IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 3, September 2024, pp. 3052∼3060
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i3.pp3052-3060

Machine learning-based decision-making approach for predicting defects detection: a case study

Elena Barzizza, Nicoló Biasetton, Riccardo Ceccato, Alberto Molena
Department of Management Engineering, Universitá degli Studi di Padova, Vicenza, Italy

Article Info

Article history:

Received May 10, 2023
Revised Nov 10, 2023
Accepted Dec 3, 2023

Keywords:

Defect production
Fashion industry
Machine learning
Predictive quality
Quality control

ABSTRACT

In today's highly competitive global market, industries must produce faultless products to achieve profitability. Machine learning (ML) algorithms provide a possible method to improve quality standards by enabling the prediction of the outcome of quality control processes. This article presents a real case study based on ML algorithms suggested to develop a knowledge-based intelligent supervisory system to predict defect products in the fashion industry. Defect detection is formulated as a binary classification problem, and several ML algorithms have been compared to determine the most suitable one on the available data. The random forest (RF), LightGBM, and C5.0 algorithms exhibit comparable high-end performances on the pre-processed dataset made available by the company. Nevertheless, since the aim of the analysed industry is to reduce the rate of false negative observations (i.e., the proportion of defected-free products wrongly classified), the best method results is RF, as it minimizes this metric.

Corresponding Author:

Biasetton Nicoló
Department of Management Engineering, Universitá degli Studi di Padova
Stradella S. Nicola, 3 Vicenza 36100, Italy
Email: nicolo.biasetton@unipd.it

1. INTRODUCTION

Nowadays, quality represents a strategic factor in all industries that contributes to classifying companies as profitable and non-profitable. Quality can be considered as an index of internal performance, as it allows one to develop diversifying product or service differentiation policies. In recent years, the quality of products and services has even increased in importance due to the massive spread of timely written online reviews by customers at no cost who want to share their feelings and opinions after use or after experience. Therefore, the provision of defective or poor quality products can significantly affect both the reputation of the brand and the loyalty of customers to a certain brand; the aforementioned are some of the reasons that have led numerous companies to adopt a zero-defect policy.

The definition of a quality management process has changed significantly over the years [1]. The first quality control (QC) process was developed during the mass production period (1900-1940), and it simply consisted of the inspection of the final products. However, increasing pressure from the market led to a major theoretical shift in the first place toward the concept of process quality. The idea underlying this theory was that looking for errors was much less efficient than finding the source of the errors and removing them. In the early 60s, quality assurance theory emerged. The process quality, i.e. a posteriori reaction to QC of either products or processes, was not enough anymore, and industries wanted to prevent risks of failure products or services identifying their major causes before their appearance. With increasing complexity of products and interdependencies with the supply chain, ISO 9000 certification has been introduced to define quality

management standards to help industries ensure that they meet the basic requirements of the customer and other stakeholders related to products or services. In the last ten years, the quality management process shifted from being implemented due to market pressure to being implemented due to the general importance of delivering high-quality results.

Improving product and service quality is one of the main goals of the fourth industrial revolution [2], and the interconnection between automation, machine learning (ML) algorithms, and real-time data is the key to its completion. From a manufacturing perspective, the ability to efficiently collect and analyse huge amount of data using powerful ML algorithms can improve QC [3] revealing hidden patterns in the data and thus allowing for a broader analysis of the impact of all factors that contribute to manufacturing processes. Within the QC field, ML applications are highly diverse as they range from regular classification based on collected data to computer vision and image recognition. Similarly, there is no universal ML approach, algorithm, or method to solve all different QC problems, as the specific features of production systems have a substantial impact on the algorithms to be implemented [4]. The application of ML algorithms to predict or classify product quality has a relevant historical background. One of the earliest examples can be traced back to 1993, and concerns the prediction of the outer diameter of parts produced through injection moulding using backpropagation artificial neural networks [5].

In recent years, we also assisted to a growing literature on ML algorithms applications to classify or predict product quality thanks to the exponential increases in computer power, database technologies, optimisation methods, new efficient and robust ML algorithms, and the availability of big data. From the review of the literature on ML and deep learning (DL) for quality prediction in manufacturing [6], it emerges that more than 80 papers have been published on this topic between 2012 and 2021, with an exponential increase in the last few years. The success of ML algorithms can be attributed to their ability to handle high-dimensional, multivariate data, greatly reducing the need for human effort and improving the quality of the product [7]. These algorithms can provide suitable solutions for collecting fast and reliable information and understanding the implicit relationships existing within large data sets collected in complex and dynamic environments [8]. An example of an application of ML to optical inspections on finished products adopting a decision tree for the optimal classification of battery separator defects can be found in [9].

ML algorithms have also been adopted to predict the final quality of products in the early stages of the manufacturing process see [10]-[13] and to detect defects in mature organisations (i.e. organisations that merge different tools to improve the quality of manufacturing processes, such as lean production, standards conformity, six sigma, design for six sigma) that only generate a few defects per million opportunities [3]. Several ML algorithms have been trained and evaluated to predict dimensional defects in a real multistage automotive assembly line in [14] while [15] analysed a welding process using both the random forest (RF) and the C4.5 decision tree algorithm for classification [16] to correlate the sound of the arc with the quality of the weld. Schorr *et al.* [17] predicted the quality of drilled and reamed hydraulic valve bores, by estimating the diameter and concentricity of re-drilled reamers, starting from torque measurements. Lastly, one of the most intriguing research areas involves geospatial big data, which comprises vast amounts of data including information about geographic locations. The utilization of geospatial big data can enhance the quality and reliability of production processes by monitoring resources and, most importantly, preventing production anomalies to avoid the manufacturing of defective products. For a discussion on such topics, you can refer to [18]-[20].

This study presents the learning process and pattern recognition strategy for a knowledge-based intelligent supervisory system, in which the main goal is the prediction of defective products. Defect detection is formulated as a binary problem, i.e. OK for defect-free products and KO for defect products. To achieve the aforementioned aim, different supervised ML algorithms are compared, in order to identify the most suitable one. A real case study of a company working in the fashion industry is presented. The paper is structured as follows. Section 2 briefly describes the ML algorithms used to predict defective products. In section 3 the case study is presented, while in section 4 the results are presented and discussed. Section 5 concludes the article by highlighting the main contributions of this study.

2. METHOD

Classification is the assignment of an object defined by a set of features to one of several predetermined classes using a learning rule. Thus, it belongs to the macro area of supervised ML algorithms that aims to infer a function to describe labelled training data (e.g., data with classification class) to predict the output of categorical

variable [21]. As stated previously, the goal of this study is to predict the QC outcome, and since this output is encoded as a binary variable (OK/KO), the following classification ML algorithms have been implemented: logistic regression; multivariate adaptive regression splines (MARS); generalized additive model (GAM) using splines; linear discriminant analysis (LDA); quadratic discriminant analysis (QDA); neural network (nnet); RF (ranger); support vector machines (SVM); light gradient boosting machine (lightGBM); elastic Net regression (glmnet); and C5.0 decision tree algorithm.

The choice of these algorithms has been driven by the literature review and the experience of other real case studies gained by the authors. Given that the input variables can encompass both qualitative and quantitative features, it is imperative that the selected models exhibit the capability to process both types of variables. The following sections provide a theoretical and concise introduction to the different algorithms that have been used.

2.1. Logistic regression

Logistic regression is a statistical modelling technique used for classification when the response variable is binary. It is part of the generalised linear model category and applies a logistic function to the linear combination of predictor variables to estimate the probability of the binary outcome of the response variable [22]. The model parameters are estimated using maximum likelihood estimation. In logistic regression, the logit transformation is applied on the odds, which are the probability of success divided by the probability of failure. So, logistic regression can be represented using (1), where p is the probability of a binary event and x_1,x_2,...,x_k are the explanatory variables:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{1}$$

2.1.1. Multivariate adaptive regression splines

MARS is a flexible nonparametric statistical procedure [23]. It approximates the nonlinearity of the model by using separate regression slopes (splines) in distinct intervals of the independent variable space, and so the gradient of the regression line may change in different intervals, which are defined by the so-called knots. MARS uses a fast but intensive search procedure to determine the variables to use and the end points of the intervals for each variable. It also allows for searching for interactions between variables. The selection of the optimal MARS model is composed of two phases: first, it constructs a large number of basis functions (BFs), which are selected to overfit data; then, in the so called backward phase, the redundant BFs (the ones which made the least contributions) are deleted. It can be considered as a white-box model, as its structure is interpretable, meaning that the relationships between variables and the response can be understood through the model's node and spline structure.

2.1.2. Generalized additive model using splines

GAM using splines is a general additive model that uses splines functions to model nonlinear relationships [24]. Each predictor variable is modelled using a separate spline function, in which a spline is a piecewise polynomial function defined over two knots. These functions smoothly connect polynomial function between the various knots, allowing GAM to model non-linear relationships between predictor variables and the response variable without assuming any particular shape of the relationship. The parameters of this model are estimated using maximum likelihood estimation. The key difference between this algorithm and the aforementioned MARS algorithm is the method used to model the nonlinearity. Although MARS uses separate slopes at different intervals in the independent variable space, GAM uses smooth functions, such as cubic splines, to model the nonlinearity.

2.1.3. Linear discriminant analysis

LDA [25], is a statistical technique used to predict the class membership of an observation. It is generally used to classify observations between two classes, but it can also be extended to handle multiple classes. LDA assumes that all classes can be separated by linear boundaries, so it constructs multiple linear discriminant functions, each representing a hyperplane in the feature space, to differentiate classes. In case of two classes, as in the case study presented here, LDA draws a hyperplane and projects the data onto this hyperplane, in order to maximize the separation of the two categories. This hyperplane is created considering two criteria simultaneously: maximizing the distance between the means of two classes and minimizing the

variation within each category. It is also important to underline that LDA is based on the assumption that the covariance matrices are equal for all classes.

2.1.4. Quadratic discriminant analysis

QDA is a statistical technique used for classification, very similar to LDA, with the difference that QDA relaxes the hypothesis that all covariance matrices are equals; as a result this technique is capable of capturing more complex relationships between predictor variables and response variables. It works by estimating the parameters of a quadratic discriminant function for each class, which is a quadratic equation involving the predictor variables. When a new observation is presented, QDA computes its discriminant scores based on these quadratic functions and assigns it to the class with the highest score.

2.1.5. Neural network

Neural networks draw inspiration from the architecture and functioning of biological neural networks, and possess the capacity to model intricate relationships between inputs and outputs [26]. They consist of a series of interconnected nodes that are arranged in layers. This multitude of nodes is known as perceptrons or neurons, which are responsible for transmitting signals to other neurons located in the subsequent layer. The signal is transmitted when a threshold value is exceeded, so the neurons have to compute an activation function to send the output to the subsequent neurons. The training phase of these complex structures involves forward and backward propagation: once the network was fed with values in a forward direction, the backward step optimizes the weights, to enhance the accuracy of the model.

2.1.6. Random forest

RF [27], is a class of statistical algorithms that can be applied for both regression and classification. It consists of n decision trees that are typically grown on an individual training data set extracted from the original data set. To address the issue of correlated trees, a random sub-sample of m predictors, from the full set of p available predictors, is selected as split candidates at each split point of a tree. This technique helps to ensure that each tree is built using a slightly different set of predictors, thus increasing the diversity among the trees in the forest, allowing a greater reduction in the prediction variability.

2.1.7. Support vector machines

SVM is a supervised ML algorithm [28]. The fundamental principle of SVM involves identifying the appropriate hyperplane for classification using support vectors, which are the points of each class located at the margin. This method is effective for both linearly separable and linearly non-separable data: indeed, in the case of linearly separable data, SVM finds a hyperplane which separates the data with the largest margin, while for non-separable data it maps the input data into a high dimensional space with a kernel function to find the separating hyperplane.

2.1.8. Light gradient boosting machine

LightGBM is a recent ML algorithm [29]. It is built on the gradient boosting framework, which utilizes the combination of several weak models to create a robust model capable of capturing complex nonlinear relationships between features and target variables. LightGBM's innovative tree-building approach uses a split-point algorithm to efficiently select the optimal split point for each feature. This, along with its design for fast and efficient training and prediction, enables LightGBM to handle large data sets and excel in scenarios where training time is constrained.

2.1.9. Elastic net regression

Elastic net regression is a regularization regression model [30]. This model is a combination of lasso regression (L1) and ridge regression (L2) penalties to overcome the limitations of both methods, especially when data are affected by multicollinearity (two or more input variables are highly correlated). Similarly to lasso regression, this method can automatically perform variable selection and shrink coefficients. This regression model can predict both quantitative and binary output.

2.1.10. C5.0 Decision tree algorithm

C5.0 is an algorithm that aims to build a decision tree capable of classifying observations based on some input variables. This algorithm works by splitting the sample on the basis of the feature with the highest information gain; then, every subsample as specified by the first split is split recursively until the subsamples

become indivisible. At the lowest level, the splits are evaluated, and those that do not significantly contribute to the model are pruned until the desired level of accuracy is achieved [31].

2.1.11. Algorithms evaluation and comparison

The input variables will be used by the ML algorithms mentioned above that will return an estimate of the probability value of belonging to the class of defective products. Subsequently, the classification of observations into the OK/KO classes is obtained setting a threshold value: if the probability predicted by the algorithm is higher than the threshold, the observation will be classified as KO, otherwise it will be classified as OK. The classification success of the different models is measured using the receiver operating characteristics (ROC) curve, the area under the curve (AUC), and the following indices based on the confusion matrix: total accuracy, sensitivity, precision, recall, specificity, false negative rate, and false positive rate. It is also fundamental to underline that the tuning parameters of the ML algorithms have been optimized using a grid search to minimize the false negative rate (i.e. the proportion of defected products wrongly classified as defected-free ones). The reason for this strategy lies in the fact that, from an industry perspective, it is worse to classify a product as free from defects when it is not than to misclassify a defect-free product and reprocess it. Furthermore, for each ML algorithm, different threshold values were tested to find the one that minimized the false negative rate.

3. DESCRIPTION OF THE CASE STUDY

The case study we are going to present in this paper concerns the tumbling of a fashion accessory made of acetate, and more specifically, the unit of analysis we are going to investigate will be a batch of identical products. To produce this product, the manufacturing process is characterised by five consecutive steps. At the end of each step, an operator evaluates the final output and decides whether the batch of products can move on to the next step or must redo the step. Therefore, each batch of products can run each step of the manufacturing process multiple times until the operator is satisfied with the final result. Consequently, the amount of time necessary to complete a process varies according to several subjective factors. At the end of the manufacturing process, an operator visually inspects 10% of the batch of products and decides, according to a series of internal quality standard requirements, whether the batch must be classified as a defected or a defected-free one. Speaking about data, it is important to underline that the source of the data we analyzed in this case study is the manufacturing execution system (MES) of the company we worked with, that is a software designed to optimize the manufacturing process by tracking and documenting the entire production cycle. From the MES, we obtained 38,743 observations, representing different batches, and after a careful data cleaning phase, performed in compliance with the company and consisting in the elimination of the observations that had a null processing or waiting time, mostly due to some technical issues, we ended up with a final data set consisting of 28,009 observations. In Figure 1, we report the boxplot of the waiting times for the five steps that compose the productive process before the data cleaning phase: from this figure. It is possible to observe that there are some negative values.

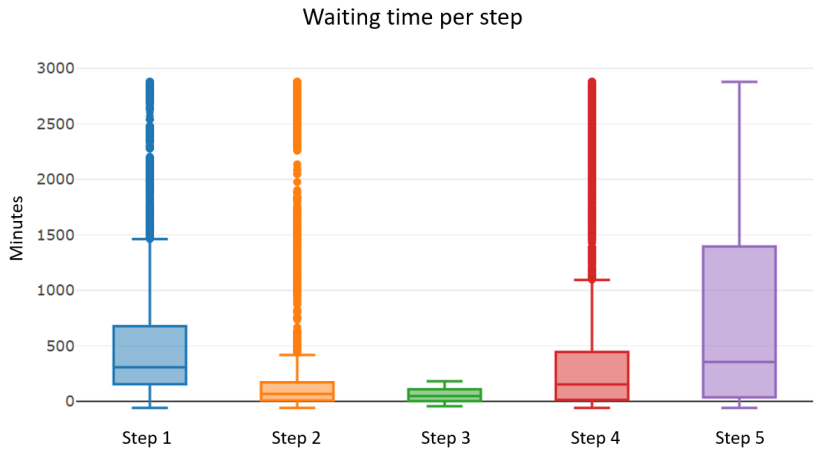


Figure 1. The waiting times for the five steps that constitute the production process

In this case study we used a total of 22 predictive variables, which can be divided into two main categories: product features and process characteristics. More in details, among the former, which are a total of 11 out of 22 predictive variables, we can identify features such as color, model, and some typical measurements of the product under study, such as the design of certain parts of the product itself. On the other hand, as far as variables related to the production process are concerned, we can include, for example, the processing time for each of the five steps already mentioned, the percentage of time spent in a certain machinery positioned in a certain area of the production chain and also the wear of the abrasive paste, being one of the most important variables in the process and thus in determining the outcome of QC. The values of mean, Q1, median, Q3 and standard deviation for the last mentioned variable are reported in Table 1. Clearly, the target variable, on the other hand, is represented by the outcome of QC downstream of each processing step.

Table 1. Mean, Q1, Median, Q3, and standard deviation values for the variable related to wear of the abrasive paste

Productive step	Mean	Q1	Median	Q3	StDev
Step 1	174.5607	83.40968	172.2402	259.9114	106.5447
Step 2	249.0394	119.9143	241.9183	369.3094	150.1812
Step 3	515.7192	238.9415	543.8706	773.4081	300.9195
Step 4	441.7456	243.1367	399.2906	600.9616	246.1685

In this case study we used R, and we went to test the different ML models described in section 2. Regarding the settings we used the common training and testing approach, where training data set are used to train a set of candidate algorithms trying different tuning parameters and testing data are used to evaluate the performances of the models on never-seen-before data. So, adopting this approach, we divided the observations and used 75%, so 21,007 observations in the training phase and 25%, so the remaining 7,002 observations in the testing phase. It is also important to underline that we have used the 10-fold cross validation approach, that is used to evaluate the generalisation ability of each candidate algorithm and select the best, according to a relevant model selection criterion, which in our case was the false negative rate.

Moreover, as already mentioned, we evaluated different threshold values, in order to find the one that minimizes the false negative rate. So, it is possible to affirm that a two-step approach was implemented: in the first step, various ML models were tested while keeping the threshold fixed—starting, for example, with a value of 0.5. After testing these ML models, we also explored four different thresholds values (0.5, 0.4, 0.3, 0.2) to find the ML model-threshold combination that minimized the false negative rate. It's crucial to emphasize one more time that minimizing this rate is highly important, indeed, for the company, a false negative (a product classified as OK when it's not) is much more it's much more problematic than a false positive (a product classified as KO when it's not). Lastly, an issues that arises is the imbalance in the data. Indeed, as one might expect, the number of batches not conforming to QC is much lower than those that are compliant. In our specific case, the total number of observations is 28,009, but only 21.2%, so 5,936 are classified as KO.

In such a situation, the classification might be influenced in favour of the most represented class, i.e. the OK ones [32]. Therefore, the ML algorithms described in the previous subsections have been implemented twice: on the original data set provided by the company and on a balanced data set extracted from the original one using the random undersampling technique that results on a data set made up of 11,545 observations, of which 5,936 KO and 5,518 OK. Lastly, we rated the different ML algorithms accordingly to their capacity of minimizing the false negative rate.

4. RESULTS AND DISCUSSION

From the performances we obtained by testing the ten ML models, and considering the false negative rate as a primary importance index, for the reasons previously discussed, we found that, regardless of the chosen threshold and the type of dataset used (balanced or unbalanced), three models always emerged as the best among the ten tested. These three models are: RF, LightGBM, and C5.0. As for the choice of the threshold, we tested four different levels: 0.5, 0.4, 0.3, and finally 0.2. Regardless of the model and the type of dataset used, the analysis revealed that the threshold minimizing the false negative rate was 0.2. So, we decided to set that threshold for a final comparison among the various ML models. We compared RF, LightGBM, and C5.0, using both datasets and fixing the threshold at 0.2. Table 2 shows the performance of various tested indices for these three models. The left side of the table represents the performances obtained in the unbalanced dataset,

while the right side represents the performances obtained on the balanced dataset. Looking at the table, it can be noted that RF always stands out as the best model since it minimizes the false negative rate in both cases.

Table 2. Performance of RF, LightGBM and C5.0, threshold equal to 0.2, unbalanced and balanced data

	Unbalanced data			Balanced data		
ML algorithm	RF	LightGBM	C5.0	RF	LightGBM	C5.0
AUC	0.7447	0.7209	0.7110	0.7294	0.6979	0.6875
Accuracy	0.6659	0.6662	0.6807	0.5377	0.5733	0.5584
Recall	0.7076	0.6490	0.6029	0.9929	0.9572	0.9608
Precision	0.3552	0.3465	0.3520	0.5288	0.5508	0.5420
Sensitivity	0.7076	0.6490	0.6029	0.9929	0.9572	0.9608
Specificity	0.6547	0.6709	0.7014	0.0480	0.1602	0.1259
FNR	**0.2924**	0.3510	0.3971	**0.0071**	0.0428	0.0392
FPR	0.3453	0.3291	0.2986	0.9520	0.8398	0.8741

5. CONCLUSION

In this study, conducted in collaboration with a company operating in the fashion industry, we compared various ML models and different settings to identify the one that allows a reliable prediction of QC outcomes. Specifically, we evaluated the eleven models presented in section 2 to identify the one that, based on some predictive variables related to product features or process characteristics, minimizes the false negative rate. Since the data was unbalanced, with approximately 80% of observations being OK (having passed QC), the study was conducted on both the original data set, consisting of 28,009 observations, and a balanced data set obtained through undersampling. In our analysis, we also experimented various threshold values to enhance performance on the false negative rate. Considering both data sets, our analysis revealed that the optimal threshold value to minimize the false negative rate is 0.2. Regarding ML models, our analysis showed that three models outperform the others: RF, LightGBM, and C5.0. Ultimately, the best ML-threshold combination was found to be RF with a threshold set at 0.2, holding true for both the unbalanced and balanced datasets. Through this analysis, we obtained a model that accurately predicts QC outcomes while minimizing the risk of false negatives observations, i.e. the number of products that are deemed defect-free but are defective. While the findings cannot be broadly generalized due to the case study nature, we can assert that, much like in previous cases, RF demonstrates satisfactory performance. Future research will be devoted to comparing our results with others in terms of concrete data for better research integrative value. The implementation of ML algorithms that are effective in predicting the outcome of QC based on product or process variables can help companies in all fields understand what to expect from the production process regarding product quality. This subject (and, therefore, this study) merely represents the initial stage of a possible prospective improvement journey that could aspire to establish production lines that culminate in zero defective products by understanding, for instance, whether production defects are associated with specific product or production line factors, through further analysis.

ACKNOWLEDGEMENTS

Authors would like to thank the leading multinational company in the fashion industry that funded this project and provided the data analysed in this project. In particular, the authors thank Barbieri Pietro, Bortolan Luca, Canale Matteo, Giada Marco, Longi Davide, Napol Fabio, Paganin Danny for their invaluable suggestions. This study was carried out within the MICS (Made in Italy—Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3—D.D. 1551.11-10-2022, PE00000004). This study was carried out within the MOST—Sustainable Mobility National Research Centre and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D.D. 1033 17 June 2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] A. Weckenmann, G. Akkasoglu, and T. Werner, "Quality management - history and trends," *TQM Journal*, vol. 27, no. 3, pp. 281–293, 2015, doi: 10.1108/TQM-11-2013-0125.

[2] R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman, "Intelligent manufacturing in the context of industry 4.0: A review," *Engineering*, vol. 3, no. 5, pp. 616–630, 2017, doi: 10.1016/J.ENG.2017.05.015.

[3] C. A. Escobar and R. M. -Menendez, "Machine learning techniques for quality control in high conformance manufacturing environment," *Advances in Mechanical Engineering*, vol. 10, no. 2, 2018, doi: 10.1177/1687814018755519.

[4] A. Frankó and P. Varga, "A survey on machine learning based smart maintenance and quality control solutions," *Infocommunications Journal*, vol. 13, no. 4, pp. 28–35, 2021, doi: 10.36244/ICJ.2021.4.4.

[5] A. E. Smith, "Predicting product quality with backpropagation: A thermoplastic injection moulding case study," *The International Journal of Advanced Manufacturing Technology*, vol. 8, no. 4, pp. 252–257, 1993, doi: 10.1007/BF01748635.

[6] H. Tercan and T. Meisen, "Machine learning and deep learning based predictive quality in manufacturing: a systematic review," *Journal of Intelligent Manufacturing*, vol. 33, no. 7, pp. 1879–1905, 2022, doi: 10.1007/s10845-022-01963-8.

[7] H. S. Deshmukh, V. S. Jadhav, S. P. Deshmukh, and K. K. Rane, "Towards applicability of machine learning in quality control–a review," *ECS Transactions*, vol. 107, no. 1, pp. 14729–14737, 2022, doi: 10.1149/10701.14729ecst.

[8] G. Köksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13448–13467, 2011, doi: 10.1016/j.eswa.2011.04.063.

[9] J. Huber et al., "Method for classification of battery separator defects using optical inspection," *Procedia CIRP*, vol. 57, pp. 585–590, 2016, doi: 10.1016/j.procir.2016.11.101.

[10] R. Wagner, B. Haefner, and G. Lanza, "Function-oriented quality control strategies for high precision products," *Procedia CIRP*, vol. 75, pp. 57–62, 2018, doi: 10.1016/j.procir.2018.04.069.

[11] F. C. Neto, T. M. Gerônimo, C. E. D. Cruz, P. R. Aguiar, and E. E. C. Bianchi, "Neural models for predicting hole diameters in drilling processes," *Procedia CIRP*, vol. 12, pp. 49–54, 2013, doi: 10.1016/j.procir.2013.09.010.

[12] P. G. Mathews and M. S. Shunmugam, "Neural-network approach for predicting hole quality in reaming," *International Journal of Machine Tools and Manufacture*, vol. 39, no. 5, pp. 723–730, 1999, doi: 10.1016/S0890-6955(98)00061-3.

[13] R. Wagner, J. Fischer, D. Gauder, B. Haefner, and G. Lanza, "Virtual in-line inspection for function verification in serial production by means of artificial intelligence," *Procedia CIRP*, vol. 92, pp. 63–68, 2020, doi: 10.1016/j.procir.2020.03.126.

[14] R. S. Peres, J. Barata, P. Leitao, and G. Garcia, "Multistage quality control using machine learning in the automotive industry," *IEEE Access*, vol. 7, pp. 79908–79916, 2019, doi: 10.1109/ACCESS.2019.2923405.

[15] A. Sumesh, K. Rameshkumar, K. Mohandas, and R. S. Babu, "Use of machine learning algorithms for weld quality monitoring using acoustic signature," *Procedia Computer Science*, vol. 50, pp. 316–322, 2015, doi: 10.1016/j.procs.2015.04.042.

[16] J. R. Quinlan, *C4. 5: programs for machine learning*. San Mateo, California: Morgan Kaufmann Publishers, 2014.

[17] S. Schorr, M. Möller, J. Heib, and D. Bähre, "Quality prediction of drilled and reamed bores based on torque measurements and the machine learning method of random forest," *Procedia Manufacturing*, vol. 48, pp. 894–901, 2020, doi: 10.1016/j.promfg.2020.05.127.

[18] E. Clayton, "Predictive modeling and visual perception algorithms, ambient sound recognition and processing tools, and deep neural network and vision sensing technologies in the decentralized and interconnected metaverse," *Linguistic and Philosophical Investigations*, vol. 22, pp. 145–161, 2023, doi: 10.22381/lpi2220239.

[19] G. Lăzăroiu, M. Andronie, M. Iatagan, M. Geamănu, R. Ștefănescu, and I. Dijmărescu, "Deep learning-assisted smart process planning, robotic wireless sensor networks, and geospatial big data management algorithms in the internet of manufacturing things," *ISPRS International Journal of Geo-Information*, vol. 11, no. 5, 2022, doi: 10.3390/ijgi11050277.

[20] G. Duncan, "Motion planning and remote sensing algorithms, predictive geospatial modeling and deep learning artificial intelligence tools, and machine perception and image recognition technologies in the blockchain-based virtual economy," *Analysis and Metaphysics*, vol. 21, pp. 193–209, 2022, doi: 10.22381/AM21202212.

[21] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. Cambridge, Massachusetts: The MIT Press, 2018.

[22] J. A. Bunge and D. H. Judson, "Data mining," in *Encyclopedia of Social Measurement*, Elsevier, 2005, pp. 617–624 , doi: 10.1016/B0-12-369398-5/00159-6.

[23] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, 2007, doi: 10.1214/aos/1176347963.

[24] S. N. Wood, *Generalized additive models: an introduction with R*. Boca Raton, Florida: Chapman and Hall/CRC, 2017 , doi: 10.1201/9781315370279.

[25] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[26] C. M. Bishop, *Neural networks for pattern recognition*. Oxford, United Kingdom: Clarendon Press, 1995.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[28] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152 , doi: 10.1145/130385.130401.

[29] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3147–3155, 2017.

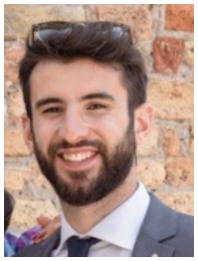
[30] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[31] S. Rajeswari and K. Suthendran, "C5.0: advanced decision tree (ADT) classification model for agricultural data analysis on cloud," *Computers and Electronics in Agriculture*, vol. 156, pp. 530–539, 2019, doi: 10.1016/j.compag.2018.12.013.

[32] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002, doi: 10.3233/ida-2002-6504.

BIOGRAPHIES OF AUTHORS

Elena Barzizza has a Master's degree in Management engineering at the University of Padua, Department of Technique and Management of Industrial systems. She is research and teaching assistant, and second year Ph.D. student in Management Engineering at the University of Padua. Her research interests include big data analysis, machine learning, and non parametric statistics. She can be contacted at email: elena.barzizza@phd.unipd.it.



Nicolò Biasetton holds a Master's degree in Management engineering at the University of Padua, Department of Technique and Management of Industrial systems. He is research and teaching assistant, and final year Ph.D. student in Management Engineering. His research interests include supervised and unsupervised machine learning algorithms, sentiment and emotional analysis. He can be contacted at email: nicolo.biasetton@phd.unipd.it.

Riccardo Ceccato holds a Ph.D. in Management Engineering at the University of Padua, Department of Technique and Management of Industrial systems. He is a lecturer in Machine Learning and Big Data, a member of the editorial board of the Journal of Applied Statistics and a reviewer for several international scientific journals. His research interests include machine learning algorithms and multivariate nonparametric techniques. He can be contacted at email: riccardo.ceccato.1@unipd.it.

Alberto Molena holds a Master's degree in Management Engineering at the University of Padua, Department of Technique and Management of Industrial systems. He is research and teaching assistant, and first year Ph.D. student in Management Engineering. His research interests include industrial statistics and design of experiment. He can be contacted at email: alberto.molena.1@phd.unipd.it.