Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 72 (2018) 1333–1338

www.elsevier.com/locate/procedia

51st CIRP Conference on Manufacturing Systems

Analytical Approach to Support Fault Diagnosis and Quality Control in End-Of-Line Testing

Vitali Hirsch[a,b]*, Peter Reimann[a], Oliver Kirn[b], Bernhard Mitschang[a]

[a]*Graduate School advanced Manufacturing Engineering, University of Stuttgart, Nobelstr. 12, 70569 Stuttgart, Germany*
[b]*Daimler AG, Hanns-Martin-Schleyer-Str. 21-57, 68305 Mannheim, Germany*

* Corresponding author. Tel.: +49 621 393 7662. *E-mail address:* vitali.hirsch@gsame.uni-stuttgart.de

---

Abstract

Operators in end-of-line testing of assembly lines often try out multiple solutions until they can solve a product quality issue. This calls for a decision support system based on data analytics that effectively helps operators in fault diagnosis and quality control. However, existing analytical approaches do not consider the specific data characteristics being prevalent in the area of End-of-Line (EoL) testing. We address this issue by proposing an analytical approach that is tailored to EoL testing. We show how to implement this approach in a real-world use case of a large automotive manufacturer, which reveals its potential to reduce unnecessary rework.

Keywords: Analytics; decision support; recommendation system; fault diagnosis; quality control; optimization

---

1. Introduction

Already in 2004, vehicle manufacturers spent more than 1.000€ per vehicle for reworking on or even scrapping erroneous product parts during quality control [1]. Considering the total number of manufactured vehicles, one original equipment manufacturer (OEM) wasted several millions for such error costs. Based on an error cost analysis at an industry partner, we revealed that these costs are even much higher today. This is mainly due to substantial increases in complexity and variety of both products and production processes.

In particular, the final phase of the production process, the End-of-Line (EoL) testing, is very complex and cost-intensive [2]. Due to misdiagnoses, operators usually carry out several expensive repair attempts and re-tests until they can solve a quality issue. This calls for a decision support system based on data analytics that increases the efficiency of operators in fault diagnosis and quality control.

Existing analytical approaches do not take into account the data characteristics and related analytical challenges that are prevalent in the area of EoL testing. Most related approaches are struggling with two major shortcomings, which are considerably limiting practical feasibility. Firstly, they usually assume a holistic data warehouse that provides huge amounts of high-quality data (e.g., [3,4]). However, the amount and quality of data is significantly limited in the EoL domain. This limited data set often shows a high complexity in real-world applications, as the data describes a heterogeneous product space and a wide range of unevenly distributed quality issues. Secondly, related work does not offer any means to evaluate and adapt the results of data analytics considering economical aspects (e.g., [4,5]). However, this is necessary to reduce costs for fault diagnosis and reworking, especially in EoL testing.

To address these issues, we propose an analytical approach as a blueprint for increasing the efficiency in the domain of EoL testing. The major contributions of this paper are:

- We propose an analytical approach that considers the domain-specific data characteristics and challenges of EoL testing. This approach provides a tailored decision support that effectively helps operators in fault diagnosis and quality control. It offers them a small list of faulty parts that are ranked according to their likelihood of being the cause of a quality issue. Furthermore, the faulty parts are evaluated

2212-8271
10.1016/j.procir.2018.03.024

regarding their economical relevance. Altogether, this offers a great potential to reduce both the number and costs of rework steps caused by misdiagnoses.

- We show how to use the results of this enhanced fault diagnosis in EoL testing to effectively monitor the upstream assembly line. This allows for an early identification and evaluation of problems occurring in the assembly.

The remainder of this paper is structured as follows: In Section 2, we survey the domain of EoL testing and discuss prevalent data characteristics, as well as corresponding challenges for a domain-specific analytical approach. In Section 3, we discuss related work and their limitations regarding the considered domain. In Section 4, we present our approach to support operators in their fault diagnosis during EoL testing, as well as to draw the right conclusions for improving the assembly line. Section 5 focuses on our prototype and on the corresponding validation of our approach. We conclude in Section 6 and list possible future work.

2. Analytics and Characteristics in End-of-Line Testing

In this section, we present background information regarding EoL testing (2.1) and discuss the underlying data characteristics and the resulting analytical challenges (2.2).

2.1. *End-of-Line Testing in the Automotive Industry*

For OEMs, EoL testing is an important part of total quality control and represents the final functional check of assembled products. It simulates real operating environments under reproducible conditions in order to determine whether a product performs its function within predefined tolerances. Thereby, the product goes through a test bench that measures values of various sensors, e.g., power, pressure, or temperature.

To pass an EoL test, all values measured by key sensors must comply with the tolerances specified in the testing scheme. In case of limit violations, a product fails the test, i.e., it is declared as defective. Before such a defective product is scrapped, operators carry out repair attempts in a rework step and try to solve the quality issue. Thereby, the operators' most challenging task is to perform a fault diagnosis, to identify the faulty part that has to be repaired or replaced. The only available information they have are symptoms in the form of suspicious sensor values. Thus, the fault diagnosis and rework highly depends on the operators' subjective knowledge regarding the causes of these suspicious sensor values. Usually, they carry out up to four repair attempts and re-tests until they can solve a particular quality issue or until they scrap the product after all. Thereby, the chance of misdiagnosis increases with the complexity of the product. Powertrain aggregates, for instance, usually require multiple repair attempts. This is because they consist of numerous interdependent parts, resulting in a high product complexity. Hence, operators' waste a substantial amount of time on fault diagnosis and on carrying out ineffective repair attempts.

2.2. *Data Characteristics and Analytical Challenges*

We want to support the operators in their fault diagnosis by offering them a small list of most likely faulty parts. We create this list via data analytics by using data mining methods.

Fig. 1 shows the process of applying data analytics for a decision support. The input of the process are *historical source data*, which need to be *prepared* and *preprocessed* before they can actually be used for data analytics. *Data preparation* includes steps for *extracting* relevant data from the sources, *cleaning* these data to increase their quality, and *integrating* the data into one common view (*prepared data set*). The *preprocessing* step transform the prepared data into a suitable format to apply analytics on it. The descriptions of the data sources are given below. In the subsequent *training phase*, the preprocessed data is divided into two parts: the *training data* and the *testing data*. The training data serves as input for data mining algorithms (e.g., C4.5 [6] for classification). The output of such an algorithm is a data mining *model*, whose *performance* is then *evaluated* by applying the model to the *testing data*. If the model performance is insufficient, the training data and/or the data mining algorithm need be adjusted and applied again. Otherwise, the model can finally be applied to *productive data* in order to perform the actual *prediction* or recommendation task. In our case, we *apply* the final *model* to recommend the top k faulty parts, this way supporting an operator in his/her fault diagnosis.

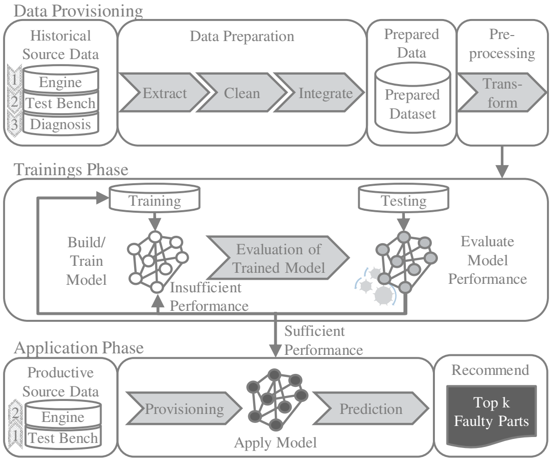


Fig. 1. Process of Data Analytics for Decision Support.

A representative use case to describe the prevalent data characteristics in the EoL testing area are engines for trucks. Engines are complex aggregates, with different construction types and a wide range of possible root causes for quality issues. For this reason, these aggregates also have a complex test setup, with numerous sensors in different torque levels.

We have collected the data from different data sources of a large automotive OEM: Engine data, test bench data, and diagnosis data.

- *Engine*: This data originates from an Enterprise-Resource-Planning (ERP) system for production planning. It includes IDs for the tested engines with their product specifications, such as a construction type describing technical properties like the engine's power. This data source includes more than 90 different construction types for 330 thousand engines.

- *Test bench*: This data originates from a test bench software and describes the test runs for engines. This means, it includes all test runs with the testing schema, values of different sensors, and the pass/fail decision. Data cleaning comprises steps for selecting failed test runs only and eliminating inconsistent data, such as negative values for the power sensor. This cleaning step significantly reduces the final amount of data, especially because the original data set consists of a high amount of inconsistent data. In the considered use case, for example, the final amount of data is reduced to only 2.7% of all test runs.
- *Diagnosis*: This data originates from a quality management system for recording product errors in assembly. Each entry includes the affected engine, the relevant stations in the assembly, and the detected error. The detected error describes the faulty part, which is encoded by an error code. Data cleaning comprises steps for filtering error codes of stations that are relevant to EoL rework. Furthermore, only meaningful error codes are of interest. For instance, error codes like “engine broken” are excluded, as they do not provide any valuable information for fault diagnosis. After cleaning and integration, less than 1% instances remain.

Diagnostic data can only be assigned to products, but not to test runs. This makes it difficult to find data entries in all three data sources that belong together and that may thus be integrated into the input of data analytics. In fact, this data integration has to be based on logical assumptions, again reducing the final amount of data to around one thousand data entries. These entries comprise 123 relevant error codes. Thereby, top 10 error codes represent approximately 40% of all available instances, leading to an extremely uneven distribution of error codes among the data.

Based on these data characteristics, we can derive the following challenges for developing data analytics solutions:

- **C1:** The first challenge arises from using historically grown and proprietary data sources, which can only be partially linked together. For this reason, both the amount and quality of data serving as input for data analytics is limited. However, data analytics requires a minimum number and quality of data to produce reliable analytical results.
- **C2:** Our second challenge results from the heterogeneous product space in our data, represented by the 90 different construction types of engines. From an analytical perspective, we have different facts (i.e., different sensors) in our data, which must be treated separately. Therefore, each construction requires one separate analytical model.
- **C3:** Our third challenge is related to the goal of building reliable analytical models based on our data characteristics. We have only few usable data with numerous error codes, which are the classes we want to predict with our final model. These error codes are unevenly distributed and, thus, the data is biased in favor of the top 10 error codes.

3. Data Mining and its Limitation in Manufacturing

Based on a literature survey of existing data mining reviews in manufacturing [7–11], we identified research gaps that are relevant to the data characteristics and analytical challenges described in the previous section. Table 1 shows the allocation of these research gaps to the related analytical challenges.

- **RG1:** A multiplicity of approaches propose a holistic data warehouse, without addressing the data sources with their inherent characteristics of the manufacturing domain. This especially holds for the data of the EoL testing area.
- **RG2:** For most approaches, the prediction of classes is limited to two classes: OK vs. not OK. Our data set however comprises a total amount of 123 error codes representing the classes. This restricts the selection of classification methods. In particular, these methods and resulting classification models must be able to deal with a multiplicity of classes and provide probabilistic results in order to get a top k list.
- **RG3:** During model evaluation, most related approaches try to maximize the classification accuracy by comparing different algorithms against each other. They however do not consider scenarios, where classes – in our case error codes – are distributed unevenly in the data. This means that most of these error codes are described by a relatively small amount of data. Usually, related approaches only recommend classes that belong to the small set that is represented in the majority of the data (the top 10 error codes in our case). Depending on the objective, it may however be important to suggest also seldom error codes. Furthermore, none of the approaches considers economic aspects of the prediction results from a business perspective.
- **RG4:** No approach adequately considers the analytical skills of the involved users, especially with respect to users from manufacturing. Hence, these users have difficulties to use corresponding analytical systems and to get the desired results. Thereby, the most tedious task is the selection of the correct model to be applied, as well as the regular update of this model. Challenge C2 refers to this gap, since the increased complexity through the heterogeneous product space results in a multiplicity of different models.

In addition to the reviews, we have examined three implemented system approaches. These approaches are designed to support human experts in their decision making for quality-related tasks. Table 1 shows to what extent the approaches meet the research gaps, respectively the challenges.

Table 1. Research gaps and challenges met by related approaches.

Research Gap	References	Challenges			Examined Approaches		
		C1	C2	C3	[4]	[5]	[3]
RG1	[7–10]	✓	-	-	◑	◑	○
RG2	[7,9,10]	-	-	✓	○	●	○
RG3	[7,9,10]	-	-	✓	○	○	-
RG4	[7–11]	-	✓	-	◑	◑	◑

✓: Challenge related to research gap | ●/◑/○: Research gap completely/partially/not fulfilled

The Advanced Manufacturing Analytics (AdMA) platform focuses on optimizing manufacturing processes through adjustments of process characteristics [4]. The AdMA platform proposes a general and holistic data warehouse that only integrates data from well-defined information systems (e.g., from an ERP). However, it does not address the complex data characteristics that are inherent in manufacturing (RG1). Furthermore, the data mining task is only a two-class classification (OK vs. not OK) (RG2). The author does not consider any evaluation of the results at all (RG3). The user perspective is considered, but without taking into account updates of models (RG4).

Kassner et al. define a Quality Analytics Toolkit (QATK) to recommend likely error codes based on the automatic recognition of errors mentioned in textual quality reports [5]. This implementation integrates unstructured text data from different tables stored in a single data source, i.e., it does not integrate several heterogeneous data sources at all (RG1). The classification task is a complex one with more than 500 classes and with a relatively large data set to build models (RG2). However, the authors do neither consider the uneven class distribution nor do they offer any economic evaluation of the predicted results (RG3). Like the AdMA platform, the update of models is not considered (RG4).

The ProDaMi is a modular data mining application with a quality management component [3]. This application mentions the importance of data preparation and refers to a data warehouse, but without addressing further details regarding the heterogeneity and complexity of data (RG1). The author do not provide any information regarding the number of classes or the way how analytical results are validated (RG2 and RG3). Nevertheless, the application offers a user interface for viewing results, but it still does not remove the burden from users to select the models that shall be applied and update them (RG4).

In summary, none of the approaches fully covers all data characteristics and challenges. Instead, each approach focuses on individual aspects that are of particular interest to the respective research goals.

4. Domain-Specific Approach to Support Fault Diagnosis and Quality Control in End-of-Line Testing

Now, we present an extended process for the EoL testing (4.1). Furthermore, we propose an architecture for a fault diagnosis and quality control system (4.2).

4.1. *Extended Process for End-of-Line Testing*

To implement a domain-specific approach, we extend an EoL process by novel analytical components for *data provisioning*, *fault diagnosis* and *quality control*, as shown in Fig. 2. The components *data provisioning* and *fault diagnosis* mainly support the application phase shown at the bottom of Fig. 1. The component *quality control* additionally helps to monitor and improve the assembly line. The overall process is now based on five major steps:

1. The *pass/fail decision* is based on sensor values from the test bench (cf. Section 2.1).
2. The *data provisioning* extracts and integrates relevant data from the productive data sources, i.e., execution data of failed test runs and data of the affected engines.
3. The *fault diagnosis* applies a data mining model on the provisioned data to recommend a list of top k error codes (possibly faulty parts) to the operator. Furthermore, this step documents the error code an operator finally selects for an affected test run.
4. Based on the fault diagnosis, the operator repairs or replaces the faulty part to fix the quality issue. After s/he has completed this *rework* step, the engine is tested again. In case of a total economic loss, the unit is scrapped. Therefore, the operator has to document the real findings and the conducted repair with its effectiveness.
5. The *quality control* is based on real findings from the rework step. It enables the identification and the initiation of countermeasures in the assembly.

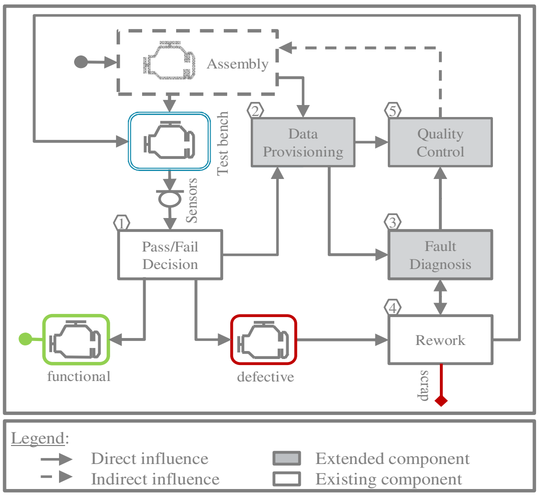


Fig. 2. Extended Process for EoL Testing Area.

4.2. *Fault Diagnosis and Quality Control Architecture*

Fig. 3 shows our analytical architecture that supports operators in fault diagnosis and quality control. Besides the source data and analytical data (4.2.1), the architecture compromises the main component recommendation and assembly monitoring (4.2.2).

4.2.1. *Source Data and Analytical Data*

Source Data: Except for the assembly data, all source data are described in Section 2.2. *Assembly* data is necessary for monitoring the assembly during the quality control step in Fig. 2. This data originates from a production process system. It defines the structure of an assembly line by associating assembly stations with their upstream and downstream stations.

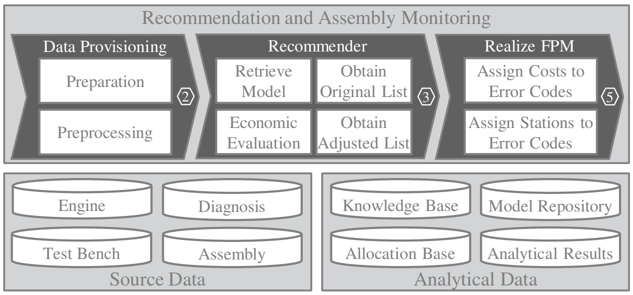
Analytical Data: These data do not originate from a source system. They are generated either by the analytical process or by the testing engineer or manager. The *knowledge base* results from data preparation and stores the prepared data set. The prepared data set is being preprocessed and used as an input for the training phase (cf. Fig. 1). The *model repository* manages different data mining models that are generated by this training phase. The *analytical results* store the recommended top k faulty parts and the operator's selected error code. The *allocation base* store the allocation of costs and stations for error codes to implement the quality monitoring.

4.2.2. *Recommendation and Assembly Monitoring*

Data Provisioning: This component implements step 2 in Fig. 2 and addresses research gap RG1. The scientific community usually assumes a holistic data warehouse, e.g., as offered by Gröger et al. [4]. However, the data schema of a data warehouse is often tailored to support only predefined analyses.

In case different kinds of analyses are needed, the data schema and corresponding data integration processes have to be adjusted. This leads to a high effort for implementing and maintaining a data warehouse. In our case, it is more efficient to extract and prepare relevant data from the source systems and to store them in a *knowledge base*. The *knowledge base* is a large table without a rigid schema, which makes the approach more flexible with respect to additional data analyses.

Mechanisms for extracting, cleaning, and storing data are reflected in the *preparation* subcomponent. The *preprocessing* subcomponent allows for transforming the data into a suitable format as required for the training and application phases.



FPM: Failure Process Matrix

Fig. 3. Architecture of Recommendation and Monitoring System. {2}/{3}/{5}: Represent the steps from Fig. 2.

Recommender: This component implements step 3 in Fig. 2 and addresses research gaps RG2, RG3, and RG4. The models are subdivided by a differentiation criterion, in our case the construction type (cf. Section 2.2). To address research gap RG4, the subcomponent *retrieve model* uses this criterion to select and apply the correct model for the relevant test run that has been performed. The applied model *obtains* the *original list* of recommendations. This *original list* of top k error codes contains most likely error codes in descending order of their likelihood. However, this list does not consider any economic aspects so far. For instance, it is economically unreasonable to suggest parts that have a marginal higher probability than other parts, but that cause much higher costs to replace or repair them (e.g., crankshafts). Thus, we *evaluate* the ranking of top k error codes *economically* according to a cost function to *obtain* an *adjusted list*. This addresses one aspect of research gap RG3. Both the *original* and the *adjusted list* are stored in the database for *analytical results*, together with the underlying test run and the error code finally selected by the operator.

Research gaps RG2 and RG3 influence each other, i.e., the fact that the data contains numerous classes also increases the uneven distribution of these classes and vice versa. To address this, we adopt an established method for data over- and undersampling: the SMOTE algorithm [12]. Instead of simply duplicating entries, this algorithm creates new synthetic data instances that are interpolations of the seldom error codes. Note that this method is also able to boost the original small data set.

To support regular updates of models (cf. RG4), we train the models after a certain amount of new *diagnosis* data becomes available. Subsequently, the *knowledge base* is updated and the newly trained models are stored in the *model repository*.

Failure Process Matrix: This component implements step 5 in Fig. 2. It monitors the assembly and also allow us to consider economic aspects in the recommendation of error codes (cf. RG3). Schloske et al. propose the failure process matrix (FPM) for quick analysis and economical optimization of complex assembly processes [13].

For our approach, we have adapted the FPM to tailor it to the domain of EoL testing. The columns of the matrix describe the chronological order of the process steps in the assembly line (see the right side of Table 2). The rows show the error codes detected in EoL testing. In case a matrix field is set with a value, the relevant error code is caused by the associated process step. The actual value of the matrix field specifies how often the process step has caused this error during a certain period of time. This period of time describes the validly of the assignments to error codes and is defined on the left side of the matrix (valid_from, valid_to). Furthermore, this left side associates each error code with the costs for reworking on the faulty part. As a result, the matrix indicates the process steps having quality issues and the respective costs.

Table 2. Failure Process Matrix.

Vaild_From	**Valid_To**	**Rework (€/Total)**	**Rework (€/Failure)**	**Failure (Error Code)**	**Process Step 1**	**Process Step 2**	**Process Step …**
12/11/2017	12/31/2018	€ 300	€ 100	Failure A	3		
12/7/2017	12/31/2017	€ 600	€ 300	Failure B		2	
				Failure …			

5. Prototypical Implementation and Validation

As discussed in Section 4.2, different components of the architecture deal with different research gaps being relevant to the domain of EoL testing. For example, the *data provisioning* component provides data for the *knowledge base*, on which models can be trained without requiring a holistic data warehouse as data source (cf. RG1). Furthermore, we use the SMOTE algorithm in the component *obtain original list* to overcome the challenges with numerous (cf. RG2) and unevenly distributed classes (cf. RG3). Furthermore, both the subcomponent *economic evaluation* and the failure process matrix consider rework costs and thus economical aspects while predicting a list of top k error codes (cf. RG3). The proper model to be applied is selected by the subcomponent *retrieve model* to unburden the operator from this task (cf. RG4). For the first time, this introduces an approach that addresses all EoL-specific challenges in a holistic way.

We have implemented our approach for the use case and data from the real-world OEM mentioned in Section 2.2. We thereby followed the well-known cross-industry standard process for data mining (CRISP-DM) [14]. Especially in the first three steps of CRISP-DM, we closely collaborated with the domain experts of the OEM to understand their business and to prepare the data properly. For data storage, we use Microsoft SQL Server 2014[1]. The *preparation* subcomponent is implemented in IBM SPSS Modeler 16.0[2]. We have built two

---

[1] https://www.microsoft.com/en-us/evalcenter/evaluate-sql-server-2014-sp2

[2] https://www.ibm.com/us-en/marketplace/spss-modeler

preparation steps, one for the training phase and one for the application phase. Furthermore, we use R[3] to implement the *preprocessing* subcomponent.

The *recommender* focuses on data mining and is thus implemented in R. We have used different data mining algorithms from the "caret" package [4] to build models recommending error codes. To overcome our challenges, we have decomposed this problem into sub-problems, i.e., each error code corresponds to a model (class model) that indicates the likelihood of this particular error code. For each class model, we have trained, tested, and evaluated several algorithms. We have then selected the algorithm with the highest performance measure, in our case the F_1 score, for each class model. The F_1 score estimates how *exact* and *complete* a model can predict the error code of interest. All available class models for one construction type result in an ensemble that is used for predicting the top k error codes.

To evaluate the resulting ensemble, we use the F_1 score as a primary and accuracy as secondary performance measure. Accuracy is defined as the percentage of test cases, where the ensemble predicts the correct error code. We have measured F_1 score and accuracy (Acc.) for recommendation lists containing the top k = 1, 5, and 10 error codes (PM@k). We compare the results with a baseline, i.e., the heuristic frequency metrics. This means the baseline always uses the error code that occurs most often in the whole preprocessed data set.

Table 3 shows that the ensemble delivers better accuracy results than the baseline only for a higher value of k. Nevertheless, the F_1 score is significantly better than the baseline for all values of k. Thus, the ensemble is more *exact* and *complete* in the prediction of error codes as the baseline. This shows the potential of our analytical approach and the underlying data mining algorithms to recommend correct error codes. This in turn helps operators during fault diagnosis to better identify the faulty part that has to be replaced or repaired.

Table 3. Results of experiment for one single construction type.

Ensemble / Base line	PM@1		PM@5		PM@10	
	F_1	Acc.	F_1	Acc.	F_1	Acc.
Ensemble	~3%	~8%	~62%	~64%	~77%	~79%
Code Frequency	~1%	~13%	~15%	~41%	~38%	~58%

We also performed a qualitative validation with domain experts and managers to show the advantages of our analytical approach. The proposed error codes have especially helped the testing engineers in cases, where there was no evidence for a cause of an error at all. Furthermore, managers can now use the FPM matrix to quantify the costs and point out possibilities to optimize the assembly.

6. Conclusion and Future Work

In this paper, we introduced an analytical approach that helps operators in fault diagnosis and quality control in the EoL testing area by using data mining. The approach especially considers the data characteristics being prevalent in industrial settings. So, it can be reused as a blueprint for such scenarios.

We have also shown how to implement this approach in a real-world use case of a big OEM. The statistical evaluation shows that the prediction performance gets much better when using data mining algorithms compared to the baseline approach. This way, our appraoch offers a robust solution to fault diagnosis and quality control, as well as to reduce the amount of unnecessary rework.

So far, the analytical approach does not consider operators' feedback in order to process them for further recommendation improvements. These and other topics, such as the enhancement of the FPM with additional features like audit costs, will be the subject of future work.

References

[1] Töpfer A. Six Sigma: Projektmanagement für Null-Fehler-Qualität in der Automobilindustrie. [December 10, 2017]; Available from: http://www.sixsigma-akademie.de/_pdf/projektmanagement.pdf.
[2] Henke J. Eine Methodik zur Steigerung der Wertschöpfung in der manuellen Montage komplexer Systeme. Stuttgart: Fraunhofer Verlag; 2016.
[3] Bernard T. Datenbestände optimal nutzen - Entscheidungsunterstützung im Produktionsumfeld mit Data-Mining-Werkzeugen. In: Müller K, editor. Prozesstechnik & Automation: P&A Kompendium 2011-2012. München: publish-industry; 2011, p. 115–117.
[4] Gröger C, Niedermann F, Mitschang B. Data Mining-driven Manufacturing Process Optimization. In: Proceedings of the World Congress on Engineering 2012 Vol III (WCE). Hong Kong: U.K. International Association of Engineers (IAENG); 2012, p. 1475–1481.
[5] Kassner L, Mitschang B. Exploring Text Classification for Messy Data: An Industry Use Case for Domain-Specific Analytics Technology. In: Pitoura E, Maabout S, Koutrika G, Marian A, Tanca L, Manolescu I et al., editors. Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016. OpenProceedings.org; 2016, p. 491–502.
[6] Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 3rd ed. Waltham Mass. u.a.: Elsevier Morgan Kaufmann; 2012.
[7] Cheng Y, Chen K, Sun H, Zhang Y, Tao F. Data and Knowledge Mining with Big Data towards Smart Production. Journal of Industrial Information Integration 2017.
[8] Choudhary AK, Harding JA, Tiwari MK. Data mining in manufacturing: A review based on the kind of knowledge. J Intell Manuf 2009;20(5):501–21.
[9] Harding JA, Shahbaz M, Srinivas, Kusiak A. Data Mining in Manufacturing: A Review. J. Manuf. Sci. Eng. 2006;128(4):969.
[10] Köksal G, Batmaz İ, Testik MC. A review of data mining applications for quality improvement in manufacturing industry. Expert Systems with Applications 2011;38(10):13448–67.
[11] Wang K, Tong S, Eynard B, Roucoules L, Matta N. Review on Application of Data Mining in Product Design and Manufacturing. In: Lei J, Yu J, Zhou S, editors. FSKD 2007: Proceedings. Los Alamitos, California: IEEE; 2007, p. 613–618.
[12] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research 2002(16):321–57.
[13] Schloske A, Henke J. Failure Process Matrix (FPM) - a new approach for the optimization of assembly lines. In: Westkämper E, editor. The 1st CIRP International Seminar on Assembly Systems. Stuttgart; 2006, p. 257–260.
[14] Shearer C. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing 2000;5(4):13–22.

---

[3] https://www.r-project.org/

[4] https://topepo.github.io/caret/index.html