Annals of Operations Research
https://doi.org/10.1007/s10479-021-04215-9

ORIGINAL RESEARCH

Data analytics for quality management in Industry 4.0 from a MSME perspective

Gorkem Sariyer[1] · Sachin Kumar Mangla[2] · Yigit Kazancoglu[3] · Ceren Ocal Tasar[4] · Sunil Luthra[5]

Accepted: 20 July 2021

Abstract
Advances in smart technologies (Industry 4.0) assist managers of Micro Small and Medium Enterprises (MSME) to control quality in manufacturing using sophisticated data-driven techniques. This study presents a 3-stage model that classifies products depending on defects (defects or non-defects) and defect type according to their levels. This article seeks to detect potential errors to ensure superior quality through machine learning and data mining. The proposed model is tested in a medium enterprise—a kitchenware company in Turkey. Using the main features of data set, product, customer, country, production line, production volume, sample quantity and defect code, a Multilayer Perceptron algorithm for product quality level classification was developed with 96% accuracy. Once a defect is detected, an estimation is made of how many re-works are required. Thus, considering the attributes of product, production line, production volume, sample quantity and product quality level, a Multilayer Perceptron algorithm for re-work quantity prediction model was developed with 98% performance. From the findings, re-work quantity has the highest relation with product quality level where re-work quantities were higher for major defects compared to minor/moderate defects. Finally, this work explores the root causes of defects considering production line and product quality level through association rule mining. The top mined rule achieves a confidence level of 80% where assembly and material were identified as main root causes.

Keywords MSME · Machine learning · Quality control · Industry 4.0 · Data analytics · Manufacturing · Association rule mining · Re-work and root causes of defect

1 Introduction

Micro Small and Medium Enterprises (MSME) play a key role in the economic growth of a country by generating a huge pool for employment opportunities (Savlovschi & Robu, 2011). MSMEs account for 99% of all firms and over half of total employment in the global economy (Ferrando et al., 2017; International Monetary Fund, 2019). In the globalization

✉ Sunil Luthra
sunilluthra1977@gmail.com

Extended author information available on the last page of the article

era, MSMEs are strained to overcome barriers to sustaining their competitive advantage and aim to enhance the competency to succeed (Ibrahim et al., 2016). Since emerging technologies have enabled increased capabilities and reduced cost of information in the global arena, taking advantage of these advances becomes a necessity for MSMEs in their manufacturing operations.

Recent advances in emerging technologies have transformed industrial systems and manufacturing processes to become smarter, flexible and automatic to satisfy increasingly competitive business requirements and demands (Lee et al., 2014). Thus, practically all areas of today's industrial activities are moving towards process automation. Such a transition into automation should assure product quality while minimising the production cost and optimising the resources available (Ferreiro et al., 2011). Industry 4.0 technologies can affect the industrial environment in terms of producing a flexible, traceable, data-driven networking system. During the process of adoption of these innovative technologies, companies may face many challenges, specifically addressing big data issues in decision-making in their efforts to improve quality and increase productivity (Lee et al., 2014). Increasing complexity in manufacturing operations and correspondingly an increase in the amount of data may cause problems for the user during process monitoring, data analysis and fault detection (Windmann et al., 2015). Due to a lack of smart analytic tools, many manufacturing processes may not be able to manage big data in a manufacturing system (Lee et al., 2014). Further, the process industry has to be re-formulated as an essential component of manufacturing in Industry 4.0 (Ge et al., 2017). For this purpose, data management and distribution are key aspects to be used as self-learning machines in a data driven manufacturing environment (Lee et al., 2014). Therefore, modern manufacturing factories operating in data-rich environments can deliver the distribution, sharing and analysis of information through sophisticated networks to embed intelligence into manufacturing (Davis et al., 2012; Lee et al., 2014). Hence, the application of data analytics tools in manufacturing control processes gains importance within industrial facilities and correspondingly, these manufacturing control processes prove to be highly efficient in quality control and Total Quality Management (TQM).

Innovative development leads to having big data analytics capabilities for businesses, particularly MSMEs, to ensure success in dynamically changing big data environments. In a world of rapid change, big data analytics is considered as a game changer enabling better business efficiency based on high operational and strategic potential (Wamba et al., 2017). Data analytics enables companies to prepare adequate contingency plans to address dynamically changing conditions (Sun et al., 2018) and has become a main component of decision making in companies (Chen et al., 2018). According to Liu (2014), data analytics is defined as a major differentiator between high and low performing organizations since it enables companies to be more proactive and forward looking, reducing costs by about 47% and improving profits by about 8%. From this perspective, data analytics capabilities have received increasing attention in literature. Kiron et al. (2014) defined data analytics capability as the competence to provide business insights using data management, technology and personnel (labor) capabilities to transform business by delivering a competitive advantage. Through data analytics capabilities, manufacturing companies can sense alterations in the rapidly changing business environments, collect and record critical and huge amounts of data; this can be analysed, turned into useful information, which can then be turned into action.

As companies focus on maximizing the quality of their goods, services and internal operations to improve their competitiveness (Chin et al., 2002), TQM has become more prevalent in operations and supply chain management literature; big data analytics

capabilities have become a more prominent issue in this context. TQM can provide many benefits through the accomplishment of specific quality principles and practices for manufacturing and service industries. There is a strong connection between adoption of TQM principles and manufacturing system performance (Chahal, 2015). The detection of possible errors at an early stage and the introduction of preventative measures is one of the fundamental functions of TQM; however, after detecting a defect, the "re-work" process is a real burden on businesses. From an industrial perspective, companies are facing challenges such as detection of the source of the problem, how much re-work is needed etc. (Chien et al., 2017; Wulfsberg et al., 2019). All of these challenges cause loss of time and money. In this sense, effective methods that enable early detection of potential errors, predicting the number of required re-works and identifying causes of potential defects would be highly advantageous to manufacturers. Existing research indicates that implementing emerging technologies results in better quality and argue that quality management is a prerequisite in MSMEs' (Bagodi et al., 2020; Kamble et al., 2020). However, there exists a gap in existing literature in developing a multi-stage model that aims to focus on different problems from a quality management context at each step by implementing big data analytics technologies.

Thus, this research proposes a 3 stage model with the aims of automated defect detection, predicting required numbers of re-works and identifying root causes of defects in each step. The integration of data analytics techniques can contribute to the area of quality management in MSME literature. Specifically, this study sets the following two research questions.

RQ1: How can big data analytics facilitate effective decision making in different quality management problems of MSMEs?
RQ2: Which features of the manufacturing process should be considered to enable managers to make predictions on their quality levels and examine the consequences?

This paper aims to propose a model that predicts any potential defects by grouping products depending on their defects (i.e. defect free or not) and then, classifying defects according to their levels. By this means, an increase in quality in the process will be achieved. This article seeks to detect any possible errors to achieve superior quality through data mining (DM) and machine learning (ML) techniques. This work further evaluates the re-work processes with proper estimations and provides estimates for how many re-works are required for the given features and defect level of the products generated. Finally, this work involves identification of the root causes in combination with specific features and defect levels by using association rule mining. From a holistic view, with the use of emerging data driven techniques, implementing such a model to improve quality in a manufacturing system is significant from an industrial viewpoint for the success of MSMEs. Therefore, this research contributes to current literature by proposing a novel multi-stage model which provides efficient solutions to different problems in quality management systems at each step. In addition, in each step of the proposed model, either by approaching the problem from a different perspective or by implementation of a different technique compared to existing studies, this study contributes to a development of the literature in this field. This research is contextualized in the kitchenware industry of a medium enterprise, where product quality depends on the continuous monitoring of any problems, identifying the reasons for them and dealing with the consequences within the manufacturing system. This industry is specifically chosen since due to high volumes and varieties of the product tree, the kitchenware

industry includes big data by its very nature; this can be properly analyzed by data analytics techniques. Thus, the model is also validated in a real-life data study.

The study is structured as follows. The literature review is given in Sect. 2. In Sect. 3, the proposed model is introduced. Section 4 presents background to the research methods. Section 5 shows implementation of the model in a case study. Results are further discussed in Sect. 6. Theoretical and managerial implications are presented in Sect. 7. Finally, Sect. 8 provides the conclusions.

2 Literature review

Following the advances in smart technologies, researchers have started to show an increasing interest in big data analytics in operation and supply chain management. Various studies have investigated the effects of these technologies in operational performance of manufacturing systems (Ahuett-Garza & Kurfess, 2018; Akter et al., 2020; Belhadi et al., 2019, 2020; Dubey et al., 2020; Fahmideh & Beydoun, 2019; Gu et al., 2021; Kamble et al., 2020; Wamba et al., 2020; Yadav et al., 2020; Yadegaridehkordi et al., 2018). Many other studies have analyzed the role of big data and big data analytics in supply chain management (Arunachalam et al., 2018; Chehbi-Gamoura et al., 2020; Hazen et al., 2018; Liu & Yi, 2018; Mishra et al., 2018; Wamba et al., 2018). All of these studies have demonstrated that use of big data analytics positively affects operation and/or supply chain management by leading improvements in system performance. However, these studies have methodologically presented empirical investigations, case studies or review and bibliometric analysis; there is a lack of proposing models by the implementation of big data analytics techniques to a specific context or problem in operations and supply chain management.

This study specifically focuses on quality management, not only since it is one of the most important contexts in operations and supply chain management literature but also since quality improvement programs require huge data in order to solve the related quality problem. Due to increased market pressure, new quality management systems are required in manufacturing ecosystems for MSMEs (Lee et al., 2019). With the developments in data management tools, big data analytics have gained importance for quality management and improvement in manufacturing.

A significant point receiving considerable attention in quality management is automated detection of defects. As large number of variables are involved in the manufacturing process, predicting defects has always been challenging for MSMEs. Chen et al. (2019) have analyzed quality management for big data systems. They proposed a dynamic coherent quality measure by characterizing the probability of critical errors. To diagnose process defects, Perzyk et al. (2014) implemented different big data analytic tools and compared their performances in terms of accuracies, robustness and applicability. They showed that simple statistical methods reveal the best performance, whereas advanced machine learning models such as neural networks and support vector machines appeared to have less value. Çiflikli and Kahya-Özyirmidokuz (2010) explored DM development related with manufacturing process specifically, through the detection and isolation of possible machine breakdowns in carpet production. A decision tree model was also conducted to further improve the manufacturing processes. Essa et al. (2019) adopted ML in a textile industry with the aim of classifying products as defect or normal, achieving 98.07% accuracy. Yapi et al. (2015) also addressed the textile defect detection problem by using ML approaches and obtained an error rate

from 0.90 to 2.80%. Law et al. (2017) analyzed automated defect discovery for dishwasher appliances from online consumer reviews; their logistic regression and neural network classifiers had an Area under Curve (AUC) around 94%. Peres et al. (2019) analyzed quality control problems using ML, achieving around 93% accuracy in classifying cars as "ok" or "nok" in an automotive industry. For early detection of product failures, Carvajal Soto et al. (2019) adopted an online ML framework and achieved 98% and 96% accuracies in train and test sets. Zhang et al. (2020) proposed deep learning models for defect detection and recognition of multivariate processes. All of these researchers show that the use of big data analytics technologies can effectively classify produced items based on their quality status as defect or non-defect. The provision of early detection of defects will save companies from additional losses. However, most of these existing studies are limited; they design the output variable of interest, the quality status of the item and have only two levels such as defect or non-defect. They do not make any further categorization on defect level. Identifying this as a gap in literature, this study proposes a five-level classification model for automated defect detection problems in which the defects are further classified based on their defect categories.

The concept of re-work is also very important in quality management. Defects in production cost time and money. Producing items with high values of re-work and scrap is not only costly but also damages the reputation of the company (Carletti et al., 2019; Chien et al., 2017). Further, it is also important to evaluate the main actors of a production system (such as machines, production lines, materials etc.) which may be responsible for more re-works while also generating predictions in re-work quantities to assist in planning of related operations (supply, budget, time scheduling etc.). Managers and practitioners might use these technologies to evaluate those actors, associated re-works and related operations. However, to the best of our knowledge, there is a lack of research in developing predictions on the required re-work levels in manufacturing systems by the use of big data analytics technologies.

Another major concern in quality management is understanding the root causes of defects. This can help companies to eliminate the potential causes and therefore provides a significant improvement in quality. Chien et al. (2017) proposed a novel data driven approach for analyzing semi-conductor manufacturing big data for low yield diagnosis to detect process root causes for yield enhancement. Carletti et al. (2019) proposed an approach for defining a feature importance in anomaly detection problems—an important ML task and extremely relevant for the purposes of quality monitoring. Dey and Stori (2005) developed and presented a process monitoring approach based on the Bayesian belief network for incorporating multiple process metrics in sequential machining operations to identify root causes of process variations. Lokrantz et al. (2018) proposed a ML framework using Bayesian networks to model the causal relationships between manufacturing stages using expert knowledge and to demonstrate the usefulness of the framework on two simulated manufacturing processes. Although these studies identified various root causes for quality problems specifically for the related industry which they analyzed, they all concluded that automatization of identification of root causes of defects would benefit MSMEs and produce knowledge for future use. However, these studies only analyzed root cause detection problems in a quality management context by using ML techniques to identify the root causes. By analyzing the root causes of defects in manufacturing systems with another big data analytics technique, namely Association Rule Mining, and additionally focusing on two of the previously mentioned problems in the field of quality management, this study differs from previous research papers.

A summary of literature studies is presented in Table 1.

Table 1 Review of existing studies

Study	Aim: Automated detection of defects			Aim: Prediction of required reworks		Aim: Identification of root causes	
	Yes/no	Method	Classification level	Yes/no	Method	Yes/no	Method
Chen et al. (2018)	Yes	Computational methods	2 (defect/non-defect)	No		No	
Perzyk et al. (2014)	Yes	DM & ML	2 (defect/non-defect)	No		No	
Çiflikli and Kahya-Özyirmidokuz (2010)	Yes	DM	2 (defect/non-defect)	No		No	
Essa et al. (2019)	Yes	ML	2 (defect/non-defect)	No		No	
Yapi et al. (2015)	Yes	ML	2 (defect/non-defect)	No		No	
Law et al. (2017)	Yes	DM & ML	2 (defect/non-defect)	No		No	
Peres et al. (2019)	Yes	ML	2 (defect/non-defect)	No		No	
Carvajal Soto et al. (2019)	Yes	ML	2 (defect/non-defect)	No		No	
Zhang et al. (2020)	Yes	ML	2 (defect/non-defect)	No		No	
Chien et al. (2017)	No			No		Yes	ML
Carletti et al. (2019)	No			No		Yes	ML
Dey and Stori (2005)	No			No		Yes	ML
Lokrantz et al. (2018)	No			No		Yes	ML
This study	Yes	ML	5 (non-defect/4 different defect categories)	Yes	ML	Yes	Association rule mining

As presented in Table 1, this study mainly contributes to existing literature by focusing not only on a specific problem, but three different and related problems in quality management. To achieve this aim, a 3 stage model is proposed. This model initially predicts the potential defects (defect or non-defect) and then classifies defects according to their levels. In the second stage, for those items classified as defects, the required level of re-work is predicted. Thus, re-work processes with proper estimations are generated. Next, the root causes are combined with specific features and defect levels using association rule mining. To realize the above mentioned aims, this paper implements ML and DM based emerging technologies in the proposed model. In this way, this study may serve as a template for MSMEs in integrating different data driven technologies to improve quality levels from different aspects.

3 Proposed model

In this paper, a 3 stage model is proposed to examine and evaluate the research questions. A widely recognized problem in quality control is defect detection in which produced items are labeled as defects or non-defects. However, when a defect exists, identifying the type of defect is also very important as this provides a detailed description of production quality levels of the company. Therefore, in the first stage, items are classified not only as defect or non-defect but also according to Product Quality Levels (PQL).

When defect and also type or level of this defect is detected during a quality control step, the production department is faced with an important challenge: "How many of the items produced in this order will require re-work?" Thus, the second stage of the model develops accurate predictions for Re-work Quantities (RQ).

In addition to predicting re-work quantities when a defect appears in production, another important challenge for companies is to understand the root causes of this defect. Understanding root causes provides deeper insights and guidance for managers in dealing with quality management. With this goal, the third stage of the model aims to discover Root Causes (R–C) of detected defects.

In the first two stages, a Multi-Layer Perceptron (MLP) algorithm for classification and prediction of outcome of interest is developed; this is underpinned by ML. The third stage utilizes one of the well-known DM techniques, Association Rule Mining (ARM), to uncover hidden rules for root causes. The proposed research model is presented in Fig. 1.

The proposed 3 stage model was implemented and tested in a case study of a medium enterprise in the kitchenware industry. The case study, data set specifications as well as the model results are given in Sect. 6.

4 Research methods

4.1 Multilayer perceptron

Due to increased research attention on artificial neural networks, both researchers and practitioners are focusing on gaining insights from data by applying neural network learning algorithms. Perceptron, the initial model of artificial neural networks, is defined with only one input and output layer (Hu, 2014). A major limitation of the perceptron model is about handling complex data patterns, nonlinear tasks and serving as a linear model. To

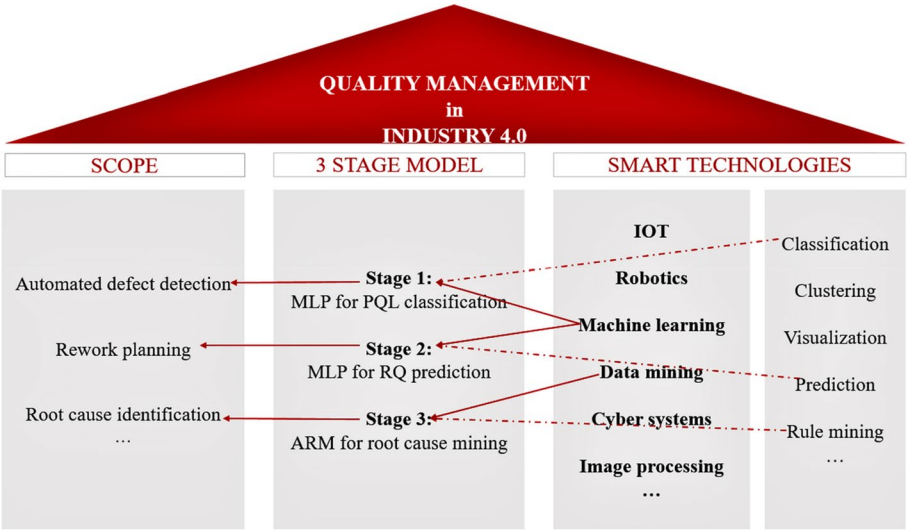


Fig. 1 Proposed model

overcome these limitations, a MLP model that positions hidden layers between input and output layers and incorporates the fundamental principle of feed-forward flow of information has been developed (Chen et al., 2015; Tsai & Huang, 2017). While building up a train model, a set of connected weights is computed considering the input features (x_1, x_2, x_3, …, x_n); these are iteratively learned in each layer until achieving the weights (w_{1j}, w_{2j}, w_{3j}, …, w_{nj}) with best score. A MLP network with one hidden layer is illustrated in Fig. 2.

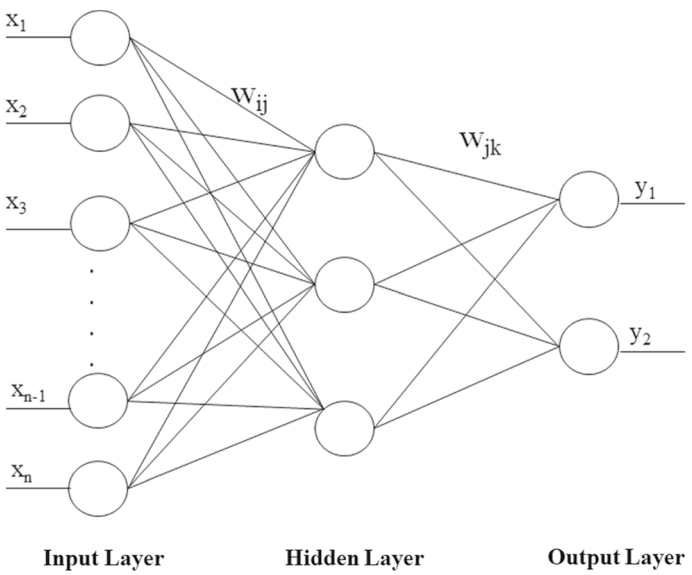


Fig. 2 MLP network

A high variety in data volume and dimensions results in the need for customisation of three main parameters; these are learning rate, momentum and number of hidden layers in MLP networks. Learning rate represents the degree of training speed of the network. This means that increased learning rate directly improves training speed of the network at a cost of an unstable network. Momentum is used to prevent building an unstable network caused by high learning rate by balancing the network. The number of hidden layers can be extended due to the nature of input data. As the number of hidden layers increases, the number of input feature combinations is extended as well.

MLP techniques are superior in pattern classification and capturing the complex behaviours of the input data set. Another important characteristic of MLP networks is their abilities in approximation of any linear or non-linear functions. Neural network architecture in a dense approach is critical for improving accuracy (specifically in the case when dealing with high numbers of input features) in comparison to other classification techniques in machine learning. With a sequential MLP approach, it is possible to extend analysis for domain specific problems by adding different types of ANNs such as convolutional, sequential or recurrence models. As a disadvantage, hyper parameter tuning takes on an important role to avoid over fitting/under fitting and exploding gradient problems; the requirement of time to train the data set is another disadvantage.

Highly correlated values exist in the model, making MLP a good fit for this 3 stage model. Since the data set of this study included a complex structure with different variables and measurements (such as nominal and numeric), linearity cannot be assumed due to such a complex structure. Therefore, the preference was made to use MLP neural networks in this study. Besides, time complexity was not a concern in this data set as the algorithm was trained rapidly.

4.2 Association rule mining

Association Rule Mining (ARM) and frequent item set mining are two of the most widely used DM techniques for a variety of applications (Li et al., 2016; Viet et al., 2020). ARM is a method of extracting relations and dependencies between input features in a structured data set based on transactional instances. By discovering patterns, rules are framed for corresponding values of input features.

Let 'T' be a set of n transactions represented by $\{T_1, T_2,\ldots,T_n\}$ and 'I' be a set of items represented by $\{i_1, i_2,\ldots,i_n\}$ where $T_i \subseteq I$. In the case of, $X,Y \subseteq I$ and $X \cap Y = \varnothing$, $X => Y$ represents an association rule as X is the antecedent (if-part) and Y is the consequent (then-part) of the rule. Antecedents or consequents of the association rules are known as items in the data set.

To evaluate quality of the association rules, two primary metrics, support and confidence, are used. For $X \subseteq I$, support(X) is the measure of how frequently X appears in the data set. Similarly, support of the rule is represented as $\text{support}(X => Y) = \text{support}(X \cup Y)$ and indicates how frequently both X and Y appear together in the data set. Confidence is the measure of reliability of a rule and computed as $\text{confidence}(X => Y) = \text{support}(X \cup Y)$ / support(X). Additional significant metrices for support and confidence to indicate quantification of prediction or classification capability of the implication are evaluated as $\text{lift}(X => Y) = \text{confidence}(X => Y)$ / support(Y).

A well-known algorithm in association rule mining, Apriori, is used in this study. Association rules are extracted from frequent item set combinations and filtered through the given support and confidence threshold values. Let 'I' represent the item set with length n.

I is frequent if and only if every subset with length n-1 is frequent as well. Thus, computational space and time complexity are significantly reduced to generate association rules. Although this algorithm has two problems, generating huge numbers of candidate data sets and constantly scanning this database to identify large sets of candidates, it is superior for identifying frequent patterns, associations and correlations from various data types such as transactional or relational data. Apriori has been widely implemented in DM literature (Soni et al., 2016).

5 Industrial case application

5.1 Case study specification

Empirical data for this study was collected from a multi-national company operating in the kitchenware industry. The company is involved in manufacturing and commercialisation of kitchen and bath products. The company has many manufacturing sites in Europe, America and Asia and promotes its products in more than 100 countries. It has a wide product tree ranging from sinks and taps to ovens, washing machines, extractor hoods and induction hobs. Referring to the OECD definition, the company is a medium enterprise since it has 50–249 employees. The company uses advanced manufacturing technologies in their business.

The data used in this study considers one manufacturing site of this company with 12 different lines of production and one additional line for re-working on defects. This company currently has a production capacity of more than 400,000 units which are produced for more than 3000 orders. However, on average, around 5% of the produced items require re-work, creating high additional costs for the company. The company uses acceptance sampling and depending on the produced item, it samples between 1 and 5% of the production volume of that item, checking the quality of sample. If no significant defect is detected in the sample, the batch passes the quality control stage. Otherwise, the company decides on how many items in the batch requires re-work. These items pass through a re-work line where remedial action is taken before shipping products to customers.

5.2 Data set characteristics and preprocessing

5.2.1 Data set characteristics

In this case study, the data set encompasses one year of data from 2019 with 3175 instances (rows) in total. In this year, the company produced 385 different types of product for 36 different customers located in 28 countries. Production was realised in 12 different lines. For each order, the number of produced items, production volume, number of items which are checked for quality and sample quantity are considered. The frequency and percentage distributions of number of orders and production volume based on country, customer and production line are summarised in Table 2.

From Table 2, it can be seen that the number of orders and production volumes differ significantly between countries, customers and production lines. The vast majority of production that covers not only number of orders, but also production volume, is delivered to Spain. Customer_1 represents the largest customer group of the company. While the number of orders produced in Line 308 is the highest and total production volume is highest

Table 2 Distributions of number of orders and production volume

	Number of orders		Production volume			Number of orders		Production volume	
	Freq	%	Freq	%		Freq	%	Freq	%
Country					*Customer type*				
Azov	2	0.063	100	0.023	Customer-1	1730	54.488	261,312	59.334
Germany	44	1.386	6375	1.448	Customer-2	406	12.787	47,088	10.692
USA	62	1.953	11,319	2.570	Customer-3	174	5.480	16,895	3.836
Australia	7	0.220	576	0.131	Customer-4	157	4.945	12,368	2.808
Belgium	33	1.039	2164	0.491	Customer-5	95	2.992	7421	1.685
Arab Emir	14	0.441	1057	0.240	Customer-6	85	2.677	19,837	4.504
Bulgaria	74	2.331	8387	1.904	Customer-7	58	1.827	8501	1.930
Britain	31	0.976	2670	0.606	Customer-8	84	2.646	15,054	3.418
Algeria	70	2.205	13,759	3.124	Customer-9	68	2.142	7687	1.745
Denmark	3	0.094	161	0.037	Customer-10	42	1.323	3542	0.804
Ecuador	2	0.063	424	0.096	Customer-11	62	1.953	11,319	2.570
France	136	4.283	31,011	7.041	Customer-12	33	1.039	2177	0.494
General	15	0.472	2128	0.483	Customer-14	52	1.638	8573	1.947
England	14	0.441	763	0.173	Customer-15	23	0.724	3913	0.888
Spain	1955	61.575	290,464	65.953	Customer-18	5	0.157	819	0.186
Israel	8	0.252	580	0.132	Customer-19	9	0.283	2714	0.616
Italy	44	1.386	5955	1.352	Customer-20	5	0.157	262	0.059
Hungary	4	0.126	158	0.036	Customer-23	5	0.157	1395	0.317
Mexico	37	1.165	4921	1.117	Customer-24	8	0.252	331	0.075
Mid.East	1	0.031	74	0.017	Customer-27	1	0.031	50	0.011
Poland	11	0.346	472	0.107	Customer-28	21	0.661	2808	0.638
Portugal	134	4.220	10,725	2.435	Customer-29	7	0.220	460	0.104
Romania	9	0.283	436	0.099	Customer-30	12	0.378	2600	0.590
Russia	199	6.268	17,852	4.053	Customer-32	1	0.031	100	0.023
Chile	60	1.890	10,171	2.309	Customer-33	7	0.220	735	0.167
Thailand	12	0.378	676	0.153	Customer-34	5	0.157	650	0.148
Turkey	131	4.126	13,112	2.977	Customer-35	3	0.094	130	0.030
Ukraine	32	1.008	2077	0.472	Customer-37	2	0.063	100	0.023
Greece	31	0.976	1843	0.418	Customer-40	4	0.126	161	0.037
Production line									
Line-101	144	4.535	13,312	4.223	Customer-42	3	0.094	390	0.089
Line-106	106	3.339	28,798	3.023	Customer-48	3	0.094	432	0.098
Line-107	165	5.197	17,140	6.539	Customer-54	1	0.031	50	0.011
Line-108	182	5.732	13,143	3.892	Customer-56	1	0.031	84	0.019
Line-115	104	3.276	54,865	2.984	Customer-57	1	0.031	30	0.007
Line-301	364	11.465	12,998	12.458	Customer-59	1	0.031	206	0.047
Line-303	186	5.858	4492	2.951	Customer-65	1	0.031	216	0.049
Line-305	37	1.165	45,506	1.020					
Line-306	382	12.031	111,402	10.333					
Line-307	545	17.165	81,587	25.295					
Line-308	615	19.370	38,567	18.525					
Line-315	345	10.866	18,600	8.757					

Table 2 (continued)

Total # of orders: 3175
Total Production volume: 440,410

in Line 307. Lines 301, 306 and 315 follow these lines in terms of number of orders and production volume. On the other hand, Line 305 has the lowest frequencies in terms of number of orders and production volume.

Box plots are given to summarize descriptive statistics on production volumes and sample quantities of orders produced in 12 production lines respectively in Fig. 3.

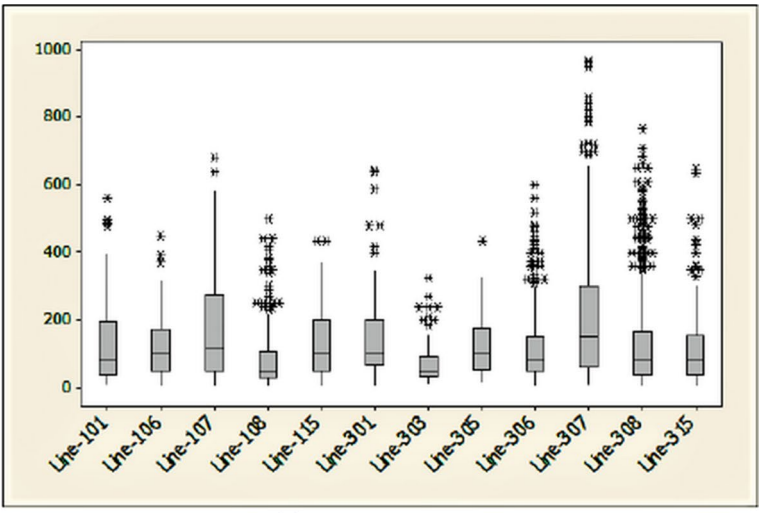
Figure 3 shows that box plots based on production volumes (Fig. 3a) and sample quantities (Fig. 3b) have very similar characteristics for each production line. This is due to the acceptance sampling technique where one item for each produced 50 items is checked. Besides, many outliers exist in the data set showing that high variability exists in order sizes. Box plots of many of the production lines are right skewed, meaning that a higher portion of order lines have relatively higher production volumes. Line 307, referred to as the line having highest total production volume in Table 2, has higher descriptive statistics (such as min., median, max.) as expected. For line 308, although the number of produced orders is high (Table 2), yet the volumes of production in these orders are observed as low when descriptive statistics of the related box plot are checked. Another observation worthy of note is related to Line 107. Although, it was not perceived to be a busy line in terms of number and volumes of produced orders, the descriptive statistics of production volumes in this line were relatively higher.

For each order, the data set includes some additional quality-based features such as defect code, defect type, defect score, number of items which are re-worked and cause of defect. In this study, defect type and defect score are merged in a variable labeled as product quality level; for details, please refer to Sect. 5.3.

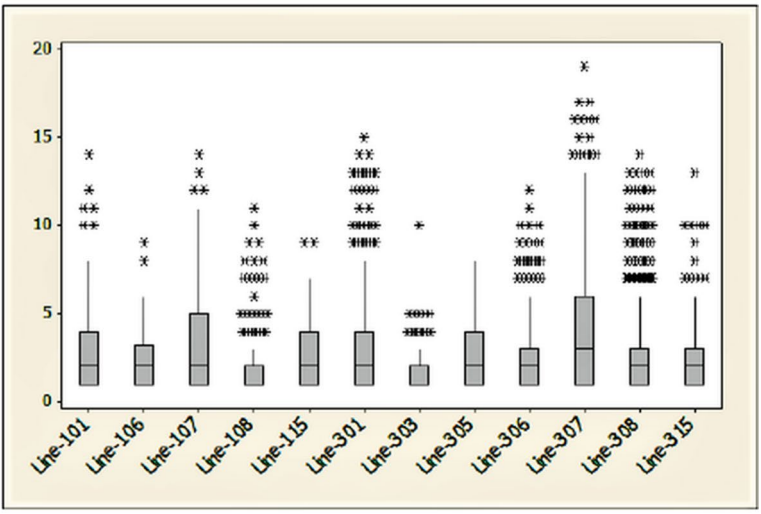
5.2.2 Preprocessing

Data preprocessing is initially performed by dropping missing values in the data set with the use of the *dropna ()* function of *pandas* module in Python. After rows with missing values are eliminated, attribute types which are represented with numerical values but categorical in nature are focused for conversion. Categorical conversion of the attributes is implemented with the *Categorical* class initializer of *"pandas"* module in Python. *Categorical* class is used to encode numerical values as categorised by providing capability of initializing corresponding attributes with categorical values. Product Id, Customer Id, Production Line, Defect Code, Product Quality Level and Root Cause are represented with numerical values in the raw data. However, they are categorical in nature and are therefore encoded as categorical attributes in preprocessing. Further preprocessing is performed on corresponding attributes of Stage 3: ARM is used for root cause identification (Production Line, Product Quality Level, Root Cause) by the transaction encoding method using *preprocessing* package of *mlxtend* module in Python. The proper input form of the Apriori algorithm is used as the transaction encoded version of input attributes. Transaction encoding performs the conversion of transactional data into list forms which are represented in binary. Table 3 illustrates a sample of original data and the transaction encoded version.

After preprocessing, the structured data set is formed. Attributes of the structured data set are summarised in Table 4.



(a) production volume



(b) sample quantity

Fig. 3 Box plots on production volumes and sample quantities in lines

Next, a step-by-step real life execution of the 3 stage model process is shown in Fig. 4.

5.3 Stage 1—MLP for product quality level classification

5.3.1 Attribute characteristics

In order to represent quality level, defect type and defect score, features of the data set are used and combined. In total, 128 different defect codes exist; they are mainly categorised under four different defect types where each type is assigned a defect score of 1, 5 or 10

Table 3 Transaction encoding (an instance data set)

Random slice of data			Transaction encoded version									
Production line (PL)	Product quality level (PQL)	Root cause (R–C) Id	PL 301	PL 303	PL 305	PQL 1	PQL 2	PQL 3	R–C 9	R–C 11	R–C 12	R–C 15
305	3	9	False	False	True	False	False	True	True	False	False	False
301	3	11	True	False	False	False	False	True	False	True	False	False
305	1	12	False	False	True	True	False	False	False	False	True	False
301	3	9	True	False	False	False	False	True	True	False	False	False
303	2	12	False	True	False	False	True	False	False	False	True	False
301	3	15	True	False	False	False	False	True	False	False	False	True
303	3	12	False	True	False	False	False	True	False	False	True	False

Table 4 Description of attributes

Attribute label	Attribute name	Attribute type	Description	Number of categories	Used in
PId	Product Id	Categorical	Represents different product types	385	Stage 1 Stage 2
Cid	Customer Id	Categorical	Represents different customers	36	Stage 1
Cnty	Country	Categorical	Represents countries which products are sent	29	Stage 1
PV	Production volume	integer	Represents lot size (number of items produced)	na	Stage 1 Stage 2
PL	Production Line	categorical	Represents different lines which products can be produced	12	Stage 1 Stage 2 Stage 3
SQ	Sample Quantity	integer	Represents number of checked items	na	Stage 1 Stage 2
DC	Defect Code	categorical	Represents different types of defects which can be seen in production	128	Stage 1
PQL	Product quality level	categorical	Represents quality level of the product	5	Stage 1 Stage 2 Stage 3
RQ	Rework quantity	integer	Represents number of reworked items when defect is detected	na	Stage 2
R–C	Root cause	categorical	Represents different causes which may lead defects in production	21	Stage 3

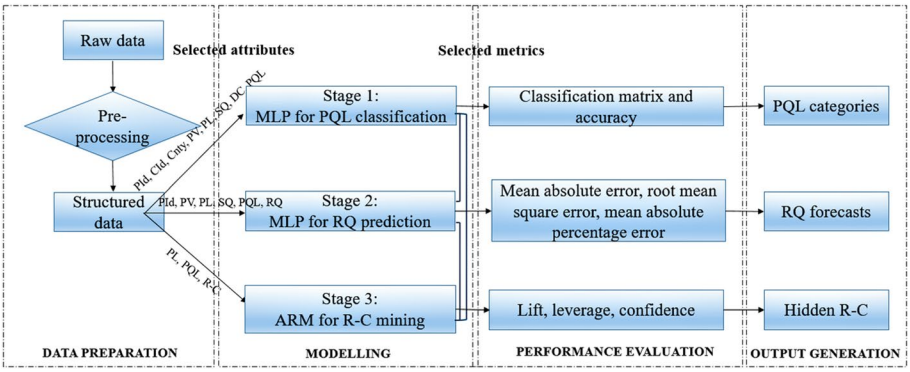


Fig. 4 Execution of 3 stage model at case company

in this data set. The majority of orders are produced without any defect. In this sense, the output variable of this model is defined as a product quality level in the following five categories:

- PQL-1: defect free orders
- PQL-2: orders with minor defects having a score of 1 out of 10
- PQL-3: orders with moderate defects having a score of 5 out of 10
- PQL-4: orders with frequent major defects having a score of 10 out of 10
- PQL-5: orders with rare major defects having a score of 10 out of 10

It was observed that 2790 (87.874%) of the products produced were defect free orders. Frequencies (percentages) of PQL-2, PQL-3, PQL-4 and PQL-5 type orders were 122 (3.843%), 185 (5.827%), 64 (2.016%) and 14 (0.440%) respectively. The defect types PQL-4 and PQL-5 are major defects (frequent and rare) represented with their defect scores and frequency of occurrence.

The input variables of this model were identified as product id, customer id, country, production volume, production line, sample quantity and defect code. For categorical input, variables having a limited number of categories and quality level distributions are presented in Fig. 5.

In Fig. 5, it can be seen that distributions of product quality levels significantly differed between levels of country, customer and production line. From Fig. 5a, it is noted that for some countries with low numbers and volumes of orders, such as Arab Emirates, Equator, Middle East and Romania, all produced orders were defect free. In Spain, which represents more than 60% of the production capacity, defect free orders were found to be 92%, with moderate defects the most frequently occurring defect types. For some other countries such as Azov Republic, Australia, Germany, Algeria and Denmark, the percentage of defect free orders was very low. In countries like Denmark, Hungary and Germany, moderate defects were frequently seen, while in Australia, Israel and Mexico, frequent major defects were widely seen. Based on customer type (Fig. 5b), the percentage of defect free orders in the two largest customer groups (Customer-1 and Customer-2) were found to be 93% and 89%; here, minor and moderate defects were seen more frequently compared to major defects. While the unique orders of customers having id of 27, 32, 54, 56, 57 and 65 were defect free, the unique order of Customer-59 has minor defects. Orders of customers with ids 29,

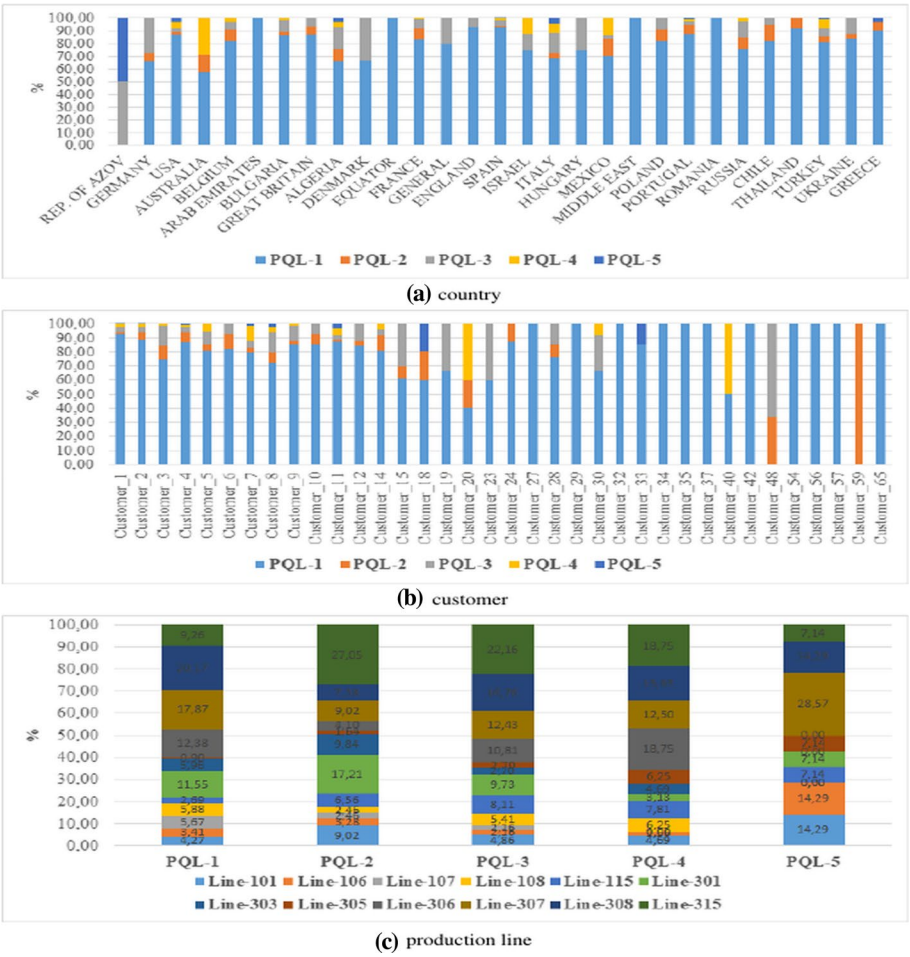


Fig. 5 Product quality level distributions

34, 35, 37 and 42 were all defect free. In addition, orders of Customer-48, Customer-23, Customer-19 and Customer-45 were found to have moderate defects. Frequent major defects were observed in the orders of Customer-40 and Customer-20. Based on production line (Fig. 5c), the majority of defect free products were produced in Lines 307, 308, 306 and 301 where higher numbers and volumes of orders were produced compared to other lines. Another important observation is that both the produced number and the volumes of orders were high in lines 301 and 315; yet, minor defects were frequently seen in these lines. In line 315, moderate and frequent major defects were also noticeable. Moderate and frequent major defects were also prevalent in Lines 306, 307 and 308. Finally, rare major defects were frequently observed in Line 307.

The correlation matrix between all attributes of stage 1 of the proposed model are presented in Fig. 6.

Figure 6 shows that the model output variable, PQL, has at least some degree of correlation with model attributes. It is clearly seen that the model output is significantly associated

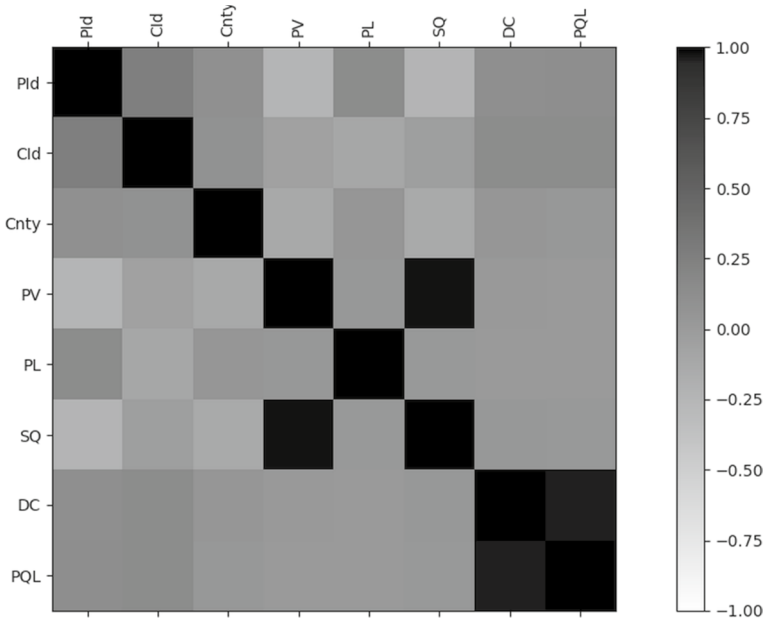


Fig. 6 Correlation matrix between attributes of stage 1 of model based on Spearman coefficient

with the model input of error code. There is some evidence of multi-collinearity, with cases of high correlation between some of the input attributes (such as PV and SQ). It should be noted that multi-collinearity is expected and as the data set has several dimensional characteristics of quality level, those may be related to each other.

5.3.2 Parameter setting

For classifying OQL, the MLP algorithm is implemented with *MLPClassifier* in the *neural network* package of *sklearn* module in Python. Fundamental parameters of MLPClassifier are those customized for the experiment—learning rate, momentum, number of hidden layers, solver function and activation function. Learning rate is set to be constant with an initial value of 0.01. Momentum is taken as 0.5 after a series of parameter tuning experiments to avoid the problem of unstable network. Number of hidden layers is set as 100 to extend the number of input feature combinations. Solver function is chosen as "adam"; this refers to the gradient-based optimizer solution (Kingma & Ba, 2014). Activation function is applied as *"relu"*; this stands for rectified linear unit function (f(x) = max (0, x)). The experiment is performed with a train/test split value of 0.33 and split is applied randomly.

5.3.3 Results

Results of MLP for PQL classification are presented in Table 5.

From Table 5, it was observed that MLP classified all defect free products correctly in the test data set; classification performance for PQL-1 level was therefore perfect. The model also provides very high performance for classifying orders with minor defects.

Table 5 MLP for PQL classification table

Actual class	Predicted class					Actual total
	PQL-1	PQL-2	PQL-3	PQL-4	PQL-5	
PQL-1	910	0	0	0	0	910
PQL-2	0	37	2	0	0	39
PQL-3	0	23	51	0	0	74
PQL-4	2	0	4	14	2	22
PQL-5	3	0	0	0	0	3
Predicted total	915	60	57	14	2	1048
Correct (%)	100	94.872	68.919	63.636	0	

Total number of correctly classified instances = 1012

Accuracy: 96.565%

However, although model performances for moderate defects and frequent major defects were acceptable, this still requires some improvement. Finally, the model was not able to classify rare major defects, as this contains really low frequencies; hence, the model has limited capability to improve its learning for this defect level.

5.4 Stage 2—MLP for rework quantity prediction

5.4.1 Attribute characteristics

This stage aims to predict the number of re-works based on considered model parameters, product id, production volume and line, sample quantity and product quality level. Since defect free products do not require re-working, these are removed from analysis. Thus, the product quality level has four categories at this stage, representing four types of defects as mentioned earlier.

For each production line, frequency and percentage distributions of number of orders requiring re-work as well as the total number of items which are re-worked are shown in Table 6.

Recalling the total number of orders presented in Table 2, Table 6 shows that 5.480% of the total order was re-worked by the company in 2019. However, the ratio of re-worked items to total production volume was lower, giving a percentage of 2.107%. When the re-work numbers based on production line are analysed, it is observed that lines 307, 308 and 306 have high numbers of orders and volumes, and therefore their defects and corresponding re-work rates are high. However, in line 315, in which minor defects were seen, the re-work rates are also observed as high. On the other hand, lines 101 and 107, in which the produced number of orders and production volume were lower, are observed as categories requiring relatively lower re-works.

The correlation matrix between the considered attributes of re-work quantity prediction stage is presented in Fig. 7.

From the correlation matrix, it can be seen that the output variable of this stage and re-work quantity are related with selected inputs. Additionally, the strength of the relation is strongest between product quality levels and re-work quantity, as they are positively co-related. This shows that product quality, or in other words type of defect, significantly

Table 6 Distribution of number of orders and items requiring reworks based on line

	# of rework required orders		# of reworked items	
	Freq	%	Freq	%
Line-101	2	1.149	94	1.013
Line-106	4	2.299	265	2.856
Line-107	1	0.575	62	0.668
Line-108	3	1.724	86	0.927
Line-115	11	6.322	547	5.894
Line-301	17	9.770	1472	15.862
Line-303	5	2.874	151	1.627
Line-305	10	5.747	451	4.860
Line-306	26	14.943	1192	12.845
Line-307	30	17.241	2090	22.522
Line-308	33	18.966	1349	14.537
Line-315	32	18.391	1521	16.390
Total	174		9280	

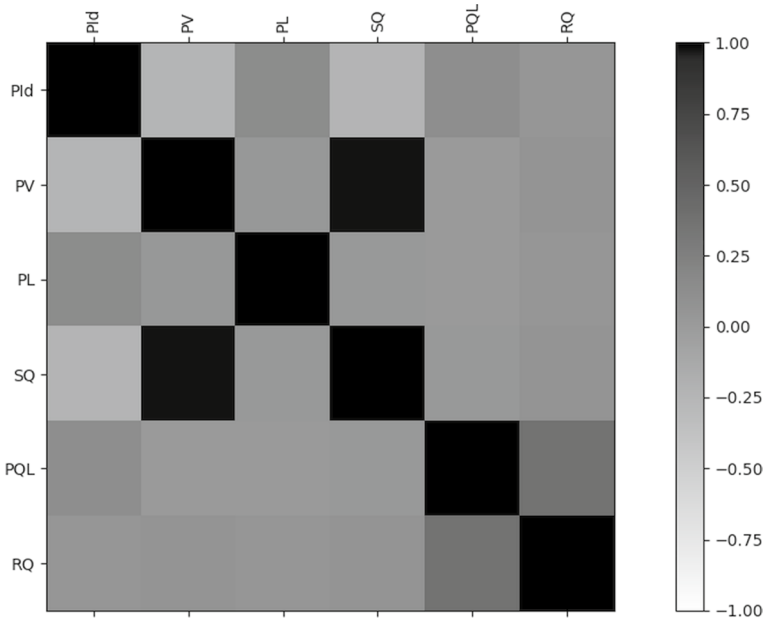


Fig. 7 Correlation matrix between attributes of stage 2 based on Spearman coefficient

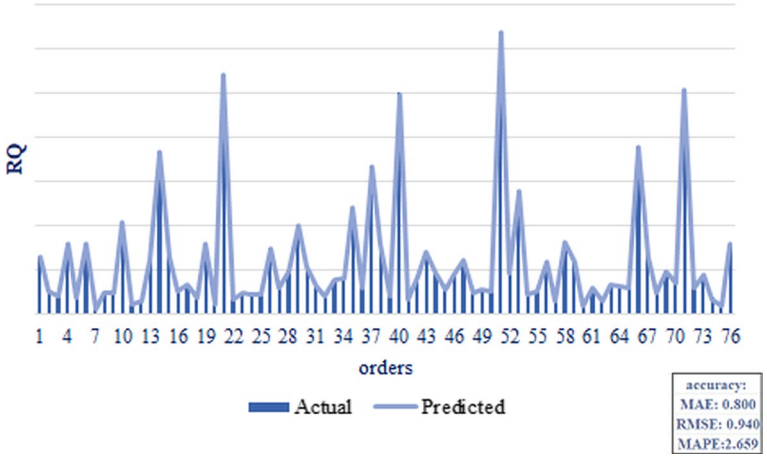


Fig. 8 MLP for RQ prediction results

affects re-work quantities. Whereas, for major defects, re-work quantities are higher if compared to minor or moderate defects.

5.4.2 Parameter setting

MLP is applied to obtain predicted values of re-work quantities. *MLPRegressor* in the neural network package of *sklearn* module in *Python* is initialized with the following customized parameters: learning rate is set to constant and initialized as 0.05, momentum is set as 0.6, number of hidden layers are taken as 100, solver function is *adam* (Kingma & Ba, 2014) and activation function is *relu*. Train/test split value of 0.33 is applied for the experiment, with split for this stage applied randomly.

5.4.3 Results

Figure 8 shows actual and predicted re-work quantity values in the test data set. Since this stage of the model presents a numeric prediction, the performance is summarised based on the metrics of mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE).

In Fig. 8, it is observed that predicted values of re-work quantity are very close to their actual values. Based on the MAPE value and given the model attributes, the MLP algorithm is able to predict required re-work quantities with around 98% of accuracy level.

5.5 Stage 3—ARM for root cause identification

5.5.1 Attribute characteristics

In the data set, recorded root causes were analysed and categorized under 16 categories; these are Research & Development (5), Print design (6), Glass process (7), Enamel process (8), Lot check (9), Material (10), Assembly (11), Planning (12), Press process (13),

Table 7 Mined rules for root causes

Rule #	Antecedents	Consequents	Confidence	Lift	Leverage
1	frozenset({108, 3})	frozenset({10})	1.000	6.367	0.009
2	frozenset({308, 4})	frozenset({11})	1.000	4.548	0.008
3	frozenset({4})	frozenset({11})	1.000	4.548	0.008
4	frozenset({115, 2})	frozenset({17})	0.800	7.640	0.018
5	frozenset({305, 2})	frozenset({15})	0.750	11.938	0.014
6	frozenset({108})	frozenset({10})	0.750	4.775	0.012
7	frozenset({308, 2})	frozenset({11})	0.667	3.032	0.014
8	frozenset({101, 2})	frozenset({17})	0.667	6.367	0.018
9	frozenset({106, 3})	frozenset({15})	0.667	6.367	0.009
10	frozenset({115})	frozenset({8})	0.636	6.077	0.031
11	frozenset({106})	frozenset({6})	0.500	4.775	0.008
12	frozenset({301, 1})	frozenset({20})	0.500	4.775	0.008
13	frozenset({108})	frozenset({3, 10})	0.500	5.026	0.500
14	frozenset({315, 2})	frozenset({13})	0.500	10.611	0.500
15	frozenset({306, 2})	frozenset({11})	0.444	2.021	0.444
16	frozenset({303})	frozenset({10})	0.400	2.547	0.006
17	frozenset({303})	frozenset({11})	0.400	1.819	0.005
18	frozenset({107, 2})	frozenset({6})	0.400	3.820	0.008
19	frozenset({307, 2})	frozenset({10})	0.333	2.122	0.006
20	frozenset({307, 2})	frozenset({11})	0.333	1.516	0.004

Procurement (14), Design (15), Supply (16), Technical problem (17), Product tree (18), Sub-industry (19) and Electrical problem (20).

Numbers in parenthesis show the assigned codes of recorded root causes. Among these listed 16 root causes, the most frequently observed causes in production were noted as 'assembly' and 'material'.

Production line and quality level are used as input attributes. In this stage, the first category of PQL is eliminated since it represents defect free items. ARM is applied to identify hidden root causes behind various production defects.

5.5.2 Parameter setting

Association rules are mined applying the Apriori algorithm in *frequent patterns* package of *mlxtend* module in Python. Initially, frequent item sets are generated with minimum support value of 0.02. After extracting the frequent items, ARM is performed with minimum confidence value of 0.3.

5.5.3 Results

Based on the confidence metric, the top 20 hidden causes identified are summarised in Table 7.

While interpreting the rules presented in Table 7, three digit numbers are used for production lines, numbers 1 to 4 are used for quality levels (OQL-2 to OQL-5) respectively

and numbers 5 to 20 are used for coding their root causes. Thus, rule # 1, shows that the root cause of frequent major defects observed in production line 108 was material related; rule #4 shows that the root cause of moderate defects observed in Line 115 was related to technical problem. Similarly, rules #16 and # 17 show that many of the problems observed in Line 303 were related to material and assembly. Further, according to rule #11, defects realised in Line 106 were caused by errors in print design process. Finally, for Lines 303 and 307, both material and assembly were mined as root causes of defects. While the root cause of defects for Lines 306 and 308 were identified as assembly, material related defects were more frequent in Line 108. In the majority of the top 20 rules, problems were related with assembly and material. However, in some rules, other primary root causes were discovered—technical problem, design, print design, enamel and press processes plus electrical problem.

6 Discussion

Due to the ever-increasing complexity of the global business environment and the advances in digital technologies, new quality management systems are required in this Industry 4.0 era. Companies are expected to be agile, flexible and possess dynamic capabilities in this competitive environment (Jacob, 2017). Advances in digital technologies are used in improving and managing various processes of businesses and make it possible for companies to be successful. In the big data era, to achieve success in this environment, MSMEs are required to have big data analytics capabilities to transform big data into actionable and valuable knowledge (Bumblauskas et al., 2017).

Quality management has been the most important concept among many of the other processes in which managers are willing to innovate. Thus, implementation of big data analytics techniques for quality management in MSMEs becomes more important. By the integration of big data analytics techniques, this study has proposed a 3 stage model to analyze three different problems in the quality management of MSMEs.

The first stage of the proposed model aims not only to automatically detect potential defects, but also to classify defects based on their categories once the defect is detected. In this stage, the target variables (order quality level) were categorized into five sections—defect free orders, orders with minor defects, orders with moderate defects, orders with frequent major defects and orders with rare major defects. A MLP neural network was then implemented to classify the produced orders based on their quality levels. The achieved 96% accuracy of this model for defect detection is seen as outperforming any previous studies in this context. By implementation of similar techniques, MLP, Essa et al. (2019) achieved 98.07%, Yapi et al. (2015) achieved 97% to 99%, & Carvajal Soto et al. (2019) achieved 98% accuracies for defect detection problems in different industries. These studies proposed a two-level classification model which could only classify items as defects or not; they were unable to provide any further details in defect types. Compared to these studies, the proposed model, although having slightly lower accuracy, can still be regarded as superior since it achieves this performance to include a 5-level classification solution. In addition, the proposed model has already outperformed the previous studies of Law et al. (2017) and Peres et al. (2019) who recorded respective accuracies of 94% and 93% for only two-level classification models.

The second stage of the proposed model aims to generate efficient predictions on required levels of re-working once a defect is detected. The concept of re-work is also very

important in quality management since orders requiring re-work and scrap are not only very costly but also damage the reputation of the company within its network; this can lead to customer losses. By identifying product id, production volume and line, sample quantity and product quality level as important features of the model and implementing the MLP network, the model achieved 98% accuracy for predicting required numbers of re-work for orders which were initially detected as defects. Although previous studies argued the criticality of re-work in quality management (Carletti et al., 2019; Chien et al., 2017), to the best of our knowledge there is a gap in literature for modelling required re-work levels using big data analytics. Thus, the weakness or superiority of the proposed model cannot be compared fairly and discussed within existing literature.

The third stage of the model aimed to identify root causes of the defects once a defect is detected in the first stage of the model. This is also a critical context in quality management for MSMEs. The elimination of the causes of defects provides immediate quality improvements; this problem has therefore been frequently analyzed in literature. With the purpose of identifying root causes of defects, Chien et al. (2017), Carletti et al. (2019), Dey and Stori (2005), Lokrantz et al. (2018) have all used MLP networks to identify various root causes of production defects which were specific to a considered industry. However, this study differs from previous pieces of research based on the implemented big data analytics technique in the root cause identification problem. Since root causes are hidden within big data sets including various features of the production system, in this study it was decided to use association rule mining to approach this problem. Association rule mining is a superior big data analytics technique for extracting relations and dependencies between input features in a data set based on transactional instances. By analyzing these transactions, ARM was able to discover hidden patterns or rules within the data set based on corresponding input features. When a defect was detected, by identifying the production line and four different categories of defect quality levels as input features, the Apriori algorithm, a widely used ARM algorithm, was implemented to discover hidden root causes in the data set. The majority of the top 20 rules were found to be related with assembly and material. Technical problems, design, enamel and press processes and electrical problems were also discovered to be some of the other causes of defects in the production system of this kitchenware company. Although these rules are industry and data specific, it was concluded that as an emerging big data analytics technique, the use of ARM is fit and proper for root cause identification.

7 Research implications

7.1 Theoretical implications

MSMEs are expected to be agile, flexible and equipped with dynamic capabilities in the ever-increasing competitive environment of Industry 4.0 (Jacob, 2017). Besides, since recent advances in technologies provide opportunities to collect, store and manage huge volumes of data, MSMEs are also expected to have big data analytics capabilities to transform dynamically changing raw data into actionable knowledge (Bumblauskas et al., 2017; Kiron et al., 2014). Grounded in dynamic capabilities theory and big data analytics capability, this study makes an attempt to contribute to these theories by proposing a multi-stage model for managing quality in the dynamically changing big data environment of MSMEs. Given a real-life data set, this study implements different big data analytics techniques to

tackle three critical problems in quality management. Thus, methodologically it differs from empirical investigations, review or bibliometric analysis or discussion papers. This model is also believed to widen application areas of these theories and to offer different perspective for researchers.

7.2 Managerial implications

For quality management in Industry 4.0, this paper proposes a 3 stage model based on emerging technologies and provides valuable advice to MSMEs on different aspects under investigation; these are detection of defect (as well as level of defect), prediction on required reworks and identification of root causes of defects.

The proposed model is capable of detecting defects automatically and as such, has many implications. Defects can have a damaging effect on companies' sales and reputation. Such impacts have significantly increased in the era of information technology. Additionally, if managers are able to detect any defect along with its type, this may further lead to significant time saving in the production system. Notably, different defect types require different processing and decision-making. For example, minor defects may require little extra effort, major defects require much more while plans on how to deal with and manage the detected defect can be made more rapidly. One other important implication of our model is that it allows MSME managers to assess system inputs for increasing defect frequencies in different defect types; this provides useful insights for managers into improving system quality. For instance, according to our findings, if minor defects are frequently seen in one production line X while major defects are detected more in line Y, a manager should focus on improving the line Y first. Since the proposed model had 96% accuracy for detecting defects as well as defect levels automatically, in line with this finding, adaptation of big data analytics into production systems for detecting defects efficiently is advised for the managers of MSMEs.

The proposed model for re-work quantity prediction is also useful for companies in planning not only the production processes, but also many of the other operations such as capacity, budget, supply of material, scheduling orders, human resources management, procurement etc. The ultimate goal for companies is to achieve zero defects in production; ths is challenging in a practical operation. Thus, preparing plans in advance by considering the possibility of defects and the required amount of re-works when defects are identified, can further improve operational performance of MSMEs. In line with the 98% accuracy achieved in predicting required re-work levels, once again, integration of big data analytics such as machine learning is highly recommended for MSME managers.

Identifying root causes of defects is also valuable in practice. Traditionally, root causes in production have generally been identified using expert knowledge at different stages of manufacturing. However, this approach has two basic problems: generated knowledge on root causes is not always transferred and shared between experts in different sites while some experts may be biased in their assessment (Lokrantz et al., 2018). Advances in ML and big data analytics leads to creating models that can discover hidden causal relations in production. Notably, the mined rules are purely data-driven, so they can present deeper insights for managers into identifying root causes. Further, mined rules do not include any bias, can be stored for future use and can be shared and generalized in different sites. Relying on data driven understanding of root causes in the system, MSMEs managers can make more accurate decisions in identifying and eliminating such causes to ensure highest levels of quality. Examination of the model's third

stage results show that generated rules were highly valuable, especially top rules with high confidence levels, meaning that they were very frequent within the data set. This is an additional reason to suggest to MSMEs' managers to implement ARM techniques for root cause identification.

8 Conclusion

Advances in digital technologies are used in improving and managing various processes of businesses, making it possible for MSMEs to be successful. Quality management has been the most remarkable concept among many of the other processes in which managers are willing to innovate. Thus, adaptation of these advances in technologies in managing quality levels becomes more important. For this reason, this study has proposed a model implementing big data analytics technologies for quality management in MSMEs. The proposed model not only enables detection of defects and identifying defect types, but also generates predictions for re-work quantities while additionally discovering root causes of the detected problems. The proposed model is tested and validated taking empirical data from a medium enterprise in Turkey serving in the kitchenware industry. Using main features of data set, product, customer, country, production line, production volume, sample quantity and defect code, MLP for product quality level classification achieved 96% of accuracy levels. Additionally, by considering the product, production line, production volume, sample quantity, product quality level, MLP for a re-work quantity prediction model was achieved with 98% performance. Finally, using production line and product quality level, ARM for root cause mining model was identified. The root cause identification was underpinned by mining their association rules, where top rule achieves a confidence of 80%.

Therefore, this research contributes to literature by proposing a three-stage model where each step provides solutions to three crucial concerns in quality management with the implementation of big data analytics technologies. In addition, in each step of the proposed model, either by approaching the problem from a different perspective or by implementation of a different technique compared to existing studies, this study also contributes to current literature.

Based on data collected from the case company, this study is somewhat limited in scope in utilizing real time data. In the context of large quantities and frequently changing products, it is important to use real time data in a manufacturing system. Since real systems are dynamic, so utilizing real time data has a great potential to enhance richness and accuracy of the developed model. It may also enable data visualization and image processing for higher manufacturing efficiency. With the increased dimensionality of data according to product and time based features, the proposed model can be extended in real time data in future research. With more product specific features such as raw material specifications and suppliers, future models can provide further analysis on different product types. Similarly, in future, smart approaches can also be extended in modelling and predicting time-based features such as lost time for re-works or lead times. By applying deep learning techniques, further implicit patterns might be discovered as future scope for research. Additionally, techniques for synthetic data generation might be applied to overcome limitations in data size.

References

Ahuett-Garza, H., & Kurfess, T. (2018). A brief discussion on the trends of habilitating technologies for Industry 4.0 and Smart manufacturing. *Manufacturing Letters, 15*, 60–63.

Akter, S., Michael, K., Uddin, M. R., McCarthy, G., & Rahman, M. (2020). Transforming business using digital innovations: The application of AI, blockchain, cloud and data analytics. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03620-w

Arunachalam, D., Kumar, N., & Kawalek, J. P. (2018). Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E: Logistics and Transportation Review, 114*, 416–436.

Bagodi, V., Venkatesh, S. T., & Sinha, D. (2020). A study of performance measures and quality management system in small and medium enterprises in India. *Benchmarking an International Journal*. https://doi.org/10.1108/BIJ-08-2020-0444

Belhadi, A., Zkik, K., Cherrafi, A., & Sha'ri, M. Y. (2019). Understanding big data analytics for manufacturing processes: Insights from literature review and multiple case studies. *Computers & Industrial Engineering, 137*, 106099.

Belhadi, A., Kamble, S. S., Zkik, K., Cherrafi, A., & Touriki, F. E. (2020). The integrated effect of Big Data Analytics, Lean Six Sigma and Green Manufacturing on the environmental performance of manufacturing companies: The case of North Africa. *Journal of Cleaner Production, 252*, 119903.

Bumblauskas, D., Nold, H., Bumblauskas, P., & Igou, A. (2017). Big data analytics: Transforming data to action. *Business Process Management Journal, 23*(3), 703–720.

Carletti, M., Masiero, C., Beghi, A., & Susto, G. A. (2019). Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In *2019 IEEE international conference on systems, man and cybernetics (SMC)* (pp. 21–26).

Carvajal Soto, J. A., Tavakolizadeh, F., & Gyulai, D. (2019). An online machine learning framework for early detection of product failures in an Industry 4.0 context. *International Journal of Computer Integrated Manufacturing, 32*(4–5), 452–465.

Chahal, A. (2015). The effectiveness of Total Quality Management in the manufacturing industries. *International Journal of Management, IT and Engineering, 5*(10), 210–225.

Chehbi-Gamoura, S., Derrouiche, R., Damand, D., & Barth, M. (2020). Insights from big Data Analytics in supply chain management: An all-inclusive literature review using the SCOR model. *Production Planning & Control, 31*(5), 355–382.

Chen, J. F., Do, Q. H., & Hsieh, H. N. (2015). Training artificial neural networks by a hybrid PSO-CS algorithm. *Algorithms, 8*(2), 292–308.

Chen, V. C. P., Kim, S. B., Oztekin, A., & Duraikannan, S. (2018). Preface: Data mining and analytics. *Annals of Operations Research, 263*, 1–3.

Chen, Y. T., Sun, E. W., & Lin, Y. B. (2019). Coherent quality management for big data systems: A dynamic approach for stochastic time consistency. *Annals of Operations Research, 277*(1), 3–32.

Chien, C. F., Liu, C. W., & Chuang, S. C. (2017). Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement. *International Journal of Production Research, 55*(17), 5095–5107.

Chin, K. S., Tummala, V. R., & Chan, K. M. (2002). Quality management practices based on seven core elements in Hong Kong manufacturing industries. *Technovation, 22*(4), 213–230.

Çiflikli, C., & Kahya-Özyirmidokuz, E. (2010). Implementing a data mining solution for enhancing carpet manufacturing productivity. *Knowledge-Based Systems, 23*(8), 783–788.

Davis, J., Edgar, T., Porter, J., Bernaden, J., & Sarli, M. (2012). Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Computers & Chemical Engineering, 47*, 145–156.

Dey, S., & Stori, J. A. (2005). A Bayesian network approach to root cause diagnosis of process variations. *International Journal of Machine Tools and Manufacture, 45*(1), 75–91.

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., et al. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics, 226*, 107599.

Essa, E., Hossain, M. S., Tolba, A. S., Raafat, H. M., Elmogy, S., & Muahmmad, G. (2019). Toward cognitive support for automated defect detection. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-018-03969-x

Fahmideh, M., & Beydoun, G. (2019). Big data analytics architecture design—An application in manufacturing systems. *Computers & Industrial Engineering, 128*, 948–963.

Ferrando, A., Popov, A., & Udell, G. F. (2017). Sovereign stress and SMEs' access to finance: Evidence from the ECB's SAFE survey. *Journal of Banking & Finance, 81*, 65–80.

Ferreiro, S., Sierra, B., Irigoien, I., & Gorritxategi, E. (2011). Data mining for quality control: Burr detection in the drilling process. *Computers & Industrial Engineering, 60*(4), 801–810.

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. *IEEE Access, 5*, 20590–20616.

Gu, V. C., Zhou, B., Cao, Q., & Adams, J. (2021). Exploring the relationship between supplier development, big data analytics capability, and firm performance. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-03976-7

Hazen, B. T., Skipper, J. B., Boone, C. A., & Hill, R. R. (2018). Back in business: Operations research in support of big data analytics for operations and supply chain management. *Annals of Operations Research, 270*, 201–211.

Hu, Y. C. (2014). Nonadditive similarity-based single-layer perceptron for multi-criteria collaborative filtering. *Neurocomputing, 129*, 306–314.

Ibrahim, Z., Abdullahb, F., & Ismailc, A. (2016). International business competence and small and medium enterprises. *Procedia-Social and Behavioral Sciences, 224*, 393–400.

International Monetary Fund. (2019). Financial inclusion of small and medium-sized enterprises in the Middle East and Central Asia. Departmental Paper No: 19/02

Jacob, D. (2017). *Quality 4.0 impact and strategy handbook: Getting digitally connected to transform quality management*. LNS Research.

Kamble, S. S., Gunasekaran, A., Ghadge, A., & Raut, R. (2020). A performance measurement system for industry 4.0 enabled smart manufacturing system in SMMEs—A review and empirical investigation. *International Journal of Production Economics, 229*, 107853.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International conference on learning representations* (pp. 1–15).

Kiron, D., Prentice, P. K., & Ferguson, R. B. (2014). The analytics mandate. *MIT Sloan Management Review, 55*(4), 1–25.

Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications, 67*, 84–94.

Lee, J., Kao, H. A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp, 16*(1), 3–8.

Lee, S. M., Lee, D., & Kim, Y. S. (2019). The quality management ecosystem for predictive maintenance in the Industry 4.0 era. *International Journal of Quality Innovation, 5*(1), 4.

Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016). Privacy-preserving-outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security, 11*(8), 1847–1861.

Liu, Y. (2014). Big data and predictive business analytics. *The Journal of Business Forecasting, 33*(4), 40.

Liu, P., & Yi, S. P. (2018). Investment decision-making and coordination of a three-stage supply chain considering Data Company in the Big Data era. *Annals of Operations Research, 270*(1), 255–271.

Lokrantz, A., Gustavsson, E., & Jirstrand, M. (2018). Root cause analysis of failures and quality deviations in manufacturing using machine learning. *Procedia Cirp, 72*, 1057–1062.

Mishra, D., Gunasekaran, A., Papadopoulos, T., & Childe, S. J. (2018). Big Data and supply chain management: A review and bibliometric analysis. *Annals of Operations Research, 270*(1), 313–336.

Peres, R. S., Barata, J., Leitao, P., & Garcia, G. (2019). Multistage quality control using machine learning in the automotive industry. *IEEE Access, 7*, 79908–79916.

Perzyk, M., Kochanski, A., Kozlowski, J., Soroczynski, A., & Biernacki, R. (2014). Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis. *Information Sciences, 259*, 380–392.

Savlovschi, L. I., & Robu, N. R. (2011). The role of SMEs in modern economy. *Economia, Seria Management, 14*(1), 277–281.

Soni, H. K., Sharma, S., & Jain, M. (2016). Frequent pattern generation algorithms for association rule mining: Strength and challenges. In *2016 International conference on electrical, electronics, and optimization techniques (ICEEOT)* (pp. 3744–3747).

Sun, Z., Sun, L., & Strang, K. (2018). Big data analytics services for enhancing business intelligence. *Journal of Computer Information Systems, 58*(2), 162–169.

Tsai, F. M., & Huang, L. J. (2017). Using artificial neural networks to predict container flows between the major ports of Asia. *International Journal of Production Research, 55*(17), 5001–5010.

Viet, N. Q., Behdani, B., & Bloemhof, J. (2020). Data-driven process redesign: Anticipatory shipping in agro-food supply chains. *International Journal of Production Research, 58*(5), 1302–1318.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research, 70*, 356–365.

Wamba, S. F., Gunasekaran, A., Dubey, R., & Ngai, E. W. (2018). Big data analytics in operations and supply chain management. *Annals of Operations Research, 270*(1), 1–4.

Wamba, S. F., Queiroz, M. M., Wu, L., & Sivarajah, U. (2020). Big data analytics-enabled sensing capability and organizational outcomes: Assessing the mediating effects of business analytics culture. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03812-4

Windmann, S., Maier, A., Niggemann, O., Frey, C., Bernardi, A., Gu, Y., Pfrommer, H., Steckel, T., Krüger, M., & Kraus, R. (2015). Big data analysis of manufacturing processes. In *Journal of physics: Conference series* (Vol. 659, No. 1, p. 012055). *IOP Publishing*.

Wulfsberg, J. P., Hintze, W., & Behrens, B. A. (Eds.). (2019). Machine learning and artificial intelligence in production: Application areas and publicly available data sets. In *Production at the leading edge of technology* (pp. 493–501). Springer Vieweg, Berlin, Heidelberg.

Yadav, N., Shankar, R., & Singh, S. P. (2020). Impact of Industry4. 0/ICTs, Lean Six Sigma and quality management systems on organisational performance. *The TQM Journal*. https://doi.org/10.1108/BIJ-08-2020-0444

Yadegaridehkordi, E., Hourmand, M., Nilashi, M., Shuib, L., Ahani, A., & Ibrahim, O. (2018). Influence of big data adoption on manufacturing companies' performance: An integrated DEMATEL-ANFIS approach. *Technological Forecasting and Social Change, 137*, 199–210.

Yapi, D., Mejri, M., Allili, M. S., & Baaziz, N. (2015). A learning-based approach for automatic defect detection in textile images. *IFAC-PapersOnLine, 48*(3), 2423–2428.

Zhang, C., Yu, J., & Wang, S. (2020). Fault detection and recognition of multivariate process based on feature learning of one-dimensional convolutional neural network and stacked denoised auto encoder. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2020.1733701

Authors and Affiliations

Gorkem Sariyer[1] · Sachin Kumar Mangla[2] · Yigit Kazancoglu[3] · Ceren Ocal Tasar[4] · Sunil Luthra[5]

Gorkem Sariyer
gorkem.ataman@yasar.edu.tr

Sachin Kumar Mangla
sachinmangl@gmail.com; sachin.kumar@plymouth.ac.uk

Yigit Kazancoglu
yigit.kazancoglu@yasar.edu.tr

Ceren Ocal Tasar
ceren.ocal@gmail.com

1 Department of Business, Yasar University, İzmir, Turkey

2 Operations Management, Jindal Global Business School, O P Jindal Global University, Haryana, India

3 Department of International Logistics Management, Yasar University, İzmir, Turkey

4 Independent Researcher, İzmir, Turkey

5 Department of Mechanical Engineering, Ch. Ranbir Singh State Institute of Engineering & Technology, Jhajjar, Haryana 124103, India