our test set.

5 Conclusions

We proposed a novel method to flexibly combine categorical labels with positional labels during training, and demonstrated its applicability in four knee MRI pathology detection tasks. Our method leverages available positional-annotated data to attach location to categorical labels, which improves the overall model performance. In addition, it reliably localizes the defects, which is useful in several potential applications, such as computer aided diagnosis and AI-based quality assurance. We show that without any re-training, our model, which was trained to use either one or two MRI orientations, can generalize well to a publicly available, which included one orientation (Sag) only. Notably, our method can be employed in other computer vision domains, such as captioning, where similar mixed-format label types are often available during training.

References

1. Astuto, B., et al.: Automatic deep learning-assisted detection and grading of abnormalities in knee MRI studies. Radiol. Artif. Intell. **3**(3), e200165 (2021)
2. Browning, J., et al.: Uncertainty aware deep reinforcement learning for anatomical landmark detection in medical images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 636–644. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_60
3. Crocker, J.C., Grier, D.G.: Methods of digital video microscopy for colloidal studies. J. Colloid Interface Sci. **179**(1), 298–310 (1996)
4. Everitt, B.S.: The Analysis of Contingency Tables. CRC Press, New York (1992)
5. Felson, D.T., et al.: Bone marrow edema and its relation to progression of knee osteoarthritis. Ann. Internal Med. **139**(5(1)), 330–336 (2003)
6. Fritz, B., Fritz, J.: Artificial intelligence for MRI diagnosis of joints: a scoping review of the current state-of-the-art of deep learning-based approaches. Skeletal Radiol. **51**(2), 315–329 (2021)
7. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., Celi, L.A.: The myth of generalisability in clinical research and machine learning in health care. Lancet Digital Health **2**(9), e489–e492 (2020)
8. Li, Z., et al.: Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8290–8299 (2018)
9. Liu, F., et al.: Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. Radiol. Artif. Intell. **1**(3), 180091 (2019)
10. Liu, F., et al.: Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. Radiology **289**(1), 160–169 (2018)
11. Namiri, N.K., et al.: Deep learning for hierarchical severity staging of anterior cruciate ligament injuries from MRI. Radiol. Artif. Intell. **2**(4), e190207 (2020)
12. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: a review. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3388–3415 (2020)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

14. Štajduhar, I., Mamula, M., Miletić, D., Uenal, G.: Semi-automated detection of anterior cruciate ligament injury from MRI. Comput. Methods Programs Biomed. **140**, 151–164 (2017)
15. Yang, D., et al.: Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 633–644. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_50