IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 3, 2019, accepted June 13, 2019, date of publication June 17, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923405

Multistage Quality Control Using Machine Learning in the Automotive Industry

RICARDO SILVA PERES[1,2], JOSE BARATA[1,2], (Member, IEEE), PAULO LEITAO[3], (Member, IEEE), AND GISELA GARCIA[4]

[1]UNINOVA—Centre of Technology and Systems (CTS), FCT Campus, 2829-516 Caparica, Portugal
[2]Faculdade de Ciencias e Tecnologia, Departamento de Engenharia Electrotecnica, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[3]Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politecnico de Braganca, Campus de Santa Apolonia, 5300-253 Braganca, Portugal
[4]Volkswagen AutoEuropa, 2954-024 Quinta do Anjo, Portugal

Corresponding author: Ricardo Silva Peres (ricardo.peres@uninova.pt)

This work was supported in part by the FCT/MCTES (UNINOVA-CTS funding UID/EEA/00066/2019) and the GO0D MAN project from the European Union's Horizon 2020 research and innovation programme under Grant 723764.

ABSTRACT Product dimensional variability is a crucial factor in the quality control of complex multistage manufacturing processes, where undetected defects can easily be propagated downstream. The recent advances in information technologies and consequently the increased volume of data that has become readily available provide an excellent opportunity for the development of automated defect detection approaches that are capable of extracting the implicit complex relationships in these multivariate data-rich environments. In this paper, several machine learning classifiers were trained and evaluated on varied metrics to predict dimensional defects in a real automotive multistage assembly line. The line encompasses two automated inspection stages with several human-operated assembly and pre-alignment stages in between. The results show that non-linear models like XGBoost and Random Forests are capable of modelling the complexity of such an environment, achieving a high true positive rate and showing promise for the improvement of existing quality control approaches, enabling defects and deviations to be addressed earlier and thus assist in reducing scrap and repair costs.

INDEX TERMS Machine learning, quality control, predictive manufacturing system, multistage, automotive industry, industry 4.0.

I. INTRODUCTION

Product dimension variability is one of the most challenging aspects involved in multistage manufacturing processes like assembly and machining in industries such as automotive, aerospace and white goods [1]. The complexity of a Multistage Manufacturing Process (MMP) is extremely demanding across the several engineering domains involved, from process modeling to process control and fault diagnosis, particularly regarding the assurance of the product's dimensional integrity. Furthermore, this inherent complexity and the random nature of uncertainties and disturbances in manufacturing processes make it considerably difficult to guarantee the desired quality of the product. Therefore, an effective method to enable the automated and early detection of potential defects during production using online data would be highly advantageous to manufacturers.

In this light, the Predictive Manufacturing Systems (PMS) paradigm has been gaining traction as an approach to develop solutions ready to answer this need. Due to the growing adoption of Industry 4.0 [2] concepts and the Industrial Internet of Things ideology, the foundation for the realization of such systems is being laid down with smart sensor networks and smart machines, with more and more data being generated every day. Hence, the conditions are being created for the utilization of advanced prediction tools capable of systematically processing these data into information that can explain the aforementioned uncertainties and thus assist personnel in making more informed decisions [3]. An example of this is the Watchdog Agent developed at the Center for Intelligent Maintenance Systems (IMS), which consists in a toolbox of algorithms for multi-sensor performance assessment and prediction [4], [5]. Some of its tools include signal processing

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

and feature extraction, health assessment through logistic regression and Support Vector Machine (SVM) based condition diagnosis. Another recent example is the IDARTS framework, proposed in [6] and based on the Cyber-Physical Production System (CPPS) concept, which presents a generic approach for the implementation of a PMS using a combination of flexible data fusion running on the edge-level with both offline and online data analysis using Machine Learning (ML) techniques.

In this paper we address the application of a PMS solution in the automotive industry, comparing the fitness of varied binary classifiers in the prediction of dimensional defects in an MMP within the Volkswagen AutoEuropa plant in Portugal. The remainder of this paper is structured as follows. Section II presents a brief overview of related work found in current literature. Afterwards, Section III describes the case study, data set and methods applied in this work. This is followed by the presentation of the experimental results from the training and validation of the different models in Section IV, along with an identification of the limitations of the approach and possible solutions in V. Finally, some discussion and closing remarks are provided in Section VI.

II. RELATED WORK

ML is currently regarded as an extremely promising field to provide improved quality control and process optimization in PMS [7]. The reason why ML techniques are regarded as such is greatly due to their capacity to handle high-dimensional, multivariate data, along with the ability to understand the implicit relationships within large data sets in complex and dynamic environments [8].

There are several accounts of the successful employment of data mining and ML to tackle challenges in areas such as process optimization [9], fault detection [10] and predictive maintenance [11] in manufacturing environments. While algorithms like Logistic Regression are typically well suited for finding primary relationships in the data, the independent variable matrix can quickly become large when the the problem involves detecting second or third order interactions [12]. Tree based classifiers like Random Forest (RF) and Gradient Boosted Tree models can detect relationships that are not as easily picked up with linear techniques.

Concerning tree-based classifiers in the manufacturing domain, several applications can be found in the literature. Wu et al. [13] applied RFs to predict tool wear in milling operations based on several features extracted from cutting force, vibration, and acoustic emission signals, with experimental results showing RFs were capable of yielding more accurate results in this instance than SVM and artificial neural network models. In [14] decision trees and RF models are used for pattern recognition classification of ultrasonic oscillograms of resistance spot welding joints. The authors point out that while both can be employed as effective decision support tools to improve quality control, when compared with regular decision tree models RFs reduced the error rate at the expense of decision interpretability. Finally, Syafrudin et al. [15] developed a real-time monitoring system using on a combination of outlier detection and RF classifiers for fault detection based on sensor data, which was tested on an automotive assembly line in Korea.

Gradient Boosted Trees are another class of ML algorithms with many succesful applications in the manufacturing industry. In [16], Chen et al. employ data-driven models based on the XGBoost [17] algorithm for the real-time prediction of welding quality in a metal active gas welding process. According to the authors, XGBoost models were shown to be capable of capturing complex nonlinear characteristics of the sensor data and dealing with anomalies in the data set. Jabbar et al. [18] proposed a decision-making tool based on XGBoost to support operators in the manufacturing of printed circuit boards. XGBoost was shown to yield both high accuracy and high recall in the classification of defects when trained on real-world data.

In summary, the related work presented in this section builds on previous research to explore how predictive modeling can be applied to manufacturing in the context of the PMS paradigm. While ML shows promise to this effect, the suitability of its techniques and different models still needs to be assessed on a case by case basis. This paper explores the application of several ML algorithms to the prediction of dimensional defects in an automotive MMP, with the goal of mitigating the propagation of said defects downstream, enabling an earlier intervention and thus contributing to the improvement of existing quality control practices by reducing scrap and repair costs.

III. MATERIALS AND METHODS

A. CASE STUDY SPECIFICATION

This case study is focused on a multistage assembly line from Volkswagen AutoEuropa's body shop, specifically between the framing inspection stage and the body-in-white assembly inspection. The goal is to apply ML using the data from the framing inspection to predict whether or not a given car is likely to carry dimensional defects later on during the inspection downstream after the different assembly operations and alignment stages. Figure 1 provides a simplified overview of this scenario:

In reality, a series of assembly operations is performed between the framing and the finish line stages, effectively putting together the car's body-in-white. On the left side of Figure 2 the tailgate assembly operation is showcased. With the help of a jig the operator mounts the tailgate onto

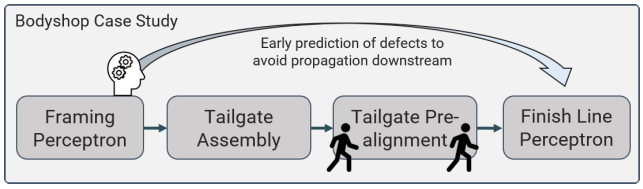

FIGURE 1. Diagram of the multistage assembly process being studied. Data originates from the automated measurements before and after the assembly of the tailgate.

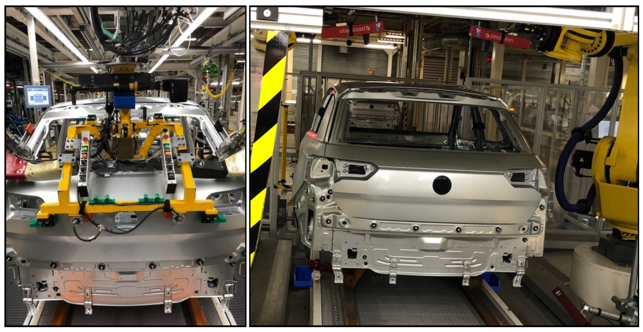

FIGURE 2. Fixture for the tailgate assembly process (left) and the Perceptron measurement station downstream (right).

the previously naked frame of the car. Once the assembly operations are concluded, a pre-inspection is conducted after which the car goes into the doors-to-body Perceptron station (right side of Figure 2).

This means that data originates from two different dimensional spaces. At the earlier stage, only the X, Y and Z features of the car's frame are available along with the respective symmetries, with the first Perceptron station measuring the deviations in relation to the product's CAD design. Once the different parts are assembled, the last Perceptron is then able to measure the gap and flush at different points between the tailgate and the sides of the car.

In this MMP defects and variations in early stages can have a significant impact in stages downstream and often remain undetected until the final inspection station. This is heavily connected to the fact that while in practice one can monitor and control the different dimensional characteristics and their boundaries on each stage, combinations of small variations within the acceptable thresholds can easily remain undetected and translate into quality problems downstream. Therefore ML can serve as the means to model the complex relationships between the framing data and the gap and flush measurements at the end of the body shop line, enabling quality control engineers to either improve the earlier stages to avoid these defects or to provide timely indications to the assembly stages in between in order to have proper alignment during their respective operations.

B. CHARACTERIZATION OF THE DATA SET

Within the context of ML, classification can be defined as the process of finding a model capable of distinguishing data classes or concepts. In supervised learning, such models are derived from the analysis of a set of labelled training data (i.e. data for which the class labels are known), which later enables the model to be used to predict the class label of previously unseen, unlabelled objects [19].

For this case study, the data set encompasses a total of 18148 unique cars with 29 dimensional features from the framing inspection station. Each car sample is labelled as 'OK' or 'NOK' according to a domain expert's assessment based on the gap and flush measurements at the last station, with 11331 and 6545 samples belonging to each class, respectively. The actual designations of the features were anonymized as requested by the manufacturer in order to protect the privacy and property of the use case.

Out of these 29 features, 10 present over 85% entries of missing values, resulting in only 19 features being used in the analysis.

Furthermore, to address the class imbalance, random under-sampling was performed on the data, generating a balanced data set with 12012 samples. Considering that the chosen sampling technique might result in some information loss, synthetic minority over-sampling was also tested as an alternative (after the train-test split) but provided significantly worse results.

Finally, 119 observations in the balanced data set still presented missing values, either due to the car still being on its way along the line between the two inspection stations, or due to some measuring or communication disturbance. Since these were relatively rare occurrences the samples with one or more missing values were discarded, although in future work it might be interesting to study the impact of different imputation techniques instead.

The absolute correlation matrix for the resulting data set can be found in Figure 3.

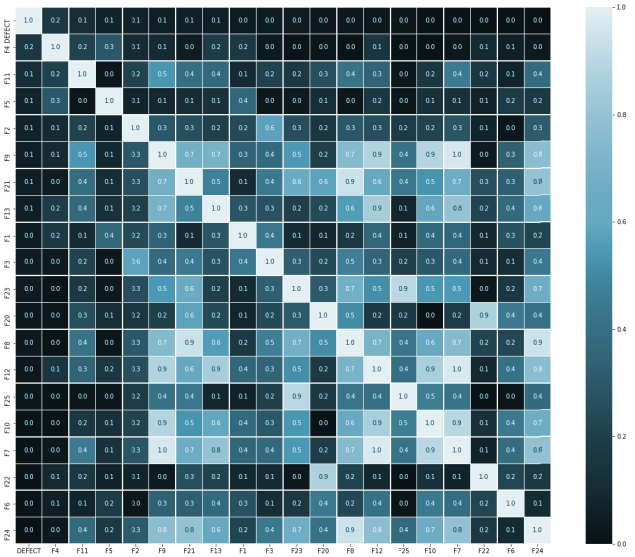

FIGURE 3. Matrix of the absolute correlations between the different features based on the Spearman coefficient.

As it can be observed, most features present low correlation coefficients with the target, suggesting that if there is in fact a relationship between the features and the car's quality downstream, non-linear classifiers might be more adequate for the case at hand. Also, there is some evidence of multicollinearity, with cases of high correlation between some of the features, which is to be expected given that the data set pertains to several dimensional characteristics of the car which are expected to be correlated.

To facilitate the visualization of the data set, Principal Component Analysis (PCA) was applied to reduce the feature

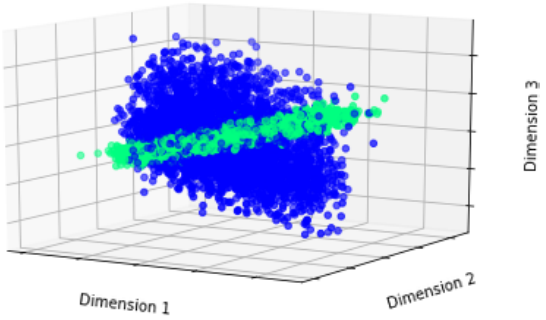

FIGURE 4. Visualization using PCA to reduce the dimensionality of the data set. Data points are colored based on the target variable.

space to only three dimensions in order to make it possible to visualize it in 3D space. The resulting plot is shown in Figure 4.

Roughly 81% of total variance is captured in the first three components resulting from the application of PCA. However, the results seem to suggest that based on the features available in the data set a reasonable class separation can be achieved. The following section provides an overview of the different algorithms employed to this effect in this study, as well as of the corresponding methods and implementation.

C. ALGORITHMS FOR DEFECT CLASSIFICATION

The implementation of the models contemplated in this study followed a fairly straightforward methodology, using Python 3.6 and the scikit-learn module [20] for all models except XGBoost [17].

Firstly, the data set was split into train and test sets. Afterwards, given that several features present skewed distributions with both positive and negative values, Yeo-Johnson transform was applied to reduce the shift followed by standardization to center and scale each feature individually using the *RobustScaler* from scikit-learn's preprocessing module. This was used instead of a standard scaler due to it being more robust to outliers in the data. Both the power transform and scaler were fitted only to the train set to avoid test set contamination. An example of the result from this preprocessing step for F12 is illustrated in Figure 5.

After this preprocessing step, several models were trained to establish a baseline with 5-fold cross validation being

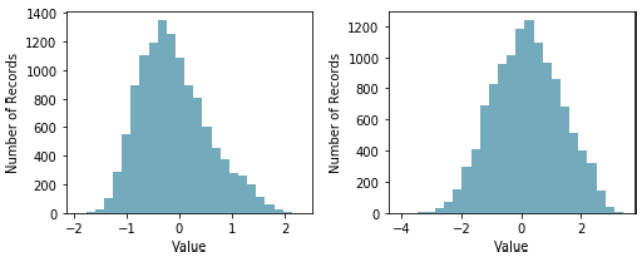

FIGURE 5. Result from the application of the Yeo-Johnson transform and robust scaler to F12. Raw distribution is presented on the left, transformed values are presented on the right.

performed on the top scoring models. Following this, the best models from this step were selected for hyperparameter tuning and finally tested on a separate holdout set, consisting of cars collected over the three days after the last sample from the original data set.

In the remainder of this section a review of the different types of models used in this work is provided, followed by a brief description of the evaluation metrics used to assess the performance of the various models.

1) GAUSSIAN NAIVE BAYES

The Gaussian Naive Bayes (GNB) method is a supervised learning algorithm based on the Bayes' theorem with the naive assumption of conditional independence between the various pairs of features given the value of the target variable. The *GaussianNB* class from scikit-learn implements GNB for classification, with the likelihood of the features assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(- \frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right) \quad (1)$$

where the parameters σ_y and μ_y are estimated using maximum likelihood.

Some practical applications of GNB include text prediction, document classification and spam filtering. It requires a relatively small amount of training data to estimate the necessary parameters, can be quite fast in comparison to more complex methods and is easy to implement, being often used as a baseline [21]. However, while its naive assumptions can make such efficiency possible, they can also adversely affect the quality of the results in several real world applications, such as the use case at hand, in which the feature pairs are unlikely to be independent.

2) K-NEAREST NEIGHBOURS

K-Nearest Neighbours (KNN) is a type of instance-based learning algorithm, meaning it does not construct a general internal model, but instead stores instances of the training data with computation being deferred until classification. Over the years it has seen several applications in both statistical estimation and pattern recognition including for instance the classification of heart disease to provide a decision-support system for clinicians [22]. Conceptually, such an approach can be carried over to the use case at hand, as we are effectively attempting to identify a condition in the cars, and furthermore, being one of the simplest ML algorithms for classification it is at least a good candidate to serve as a baseline.

For KNN, the input consists in the k closest training examples in the feature space, with the output being a class membership attributed by a simple majority vote of the nearest neighbours based on some distance metric such as the Euclidean distance.

3) XGBOOST

XGBoost [17] stands for eXtreme Gradient Boosting and is an optimized implementation of gradient boosted trees, designed to be highly efficient and flexible. It is a non-linear algorithm which typically works well with numerical features and requires relatively less feature engineering and hyperparameter tuning to yield good results.

Generally, such methods can be prone to overfitting, as they constantly involve fitting a model on the gradient. To mitigate this, one can optimize for the number of trees until the out of sample error starts increasing once more.

XGBoost models are frequently used to solve Kaggle challenges across several domains, with real world applications including for instance the identification of complex relationships between variables for rare failure prediction in manufacturing processes [12].

4) RANDOM FOREST

In the context of classification problems, RF is an ensemble learning method that operates by constructing several decision trees at training time and outputting the class that is the mode of the classes of the individual trees. While a single decision tree can easily run into overfitting problems, being also sensitive to small variations in the data, due to their nature RFs are more robust to such challenges.

5) SUPPORT VECTOR MACHINE

The SVM algorithm constructs hyperplanes in infinite-dimensional spaces to classify data into distinct classes. One can consider a good separation to be achieved by the hyperplane with the largest distance to the nearest training-data point of any class (functional margin), as typically larger margins correspond to a lower generalization error.

While this is a fairly formal approach to the classification problem, one disadvantage mentioned in the scikit-learn documentation for the *SVC* implementation is that fit time complexity is more than quadratic with the number of samples, making it hard to scale for data sets with more than a couple of 10000 samples. While this is not the case for this particular case study, it is something to keep in mind when comparing to other approaches.

D. EVALUATION METRICS

1) ACCURACY

Accuracy can be used as a statistical measure of how well a binary classifier identifies or excludes a condition. It is the proportion of true results among all the observed cases. The formula for quantifying binary accuracy is:

$$Accuracy = \frac{tp + tn}{tp + fp + fp + fn}. \tag{2}$$

where tp, tn, fp and fn refer to true positives, true negatives, false positives and false negatives, respectively.

However, while high accuracy is typically regarded as a good indicator of performance, accuracy alone can be very misleading, particularly for imbalanced cases. Also as a metric for comparison the same holds true, as two models can yield the same accuracy results while performing differently with respect to the types of correct or incorrect predictions they provide.

2) RECALL

To assist with the aforementioned challenge, one other metric that can be calculated is recall. Recall represents the proportion of true positives that was identified correctly, thus being a suitable metric to use for model selection when there is a high cost associated with false negatives. It can be calculated as follows:

$$Recall = \frac{tp}{tp + fn}. \tag{3}$$

3) PRECISION

To complement this, precision is then the proportion of the values identified as positives that was actually correct. As such, it is an adequate measure to use when the cost associated with false positives is high, being calculated as indicated in 4.

$$Precision = \frac{tp}{tp + fp}. \tag{4}$$

4) F1 SCORE

For cases in which a balance between precision and recall is preferable, and particularly when there is an uneven class distribution, the F1 score is often used as the evaluation metric. It is the harmonic average of the precision and recall, with 1 and 0 being its best and worst values, respectively, as given by the formula:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{5}$$

5) AREA UNDER THE RECEIVER OPERATING CHARACTERISTICS

Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) curve is a performance measurement for classification problems at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It provides an indication of how well a model is capable of distinguishing between classes. More specifically for this case study, the higher the AUC, the better the model is at predicting cars that are OK as OK, and cars that are NOK as NOK.

The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on y-axis and FPR is on the x-axis.

IV. RESULTS

At first, several models were implemented without any hyperparameter tuning to create a baseline. The model training was performed on a machine with an Intel Core i7-9700K, 2x8GB 4000MHz DDR4 memory and an NVIDIA GeForce RTX 2070. The results are summarized in Table 1.

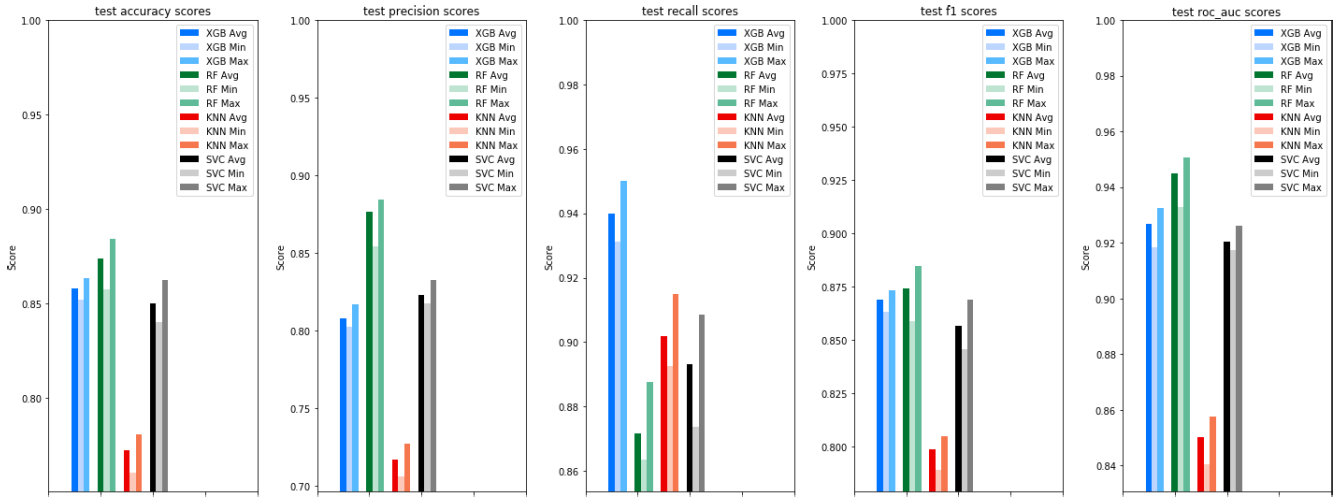

FIGURE 6. Test results from 5-Fold Cross Validation. Results are divided by accuracy, precision, recall, f1 and roc_auc scores.

TABLE 1. Baseline model results.

Model	Accuracy	Training (s)	Prediction (s)
Random Forest	87.873	0.148	0.006
SVC (RBF Kernel)	85.325	2.076	0.470
XGBoost	84.331	0.301	0.007
K-Nearest Neighbours	77.962	0.008	0.371
Logistic Regression	57.936	0.073	0.001
Naive Bayes	56.178	0.004	0.001

As hypothesized during the exploratory data analysis from Section III-B, the two linear models, Logistic Regression and Naive Bayes, performed significantly worse than the worst performing non-linear classifier, suggesting a stronger non-linear relationship between the features and the target. Based on this, 5-fold cross validation was performed on the four non-linear classifiers in order to obtain a more realistic measure of accuracy and avoid overfitting on the training data. The results from the cross validation step can be found in Figure 6.

From the observation of Figure 6, it can be said that the baseline RF model performed better on average across all metrics except for recall, for which XGBoost was considerably superior. The XGBoost and SVM models performed slightly worse, with KNN performing considerably worse overall. One particularity to take into account in this MMP is the possibility of the feature distributions to change over time. This can happen for several reasons and is further discussed in Section V, but to tackle this challenge, an approach could be to monitor the accuracy of the deployed model and retrain it if it drops below a certain threshold. This means that more computationally expensive models that take longer to train and perform cross validation on, like SVM, might not be adequate for such a scenario.

Based on this, hyperparameter tuning through randomized search was performed on the three best models, which were then compared on the test set based on the same evaluation metrics used for cross validation. The tuned parameters can be found in Table 2, where any omitted parameters are assumed to take the default values from their respective implementations.

TABLE 2. Parameters for each model resulting from the tuning through randomized search optimizing for roc_auc. Tuning was performed on 100 iterations with 5-fold cross validation.

Model	Parameters
XGBoost	colsample_bytree: 0.970, gamma: 6.079, learning_rate: 0.202, max_depth: 11, min_child_weight: 11.507, n_estimators: 59, reg_alpha: 0.232, subsample: 0.962
Random Forest	n_estimators: 500, min_samples_split: 2, min_samples_leaf: 1, max_features: auto, max_depth: 50
SVC	C=10, gamma=0.01, kernel='rbf'

TABLE 3. Tuned model results. Models are evaluated based on the same metrics used for the baseline models' cross validation.

Model	Accuracy	Recall	Precision	F1	ROC AUC
XGBoost	0.928	0.979	0.888	0.931	0.972
Random Forest	0.925	0.981	0.883	0.929	0.977
SVC	0.914	0.977	0.868	0.919	0.969

The results are summarized in Table 3. Additionally, the corresponding ROC curves can be found in Figure 7, in which the dashed diagonal line defines the reference point for which the models have no capacity to distinguish between classes.

The results are extremely close, especially for the two ensemble models, with XGBoost being superior in three out of the five evaluation metrics, if only by a slight margin when compared to the values scored by the RF model. The SVC model appears to not have generalized as well as the others as evidenced by its lower capacity to separate the target classes in the ROC curve, albeit with marginal differences and while still yielding fairly improved results over those of its baseline counterpart.

Finally, each model was tested on a new holdout data set, originating from measurements taken from 1000 cars over the three days following the last entry of the original data set (8 samples were discarded due to missing values). The resulting confusion matrices are depicted in Figure 8. The results

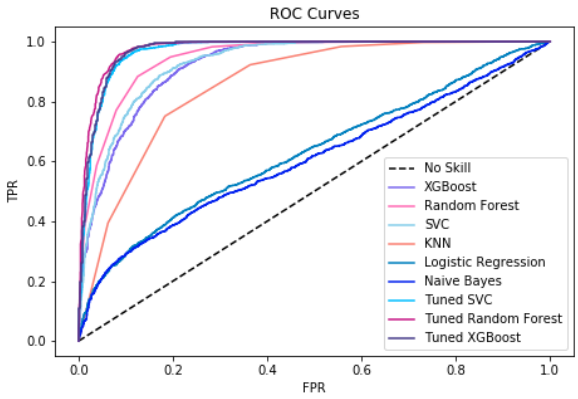

FIGURE 7. ROC curves for each of the models compared in this study. ROC is a probability curve and AUC represents a measure of separability between classes.

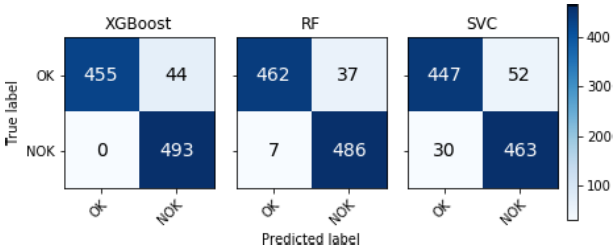

FIGURE 8. Confusion matrices for the holdout validation. The tuned XGBoost model achieved perfect recall on cars predicted over the three days after the last sample from the original data set.

suggest that the models were capable of generalizing well, being able to accurately predict the occurrence of defects in real car samples outside of the original data and thus provide important support in the earlier identification of deviations in the assembly line.

V. LIMITATIONS OF THE APPROACH

One possible barrier to the success of such a predictive solution in the long term is the possibility of drastic changes in the underlying distributions of the dimensional characteristics of the cars. This can happen for instance due to a change in the materials' suppliers or the replacement of parts in the stations before the first stage considered in this study. This is typically known as *Concept Drift*, referring to the change in relationships between the input and output data of the underlying problem over time [23].

A possible solution in the occurrence of this case during production would be through online monitoring and/or training of the models using for instance an architecture similar to the one showcased in Figure 9 based on the IDARTS framework [6].

For such an architecture, a Multi-Agent System (MAS) can be used to implement the CPPS that abstracts the MMP with one agent associated to the framing stage and another to the final one. While the framing agent can request quality predictions from a server hosting the deployed classifier and alert operators as defects are identified, the other can check for the ground truth associated with the measurements taken at the end of the line. These values can be stored

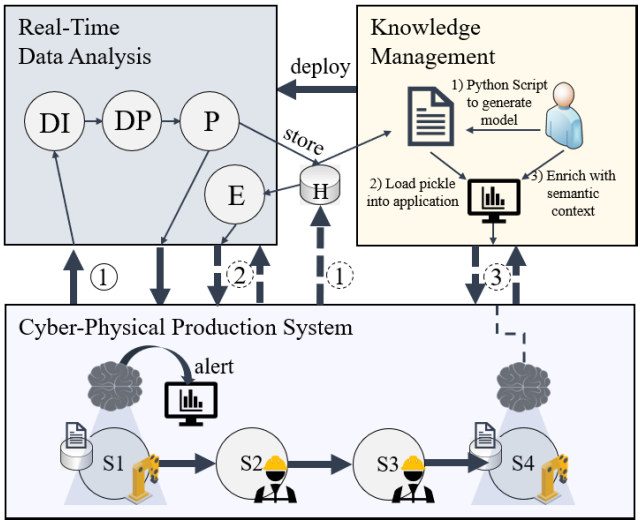

FIGURE 9. Possible deployment architecture based on the IDARTS framework [6]. Legend: DI - Data Ingestion; DP - Data Preprocessing; P - Prediction; E- Evaluation; H - Historical database; 1 - Prediction request; 1 (dashed) - Store ground truth; 2 (dashed) - Request mode evaluation; 3 (dashed) - Request updated model.

in a historical database, with the agent either periodically requesting a re-evaluation of the model to trigger a re-fit if the performance goes below a given threshold, or having the model be periodically updated using the static model as a starting point, for models that support such a functionality. This can be more efficient than the first approach, as it reuses the existing state instead of discarding it, only updating it on the most recent historical data.

The usage of a MAS also enables the system to adapt to other changes in run-time, including for instance the addition or removal of elements from the line during production without requiring additional programming effort or downtime. This means that for instance handheld smart inspection tools can be added in to provide additional measurements for the stages in between with the system being able to automatically enact a self-organized response and accommodate such devices and new data into the existing solution.

VI. DISCUSSION AND CONCLUSIONS

In this study we have addressed the application of an ML-based solution for multistage quality control. The performance of several binary classification models was evaluated and validated on data from a real automotive multistage assembly line within the Volkswagen AutoEuropa plant, encompassing two automated measurement stages on each end with human operated assembly and pre-alignment stages in between.

The analysis of this MMP is particularly challenging due to the amount of variability introduced by the human operators in the loop, responsible for the alignment and inspection of the assembled cars. However, the results suggest that there are certain dimensional variations in the early stages (even those within specification) that can be used to predict deviations at the end of the line regardless of these interventions,

indicating that some of these feature interactions are considerably hard to detect without the assistance of more complex data analytics approaches like the one being proposed.

While domain expert knowledge is critical for the correct assessment of the corrective actions that need to be carried out during the assembly operations (i.e. offsetting the jig), such an approach can provide further insights to enable an earlier intervention in the framing stages to prevent the propagation of the defects downstream, as well as a quicker identification of problematic cars for the final assembly.

We showed that non-linear algorithms like XGBoost and RFs are capable of detecting the complex relationships encompassed in this multivariate data set, providing quality estimations with a high capacity to distinguish between OK and NOK cars in an automotive multistage assembly process with high recall. We validated this results on two different test sets, one pertaining to the original data set and the other containing samples collected over the course of the 3 months following the last sample from the original data set. On both we show that the selected models are capable achieving high performance across all the evaluation metrics considered in this study, namely accuracy, recall, precision, F1 score and AUC.

Limitations and possible obstacles to the long term success of the predictive approach presented in this work were also discussed, more concretely in regards to the detection of concept drift, with possible solutions and venues for future research having been proposed in Section V. Overall we consider that the approach shows real potential in contributing towards the improvement of existing quality control strategies, with results hinting that reliable predictions can be provided to assist domain knowledge experts in making informed decisions towards the mitigation of defect propagation in multistage assembly scenarios.

ACKNOWLEDGMENT

The authors would like to thank Pedro Escorcio and Luis Peralta from Volkswagen AutoEuropa for their precious assistance with all matters pertaining to the acquisition and interpretation of the data related to the case study.

REFERENCES

[1] Y. Ding, D. Ceglarek, and J. Shi, "Fault diagnosis of multistage manufacturing processes by using state space approach," *J. Manuf. Sci. Eng.*, vol. 124, no. 2, pp. 313–322, 2002.

[2] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster, *Recommendations for implementing strategic initiative INDUSTRIE 4.0: Securing future German Manuf. industry; final Rep. Industrie 4.0 Work. Group*. Forschungsunion, Berlin, Germany, 2013. [Online]. Available: https://www.din.de/blob/76902/e8cac883f42bf28536e7e8165993f1fd/recommendations-for-implementing-industry-4-0-data.pdf

[3] J. Lee, E. Lapira, B. Bagheri, and H.-A. Kao, "Recent advances and trends in predictive manufacturing systems in big data environment," *Manuf. Lett.*, vol. 1, no. 1, pp. 38–41, Oct. 2013.

[4] D. Djurdjanovic, J. Lee, and J. Ni, "Watchdog agent-an infotronics-based prognostics approach for product performance degradation assessment and prediction," *Adv. Eng. Informat.*, vol. 17, nos. 3–4, pp. 109–125, 2003.

[5] J. Lee, J. Ni, D. Djurdjanovic, H. Qiu, and H. Liao, "Intelligent prognostics tools and e-maintenance," *Comput. Ind.*, vol. 57, no. 6, pp. 476–489, Aug. 2006.

[6] R. S. Peres, A. D. Rocha, P. Leitao, and J. Barata, "IDARTS—Towards intelligent data analysis and real-time supervision for industry 4.0," *Comput. Ind.*, vol. 101, pp. 138–146, Oct. 2018.

[7] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, Jan. 2016.

[8] G. Köksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13448–13467, Sep. 2011.

[9] D.-S. Kwak and K.-J. Kim, "A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2590–2596, Feb. 2012.

[10] D. Kim, P. Kang, S. Cho, H.-J. Lee, and S. Doh, "Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing," *Expert Syst. Appl.*, vol. 39, no. 4, pp. 4075–4083, Mar. 2012.

[11] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Trans. Ind. Informat.*, vol. 11, no. 3, pp. 812–820, Jun. 2015.

[12] J. Hebert, "Predicting rare failure events using classification trees on large scale manufacturing data with complex interactions," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 2024–2028.

[13] D. Wu, C. Jennings, J. Terpenny, R. X. Gao, and S. Kumara, "A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests," *J. Manuf. Sci. Eng.*, vol. 139, no. 7, 2017, Art. no. 071018.

[14] Ó. Martín, M. Pereda, J. I. Santos, and J. M. Galán, "Assessment of resistance spot welding quality based on ultrasonic testing and tree-based techniques," *J. Mater. Process. Technol.*, vol. 214, no. 11, pp. 2478–2487, Nov. 2014.

[15] M. Syafrudin, G. Alfian, N. Fitriyani, and J. Rhee, "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing," *Sensors*, vol. 18, no. 9, p. 2946, Sep. 2018.

[16] K. Chen, H. Chen, L. Liu, and S. Chen, "Prediction of weld bead geometry of MAG welding based on XGBoost algorithm," *Int. J. Adv. Manuf. Technol.*, vol. 101, no. 12, pp. 2283–2295, 2018.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[18] E. Jabbar, P. Besse, J.-M. Loubes, N. B. Roa, C. Merle, and R. Dettai, "Supervised learning approach for surface-mount device production," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.* New York, NY, USA: Springer, 2018, pp. 254–263.

[19] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[21] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 616–623.

[22] M. A. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using K- nearest neighbor and genetic algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, Jul. 2013.

[23] I. V. Z. E. Liobait, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications," in *Big Data Analysis: New Algorithms for A New Socity*. New York, NY, USA:Springer, 2016, pp. 91–114.

RICARDO SILVA PERES was born in Lisbon, Portugal, in 1991. He received the M.Sc. degree in electrical and computer engineering from the Nova University of Lisbon, in 2015, where he is also concluding the Ph.D. degree. Since 2014, he has been a Researcher with UNINOVA–Centre of Technology and Systems focusing on the development of predictive manufacturing systems. He has participated in several national and international research projects including FP7 PRIME, H2020 PERFoRM, H2020 OpenMOS, and H2020 GO0D MAN. He has authored over a dozen publications in high-ranked international scientific journals and conference proceedings (peer-reviewed). His research interests include predictive manufacturing, cyber-physical systems, artificial intelligence, and multi-agent systems. He has also been a member of the IEEE IES Technical Committee on Industrial Agents, since 2018.

JOSE BARATA received the Ph.D. degree in robotics and integrated manufacturing from the Nova University of Lisbon, 2004. He is a Professor with the Department of Electrical Engineering, Nova University of Lisbon, and a Senior Researcher with the UNINOVA Institute. He has participated in more than 15 international research projects involving different programmes (NMP, IST, ITEA, ESPRIT). Since 2004, he has been leading the UNINOVA participation in EU projects, namely EUPASS, self-learning, IDEAS, PRIME, RIVERWATCH, ROBO-PARTNER, and PROSECO. His main research interests are in the areas of intelligent manufacturing with particular focus on complex adaptive systems, involving intelligent manufacturing devices. In the last years, he has participated actively in the research of SOA-based approaches for the implementation of intelligent manufacturing devices (e.g., within the Inlife project). He has published over 100 original papers in international journals and conferences. He is a member of the IEEE Technical Committees on Industrial Agents (IES), Self-Organisation and Cybernetics for Informatics (SMC), and Education in Engineering and Industrial Technologies (IES). He is also a member of the IFAC Technical Committee 4.4 (Cost Oriented Automation).

PAULO LEITAO received the M.Sc. and Ph.D. degrees in electrical and computer engineering, both from the University of Porto, Portugal, in 1997 and 2004, respectively. He joined the Polytechnic Institute of Braganca, in 1995, where he is a Professor with the Department of Electrical Engineering and Coordinator of CeDRI (Research Centre in Digitalization and Intelligent Robotics). His research interests are in the field of industrial informatics, intelligent and reconfigurable systems, cyber-physical systems, the Internet of Things, distributed data analysis, factory automation, multi-agent systems, holonic systems, and self-organized systems. He participates/has participated in several national and international research projects (EU FP7 and H2020) and networks of excellence. He has published four books and more than 200 papers in high-ranked international scientific journals and conference proceedings (peer-review). He is the coauthor of three patents and received four paper awards at INCOM'06, BASYS'06, IEEE INDIN'10, and INFOCOMP'13 conferences. He served as a General Co-Chair of several international conferences, namely IEEE INDIN'18, SOHOMA'16, IEEE ICARSC'16, HoloMAS'11, and IFAC IMS'10. He is a Senior Member of the IEEE Industrial Electronics Society (IES) and Systems, Man, and Cybernetics Society (SMCS), past Chair of the IEEE IES Technical Committee on Industrial Agents, and member at-large of the IEEE IES Administrative Committee (AdCom). He is currently the Chair of the IEEE Standards Association P2660.1 Working Group.

GISELA GARCIA received the degree in environmental engineering from Technical Superior Institute, Technical University of Lisbon, Portugal, and the post-graduate degree in industrial engineering from Faculty of Science and Technology of New University of Lisbon, Portugal. She has been working in the automotive industry since 2005 as Lean Manufacturing Specialist and as Project Manager. Her current main activities are related with R&D and Innovation Project Management, Government Incentives Management, and Industry 4.0 Activities Coordinator. Her past main activities were related with shop floor management, KPIs system management, and product change management.