CLASSIFICATION OF ROLLED METAL DEFECTS USING RESIDUAL NEURAL NETWORKS

Ihor Konovalenko[1], Pavlo Maruschak[1], Lyubomyr Mosiy[1], Frantisek Duchon[2], Michal Kelemen[3]

[1] Department of Industrial Automation, Ternopil National Ivan Puluj Technical University, Rus'ka str. 56, 46001 Ternopil, Ukraine; icxxan@gmail.com (I.K.); maruschak.tu.edu@gmail.com (P.M.)
[2] Slovak University of Technology in Bratislava, Ilkovičova 3, SK-812 19, Bratislava; Slovak Republic, E-mail: frantisek.duchon@stuba.sk
[3] Technical University of Kosice, Letna 9, 04200, Kosice, Slovak Republic, E-mail: michal.kelemen@tuke.sk

Abstract: The authors investigated deep residual neural networks, which are used to detect and classify defects found on the rolled metal surface. Based on the neural network with ResNet152 architecture, a classifier for recognizing defects of three classes was built. The proposed technique allows recognizing and classifying surface damage with high accuracy in real-time based on its image. The average binary accuracy of the classification made based on the test data is 97.3%. Neuron activation fields were studied in the convolutional layers of the model. The results obtained show that areas, which correspond to those with damage in the image, are activated. False-positive and false-negative cases of classifier application are investigated. Errors were found to occur most frequently in ambiguous situations when surface artefacts of different types are similar.

Keywords: metallurgy; steel sheet; surface defects; visual inspection technology; classification; neural network.

1. Introduction

The condition of its surface layer, to a large extent, evaluates rolled metal quality. Tasks related to eliminating damage on the surface layer are most relevant for improving the quality of finished products in metallurgy (Luo, 2016; Zhou, 201; Dhua, 2019). At the same time, automated control systems are most promising, as they ensure maximum sensitivity and resolution, high performance, accuracy and ease of realisation. Neural network technologies are used to develop modern algorithms for high accuracy defects (Chen, 2020; Lee, 2019; Fang, 2020). During training, they make it possible to automatically generate datasets that describe defects of different classes and then detect them in new images effectively.

To date, regulations that provide for an unambiguous classification of rolled metal surface defects (GOST 21014-88) are limited in number. Approaches to evaluating rolled metal based on such standardized features allow for an averaged analysis of defects only. Therefore, as evidenced by Tao (2018) and Kostenetskiy (2019), finding new approaches to the analysis of surface defects is relevant.

Our research aims to develop a neural network classifier, which can detect defects of different types with high accuracy in near real-time and classify them according to a given set of defect types.

2. Training datasets and their peculiarities

The sample for training the neural network classifier was formed using photo images of rolled metal steel surfaces with various defects. Most images were taken from a dataset provided in 2019 by Severstal, a Russian public company that participated in the international Kaggle competition. In total, we selected 14830 images of steel surfaces, both undamaged and with defects of three types. All images were reviewed and marked by experts using a specially developed application – Image Labeling Tool. The dataset formed this way was divided into three parts: the training part, which contained 10482 images, the validation part (2790 images), and the test part (1558 images). Images for all three parts were selected so that the ratio of classes was the same. Training and validation parts were used during training, and the test part was used during testing the model on unknown images.

The images below show the defects of the three classes. Each of them primarily differs in occurrence (Konovalenko, 2020), which causes morphological and, consequently, visual differences. Examples of images of surface fragments with defects of different classes are shown in Fig. 1, and their description is presented in Table 1.

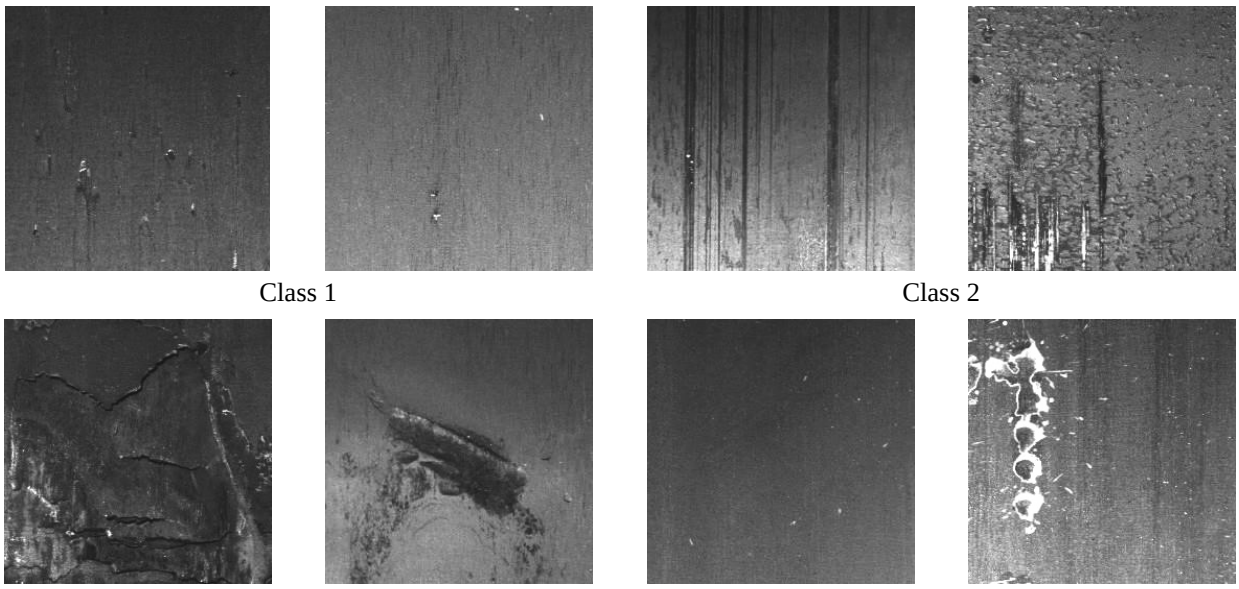


Fig. 1. Classes of surface defects

Table 1. Classes of defects found on rolled metal

Defect class	Description
I	Small round defects, in particular, rolled-in scales. Other damage of this type can be formed by pressing a solid body into the surface. Most often, such defects have no pronounced orientation. Sometimes they appear in groups. Areas of damage of the first class are usually the smallest ones compared to defects of other classes.
II	Lines, scratches and abrasions. They can be both single and multiple. Scratches are usually oriented in the direction of rolling. Defects of this type are usually caused by friction of the metallurgical equipment against the sheet surface. Defects of this class are most numerous, have different shapes, directions and staining. Scratches of the second class can be very small lines (which sometimes make them look like defects of the first class) and large ones passing through the entire surface of the image.
III	Rolling films, which have the form of surface "tears-out" with uneven edges. They usually have no orientation, can be of any shape and occupy a significant area on the surface.

It is noteworthy that undamaged surfaces are also characterized by high inhomogeneity. They are distinguished by a significant variability of morphological and visual features. Such surfaces are often texturally homogeneous but may contain glare; sometimes, they are strongly illuminated or darkened in places. Undamaged areas often contain rough formations with different structures or various coloured artefacts of different shapes. Industrial images may contain the specimen edge and background behind them, which is not informative in defectometry. Some undamaged surfaces contain corrugations.

The topology, types and shape of defects of different classes are very diverse, which causes difficulty in classifying them even at the expert level. The similarity between small defects of the first and second classes is most common. Small defects contain fewer morphological features; therefore, it is more difficult to form sets of characteristics that can be reliably distinguished. They often occur in adjacent parts of the image or even form a single structure of the defect. The elements of specific surface images of defect-free surfaces are similar to damage of a particular class, further complicating the situation. In addition, the damage is often located on a structurally inhomogeneous surface, which also makes it challenging to identify it. Furthermore, the second- and third-class defects can end up in a smooth gradient, which complicates their identification in border areas. To bring the working conditions of the classifier closer to the industrial ones, all such cases are sufficiently presented in the training sample.

The frequency of defects of different classes is different in the production environment; therefore, various defects are presented unevenly in the training dataset. In total, 1919 images with first-class defects, 6667 images with second class defects, and 6238 images with third-class defects were used for training. However, many images contain defects of several classes at a time. Most images contain defect-free surfaces.

3. Methodology for training neural network classifier

Based on the previously obtained results (Konovalenko, 2020), the neural network model ResNet152 presented in 2015 by He *et al.* was chosen as the basis for the classifier. Since 2015, networks based on ResNet architecture have shown excellent results in various areas related to image processing. The main feature of ResNet models is shortcut connections that transmit the input directly to the end of the residual block. This makes it possible to significantly reduce the vanishing gradient problem in deeper layers of the network and increase the model depth.

The input layer of the classifier receives an image of 256 × 256 pixels. Blocks of ResNet model layers follow this. However, the last ResNet model layer was replaced by a fully connected layer of 3 neurons. Each of the neurons is responsible for a defect of its class. Sigmoid activation function was used for output neurons. A defect of a specific class present on the input image causes the appearance of 1 on the corresponding output neuron (or 0 – in the opposite case). The model used is a multilabel classifier, which makes it possible to identify defects of several classes on the same image.

The augmentation technique was used to diversify the training sample. This allowed us to increase the training sample significantly and thus contribute to developing the best generalizing properties of the model (Luo, 2016). Previous research aimed at solving our problem also showed that augmentation allows attaining better performance of the model. Therefore, in addition to standard augmentation methods (horizontal and vertical flip, rotation at an angle multiple of 90º). In this case, sections measuring 256 × 256 pixels were selected randomly on the input images during training (according to input layer shape). From these sections, the tensor was formed, which was fed to the neural network input. This made it possible to significantly diversify training data (especially insufficient ones) and provide conditions under which training batches will never be repeated in practice.

Another feature of training data is a significant imbalance of images, especially "defect – defectless" ones. This is primarily due to the varying frequency of defects of different types in the production environment. Firstly, most images do not damage (this corresponds to an actual situation when production shortages are not very common). Secondly, many defects occupy a small area of the image, so frames can also get into undamaged areas when using the random crop technique. In addition, more significant defects are more likely to fall into random frames.

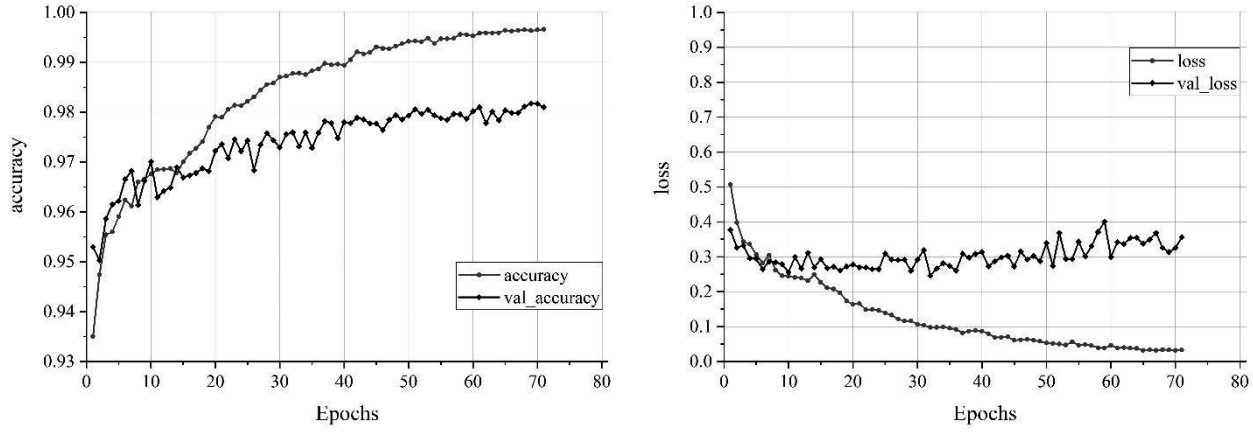
While training, the majority class (class 2 in our case) quickly becomes well-classified since we have much more data. Therefore, the focal loss function proposed by Lin *et al.* in 2017 was used to solve unbalanced classes. Thus, to ensure that we also achieve high accuracy for minority classes, we use the focal loss function to give these minority classes examples of more relative weight during training. Focal loss applies a modulating term to the cross-entropy loss to focus training on hard detective examples. It down-weights the well-classified examples and puts more training emphasis on the data that is hard to classify.

ResNet classifiers were realised in Python 3.8 language using the Keras and TensorFlow libraries. We used a workstation based on Intel Core and 7-2600 CPU for training and testing and two NVIDIA GeForce GTX 1060 GPUs with 6 GiB of video memory.

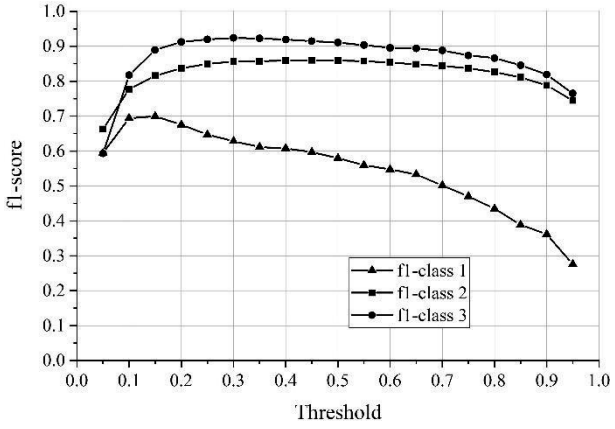
4. Best models result

About 30 neural network models based on ResNet36, ResNet50 and ResNet152 architectures were developed and trained. The training was performed at different hyperparameters of the model. Each of the trained models was investigated on test data. A model with ResNet152 architecture achieved the best result. It has a greater depth, which allows forming a complete set of features for the defects of each class.

The graph, which illustrates the model's training process, is shown in Fig. 2, a, b. Over several epochs, the model attains its maximum level of generalization, and validation loss begins increasing gradually, while training loss keeps decreasing. During the training of both models, validation binary accuracy varied from approximately 0.95 to 0.98.



a b



c

Fig. 2. Variation of binary accuracy metric (a) and focal loss function (b) during training; graph *f1-score* (c) versus threshold value on the source layer for different classes of defects

However, the accuracy metric does not fully reflect the quality of the model. Since the distribution of images by class is unbalanced, and defect-free images dominate the training sample, the accuracy metric primarily shows the success of detecting defect-free images. Therefore, other metrics were considered to study the model quality, including the recall, precision, and f1-score metrics.

The recall metric ($Recall = TP/(TP + FN)$) shows the proportion of images with defects of a specific class, which the model recognizes as defects of this class. The highest recall value was attained for defects of third class - 0.908. This means that the model detects almost 91% of the damage of this class. Defects of the second class are recognized a little worse (the recall value is 0.824). Minor defects of the first class are the most difficult to classify - the recall value is equal to 0.627. Experiments have shown that defect-free specimens are the easiest to recognize - more than 97% are detected correctly.

The precision metric ($Precision = TP/(TP + FP)$) shows what proportion of images recognised as defects of a specific class contain such defects. The highest values of this metric were obtained for defects of the third class. High values were also obtained for defects of the second class. The worst accuracy of detection was observed in the case of defects of the first class.

The sigmoid activation function is used in the initial layer of classifiers. Therefore, at the output of each neuron, the value is in the range of $[0\ldots 1]$. Whether a specific pixel of an image belongs to a specific class of damage is made if the value of the corresponding output neuron exceeds a certain threshold. The value of this threshold affects the model quality metrics for this class of defects. To select the optimal threshold value for each class, the f1-score metric was calculated.

The integral f1-score metric ($f1 = 2 \cdot (Precision \cdot Recall)/(Precision + Recall)$) is a harmonic mean of the precision and recall metrics and generally characterizes the model's ability to recognize defects of a specific class. Graphs showing f1-score for three classes of defects are presented in Fig. 2,c. Small defects of the first class are assumed to be the most difficult to identify: in addition to being located below other classes, the maximum of the f1 curve is the fastest to decrease after the maximum. This is consistent with the above-mentioned morphological features of defects of the first class: they are small and may resemble either defects of other classes or the surface formations of defect-free zones, which leads to difficulties in identifying them. Most stable is identifying defects of the second class: a significant area of the corresponding curves is almost horizontal. This indicates that the model is "confident" when recognizing defects of the second-class in a wide range of thresholds. The curve of the f1 metric attains the highest maximum for defects of the third class. Therefore, the set of features of this class must be the easiest to be detected by the model and most different from the features of other classes.

Based on the data obtained, the optimal threshold values were selected for each class.

The model quality metrics determined based on the test data are given in Tables 2 and 3. Table 2 contains the values of binary accuracy (acc), precision (prc), recall (rcl) and f1-score for defects of the classes considered. Table 3 contains the same metrics for defect-free images.

Table 2. Quality metrics of classifiers determined based on test data for defects of various kinds

Class 1				Class 2				Class 3			
acc	prc	rcl	f1	acc	prc	rcl	f1	acc	prc	rcl	f1
0.976	0.792	0.627	0.700	0.945	0.897	0.824	0.859	0.993	0.939	0.908	0.923

Table 3. Quality metrics of classifiers determined based on test data for defect-free images

Defectless			
acc	prc	rcl	f1
0.938	0.944	0.973	0.958

5. Model feature maps

To better understand the operation of the model, we investigated feature maps formed by convolutional layers of the developed classifier. As François Chollet (2017) showed, the activation pattern of intermediate neurons reflects how successfully the neural network converts the input signal. It also shows how the input image is decomposed by intermediate filters of different layers formed during training. The initial layers of the model contain the whole image but with an emphasis on specific areas. At this stage, the model retains most of the information from the input image but already focuses on its most interesting features in terms of classification. The deeper we move through the convoluted layers of the model, the more abstract the picture of neuronal activation becomes. Moreover, it acquires more and more image elements inherent in a particular class of defects.

Most interesting and illustrative are feature maps from the last convolutional layer. From it, the data are fed to the last generalizing full-connected layer, which in the long run forms the decision about the presence of a particular type of damage and activates the corresponding output neuron.

This study showed that defects of different topologies are quite fully represented on the feature maps. This indicates that the model attains good generalizing properties and makes it possible to build tools for semantic segmentation, which allow locating damage and calculating its geometric characteristics (size, area, shape, etc.) in addition to assigning a class label to the image.

6. Problems of defect detection

From the perspective of the practical application of the model, it is crucial to understand what difficulties arise in its application and how it can give the wrong result. To investigate this, a test data set was prepared based on the test images. From each test image, ten random plots measuring 256 × 256 pixels (according to the input layer size of the model) were selected at random. To ensure an unambiguous identification of the results of the model operation, only those specimens were selected from the prepared test data set, which contain defects of only one class (such specimens make up the majority). The developed classifier was applied to the prepared array of image areas, and a confusion matrix was formed based on the results obtained (Fig. 3).

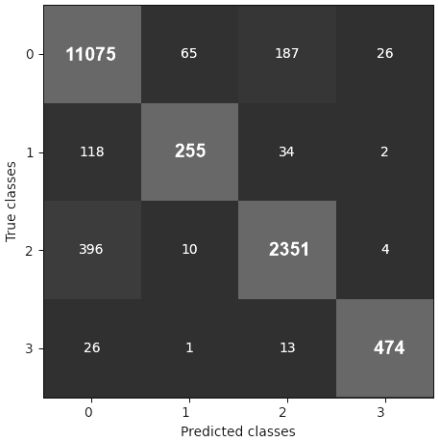
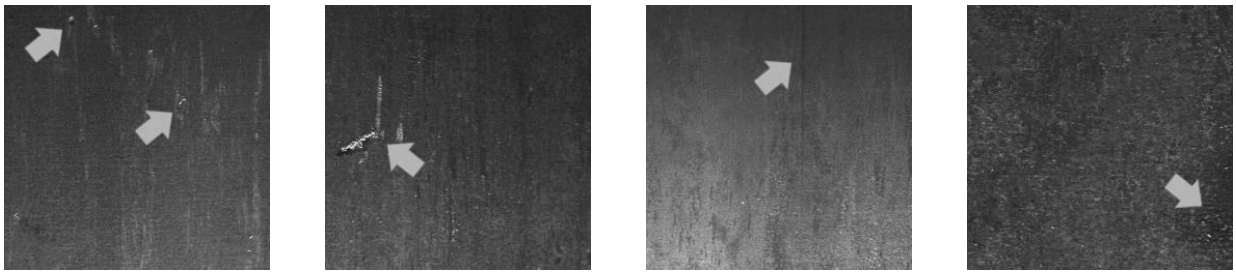


Fig. 3. Confusion matrix for images with defects of different classes

The results show that most misclassification problems arise in case of defects of the first class: a significant part of them (~29% of their total number) is recognized as a defect-free surface. Examples of such images are given in Fig. 4a. Most cases involve very small surface artefacts that merge with the surrounding background and are visible only when the image is magnified or similar to a defect-free surface's textural features. Some (~8%) defects of the first class are recognized as the second class, and very few - as the third class. This is because defects of the first and third classes differ primarily in area. Figure 4b presents examples of the first-class recognized as the second class. The vast majority of such cases involve defects similar to small scratches or abrasions of the second-class surface.

Thus, we can conclude that a relatively low classification result for defects of the first class is mainly due to their similarity with the morphological formations of defect-free surfaces and minor defects of the second class. Particularly noteworthy is that differences between these groups of surface formations are classified ambiguously even at the expert level.



a b c d

Fig. 4. Examples of incorrect recognition of defects

The model perceives a particular part of defects of the second class (~14%) as a defect-free surface. At the same time, a minimal number (less than half a per cent) of these defects are mistakenly recognized as defects of other classes. Figure 4c shows examples of images with defects of the second class recognized as defect-free surfaces. Most of such cases involve mild defects that merge with light glare or are shaded. In some cases, we deal with marginal fragments, where damage is weakly expressed. Cases where damage is clearly visible, but its morphology is atypical for defects of the second class, are the smallest in number.

The classifier is the best at identifying defects of the third class. Of these, ~5% are perceived as defect-free surfaces by the model, and ~3% - as defects of another (primarily second) class. Figure 4d shows examples of defects of the third class recognized as defect-free surfaces. A vast majority of such cases involve marginal fragments of defects, where their morphological and visual features are weakly expressed. It is much less common in cases where a defect is well pronounced, but its area is so large that it resembles the background.

Defect-free surfaces are recognized correctly in approximately 98% of cases. Most often, the model perceives artefacts of such surfaces as defects of the second class (~1.6% of cases). As evidenced by the analysis of results, in such cases, the image contains artefacts that resemble the second class's defects. Experts classify such morphological formations on the image ambiguously (without investigating an actual surface).

Conclusions

A classifier for recognizing defects of various classes that occur on rolled metal surfaces has been developed and studied on a test dataset. The classifier is based on a deep convolutional neural network with ResNet152 architecture. The proposed technique allows classifying images with high accuracy in near real-time while recognizing three classes of defects. The average binary accuracy of the classification made based on test data is 97.3% for all images (including those containing defect-free surfaces). The model was found to be the best at detecting defects of the third class and undamaged surfaces.

The study of false negative and false positive classification cases showed that errors most often occur in the case of significant visual similarity of surface artefacts of different types or very small defects. However, even the expert assessment of images is ambiguous in such cases, and different experts may come to different conclusions. Such shortcomings can be eliminated by training the model on images with higher resolution.

The study of neuron activation fields in the convolutional layers of the model has revealed that feature maps reflect the location, size and shape of the objects of interest very well. This makes it possible to develop a model for semantic segmentation of defective images based on the architecture proposed. In addition to detecting and classifying defects, this model can localize them in the image, thus allowing the calculation of their area and other spatial characteristics.

References

1. Chen, H., Hu, Q., Zhai, B. et al. (2020). A robust weakly supervised learning of deep Conv-Nets for surface defect inspection. *Neural Comput & Applic, 32*, 11229–11244. doi:10.1007/s00521-020-04819-5

2. Dhua, S.K. (2019). Metallurgical analyses of surface defects in cold-rolled steel sheets. *J Fail. Anal. and Preven, 19*, 1023–1033. doi:10.1007/s11668-019-00690-2

3. Fang, X., Luo, Q., Zhou, B., Li, C., Tian, L. (2020). Research progress of automated visual surface defect detection for industrial metal planar materials. *Sensors , 20*, 5136, doi:10.3390/s20185136

4. François Chollet. (2017). *Deep Learning with Python*. Manning Publications.

5. GOST 21014-88. (1989). Rolled Products of Ferrous Metals. Surface Defects. Terms and Definitions; Izd. Stand.: Moscow, USSR; p. 61. (In Russian)

6. He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv, arXiv:1512.03385v1.

7. Kaggle Severstal: Steel Defect Detection. Can You Detect and Classify Defects in Steel? (2019). *Kaggle*. Retrieved from https://www.kaggle.com/c/severstal-steel-defect-detection.

8. Konovalenko, I., Maruschak, P., Brezinová, J., Viňáš, J., Brezina, J. (2020). Steel Surface Defect Classification Using Deep Residual Neural Network. *Metals*, *10*, 846.

9. Kostenetskiy, P., Alkapov, R., Vetoshkin, N., Chulkevich, R., Napolskikh, I., Poponin, O. (2019). Real-time system for automatic cold strip surface defect detection. *FME Trans. 47*, 765–774. doi:10.5937/fmet1904765K.

10. Lee, S.Y., Tama, B.A., Moon, S.J., Lee, S. (2019). Steel Surface Defect Diagnostics Using Deep Convolutional Neural Network and Class Activation Map. *Appl. Sci. 9*, 5449, doi:10.3390/app9245449.

11. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal Loss for Dense Object Detection. arXiv, arXiv:1708.02002v2.

12. Luo, Q., He, Y. (2016). A cost-effective and automatic surface defect inspection system for hot-rolled flat steel, *Robotics and Computer-Integrated Manufacturing*, *38*, 16-30.

13. Takahashi, R., Matsubara, T., Uehara, K. (2018). RICAP: Random Image Cropping and Patching Data Augmentation for Deep CNNs. *Conference on Machine Learning*, Proceedings of The 10th Asian Conference.

14. Tao, X., Zhang, D., Ma, W., Liu, X., Xu, D. (2018). Automatic metallic surface defect detection and recognition with convolutional neural networks. *Appl. Sci. 8*, 1575, doi:10.3390/app8091575.

15. Zhou, S., Chen, Y., Zhang, D., Xie, J., Zhou, Y. (2017). Classification of surface defects on steel sheet using convolutional neural networks. *Mater Technology*, *51*(1):123.