Shapley Values based Regional Feature Importance Measures Driving Error Analysis in Manufacturing

Valentin Göttisheim, Holger Ziekow, Ulf Schreier, Alexander Gerling
Furtwangen University
78120 Furtwangen, Germany
email: {valentin.goettisheim, holger.ziekow, ulf.schreier, alexander.gerling}@hs-furtwangen.de

***Abstract* – Data driven manufacturing quality management using machine learning for error detection can leverage predictive models for error analysis. Quality engineer experts evaluate the models input and interpret important features in the context of the specific manufacturing domain. In this paper, we propose three heuristics to determine the importance of features leading to actionable insights for error analysis. All proposed metrics are illustrated on synthetic data and evaluated on a real-world dataset.**

Keywords – manufacturing quality management; error analysis; feature importance; Shapley Values; xAI; machine learning.

I. Introduction

Quality management in modern manufacturing processes involves extensive testing and collection of detailed measurements along production lines. This provides the basis for data driven error analysis. However, quality managers struggle with finding error causes in large sets of quality data [1]. Artificial Intelligence (AI) methods can help to analyze such data and predict errors [2]. In combination with eXplainable AI (xAI), predictive models can further put quality engineers on the right track for finding causes of production errors. That is, feature importance metrics can reveal features that hint at error causes [3]. However, well known feature importance measures are not tailored to this task. In this paper, we expand our earlier work [4] and introduce new feature importance measures which are designed to reveal features of interest for quality management in manufacturing.

Our work is rooted in a research project with a German manufacturer [5]. Here, we found that combining human expertise with AI-based data analysis is desirable for error analysis in production lines. This is because (a) quality managers seek to understand the error causes and may not blindly trust AI-based results and (b) human experts have background knowledge and a deep understanding of the production process, that the AI does not have access to. Hence, this work explicitly keeps the human experts in the loop and focuses on using AI models for providing input to human analysts.

This work targets typical manufacturing setups, where production lines comprise a sequence of production steps and several test stations along the production line. Test stations perform measurements on each product at different steps of the production. This leads to detailed records of individual product instances that can include hundreds of thousands of measurements per product [6]. However, the high number of different measurements poses challenges for finding causes of errors in the data. Another challenge in error cause analysis is that errors are rare [7]. Modern manufacturing processes are usually highly optimized and quality management is often about driving down rare – but still costly – errors. Yet, existing applications have successfully used such high dimensional and imbalanced data to build AI models for predicting production errors [6][8]. The aim of such models is to take measurements from test stations early in the production sequence and predict errors that occur downstream in the production line. If errors can be predicted early with sufficient reliability, products can be removed early in the process and costs for downstream production steps can be avoided [2].

Furthermore, such AI models can be analyzed to hint at the cause of errors. We leverage this to provide insights to human experts in quality management. Existing works use feature importance measures to identify quality measurements that are relevant in predicting and explaining errors. For example, if a heat measurement of an oven is important in predicting errors, then errors may be avoided by adjusting the temperature setting. Identifying such interesting measurements amongst the thousands of data points can help quality managers to find error causes and improve production [9]. However, existing importance measures are not tailored to find features that are interesting for inspection in error cause analysis. Instead, they take a global view and capture how much a model relies on a given feature on average. As we show in this paper, such a global view is often not useful when it comes to spotting rare but strong relations that lead to actionable insight in error analysis.

In contrast to global importance measures, xAI methods like Shapley Additive Explanations (SHAP) [10] and LIME [12] provide local explanations for the impact of features on a prediction. That is, the impact of features can be estimated for individual data instances. However, analyzing a data instance in isolation is of little use for quality management in manufacturing. That is, a single data instance is not enough to draw conclusions for actionable insights.

With this work, we provide feature importance measures that are conceptually in between global and local feature importance. We refer to this as regional feature importance. With this concept we setup on and expand our earlier work [4]. That is, we analyze sets of local feature importance values for interesting effects. The result of this analysis is captured in new importance measures that capture different interesting aspects. In this paper, we mathematically define our applied notion of interestingness. Intuitively, we consider a feature interesting if it hints at actionable insight for quality managers. Such an action could be setting a threshold in a quality check or adjusting the process to avoid certain value ranges. Intuitively, drastic changes in error rates and high error rates in well-defined parts of a value range make features interesting. In this paper, we provide importance measures that formally capture such notions of interestingness and map them to an importance score. In summary, we make the following key contributions:

1) We introduce and formally define novel feature importance measures that are tailored to find relevant features in manufacturing quality management.
2) We test and illustrate the benefits of our proposed measures with synthetic data.
3) We evaluate the proposed measures on real-world data and compare them with established importance measures.

With these contributions, we help human experts in quality management better leverage results from AI models for driving their analysis.

The remainder of this paper is structured as follows. In Section II, we briefly summarize the corresponding background. In Section III, approaches to derive the regional feature importance are proposed which then are evaluated in a real-world dataset in Section IV. In Section V, we discuss related work and conclude in Section VI.

II. Background

When using Machine Learning (ML) support for error analysis in quality management processes, feature importance metrics can become a tool to rank and identify features that are suitable to guide Quality Engineers (QE) in finding error causes in production. Such a process inspired the present work is carried out in the production of an industry partner in the research project [4]. ML-driven quality management processes here focus on QEs as primary actors. Using ML support, QEs are intended to analyze production and take corrective maintenance steps in production. However, the development and deployment of models for the ML support system are embedded in automated pipelines and maintained by data scientists. The automated ML pipeline includes several steps like data preprocessing, i.e., feature selection or evaluation of model performances through cost-sensitive metrics [2]. As such, the system is designed to enable QEs to use ML support for error causes analysis, but not to be engaged with the technical depth of the ML system.

A reference process focusing on QEs intended to investigate errors in production is laid out in [1]. Key steps include the selection of production data for the automated ML pipeline. Later steps involve error identification and correction in production using ML support. To identify error causes the QE is intended to use feature importance to find features that suits as explanations for error causes.

SHAP is one of the more recent advancements in the field of xAI, focusing on the interpretability of ML models. SHAP targets instance-based as opposed to global model explanation. By aggregating explanations of instances, it is possible to evaluate the importance of features incorporating aspects of interest to guide QEs in error cause analysis [3]. SHAP evaluates the marginal contribution a feature has on its model output. The contribution $\phi_f \in \mathbb{R}$ for a feature f with model m is attributed using Shapley Values from game theory:

$$\phi_f = \sum_{S \subseteq N \setminus \{f\}} \frac{|S|!\,(M-|S|-1)!}{M!} [m_x(S \cup \{f\}) - m_x(S)],$$

where M is the number of all features, S is the set of input values, and |S| is the magnitude of S (for example, $S = x_1, x_2, \dots, x_{f-1}, x_{f+1}, x_n$ and $|S| = n-1$). To compute the feature contribution usually the explanation model $e(z') = \phi_0 + \sum_{f=1}^{M} \phi_f z'_f$ where $z' \in \{0,1\}^M$ is used. This is the weighted average over all feature contributions. The explanation model is computed using the mapping $m_x(S) = m(e_x(z'))$ which maps all input values S to whether the feature is being used ($z' = 1$) or not known ($z' = 0$).

In an earlier work [4], we already picked up the idea of Shapley Value based heuristics to determine importance measures leading to helpful insights for quality engineering. For completeness, the main ideas are briefly summarized below.

Max-SHAP: The Max-SHAP heuristic focuses on the maximal SHAP values and ranks the features according to their maximum SHAP value. The intuition behind this metric is that a feature showing a high SHAP value for an error instance is a good explanation. Formally, Max-SHAP is defined as:

$$Max\ SHAP_f(m,S) = \max\{\phi_f(m,x) | x \in S\}$$

Max-Main: This metric focuses similarly to the Max-SHAP on maximal values but in contrast, the maximum of the SHAP main effects is assessed. The intuition is that a feature with high main effects is considered a good explanation. SHAP main effects do not include interactions between features and therefore are the single most simple explanation for an error case. The Max-Main metric is defined as:

$$Max\ Main\ Effect_f(m,S) = \max\{\phi_f(m,x) - \sum_{j \neq f} \phi_{f,j}(m,x)| \,|x \in S\}$$

Range-SHAP: Considering a change in SHAP values over the feature value range, the feature with a bigger change is considered more interesting. The intuition is that features with varying contributions are more likely to indicate error cases. The Range-SHAP metric is defined as:

$$Range\ SHAP_f(m,S) = \max\{\phi_f(m,x)|x \in S\} - \min\{\phi_f(m,x)|x \in S\}$$

III. Aggregations of SHAP values to assess regional feature importance

We aim to identify features that are helpful in finding error causes in production. Production errors are rare events because of the optimized production process. Many error occurrences are of random nature. However, some have a clear cause which is captured in the data. Features that reflect these error events are hints to causes and therefore should be rewarded with high importance scores. The fundamental idea is a scoring-function g: g(*f, X, ...*)→ ℝ that aggregates SHAP values and scores „interesting“ features high. Here, we refer to f as the target feature used in the analysis with the dataset X, and ... represents the use of additional parameters which are unique to each approach later proposed. We further refer to $\phi_f(x)$ as SHAP values of the feature f computed on the data instance $x \in X$ and define $\bar{\phi}_f(X) = \frac{1}{|X|}\sum_{x \in X} \phi(x)$ as the SHAP value mean.

In the following, three approaches to assess the importance score g are proposed. For each approach first, the basic concept is described and then the idea is illustrated using synthetic sample data where the ground truth is known. In all illustrations, the error cause can be attributed to features A, B or random occurrences affecting 1% of the data. We then train an XGBoost classifier [12] for error prediction on the 110,000 data points and compare the importance metrics g with the classic state-of-the-art importance measures implemented in the XGBoost library (v1.3.3). All illustration results are listed in Tables 1 to 3 and reported as the rank of the feature with mean score and standard deviations (mean/standard deviation) in brackets. Mean and standard deviations are based on the 15 repetitions the illustrations were conducted.

A) Outlier-Approach

The intuition behind this approach is that features with abnormal SHAP values are potentially interesting. Therefore, we perform anomaly detection on the distribution of SHAP values and assess the SHAP values for abnormal data points. For simplicity, we assume that the SHAP values approximately form a normal distribution. The outliers then can be detected using mean and standard deviation where the less interesting SHAP values are around the expectation value. However, other anomaly detection methods may be used as alternatives. Assuming a normal distribution, we detect increased SHAP values by determining the SHAP values outside of the normal distribution with at least λ standard deviations $\sigma(\phi_f(X))$ above the SHAP value mean $\bar{\phi}_f(X)$. The importance score g is formally defined as the sum over all outliers $\mathrm{outl}(\lambda,\ \mathrm{X}) = \{\, x \in X \mid \phi_f(x) \geq \bar{\phi}_f(X) + \lambda\, \sigma(\phi_f(X))\}$ as:

$$g(f, X, \lambda\,) = \sum_{x' \in\, \mathrm{outl}(\lambda, \mathrm{X})} \phi_f(x')$$

Note, $\phi_f(x')$ refers to the SHAP values for feature f computed on data instances x' where the set $\mathrm{outl}(\lambda,\ \mathrm{X})$ only includes outliers λ standard deviations above the SHAP value mean. The concept is that we consider high SHAP values as interesting because high SHAP values provide a strong indication for errors and therefore hint at error causes.

To illustrate the Outlier-Approach, we now compare the approach with state-of-the-art feature importance measures implemented in the XGBoost library using synthetic data. Metric g is computed with $\lambda = 2$ considering SHAP values 2σ above the features SHAP value mean. As data, we consider the following dataset: Feature A causes a small number of errors within a small value range, i.e., an 8% error rate within a 0.025 quantile range. Feature B causes a decreasing error rate from 4% to 3% over the feature value in a 0.8 quantile range. Thus, feature B has a lower error rate but within a much broader range than feature A. In this situation, we consider feature A with its strong (albeit rare) relation to error events- as more interesting.

The illustration result listed in Table 1 shows the proposed approach Outlier Shap ranks feature A as more important while the classic importance measures do not. That is, all measures except Weight rank feature B higher than A. Note, Weight scores feature A and B very similar. Both features show a difference in mean scores of 19.8 and considering the standard deviations of more than 33.6 both scores of feature A and B are almost equal. Therefore, Weight does not rank feature A clearly higher. Thus, only the proposed Outlier Shap is reliable and detects the more important feature A correctly.

TABLE 1. ”ILLUSTRATION RESULT”: EVALUATED FEATURE IMPORTANCE SHOWING “OUTLIER SHAP” CORRECTLY RANKS FEATURE A AS MORE IMPORTANT – NOTATED: FEATURE (MEAN/STD).

Rank	**Weight**	**Gain**	**Cover**	**Total Gain**	**Total Cover**	**Avg. Abs. SHAP**	**Outlier Shap**
1	A (921.2/ 33.66)	B (5.0/ 0.21)	B (1065/ 63.5)	B (4566/ 141)	B (960039/ 50756)	B (0.77/ 0.02)	A (4660/ 137)
2	B (902.3/ 36.85)	A (3.33/ 0.07)	A (781/ 58.6)	A (3072/ 137)	A (720332/ 63755)	A (0.21/ 0.008)	B (429/ 161)

B) Micro-Average Approach

The intuition behind this approach is that features are of interest if they show high SHAP values within a small feature value range. Therefore, we divide the feature value range into equally sized partitions and determine the average SHAP value of each interval. The importance score g is then determined as the maximum average SHAP value of all intervals belonging to the given feature. Formally the importance score is defined as the maximal average SHAP value $\max\{\bar{\phi}_f(X_i)\}$ where $x \in X_i$ of interval i comprises the datapoints $X_\mathrm{i} = \{x \in X \mid (i * d) \leq x < (i * d) + d\}$ over the equally sized feature range $d = \frac{1}{n} * range(X_f)$ for the set of feature values X_f and given number of intervals n:

$$g(f, X, n) = \max\{\bar{\phi}_f(X_i) \mid i = 0, \ldots, n-1\}$$

Note, $\bar{\phi}_f(X_i)$ refers here to the mean SHAP value of interval i for given number of intervals n. We consider intervals with higher SHAP values as more important. The concept is that the higher the average SHAP value within an interval, the more important this interval is for the contribution to error events.

An illustration of this approach is an increased error ratio in a tail of a normal distributed feature. Consider the normally distributed $N(0,1)$ features A and B. Feature A causes a 10% error rate in the upper tail of the feature range. Feature B causes a slightly decreasing error rate of 4% to 3% from the 0.3 to the 0.7 feature value quantile. We argue that feature A – even though it explains fewer errors - is more interesting because it leads to more actionable insights. This type of error events is often caused by exceeding a threshold that could be checked in production without much effort.

For this illustration, metric g is computed with a number of intervals $n = 100$. The results listed in Table 2 show that the classical feature importance ranks the less interesting feature B as more important, whereas metric g the Micro-avg Shap correctly ranks feature A as more important. Considering the results for Weight with an absolute difference in scores of A and B of 25 and a standard deviation of 53 for B, both features A and B are attributed with similar importance, with B being ranked slightly higher.

TABLE 2. "ILLUSTRATION RESULT": EVALUATED FEATURE IMPORTANCE SHOWING "MICRO-AVG SHAP" CORRECTLY RANKS FEATURE A AS MORE IMPORTANT – NOTATED: FEATURE (MEAN/STD).

Rank	**Weight**	**Gain**	**Cover**	**Total Gain**	**Total Cover**	**Avg. Abs. SHAP**	**Micro-Avg Shap**
1	B (921/ 53.6)	B (4.75/ 0.168)	B (1029/ 64.5)	B (4362/ 177)	B (944996/ 48209)	B (0.8/ 0.02)	A (2.62/ 0.579)
2	A (896/ 35.9)	A (3.34/ 0.12)	A (722/ 53.0)	A (2993/ 151)	A (646955/ 46820)	A (0.22/ 0.01)	B (0.997/ 0.294)

C) Slope-Approach

The intuition behind this approach is that rapid changes in SHAP values are of interest. A rapid change of SHAP values within a small interval of feature values indicates a clear threshold at which interpretation as the cause of error events is possible. The rate of SHAP value change can be represented by a slope. Like the former approach equally sized intervals of a feature are constructed but then regression slopes on the means of SHAP values over multiple intervals are computed. To make the approach more precise, further conditions are imposed to accept the slope as admissible solution or the computed slope is not considered for the importance score. Formally, the importance score g is defined:

$$g(f,X,n,w,t) = max\{|slope(w_j)|\,|\, j = 0,\dots,n-1\}$$

The slope $slope(w_j) = \beta$ is computed over the rolling window $w_j = \{\bar{\phi}_f(\mathrm{X_j}),\dots,\bar{\phi}_f(\mathrm{X_{j+w}})\}$ of size w where β minimizes $\epsilon = \bar{\phi}_f(X_i) - i * \beta$ for $\bar{\phi}_f(X_i) \in w_j$. Here β represents the slope over the averaged shap values of multiple windows w_j. To compute β we used simple OLS regression. Also note that X_i for interval i is similar defined to the previous micro-average approach. The slopes $slope(w_j)$ is considered or discarded, i.e., set to zero, for some threshold t if either condition is fulfilled:

$$\exists\ i = j,\dots,j+w\colon q_{75}\left(\{\bar{\phi}_f(X_i) \in w_j\}\right) - q_{25}\left(\{\bar{\phi}_f(X_i) \in w_j\}\right) \geq t \qquad (1)$$

$$\max\left(\{\bar{\phi}_f(X_i) \in w_j |\ i = j,\dots,j+w\}\right) \geq t \qquad (2)$$

Condition (1), called Interquartile (IQR) shap variation, is chosen such that the mean shap values in the window w_j in between the 0.75 quantile q_{75} and the 0.25 quantile q_{25} has to be greater than or equal to given threshold t. For condition (2) the Max Shap variation the maximal mean shap value in the window w_j has to be greater than or equal to a threshold t. If the corresponding condition is fulfilled, the $slope(w_j)$ is accepted. Both conditions suppress steep local slopes covering a too small value range.

To illustrate this metric on a sample case, we consider a change in error rate within a small feature range. We assume two uniform distributed features A and B. Feature A causes an error rate of 15% in a 0.01 feature value quantile range. Feature B causes a decreasing error rate from 10% to 2% over the entire feature value range. Although feature B may be more important globally, we consider feature A as more important to identify actionable insights. To compute the importance score g, we set the numbers of intervals to $n = 300$ and the rolling window of size $w = 10$. The thresholds t are chosen for the IQR variation (1) as $t = 0.1$ and to $t = 0.4$ for the Max Shap variation (2). As such, the interquartile range or the maximum value respectively of the mean SHAP values over the intervals in window w_j has to be greater than or equal to t.

The illustration results listed in Table 3 show the classic importance measures rank feature B as more important. In contrast, both proposed Slope variations correctly rank feature A as more important. Observing Weight with an absolute difference in scores of 2 feature B is slightly ranked better than feature A. However, the values are very similar, given a standard deviation of about 37.

TABLE 3. "ILLUSTRATION RESULT": EVALUATED FEATURE IMPORTANCE SHOWING "SLOPE IQR SHAP" & "SLOPE MAX SHAP" CORRECTLY RANKED FEATURE A AS MORE IMPORTANT – NOTATED: FEATURE (MEAN/STD).

Rank	**Weight**	**Gain**	**Cover**	**Total Gain**	**Total Cover**	**Avg. Abs. SHAP**	**Slope IQR Shap**	**Slope Max Shap**
1	B (971/ 37.0)	B (4.41 /0.127)	B (1425 /56.1)	B (4278 /185)	B (1385121/ 84941)	B (0.362 /0.009)	A (0.163 /0.018)	A (0.162 /0.019)
2	A (969/ 37.1)	A (3.32 /0.097)	A (1078 /83.2)	A (3213 /155)	A (1043942/ 73457)	A (0.151 /0.009)	B (0.089 /0.014)	B (0.089 /0.014)

IV. EXPERIMENTS

With the aim to identify features that are interesting in error cause analysis, we conduct an experiment with the Secom dataset [13]. The real-world dataset originates from a semiconductor manufacturing process containing 591 features and 1667 instances of which 106 are error instances.

The learning problem is formulated as a binary classification task and a XGB gradient boosting model with default parameters used for training. For model training and the evaluation of importance metrics, we used the whole dataset. This may resemble an idealized case where the training data perfectly meets the data distribution during prediction but also prevents uncertainty being induced by the model for the evaluation of the importance metrics. The model achieved a perfect training score according to the f1 score (f1=1.0). The SHAP values are computed using the SHAP package (v0.40.0) [14]. The proposed importance metrics were computed using the same parameters described in the illustrations on the synthetic datasets for each proposed approach.

We use the experiments to compare existing and our proposed importance metrics and discuss exemplary features in detail along with their SHAP plots. Figures 1 to 4 show the

TABLE 4. "SECOM EXPERIMENT RESULTS": TOP 5 IMPORTANCE RANKINGS GROUPED COMPARISON WITH HIGHLIGHTED AGREEMENTS TO THE PROPOSED METRICS (GREY) AND UNIQUE FINDINGS (BOLD).

Rank	Classic Metrics					SHAP-based Metrics				Proposed Metrics			
	Weight	*Gain*	*Cover*	*Total Gain*	*Total Cover*	*Average Abs Shap*	*Max Main*	*Max Shap*	*Range Value*	*Slope Approaches*		*Micro-Avg Approach*	*Outlier Approach*
										Max Shap	*IQR Shap*		
1	F59	F210	F168	F59	F59	F59	F59	F59	F59	F59	F59	F64	F59
2	F333	F539	F429	F333	F64	F21	F64	F64	F64	F64	F64	F59	**F423**
3	F103	F29	F426	F64	F426	F333	F40	F426	F333	F429	F33	F103	F64
4	F2	F109	F100	F132	F121	F488	F426	F333	F103	F333	**F130**	F40	F333
5	F33	F304	F331	F33	F574	F103	F153	F40	F33	**F475**	F429	F121	F2

discussed SHAP plots. However, because of the limited space, not all SHAP plots of every feature can be shown. The SHAP plots have the feature values on the x-axes and the corresponding SHAP value on the y-axes. Visually distinguishable are error instances colored red and non-errors colored blue. Colored grey in the background are histograms of the corresponding feature values.

For the following discussion of the results, the metrics are grouped into (1) Classic, (2) SHAP-based and (3) our proposed importance metrics. Table 4 shows the resulting rankings. Highlighted in grey are the agreements of the corresponding group with the proposed metrics. At the first sight, the metrics Gain and Cover have the least overlapping results with the proposed metrics. Since both metrics are commonly used as importance measures this is surprising. On the other hand, the metric Weight has quite a lot in common with the proposed metrics but the previous illustrations on synthetic data already showed that Weight is not a reliable measure. In the following, agreements and disagreements over the grouped rankings of the top three features are discussed in more detail.

A) Agreements of importance

Across all groups of metrics, some features have identical importance. All groups assign similar importance to the features F59, F64, F333 and F103.

Features F59, F64 and 103 are shown in Figure 1 and are examples of interesting features. F59 has increased SHAP values on the right-hand side. A point of interest is the threshold of about 10 of the feature value where the SHAP values are increasing. This hints at a threshold that can be useful for error cause analysis and error explanation. An advanced understanding of such thresholds enables correction in production, like adjustment of parameters or setting machinery alarms. Further assessing the right-hand side of F59 shows high SHAP values, which is an indicator for a cause explanation. Note, this area of increased SHAP values also shows an increased error rate, rendering this an interesting feature. F64 shows these interesting properties too: A distinguishable threshold at which SHAP values are increasing, strong effects, i.e., high SHAP values, and a higher error ratio in the area of increased SHAP values.

Ranked as equally important across all groups – with minor differences – are F103 and F333. Both features are of interest and show equivalent properties as described above. Due to space limitations, just F333 is shown in Figure 1.

Out of the Classic feature importance, F429 (Figure 2) is interesting. It is ranked second by Cover and in the top 3 of the proposed metrics. It shows a moderate effect (more than 0.8) and a clear threshold at which the SHAP values are increasing. This threshold defines the area of an increased error ratio, rendering the feature interesting.

F40 (Figure 2) is ranked as identically important both by the SHAP-based and by the proposed metrics. It has high effects and a good error ratio in an area of increased SHAP values. Note that the feature comprises only a few instances with high effects. It depends on the cost structures and specifics of the production process whether the amount justifies the considerations for a deeper error cause analysis. However, we argue that the observation hints at an interesting phenomenon.

B) Disagreements of importance

Besides commonalities, there are also differences in rankings which are discussed next. Comparing the Classic with the proposed metrics, major differences are revealed by Gain and Cover. Here, we focus on the features F210 and F539 shown in Figure 3 at the bottom of the next page, and F29 as well. Gain ranks all three features high, but we argue that these features are of less interest to determining cause

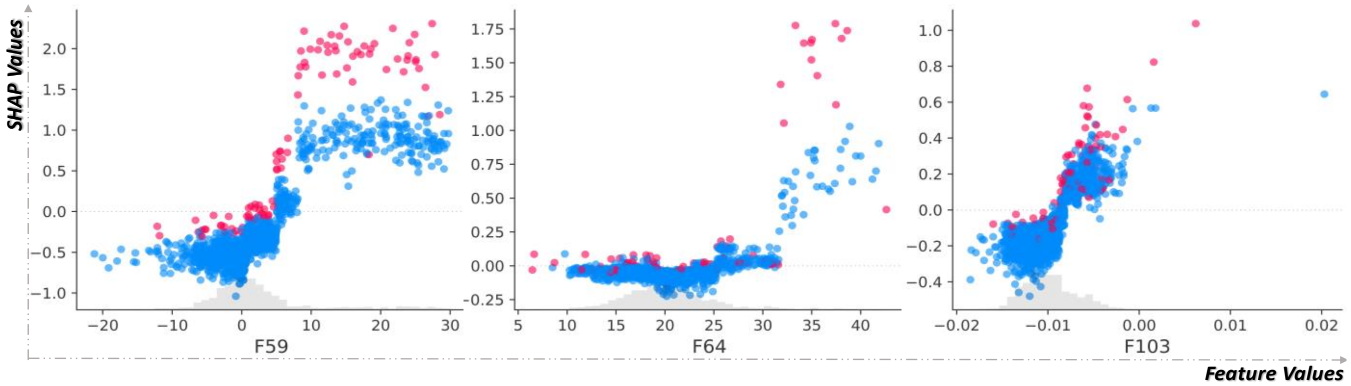

Figure 1. Agreements between importance measures: SHAP plots of features F59, F64 and F103 which have similar rankings across all importance measures and show interesting patterns for cause anlysis – errors-instances (red) and non-errors (blue).

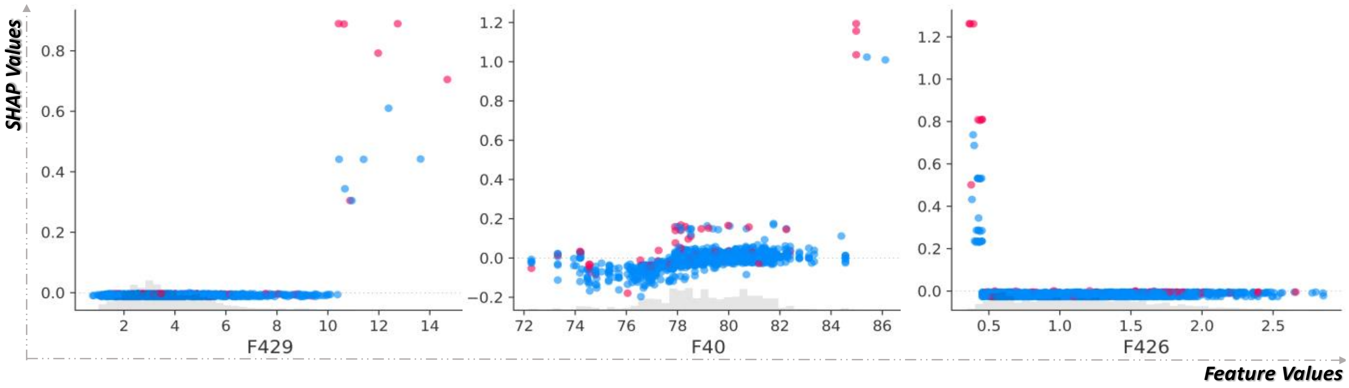

Figure 2. Agreements between importance measures ff: SHAP plots for features F429, F40 with similar rankings over all importance measures. Except feature F426 with low mean average effect falls behind in importance of the proposed metrics – errors-instances (red) and non-errors (blue).

explanations. F210 has an overall weak effect, i.e., the maximal SHAP value is around 0.1. It is also not possible to determine a clear threshold at which SHAP values are increasing, nor to specify a clear threshold of higher error rates. F539 shows a few instances with increased error rate but also with weak effect, i.e., a SHAP value of 0.3 and therefore is not interesting enough for further investigations. Feature F29 is not shown here due to the space limitations but has a similar appearance as both features described above and thus is of less interest.

F168 is shown in Figure 3 and ranked first by Cover, but not ranked in the top 5 by any of the other groups. It shows an area of increased SHAP values with a suitable error ratio but the effect, i.e., of 0.35 is quite low and therefore we argue that this feature also is of low interest.

Feature F426, shown in Figure 2, has interesting properties and ranked as important by the Classic importance but is not ranked in the top five using the proposed metrics. F426 shows a step increase in SHAP values on the left-hand side with a maximal effect of 1.2 which looks sufficient at first glance. Further examinations showed that F426 was ranked seventh by the proposed Slope Max Shap approach. Investigating the region of increased SHAP values revealed also that F426 has a few high effects but the average effect for this region is quite low. This indicates over plotting of the SHAP plot and relativizes the interestingness of F426.

Comparing SHAP-based and the proposed metrics, feature F21 stands out as it ranked as important in the group of SHAP-based measures but neither by the Classic nor by the proposed metrics. F21 (not shown) has a decent effect of 0.6 and a steep increase in SHAP values. Yet, it does not have an area with a high error rate and falls behind in the proposed metrics.

C) *Highlights of proposed metrics*

F423, F475 and F130, as shown in Figure 4, are only ranked as important by the proposed metrics. F423 has strong positive effects, clear threshold upon which SHAP values increase and a relatively high error ratio in in a specific area. We therefore argue that the feature is interesting. F475 has strong effects, a clear threshold where effects increase, and a good error ratio. Therefore, we argue that it is also interesting for further investigations. F130 shows negative SHAP values in the bottom left corner which indicates a value range where the error rate is much lower. It is also possible to determine a threshold at which SHAP values decrease to a zero effect. This might hint at means to prevent errors and we argue that the feature is therefore interesting.

Overall, the experiments support the usefulness of the proposed metrics. They show that high ranked features have properties that are interesting for a cause analysis of production errors. Furthermore, we have shown that established metrics may rank features high, that do not have such interesting properties. In contrast, the proposed metrics rank features high that are interesting but considered not important by existing metrics. This demonstrates that cause analysis in manufacturing can benefit from the proposed metrics.

V. RELATED WORK

Root cause analysis in the production environment has been well studied [15] and several methods for model

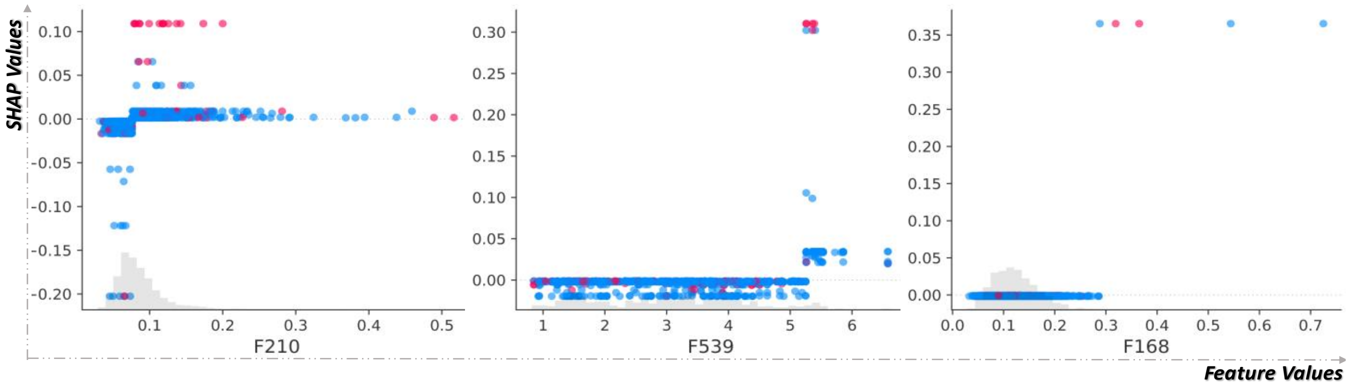

Figure 3. Disagreements between importance measures: SHAP plots for features F210, F539 and F168 ranked by the Classic importance measures as important – errors-instances (red) and non-errors (blue).

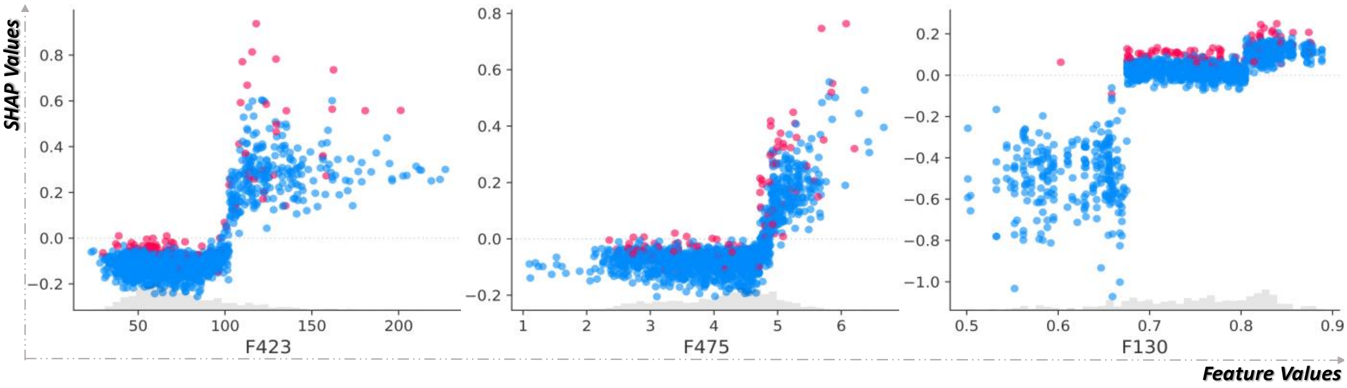

Figure 4. Important features according to proposed metrics: SHAP plots for features F423, F475 and F130 ranked by the proposed importance measures as important – errors-instances (red) and non-errors (blue).

interpretability through xAI have been reported [16]. However, we argue that the proposed metrics are more related to feature importance measures. The metrics may be used in root cause analysis to incorporate expert knowledge. Applied xAI in the manufacturing domain is used to extract explanations from a machine learning model to, e.g., enhance trust in the model, used for model optimization or to assist domain experts taking actions according to the model insights. In [17], saliency maps and class activation maps are extracted from a deep learning model. In [3], the authors use an isolation forest as model to determine normal production line behavior and feature importance to explain the model. Mehdiyev and Fettke apply local and global explanations to examine the impact of different views on the generated insights [18]. However, neither work addresses the problem of which feature provides the most promising insights given the possible tremendous feature space and the corresponding effort required to examine all explanations. To the best of our knowledge, we are the first to provide SHAP-based importance measures tailored to the task for quality management. Lundberg et al., introduced the idea of SHAP-based feature importance [14]. To determine a feature’s overall effect, the absolute SHAP value across all considered instances is averaged and thus a global importance measure. In contrast, our proposed measures just consider instances that possibly encompass interesting properties for quality management. Other global importance measures used in the domain have a broad history. A detailed description of the following global importance measures is laid out by Molnar [19]. In [20], Permutation Feature Importance is introduced. A global measure of where the features are perturbated and the resulting performance loss of the model is taken as a measure of the features importance. Mehdiyev and Fettke [18] used Individual Conditional Expectation (ICE) [21] as the global importance. Also possibly used are Partial Dependence Plots (PDP) [22]. However, neither ICE nor PDP accumulates a single importance score. Both are used as visualizations of global model behavior. Overall, one of the most influential global importance measures is the Gini index [23]. According to Lundberg [24], the Gini index is equivalent to the in XGBoost [12] implemented importance measure Gain which *uses the average training loss reduction gained when using a feature for splitting*. Lundberg [24] also describes Weight as *the number of times a feature is used to split the data across all trees* and Cover as *the number of times a feature is used to split the data across all trees weighted by the number of training data points that go through those splits.* Both the total importance scores used for comparison are described in the XGBoost documentation [25] for Total Gain as the *total gain across all splits the feature is used in* and Total Cover as *the total coverage across all splits the feature is used in.* For local feature importance, LIME [11] could also be considered. However, to compute explanations LIME uses sampling which is not restricted to solely interesting areas.

VI. CONCLUSION

In this paper, we have introduced regional feature importance measures that aim at identifying interesting features for quality management in manufacturing. We discussed the underlying notion of interest and provided corresponding formal definitions. Conceptually, our importance measures are between established global and local feature importance measures and highlight regional effects which are helpful in finding production error causes. We illustrate the usefulness of the new measures through experiments using synthetic and real-world data.

Our experiments show that the prosed measures successfully elicit features that – based on our experience [5] – are interesting, but are missed by established methods. Therefore, we conclude that quality managers benefit from adding our proposed importance measures to the pool of xAI methods. We thereby improve xAI for error prediction models in manufacturing. With the help of our importance measures, quality managers get hints about interesting relations that are reflected in the prediction model and drive deeper analysis accordingly.

Subject to future work are questions about the integration of the importance measure in the machine learning pipeline. In this work, we assume that the measures are applied at the end of the pipeline, potentially after automated feature engineering, and model optimization. However, the proposed measures may drive the analysis of features earlier in the pipeline as well. Furthermore, future work may expand on the idea of providing importance measures between global and local measures. With our work, we have presented several heuristics which follow this concept for capturing interesting patterns in features.

REFERENCES

[1] C. Seiffer, A. Gerling, U. Schreier and H. Ziekow, "A Reference Process and Domain Model for Machine Learning Based Production Fault Analysis", *Enterprise Information*

Systems: 22nd International Conference, ICEIS 2020, Springer International Publishing, pp. 140–157, 2021.

[2] A. Gerling et al., "Comparison of algorithms for error prediction in manufacturing with automl and a cost-based metric", *Journal of Intelligent Manufacturing*, 33.2022(2), pp. 555–573, 2022.

[3] M. Carletti, C. Masiero, A. Beghi and G.A. Susto, "Explainable Machine Learning in Industry 4.0: Evaluating Feature Importance in Anomaly Detection to Enable Root Cause Analysis", *IEEE International Conference 2019*, pp. 21–26, 2019.

[4] H. Ziekow, U. Schreier, A. Gerling and A. Saleh, "Interpretable Machine Learning for Quality Engineering in Manufacturing - Importance Measures that Reveal Insights on Errors", *The Upper-Rhine Artificial Intelligence Symposium, UR-AI 2021, Artificial Intelligence - Application in Life Sciences and Beyond*, Germany, Kaiserslautern: Hochschule Kaiserslautern, University of Applied Sciences, pp. 96–105, October 2021.

[5] H. Ziekow et al., "Proactive Error Prevention in Manufacturing Based on an Adaptable Machine Learning Environment", *Artificial Intelligence: From Research to Application: The Upper-Rhine Artificial Intelligence Symposium UR-AI 2019*, Offenburg, Germany, Karlsruhe: Hochschule Karlsruhe - Technik und Wirtschaft, pp. 113–117, March 2019.

[6] R. S. Peres, J. Barata, P. Leitao and G. Garcia, "Multistage Quality Control Using Machine Learning in the Automotive Industry", *IEEE Access*, vol. 7, pp. 79908–79916, 2019.

[7] Y. Wilhelm, U. Schreier, P. Reimann, B. Mitschang and H. Ziekow, "Data Science Approaches to Quality Control in Manufacturing: A Review of Problems, Challenges and Architecture", *Symposium and Summer School on Service-Oriented Computing*, Springer, pp. 45–65, 2020.

[8] A. Gerling et al., "Results from using an Automl Tool for Error Analysis in Manufacturing", *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume* 1, pp. 100–111, 2022.

[9] C. Seiffer, A. Gerling, U. Schreier and H. Ziekow, "A Reference Process and Domain Model for Machine Learning Based Production Fault Analysis", *Enterprise Information Systems: 22nd International Conference, ICEIS 2020*, Springer International Publishing, pp. 140–157, 2021.

[10] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, NY, USA, pp. 4768–4777, 2017.

[11] M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144, 2016.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, pp. 785–794, 2016.

[13] D. Dua and C. Graff, "UCI Machine Learning Repository", Irvine, CA: *University of California, School of Information and Computer Science*, 2019. [Online]. Available from: http://archive.ics.uci.edu/ml.

[14] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees", *Nature machine intelligence, 2(1)*, pp. 56–67, 2020.

[15] E. Oliveira, V. L. Miguéis and, J. L. Borges, "Automatic root cause analysis in manufacturing: an overview & conceptualization", *Journal of Intelligent Manufacturing*, 2022.

[16] G. Sofianidis, J. M. Rožanec, D. Mladenić and D. Kyriazis, "A Review of Explainable Artificial Intelligence in Manufacturing", *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production*, pp. 93–113, 2021.

[17] C. V. Goldman, M. Baltaxe, D. Chakraborty and J. Arinez, "Explaining Learning Models in Manufacturing Processes", *Procedia Computer Science, 180*, pp. 259–268, 2021.

[18] N. Mehdiyev and P. Fettke, "Local Post-Hoc Explanations for predictive Process Monitoring in manufacturing", *29th European Conference on Information Systems - Human Values Crisis in a Digitizing World, ECIS 2021*, Marrakech, Morocco, 2020.

[19] C. Molnar, “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable”, *2nd edn.*, 2022. [Online]. Available from: https://christophm.github.io/interpretable-ml-book, retrieved on 08/26/2022.

[20] L. Breiman, "Random Forests", *Machine Learning 45*, pp. 5–32, 2001.

[21] A. Goldstein, A. Kapelner, J. Bleich and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation", *journal of Computational and Graphical Statistics,* 24, pp. 44–65, 2015.

[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *The Annals of Statistics, 29 (5),* pp. 1189–1232, October 2001.

[23] T. Hastie, R. Tibshirani and J. Friedman, "Random forests", *The elements of statistical learning*, Springer, pp. 587–604, 2009.

[24] S. M. Lundberg, “Interpretable Machine Learning with XGBoost”, April, 2018. [Online] Available from: https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27, retrieved on 08/26/2022.

[25] XGBoost Documentation, “Python API”, *Reference. xgboost developers*. [Online]. Available from: https://xgboost.readthedocs.io/en/latest/python/python_api.html, retrieved on 08/26/2022