IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 2, 2019, accepted January 15, 2020, date of publication January 27, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969816

Predicting Enamel Layer Defects in an Automotive Paint Shop

JAVIER SALCEDO-HERNÁNDEZ[1], JON GARCÍA-BARRUETABEÑA[1], IKER PASTOR-LÓPEZ[2], AND BORJA SANZ-URQUIJO[2]

[1]Department of Applied Mechanics, Faculty of Engineering, University of Deusto, 48014 Bilbao, Spain
[2]Department of Computer, Electronic and Communication Technologies, Faculty of Engineering, University of Deusto, 48014 Bilbao, Spain

Corresponding author: Javier Salcedo-Hernández (javiersalcedo@deusto.es)

ABSTRACT The appearance of the painted surface of the vehicle is key in the quality that the automotive customer perceives. The assurance of this quality starts in the automotive paint shop and compromises the effectiveness of the painting process as every paint defect is reworked. This entails material and labour costs, reducing the efficiency of the process and affecting the competitiveness of the product. To improve the efficiency while guaranteeing the quality, predictive control rather than corrective must be implemented. In order to achieve this control, a predictive model of quality is needed. As a first step to generate said model, this article demonstrates the correlation between the variables of the enamel coating process and the quality of the paint film of the vehicle. As there are no available application examples in the industry, a procedure is proposed in which the necessary steps for the creation of an industrial data set and a predictive model of quality are defined. The procedure is tested in an automotive paint shop. As a result, relevant variables for the quality assurance are identified and the correlation between process variables and the resulting quality is verified, concluding that the implementation of predictive control in the process is feasible.

INDEX TERMS Predictive modelling, automotive industry, manufacturing digitization, industrial data set.

I. INTRODUCTION

The automotive manufacturing process is divided into three main tasks. Firstly, in the body shop, the metallic structure of the vehicle is created. Secondly, in the paint shop, a corrosion prevention layer, a coloured layer and a bright protective layer are applied to the metallic surface. Finally, in the final assembly, the powertrain components, drivetrain elements and driver comfort equipment are installed to obtain the finished product. The processes carried out in the paint shop, are the most delicate steps in the vehicle manufacturing. A paint shop is a manufacturing bottleneck in many plants due to the complexity of the car body painting process, tight production management labours and rigorous quality requirements. As defined above, the paint film is composed of a superposition of layers that are applied to the metallic surface throughout the painting process. In order to achieve a solid paint film, a series of world wide standardized activities are performed consecutively. As presented in [1], historically, these activities have been: washing the metal surface to remove dirt and metal remains from the body shop, phosphating as metal surface pretreatment, cathodic electrodeposition of an anticorrosion layer, sealing and underbody protection, application of a preparation layer or primer surfacer, application of coloured enamel, application of a protective varnish layer and waxing. Along the process, quality controls are performed, normally to random car bodies, so the process is therefore controlled and adjustments are made because, for example, each layer thickness variations or repetitive appearance of defects. An exhaustive control of the painted body is also carried out after applying the varnish so imperfections in the paint film layer can be corrected ensuring compliance with the paint shop quality criteria. These reworks are often required and can significantly increase the manufacturing costs. To help evolve the vehicle manufacturing process, the automotive industry has always been following the latest evolutions of technology such as manufacturing software [2], manufacturing equipment [3], materials [4], [5], tools [6] or automated systems [7]. In the paint shop, the development of coating machines, robots and automation makes craftsmanship no longer needed. This early process digitisation makes automotive industry inclined to adopt the proposals made by new technological paradigms as European Industry 4.0 [8] in order to increase process efficiency while maintaining the quality of the product.

The associate editor coordinating the review of this manuscript and approving it for publication was Yucong Duan.

The quality of the product of a paint shop is related to corrosion protection and, the final appearance and the long lasting ability of its paint film layer. Related to the final appearance, many different quality issues can be detected [9]. Among these, inclusions in the paint film (lifts in the paint film caused by the presence of a underlying contaminating element such as dirt or dust) and imperfections in the shape of craters (usually, as round depressions in the paint film) are the most common. These imperfections are due to the combination of multiple reasons, changes in air flows that allow to approach impurities contained in the environment to the body, changes in temperatures and humidity that change the speed of evaporation of solvents, wear of paint application elements and others [9]. Although the process is profusely controlled and known, analysing the origin of each defect is a difficult task to perform because, on the one hand, they are usually the result of combinations of different factors even though process parameters are within the operating margins and, on the other hand, the defect cause can be only identified using destructive methods that can not be applied to every sample.

Historically, to detect and correct these defects, inspection zones have been implemented at the end of each relevant paint layer application. Here, specialized personnel carefully examine the surface of the body. The weak point of this method is the consistency of detection since human vision differs between individuals and, in addition, fatigues. Due to the advancement of artificial vision technology, these defect detection activities have been automated [10] in such a way that they allow for constant evaluation criteria.

In this research, the quality criteria has been defined by using the data provided by an automatic fault detection process that has been applied to the entire production and in which the sensitivity has been constant throughout the study. This allows (applying the precepts proposed by Industry 4.0 [11] about analysis of large amounts of data) to venture into the development of solutions, such as predictive models, that allow improving the efficiency of processes. The aim of this research line is to help with the development of a predictive model that will enable to evolve from a corrective quality control to a preventive one. Hence, the objective of this particular study is to verify the predictive capabilities of the paint process data on the appearance of defects in the paint film layer and, also, to identify the factors that are going to lead to the appearance of a defect and being able to prevent it. The predictive models applied to the manufacturing industry are being applied mainly to predictive maintenance of equipment [12] and logistics estimates and scheduling [13], movement of materials and products to optimize lots. The study to make the model, has been divided into three main steps. First, the definition, in which the quality criteria to be predicted are fixed and the relevant variables for the prediction are identified. Second, the generation of the data set, in which the data of the variables identified in the previous step are extracted and synchronized. Last but not least, the analytical part in which the model is developed.

II. RESEARCH METHOD

The research was carried out in an European automotive paint shop. The main steps of the painting process are described in 1a. For this research, only a small section of the paint shop work flow is considered (see Figure 1b).

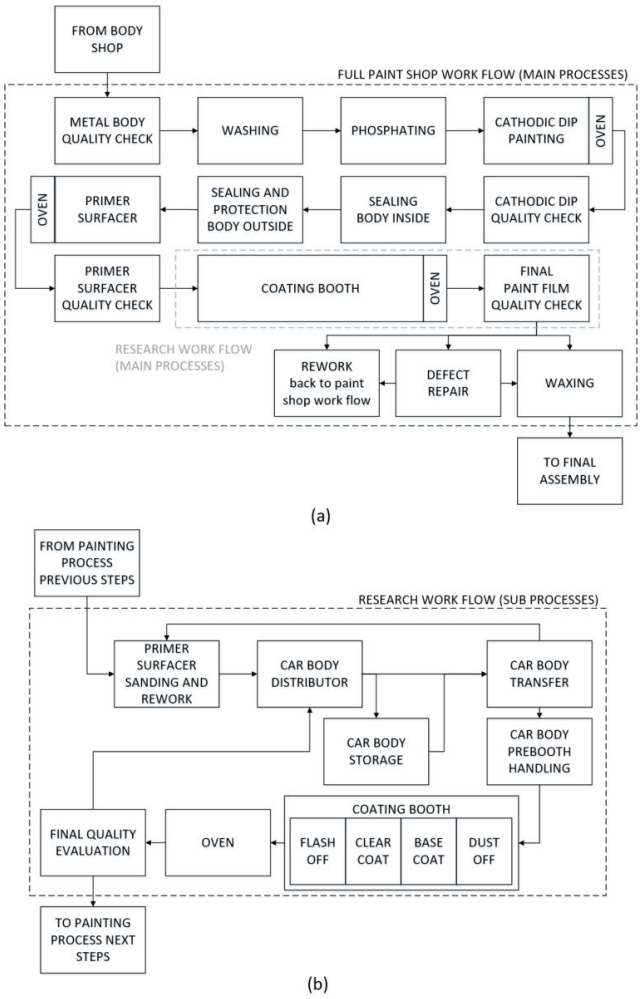

FIGURE 1. (a)Full paint shop work flow.(b)Sub processes considered in the research. Research work flow and subprocesses.

Figure 1 shows the process activities that are considered in order to predict quality results. From the area of quality assurance of the previous layer (primer sanding and rework), the stages of displacement and stay of the car body through the facilities (distributor, storage, transfer and pre booth handling) to the enamel coating process (booth and oven) and the final evaluation of paint quality.

Here, the quality data of the product that is considered for the model development is gathered at the end of the work flow in order to simplify the quality analysis of the product. This is due to the inaccuracy of the quality data collected in previous steps of the painting process (the product quality analysis consists of few samples and is evaluated manually). Thus, the quality of the product up to the entry point of said work flow is assumed to be without defects, although this supposes

suppressing the previous variability due to previous painting sub processes.

The samples considered for the research are the metal parts that make up each vehicle body. This is because a distinction in the variants of the vehicle's chassis configuration that are possible is not considered, such as, the kind of tailgate (which has two options), the presence or absence of rear side doors or the installation of a panoramic roof. So, for example, if a chassis has a panoramic roof (the roof-metal-sheet has a cut all along the piece), the roof-piece of that chassis will not be part of the population of the group of roof-pieces. Another reason for this piece-level evaluation is that, due to the rigorous quality controls motivated by the demand of Total Quality [14] requirement, the final quality control detects defects in all the samples considered at the single-ID chassis level, providing only examples of negative quality and thus, compromising the viability when training the model with data at the single-ID chassis level. The automatic paint film quality evaluation divides the results of each chassis into a certain number of body surfaces that correspond to body parts (both external and internal). Only the external results are used, as the internal surfaces evaluation is not completely automated. Thus, the same number of models have been developed, one for each of the said external surfaces, so the same number of prediction results have been obtained. From these, only the most significant ones are shown. Another relevant consideration is that only vehicles painted in an specific silver enamel have been taken into account since it is, by far, the most applied color in this particular paint shop process and discards the variability due to the chemistry of each enamel. Therefore, the variable regarding the color is not used. The amount of input variables used to train the model is the maximum available (all of which data could be obtained) on the dates the research was carried out. There are three rounds of predictive analysis with variations in the number of samples used to train the model. In the first round, 2 697 valid samples are used to train the model, in the second round 27 243 and in the third round 68 558.

Waikato Environment for Knowledge Analysis (WEKA) [15], [16] is used as the model developing software and multiple algorithms are considered. WEKA integrates the necessary resources for the development of data mining applications in a simple way (data preparation methods, data mining algorithms and visualisation possibilities of data and results are already available in the software) and the solution with the best performance that is provided by WEKA is easily exportable to other languages of industrial interest. The research has been developed in three consecutive stages, the definition of the data set, the generation of the data set and the analysis of gathered data to develop the model (see Figure2).

A. DATA SET DEFINITION

A series of steps are carried out in this first stage, such as: understanding the process, studying its production steps and the sequence of actions that make up the body painting process. Then, identify the Key Performance Indicators (KPI) of the quality criteria that are the objective of the predictive model, as well as the process variables that influence them. For each step, the process input variables and disturbances that can influence the KPI of the quality expected in that sub process are identified. For the identification of the relevant variables for the prediction, a functional analysis of the process is performed, identifying the work flow, defined as the main steps, sub-processes, and actions that, without being a relevant process activity, may have an influence on the result to be evaluated.

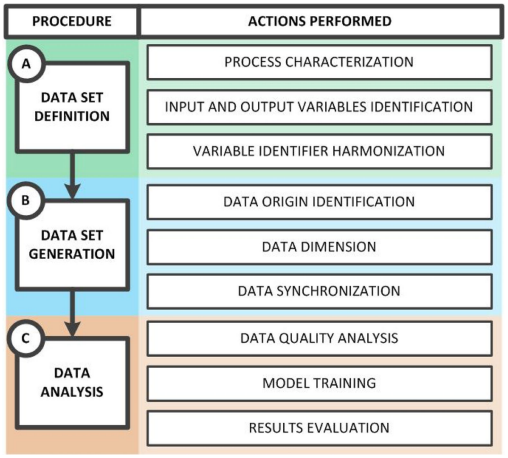

FIGURE 2. Actions performed in the research.

For this research, in order to understand the process using the functional decomposition of the manufacturing work flow, an analysis is carried out by standing at the beginning of the production line and following the car body throughout its painting process so the manufacturing sequence can be defined. Here, the relevant manufacturing steps of the work flow are identified through the study of the factory documentation of the process, the documentation about the control systems, the study about the state of the art of the process [1] and the expert knowledge. Information is collected for each of these steps of the process, identifying, through the said knowledge sources, the relevant variables, the disturbances and an evaluation of the process step performance related to, in this research, paint film quality parameters (as KPI).

Another additional action to deploy in this first step is to develop a procedure for the uni vocal identification of variables. This is because the reference to each variable by any work team involved in the project must be clear. In this particular case, a procedure for naming variables has been developed depending on the location of the measurement (see Figure 3).

As observed in the figure, the uni vocal code designation depends on the location of the variable in the process. The code's multilevel designation is arranged using the study of the paint shop layout, by identifying the main actions that form the painting process and within each action, as consecutive layers, the sub-processes, stations and equipment until

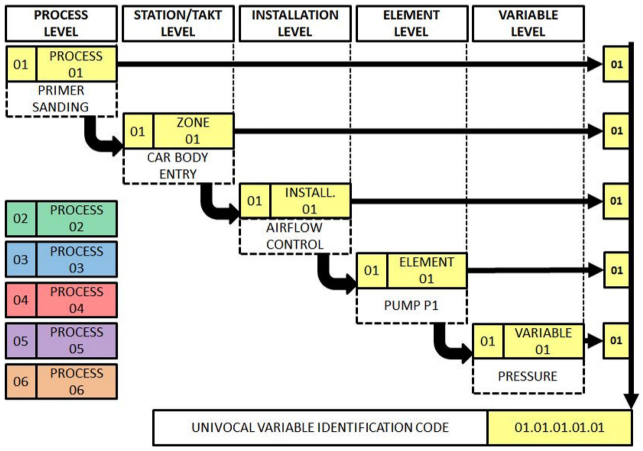

FIGURE 3. Variable unique ID generation procedure.

the variable that has to be identified is reached. The classification is done numerically and, also, assigning a color gradient to each process and related sub-processes and machinery, being position dependent. Thus, any participant of the project can figure out what is referred in the code just by following the path through the layout of the plant. In the example, the code 01.01.01.01.01 refers to, firstly, the initial step considered in the research corresponding to the primer sanding sub-process, secondly, to the car body entry station that is the first recognizable step of the said sub process, thirdly, to the airflow control that is agreed to be the first installation of the station and, lastly, to the P1 pump, being the last code referred to the most relevant variable of that element. If there are others, such as the temperature of the pump, they have the code 01.01.01.01.02 and so on.

B. DATA SET GENERATION

The objective of the data generation procedure is to build a data board in which the rows are each unique chassis and each column is one of the variables considered relevant to the process (see Figure 4).

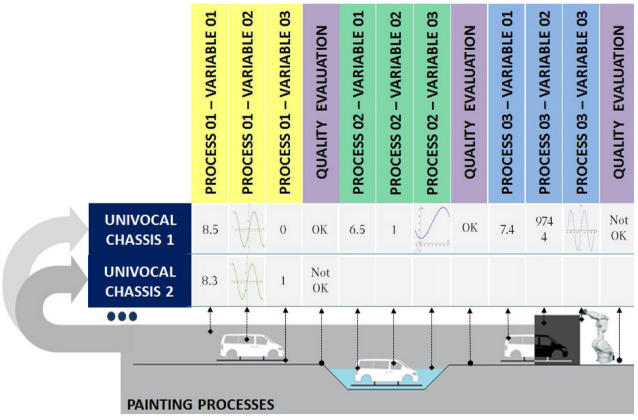

FIGURE 4. Data set synchronisation procedure.

Each chassis unique ID number represents a group of process data with the values of the variables at the time they affected the corresponding paint work. In this study, it has not been stored in relation to the pair chassis-part since it has not been able to perform such a precise synchronization between variables and body parts. Hence, as a simplification, the variable values are assigned to each complete chassis, and so they are to each of its parts. Both categorical and numeric variables are recorded. The numerical values stored in each cell of the data board correspond to the entire time series defined for each variable. In this research, the statistical mean is used to represent the value of each group of time series corresponding values.

To obtain the values of the variables that were identified in the data set definition procedure, a search of the data sources is carried out. Here, the identified variables that are considered relevant for the model by the expert knowledge are filtered in order to save efforts in the subsequent procedural steps. This research considers a source of data to any form of data acquisition and registration of the identified variables. Taking into account the characteristics of each data source, a classification of the relevant variables is made according to several considerations about the nature of the measurement (see Figure 5).

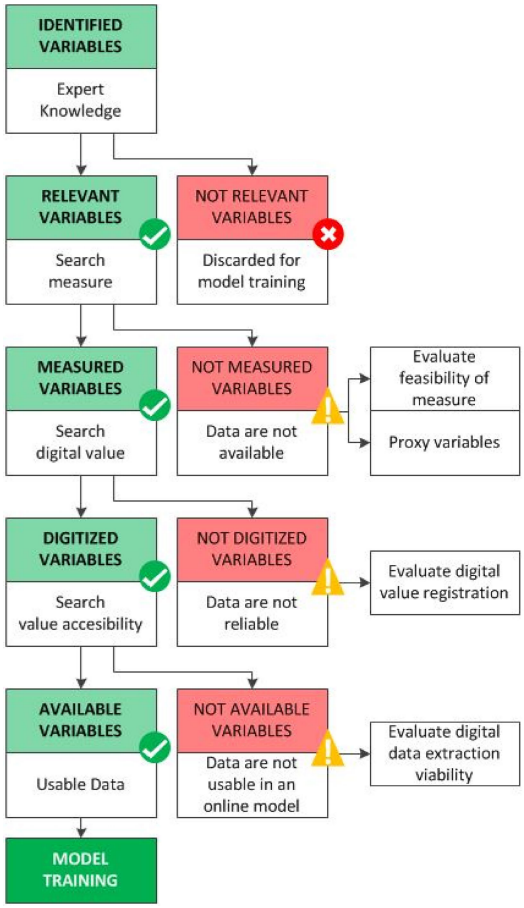

FIGURE 5. Variable classification methodology.

Firstly, a difference is made between measured variables, by any means, or not measured. As a consequence of this first filtering, procedures that lead to obtaining data from the not measured variables are developed, as, for example, the study of the physical or logical dimension of the variable in order to define the appropriate measuring method, or, the consideration of proxy variables that are already available. Secondly, from the list of measured variables, a differentiation is made attending to the variable measurement digitization. That is, to analyse the variable sensor or indicator and define it as mechanic or digital, so, in the second case, the digital data can be somewhere within the paint shop data systems. This analysis leads to define procedures for the digitization of variables whose data can not be exploited due to the use of mechanical instrumentation. In the third and last filter, the digital measurement availability is considered. This evaluates whether the digital data can be consulted directly from the data source or downloaded or it is in an unreachable environment. Thus, methodologies for extracting data from such unreachable sources can be studied [17], [18].

Once the already available data and their related data sources are identified, they are synchronized. The primary synchronization key is the chassis number [19] and the synchronization criteria, as mentioned above, is to store the values of the variables while they affect the paint work. In the paint shop analysed for the research, data capture and storage systems contain different information fields, each data source has its own data structure. Because of this, the synchronization has been done by assembling the available data in each usable data source. To achieve this synchronization sequence, a special algorithm has been developed that consults the production data to know the painted bodies IDs, their colour characteristics and their production times, with regard to the passing times through control points of the process and stay times in stations. Thus, the timestamps of the variables stored in other data sources are identified so that they are assigned to the chassis at the corresponding time. This information is completed with the paint film quality data, so that these quality results are linked to their corresponding process data (see Figure 6).

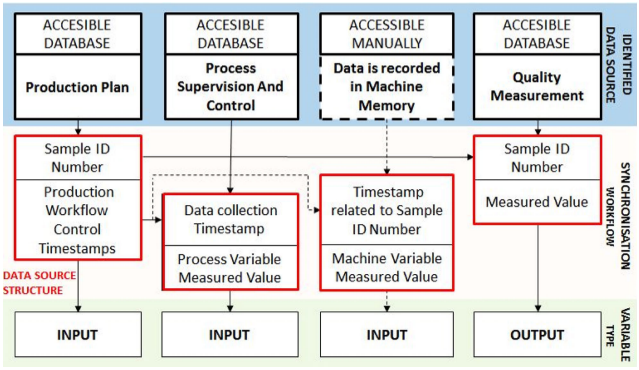

FIGURE 6. Use case data set synchronisation procedure.

C. DATA ANALYSIS

Once the data of the variables are synchronized with the chassis ID as the primary key, a series of steps are carried out prior to the training of the models. In this research step, an evaluation of the quality of the sample's data has been made [20] and the balancing of the samples' good/bad quality cases has been evaluated.

First, to validate the data of the samples, a series of verified indicators have been taken into account for the data quality of the different data sources [21]:

- The *completeness* consists in counting the number of null values in the data source, indicated as a percentage of the total number of values.
- The *conformance* is assessed in two steps, first, the correct format for each variable is established, then, the number of instances in the correct format, that has been defined, is accounted.
- The *consistency* determines if the value studied is far from the average of the normal values.
- The *accuracy/precision* studies the dispersion of the sensor values.
- The *duplicity* shows the number of duplicated data among the files that form the data set.
- The *integrity* determines the expected number of variables for each of the lines in the data set. It verifies that the number of variables found corresponds to those expected.

A car body whose data variables do not meet the data quality criteria has been discarded. This led the research to have a very relevant loss of samples at first (see Figure 7a) since incomplete or non-conforming data have appeared as the enamel coating process progressed (the enamel coating process of the paint shop analysed is divided into base coat BC1-BC2-BC3 and clear coat CC1-CC2 application sub-processes). Once the first data quality problem appears, the sample is discarded. Subsequently, data quality improvement procedures were implemented (as the redefinition of the date format, since there were inconsistencies between dd/mm or mm/dd for the same data source) which managed to correct the situation to a great extent (see Figure 7b).

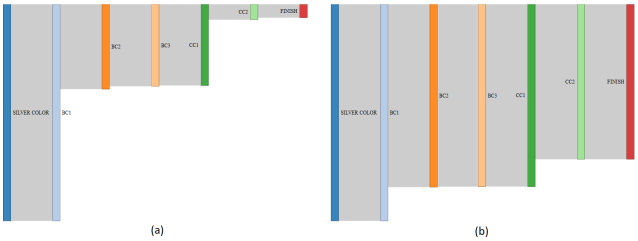

FIGURE 7. (a)Valid samples before data quality enhancement.(b)Valid samples after data quality enhancement.Sample loss due to data quality analysis results.

As mentioned in previous sections, each sample, which corresponds to a unique chassis ID in the data table (see Figure 4), has been decomposed in a certain number of

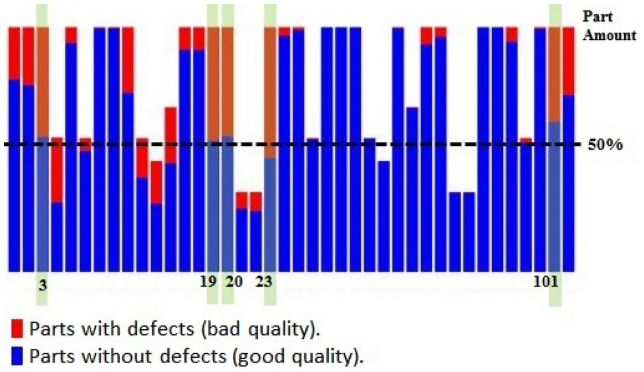

FIGURE 8. Sample good/bad quality distribution. Per PartID.

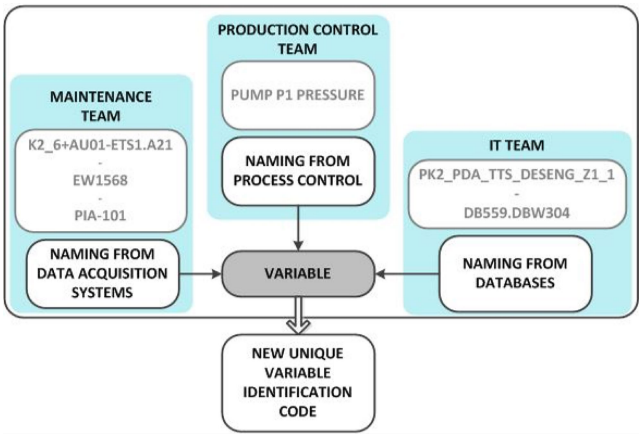

FIGURE 9. Variable unique identification.

body parts. A second step, prior to the training of the model, has been the measurement of the balance between positive/negative quality results for each of the body parts that form a sample (see Figure 8). Here, only the part data groups whose amount is the same as chassis IDs have been considered to maximize the model training quality. So, part groups like the roof-pieces do not have the same amount of part-samples as chassis ID-samples, as some of the chassis have a panoramic roof that is not taken into account for the part-samples.

Over-sampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) [22] have been applied to balance the data sets. The data set balancing results that are valid for the training of the model have been obtained in some of the parts that are marked in green in Figure 8. Specifically, the model training efforts are applied to these parts (see Table 1).

TABLE 1. Vehicle parts codification.

PART NUMBER	PART DEFINITION
3	External-right side hinge.
19	Front right door. External side.
20	Front left door. External side.
23	Hood. External side.
101	Front right cavity w/out post

Once the data have been separated and validated in the corresponding data sets, a series of Supervised Machine Learning Algorithms are applied in order to discover which one gives the best predictive result [23], [24]. In this research, the algorithms were implemented with the *WEKA* software as it allows rapid model prototyping and the best solution can be easily translated into the programming language preferred by the company in order to make it integrable in the existing systems. Precision [25] and Area Under ROC Curve (AUROC) [26] metrics are used to evaluate the model performance.

III. RESULTS

The results of the research are presented at several levels: definition, generation and analysis. Regarding the level of the definition, the result is the improvement of the knowledge about the process, including: an in-depth analysis of the process that provides, for each important step, the input, disturbances and expected results related to paint quality KPIs. A variable definition limit has also been identified, in which the denomination of each variable by each plant team is made according to their particular interests (see Figure 9). To overcome the limit, a coding system is proposed for the uni vocal identification of variables (see Figure 3).

In the generation, the data sources of the plant have been analysed, searching the data of the variables and making a classification of them according to the availability of their data (see Table 2). In this way, procedures have been developed so that all data, whatever their origin, will be available in the near future. These generation procedures have been based on a study of the state of the art in the generation of data from the painting process within the vehicle manufacturing group.

For synchronization, a schema has been generated (see Figure 10) that orders, in the process sequence (spatial synchronization), all the digitized and available data sources, so that the data can be assigned to the variables according to the primary key, the chassis unique number, as accurately as possible (temporary synchronization).

According to the Figure 10, due to the data of the production plan (in pink), it is known, for each chassis ID, the times of entry (in violet) and exit (in navy blue) of each process. In this way, knowing the data sources that exist in each process (material data -in brown-, application data -in grey-, environmental and process control data -in green-), the values of the variables in that time range are assigned to the corresponding sample (chassis ID). In the coating booth data sources, there are intermediate times of exit of the sub processes (in navy blue) that allow assigning the values with greater precision. For the oven data sources, another synchronization strategy has been used. In this case, only the oven entry and exit times are known, but, as the displacement of the chassis through the installation is constant, it has been possible to divide the sections in which each body is located so the curing temperatures of the paint layers corresponding

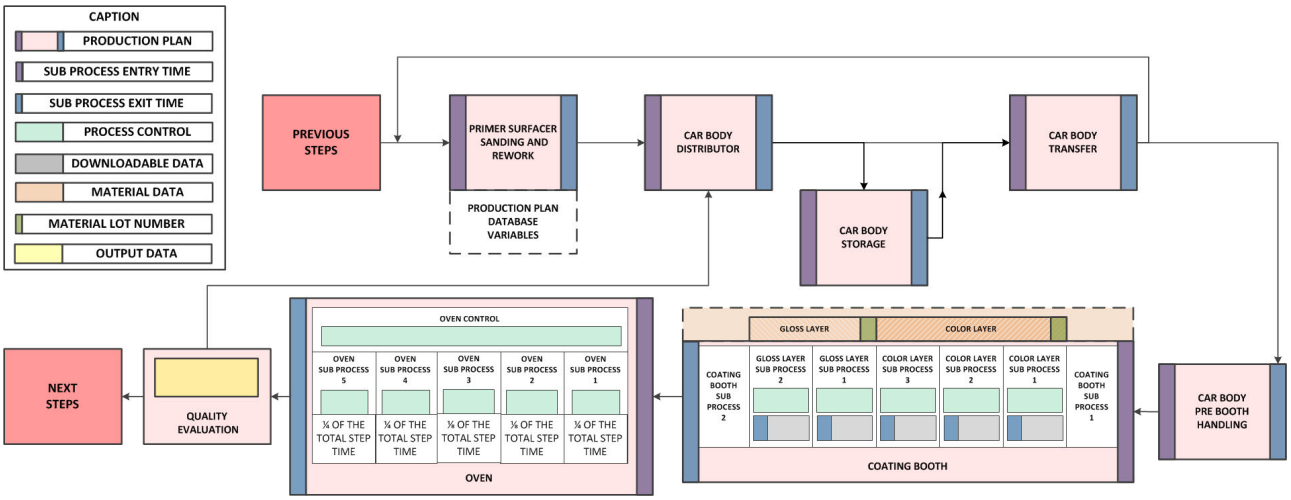

FIGURE 10. Synchronization procedure of the identified data sources.

TABLE 2. Research variables.

	AMOUNT	PERCENTAGE
Identified Variables	679	100
Relevant Variables	402	59
Measured Variables	349	51
Digitized Variables	121	18
Available Variables	104	15

TABLE 3. Model results.

	FIRST ROUND *2697 samples*		SECOND ROUND *27243 samples*		THIRD ROUND *68558 samples*	
PART ID	PRECISION	AUROC	PRECISION	AUROC	PRECISION	AUROC
3	50.58%	0.511	60.52%	0.602	62.72%	0.681
19	50.92%	0.518	55.89%	0.550	61.40%	0.671
20	52.18%	0.520	61.18%	0.652	65.03%	0.690
23	49.69%	0.507	53.56%	0.537	58.48%	0.604
101	65.86%	0.521	71.19%	0.554	66.07%	0.672

to each section have been assigned to each chassis in a precise way. Quality results are directly assigned to the chassis ID.

Data have been obtained for 15% of the variables identified (see Table 2) as relevant. In these 104 variables used to train the model, environmental variables (temperatures and humidity), production variables (such as dwell times of the bodies in the sub processes, the passage through unfinished body storage areas or oven time) or application variables (such as paint product volumes applied by the robots) are included.

In the data analysis part, as a result of the model performance, precision and AUROC have been obtained for three rounds of data. The best performance algorithms have been, decision tree in the first and second round and simple logistic regression in the third round. The difference between rounds is the number of samples used to train the model. Related to the number of samples, another valuable result is the development of procedures that improve data quality and allow to have as many useful samples as possible. The model results have been considered for the part ID with the best balanced datasets, which are parts 3, 19, 20, 23 and 101 (see Figure 8). In the first and second round, the best performing algorithm has been the decision tree and in the third one a simple logistic regression. The most promising results have been in part number 20, with a precision of 65% and AUROC of 0.69 (see Table 3). It must be highlighted that these results have been achieved using only the 15% of the identified process variables.

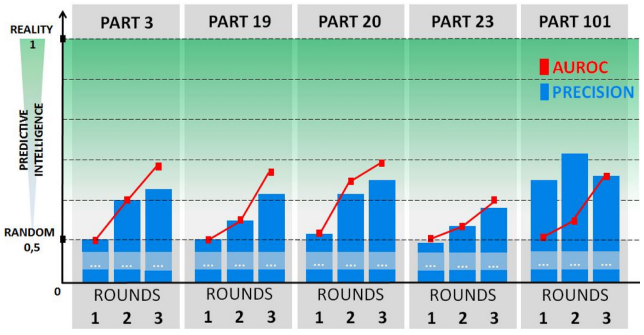

FIGURE 11. Evolution of the model's accuracy and AUROC through three iterations. In each iteration the number of samples is increased according to availability.

Accuracy and AUROC have been increasing as the number of samples increased (see Figure 11) except for the part number 101, in which accuracy has dropped between the second and third round of analysis despite increasing the AUROC.

In the third round of data analysis, the variables considered most relevant for the prediction of the model have been, in general, the application variables related to the amount of material used for coating (see Figure 12). This makes sense since, for example, an excess in the amount of applied material can increase the risk of contamination due to paint atomized particles or, a lack of applied material can worsen the surface covering capabilities of the coloured or glossy layers, making surface defects easier to appear.

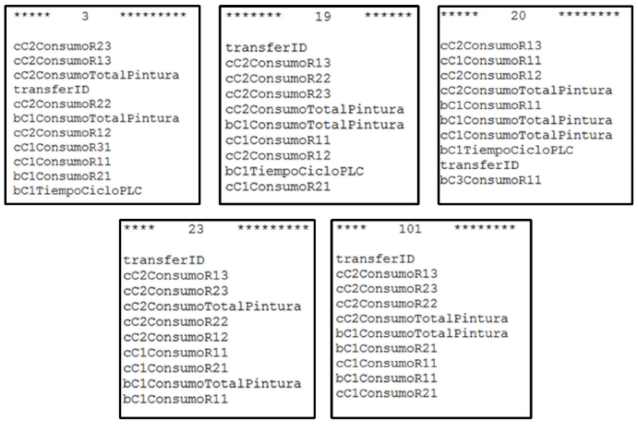

FIGURE 12. Input variables with the highest predictive potential.

As an additional detail to check the validity of the results, analysing the most relevant variables for defect appearance prediction of the chassis part number 20, that corresponds to the left front door, the material application variables correspond to application robots of that same side, the left side, which can be considered as coherent.

IV. DISCUSSION

In the light of the results, it can be inferred that there is a correlation between the input data of the process and the paint film quality output. This implies that it is possible to implement a predictive control in the plant.

Regarding the model, it is able to identify variables that are relevant for the KPI. These variables are related to the amount of paint used by the application robots in the paint booth. This can be explained by the fact that an excessive amount of paint can cause problems of sags or over-spray contamination; while an amount of paint smaller than necessary causes a reduction in coating capability. It should be pointed out that in both cases these are values within the operating regime of the machine, since values outside the range would cause an error. As these errors are critical, those chassis IDs with these defects are not used to train the model. In addition, it is verified that there is a relation between the amount of paint that is applied in the coating booth and the paint film quality as the body pieces on the left are more affected by the paint applicators on the left and vice versa.

Increasing the sample size improves both the accuracy and the AUROC of the model. It must be taken into account that the relation between the increase in the sample size and the improvement in accuracy is not linear, since the increase in accuracy caused by the addition of a new sample decreases as the sample size is larger. There is also a limit on the predictive capacity of the model: increasing the accuracy that has been reached requires an exaggerated number of samples. This implies that it would only be possible to significantly increase the accuracy of the model by including additional input variables, selecting them from those identified but not used yet.

As mentioned above, not all the identified input variables were accessible at the time of the research and, therefore, it was not possible to collect their values, either because there was no measurement system or the data collection system was not appropriate. An example of this second category is a needle pressure gauge or data whose value is registered on paper.

It is therefore concluded that in order to improve the accuracy of the model it is essential to increase the degree of digitalization of the plant [28] so that values from all, or at least a greater percentage, of the relevant identified variables can be extracted.

It should be taken into account that, even if values were obtained for all the relevant variables, the painting process presents a great variability since it is subjected to what is known as special causes [27], i. e., uncontrollable events, such as, contamination due to changes in the atmosphere of the paint booth.

It should be remembered that, in this research, the defects inherited from the previous processes to the enamel coating have not been considered, since in these previous steps, the quality controls are not as exhaustive as the final quality control and many of these defects can only be discovered through destructive tests that are performed only if a defect appears on a recurring basis. The limitation introduced by this simplification could only be avoided by measuring the quality of each of the previous layers with the same level of detail, both in terms of the number of samples and the number of registered variables.

V. CONCLUSION

This research presents an analysis of the vehicle body painting process of a European vehicle manufacturing plant with the aim of verifying the correlation between the process variables and the paint film layer quality of the vehicle.

With this aim, a procedure that consists of three phases is proposed: definition, in which the variables are identified; generation, in which the values of these variables are collected; and analysis, in which the collected data is used to train a predictive model of quality.

From the definition phase, it is concluded that, in a complex environment such as an industrial plant, before facing the collection or processing of data, it is necessary to develop a methodology of univocal denomination of variables so that any member of a work of the plant could be able to easily locate the source that produces each of the variables.

With regard to the generation phase, the main conclusion is that, in order to obtain sufficient data to successfully train a predictive model of paint film quality, it is necessary, on the one hand, that the factory data is accessible, in terms of the number of variables measured and the speed of data collection; and, on the other hand, the spacial and temporal synchronization of the variables, that is, the possibility of identifying the location where a variable occurs and the option of selecting the precise timestamp of the variable when it was relevant for the quality of the paint layer.

Finally, from the analysis phase, it is concluded that it is essential to guarantee the quality of the data and to this aim, it is necessary to define a series of parameters (completeness, conformance, consistency, accuracy, duplicity and integrity) that data must meet to be accepted as valid for training the model. In addition, for the correct training of the model, the number of samples must be balanced with regard to cases of good and poor quality, so that the model is not skewed towards the case overrepresented in the sample.

For the specific case of the paint shop, in view of the results of precision and AUROC values obtained from the prediction model of quality for different sample sizes (in the first round, 2 697 valid samples are used, in the second round 27 243 and in the third round 68 558) and supervised machine learning algorithms (the best performing algorithms were decision tree in the first and second round and simple logistic regression in the third round), it is concluded that the correlation between the input and output data exists. Therefore, it is possible to implement a predictive control in the paint shop that helps improve the efficiency of the process, in terms of rework savings, while maintaining the product quality.

It should be mentioned that to avoid the limitations found through the research and improve the results of precision and AUROC of the predictive model, it is necessary to increase the number of available variable values on the generation of industrial data on legacy machinery, not designed for data streaming in first place. Also, further work is needed with regard to the spatial and temporal synchronization of the process data that is generated from different types of sources, so that the precision with which the measurements are assigned to the variables in the data set is increased.

REFERENCES

[1] H. J. Streitberger and K. F. Dossel, *Automotive Paints and Coatings*, 2nd ed. Hoboken, NJ, USA: Wiley, 2008.
[2] D. Rambabua, G. Bhaskarb, A. Leninc, and K. Pazhaniveld, "Advanced product design and optimization using SAP PLM and integration with supply chain processes in automotive manufacturing," *J. Web Eng.*, vol. 17, no. 6, pp. 2089–2103, 2018.
[3] M. A. B. Marzuki, M. F. M. Azmi, and R. L. Jaswadi, "Design optimization of automotive component through numerical investigation for additive manufacturing," *J. Built Environ., Technol. Eng.*, vol. 6, pp. 19–26, May 2019.
[4] G. Cole and A. Sherman, "Light weight materials for automotive applications," *Mater. Characterization*, vol. 35, no. 1, pp. 3–9, Jul. 1995.
[5] J. Galán, L. Samek, P. Verleysen, K. Verbeken, and Y. Houbaert, "Advanced high strength steels for automotive industry," *Arch. Civil Mech. Eng.*, vol. 8, no. 2, pp. 103–117, 2008.
[6] M. Garavaglia, A. G. Demir, S. Zarini, B. M. Victor, and B. Previtali, "Process development and coaxial sensing in fiber laser welding of 5754 Al-alloy," *J. Laser Appl.*, vol. 31, no. 2, May 2019, Art. no. 022419.
[7] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the Internet of Things and industry 4.0," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 17–27, Mar. 2017.
[8] M. Hermann, T. Pentek, and B. Otto, "Design principles for industrie 4.0 scenarios: A literature review," Technische Univ. Dortmund, Dortmund, Germany, Working Paper No. 01/2015, 2015. Accessed: Jun. 20, 2015, doi: 10.13140/RG.2.2.29269.22248.
[9] Paint Defects. PPG Industries. (Jul. 4, 2019). *PPG Industries 2019*. [Online]. Available: https://uk.ppgrefinish.com/en/paint-defects/
[10] J. Tornero, L. Armesto, M. C. Mora, N. Montes, Á. Herráez, and J. Asensio, "Detección de defectos en carrocerías de vehículos basado en visión artificial: Diseño e implantación," *Revista Iberoamericana de Automática e Informática Ind. RIAI*, vol. 9, no. 1, pp. 93–104, Jan. 2012.
[11] T. Bauernhansl, "Die vierte industrielle revolution—Der Weg in ein wertschaffendes produktionsparadigma," in *Industrie 4.0 in Produktion, Automatisierung und Logistik*. T. Bauernhansl, M. ten Hompel, and B. Vogel-Heuser, Eds. Wiesbaden, Germany: Springer, 2014.
[12] Y. Peng, M. Dong, and M. J. Zuo, "Current status of machine prognostics in condition-based maintenance: A review," *Int. J. Adv. Manuf. Technol.*, vol. 50, nos. 1–4, pp. 297–313, Sep. 2010.
[13] L. Mönch, J. Zimmermann, and P. Otto, "Machine learning techniques for scheduling jobs with incompatible families and unequal ready times on parallel batch machines," *Eng. Appl. Artif. Intell.*, vol. 19, no. 3, pp. 235–245, Apr. 2006.
[14] R. H. Chenhall, "Reliance on manufacturing performance measures, total quality management and organizational performance," *Manage. Accounting Res.*, vol. 8, no. 2, pp. 187–206, Jun. 1997.
[15] S. R. Garner, "WEKA: The waikato environment for knowledge analysis,"in *Proc. New Zealand Comput. Sci. Res. Students Conf.*, 1995, pp. 57–64.
[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[17] L. Wang and J. Qi, "The real-time networked data gathering systems based on ethercat," in *Proc. Int. Conf. Environ. Sci. Inf. Appl. Technol.*, vol. 3, Jul. 2009, pp. 513–515.
[18] C. Zhu, S. Wu, G. Han, L. Shu, and H. Wu, "A tree-cluster-based data-gathering algorithm for industrial WSNs with a mobile sink," *IEEE Access*, vol. 3, pp. 381–396, 2015.
[19] H. J. Grundig and M. Klein, *Business Intelligence Recording of Process Data at GM Paint Shops*. Berlin, Germany: Strategies Car Body Painting, 2016.
[20] G. Press. (2019). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Forbes*, Accessed: Feb. 15, 2019. [Online]. Available: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparationmost-time-consuming-least-enjoyable-data-science-task-survey-says/
[21] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 381–396, 2002.
[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jul. 2018.
[23] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.
[24] D. Wolpert and W. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.
[25] E. W. Steyerberg, S. E. Bleeker, H. A. Moll, D. E. Grobbee, and K. G. Moons, "Internal and external validation of predictive models: A simulation study of bias and precision in small samples," *J. Clin. Epidemiology*, vol. 56, no. 5, pp. 441–447, May 2003.
[26] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[27] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*. New York, NY, USA: Random House Group, 2007.
[28] K. Bley, C. Leyh, and T. Schäffer, "Digitization of german enterprises in the production sector—Do they know how 'digitized' they are?" in *Proc. 22nd Amer. Conf. Inf. Syst. (AMCIS)*, San Diego, CA, USA, Aug. 2016. [Online]. Available: https://aisel.aisnet.org/amcis2016/EntSys/Presentations/9/

JAVIER SALCEDO-HERNÁNDEZ received the master's degrees in automotive engineering and big data environment. He is currently pursuing the Ph.D. with the University of Deusto, Spain. He is an Industrial Engineer. He is conducting research on the digitization of automotive paint shops for the development and implementation of the proposals made in the European paradigm of Industry 4.0. He has participated in applied research projects related to the maintenance of electricity transport infrastructures and the development of electric vehicles, specifically in the design of applications and prototypes of onboard embedded systems. He also has experience in the field of industrial maintenance. His research interests include advanced manufacturing, data life cycle management, and business intelligence.

JON GARCÍA-BARRUETABEÑA is an Industrial Engineer, specialized in structural mechanics. He joined the University of Deusto, in 2013, from the Ikerlan IK4 Research Center. He began his teaching career, in 2004, at Mondragon Unibertsitatea, and his area of specialization is structural dynamics. His research is focused on the vibroacoustic analysis of mechanical systems from the experimental point of view as well as modeling and simulation. He currently directs three doctoral theses and is the Director of two masters in automotive engineering, both official and one of them international. He has a recognized research merit, from 2008 to 2013, and his latest research focuses on the field of automotive comfort. He has ten high impact JCR publications, two book chapters, more than 15 research projects as well as numerous contributions in international congresses.

IKER PASTOR-LÓPEZ received the degree in computer engineer, in 2007, the master's degree in information security, in 2010, and the Ph.D. degree *(cum laude)* in computer science, in 2013. He completed a program in big data and business intelligence, in 2016. He worked in the Industry Unit, DeustoTech, and focuses its scientific interests in the areas of big data analytics, opinion mining, and computer vision. Since 2019, he has been a Researcher with the Faculty of Engineering, University of Deusto. He is the author of several scientific articles reviewed by peers in conferences and indexed journals. He has participated in the gestation, scientific development, and technical development in numerous competitive projects and contracts with companies, the latter with several successful cases of knowledge transfer actions. He is a member of the scientific committee of several congresses, such as CISIS, SOCO, and ICEUTE, and a Reviewer of journals included in the JCR as Magazine of Engineering and Industry—DYNA.

BORJA SANZ-URQUIJO received the Ph.D. degree *(cum laude)* in information systems, in 2012, in the topic of malware detection in Android mobile devices. He was a Researcher with the Computing Research Unit, DeustoTech, from 2008 to 2018, and was the Head Researcher, from 2015 to 2018. Since 2019, he is a Researcher with the Faculty of Engineering, University of Deusto. His main research skills and interests include machine learning, big data, knowledge discovery, and information retrieval. He has a long track-record leading national and international projects in the research areas previously mentioned and has worked closely with different social agents, enterprises, and other research centers. He has participated in more than 20 research projects, between H2020, national, and private projects, being the Project Manager in several of them. He has published several book chapters and more than 35 articles in specialized national and international journals of impact, such as *Logic Journal of IGPL, Electronic Commerce Research and Applications* or *Expert Systems with Applications*.