Reliable BOF endpoint prediction by novel data-driven modeling

Jochen Schlüter[1], Hans-Jürgen Odenthal[1], Norbert Uebber[1]

Hendrik Blom[2], Tobias Beckers[2], Katharina Morik[2]

[1)] SMS Siemag AG,
Eduard-Schloemann-Straße 4, D-40237 Düsseldorf, Germany
Tel.: +49 (0)211-881-6521
Fax: +49 (0)211-881-4997
E-mail: norbert.uebber@sms-siemag.com
Web page: http://www.sms-siemag.com

[2)] TU Dortmund University,
Joseph-von-Fraunhofer-Straße 23, D-44221 Dortmund, Germany
Tel.: +49 (0)231-755-8126
Fax: +49 (0)231-755-5105
E-mail: hendrik.blom@tu-dortmund.de
Web page: http://www-ai.cs.tu-dortmund.de

Key words: BOF, endpoint prediction, data-driven model, machine learning

In a cooperative effort between SMS Siemag AG (SMS), AG der Dillinger Hüttenwerke (DH), and the Chair for Artificial Intelligence of the TU Dortmund University (TUD), modern data mining and machine learning algorithms have been closely integrated with steel processing by a newly developed modular, distributed and scalable data-stream processing system. It can be easily extended to read and analyze data from all kinds of sensors and data sources. Its analysis can readily be applied to metallurgical and manufacturing processes.
The new approach makes use of a variety of static and dynamic process and measurement variables and uses the precision of data mining and its adaptability to the 190 t converters of DH in order to predict target variables, such as the temperature or the phosphorus content of the melt at the end of blowing.
Rough physical conditions can lead to sensor failures and deterioration. Use is made of statistical methods and multiple models based on different process in order to cope with these technical issues.

INTRODUCTION

Today's steel industry is characterized by overcapacity and increasing competitive pressure. This makes changes necessary. There is a need for continuously improving processes, with a focus on consistent enhancement of efficiency, improvement of quality and thereby better competitiveness.
In a cooperative effort between SMS, DH and the Chair for Artificial Intelligence of the TUD, a real-time data-driven model based on modern high-efficiency algorithms was used to predict the conditions at the BOF end point [1]. The results of the first approach were promising. Nevertheless, many of the steps for creating, applying and evaluating the models involved prototypes and manual procedures.
Advances in plant and process technologies require flexibility. In particular, this flexibility is also required for modeling these processes. Classical metallurgical and thermodynamic models are restricted in their flexibility. Either these models are provided with numerous parameters and are therefore complex to handle or a too small number of parameters limits the reliability of the models. In this paper we argue that machine learning and data mining techniques can provide the flexibility that is required to improve the quality of the BOF-process.

The need for flexibility, reliability and scalability made it necessary to develop and implement a completely new software architecture. This new architecture enables improvements in the quality and reliability of predictions and control of BOF process. The reduction of error-prone manual steps and the extensive use of automation decisively increase the usability of the system.
The improvements in the software architecture were only possible with the in-depth knowledge about the streaming nature of the given data, which was gained in the first part of the project. The importance of the synchronization of multiple data sources and sensors cannot be overestimated. The use of stream mining algorithms to detect concept drifts [2,3] and to schedule the calculation of new prediction models enables the system to adapt to changing operational conditions as reliably and autonomously as possible.
Besides the choice of the machine learning and data mining algorithms, the manner of modeling and the type of variables and features used are also of great importance. This is especially true if the predictions of target variables, such as the temperature or the phosphorus content of the melt, are to be used to predict the end point of the BOF-process and to control the addition of heating or slagging agents. The modeling enhances the understanding of the process by also exposing relationships between input and target variables.
At the end of the day, the most important feature of the system is the supporting of the operator. The operator can only rely on the system if the information is represented as intuitively and accessibly as possible [4,5]. The use of machine learning and data mining techniques in combination with a modular, distributed and scalable system offers the best opportunities for improving the quality of the BOF process and for tackling the future requirements of the steel industry.

DATA MINING AND MACHINE LEARNING

The day-to-day operations of all big Internet companies such as Google, Amazon and Facebook rely heavily on scalable and reliable machine learning and data mining techniques [6,7]. They use efficient algorithms to process huge amounts of data for personalized recommendations, advertisement or search. Even though the prediction tasks of these companies are quite comparable to the tasks in the steel industry, the use of machine learning and data mining techniques is not yet very widespread. These techniques still have a large potential to increase the performance of efficient modeling and control of steelwork processes. The principal characteristics and potential of these methods are explained in Figure 1.

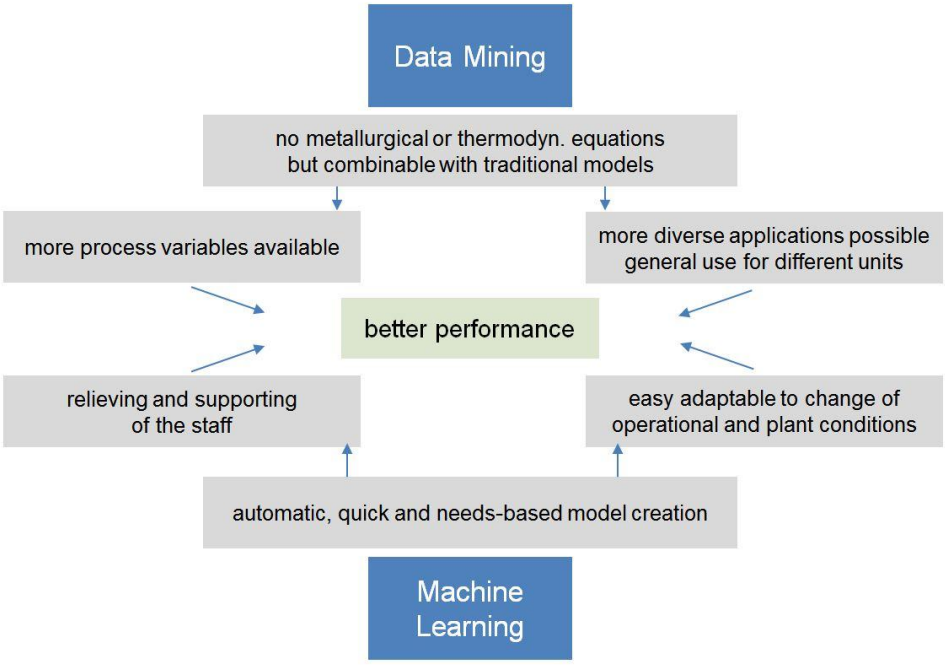

Fig.1. Features and impacts of data mining and machine learning for modeling applications

While the established methods of the process modeling are based on metallurgical and thermodynamic rules for idealized systems, data-driven or statistical models are based on partly hidden relationships (data mining) that are systematically determined by applying certain algorithms to a dataset. This has the advantage that not only process variables are able to be used, which are part of the metallurgical equations, but also data that are not directly related to the target values. They may still contain information about the process and help to improve the prediction.
Some of this additional information, such as converter sound, character of the converter flame, etc., was previously used by the operators to control the BOF process. The demarcation between the control room and the melting units reduces the ability to use a lot of the

information from the metallurgical processes. However, this lost information could be made accessible through the rapid developments in the field of advanced measurement techniques. The use of data-based methods makes it possible to introduce even more, and more diverse, process information.
Furthermore, the data-centric approach is not limited to the common target variables. The same algorithmic approach can be used to predict the temperature at the end of blow or the required amount of deoxidation agents. Due to the data-driven nature of the methodology, the model can be adapted to the current process conditions easily, such as the wear on the converter or the lance level, by relearning the model with the current data. Common models are harder to expand or modify. Usually, the lack of trained staff or time will delay or prevent these necessary adjustments. Innovation processes will be shortened by adopting a semi- or fully-automated model. This flexibility allows widespread use of machine learning and data mining techniques on all the different systems involved in the steel production process.

THE BOF PROCESS FROM A DATA POINT OF VIEW

The data of the BOF process comprise continuous, event-based and static data, which will occur simultaneously. This stream of data can be partitioned in an infinite sequence or stream of BOF processes. Multiple sensors at different sample rates measure the continuous data, such as the off-gas temperature and off-gas analysis. Event-based data, such as hot metal analysis or scrap weights, always have a timestamp and occur in groups of multiple variables. Static data, like the plant conditions, define the context of the BOF process. Only if it is possible to synchronize and aggregate these different data streams, the data can be used in a single analysis. This aggregated data stream will attain the highest sample rate of all the sub-streams and consist of different sets of variables at every point in time [8]. The spatial dimensions of the sensor deployment and the resulting delay in sensor measurements make it necessary to analyze the sensor deployment more carefully.
The operational conditions or the context in which steelwork processes take place are not constant but change over time with modifications of the equipment, wear of components or drifts of sensor accuracy. However, necessary maintenance cycles or converter campaigns may change the context abruptly. The knowledge about the continuous nature of hot metal production and the slow change in the context and the cyclic behavior of the process can be used to adapt the prediction more precisely by selecting data from similar historical processes.
Additionally, the conditions in a steelworks are rough and it cannot be guaranteed that the complete measurement infrastructure will work reliably at all times. There are multiple possibilities for coping with this problem. Either a statistical model of the process variable can assign the missing values or multiple prediction models can be created for every target variable, depending on different sensor data.

DATA MINING AND MACHINE LEARNING FOR THE BOF PROCESS

In the context of machine learning, the prediction of the endpoint is to be seen as a supervised learning problem. From a given training set $T = X \times Y = \{ x_i , y_i \}, i = 1,\ldots,N$ of examples with learning features $x_i \in X$ and label or target variable $y_i \in Y$ a function $f(x) = \hat{y}$ should be learned, which has a minimal error $L(y_j , \hat{y}_j)$ on a given test set Z .
However, due to the given data and circumstances of the BOF-process, the common approach cannot be used directly. To combine static data, such as hot metal analysis, and time-series data, such as off-gas temperature, it is necessary to extract learning features from the time-series data [7]. Learning is further complicated by the different length of every BOF-process and the fact that the true label or target variable is only known at the end of process. As can be seen in our experiments, the choice of the right features is more important than the choice of the learning algorithm used. As long as the algorithm supports non-linear relationships between the learning features and the label, the results will be acceptable. The Support Vector Regression showed good results and has therefore been used in most of our cases [9].
Figure 2 illustrates a situation where half of the BOF process is finished. All of the target values T, [%P], [%C], (%Fe) and the expected oxygen consumption is predicted every second. The optimum end point is derived from the analysis of all the predicted target values. To be able to intervene as early as possible in the process and to perform counter measurements, it is important to have a reliable prediction as early as possible. The prediction rate of 1 Hz enables the system to react within a small time frame.

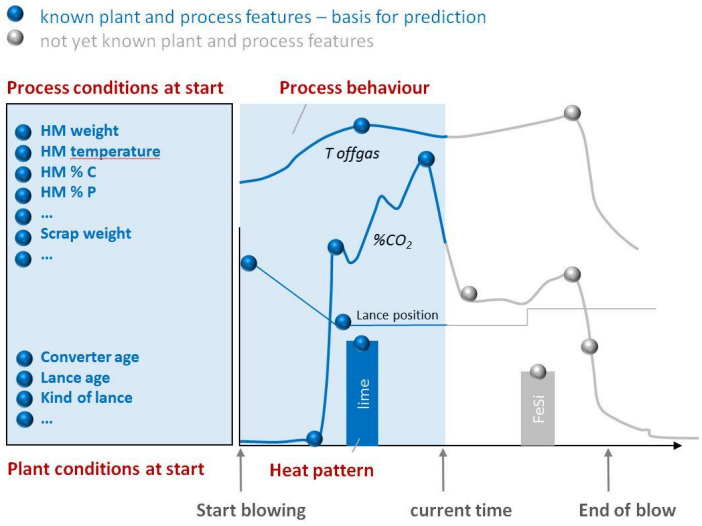

Fig. 2. Illustration of continuous, event-based and static features and their availability for BOF model prediction

The BOF process is characterized by the fact that target variables usually depend on each other. It is, for example, difficult to achieve a low phosphorus content in the melt in combination with high tapping temperatures. From a mathematical point of view, this defines the prediction of the end point as a multi-objective optimization problem [10]. Within the scope of this common project, a multi-objective control tool has been developed that corrects the running BOF process as early as possible, so that several target values are optimized.

SCHEDULING AND MANAGEMENT OF LEARN-PREDICT CYCLES

In the given setting, the construction of prediction models, the so-called Learn Prediction Cycle, consists of three phases (Figure 3). The first phase is the collection of raw data and the extraction of learning features. The set of learning features and the given label or target variables define the prediction model. The second phase is the selection of learning features and data of BOF processes respectively and the learning of the prediction model. The last phase is the online application and evaluation of the prediction model on the actual data stream. The last phase should not be confused with the evaluation on the test set of the model in the learning phase. Depending on the size of the data set used and the algorithms used, the first two phases can take a non-negligible amount of time. Therefore, multiple models have to be learned continuously so as to be available without too much delay.

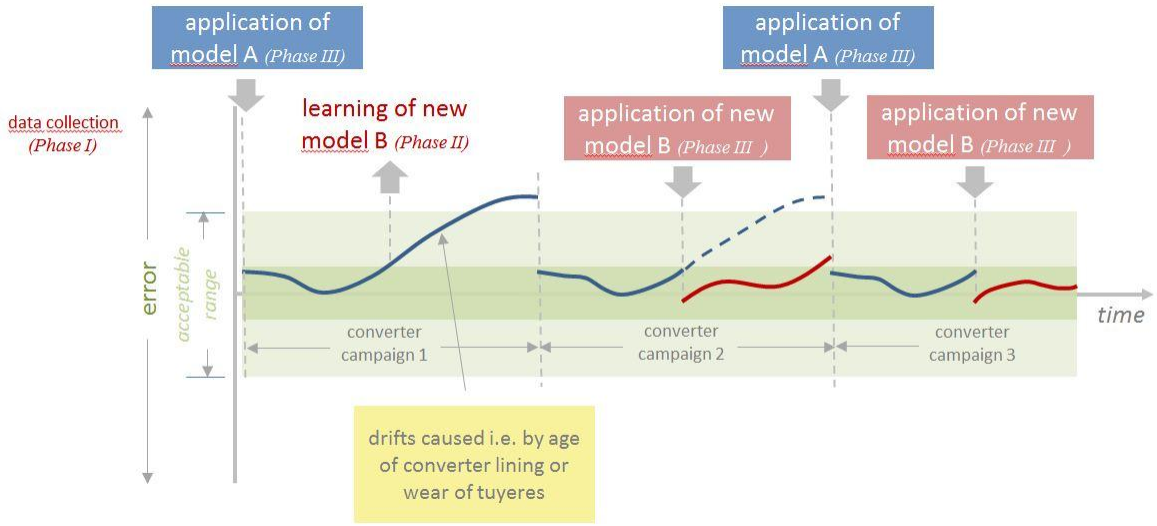

Fig. 3. Illustration of possible progress of model prediction error (in case of cyclic changes)

The quality and the reliability of the input data are monitored continuously and are considered, when a set of learning features is selected. The system has to deal with changing operational conditions and data supply. As mentioned before, the operational conditions of the process will drift, due to the wear of components, over time or even change abruptly (shift) due to necessary maintenance cycles or modifications of equipment.
In the new BOF prediction model, the selection of learning features and the decision to learn a new model is done automatically by the system. There exist multiple algorithms for detecting concept drifts and shifts [2,3]. If the algorithm detects a shift, the error of the actual model will be compared to the error of the alternative models. If a better model is found, the actual model will be replaced. When no better model is available, a new model is learned as fast as possible. Some algorithms have a warning level, which are used to accelerate the model learning process. The combination of concept drift algorithms, automated model learning and automated model replacement in cases of missing values will improve the average performance and reliability of the predictions. If the system is aware of the operational conditions and the cyclic behavior of the BOF process, it will be able to select similar BOF processes and the corresponding learning features. This reduces the average complexity and runtime of the model learning process so that more and better suited models can be learned.

SOFTWARE ARCHITECTURE

The most important architectonic feature of the new software is its modular structure, shown in Figure 4. Every module, for example the application of a prediction model, can be started, stopped and executed independently. This means that multiple models, feature extractions, model evaluations and even graphical user interfaces are executed in parallel not only for one but also for multiple melting units.

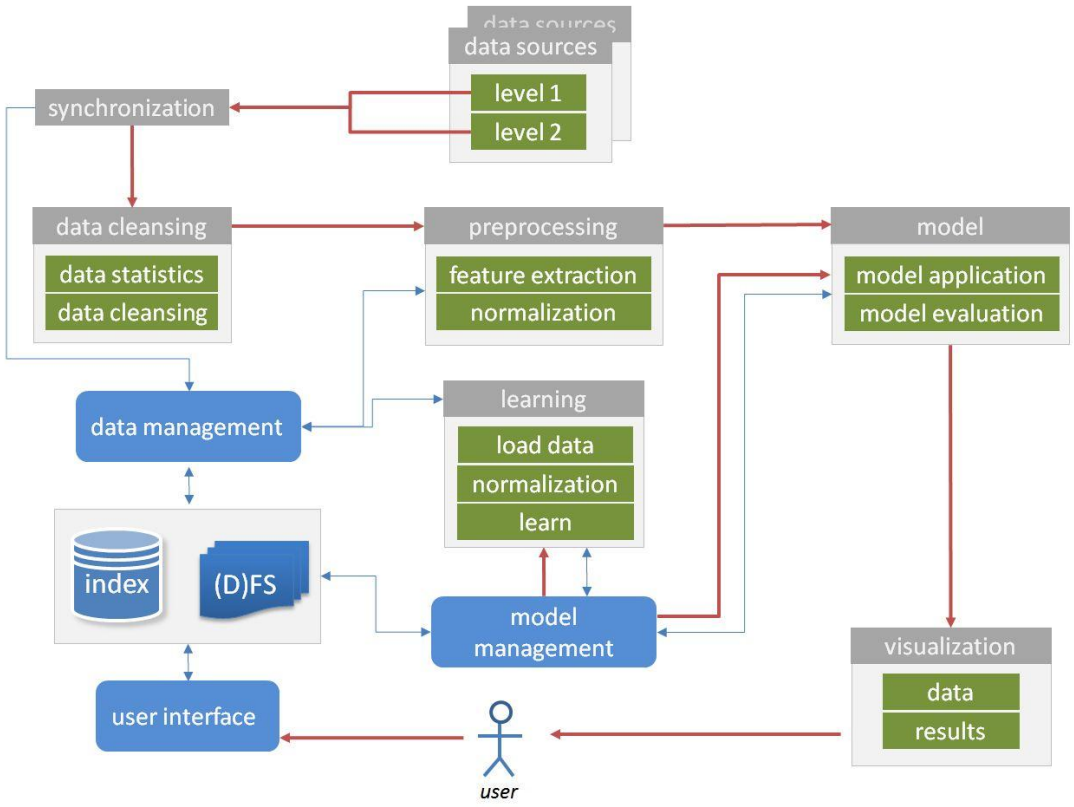

Fig. 4. Overview - New Data Mining software architecture applied to DH BOF

The communication between modules is standardized and relies on network protocols. Therefore, modules can be executed on different computer systems and interact with each other seamlessly. The system is based on the *streams*-framework [8]. The well balanced level of abstractions of data sources make it possible to connect almost every data source to the system without changing any implementation of the succeeding modules. At this present time, the simple key-value files are used for Level 2 data. High-velocity Level 1 data, with a maximum sample rate of 20 kHz, are acquired via the IBA-PDA system. For this application a powerful PC is useful. Fur-

thermore, it is possible to process high-velocity and high-volume data streams, such as video or audio streams, if the I/O hardware will support it. Usually, only the I/O hardware limits the sample rate.
The data management comprises the raw data storage and an index. Until now, the raw data have been stored on the file system but they can be extended to internet scale by using modern distributed No-SQL databases like Googles BigTable [11] or distributed file systems like HDFS [12]. The index contains all the extracted features, so as to make them accessible for automated model creation. In the current version the index is implemented in a simple SQL database. The models, the prediction results and the evaluation results of the models are also stored in the index. This enables the system to start new models autonomously.
Due to decreasing prices and increasing capabilities of data storage, the system is designed to store as much data as possible, but also in the most compressed way possible. This enables the system to access as much historical data as possible. The user of the system is able to emulate a sequence of thousands of BOF processes to tune and optimize the system and to apply new machine learning and data mining algorithms to the historical data under the same conditions as in the real steel plant.

GRAPHICAL USER INTERFACE

A graphical user interface was designed to support the operator in controlling the BOF process by representing the results of the Data Mining Prediction Model in the control room (Figure 5). Although a large number of influencing parameters are used for the model calculations, only the significant variables of the process are shown in the GUI. These include the actual quantities of the input materials used as well as additions, heating and cooling agents. The number of analytical values of the hot metal is limited to essential components.

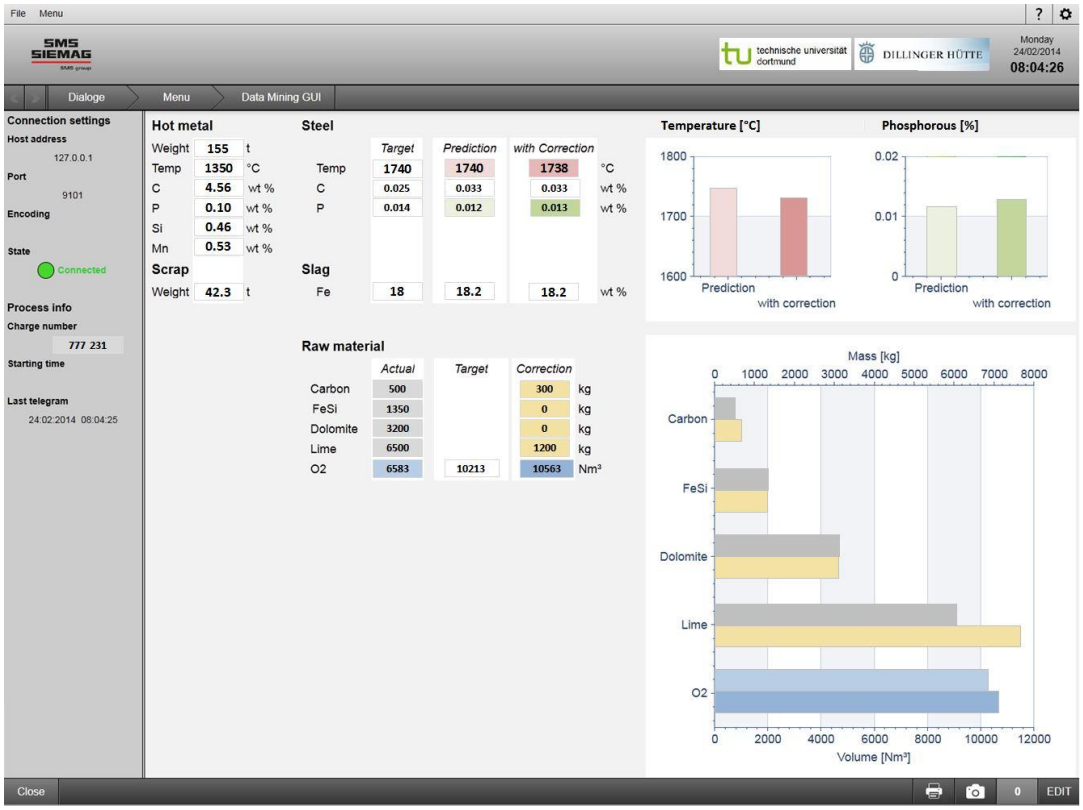

Fig. 5. Graphical user interface for use in the DH BOF control room

In addition to the target values T, [%P], [%C], (%Fe) and the expected oxygen consumption, the current prediction results are displayed and updated every second. The individual prediction values are based on the actual prediction model, which is chosen by the system autonomously.
In order to reach the target temperature, corrective suggestions are made by the Data Mining Prediction Model for heating and cooling materials and additions. They are displayed both numerically and in the form of bar charts. The final states of temperature and phosphorus, predicted on the basis of these corrective suggestions, are listed and illustrated graphically. Thus, the operator is always informed about the impact of the suggested process changes.

If the endpoint of the BOF process is reached and actual values for the steel temperature and steel analysis are available, the diagrams will be replaced by a comparison between the actual and the predicted values. This will help the operator to gain an impression of the current reliability and quality of the model used.

CONCLUSION AND SUMMARY

The application of modern data mining and machine learning techniques to metallurgical processes, such as BOF, is still at the initial stage. Within the last 8 month a completely new software architecture for Data Mining prediction with a flexible and reliable usability was developed and implemented. The commissioning and start of use at the Dillinger Hütte will be at the beginning of March 2014, so that online results are presented in the conference presentation. A Subsystem to store continuous and event based raw data of the two 190 t converters was installed in Oct. 2013, to have authentic data for model learning.
We have shown that the usage of these techniques will lead to better overall performance. The combination of automatic model monitoring, learning and replacement could even increase the average performance and ensure greater acceptance. However, there are a lot of possibilities for monitoring the models and stimulating their updating and application. Hence, it is necessary to gain further experience in this field. For this, it is also useful, that there is the possibility to run the Data Mining prediction model on a conventional PC. After a certain period of gaining experience proving reliability there is in principle the possibility to use the Data Mining prediction model and the corrective suggestions for process control.
The general character of data-based modeling enables a wide range of application without making great efforts to adapt rules and parameters. On the other hand, rule-based modeling is sometimes associated with a certain degree of confidence. There is in principle no need to restrict the modeling to pure utilization of data and to ignore fundamental physical rules. In the future there will be huge potential to combine the advantages of both modeling technologies. We are convinced that now that this development input has been triggered, we can boost the reliability, diversity and efficiency of the automation and, by means of this, enhance the economy of steelworks production.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the AG der Dillinger Hüttenwerke for the excellent cooperation within the scope of this project.

1. N. Uebber; H.J. Odenthal; J. Schlüter; H. Blom; K. Morik: A novel data-driven prediction model for BOF endpoint, JSI Paris 2012 30th Intern. Steel Industry Conf., 18.-19.12.2012, Paris (F), pp.28-29.
2. R. Klinkenberg; T. Joachims: Detecting Concept Drift with Support Vector Machines. In ICML (2000), pp. 487-494.
3. M. Baena-García et al.: Early drift detection method (2006).
4. J. Rasmussen: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. Systems, Man and Cybernetics, IEEE Transactions on 3 (1983): pp. 257-266.
5. R. Parasuraman et al.: A model for types and levels of human interaction with automation. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on30.3 (2000): pp. 286-297.
6. M. A. Russell: Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More, O'Reilly Media, Inc.(2013).
7. G. Linden; B. Smith; J. York: Amazon. com recommendations: Item-to-item collaborative filtering, Internet Computing, IEEE 7.1 (2003): pp. 76-80.
8. C. Bockermann, H. Blom: The streams framework. Technical Report 5, TU Dortmund University, 12 2012.
9. A.J. Smola; B. Schölkopf: A tutorial on support vector regression, Statistics and Computing (2004), pp.199-222.
10. M. Ehrgott: Multicriteria optimization, Springer-Verlag (2005).
11. F. Chang et al.: Bigtable: A distributed storage system for structured data, ACM Transactions on Computer Systems (2008) United Kingdom, 1981, pp 188.
12. K. Shvachko et al.: The hadoop distributed file system, Mass Storage Systems and Technologies (2010), IEEE 26th Symposium, pp. 1-10.