Extract from STEEL + TECHNOLOGY (2019), issue-no. 4, page 64 – 67.

Practical improvements in surface inspection

Classifying defects more reliably

While AI-based technologies can now outperform humans in many categories of image processing, the "experienced" reliability of surface inspection systems in the steel industry often leaves much to be desired. In apparent contradiction to this, the very same systems are able to achieve high classification accuracy on controlled test data. A structured optimization of datasets and classifiers using deep learning technology drastically increases the practical performance of existing systems. For the best possible performance, surface defects must be classified multiple times throughout production.

Modern requirements for cold-rolled flat steel products necessitate ever-higher production quality, especially for applications in the automotive industry. Manual inspection through the complete production process is often not possible due to both cost and organizational restrictions. For this reason, large steel manufacturers are dependent on the use of automated visual surface inspection systems. These systems consist of cameras and illumination devices located above and below the steel strip, hardware and software for defect detection and classification, as well as a software interface for the user. The correct classification, i.e. the determination of an image's surface defect category, is the critical operational piece of the system. This classification itself is ultimately what enables the system to, for example, distinguish non-metallic inclusions from superficial dirt.

Typically, the manufacturers of the surface inspection systems follow the steps of installing the systems on-site, optimizing the classifiers, and training future users. After acceptance, factory staff is then responsible for further optimization of these systems, with occasional support from system manufacturers. Frustration concerning poor classifier performance is often directed towards these internal employees, despite that they have tuned the classifiers optimally given their information constraints. To make matters worse, the technologies behind the operation of these classification systems are trade secrets of the inspection system manufacturers and only the most rudimentary optimization tools are made available for use. Making improvements to the identified weak points of the systems is practically impossible in many cases.

Both the surface inspection system manufacturer's specifications and scientific literature give consistently high values for classification accuracies. Often, accuracy levels of 95% or more are declared. However, the employees who are responsible for quality control find that, in production, the classification results tend to be much more unreliable.

In the following section, the most important reasons behind this discrepancy are explained and possible solutions are presented. Smart Steel Technologies (SST) offers a wide range of software tools that significantly improve classification performance in practice without the need to replace existing inspection systems consisting of expensive integrated imaging systems (**figure 1**).

Training and test sets

To build a classifier that can properly classify all types of surface defects, training sets and test sets that contain represent-

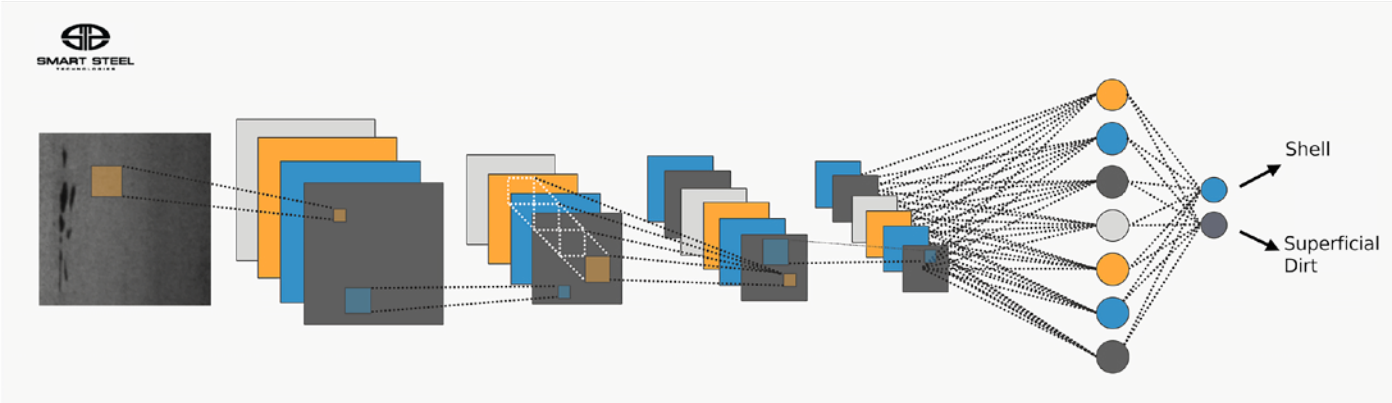

Figure 1. Deep learning technology is superior to conventional methods of surface defect classification (Picture: SST)

Authors: Dr. Falk-Florian Henrich, Founder and CEO; Dr. Otmar Jannasch, VP Metallurgy; Dr. Jan Daldrop, Machine Learning Engineer; Selim Arikan, Machine Learning Engineer; Matt Fabina, Director of US Operations; Smart Steel Technologies, Berlin, Germany – Contact: henrich@smart-steel-technologies.com

ative images of all relevant defect types are required. The classifier is then trained using the images in the training set, and the classification performance is validated by the images in the test set.

Individual surface defect types on hot and cold rolled strips come in a variety of shapes and structures. At the same time, disparate defect categories (e.g. shells and scratches) sometimes have similar appearances and visual characteristics. Additionally, if individual variants (subclasses) are missing in the training set, they are not reliably learned and detected by the classifiers.

The physical appearance of defect images depends on many metallurgical factors. One example is the steel grade of a coil, which directly affects the visual characteristics of both the steel surface and the actual defects. Additionally, factors such as the environmental conditions of the cameras (e.g. dampness, condensation, dirt) may affect the visibility of the defect in the images. Ideally, each defect category should be represented by images captured in each of these environmental variations.

Another important aspect of training set construction is the selection of images over an extensive production period. A typical confounding effect is a concentration of images for a defect category among only a few coils. When a classifier is trained using this unbalanced and biased data, there is a danger that the classifier will "learn" the characteristics of the defect only within the specific context of the few sampled coils.

When training and test sets are sampled from an overly-homogenous data set, the validation accuracy of the classifier on the test data is artificially inflated. If for example, exceptional visual cases are missing in the training set, the classifier will have never been exposed to them. One way to mitigate this effect is to use random samples of images for both the training set and test set. However, this naive randomization may introduce unintended consequences. For example, if in this random scheme a classifier is by chance only shown oil flecks on certain steel grades, it may naively associate oil stains with visual characteristics related only to those steel grades. The classifier would also perform artificially well on such a test set since the training set and test set would be randomly sampled from the same base dataset. This phenomenon poses a major problem in practice. Real-world cases of training set and test set splits have produced classification accuracies of up to 80% using only the metadata associated with the defect image, such as steel grade and production time.

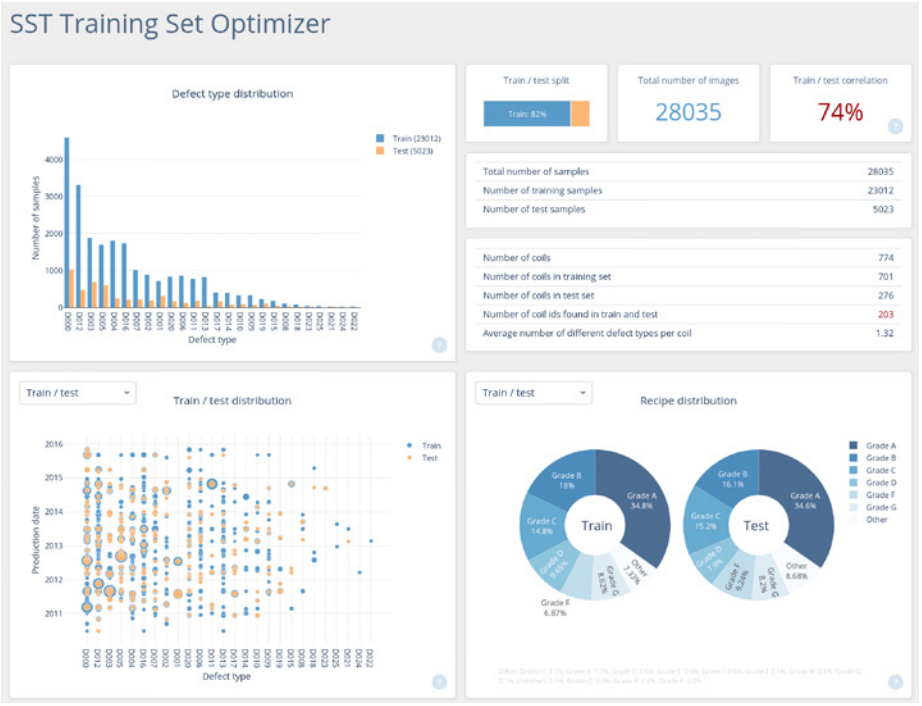

Figure 2. The SST Training Set Explorer can be used to create high-quality and representative training and test sets (Picture: SST)

Even though the classifiers in these examples may achieve high accuracy levels in the validation phase, this performance will fail to be reproduced under real, long-term production conditions.

Last but not least, the size of the training set is a decisive factor for classification accuracy. Yet, in practice, many training and test sets are relatively small. Investing time and effort to create a large dataset is always worthwhile, especially when modern classifier technologies are used, as they can capture both the large variance within classes and the subtle differences between different defect types.

Smart Steel Technologies has developed several software tools for properly creating effective training sets and test sets. The SST Training Set Optimizer (**figure 2**) is a statistical analysis tool that monitors the composition of training sets and test sets, enabling the efficient addition of defect images. In addition to the class distribution, the chosen production period and steel grades per class can be monitored. A train-test correlation is automatically calculated based on metadata. Instead of evaluating the datasets by naive statistics (e.g. total number of images), a comprehensive overview of the training set and test set condition is presented instead, including potential steps to take for improvement of classification systems.

To quickly find and identify suitable images to training sets and test sets, Smart Steel Technologies has developed a deep learning based image search engine (**figure 3**) that allows millions of defect images to be searched for visual similarity in fractions of a second. Using the built-in filters for specific steel grades and production periods, hundreds of images can be found and added to the training or test set within a few hours. The image search effectively finds suitable defect images even for the rarest surface defects.

Defect distributions

There exists a natural, significant class imbalance in steel production: some defect image categories such as pseudo-defects or superficial dirt occur very frequently, while more severe defects are relatively rare. Classification systems must be able to effectively identify the few truly

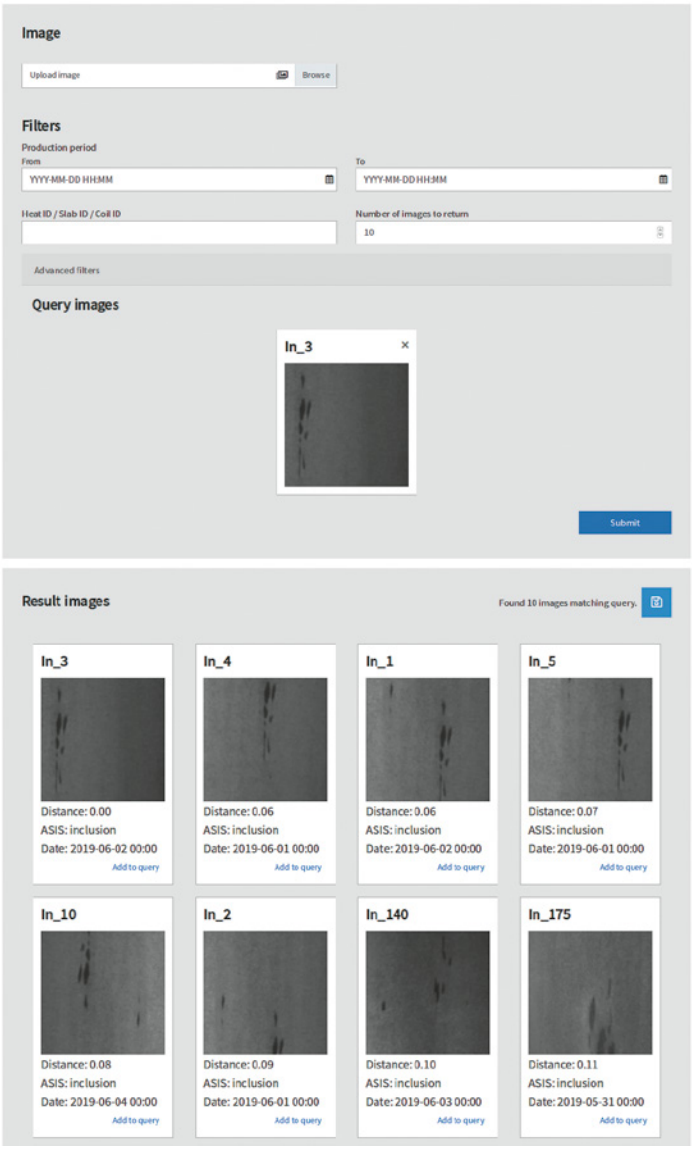

Figure 3. With the deep learning based SST Image Search millions of defect images can be searched in fractions of a second (Picture: SST; Example defects: NEW dataset [10])

severe defects in the relatively large corpus of images.

To illustrate the problem, consider a classifier that correctly classifies shell defects against other defects with a 90% probability. Applied to 1,000 defect images of a coil that contains 50 shells, such a classifier would report only 45 correctly as shells. Applying the 90% probability to the remaining 950 images that do not contain a shell would yield 95 "false positive" shells. In total, such a classifier would assign 140 images to the shell category, of which only about one third would correctly be in the shell defect category. The user now has a reasonable impression that the classifier does not work.

Such a system would detect severe defects for nearly every coil, triggering a mandatory manual follow-up inspection process. The problem cannot simply be solved by changing the sensitivity thresholds for classification to reduce false negatives, as the accuracy (precision) is always increased at the expense of the sensitivity (recall). Insufficient recall, however, means that serious errors may not be detected and defective coils are released as a result.

The defect distribution of the test set must be chosen carefully. Consider that the test set should mirror the distribution of production in order to increase the likelihood that measured performance values are matched in production. However, the accuracy measure of a classifier on such a test set would not reflect the performance of the classifier with respect to rare and severe defects, though it is exactly these rare and severe defects that are often the central focus of classification systems. As a rule, a compromise between the two extremes is recommended: test sets should represent both defects that are frequent in production as well as those that are metallurgically relevant. The test set should also challenge the classifier to the extent that improvements in accuracy can be detected when further training data is added. With the assistance of the SST Training Set Optimizer, such an effective test set can be generated. These test sets can then be used to realistically assess the classifier performance in production.

Classification technology

A well-performing classifier is the most critical element of a reliable surface inspection system. Typical surface inspection systems only make use of conventional classifiers which extract manually-designed features for use during classification [1]. These features are optimized for use on many different materials such as paper, glass, fabric, steel, foil and plastic. If surface defects differ due to a property that is not mapped within this relatively limited feature set, such defects cannot be distinguished by the classifier. Furthermore, the features are often considered as trade secrets and the manual addition of features is simply impossible.

After the publication of the AlexNet architecture in 2012 by Krizhevsky, Sutskever and Hinton [2], which has since been cited by over 40,000 scientific publications, a new era of image processing began. Deep neural networks achieved 85% top-5 classification accuracy in the ImageNet Large Scale Visual Recognition Challenge, over 10% more than traditional image recognition systems. The latest deep convolutional neural networks (CNN) achieve a top-5 accuracy of up to 98% in the same benchmark, a performance that is superior to human test subjects. Due to CNNs' unprecedented success in the ImageNet challenge, it should not come as a surprise that they also provide significantly higher classification accuracy levels in steel surface inspection than the conventional methods described above [4 – 9].

Smart Steel Technologies uses deep learning technology for automated surface inspection and precise defect classification on GPU servers. The company has been developing custom-built CNN architectures for classification of steel defects for several years. These deep learning systems are often more than 10% better than traditional systems on the same training and test data. The software can be integrated directly into production with read-only access to the databases of an existing surface inspection systems without any need for expensive hardware upgrades. The clear advan-

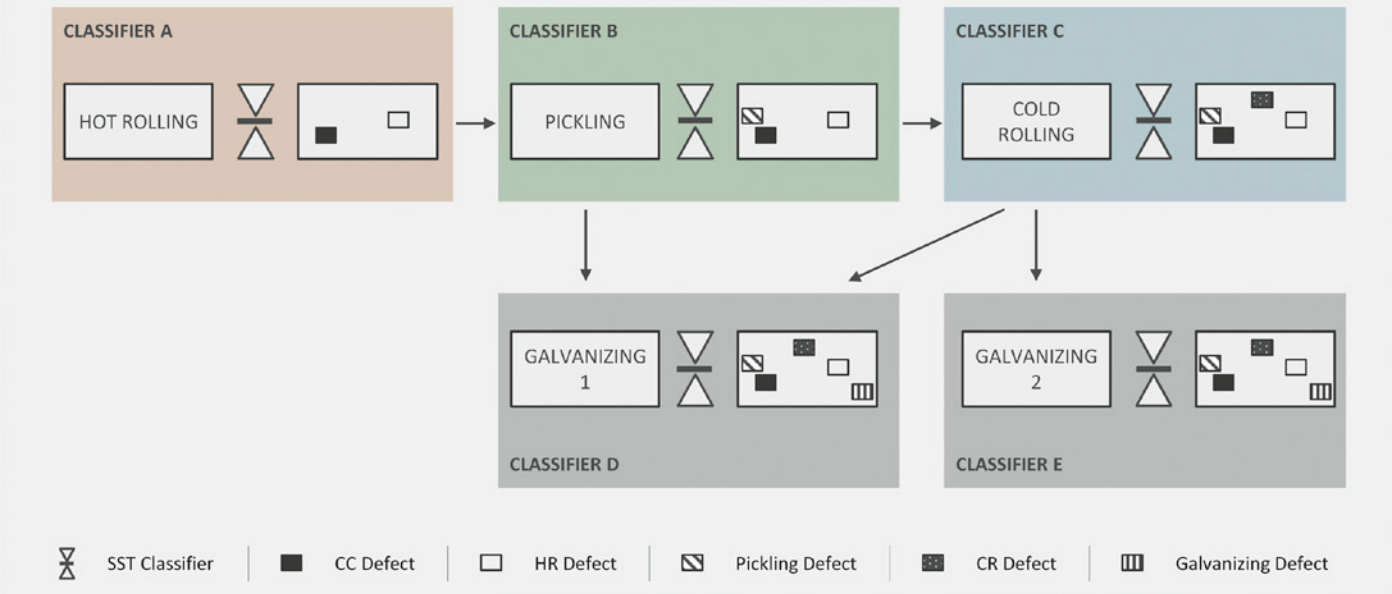

Figure 4. With SST software for material tracking, the surface inspection results of different production lines are brought together in precise positions (Picture: SST)

tage of this approach is the quick and cost-effective improvement of these existing systems. The classifier performance is evaluated and optimized in close cooperation with the customer during production.

Furthermore, unlike many traditional methods, the increase in data volume when using deep learning technology frequently leads to an improvement in the classification result. While users of classical systems sometimes find it frustrating when no improvement of the classification is observed after adding a larger amount of data, deep learning systems can always be improved by the addition of training data. In order to alleviate a deep learning classifier's confusion of two similar defect categories, it is often sufficient to simply increase the number of representative images of those two categories within the training set.

Various production lines

In most plants, multiple independent surface inspection systems can be found along each production line (**figure 4**), and in some cases, provisioned by multiple suppliers. While a few manufacturers have made efforts to compare performance across plants and production lines, many surface inspection systems are evaluated completely independently. In particular, classification systems within a single plant do not communicate with other upstream or downstream classification systems. This lack of communication prevents feedback loops that would improve classification results, for example, one system communicating high-probability shell defect classifications to a downstream production line. Additionally, global coil information, e.g. that a high rate of non-metallic inclusions was observed on a certain coil in a previous processing step, improves classification in later production lines.

Smart Steel Technologies offers flexible material tracking software that allows combining inspection results from different systems even in extreme cases. The defect positions are accurately superimposed taking into account strip position transformations, cropping and trimming shears as well as the cutting and welding of coils (**figure 4**). This enables the detected defects of all relevant surface inspection systems to be viewed together in one central coil map, as well as the SST classifiers to access the inspection results of upstream production lines, resulting in improved classification results.

References

[1] Neogi, N.; Mohanta, D. K.; Dutta, P. K.: Review of vision-based steel surface inspection systems, EURASIP J. Image Vide. 2014:50, 2014
[2] Krizhevsky, A.; Sutskever, I.; Hinton, G.E.: Imagenet classification with deep convolutional neural networks, NeurIPS, 2012, pp. 1097 – 1105
[3] He, K.; Zhang, X.; Ren, S.; Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proc. IEEE Comput. Soc. Conf. 2015, pp. 1026 – 1034
[4] Masci, J.; Meier, U.; Ciresan, D.; Schmidhuber, J.: Fricout, G.: Steel defect classification with max-pooling convolutional neural networks, IJCNN 2012, pp. 1 – 6
[5] Masci, J.; Meier, U.; Fricout, G.; Schmidhuber, J.: Multi-scale pyramidal pooling network for generic steel defect classification, IJCNN 2013, pp. 1 – 8
[6] Yi, L., Li, G.; Jiang, M.: An End-to-End Steel Strip Surface Defects Recognition System Based on Convolutional Neural Networks, Steel Res. Int. 88(2), 1600068, 2017
[7] Zhou, S.; Chen, Y.; Zhang, D.; Xie, J.; Zhou, Y.: Classification of surface defects on steel sheet using convolutional neural networks, Mater. Technol. 51(1), 2017, pp. 123 – 131
[8] Arikan, S.; Varanasi, K.; Stricker, D: Surface Defect Classification in Real-Time Using Convolutional Neural Networks, arXiv:1904.04671, 2019
[9] Kostenetskiy, P.; Alkapov, R.; Vetoshkin, N.; Chulkevich, R.; Napolskikh, I.; Poponin, O.: Real-Time System for Automatic Cold Strip Surface Defect Detection, FME Transactions 47, 2019, pp. 765 – 774
[10] Song, K.; Yan, Y.: A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects, Appl. Surf. Sci. 285, 2013, pp. 858 – 864