

52nd CIRP Conference on Manufacturing Systems

Characterizing Strip Snap in Cold Rolling Process Using Advanced Data Analytics

Zheyuan Chen^{a*}, Ying Liu^a, Agustin Valera-Medina^b, Fiona Robinson^c

^a Institute of Mechanical and Manufacturing Engineering, School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK

^b School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK

^c Cogent Power Ltd, Orb Electrical Steels, P.O. Box 30, Stephenson Street, Newport, NP19 0RB, UK

* Corresponding author. Tel.: +44 7513646692. E-mail address: chenz57@cardiff.ac.uk

Abstract

Among the undesirable quality incidents in the cold rolling process of strip products, strip snap could result in yield loss and reduced work speed. Therefore, it is necessary to reveal the factors influencing the occurrence of this failure for quality improvement. In this study, a data analytics approach was applied with the aim of determining relevant variables affecting snap occurrence. To validate this approach, a case study was conducted based on real-world data collected from an electrical steel reversing mill. The results suggested a selection of variables to characterize the quality issue of strip snap in the cold rolling process. This quality characterization study was performed as the preliminary stage of a quality improvement task.

© 2019 The Authors. Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>)

Peer-review under responsibility of the scientific committee of the 52nd CIRP Conference on Manufacturing Systems.

Keywords: strip breakage; cold rolling; quality improvement; data analytics; machine learning

1. Introduction

Assumed a plain-strain deformation process, cold rolling compresses and squeezes incoming strip feedstock into thinner strips between the working rolls. With the rapid development of cold rolling processes, a wide variety of sensors are deployed which enable opportunities for quality improvement through data analytics under this data-rich environment. For the cold rolling process of High Silicon strip products, a typical incident of strip snaps frequently occurs. Strip snap, also known as strip breakage or strip tearing, is one of the most common quality issues in the cold rolling process [1]. This incident results in damage on rolls, the steel strip and loss of yield. Therefore, research to identify and determine the causes of strip snap is of great significance in production yield improvement, cost reduction and mill service life extension [2].

The causes of strip snaps have previously been studied in a handful of works [3-5] which focused on retrospective analyses after the occurrence of this incident using conventional metallurgy or mechanical theories. According to these works,

causes of strip snap in cold rolling are various: equipment factors, material defects, improper operation, sensor malfunction and production adjustment. Recently, researches of these strip snap cause analysis have been conducted by employing data analytics [6, 7]. The studies carried out based on a subset of selected variables from hundreds of process measurements to analyse the correlations between these selected variables and strip snap. However, before data analytics were conducted in these works, variables had already been selected based on previous domain knowledge and expertise, thus leaving out of other important causes.

The scope of this paper is to investigate the relation between cold rolling process variables and strip snap using data analytics methods of feature selection and classification. Feature selection, which is to the benefit of understanding data and gaining knowledge from data mining [8], was applied to understand the potential reasons of strip snap. *k*-nearest neighbours (*k*-NN) classification models were built to evaluate the results of feature selection. With the determination of process variables that affect the strip snap quality issue, this

study was performed as a quality characterization work which stands as the preliminary stage of quality improvement [9].

The subsequent sections are structured as follows. In Section 2, a review of both the cold rolling process and strip snap cause analyses using different methods is addressed, followed by a review on cold rolling process characterization. Subsequently, the research problem of this study is stated in Section 3, and the methodology of strip snap characterization in the cold rolling process is proposed in Section 4. Before conclusions, a case study is revealed in Section 5. The case study aims at determining the variables relevant to strip snap. Finally, the findings and further works are concluded in Section 6.

2. Literature review

2.1. Cold rolling process and research on the cause of strip snap

The rolling process has an essential role in the manufacture of a wide variety of products because of its high accuracy, efficiency and production rate. Sheets and strips can be rolled either in the single stand or tandem mills [1].

As one of the main processes in electrical steel strip production, cold rolling enhances strip properties by changing the microstructure and thickness of the steel. These enhanced properties include yield strength, tensile strength, surface smoothness and hardness [6].

Similar to other metal forming processes, the final product of cold rolling can exhibit some mechanical defects. Various defects were observed in industrial metal forming processes, including plane strain rolling. Based on technical reports, common defects in the sheet metal rolling process are edge cracking, central burst, surface defects and buckling of the strip. Among these defects, strip tearing requires special consideration, because it does not only significantly increase production costs but can also cause serious damage to rolls and mill accessories [10].

As one of the most common production incidents in the cold rolling process, plenty of research has been conducted on the causes of strip snap. It has been summarized [4] that strip snaps could result from material defect, equipment malfunction, improper operation and improper rolling parameter settings.

Among these possible causes, equipment factor has been analyzed as the primary one. In a real case study [5], servo-valve malfunction resulted in high-pressure fluctuation, thus leading to inter-frame tension deviations, crushing the strip on one side. Other equipment malfunctions such as the piston rod protrusion of HGC (Hydraulic Gap Control) and tension meter detection accuracy have also been discussed by [11].

Apart from equipment factors, inappropriate operation and parameter settings also account for the occurrence of strip snap. Several operating parameters related to the working roll such as diameter disparity between top and bottom working rolls, levelness of the bottom working roll and convexity degree of the working rolls have been discussed to be significant strip snap causes. Apart from working rolls, levelness and perpendicularity of the deflector rolls have also been proved to generate strip snaps as well [3].

2.2. Characterizing the cold rolling process

Regarding the modelling of the cold rolling process, many different models have been developed and presented over the past decade. These models generally consist of rolling parameters such as tension, roll force, torque and yield strength of the strip as well as several operating parameters. A model developed by Orowan was considered to be one of the most comprehensive among these cold rolling process models. Conventional rolling force formulas, however, provide not more than reasonably accurate approximations [12, 13].

Recognized to be a desirable approach to investigate the design of mill equipment and rolling operation practices, the mathematical modelling of the cold rolling process is conducive to productivity and quality improvement [14].

However, many factors such as friction conditions, roll flattening, deflection of the rolling mill and temperature make the theoretical analysis of the rolling process very complicated and time-consuming. Since the exact values of these variables cannot be measured during the rolling process, there are many parameters needed for better accuracy of the mathematical model [15]. For example, strip snap has been analyzed using a strip deformation model. Nevertheless, the strip deformation models are dependent on parameters which are determined from a combination of approximations to existing rolling theory [16].

3. Problem statement

This paper deals with the occurrence of strip snap during a grain-oriented electrical steel cold rolling process. According to the domain experts, this type of steel is fragile due to a high silicon content which leads to a higher rate of strip snaps during cold reduction to 10% of the starting thickness. The occurrence of this undesired incident can result in yield loss due to the failure to achieve the final target thickness. In addition to this, when the cold rolling production resumes from this incident, an altered rolling performance may occur due to the unexpected disruption caused by strip snaps. This disruption may result in a variation of strip thickness, thus influencing the magnetic loss which is proportional to strip thickness. It is therefore of great significance to identify the causes of strip snap regarding production yield improvement and cost reduction for cold rolling processes.

The objective of this study is to determine the causes of strip snap from multiple measurements in a cold rolling process. This study focuses on the cold rolling process rather than on the incoming material; therefore, variables related to materials are not in the scope of this research. In this context, two historical cold rolling databases that would have a significant influence on this strip snap problem were selected after discussions with the shop floor engineers at the mill of interest. In this paper, data analytics are used to infer possible causes of strip snap incident and cold rolling process variations.

4. Methodology

With the aim of discovering relevant features and discarding those that are irrelevant, data analytics were carried out to

discover potential variables influencing the occurrence of strip snaps. These tasks consist of discovering novel, interesting, and potentially useful patterns from large data sets and applying algorithms to extract information. In the scope of this study, the process data archived in the databases are massive due to the diversity of the cold rolling process measurements. A typical procedure of process data analytics includes the following steps: the collection of both normal and fault data which covers most of the operating regions is firstly carried out, following by data reconciliation. Subsequently, latent variables methods are applied to model the data. Fault detection indices and control limits are generated in the final step [17]. However, these conventional data analytics approaches are not connected to the recent progress in machine learning since the conventional data analytics require a set of carefully collected clean data with irregularly measured and unstructured data mostly unused.

With reference to the procedure of both process data analytics and machine learning [18], the methodology in this study was conducted following this procedure: the first step of this data analytic approach is the collection of data from different data sources that are potentially related with strip snaps and then transforming the data into the same structure and format. Following this, data cleaning is conducted to filter out the noise and to discard idle states of the rolling mill. Subsequently, feature selection techniques are applied in order to find the relevant variables from the training dataset. Feature selection technique is used in conjunction with the classifier. The performance of this classifier is used to determine the preferable feature selection algorithm. Finally, by summarizing selected variables, potential variables affecting the quality issues are selected. Further investigation and knowledge discovery are conducted based on these selected variables. The flowchart of proposed methodology is illustrated in Figure 1.

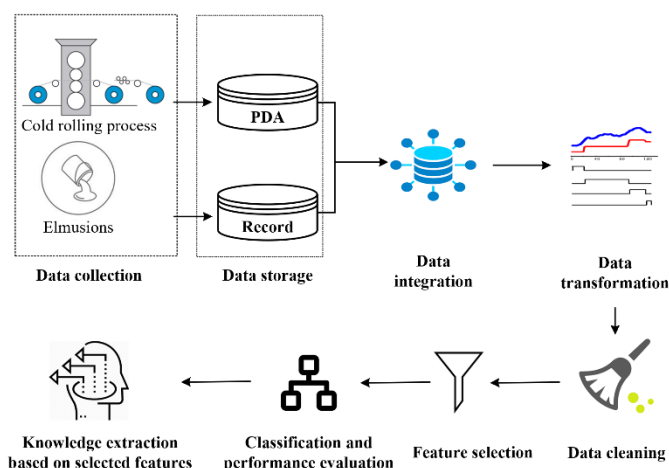


Fig. 1. The methodology flowchart.

5. Case study

5.1. Experiment setup

This case study was conducted based on the historical data provided by an electrical steel mill. The company produces strip coil products which are cold rolled five passes back and forth through a reversing mill. With the silicon exceeding 3%

and the thickness decreasing by up to 90%, these fragile strips often break undesirably during the cold rolling process. The broken strips were marked with a specific pass number by the shop floor engineer. Meanwhile, each strip coil was measured and recorded with detailed data gathering all types of variables.

The following experiment is set up with the aim of determining variables that show the most discrimination degree to differentiate snap and good coils. In this context, with the advantage of gaining knowledge regarding the process that generated the data [8], a set of comparative experiments were conducted based on k -NN models built from different feature selection techniques. six feature selection techniques were used and compared to obtain the preferable feature selection result. The following feature selection techniques and k -NN classification experiments were carried out on Weka (Waikato Environment for Knowledge Analysis) with version 3.8 [19].

Feature selection approaches are classified as wrapper approaches and filter approaches. The filter approaches focus on the evaluation of every single attribute [20]. Meanwhile, the wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected [21]. This approach tends to find features better suited to the predetermined learning algorithm, thus resulting in superior learning performance, but it also tends to be more computationally expensive than the filter model. When there is a high dimensional dataset, the filter model is usually chosen due to its computational efficiency [22]. In this study, considering both the high dimensionality of variables and the objective to select features based on general characteristics of the training dataset, six prevailing filter approaches were applied as listed in Table 1.

Table 1. List of filter-based feature selection techniques applied in this study.

Abbreviation	Name
<i>subset search algorithms</i>	
CFSsub	Correlation-based subset feature selection [23]
<i>feature weighting algorithms</i>	
GR	Gain ratio [24]
IG	Information Gain [24]
SU	Symmetric uncertainty [24]
RF	ReliefF [25]
CFS	Correlation-based feature selection [23]

With the aim of measuring the correlation between variables and class concept, feature selection techniques with correlation-based evaluation criteria were selected. There are two approaches to measure this correlation [26]: one is based on linear correlation including CFS, the other one is based on information theory including GR, IG and SU. As a contrast with correlation-based evaluation criteria, the RF feature selection technique with distance evaluation criteria was also chosen.

Within filter approaches, different feature selection techniques can be categorized into feature weighting algorithms and subset search algorithms [22]. For the subset search algorithms, the number of selected features is automatically generated; while this number is required to be set manually for feature weighting algorithms by setting a

feasibility threshold. There is the possibility that critical attributes may be omitted if the threshold is set too high [27].

In this case, the CFSsub algorithm was first applied to obtain the number of selected features for this subset search algorithm. Subsequently, by referring to this number, four comparative experiments were conducted for each feature weighting algorithms. These experiments were set with a different number of selected features in order to obtain the desired number for this parameter.

With the determination of a desired number of selected features for each feature weighting algorithm, these six feature selection techniques were compared through the classification performance of *k*-Nearest Neighbors (*k*-NN) [28].

In the binary classification problem domain, the confusion matrix (shown in Table 2), also known as an error matrix, allows the visualization of the results of correctly and incorrectly recognized samples [29].

Table 2. Confusion matrix.

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

Conventionally, the accuracy rate is a common metric for binary classification. Nevertheless, regarding the unbalanced dataset in this study, the accuracy rate is no longer an appropriate index since it does not distinguish between the numbers of correctly classified instances of different classes [30]. To evaluate the classification performance of positive and negative classes independently, we can obtain the following two metrics from the confusion matrix [31]:

True positive rate (TPR) is the percentage of positive instances correctly classified:

$$TPR = \frac{TP}{TP+FN} \quad (1)$$

True negative rate (TNR) is the percentage of negative instances correctly classified:

$$TNR = \frac{TN}{FP+TN} \quad (2)$$

However, none of these measures is adequate to evaluate the classification performance since classification intends to achieve good results for both positive and negative classes. In this case, the receiver operating characteristic (the ROC curve) graph is a way to combine these measures to produce an evaluation measurement [32]. The ROC allows the visualization of the trade-off between the FPR (x-axis) and the TPR (y-axis). The area under the ROC curve (AUC) provides a single measure of a classifier's performance for the evaluation regarding which model is better on average.

5.2. Data collection and preprocessing

The case study is conducted by the support of a real reversing cold rolling mill with Production Data Acquisition (PDA) equipment installed on site. PDA is a system which measures more than 1000 cold rolling parameters including

setup values, operation variables and measured process variables. These parameters of cold rolling are monitored and recorded in real-time at the frequency of 100 Hz.

Although the PDA system records more than 1,000 process variables, discarding the non-informative variables delivered, only 454 variables relevant to the reversing mill were selected. With the aim to record the continuous working condition of the mill, the PDA system recorded these variables in continuous manner. There are lots of data being recorded even though the mill is not in operation. Thus, these data were discarded.

The cold rolling process is analyzed under the circumstance that lubricating performance can be tolerated while the high heat capacity of water is required. The emulsion, therefore, acts as both the coolant and lubricant which influence the rolling friction of the cold rolling process [33]. Therefore, apart from PDA data, the emulsion record which keeps track of the daily emulsion information was another source of data.

Since this study focuses on the reversing cold rolling process rather than on the incoming material, the training dataset contains 288 coils with the same incoming material grade, covering three months of production.

As a batch process, the collected dataset from the cold rolling process was a three-dimensional array (coils×variables×time) [34] which is not compatible with conventional feature selection techniques.

For the cold rolling procedure conducted on this reversing mill, incoming coils are passed back and forth for at least five passes depending on the final thickness required. An example demonstrating the different gauge targets set for each pass is listed in Table 3. Apart from the gauge, many other variables such as speed, tension and load are all varying within certain ranges among each pass. In the scope of a reversing rolling process, different passes can be considered as different rolling stages. Process variables are more similar in the same pass than other passes due to the different rolling characteristic.

Table 3. An example of rolling passes with gauge target.

Input gauge	Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Output gauge	
Gauge[35]	2.300	1.450	0.980	0.550	0.360	0.205	0.230

Therefore, considering the similarity of process variables in each pass, the mean values of each variable during different passes are extracted in order to transform the three-dimensional array into two-dimensional by the following procedure,

given a time series of a variable in a single coil:

$$S = \{c_1, c_2, \dots, c_n\} \quad (3)$$

S in this study is represented as:

$$S = \{ \langle c_{is}, c_{ie} \rangle \} \quad (4)$$

i denotes the pass number in the cold rolling process; *is* and *ie* denotes the start and end points of *i* pass respectively;

$\langle c_{is}, c_{ie} \rangle$ is the mean value of data points in the *i*th pass segment.

After data transformation, the three-dimensional array (coils×variables×time) was transformed into a two-dimensional array (passes×variables) which acted as the training dataset in this experiment. A pass in this array acted as an instance, the mean value of each variable within this pass is taken as the value of this instances. According to the running log from the shop floor, every pass is marked with a note indicating whether there was a snap occurred during this snap or not.

The original training dataset (1,297×464) consists of 1,297 instances and 464 attributes (459 variables collected from the PDA system, nine variables collected from emulsion record and the snap label). Among these 1,297 instances, 106 were labelled as “snap”.

It should be noted that within this dataset, there are negative values in certain variables, thus indicating the vector direction related to the reversing mill process. The absolute value of these negative values was taken before a Min-Max normalization was conducted on this dataset.

5.3. Experiments

Firstly, by applying the CFSsub algorithm on the training set, 18 attributes were selected. The results from BestFirst and GreedyStepwise searching methods are quite similar. There were 18 variables selected after applying this algorithm. Concerning this number, four different numbers of selected features (10, 15, 20 and 25) were tested on the five chosen feature weighting algorithms.

Subsequently, starting from correlation-based feature selection (CFS) technique, four subsets of variables were selected by applying this method using different numbers of selected features. The performance of the *k*-NN classification models built on these four subsets was compared in Table 4. The metrics to evaluate the performances were TPR (positive class = snap) and AUC which is a quantitative representation of the ROC curve. TPR was selected to indicate the probability of snap detection while AUC was selected to minimize the negative influence of the classifier performance resulted from skewed classes distributions [36].

Table 4. TPR (positive class=snap) and AUC of the classification model built on different numbers of selected features using CFS.

Number of selected features	10	15	20	25
TPR	0.736	0.802	0.783	0.755
AUC	0.872	0.903	0.897	0.881

The desired number of selected features was selected based on the results shown in Table 4. The best result 15 was obtained by selecting the highest TPR and AUC.

Table 5. The desired number of selected features.

Feature selection techniques	Desired number of selected features
GR	10
IG	10
RF	20
SU	10
CFS	15

Similarly, by comparing the TPR and AUC for other feature weighing algorithms, the desired numbers of selected features are listed in Table 5.

The classification was conducted for comparing these feature selection techniques. For a start, a *k*-NN classification model was built on the training dataset without applying any feature selection techniques. The algorithm yielded the best performance when *k* was set to 1 under 5-fold cross-validation. Afterwards, the six filter-based feature selection techniques (shown in Table 1) were used in conjunction with a *k*-NN classifier with the desired number of selected features. Table 6 shows the classification performance based on selected features for each feature selection technique.

5.4. Results and Discussions

Table 6. TPR (positive class=snap) and AUC of *k*-NN on selected features for each feature selection technique.

Title	Full set	CFSsub	GR	IG	SU	RF	CFS
TPR	0.538	0.717	0.811	0.792	0.820	0.717	0.802
AUC	0.762	0.860	0.890	0.879	0.891	0.854	0.903

As shown in Table 6, correlation-based feature selection, Symmetric uncertainty and Gain ratio generally performed better than other feature selection techniques. Selected features of these three feature selection techniques were summarized in Table 7. Regarding AUC, without applying feature selection, the AUC of the *k*-NN algorithm is 0.762. In contrast, when feature selection techniques were applied, the AUCs of the *k*-NN algorithms were obviously higher than 0.762.

Table 7. Selected features from better-performed feature selection techniques.

Name	Name
1. Bottom Hydraulic pressure feedback	11. Bottom WR diameter
2. SU mass flow Ki	12. DS capsule servo input supply pressure
3. SU setup counter	13. The rate of change of total load reference
4. Left-hand deflector roll diameter	14. Gap control on permits
5. Right-hand deflector roll diameter	15. Run mode
6. Servo back position B	16. Back capsule tilt
7. Mean Trim 1_4	17. OS screw firing angle
8. Top BUR diameter	18. DS screw firing angle
9. Bottom BUR diameter	19. Gemlant OP GEM RX2 Receive last argument error
10. Top WR diameter	20. Gemlant OP GEM RX2 Receive last argument error

CFS and CFSsub share the same evaluation for feature goodness for classification, while the CFS with desired number of selected features performed better than CFSsub regarding both AUC and TPR. This indicated that the automatically generated number of selected features for subset search algorithm could be deficient.

However, the filter approaches of feature selection techniques are incapable to remove the redundant features. Even if many of the selected features are highly correlated to each other, these features are still selected so long as they are

deemed relevant to the class concept [37]. This can be observed from the selected features in Table 7 that “Left-hand deflector roll diameter” and “Right-hand deflector roll diameter” are somehow related since the diameter of the left-hand deflector should match the diameter of the right-hand deflector roll. The filter approach feature selection technique here only selects the relevant variables without removing the redundant variables.

6. Conclusions

The result of the final evaluation test shows that the 20 variables shown in Table 7 are the main critical parameter for snap coils suggesting that there is no uniform cause or structure for snap coils. This coincides with the real situation that strip snaps could be classified into various types according to different manifestation such as pinch, stress, straight line and so on.

Among these 20 variables, there are some variables usually not covered under the conventional analysis while they could be relevant and important. For example, setup counter which records the number of attempts for setting up the reserving mill. Further investigation is needed to understand the influence of such variable.

In addition, the representation of time series process data in this study has not considered the dynamic change of variables. Our further work will focus on a more approximate way to represent this process data with less distortion.

References

- [1] M. Mashayekhi, N. Torabian, M. Poursina, Continuum damage mechanics analysis of strip tearing in a tandem cold rolling process, *Simulation Modelling Practice and Theory*, 19 (2011) 612-625.
- [2] S. Iwadoh, T. Mori, Effect of work roll materials and progress of manufacturing technology on cold rolling and future developments in Japan, *ISIJ international*, 32 (1992) 1131-1140.
- [3] X. Cui, A. Zhao, Analysis the Causes of Deviation Breaking of Cold Rolling Strip, *Xinjiang Steel*, (2013) 44-47.
- [4] J. Xu, Cause Analysis for Strip-breaking of Non-oriented Silicon Steel, *JIANGXI METALLURGY*, (2015) 26-28.
- [5] B. Liu, Study and application of decoupling control of roll tilt and rolling force difference to inhibit strip break during cold rolling, *Metallurgical Industry Automation*, (2015) 40-44.
- [6] K. Takami, J. Mahmoudi, E. Dahlquist, M. Lindenmo, Multivariable data analysis of a cold rolling control system to minimise defects, *The International Journal of Advanced Manufacturing Technology*, 54 (2011) 553-565.
- [7] C. Wang, Research on Fault Diagnosis of Belt Tearing in the Cold Rolling Process, *Instrumentation Technology*, (2014) 16-20.
- [8] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *Journal of machine learning research*, 3 (2003) 1157-1182.
- [9] G. Köksal, İ. Batmaz, M.C. Testik, A review of data mining applications for quality improvement in manufacturing industry, *Expert systems with Applications*, 38 (2011) 13448-13467.
- [10] W. Johnson, A. Mamalis, A survey of some physical defects arising in metal working processes, in: *Proceedings of the Seventeenth International Machine Tool Design and Research Conference*, Springer, 1977, pp. 607-621.
- [11] K. Yan, H. Li, Causes and Countermeasures of Cold Strip Break in Masteel, *Anhui Metallurgy*, (2006) 36-37.
- [12] F. Hou, J. Zhang, J. Cao, X. Shi, Review of chatter studies in cold rolling, *Gangtie Yanjiu Xuebao (Journal of Iron and Steel Research)*, 19 (2007) 6-10.
- [13] E. Orowan, The calculation of roll pressure in hot and cold flat rolling, *Proceedings of the Institution of Mechanical Engineers*, 150 (1943) 140-167.
- [14] J. Son, D. Lee, I. Kim, S. Choi, A study on genetic algorithm to select architecture of a optimal neural network in the hot rolling process, *Journal of Materials Processing Technology*, 153 (2004) 643-648.
- [15] E.J.M. Geddes, I. Postlethwaite, Improvements in product quality in tandem cold rolling using robust multivariable control, *IEEE transactions on control systems technology*, 6 (1998) 257-269.
- [16] H. Furumoto, K. Yamada, J. Yanagimoto, Effect of the number of work-roll surface division on prediction of contact length in coupled analysis of roll and strip deformation during sheet rolling, *ISIJ international*, 42 (2002) 736-743.
- [17] S.J. Qin, Process data analytics in the era of big data, *AIChE Journal*, 60 (2014) 3092-3100.
- [18] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science*, 349 (2015) 255-260.
- [19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter*, 11 (2009) 10-18.
- [20] L. Talavera, An evaluation of filter and wrapper methods for feature selection in categorical clustering, in: *International Symposium on Intelligent Data Analysis*, Springer, 2005, pp. 440-451.
- [21] R. Kohavi, G.H. John, Wrappers for feature subset selection, *Artificial intelligence*, 97 (1997) 273-324.
- [22] P. Langley, Selection of relevant features in machine learning, in: *Proceedings of the AAAI Fall symposium on relevance*, 1994, pp. 245-271.
- [23] M.A. Hall, Correlation-based feature subset selection for machine learning, Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato, (1998).
- [24] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.
- [25] N. Sánchez-Maróño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter methods for feature selection—a comparative study, in: *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, 2007, pp. 178-187.
- [26] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856-863.
- [27] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, R. Wald, Threshold-based feature selection techniques for high-dimensional bioinformatics data, *Network modeling analysis in health informatics and bioinformatics*, 1 (2012) 47-61.
- [28] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is “nearest neighbor” meaningful?, in: *International conference on database theory*, Springer, 1999, pp. 217-235.
- [29] S.V. Stehman, Selecting and interpreting measures of thematic classification accuracy, *Remote sensing of Environment*, 62 (1997) 77-89.
- [30] N.V. Chawla, N. Japkowicz, A. Kotcz, Special issue on learning from imbalanced data sets, *ACM Sigkdd Explorations Newsletter*, 6 (2004) 1-6.
- [31] J.D. Storey, The positive false discovery rate: a Bayesian interpretation and the q-value, *The Annals of Statistics*, 31 (2003) 2013-2035.
- [32] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern recognition*, 30 (1997) 1145-1159.
- [33] S. Wang, A. Szeri, K. Rajagopal, Lubrication with emulsion in cold rolling, *Journal of Tribology*, 115 (1993) 523-531.
- [34] M. Kano, Y. Nakagawa, Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry, *Computers & Chemical Engineering*, 32 (2008) 12-24.
- [35] S.B. Naceur, M. Omran, The effects of bank regulations, competition, and financial reforms on banks' performance, *Emerging markets review*, 12 (2011) 1-20.
- [36] V. García, R.A. Mollineda, J.S. Sánchez, Index of balanced accuracy: A performance measure for skewed class distributions, in: *Iberian Conference on Pattern Recognition and Image Analysis*, Springer, 2009, pp. 441-448.
- [37] K. Kira, L.A. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: *Aaai*, 1992, pp. 129-134.