Detection of Multidrug-Resistant Tuberculosis Using Convolutional Neural Networks and Decision Trees

Martha Tatusch and Stefan Conrad

Heinrich Heine University, 40225 Düsseldorf, Germany
{tatusch, conrad}@cs.uni-duesseldorf.de

Abstract. In 2018, tuberculosis was one of the top 10 causes of death worldwide. Especially patients that develop a multidrug-resistance are endangered and need special medical treatment. Within the ImageCLEF 2018 challenge the automatic distinction between drug-sensitive and multidrug-resistant tuberculosis was investigated by only using the CT scan, age and gender of a patient. In this paper, we present different approaches using convolutional neural networks, decision trees and the combination of both classifiers. We show that our models achieve competitive results regarding the other participants of the challenge and show an improvement with respect to our last year's results. All of them are represented in the ranks between 4th and 26th of 39. Our best method regarding the AUC measure reached a score of 0.5810. In regard of accuracy the best approach got a result of 0.572.

Keywords: Convolutional Neural Networks · Image Processing · Classification · Tuberculosis · Multidrug-resistance · ImageCLEF 2018

1 Introduction

Tuberculosis is a disease that is caused by an infection with the mycobacterium tuberculosis. Although the bacterium was found about 135 years ago, it still is one of the top ten causes of death worldwide according to the World Health Organization (WHO)[1]. Due to medical progress the disease can be cured. Some patients, however, can develop a resistance to several drugs. This circumstance complicates the medical treatment and must therefore be recognised as soon as possible. Since the distinction between drug-sensitive (DS) and multidrug-resistant (MDR) tuberculosis is difficult and necessitates several expensive tests, it would be helpful to find an automated solution that only requires the CT scans that are usually done anyway.

This year, the ImageCLEF 2018 tuberculosis Task 1 [9] took up this challenge once again. As last year's results as a whole were not satisfying for us yet, we wanted to participate in the competition one more time. The goal was to elaborate an automatic model that can predict a score for the probability whether a

[1] http://www.who.int/en/news-room/fact-sheets/detail/tuberculosis date: 28.06.18

patient suffers from MDR.
In contrast to the last year's challenge, additionally to the images the age and gender of the patients were given. Also the number of images was increased. Nevertheless, the amount of training data is still relatively small and therefore another challenging factor.
In this paper we introduce our approaches for the task which include the usage of convolutional neural networks, decision trees and the combination of both classifiers.

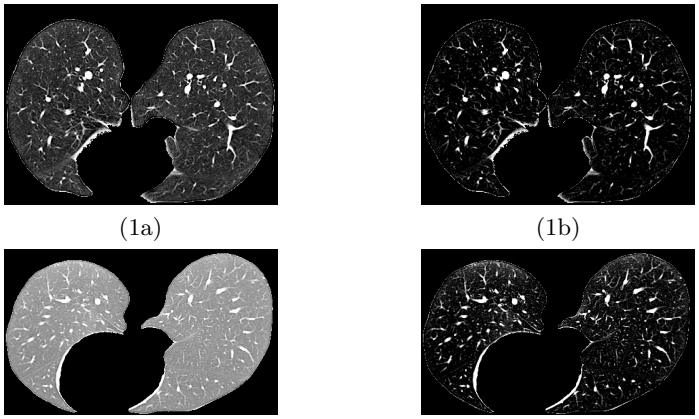
2 Related Work

In the ImageCLEF 2017 tuberculosis challenge the same task was set for the first time[8]. Only the provided dataset varied slightly. There were around 10% fewer training images and neither the age nor the gender of the patients was given. In 2017, the best approach regarding the AUC score was a graph-based model that is described in [7]. The authors divided the lung into 36 regions that were represented by nodes. These nodes could then be connected by edges using different methods. The goal was to describe the lungs by a feature vector extracted from the underlying graph. These features could then be used to train a support vector machine. This model achieved ranks 1 to 3 with AUC scores between 0.5825 and 0.5624.
The second best team used a combination of convolutional and recurrent neural networks and achieved results between 0.562 and 0.5501 on the ranks 4 – 6. The model was published in [15]. The UIIP team was ranked 7th and thus represented the third best team. In [13] the authors introduced their own segmentation algorithm and their method based on feature extraction by considering supervoxels. This team used external sources to segment the lungs.

Regarding the accuracy, however, our team aimed the first rank with 0.5681 using convolutional neural networks with a flat network architecture [2]. The score was around 4% better than a model that only classifies into the most represented class. This classifier would reach an accuracy of 0.528.

3 Methods

Since the manual distinction between DS and MDR tuberculosis using only CT Scans is not possible until now, it is very difficult to determine relevant features of the images. For this reason, the usage of neural networks, which do not require feature extraction, is reasonable. On the one hand a convolutional neural network can be used, which only considers the images, on the other hand an architecture can be developed, which examines the images as well as the additional features *age* and *gender*. The combination of the images and the textual features cannot only be done by a CNN with multiple inputs, but also by creating a feature vector of the CNN's result, the age and the gender, and using it as the input of another classifier. In this work, a decision tree has been chosen as the second classifier. Before discussing the different approaches, the preprocessing of the



(1a) (1b)

(2a) (2b)

Fig. 1: Illustration of two example scans before (left) and after (right) the denoising step from [2].

images, which has a great impact on the results, will be explained, since it is the same for all models.

3.1 Preprocessing

The CT scans are three-dimensional grayscale images whose intensity values are specified by Hounsfield Units (HU) [3] – a uniform measure for CT scans. Although the Hounsfield Units are fixed, they leave a little scope for the actual values, which can vary due to different computer tomograph settings [6]. These circumstances can explain the high diversity of the given scans. In Figure 1 two lungs of the given CT scans are shown in 1a) and 2a). In 1a) the Hounsfield Units are in the range of $[-1024, 2017]$. In 2a) the values vary between -1582 and 1941. Due to the different size of the HU ranges, the intensity of the scans varies a lot. This is also reflected in the representation of the images in Figure 1. Because of this, a preprocessing of the images is essential. As the preprocessing method developed in [2] led to an improvement of the results, it was used this year, too. The procedure is explained in [2] and can be summarized in 5 steps:

1. Set the smallest Hounsfield Unit value of the images to the second smallest value -1.
2. Normalize all values to the range of $[0, 1]$.
3. Segment the scans using the provided masks that were computed by the method presented in [10].
4. Denoise the image by creating an intensity histogram with 256 even distributed bins. Set all values v with $v \leq u$ to u, where u is the upper border of the bin with the highest number of occurrences.
5. Increase the contrast by normalizing the range from $[u, 1]$ to $[0, 1]$ again.

The first step is necessary, as the smallest value represents the background of the scan and causes a large gap between the smallest and the second smallest value of it. The fourth step is done to decrease the noise of the image. As illustrated in Figure 1, the diversity of the segmented and normalized scans is very high. According to [4] the relevant features in an image are mainly represented in the properties of the nodules. These regions contain relatively high Hounsfield Units. Since the noise mostly occurs in the darker parts of the scan and the relevant regions are represented by high values, it is reasonable to reduce the noise which occurs beneath a certain threshold. In Figure 1, the lower bound of image 1a) has been increased from -1024 to -888 in image 1b). Thus, the value range has been decreased from 3041 values to 2905. Regarding the image 2a), the lower bound went from -1582 to -860 in 2b), so that the value range only contained 2801 values instead of 3523. The difference of the lower bounds of the two considered images has therefore been decreased from 558 to 28. Highlighting bright areas in the resulting scans by increasing the contrast is also helpful, since relevant regions get greater emphasis.

3.2 Convolutional Neural Networks

The first approach of this work is a classification of the images using a convolutional neural network, that takes the three dimensional scans as input. This was done by using the Keras API [5] with Tensorflow [1] backend. Five quite similar architectures have been submitted. The architecture of two networks using the Spatial Pyramid Pooling Layer (SPP) [11] is shown in Table 1.

Since the usual SPP was made for 2D images and the used CNNs were three dimensional, a few modifications had to be made. These were already carried out and described last year in [2] and could therefore be reused. The CNN named *Conv48* creates 4 feature maps in the first convolutional layer. It has the binary cross entropy as its loss function and the stochastic gradient descent as optimizer. *Conv68*, however, determines 6 filters in the first block and uses the

Table 1: Architecture of the CNNs using the SPP.

Layer	Number of Filters	Filter Size	Stride
MaxPooling 0	n.a.	(4, 4, 1)	(4, 4, 1)
Convolution 1	4 or 6	(3, 3, 3)	(1, 1, 1)
MaxPooling 1	n.a.	(2, 2, 2)	(1, 1, 1)
Convolution 2	8	(3, 3, 3)	(1, 1, 1)
MaxPooling 2	n.a.	(2, 2, 2)	(1, 1, 1)
Dropout 0.25	n.a.	n.a.	n.a.
Spatial Pyramid	[1,2,4,8]	n.a.	n.a.
Dense	n.a.	n.a.	n.a.

Table 2: Architecture of the CNNs using the flatten layer.

Layer	Number of Filters	Filter Size	Stride
MaxPooling 0	n.a.	(4, 4, 1)	(4, 4, 1)
Convolution 1	16	(3, 3, 3)	(1, 1, 1)
MaxPooling 1	n.a.	(2, 2, 2)	(1, 1, 1)
Convolution 2	8	(3, 3, 3) or (3, 3, 1)	(1, 1, 1)
MaxPooling 2	n.a.	(2, 2, 2)	(1, 1, 1)
Convolution 3	8	(3, 3, 3) or (3, 3, 1)	(1, 1, 1)
MaxPooling 3	n.a.	(2, 2, 2)	(1, 1, 1)
Dropout 0.25	n.a.	n.a.	n.a.
Flatten	n.a.	n.a.	n.a.
Dense	n.a.	n.a.	n.a.

categorical cross entropy and the adam optimizer. Both networks can handle an unfixed input size because of the SPP. Nevertheless, the nets have been trained and tested with images of different as well as uniform sizes.

In Table 2 the architecture of the second type of CNNs is shown. Instead of the SPP, the flatten layer is used. Besides, one additional block of a maxpooling and convolution layer is inserted. The difference between the structure of *Flatten* and *Flatten3* is in the second and third block, where Flatten performs the convolution only over the X- and Y-axis of the images while Flatten3 considers all axes. Both networks use binary cross entropy and stochastic gradient descent. Another submitted architecture is called *FlattenX*. It is very similar to Flatten.

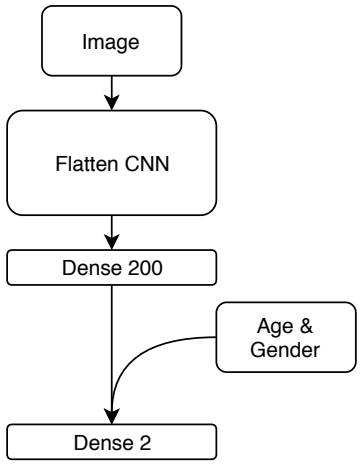


Fig. 2: Architecture of the CNN with multiple inputs.

Instead of the first maxpooling layer, however, it has a convolutional layer with 8 filters and a filter size of $(5, 5, 3)$. The intention of this modification was to avoid a possible loss of information in the first layer.
The last submitted CNN is the *MultiInputCNN*, which takes an image and a two dimensional vector as inputs. As can be seen in Figure 2, the first part of the net is identical to the Flatten CNN. After the flatten layer, however, a dense layer with 200 units is used. Its result is merged with the second input which contains the age and gender of the considered patient. This data is then processed by a dense layer with the size of 2.

3.3 Decision Trees

As already mentioned in the introduction of Section 3, all given information can be combined by using multiple classifiers. To refine the results of the CNN, a decision tree was trained using the CNN's binarized output, the age and the gender. The decision tree was chosen because the considered features and decisions are easy to track. This is very helpful to understand the impact of CNN's results.
Depending on the selection of the network the classifier learned whether to consider the CNN's result or not. In Figure 3 the structures of the two best decision trees are illustrated. On the right, the result of using the Flatten net is displayed. After checking age and gender the CNN's output is considered, as well. In contrast to this, the results of the Flatten3 net have no influence on the decision tree's assignment at all. The structure is shown on the left of Figure 3. It is noticeable that the tree in a) exactly corresponds to the first part of the one in b). The models have been retrieved by using the *DecisionTreeClassifier* from Scikit-Learn [14]. The parameters of both trees are the same: the minimum

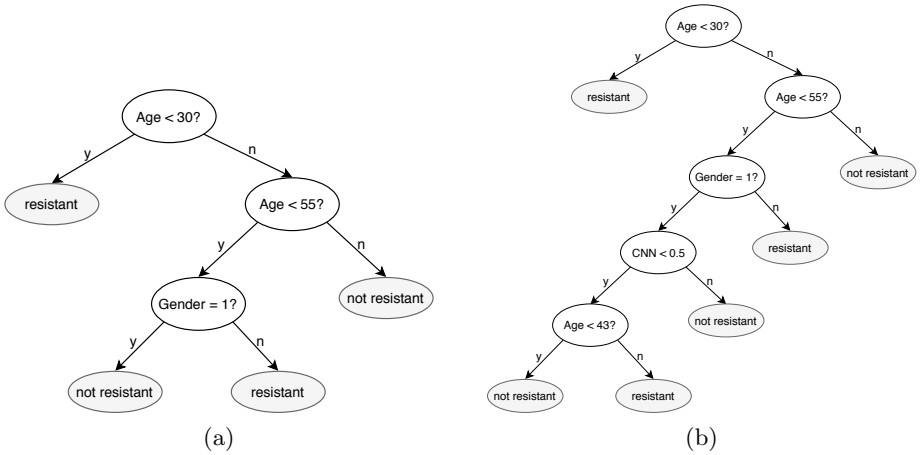


Fig. 3: Illustration of the two best decision trees' structures.

fraction per leaf was set to 10%, the minimum impurity decrease was 0.01. All architectures used the Gini Impurity as a measure for the information gain.

4 Experimental Results

The training set that is provided by the ImageCLEF 2018 [12] contains 259 training images. 134 of them belong to patients with DS tuberculosis, the other 125 scans represent MDR. The test set is composed of 236 images, 99 of which have DS and 137 MDR. The ratio of the types is therefore very different in the two data sets. The CT scans consist of 50 to 400 slices with 512×512 pixels. The images have been resized by using three-dimensional bounding boxes around the lungs. That means, the side lengths of the two-dimensional slices have been cropped to the outermost points of the whole lung and empty slices have been removed from the scan.
When using a fixed input size, all images have then been resized to a X- and Y-length of 250 pixels and a total number of 100 slices. If the length of a side (x, y or z) was too small, it was enlarged by adding equally black borders on both sides of the axis. If the length was too big, interpolation was used to downsize the side with the biggest variance regarding the desired dimension, so that the proportions of the side lengths remained the same. Afterwards the sides that became too small were enlarged as described.
All networks have been trained with the complete training set and for at least 30 epochs. In the combination of CNN and decision tree, the network first had been trained with 180 and the tree with 40 randomly selected CT scans which had an even type distribution. Afterwards the network was post-trained with all given training images.

In Table 3 the preliminary results of the challenge are displayed as published on the website[2]. For a better understanding a few files of our team have been renamed. As our results are among the top 20 regarding the AUC score, only an excerpt of the top 25 of 39 runs was selected. The combination of the Flatten3 CNN and the decision tree achieved the best AUC as well as Accuracy score regarding the complete set of our submitted runs. This method was ranked 6th with respect to the AUC measure and 4th in regard of the Accuracy. The run is named "MDR_Flatten3_DTree.txt". "MDR_Flatten_DTree.txt" represents the predictions of the combined approach with the Flatten CNN. It is noticeable that both runs retrieve exactly the same results. That is, because the predictions are the same, as well. This phenomenon can be explained by the fact that the network classified all scans of patients of gender 1 and the age between 43 and 55 as *not resistant*. The additional branch of the tree (in regard to the architecture without CNN) has therefore never been reached using the test set.
Just behind our best runs we have the results of Conv68 which was trained with images of a fixed size. The networks Conv48, Flatten and Flatten3 reached with similar results the AUC ranks 11, 12 and 13. Regarding the accuracy, the Flatten architecture achieved the second best rank for our team. Unexpectedly,

[2] http://www.imageclef.org/2018/tuberculosis date: 28.06.18

Table 3: The top 25 of 39 results of the MDR detection tuberculosis task of the ImageCLEF 2018 challenge ranked by the AUC (R1) and accuracy score (R2).

Group Name	Run	AUC	R1	Accuracy	R2
VISTA@UEvora	MDR-Run-06-Mohan-SL-F3-Personal.txt	**0.6178**	1	0.5593	8
San Diego VA HCS/UCSD	MDSTest1a.csv	0.6114	2	**0.6144**	1
VISTA@UEvora	MDR-Run-08-Mohan-voteLdaSmoF7-Personal.txt	0.6065	3	0.5424	17
VISTA@UEvora	MDR-Run-09-Sk-SL-F10-Personal.txt	0.5921	4	0.5763	3
VISTA@UEvora	MDR-Run-10-Mix-voteLdaSl-F7-Personal.txt	0.5824	5	0.5593	9
HHU-DBS	MDR_Flatten3_DTree.txt	**0.5810**	6	**0.5720**	4
HHU-DBS	MDR_Flatten_DTree.txt	0.5810	7	0.5720	5
HHU-DBS	MDR_Conv68adam_fl.txt	0.5768	8	0.5593	10
VISTA@UEvora	MDR-Run-07-Sk-LDA-F7-Personal.txt	0.5730	9	0.5424	18
UniversityAlicante	MDRBaseline0.csv	0.5669	10	0.4873	32
HHU-DBS	MDR_Conv48sgd.txt	0.5640	11	0.5466	16
HHU-DBS	MDR_Flatten.txt	0.5637	12	0.5678	7
HHU-DBS	MDR_Flatten3.txt	0.5575	13	0.5593	11
UIIP_BioMed	MDR_run_TBdescs2_zparts3_thrprob50_rf150.csv	0.5558	14	0.4576	36
UniversityAlicante	testSVM_SMOTE.csv	0.5509	15	0.5339	20
UniversityAlicante	testOpticalFlowwFrequencyNormalized.csv	0.5473	16	0.5127	24
HHU-DBS	MDR_Conv48sgd_fl.txt	0.5424	17	0.5508	15
HHU-DBS	MDR_Conv68_DTree.txt	0.5346	18	0.5085	26
HHU-DBS	MDR_FlattenX.txt	0.5322	19	0.5127	25
HHU-DBS	MDR_MultiInputCNN.txt	0.5274	20	0.5551	13
VISTA@UEvora	MDR-Run-01-sk-LDA.txt	0.5260	21	0.5042	28
MedGIFT	MDR_Riesz_std_correlation_TST.csv	0.5237	22	0.5593	12
MedGIFT	MDR_HOG_std_euclidean_TST.csv	0.5205	23	0.5932	2
VISTA@UEvora	MDR-Run-05-Mohan-RF-F3I650.txt	0.5116	24	0.4958	30
MedGIFT	MDR_AllFeats_std_correlation_TST.csv	0.5095	25	0.4873	33

the MultiInputCNN received the worst results. The achieved accuracy of 0.5593 is the same for the ranks 9 to 12. The FlattenX network reached the second worst AUC results for our team. This confirms the thesis from [2] that smaller network architectures are better suited for the data set than deep ones.

5 Conclusion

We have shown that our approaches achieve competitive results. Despite the small amount of data, the use of convolutional neural networks can be reasonable if the architecture is not too deep. All our runs reached AUC ranks between 6 and 20 and accuracy ranks between 4 and 26 of 39. The best classifier turned out to be the decision tree, which only takes into account the age and gender of the patient. Its AUC and Accuracy results are only about 4% worse than the best scores of the challenge. Nevertheless, it has to be said that the accuracy results are worse than those of a classifier that only categorizes into the class with the most representatives. This one would reach a score of 58.05%.

The quality of the classifiers can certainly be increased by improving the preprocessing of the images with the help of medical expertise. Furthermore, the optimization of the provided masks could lead to better results, because these do not consider relevant regions of the lungs and contain parts of bones in some cases. Also, it would be interesting to perform the classification with other known medical data besides the age and gender of the patients.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/
2. Braun, D., Singhof, M., Tatusch, M., Conrad, S.: Convolutional neural networks for multidrug-resistant and drug-sensitive tuberculosis distinction. In: CLEF2017 Working Notes. CEUR-WS (2017)
3. Brooks, R.: A quantitative theory of the hounsfield unit and its application to dual energy scanning. Journal of Computer Assisted Tomography **1**(4), 487–493 (1977)
4. Cha, J., Lee, H.Y., Lee, K.S., Koh, W.C.A., Kwon, O., Yi, C.A., Kim, T.S., Chung, M.J.: Radiological findings of extensively drug-resistant pulmonary tuberculosis in non-aids adults: Comparisons with findings of multidrug-resistant and drug-sensitive tuberculosis. Korean Journal of Radiology **10**(3) (2009)
5. Chollet, F., et al.: Keras. https://keras.io (2015)
6. Cropp, R.J., Seslija, P., Tso, D., Thakur, Y.: Scanner and kVp dependence of measured CT numbers in the ACR CT phantom. Journal of Applied Clinical Medical Physics **14**(6), 338–349 (2013)
7. Dicente Cid, Y., Batmanghelich, K., Müller, H.: Textured graph-model of the lungs for tuberculosis type classification and drug resistance prediction: Participation in imageclef 2017. In: CLEF2017 Working Notes. CEUR-WS (2017)
8. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF2017 Working Notes. CEUR-WS (2017)
9. Dicente Cid, Y., Liauchuk, V., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2018 - detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: CLEF2018 Working Notes. CEUR-WS (2018)
10. Dicente Cid, Y., Jiménez del Toro, O.A., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in ct volumes. In: Proceedings of the VISCERAL Anatomy Grand Challenge at the 2015 IEEE ISBI. pp. 31–35. CEUR-WS (2015)
11. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. vol. 8691, pp. 346–361. Springer (2014)
12. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M.,

Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. LNCS Lecture Notes in Computer Science, Springer (2018)
13. Liauchuk, V., Kovalev, V.: Imageclef 2017: Supervoxels and co-occurrence for tuberculosis ct image classification. In: CLEF2017 Working Notes. CEUR-WS (2017)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Sun, J., Chong, P., Tan, Y.X.M., Binder, A.: Imageclef 2017: Imageclef tuberculosis task - the sgeast submission. In: CLEF2017 Working Notes. CEUR-WS (2017)