Steel Quality Monitoring Using Data-Driven Approaches: ArcelorMittal Case Study

Mohamed Laib[1(✉)], Riad Aggoune[1], Rafael Crespo[2], and Pierre Hubsch[2]

[1] ITIS Department, Luxembourg Institute for Science and Technology, Esch-sur-Alzette, Luxembourg
Mohamed.Laib@list.lu

[2] ArcelorMittal Global R&D, Esch-sur-Alzette, Luxembourg

Abstract. Studying manufacturing production process via data-driven approaches needs the collection of all possible parameters that control and influence the quality of the final product. The recorded features usually come from different steps of the manufacturing process. In many cases, recorded data contains a high number of features and is collected from several stages in the production process, which makes the prediction of product quality more difficult. The paper presents a new data-driven approach to deal with such kind of issues. The proposed approach helps not only in predicting the quality, but also in finding to which stage of the production process the quality is most related. The paper proposes a challenging case study from ArcelorMittal steel industry in Luxembourg.

Keywords: Industry 4.0 · Quality monitoring · Predictive modelling · Dimension reduction · Machine learning

1 Introduction

The concept of *Industry 4.0* (also known as the 4th industrial revolution) is gaining great popularity in several sectors. It aims at helping manufacturing industry to be more competitive and efficient by improving their performance. In addition, it gives more flexibility in handling different aspects of business processes [1]. The core asset in going toward Industry 4.0, among others, is to focus on the collection of data, and the good use of the extracted information from this data. Consequently, the digital transformation in manufacturing industry covers many modern research topics in data science such as statistical and machine learning techniques, big data technologies, and the expert knowledge in factories [2–4].

Nowadays, data science research and the advancements in Applied Machine learning and Data Analytics are becoming a very attractive topic in industrial

The present work is partly supported by the Luxembourg National Research Fund, FNR (BRIDGES18/IS/13307925).

O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 63–76, 2022.
https://doi.org/10.1007/978-3-031-10536-4_5

sectors. In recent years, new methodologies and frameworks are being considered in the literature regarding the use of data-driven approaches for decision-making and quality improvement. The range of applied methods is wide, like, for instance, regularized linear regression in predicting product quality, in the semiconductor manufacturing process [5], or extreme learning machine for predicting heat effected zones of the laser cutting process [6]. Other techniques have been used and investigated, Loyer et al. did a comparison of some machine learning methods to estimate manufacturing cost of jet engine components [7]. A detailed literature review about data analytics in manufacturing is proposed by Sivri and Oztaysi [8]. Cheng et al. emphasized some limitations of the available data mining techniques in dealing with complex and unstructured data issued from manufacturing [9]. Iffat et al. suggested some methodologies and approaches on how to deal with high dimensional unstructured data, containing sequence data. They worked mainly on summarizing the sequences using some time series measures and Variational autoencoders [10].

Furthermore, the technology of collecting data is being continuously improved, and many sectors are paying attention to intelligent ways on how to collect data. Consequently, the volume of data is increasing day by day. Moreover, in several manufacturing sectors, data is collected from different stages of production. Thus, assessing quality of the final product using data-driven approaches becomes more challenging because of (1) the existing redundancy in data; and (2) the fact that there are different stages of the manufacturing process. Depending on the aspect of quality, the deviation can be linked to a specific step of production and detecting that is very important in order to improve the quality.

This paper focuses on how to extract information from manufacturing data in order to understand it and predict product quality. The quality is defined with regard to international standards, depending on the product and regions. In fact, the paper investigates one specific product, and proposes a new methodology of dealing with (1) high dimensional manufacturing data collected from different stages; and (2) finding in which stage of the production the quality can be most improved. The present work uses, as case study, ArcelorMittal (AM) steel data collected with the help of experts domain. Several data-driven studies have been conducted on manufacturing data, Konrad et al. proposed a production control concept based on data mining techniques [11]. Lieber et al. proposed a quality prediction in manufacturing processes in two steps. The first step introduces data preprocessing and feature extraction, and the second one combines several supervised and unsupervised machine learning methods to predict product quality [12]. Bai et al. gave a reinforcement learning framework based on AdaBoost algorithm to predict quality [13]. In the present paper, dimension of data is reduced in order to minimize the redundancy, which is a very important step that helps in the understanding of the data [14,15]. Then, a predictive model is applied at each stage of the production in order to monitor the quality through the different stages of production. The aim of the proposed methodology is not only to predict the quality of the product, but also to monitor the product quality through the production stages.

The rest of the paper is organized as follows. Section 2 presents data and the exploratory analysis. Section 3 explains the proposed approach and the used

techniques. Results and discussions are presented in Sect. 4, with validation and interpretation from AM experts. Finally, conclusions and future challenges are given in the last section.

2 Data and Exploratory Analysis

2.1 Data Description

First, the database contains mainly five tables (datasets): Ebauche (EBCH), Blooming (BLM), Tandem (TDM), Thermal Treatment (TT), and Finishing (FIN), which corresponds to the five main steps of the production. These datasets are extracted from the Manufacturing Execution System (MES) from Arcelor-Mittal Luxembourg and correspond to four years of production. The dataset entails the complete transformation process the steel undergoes from scrap to finished product in order to explore as broad correlations as it is currently possible. Individual products are tracked through the whole production process; each entry in the database can be associated with one individual, identified product. The step resolution grows finer, as do the process parameters linked to each step: initially every entry in the database corresponds to a ladle load of 150 tons, then reduced to semi product entries of 10–25 tons. Furthermore, TT treated beams have the highest resolution of information, with a database entry every 3.2 m.

Before the Continuous Caster Machine (CCM), the scrap is molten into recycled steel. The liquid steel's chemical composition is tested once and then undergoes several selective processes including degassing, killing and finally is brought to the desired grade on the ladle furnace. Such process parameters are recorded in the Database. Grades and mixing zones logged for each individual strand of the CCM, leading to a complete histogram of the material until solidified.

For every bar in EBCH, the exact specification of the semi-product including, amongst other the casting date, the heat number, the grade, the strand number are recorded together with the information on the final product including the rolling campaign as well as the product type and dimensions. Although the Reheating Furnace Model was excluded in this analysis, the impact is indirectly inferred as stay time in the furnace, entry state and temperature of the product are taken into account both before and after the furnace.

Concerning the Blooming Stand, it is structured under multiple passes, in which are included roll position and guiding system position. Added to these are a set of guidelines, corrections as well as sensor measurements to ensure process control. The BLM dataset contains detailed information on temperature and time along with roll geometry, some relevant groove parameters and number, guiding system positioning, rolling speed, feeding speed and bar rotation histogram.

Regarding TDM, the universal rolling mode and grooves are taken into account. The process is similar to BLM, insofar as the bar is hot rolled in back and forth passes, with gap adjustments in the roll positions. At each one of these passes, speed, force as well as temperatures on flanges and selective cooling are recorded. In contrast to BLM, universal rolling and edger stands have only one

groove, but the positioning of the rolls and their vertical or horizontal positions are all recorded. Same with calculated strand length and rolling speed.

The bar passes the finishing stand FIN, once. Recorded in that database are, amongst others, rolling speed, roll gaps, roll geometry, forces and a set of temperatures on a number of defined points on the surface of the beam. Not all bars pass through the thermal treatment unit. For the ones that do, speed, flow-rate and ratios as well as temperature of the coolant are recorded alongside with a set of temperature measurements. In TT, bars contain a series of measurements per bar, including a series of cooling parameters specific to the process and cooling models. These include more detailed temperature measurements as well as web and both sides of flanges, in addition to the evolution of temperature lengthwise. Speed variations, and water temperatures, flow rates and rations are also stored to ensure a homogeneous cooling and tailor-made crystallography.

The finished product undergoes multiple checks with regard to its surface quality, its geometry and its mechanical properties. The recorded values are associated to individual pieces and stored in a database. It is through this link that the association between the quality of the final product and the process parameters can be achieved. We aim to improve the predictability of process deviations to ensure the corrections are made before they can have an impact on the product, and instead act proactively to ensure a tight control before process discrepancies occur. This work aims at understanding the origin of deviation using data-driven approaches based on dimension reduction and classification models.

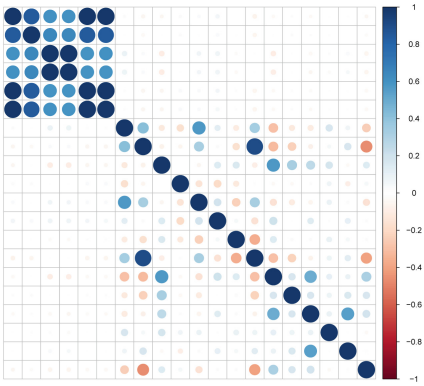
After extracting data from the production steps, a cleaning process has been done to handle missing values and/or outliers. Especially for BLM and TDM where data is structured with multiple passes. These passes have been summarized by taking values corresponding to the finale pass and also some statistics of previous passes (median and standard deviation).

2.2 Exploratory Data Analysis

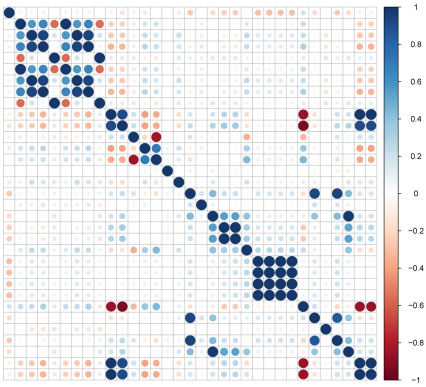
As in any data-driven study, exploratory data analysis (EDA) is a crucial step. In this part, we investigate and visualize the five datasets used in this paper. The panel of available EDA tools is very large, and since the datasets have a high number of features, we focus first on correlations between them. As mentioned above, features are highly correlated. Figure 1 shows the correlations, each dataset has grouped features.

In order to investigate more the used datasets, and since the features are continuous, principal component analysis (PCA) is applied as dimension reduction method to explore the most important part of each dataset. EBCH dataset is used as example to visualize the existing relationship between feature. The correlation matrices are difficult to inspect due to the large number of features. However, one can already see some grouping among the feature of each dataset.

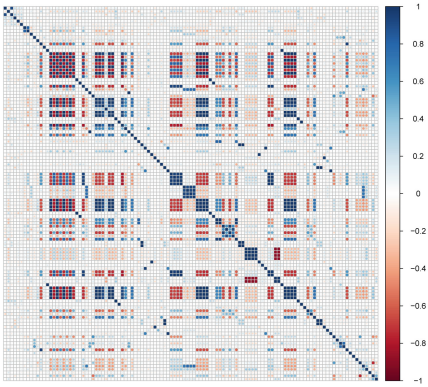
Figure 2 shows results of the PCA, 75% as information is gathered in the first 7 principal components (PCs), where the elbow (4 PCs) shown in the left panel of Fig. 2 is equivalent to 60% of the total variance. The correlation circle (on the 2 first PCs) highlights the redundancy showed in Fig. 1a. Further, it shows mainly



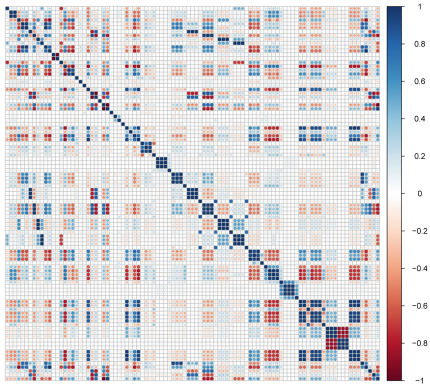
(a) EBCH (20 features)



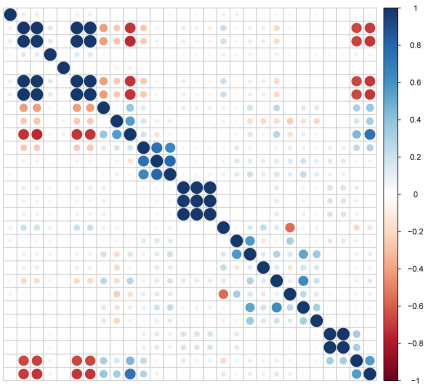
(b) BLM (33 features)



(c) TDM (115 features)



(d) TT (97 features)



(e) FIN (28 features)

Fig. 1. Correlation matrices for the five datasets, which represent the production steps.

3 groups of features, where two of them are correlated negatively. Following the elbow shown in Fig. 2 and considering only the four first components, one can see the contribution of each input variable on Fig. 3. Further investigations using PCA results against the output show the difference between the two classes (with and without deviation), Fig. 4 visualizes the four first components versus the output. In addition, Fig. 5 shows variables that have higher contributions (according to Fig. 3). The difference between classes is hardly clear, which is expected since the production process is only at its beginning.

From EDA results, one can see that all datasets have several correlated features. Merging all datasets cannot help in monitoring the product quality, because correlated features between production steps could lead the engineer to a database that has less effect on the deviation understudy. Therefore, this merges can lead to wrong conclusions about the important features responsible for the deviation, and it would be difficult to monitor the quality through the production procedure.

3 Proposed Approach

The quality of the final product cannot be modelled with the use of only one dataset because the different datasets corresponds to different production steps, each of which contributes to the final properties of the products.

Using only one dataset is not enough in detecting the deviation in the process, mainly because this deviation may not happen in this step of production and one would need more input features from other datasets. However, before using new features from the next steps of the production, we should consider the existing redundancy. The proposed approach is divided into two part (see Fig. 6):

First, an unsupervised feature selection algorithm is applied to reduce the existing redundancy in each dataset. As in many high dimensional datasets, the information can be summarized in a subset of features, and the remainder can be either redundant or irrelevant. In other words, the useful information that explains the phenomenon, described by the data, is in a smaller subset of features [14,15]. In manufacturing data, it is hard to say which features bring more information and which of them are completely irrelevant. In fact, in the ArcelorMittal case study, the dataset contain many features collected for many reasons, not only for monitoring the quality. Some of these feature are highly correlated (see Fig. 1 for example). These high (linear/nonlinear) correlations (i.e. redundancy) may lead, apart to the mentioned problems on the introduction, to difficulty in finding which step of the production has more effect on the final product. Therefore, in the present paper, this redundancy is reduced by using the coverage-based algorithm [16]. This algorithm was used because it does not need to find hyperparameter, which can help in having an automatic procedure to choose less redundant features. Furthermore, this algorithm showed its efficiency in manufacturing data in [10]. The dimensionality reduction performed in this step consists mainly in reducing the redundancy from each dataset of the process procedures.

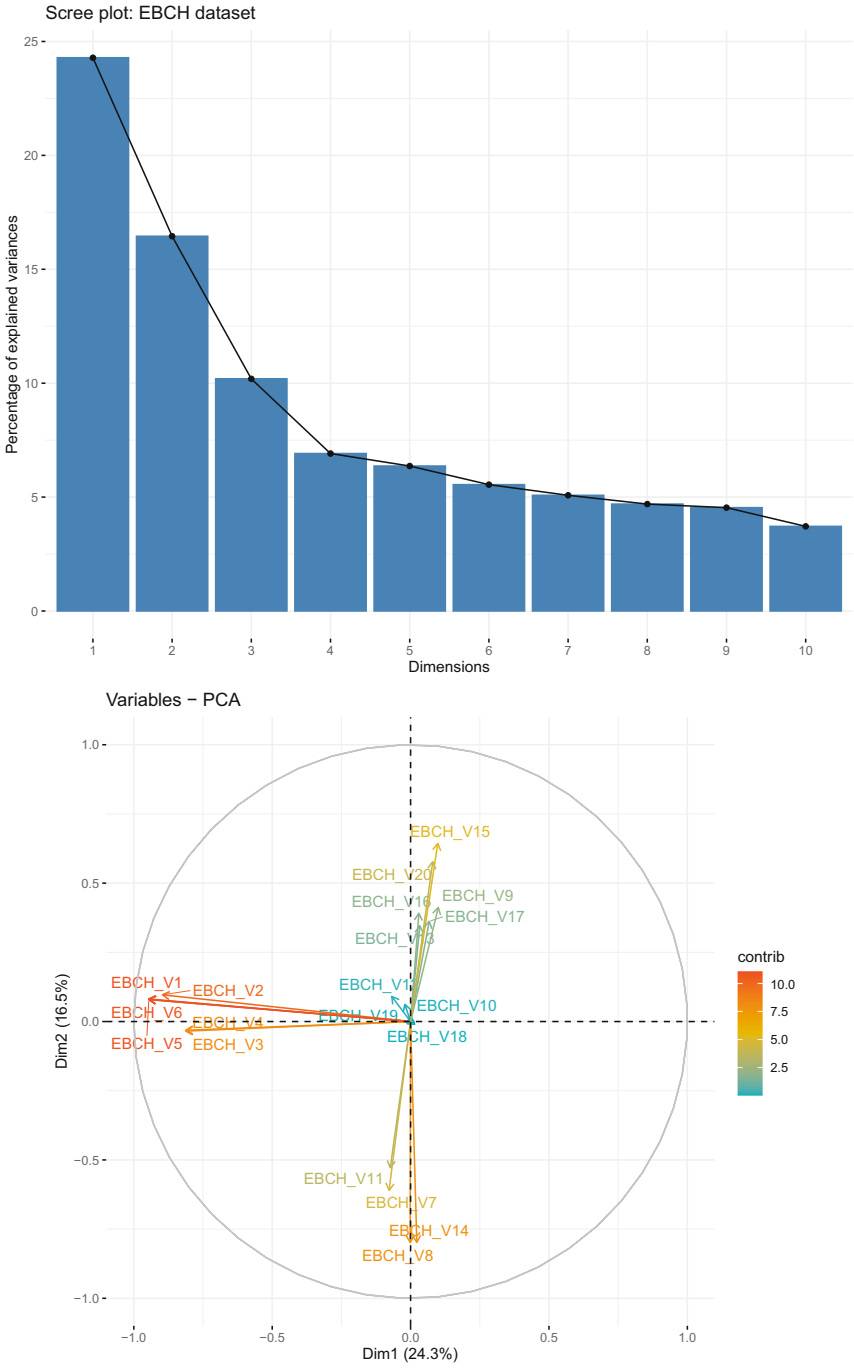


Fig. 2. PCA results applied on EBCH dataset. Top panel: eigenvalues plot representing the percentage of information of each principal component. Bottom panel: correlation circle on the 2 first components.

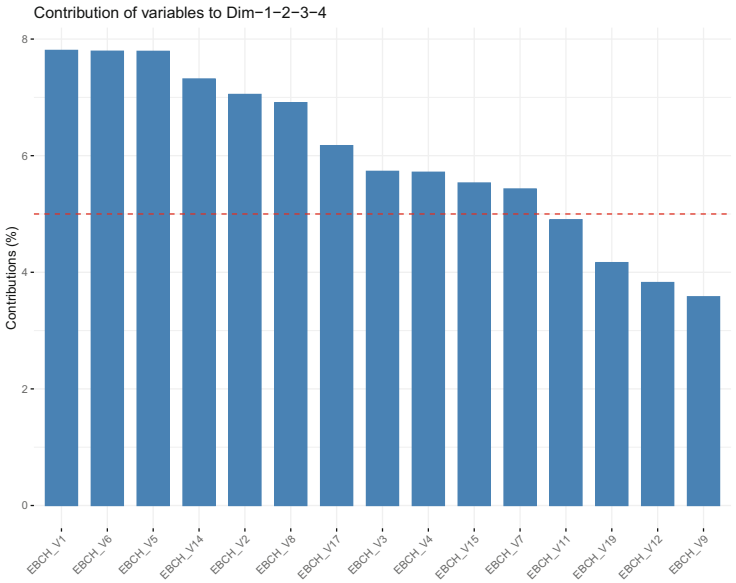


Fig. 3. Contribution of each variable from EBCH to the four first principal components. The dashed red line represents the expected average contribution, if there were no redundancy in data. (Color figure online)

The second step of the proposed approach is the modelling part, which consists in using random forest (RF) on the selected features [17]. However, the modelling is used following the process procedures, as described in Fig. 6. The goal is to model the product quality and identify which feature (and/or at which step of production) can detect deviation from the expected result can be attributed to.

4 Results and Discussions

In order to reduce the redundancy from each dataset, an unsupervised feature selection based on the coverage measure has been applied to each dataset. The results are shown in Fig. 7. As expected, and due to the existing redundancy in the data, the algorithm was able to select a smaller set of features from the used datasets.

In each dataset, the number of selected features is much lower than the original number of features. Table 1 summarizes the results for each stage of the production. Furthermore, one can see already through random forest results, applied on separated dataset, that TDM gives the best performance of classification. Therefore, TDM is responsible for this kind of deviation happened to product. However, to confirm this result, other random forest models have been applied, by adding more feature recorded from other steps of the production process. In fact, we accumulated the datasets following the production process, in order to understand better the data and improve the performance of the classifier. Figure 8 shows the comparison between using only separated dataset and accumulating them.

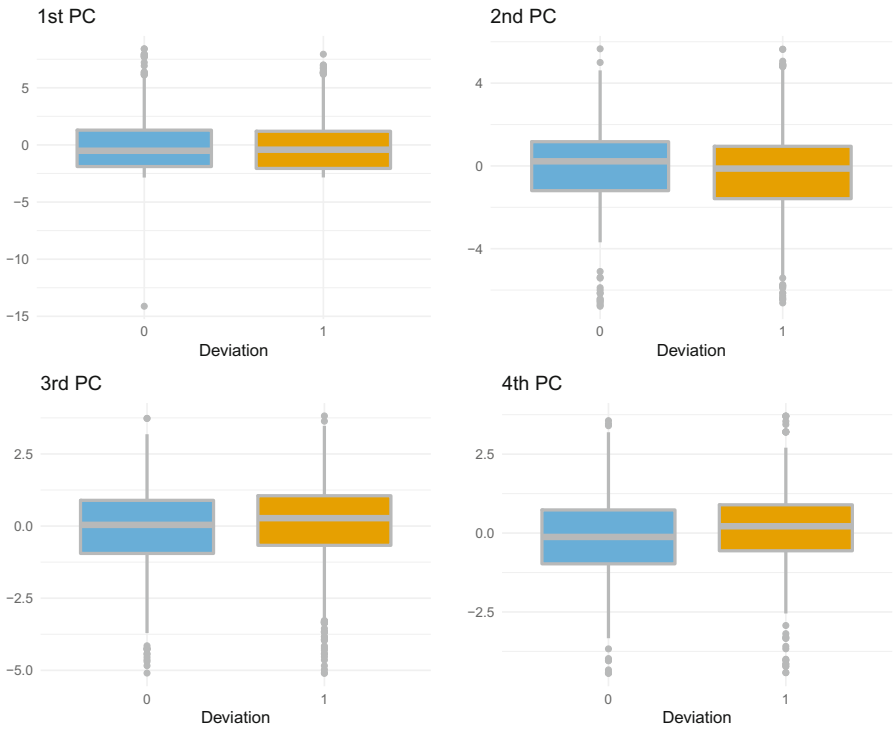


Fig. 4. Four first principal components vs output. PCA components do not a clear distinction between presence and absence of deviation.

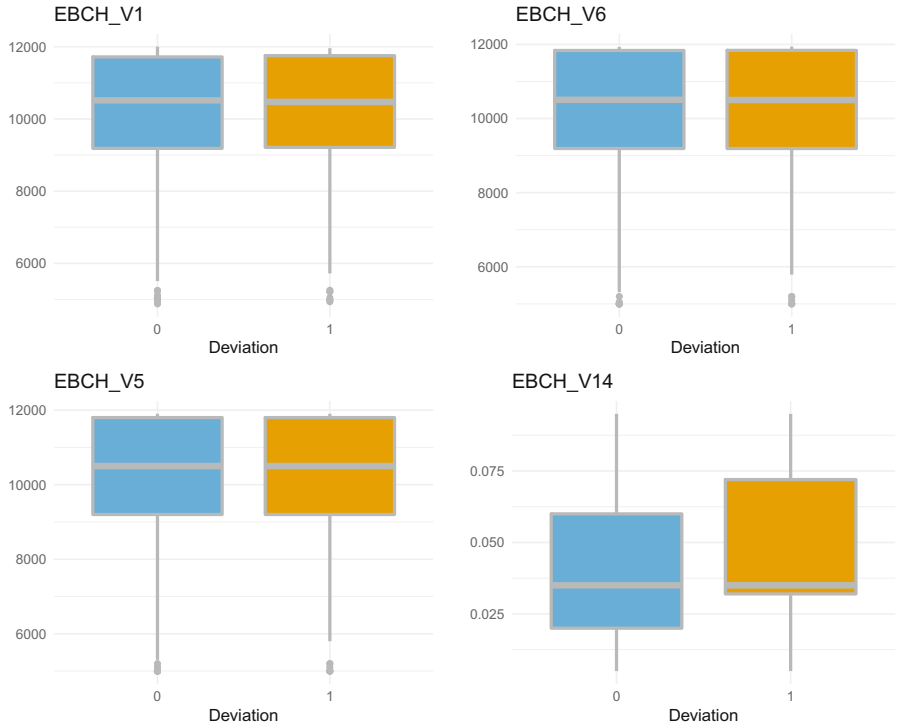


Fig. 5. Variables with higher contribution (according to Fig. 3) vs. output.

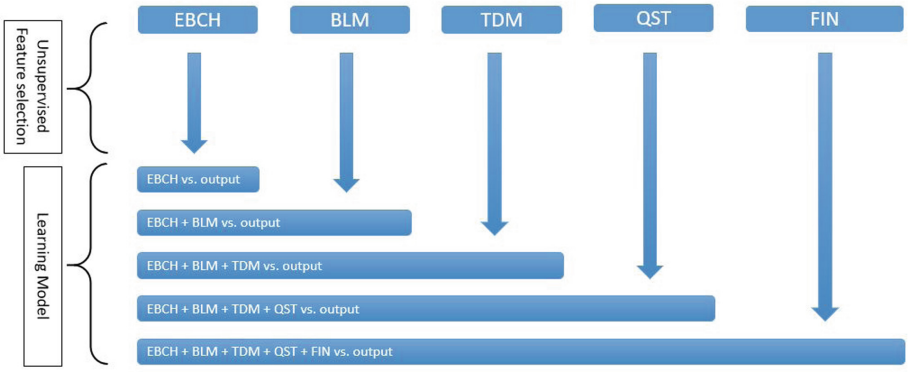


Fig. 6. The proposed approach consisted in reducing the number of features, then apply a learning algorithm at each step of the production by accumulating the features of previous steps.

In Fig. 8 the black line reaches its maximum with accumulating EBCH, BLM, and TDM. After TDM, the performance is slightly lower but stay stable for TT

and FIN (mainly because of the presence of variables from TDM). However, the red line, presenting the results of separated dataset, FIN performance drops too low, which means that this stage has less effect on the manufacturing process and does not influence the finale outcome.

As mentioned above, the quality is defined following international standards that depend on the type of product and also the region to which the product is delivered to. In this case study, we have chosen one particular measure of quality, which is adherence to the prescribed geometry (defined by international standards) and we can read from our results that features recorded in TDM have more effect in describing and predicting the final product quality.

Table 1. Summary of feature selection results and random forest classifier for each dataset.

Datasets	# Original feat.	# Selected feat.	Kappa (Separated)
EBCH	20	4	0.26
BLM	33	12	0.28
TDM	115	15	0.31
TT	97	19	0.31
FIN	28	12	0.3

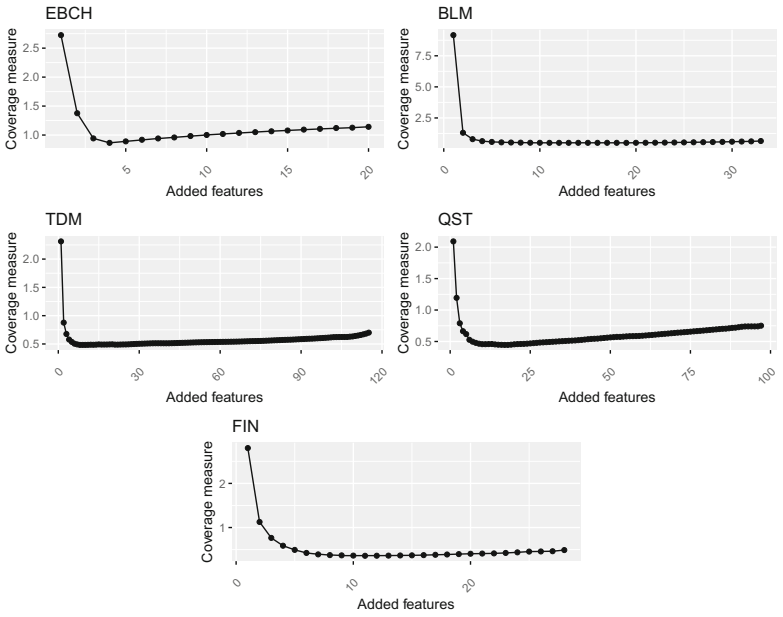


Fig. 7. Unsupervised feature selection applied on each dataset.

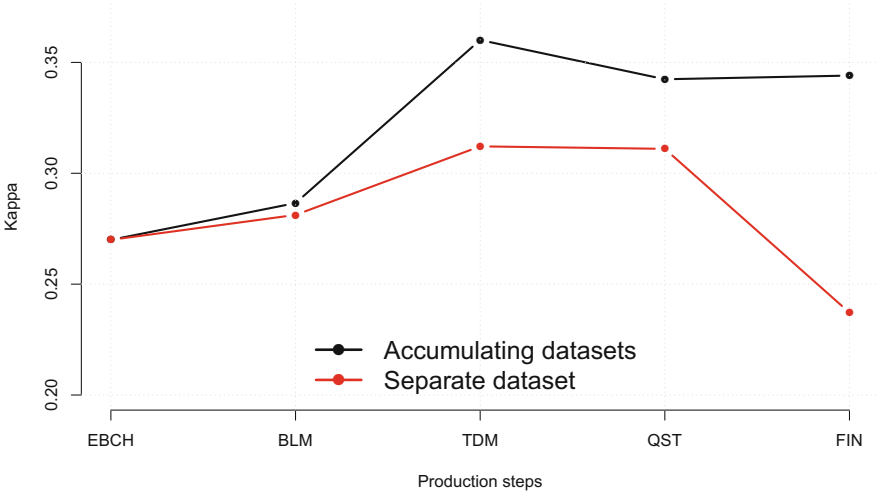


Fig. 8. Random forest results from different datasets. Black line presents the results when the datasets are accumulated following the production process. The red line is the results by using the dataset separately. (Color figure online)

5 Conclusion

This paper presents a data-driven approach for monitoring quality deviation in a manufacturing process. The proposed approach is divided into two main steps:

1. Application of unsupervised feature selection to find a small set of non-redundant features. The purpose of using such algorithm, besides its good performance, is because we know that almost all features recorded for this study are relevant to predict the output but are highly correlated. Therefore, we need only to reduce redundancy from datasets.
2. Using a random forest classifier for both separated and accumulated datasets to find out with which dataset we have the best performance. In addition, the comparison between accumulated and separated datasets allows us to conclude which stage of the production is most sensitive to deviation and must be closely monitored.

Pointing out that TDM has more effects on this deviation was expected by the R&D team from ArcelorMittal. With such results, the deviation will be corrected in the right stage of production and improve the quality of the final product. Furthermore, the results encourage the use of data-driven approaches to explore, understand manufacturing data, and improve the quality.

The present work explored several datasets collected from several stages of the production process. It presents a contribution to the understanding of ArcelorMittal data. Further works could be in using more advanced dimension reduction tools that take into account data collected from different sources, such

as integrated principal component [18], which may help in extracting more knowledge and gives more insights that can improve classification models.

Acknowledgment. The authors thank both teams from ArcelorMittal and LIST, who were involved in discussions about the PAX project.

References

1. Kagermann, H., Wahlster, W., Helbig, J.: Recommendations for implementing the strategic initiative industrie 4.0. In: Final report of the Industrie 4.0 Working Group, Federal Ministry of Education and Research, p. 84 (2013). http://forschungsunion.de/pdf/industrie_4_0_final_report.pdf
2. Khan, M., Wu, X., Xu, X., Dou, W.: Big data challenges and opportunities in the hype of industry 4.0. In: 2017 IEEE International Conference on Communications (ICC), pp. 1–6 (2017). https://doi.org/10.1109/ICC.2017.7996801
3. Lu, Y.: Industry 4.0: a survey on technologies, applications and open research issues. J. Ind. Inf. Integr. **6**, 1–10 (2017). https://doi.org/10.1016/j.jii.2017.04.005
4. Rossit, D.A., Tohmé, F., Frutos, M.: A data-driven scheduling approach to smart manufacturing. J. Ind. Inf. Integr. **15**, 69–79 (2019). https://doi.org/10.1016/j.jii.2019.04.003
5. Melhem, M., Ananou, B., Ouladsine, M., Pinaton, J.: Regression methods for predicting the product quality in the semiconductor manufacturing process. IFAC-PapersOnLine **49**(12), 83–88 (2016). 8th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2016. https://doi.org/10.1016/j.ifacol.2016.07.554
6. Anicic, O., Jović, S., Skrijelj, H., Nedić, B.: Prediction of laser cutting heat affected zone by extreme learning machine. Opt. Lasers Eng. **88**, 1–4 (2017). https://doi.org/10.1016/j.optlaseng.2016.07.005
7. Loyer, J.-L., Henriques, E., Fontul, M., Wiseall, S.: Comparison of machine learning methods applied to the estimation of manufacturing cost of jet engine components. Int. J. Prod. Econ. **178**, 109–119 (2016). https://doi.org/10.1016/j.ijpe.2016.05.006
8. Sivri, M.S., Oztaysi, B.: Data Analytics in Manufacturing, pp. 155–172. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-57870-5_9
9. Cheng, Y., Chen, K., Sun, H., Zhang, Y., Tao, F.: Data and knowledge mining with big data towards smart production. J. Ind. Inf. Integr. **9**, 1–13 (2018). https://doi.org/10.1016/j.jii.2017.08.001
10. Iffat, U., Roseren, E., Laib, M.: Dealing with high dimensional sequence data in manufacturing. Procedia CIRP **104**, 1298–1303 (2021). 54th CIRP CMS 2021 - Towards Digitalized Manufacturing 4.0. https://doi.org/10.1016/j.procir.2021.11.218
11. Konrad, B., Lieber, D., Deuse, J.: Striving for zero defect production: intelligent manufacturing control through data mining in continuous rolling mill processes. In: Windt, K. (ed.) Robust Manufacturing Control. LNPE, pp. 215–229. Springer, Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-30749-2_16
12. Lieber, D., Stolpe, M., Konrad, B., Deuse, J., Morik, K.: Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. Procedia CIRP **7**, 193–198 (2013). forty Sixth CIRP Conference on Manufacturing Systems 2013. https://doi.org/10.1016/j.procir.2013.05.033

13. Bai, Y., Xie, J., Wang, D., Zhang, W., Li, C.: A manufacturing quality prediction model based on AdaBoost-LSTM with rough knowledge. Comput. Ind. Eng. **155**, 107227 (2021)
14. Lee, J.A., Verleysen, M. (eds.): Nonlinear Dimensionality Reduction. Springer, New York (2007). https://doi.org/10.1007/978-0-387-39351-3
15. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
16. Laib, M., Kanevski, M.: A new algorithm for redundancy minimisation in geo-environmental data. Comput. Geosci. **133**, 104328 (2019)
17. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
18. Tang, T.M., Allen, G.I.: Integrated principal components analysis. arXiv:Methodology (2018)