A Reference Architecture for Quality Improvement in Steel Production

David Arnu, Edwin Yaqub
RapidMiner GmbH
Dortmund, Germany
darnu@rapidminer.com,
eyaqub@rapidminer.com

Claudio Mocci, Valentina Colla
TeCIP Institute
Scuola Superiore Sant'Anna
Pisa, Italy
c.mocci@sssup.it, v.colla@sssup.it

Marcus Neuer
VDEH-Betriebsforschungsinstitut GmbH
Düsseldorf, Germany
marcus.neuer@bfi.de

Gabriel Fricout, Xavier Renard
ArcelorMittal Maizières Research SA
Maizières-Les-Metz, France
gabriel.fricout@arcelormittal.com,
xavier.renard@arcerlormittal.com

Christophe Mozzati
PREDICT SAS
Vandoeuvre lès Nancy, France
christophe.mozzati@predict.fr

Patrick Gallinari
Universite Pierre et Marie Curie
Paris, France
patrick.gallinari@lip6.fr

***Abstract*—There is a global increase in demand for steel, but steel manufacturing is a highly sophisticated and costly process where good quality is hard to achieve. Improving the quality remains a major challenge faced by the steel industry. The EU project PRESED (Predictive Sensor Data mining for Product Quality Improvement) addresses this challenge by focusing on widespread recurring problems. The variety and veracity of data, as well as the change in properties of the observed material complicates the interpretation of data. In this paper, we present the reference architecture of PRESED, which is being purpose-built to address the vital concerns of managing and operationalizing the data. The architecture leverages big and smart data concepts with data mining algorithms. Data preprocessing and predictive analytics tasks are supported by means of a malleable data model. The approach allows to design processes and evaluate multiple algorithms pertinent to the problem at hand. The concept is to store and harness the complete production data instead of relying on aggregated values. Early results on data modeling show that fine grained preprocessing of time series data through feature extraction and predictions provide superior insights than traditionally used aggregation statistics.**

Keywords—process optimization; steel manufacturing; data mining; time series; nosql

I. INTRODUCTION

European steel industry is facing extremely challenging global markets with very strong international competition. It is a fact that costs for manpower, raw materials and energy are significantly lower in other regions of the world like China, India or Brazil. It is therefore indispensable for Europe to exploit its current production facilities in a much smarter way to keep its competitive edge. Therefore, novel methods to improve the manufacturing of steel products are essential for market success.

Steel production is a complex process comprising different steps. Each of these steps has to be very precisely mastered in terms of process conditions (temperature, casting speed, cooling flow rate, etc.), as slight deviations can lead to the occurrence of defects on the product. Mostly, physical understanding enables to target the origin of the problem. But in situations where a clear understanding is not available, data mining is a key technology for addressing and understanding the origin of the problem. Due to the complexity of the underlying processes and the data, in particular material tracking over the complete production cycle, sophisticated tools have to be developed.

In this work, we present a conceptual architecture for storing and processing the massive amount of sensor data that are created during steel manufacturing. We aim to build a reference model that is able to take advantage of new data storage methods (e.g., NoSQL) and allows the usage of predictive analytic methods on the stored data.

The layout of this paper is as follows. In Section II, we outline the state of the art for quality management and predictive analytics. An overview of current big data technologies and their relevance to the PRESED project is also presented. In Section III, we give an overview of the manufacturing process in steel plants. Section IV presents the PRESED architecture defined by the PRESED project, its data model, a summary of applicable algorithms and the embedded ontology design. We conclude the paper in Section V and highlight future directions.

This project is funded by the European Commission Research Program of the Research Fund for Coal and Steel Technical Group: TGS9/ Grant number RFSR-CT-2014-00031

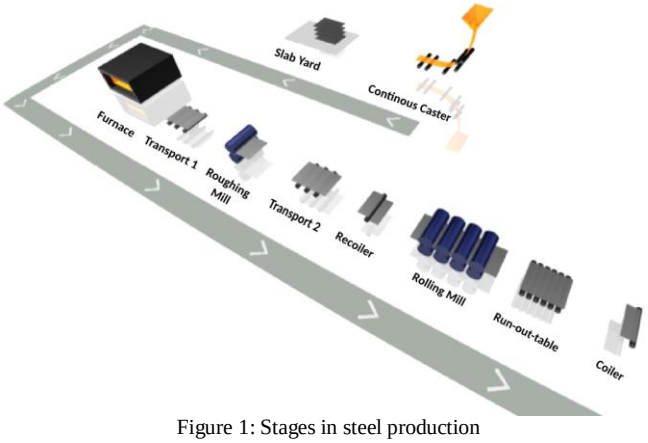

Figure 1: Stages in steel production

II. Related Work

In the past, data mining helped to perform quality surveillance for industrial processes [1] and was used along the production route of steel [2], often with a focus on surface quality [3] or root cause analysis [4]. These tools evaluated several relevant but static quantities to find the root causes of certain production issues. These approaches were limited to a small subset of information about the product, simply because the amount of available data was too large and several aggregations were applied before the actual mining took place. Another work that applied data mining to the steel industry is presented in [5].

The presented architecture approaches data mining on sensorial time series data. This emphasis on uncompressed data differs significantly from traditional methods as it requires a thorough analysis of the time series. Storing the data in a relatively unprocessed state and applying data analytical processes is a common paradigm for big data architectures, e.g., the Lambda Architecture by Nathan Marz [10], which targets key-value paired data.

In contrast to the referred works, the novelty of our work lies in *product-orientation***.** To efficiently store information belonging to a product, each product – in the following called a metal unit - has a virtual representation in a NoSQL database. The metal unit can be a heat, a slab or a coil, depending on the physical state of the product. Adopting such a view is new to steel industry and was not used in previous works. We decided to use a NoSQL database for storing the metal unit, as it does not depend on a fixed uniform data schema, but other storage models are also possible.

Also, a series of dedicated algorithms have been developed for the PRESED project, that significantly extend the common state of the art in the field of predictive data mining. Among them are outlier detection, Self-Organizing Maps and Deep Learning methods that were trained with data from the use cases. Synergies were found between the product-oriented data concept and the machine learning algorithms, as the new data concepts allowed an easy and flexible storage of multiple labels per product.

III. Domain of Steel Manufacturing

Manufacturing steel is a time-sensitive and multi-step process as illustrated in Fig. 1. A typical process flow to produce steel for flat products (sheet) involves the following steps:

- **Iron making**, where liquid iron is produced either from iron ore (blast furnace) or from scraps (electric arc furnace)
- **Steel making**, where the chemistry of the liquid metal is progressively adjusted through several reactors in order to reach the expected target.
- **Continuous casting**, where the liquid steel solidifies into one slab – a piece of solid steel.
- The slab is then re-heated and **hot-rolled** to have a first length adjustment and produce what is called a "coil". For flat products, the length of a slab is typically 20 meters, whereas the length of a coil can reach several hundreds of meters.
- Another step is **cold-rolling** for further elongation of the coil. Depending on the target thickness, the final length can reach several kilometers.
- **Annealing and galvanizing**: a thermal cycle is applied on the steel to adjust its mechanical properties before a zinc coating is applied against corrosion.

A. Use cases

For developing the PRESED architecture and validating our approach, the following use cases are considered:

(1) Predict Sliver Occurrence: Slivers are one of the major sources of quality loss on the steel surface. The defect occurs at the continuous casting step, when slight amount of slag (various liquid oxides remaining on the top of the continuous casting machine) are entrapped at the liquid steel surface during the solidification process. The occurrence of the defect is strongly impacted by the

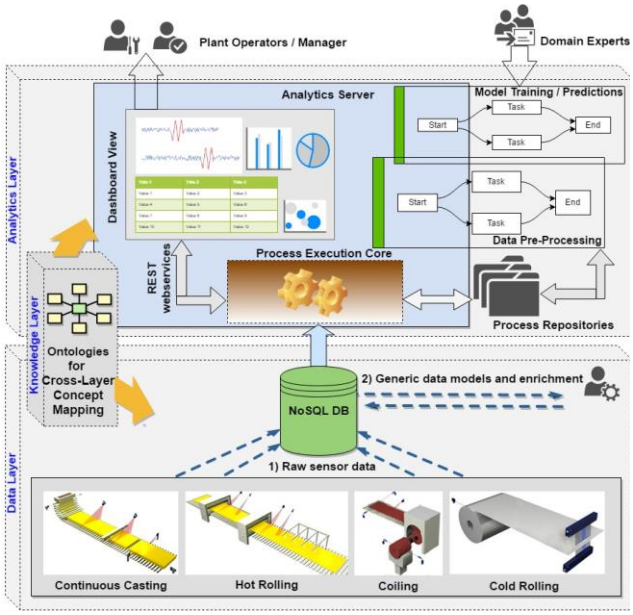

Figure 2: Architecture diagram depicting key concepts

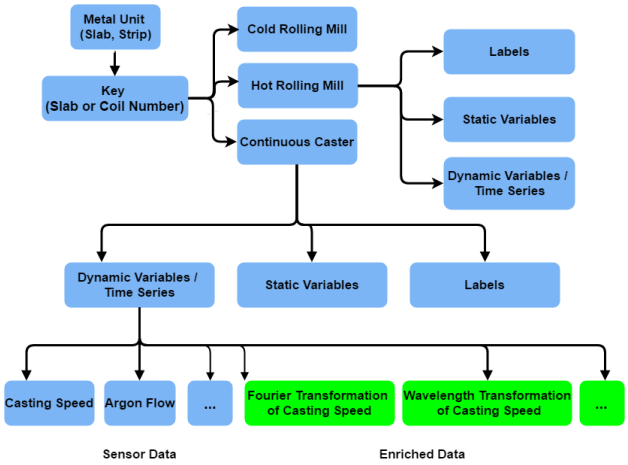

Figure 3: The generic data model for the metal unit.

liquid steel flow in the continuous casting machine, which is itself dependent in a very complex way from all process conditions (temperature, speed, continuous casting actuators, etc.). This is a major concern, because the defect is usually detected at the final step of the process, leading to very high re-allocation costs.

(2) Scattering of Mechanical Properties (MP): MP are one of the most important product criteria for the customer. If the specifications are not reached, a lot of issues can be experienced by the steel customer, in particular for stamping operations. For certain steel grades, these mechanical properties are very sensitive to slight variations of temperature during the last annealing cycle. The thermal behavior of the steel in the last furnace is itself highly dependent on process parameters ranging from continuous casting to the galvanizing lines. A bad mastering of these parameters leads to scattering in mechanical properties, but the study of the phenomenon is very complex since data from all along the production chain has to be considered.

(3) Production Process Chain: This use case aims at controlling the (non-Sliver related) surface defects and improving the inner quality of steel product by considering all phases of the production chain between electric furnace and continuous casting.

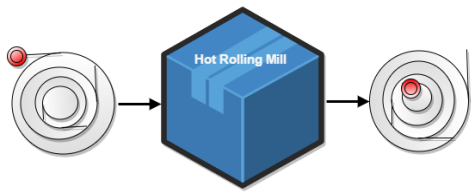

Figure 4: Effect of coiling and de-coiling on length and observed position points

IV. ARCHITECTURE

The proposed architecture addresses various concerns for data storage, data mining, visualization and concept formalism. It is quite evident from previous sections that a singular software framework cannot address the diverse requirements for our use cases. Instead, we emphasize on an extensible approach which is capable of integrating the following technical requirements:

- A generic malleable data model for preprocessing and enriching the raw sensor data.
- Application of various algorithms to the problem at hand e.g., data transformation, feature extraction, outlier detection or classification.
- Designing and executing data mining processes that encompass the above concerns in a visual and concomitant manner to support reuse and sharing.

The PRESED architecture serves the needs of three major stakeholders: **1)** The *Data Engineers*: They model the raw data into a generic model (Section A-1). This is a preparatory activity which may be done infrequently. **2)** The *Domain Experts*: They author data mining processes on the now unified data model. **3)** The *Plant Operators* and *Managers*: These actors execute the previously created processes to seek advanced insights on product quality and potential defects. The results are presented visually to assist human interpretation and enable timely corrective measures.

At the high level, the design splits these concerns into three planes: a *data layer*, an *analytics layer* and a *knowledge layer* as shown in Fig. 2. The salient features of these layers along with the challenges faced are presented below in detail.

A. Data Layer

The data layer deals with the set of sensor readings from various stages of production e.g., continuous casting, hot or cold rolling, coiling and uncoiling of the metal unit. This 'raw' time series data is encapsulated in an object representing the metal unit and stored in a NoSQL database. Although the volume of data collected so far is not as extensive as in other big data projects, it is expected to grow in future. We selected MongoDB because of the simplicity it offers to store unstructured and sparse data as objects that can be efficiently retrieved. It further allows the user to update an object or entire collection of objects with new attributes as or when the need for data enrichment arises. This is well suited to industrial settings where sensor values become dynamically available and where readings corresponding to a certain stage of the production process need to be inspected. In so, the *variety* and *veracity* of the big data paradigm play a strong role in this work.

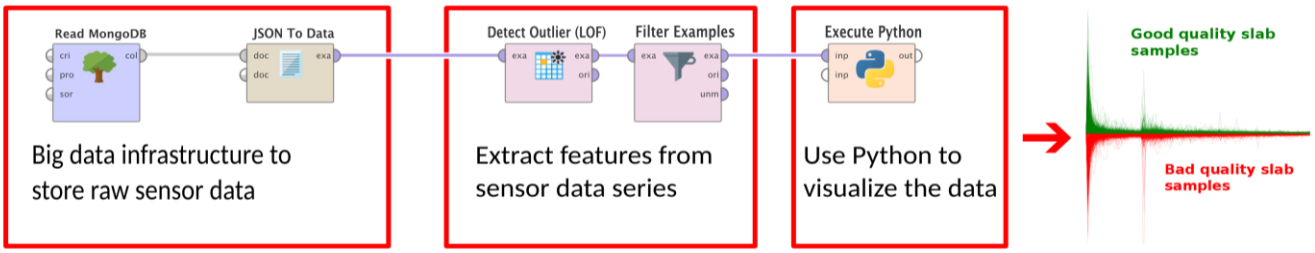

Figure 5: RapidMiner process reading data from MongoDB and performing pre-processing steps before applying a Python script for advanced visualization (left). Fourier transformation of the mould level. Overlay of 1000 slabs, divided into good (green) and bad (red) label classes (right)

1) Generic Data Model

The structure of the metal unit (a coil or slab of steel or a heat) is composed as a generic data model as shown in Fig. 3. The unit is identified by its identifier and has aggregated nodes that hold data from its production process (casting, hot or cold rolling, etc.) or data derived through enrichment (depicted as green nodes). In general, data for each of these categories exist either as static variables which hold fixed univariate values or as dynamic variables which hold time series data. Additionally, information about the unit such as inspection data, cuts, parent/child units can also be linked through identifiers.

The datasets being used are real data from steel plants in France and Italy. For the first two use cases, there exists data for 2261 coils, of which 1205 were labeled (or classified) as good and 1056 of bad quality. The data also includes 5000 slabs, with casting speed, actuator position, mould level, argon gas flow and pressure values - collected during the continuous casting phase. Further, a set of 70 static and 125 dynamic variables are considered for data preparation. The data for the third use case treats a heat as a metal unit. It encompasses the process as steel goes from the electric furnace to the continuous casting phase. A heat in this context refers to the molten metal. The data include 10,000 heats collected over a 2-year time period.

2) Achieving Smart Data through Enrichment

Before applying a learning algorithm, the data needs to be pre-processed. Concerning time series data, the volume of data is not the only key factor. To deliver qualitative reliable information, additional enrichment techniques are required. The first steps are sanity checks such as the usage of consistent units, treating missing values and normalization of data ranges. The latter also serves to anonymize the possibly sensitive data when sharing them among project partners. The metal unit may change in length due to the milling or cutting of the product. Further, as a result of coiling and de-coiling, the head of the coil transforms to the tail point and vice-versa as shown in Fig. 4.

For situations like these, where the geometry inside the metal unit changes, or wherever different sampling intervals are used for sensorial data acquisition at different stages, rescaling is done by interpolating all data with inferior sampling points onto the discretization of highest resolution. This preserves the information for the higher resolved data, which would be lost if down-sampling were to be performed.

3) Systematic Process Design

The next step is to systematically compose these repetitive operations as an executable data mining process. This activity is supported through the open-source and free-to-use RapidMiner [8] tool that allows to graphically create data mining processes. Processes resemble a pipeline composed of tasks that are wired together through drag and drop mechanism. Tasks may perform ETL (Extract, Transform and Load) or machine learning operations by invoking different algorithms and evaluating their performance.

For instance, Fig. 5 illustrates how data enrichment is applied through transformation functions on the raw sensor values. First, the process reads data from the MongoDB instance and converts the JSON format into an in-memory representation. Next, feature extraction is performed by first applying an outlier detection algorithm and then filtering the data set based on the outlier score. Finally, a Python script is invoked to generate a visualization which shows the Fourier transformation of the impurity (mould) level, where 1000 slabs are shown - classified as having good or bad quality.

Such transformation processes help to detect or highlight certain properties of signals e.g., calculating the derivation or applying a signal space transformation. Because of the malleable data schema, it is possible to pre-calculate the transformation and append the results to an existing metal unit. Thus, even the results of computationally intensive transformations are available for reuse. Complex processes for predictive analytics are designed and applied in the same way.

B. Analytics Layer

The analytics layer caters for the management and execution of data mining processes. This layer centers around the Analytics Server which provides: **1)** A repository for storing RapidMiner processes. This eases collaboration by providing shared access. **2)** An execution core to execute processes upon demand. **3)** A dashboard view that allows to browse the data and display execution results as charts (using Python and/or JavaScript libraries). These visualizations can be customized through query parameters because the Analytics Server exposes processes as REST-full web services. The latter paves the way for interoperability with legacy systems and operationalizing on predictions.

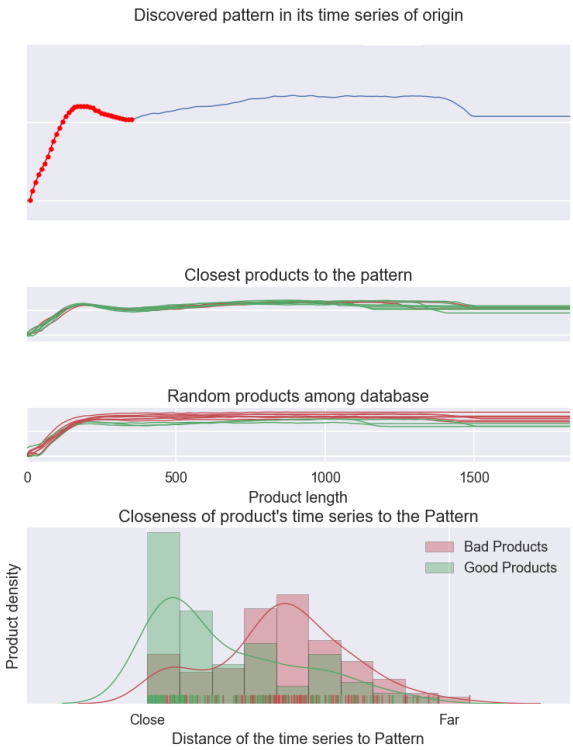

Figure 6: An example of temporal pattern discovery using the shapelet algorithm.

1) Feature Extraction

When dealing with multivariate time series, it is essential to extract the most relevant features. The information contained in the dynamic aspects of sensor data (temporal evolution of the measurements) faces specific issues, which are:

- The relevant temporal information is typically encoded in many complex ways such as segments, spikes, periodicity, drifts or often a combination of these. The relevant information can be the whole signal or only a sub-sequence.
- In an industrial context, the time series are usually multivariate: dozens of sensors measure process parameters at the same time or at the same positions of the product.
- Time series are particularly prone to noise: the typical measurement noise and process variability are duplicated by the successive observations. Time series also suffer from the time-axis noise named time warping: the same phenomenon can occur at several speeds and also suffers from local misalignment.
- When comparing time series, the high dimensionality is a prominent issue as the number of measurements is usually very large – this is often referred in mathematics as the "curse of dimensionality" [9].

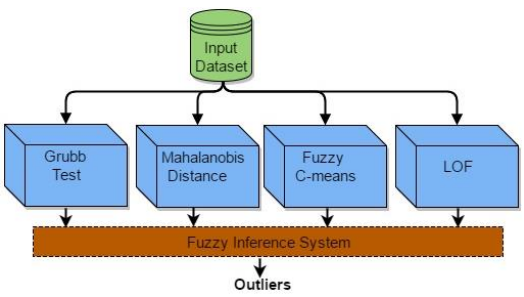

Figure 7: Outlier detection algorithm

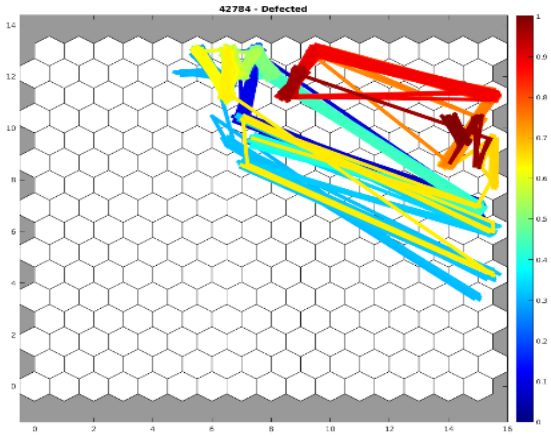

Figure 8: SOM hits plot with trajectory

Those problems are not restricted to the use cases discussed here and there are existing solutions, e.g., low-pass filters, dynamic time warping, dimensionality reduction algorithms. These solutions are incorporated into the data enrichment process.

In interaction with process experts, a supervised temporal pattern discovery approach based on the shapelet concept [12] was developed (Fig. 6). The objective was to discover localized discriminant sub-sequences in the time series. The search is driven by the product quality information. In [11], the scalability issue of the discovery process was addressed in order to apply the method on large datasets. Furthermore, a generalization of the shapelet concept was developed to efficiently discover diverse shapelets with respect to the whole context.

2) Outlier Detection

Detecting deviant observations can help to identify possible problems in the production during an early stage. The main idea of the proposed outlier detection algorithm ensemble is to combine different approaches (density, clustering, distribution and distance) by means of a fuzzy inference system (Fig. 7). Due to the automatic computation of all the parameters needed to execute the elaboration of the algorithm, no a-priori knowledge is required. The fuzzy inference system is further described in [6].

3) Predictive Models

Another aspect addressed in PRESED regards the predictive modeling for process control. For instance, the third use case applies an unsupervised learning approach to detect anomalous behavior of one or more continuous casting process variables.

This has been achieved by means of a SOM (Self Organizing Map) [7].

The cause of defected heats might be the sudden changes in values of specific process variables. This phenomenon can be visualized by looking at a customized SOM hits plot, by adding trajectories of the activated neurons to the net (Fig. 8).

C. Knowledge Layer

The objective of the Knowledge Layer is to facilitate an exchange between process experts and data mining experts. To do so, concepts relative to steel manufacturing are modeled together with concepts relative to data processing, as seen in Fig. 9. These are formalized in an ontology (using the OWL[1] language) that contains concepts, instances and rules. The list of concepts combines a list of common defects and physical concepts, characterizing a metal product (such as density or crystal structure). Rules (described using the SWRL[2] language) link a defect to its effect to the final product.

To exploit this ontology, a web portal is developed based on the KASEM [13] software. The portal allows to query the knowledge stored inside the ontology. The objective is to enable process experts to find a suitable algorithm for a given problem scenario. Using theoretical knowledge combined with the results from experiences of previous cases, the software can advise the best data processing algorithm(s) for this specific problem. The high-level model in the ontology also allows this application to be generalized in a fleet-wide approach. It is thus possible to access knowledge gained in another plant for similar situations and have more precise suggestions at hand [14]. For example, a query can show the used data enrichment processes or the applied algorithms and their parameters for a new, but similar metallurgical problem.

Figure 9: Modeling of steel and data processing concepts inside an ontology

V. Conclusion and Future Work

Improving the quality of steel production processes has been a long-term goal for the industry. The PRESED architecture addresses these goals by leveraging big and smart data technologies with data processing and mining techniques. The on-going progress on use cases is expected to lead to novel results which may serve as a niche for the steel industry in improving the product quality as well as optimizing the whole production processes. Future extensions and improvements are planned especially regarding operationalizing of results in real plants. We also plan to curate the process web services to form a unified API for broader adoption of PRESED architecture. Finally, a link between KASEM and RapidMiner will be investigated to use the knowledge stored in the ontology for context-driven algorithm instantiations.

References

[1] J. Ordieres-Meré, F. Alba-Elías, A. González-Marcos, M. Castejón-Limas, F. J. Pisón-Ascacíbar, Data mining and simulation processes as useful tools for industrial processes, Proc. 5th WSEAS Int. Conf. Simulation, modelling and optimization, pp249-255, 2005

[2] J. Ordieres-Meré, M. Castejón-Limas, Data Mining Applications in Steel Industry, IGI-Global, 2006

[3] S. Mehran Sharafi, H. Reza Esmaeily, Applying data mining methods to predict defects on steel surface, J. Th. Appl. Inf. Technology, Oct. 2010

[4] H. Peters, A. Ebel, J. Hackmann, M. Pander, Industrial data mining in steel industry, StahlEisen, Vo. 132 (2), December 2012

[5] J. Deuse, B. Konrad, D. Lieber, K. Morik, M. Stolpe, Challenges for Data Mining on Sensor Data of Interlinked Processes, SFB 876, 2011

[6] Cateni Silvia, Valentina Colla, and Gianluca Nastasi. "A multivariate fuzzy system applied for outliers detection." *Journal of Intelligent & Fuzzy Systems* 24.4 (2013): 889-903.

[7] Gianluca Nastasi, Claudio Mocci, Valentina Colla, Frenk Van Den Berg, Willem Beugeling. SOM-based analysis to relate non-uniformities in magnetic measurements to Hot Strip Mill process conditions. Proceeding of the 19th World Conference of Non-Destructive Testing (WCNDT) 13-17 June 2016, Munich, Germany

[8] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006, August). Yale: Rapid prototyping for complex data mining tasks. In*Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935-940). ACM.

[9] Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. 2001. NY Springer.

[10] Marz, N., & Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co.

[11] X. Renard, M. Rifqi, G. Fricout, M. Detyniecki : "EAST representation: fast discovery of discriminant temporal patterns from time series", ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data, Riva Del Garda, Italy (2016)

[12] X. Renard, M. Rifqi, W. Erray, M. Detyniecki : "Random-shapelet: an algorithm for fast shapelet discovery", 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), Paris

[13] Monnin, M., Léger, J. B., & Morel, D. KASEM: e-Maintenance SOA platform. In Proceedings of 24th International Congress on Condition Monitoring and Diagnostics Engineering Management, 29th May–1st June, Stavanger, Norway (2011)

[14] G. Medina-Oliva, F. Peysson, A. Voisin, M. Monnin, J-B Leger, Ships and marine diesel engines fleet-wide predictive diagnostic based on ontology, improvement feedback loop and continuous analytics, Proceedings of 26th International Congress of Condition Monitoring and Diagnostic, Engineering Management COMADEM, Helsinki, Finland, 2013

[1] Ontology Web Language

[2] Semantic Web Rule Language