Scripta Materialia 193 (2020) 33–37

Contents lists available at ScienceDirect

Scripta Materialia

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

Race against the Machine: can deep learning recognize microstructures as well as the trained human eye?

Michiel Larmuseau[a,b,c,*], Michael Sluydts[a,b], Koenraad Theuwissen[d], Lode Duprez[d], Tom Dhaene[c], Stefaan Cottenier[a,b]

[a] Center for Molecular Modeling, Ghent University, Technologiepark 46, Zwijnaarde, B-9052, Belgium
[b] Department of Electromechanical, Ghent University, Technologiepark 46, Zwijnaarde, B-9052, Belgium
[c] IDLab, Department of Information Technology, Ghent University – IMEC, Technologiepark 126, Zwijnaarde, B-9052, Belgium
[d] OCAS NV/ArcelorMittal Global R&D Gent, Pres. J. F. Kennedylaan 3, Zelzate, B-9060, Belgium

ARTICLE INFO

Article history:
Received 23 July 2020
Revised 8 October 2020
Accepted 12 October 2020

Keywords:
Image analysis
Steels
Modeling
Scanning electron microscopy (SEM)

ABSTRACT

The promising results of deep learning in image recognition suggest a huge potential for microscopic analyses in materials science. One major challenge for its adoption in the study of materials is the limited number of images that are available to train models on. Herein, we present a methodology to create accurate image recognition models with small datasets. By explicitly taking into account the magnification and by introducing appropriate transformations, we incorporate as many insights from material science in the model as possible. This allows for a highly data-efficient training of complex deep learning models. Our results indicate that a model trained with the presented methodology is able to outperform human experts.

Microscopy images are an important source of information on the small-scale structure of materials, referred to as the microstructure. However, most of the time, this information is analysed only qualitatively: domain experts for instance mainly examine these images to assess whether the processing of the material went well. While such an assessment is important, a lot of information contained in the microscopy image is not used. In recent years, approaches towards a more quantitative analysis of these images using machine learning have been thoroughly investigated in literature[1–4]. However, all these approaches rely on generic microstructure descriptors that are not tailored to the specific materials in the dataset, resulting in sub-optimal performance.

Deep learning methods[5] make it possible to learn microstructural descriptors directly from the available data. Despite the first report of deep learning outperforming humans in an image recognition task already dating back to 2011[6], its adoption in practical material science remains limited[7–9]. A possible explanation for this, is that deep learning is often only deemed to outperform classical methods when large datasets of images are available. With most commonly used datasets in material science literature containing around the order of a thousand images[2,4,10], this opinion feels reasonable. However, recent work[11] has shown that by creating tailor-made deep learning models for specific datasets, one can outperform models that use generic microstructure descriptors.

This would make it possible to investigate how well deep learning models perform in recognizing microstructures compared to expert materials scientists. Although it is clear from other fields in computer vision that deep learning models can outperform human experts[12,13] provided a sufficient number of images, we here aim to investigate the performance of models that are trained on only around a hundred microstructure images in total. A hundred images is a practical amount, as it can easily be collected in a systematic study of a class of materials.

To evaluate the performance of both a neural network that is obtained using the methodology presented in [11] and the panel of experts, we have organized two different quizzes. For the first quiz, 36 microscopy images need to be assigned to one of the five pre-defined classes of microstructures. The option “None of these” is included in case a microstructure image is shown that does not belong to one of the five pre-defined classes. The dataset on which the model was trained, the training set, contains 134 images in total, with magnifications, expressed as pixels per micrometer, ranging from 1.1 to 212 pixels per micrometer ($pp\mu m$). Both optical microscopy images and scanning electron microscopy (SEM) images have been included. An example for each type of microstructure is shown in Fig. 1 (a). For the second quiz, 21 SEM microscopy images of complex martensitic steels need to be assigned to one

* Corresponding author.
E-mail address: michiel.larmuseau@ugent.be (M. Larmuseau).

https://doi.org/10.1016/j.scriptamat.2020.10.026
1359-6462/

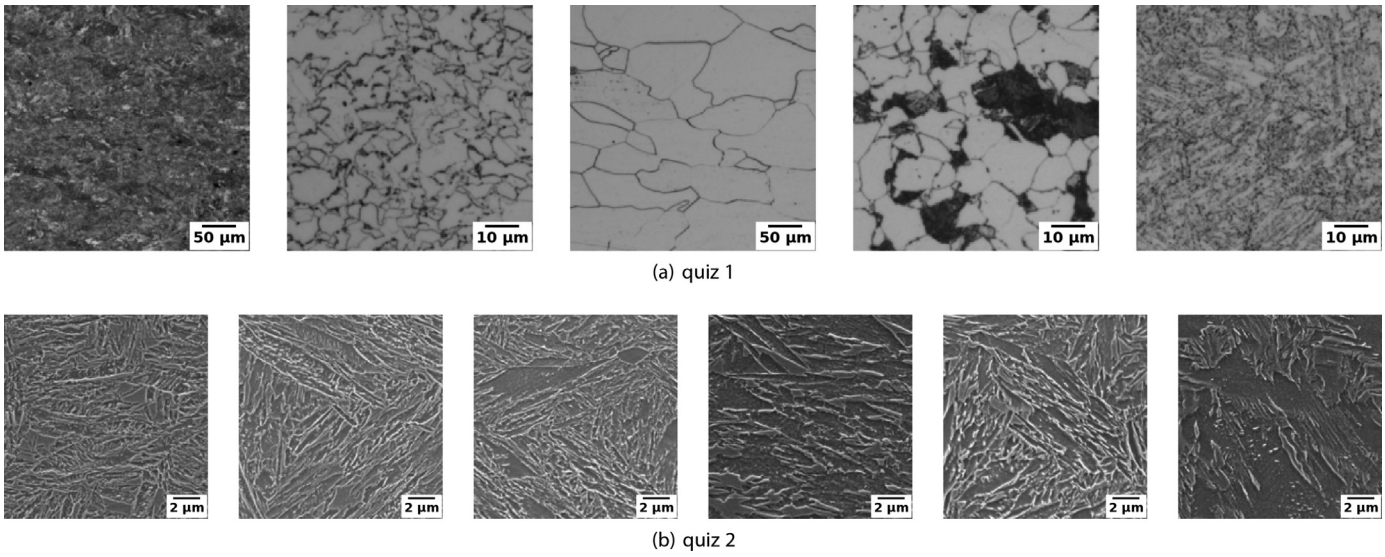

Fig. 1. An illustration of the microstructures for each of the pre-defined material classes for quiz 1 (a) and quiz 2 (b). More information on the material classes can be found in Supplementary S1 and S2.

of the six pre-defined classes of microstructures. Here, the option "None of these" is not included for the panel of experts, mainly to reduce the difficulty of the quiz. The training set contains 58 images in total, with magnifications of either 53 $ppµm$ or 106 $ppµm$. Some examples of these images are shown in Fig. 1 (b). In Supplementary S1 and S2, we give a description of each of the material classes and show the distribution of the magnifications for each microstructure type for both quizzes.

Since we want to train a model with only a few images of each microstructure class, it is necessary to include as many insights from material science as possible. For instance, the model should be robust to imaging conditions such as lighting, etching and dust particles. We achieve this by artificially modifying the images during training and in this way augmenting the information shown to the model. Concretely, we have made use of the following modifications:

Aggressive data augmentation transforms the images before they are used for training the model. These transformations include generic ones such as rotations, mirroring and changes in brightness. Additional transformations were conceived to better highlight the relevant features in a microscopy image. Local blurring of the image is used to mimic for instance a dust particle covering a certain part of the image. Edge detectors stress the importance of phase boundaries. Warping of the image explicitly includes the translational invariance of the image in combination with other transformations.

Crops at different length scales are used to further increase the number of images on which the model is trained. Given an original image, which has a pixel resolution of 1000 × 1200, a smaller crop of the image is randomly selected and either down- or up-sampled, so that the final crop used for training has a resolution of 200 × 200 pixels. Because of this, the model can learn to recognize the material at different length scales.

Explicit inclusion of the magnification serves as an additional input to the model. For each randomly selected crop, the number of dots per micrometer is computed and used as input for the deep learning model. This is deemed necessary, as in microscopy images there can be a large range of magnifications compared to classification problems in other fields of computer vision. For instance, in the training set for quiz 1 there is a factor of 265 difference between the smallest and the largest magnification. Note that this in fact mimics how human experts classify images, as they would always require a scale-bar in order to classify a material.

The machine learning approach used in this work consists of two steps: 1) obtaining the microstructural descriptors and 2) assigning these descriptors to a material class. In the first step, we train a deep neural network[5] to learn descriptors of a microstructure image through the use of triplet networks[14]. The neural network takes as input both a transformed crop $\mathbf{x_i}$ and its resolution l_i, expressed as dots per micrometer, and outputs a vector $\mathbf{f}(\mathbf{x_i}, l_i)$. We will refer to the elements of this vector as the descriptors of the microstructure image. The deep neural network has several millions of parameters, which are iteratively updated by minimizing for each crop $\mathbf{x_i^a}$ the triplet loss[15]:

$$\max\left(0, ||\mathbf{f}(\mathbf{x_i^a}, l_i^a) - \mathbf{f}(\mathbf{x_i^p}, l_i^p)|| - ||\mathbf{f}(\mathbf{x_i^a}, l_i^a) - \mathbf{f}(\mathbf{x_i^n}, l_i^n)|| + \alpha\right), \quad (1)$$

where ||.|| represents the euclidean norm, $\mathbf{x_i^a}$ and $\mathbf{x_i^p}$ are crops belonging to the same material class and $\mathbf{x_i^n}$ is a crop belonging to a different material class. The α is a positive, real number that represents the desired minimal distance between the descriptors of different material classes. Thus, this method learns to map a microstructure image to a vector representation of a desired dimensionality, so that similar types of microstructures have descriptors that lay close together and different types of microstructures have descriptors that are separated from each other. More explanation on how a triplet network works, can be found in Supplementary S3. Further details on the training procedure can be found in Supplementary S4. All the results we report are obtained using the deep learning library PyTorch[16].

Throughout this work, we will only use two-dimensional descriptors, as using such low-dimensional descriptors allows for training more robust classification models afterwards[17]. The output of the triplet network is shown in Fig. 2, where each coloured dot represents the descriptor of a crop in the training set. Dots belonging to the same class have the same colour. Clearly, the neural network has been well optimized, as descriptors of the same material class lay close together and far away from the other classes.

In the second step, we need to assign each of the descriptors to the right material class. A key challenge in the quizzes is the possibility for the images to belong to none of the pre-defined classes, so that we need to recognize when a given microstructure image differs from the images in the dataset on which the model was

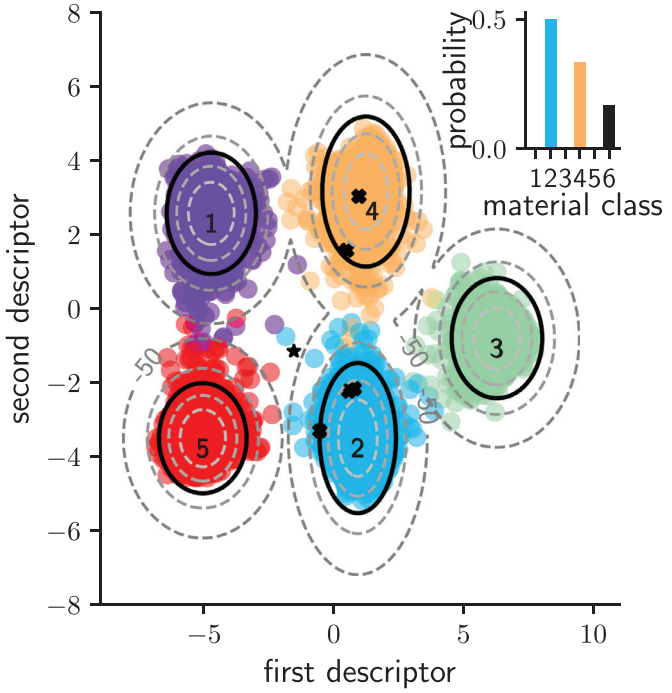

Fig. 2. Illustration of the used methodology. The coloured dots represent the descriptors of the crops used for training the model. Dots with the same colour belong to the same material class. The contours show the log-likelihood of the Gaussian mixture model that is fitted to these points. The black contour lines show the threshold for determining outliers. Points outside these contour lines are considered to belong to none of the pre-defined material classes. The five black crosses are crops from the same quiz image that are assigned to one of the classes, while the black star is a crop from that image that is considered an outlier. The resulting probability distribution for these crops is shown in the inset, where the "None of these" option is the sixth class. As most of the crops are assigned to class two, the model would predict the quiz image to belong to that class with a probability of 50%. Best viewed in colour.

trained. To do so, we fit a multivariate Gaussian distribution to each of the class centres of the descriptors using the Gaussian mixture model implementation in scikit-learn[18]. The resulting contours of the likelihood of the Gaussian mixture model are shown in Fig. 2. Crops that have a descriptor with a likelihood under a certain threshold are considered to belong to the "None of these" class. Details on the fitting procedure can be found in Supplementary S4 and details on the determination of the threshold can be found in Supplementary S5.

To be consistent with the training procedure, the model can only deal with images of size 200 × 200. During the evaluation of the quiz images, we therefore randomly select a number of crops from the image. After applying image augmentation to these crops, the microstructural descriptors are computed using the triplet network. These descriptors are then passed to the Gaussian mixture model to obtain for each crop a probability for each material class. This is shown in Fig. 2, where the black crosses and star represent crops belonging to the same image. After averaging the probabilities of the crops belonging to the same image, we obtain a probability distribution for the entire image. As averaging over more crops is preferable to obtain a meaningful probability distribution, we evaluate a thousand crops per image. By using a probability distribution rather than a hard decision, we can assess how certain the model is in its predictions.

To evaluate the performance of the proposed method, we hold two quizzes for both human participants and the deep learning model. For the first quiz, the human panel consists of 21 experienced metallurgists, to whom we will refer as "experts", and at 56 additional participants who don't necessarily have any experience in analysing microstructure images. We refer to this second group of participants as "non-experts". In the second quiz, 20 experts and 27 non-experts participated. The quiz was sent through a link and there was no time limit to answer the questions. Although we use the word "expert" for the metallurgists, we emphasize that especially for the second quiz, not everyone is equally familiar with the microstructures under consideration. Both panels were shown some representative example images for each material class, but they did not see the entire training set.

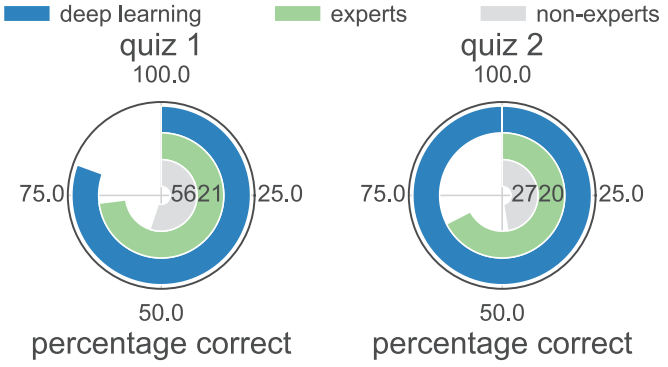

Fig. 3. Comparison of the classification accuracies of the panel of experts and the deep learning models for both quizzes. The numbers in the bars indicate the number of participants.

In Fig. 3, we compare the classification accuracy of the non-expert panel, the expert panel and the deep learning model. Unsurprisingly, the experts beat the non-experts by a large margin in both quizzes. The deep learning model achieved the highest accuracy in both cases. For the first quiz, the experts classified on average 73% of the images correctly, while the deep learning model achieves an accuracy of 81%. The best performing expert managed to obtain a score of 86%, outperforming the model by 5 percentage points. Still, we can conclude that the model is competitive with the best experts. For the second quiz, the experts obtained an average score of 67%, while the deep learning model classified all images correctly. The best expert in the panel also managed to obtain a perfect classification score. It is remarkable how well the model performs considering the small amount of data. Because deep learning models consider every single pixel in the image, they are able to discern very small details, which are important for this kind of complex microstructures. Code to reproduce the results can be found on microstructuredb.com/papers.

In Supplementary S7 and S8, we have included an overview for every quiz image of the probabilities assigned to each class by the machine learning algorithm. It is interesting to compare this to how the experts voted. It is clear for quiz 1 that both the human expert and the deep learning algorithm struggle with the "None of these" option. This option confuses the majority of experts in three out of the in total four wrongly classified images, whereas the deep learning algorithm fails because of this option in six out of the in total seven wrongly classified images. One such an image is shown in Fig. 4 (a), where we see a pearlitic-ferritic structure at a low magnification. We have also included a saliency map that shows the pixel-level predictions of the model. These maps are obtained by averaging the predictions of many crops using the procedure outlined in [11]. Since the training set only contains images that have a magnification of at least a factor $x2.5$ higher, the model has to extrapolate and fails to correctly recognize the microstructure. This is also reflected in the completely dark saliency map as the model fails to recognise any microstructural feature in the image. For the human experts, who can fall back on their much larger mental library of images, it is easier to recognize that this is indeed a low-magnification image of a pearlitic-ferritic material.

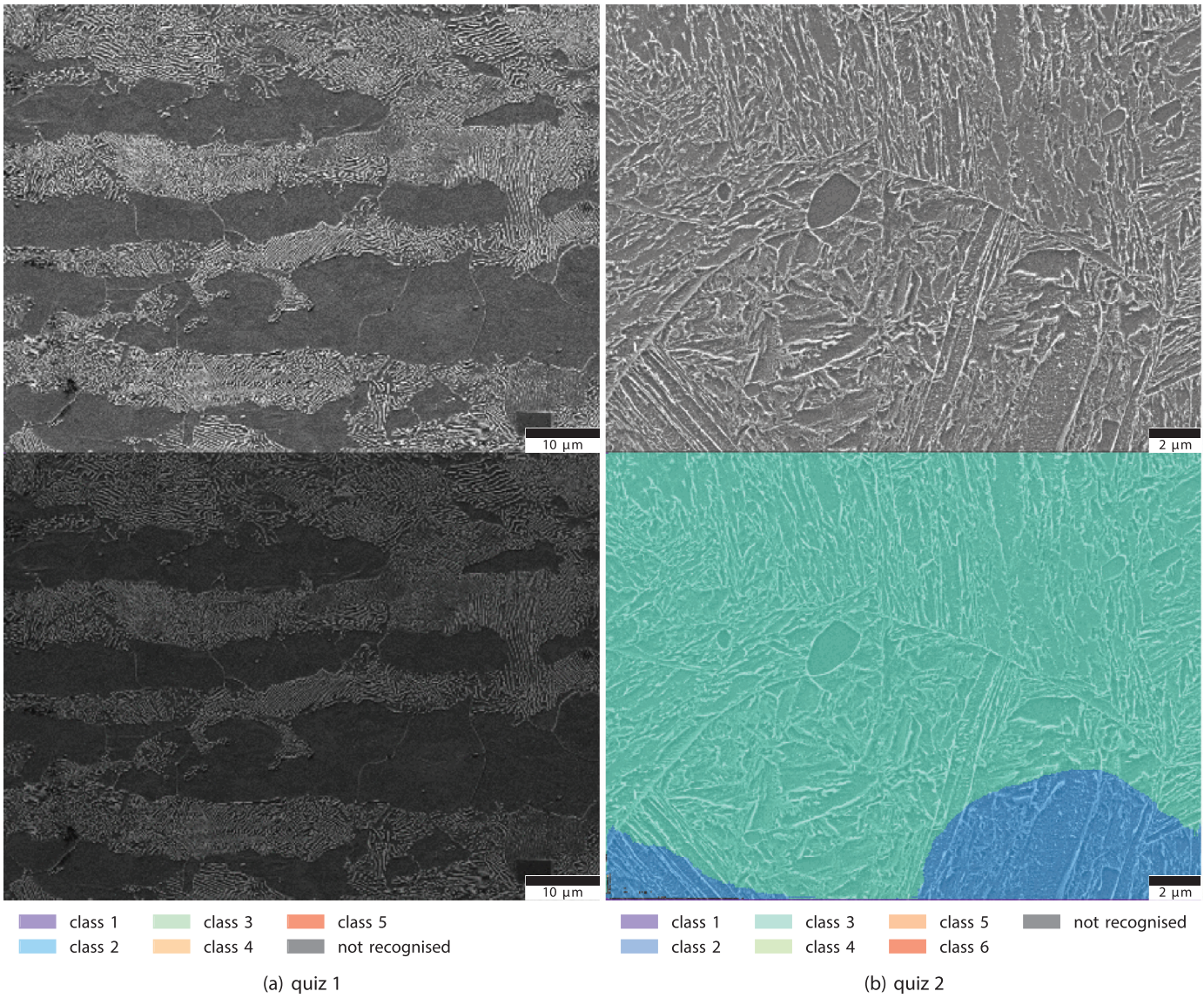

Fig. 4. Visual analysis of the model predictions using saliency maps. In both quizzes, there is a limited number of images where human experts outperform the deep learning algorithm, in contrast to the overall result. For each of the two quizzes, one of these images is shown here together with the saliency maps that provide pixel-level assignments of the model. The image on the left contains both ferrite and pearlite and hence belongs to class 4, whereas the model does not assign this image to one of the pre-defined classes. The saliency map is entirely dark, implying that no microstructural features are recognised. The image on the right features some sharp needle-like structures and hence belongs to class 3. The model correctly predicts this, although it is not very certain about its prediction, due to the presence of precipitates in the lower corners, which are typical features of materials from class 2. These regions are marked blue in the saliency map and hence assigned to class 2.

For the second quiz, we see that the machine learning model not only obtains a perfect classification score, but that is it also very confident in its predictions. The predictions in which it is least confident, are those for the images that belong to classes 3 and 5, which are the classes with the fewest training images. There is only one image where the human experts are clearly more confident in their prediction and this image shown in Fig. 4 (b). Due to the presence of some martensitic regions with a small amount of carbides in the lower right part of the image, it is indeed understandable for the model to assign some probability to class 2, which is the only class with such fine carbides. This is also shown in the saliency map, where the precipitate-containing regions are marked blue. In Supplementary S9 and S10, we discuss some more images and their saliency maps. In line with the findings in [11], we find that the model autonomously learns the importance of microstructural features such as the grain size, the grain shape and the presence of precipitates.

The results above are promising, as they show that even when small datasets of images are available deep learning models are capable of analysing microstructure images at least on par with human experts. As the deep learning methods will become even more performant, we anticipate that human experts will soon struggle to keep up with deep learning models and that such models will play a crucial role in the analysis of microstructure images.

In this report we illustrated how a deep learning model can be used to recognize structures even when only small datasets are available. We tailored practices from other fields in computer vision to the problem of microstructure recognition by introducing appropriate data augmentations and by explicitly including information about the magnification in the model. We showed how the models we trained can outperform a panel of experts and concluded that the use of deep learning is especially effective in the study of complex martensitic microstructures.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

M.L. and S.C. acknowledge financial support from OCAS NV by an OCAS-sponsored PhD position and by an OCAS-endowed chair at Ghent University, respectively. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.scriptamat.2020.10.026

References

[1] D.T. Fullwood, S.R. Niezgoda, S.R. Kalidindi, Acta Mater. 56 (2008) 942–948.
[2] A. Chowdhury, E. Kautz, B. Yener, D. Lewis, Computational Mater. Sci. 123 (2016) 176–187.
[3] B.L. DeCost, E.A. Holm, Computational Mater. Sci. 110 (2015) 126–133.
[4] J. Gola, J. Webel, D. Britz, A. Guitar, T. Staudt, M. Winter, F. Mücklich, Computational Mater. Sci. 160 (2019) 186–196.
[5] J. Schmidhuber, Neural networks 61 (2015) 85–117.
[6] D. Cireşan, U. Meier, J. Masci, J. Schmidhuber, Neural Networks 32 (2012) 333–338.
[7] S.M. Azimi, D. Britz, M. Engstler, M. Fritz, F. Mücklich, Sci. Rep. 8 (2018) 1–14.
[8] B.L. DeCost, B. Lei, T. Francis, E.A. Holm, Microsc. Microanal. 25 (2019) 21–29.
[9] F. Ajioka, Z.-L. Wang, T. Ogawa, Y. Adachi, ISIJ Int. 60 (2020) 954–959.
[10] B.L. DeCost, T. Francis, E.A. Holm, Acta Mater. 133 (2017) 30–40.
[11] M. Larmuseau, M. Sluydts, K. Theuwissen, L. Duprez, T. Dhaene, S. Cottenier, NPJ Comput. Mater. 6 (2020) 1–11.
[12] H. Tang, X. Chen, Y. Liu, Z. Lu, J. You, M. Yang, S. Yao, G. Zhao, Y. Xu, T. Chen, et al., Nature Machine Intelligence 1 (2019) 1–12.
[13] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'xDonoghue, D. Visentin, et al., Nat. Med. 24 (2018) 1342–1350.
[14] E. Hoffer, N. Ailon, in: International Workshop on Similarity-Based Pattern Recognit., Springer International Publishing, Cham, 2015, pp. 84–92.
[15] F. Schroff, D. Kalenichenko, J. Philbin, in: IEEE Conference on Computer Vision and Pattern Recognit., IEEE, New York, 2015, pp. 815–823.
[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., in: Advances in neural information processing systems, Curran Associates, Brooklyn, 2019, pp. 8026–8037.
[17] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Springer, New York, 2001.
[18] S. Van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, PeerJ 2 (2014) e453.