Regional Feature Importance for Error Analysis in Manufacturing

Importance Measure that Reveal Insights

Valentin Göttisheim, Holger Ziekow,
Peter Schanbacher
Furtwangen University
78120 Furtwangen, Germany
{valentin.goettisheim, holger.ziekow,
peter.schanbacher}@hs-furtwangen.de

Djafar Ould-Abdelsam
Université de Haute-Alsace
IRIMAS Laboratory, Université de Haute-Alsace
68100 Mulhouse, France
djaffar.ould-abdeslam@uha.fr

Abstract - **Quality management in manufacturing can benefit from integration of artificial intelligence to detect and analyze errors in production. However, finding the causes of errors requires not only accurate predictions, but also suitable explanations of the underlying data analysis processes. This paper extends our previous work on a novel approach to measure the importance of features for error analysis. This approach bridges the gap between global and local importance and introduces the concept of regional feature importance, which captures the impact of features in specific regions of the feature space, rather than globally or locally. We generalize this method as a task of partitioning the feature space and aggregating the local importance of features within these regions. Our findings demonstrate that this approach can reveal interesting and actionable insights for quality management in manufacturing.**

Keywords - eXplainable Artificial Intelligence; XAI; SHapley Additive exPlanations; SHAP; feature importance, manufacturing quality management; error analysis.

I. Introduction

In this paper, we expand our earlier work [1] on regional feature importance (RFI) measures for quality management in manufacturing. This work generalizes the concept of RFI as a two-step process of partitioning and aggregation. The partitioning step divides the feature space into regions based on certain criteria, and the aggregation step combines the feature importance values within each region. Additionally, we extend our previous work by introducing new measures and demonstrate the usefulness of the proposed RFIs on synthetic data and a real-world dataset.

Feature importance measures are valuable tools for analyzing complex, high-dimensional production data to identify the causes of errors or defects. Quality management in modern manufacturing processes involves extensive testing and collection of detailed measurements along production lines. However, quality managers often struggle to pinpoint error causes within large data sets [2]. Artificial Intelligence (AI), combined with eXplainable AI (XAI), can utilize such data to predict errors [3] and provide insights to quality engineers by identifying potential causes of production errors [4]. Feature importance metrics can reveal these features that constitute potential error causes and reveal interesting insights for quality managers. However, existing feature importance measures are not tailored to this task [4] [5].

Our work is rooted in a research project with a German manufacturer [6]. Here, 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 involves 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 [7]. However, the high number of different measurements poses challenges for finding causes of errors in the data. Moreover, errors are rare in modern manufacturing processes which usually are highly optimized [8]. Quality management is often about driving down rare – but still costly – errors. Yet, existing applications have successfully used such high-dimensional to build AI models for predicting production errors [7][9]. 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 capability 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 among 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 demonstrate in this paper, such a global view often fails 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) [11] and LIME [12] provide local explanations for the impact of features on a prediction. These methods estimate the impact of features on individual data instances. However, analyzing a single data instance in isolation may not yield enough context to draw actionable conclusions and, thereby, limiting the usefulness in quality management.

Our work introduces feature importance measures that bridges the gap between global and local feature importance. We refer to this as regional feature importance (RFI). Building on this concept, we extend and refine our previous paper [1]. 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 actions include setting thresholds in quality checks or adjusting processes to avoid specific value ranges. Intuitively, drastic changes in error rates and high error rates in well-defined parts of a value range make features interesting. This paper presents importance measures that formalize these concepts of interestingness and translate them into an importance score. Specifically in this work, we make the following key contributions:

1) Extending and formally defining novel feature importance measures that are tailored to finding relevant features for quality management in manufacturing,
2) proposing a two-step process of partitioning and aggregating to construct the RFI measures,
3) providing five metrics to determine the RFI measures based on different criteria, and
4) evaluating the proposed measures on real-world data and comparing them with established importance measures.

With these contributions, we aim to assist human experts in quality management to 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 are illustrated on synthetic data in Section IV, and evaluated on a real-world dataset in Section V. In Section VI, we discuss related work and conclude in Section VII.

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 [5]. 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 [3]. As such, the system is designed to enable QEs to use ML support for error causes analysis, but not to engage with the technical depth of the ML system.

A reference process focusing on QEs intended to investigate errors in production is laid out in [2]. 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 [5]. 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 $|S| = n - 1$). One way to compute the feature contributions is an explanation model $e(z') = \phi_0 + \sum_{f=1}^{M} \phi_f z'_f$ where $z' \in \{0,1\}^M$[11]. This is the weighted average over all feature contributions. The explanation model is computed using the mapping $m_x(S) = m\big(e_x(z')\big)$ which maps all input values S to whether the feature is being used ($z' = 1$) or not known ($z' = 0$). However, SHAP values can also be efficient computed using model specific methods such as for tree-based model exploiting the internal structure of the model [13] .

SHAP decomposes the feature contribution into main effect and interaction effect with other features [13]. The main effect of a feature is the average marginal contribution of that feature across all possible coalitions of other features. The interaction effect of a feature with another feature is the difference between the joint contribution of both features and the sum of their individual contributions. The advantage of distinguishing the main and interaction effects is that it can reveal how features influence the model output not only by themselves, but also by interacting with other features. This can help to understand the complex and nonlinear relationships between features and the model output. Therefore, SHAP values can be decomposed in main and interaction effect by:

$$\phi_f(x_i) = \phi_{f,f}(x_i) + \sum_{j \neq f} \phi_{f,j}(x_i)$$

Where $\phi_{f,j}(x_i)$ is the SHAP interaction value between a feature f and a feature j on instance x_i. Moreover, $\phi_{f,f}(x_i)$ then is the main effect which is the resulting effects of feature f with itself.

In the following section we propose SHAP-based importance measures that are tailored to the task of quality management in manufacturing.

III. Regional Feature Importance

RFI is an approach to measure the contribution of a feature to the prediction of a model in a specific region of the feature space [5]. It bridges the gap between global and local feature importance [14][15]. This is unlike global importance, which assesses features over the entire dataset, and local importance, which focuses on the individual prediction [16]. Regional importance focuses on specific regions in the data and aims to identify those that are most relevant for explaining the error cause in manufacturing processes.

We propose a two-step process of partitioning and aggregation to construct the RFI. Partitioning are methods for dividing the feature value range into clusters that capture the local patterns and behaviors of the features. The partitioning is based on criteria, such as distance of data points, output predictions or intervals of feature values. This step isolates distinct areas where feature behavior is expected to be similar and allows interpretations about different contexts and regions. Following the partitioning, SHAP values within the region are aggregated to determine an importance score. These aggregations are based on criteria such as mean SHAP values, error frequency or change in SHAP values to pinpoint interesting prediction contributions of features.

To describe the RFI we define $X = \{x_1, x_2, \dots, x_n\}$ as the set of data point in the dataset X. However, for simplicity we also refer to data points as instance $x \in X$. Correspondingly, we define the set of labels $Y = \{y_1, y_2, \dots, y_n\}$ as the real data labels. We denote Mas the machine learning model that maps instance x to a prediction $M(x) = \hat{y}$. We use SHAP values $\phi_f(x)$ as a measure to quantify the contribution of an instance x to the prediction for a feature f. We seek to evaluate a scoring-function $g: g(f, X, y, M) \rightarrow \mathbb{R}$ that aggregates SHAP values and scores "interesting" features high.

In the following sections, we provide the two-step process of partitioning and aggregations to construct the RFI in detail. Moreover, we propose several methods for partitioning and aggregations, provide mathematical formulations and intuition for interpretation and reasoning.

A. *Partitionings*

To determine the RFI, first the feature value range is partitioned into clusters that optimize the feature importance score:

$$g(f, X, y, M) = \max_{C \in P(f)} A(f, X, C)$$

Here, $A(f, X, C)$ represents the aggregations later defined, which are obtained for sets of clusters $C \in P(f)$. $P(f)$ represents the proposed partitioning functions, resulting in sets of clusters $P(f)$=$\{C_1, C_2, \dots C_m\}$ for feature f. Each cluster $c \in C_f$ is a disjunct subset of the data Xsuch that $c \subseteq X$. In the following, we describe the used partitioning methods $P(f)$ to obtain C_f in detail.

Decision-Tree Partitioning: This method divides the feature value range into segments based on the splits of a decision tree that is trained on the feature values f and the output labels y. Each split is determined by selecting a threshold that maximize the reduction in impurity only considering one feature f. As impurity measure, we used the Gini $1 - \sum_{i=1}^{k} p_i^2$, where p_i is the proportion of data points in class i at a given node. Clusters C are then determined by the node in which data points x falls. The intuition is that features that have interesting patterns of SHAP values in regions that are predictive of the output labels are more indicative of error causes. This method can handle complex feature value distributions, as it creates partitions that are based on the feature value patterns and the output labels y regardless of the feature value range.

K-means Partitioning: This method uses the k-means algorithm to segment the feature values f into clusters, aiming to reduce the within-cluster variance. It minimizes the within-cluster sum (WCSS) of squares, defined as WCSS = $\sum_{x_f \in c_j} \left|\left|x_f - \mu_j\right|\right|^2$ where μ_j is the centroid of cluster c_j. This approach groups feature values close to a segment's centroid and separates those further away, thereby identifying areas with consistent feature values and therefore, sheds light on the local structure of the data, providing valuable insights into feature behavior.

Interval-based Partitioning: This method linearly divides the feature range into equal-sized intervals. This divides the entire range of f into partitions, into intervals I_j covering an equal portion of the feature's value range. Every data point x is allocated to the cluster c_j based on the intervals I_j range in which a feature value x_f lies. This method does not account for density or distribution patterns, instead, it examines how the characteristics changes across the features spectrum, and therefore, exhibits patterns which changes across the feature value range.

Hierarchical-based Partitioning: This method partitions the feature based on hierarchical clustering. Clusters are formed based on the pairwise Euclidean distances between data points within the feature f. To obtain the clusters the resulting dendrogram is cut at a specific level or the number of partitions is specified. Each cluster $c_i \in C$ comprises data points that are closely related in terms of their feature value. Therefore, this partitioning emphasizes similarity in the feature value.

With the discussed partitioning approaches, we laid out diverse approaches to segment a feature space. Each method offers a unique insight, emphasis different aspects of data relationships and distribution patterns. We now discuss the aggregations, facilitating the interpretation and synthesis of findings to gather actionable insights from the RFI.

B. Aggregations

After partitioning, we now describe the aggregations as second step to construct the RFI score $g(f, X, y, M) = \max_{C \in P(f)} A\,(f, X, C)$ where A denotes the aggregation method applied to each cluster C that has been derived from a partitioning method $P(f)$ of feature f.

Mean-Shap: The aggregation $A_{\text{mean}}(f, X, C)$ quantifies the average SHAP value across the data points of each cluster C. Formally, it is defined as:

$$A_{\text{mean}}(f, X, C) = \frac{1}{|C|} \sum_{x \in C} \phi_f(x)$$

Here, $|C|$is the number of data points in cluster C, and $\phi_f(x)$ is the SHAP value of feature f for a data point x. This aggregation method provides an average measure of the influence of a feature within a specific cluster, giving an overall indication of its importance within its feature value range.

Main-SHAP: Main SHAP quantifies the average main effect within a specific cluster, isolating these effects from interactions with other features. This metric measures the effect caused by the feature itself rather than of interactions. The intuition is that it provides a quantification of the contribution from the intrinsic influence of the feature on error in production, offering a possibly simple to interpret explanation in the domain context. This aggregation method is crucial for pinpointing features as explanation, facilitating which features are fundamental. The Main-SHAP is defined using the main SHAP values $\phi_{f,f}(x)$ described in the background section and formally expressed as:

$$A_{Main}\,(f, X, C) = \frac{1}{|C|} \sum_{x \in C} \phi_{f,f}(x)$$

Error-Shap: This aggregation method sums up the SHAP values of instances indicating errors or faulty products in the data. The intuition is that features that have high SHAP values for the errors are more relevant for explaining the causes. Formally, let x_e be the subset of data points $x_e \subseteq X$ where the actual labels indicating errors in the products. Error-Shap is then defined as:

$$A_{\text{err}}(f, X, C) = \sum_{x_e \in C} \phi_f(x_e)$$

Here, $\phi_f(x_e)$ represents the SHAP value for the error instance x_e. This aggregation sums up the SHAP values for across all error instances, emphasizing the relevance of the feature in explaining the causes of errors.

Error-Rate-Shap: This method multiplies the shap value of a feature by the error rate of the corresponding feature value. The intuition is that features that have high SHAP values and high error rates are more interesting for finding error causes. Formally, Error-Rate-Shap is defined as:

$$A_{\text{erate}}(f, X, C) = \sum_{x \in C} \phi_f(x) \cdot e(C)$$

where $e(C)$ represents the error rate within cluster $C$and is calculated as $|x_e|/|x_{ne}|$ within the same cluster. Specifically, as the proportion of the amount of error instances $|x_e|$ compared to the amount of non-error instances $|x_{ne}|$ for x_{ne} as subset of data points $x_{ne} \subseteq X$ where the actual labels indicate a non-error label. This error ratio $e(C)$ quantifies how often that cluster is associated with errors.

Slope-Shap: This measure assesses how rapidly the SHAP values of a feature change over its value range. It aims to identify features with significant SHAP value shifts, indicative of potential error causes. After partitioning, a rolling window method is applied to calculate the slope of the mean SHAP values across the window $w \in W$. The Slope-Shap aggregation is then defined as the absolute sum of these slopes within each window.

$$A_{\text{slope}}(f, X, C) = \sum_{w \in W} \left| \frac{\overline{\phi}_f(X_{w+1}) - \overline{\phi}_f(X_w)}{\Delta} \right|$$

Here, W denotes the set of rolling windows, X_w is the subset of X for the w-th window, and Δ represents the width of each window. $\overline{\phi_f}$ is denoted as the mean SHAP value of X_f, the set of all data points where feature f is present, and calculated as $\overline{\phi}_f(X') = \frac{1}{|X'|} \sum_{x \in X'} \phi_f(x)$ with $|X'|$ representing the number of data points of X'.

Z-score SHAP: This aggregation quantifies the deviation of SHAP values within a specific region from the average contribution of a feature. It provides a measure of how abnormal extreme feature contributions are in relation to the features typical impact. A high absolute Z-score indicates a significant deviation, pointing to features that have an unusually strong impact on predictions within regions compared to their average impact across all data. The Z-score SHAP is defined as follows:

$$A_{zscore}(f,X,C)=\frac{\sum_{x\in C}\phi_f(x)-\mu_f}{\sigma_f}$$

where μ_f is the mean SHAP value and σ_f is the standard deviation of the SHAP values of feature f across the entire dataset.

In this section, we have presented the concept of RFI and proposed several partitioning and aggregation methods to capture interesting and relevant aspects of features for error analysis in manufacturing. To illustrate the usefulness and effectiveness of our proposed RFI measures, we now apply them to synthetic data scenarios that mimic common error situations in manufacturing processes.

IV. Synthetic Data Evaluation

In the following, we illustrate the proposed regional feature importance using synthetic data, specifically designed to reflect characteristics of manufacturing situations, focusing on identifying features that are most interesting for explaining error causes. We focused on two specific error scenarios (A) Tail Error, and (B) Segment Error. Each scenario was designed to mimic distinct types of errors encountered in manufacturing data.

A) Tail Error: Involves a higher error rate in the tail of a normal distribution, representing errors that are rare but relevant.

B) Segment Error: This scenario focuses on a feature that strongly impacts the label, but only within a small range.

Global feature importance (GFI) measures aggregate the importance of features across the entire data distribution. This aggregation dilutes the impact of features for error analysis that focuses on rare events which may only occur in certain regions of the feature value range. The RFI highlights these critical regions, while global importance may overlook them due to averaging effects, leading to a misrepresentation of the feature's true impact in error analysis.

The use of synthetic data has the advantage that the ground truth is known. Here the correctness can be assessed by determining if the importance metric correctly identifies the most relevant feature (i.e., pinpointing relevant features as explanations for error analysis). Each synthetic data scenario consists of three main features: Target, Trap, and Noise. The "Target" feature is directly tied to error rates, "Trap" is a deceptive feature that ranks high in global importance but is less interesting in error analysis, and "Noise" introduces a minimal error rate, serving as a control variable. Each scenario is modeled as a binary classification task with 1 labeled as an error instance indicating a faulty product in production or 0 if not. We employ XGBoost [17] as a predictive model trained on 10,000 data points on which it achieves a perfect training ROC AUC score of 1. Note, that potential overfitting is not a concern for our experiments as they target feature importance and not prediction quality. Subsequently, global importance measures, such as Weight, Gain, Cover, or Abs. Mean SHAP and the RFI is computed. This process is repeated 10 times. The findings are presented in Table 1 and Table 2 as mean importance scores alongside the standard deviations.

A. Tail Error

This scenario simulates errors occurring in the tail regions of data distributions, often representing rare yet critical error cases. These errors are challenging to detect because they occur infrequently, yet they can have a considerable impact on the overall results in manufacturing. The scenario comprises three features "Target", "Noise", "Trap", and the binary label y that is affected by the error characteristics. We introduced anomalies specifically in the tail of the Target feature, ensuring these errors are clearly noticeable but overall not frequent.

- **Target:** A normally distributed feature that causes a 10% error rate based on a 0.02 quantile in the upper tail of its value range.
- **Trap:** A normally distributed feature that causes a 4% error rate over 0.5 quantiles of its range.
- **Noise**: A normally distributed feature that introduces a small amount (1%) of random noise in the label.

We argue that feature Target, despite its infrequent yet strong relations to errors, presents a more interesting case for error analysis than the Trap feature. Although the Trap feature encompasses more errors, it's less interesting as it comprises a much broader value range.

The global importance measures, as reported in Table 1, show that the Trap feature ranks highest, while the Target

TABLE 1. "EVALUATION RESULT GFI": RANKING THE DECEPTIVE FEATURE "TRAP" IMPORTANT

GFI	Rank		
	1	*2*	*3*
Cover	Trap 122.94/7.2	Target 83.76/7.22	Noise 75.28/3.10
Gain	Trap 1.49/0.05	Target 1.33/0.05	Noise 1.32/0.06
Mean Abs. Shap	Trap 0.52/0.05	Target 0.26/0.02	Noise 0.24/0.04
Total Cover	Trap 77331.80/5177.75	Target 52595.52/3989.74	Noise 46620.20/3220.44
Total Gain	Trap 936.7/34.4	Target 839.8/54.5	Noise 816.0/58.7
Weight	Trap 623.1/34.9	Target 619.0/30.0	Noise 607.9/24.7

Notation: feature (mean/std)

TABLE 2. " EVALUATION RESULT RFI": RANKING THE INTERESTING FEATURE "TARGET" AS IMPORTANT

RFI		**Rank**		
A	*P*	*1*	*2*	*3*
Mean SHAP	Tree	Target 1.34/0.13	Trap 1.09/0.15	Noise 0.5/0.11
	Hierarch.	Target 2.37/0.61	Trap 0.88/0.15	Noise 0.53/0.34
	Interval	Target 1.07/0.14	Trap 1.01/0.13	Noise 0.4/0.11
	K-means	Target 2.05/0.42	Trap 0.93/0.12	Noise 0.42/0.2
Error Rate SHAP	Tree	Target 0.18/0.03	Trap 0.1/0.02	Noise 0.03/0.01
	Hierarch.	Target 0.76/0.35	Trap 0.07/0.07	Noise 0.07/0.14
	Interval	Target 0.11/0.02	Trap 0.08/0.02	Noise 0.02/0.01
	K-means	Target 0.56/0.28	Trap 0.06/0.01	Noise 0.02/0.01
Error SHAP	Tree	Target 0.16/0.02	Trap 0.14/0.02	Noise 0.04/0.01
	Hierarch.	Target 0.13/0.03	Trap 0.1/0.01	Noise 0.03/0.01
	Interval	Target 0.13/0.04	Trap 0.1/0.02	Noise 0.03/0.01
	K-means	Target 0.13/0.03	Trap 0.1/0.02	Noise 0.03/0.01
Main SHAP	Tree	Target 1.29/0.12	Trap 1.15/0.18	Noise 0.53/0.11
	Hierarch.	Target 1.92/0.37	Trap 0.82/0.18	Noise 0.74/0.45
	Interval	Target 1.0/0.09	Trap 0.98/0.15	Noise 0.41/0.11
	K-means	Target 1.69/0.3	Trap 0.91/0.13	Noise 0.72/0.34

Notation: feature (mean/std)

feature, which we consider most important, ranks second. As anticipated, feature Noise, which was intended to be the least important, ranks last. These results show that global feature importance assigns higher importance to features that are less relevant for error analysis, thereby illustrating the limitations of global metrics for error analysis. The RFI aims to overcome the limitations of global measures by focusing on specific regions. The RFI is computed for every feature f using the proposed partitioning methods P to determine clusters $C \in P(f)$. The importance score $g(f, X, y, M)$ is the maximum value of a proposed aggregations A among all clusters C: $\max_{C \in P(f)} A(f, X, C)$.

In this scenario, Mean SHAP, Error Rate and Error SHAP are crucial metrics. The RFI importance scores g are presented in Table 2 as importance rank with the average importance scores g and its standard deviation over the 10 repetitions. Over all metrics except the Error SHAP with hierarchical partitioning, the feature Target is consistently ranked as the most important.

The RFI can also be illustrated as a scatter plot with curves the aggregations of SHAP values across the identified clusters as in Figure 1 shown. The blue points symbolize the SHAP values of individual data instances. The red points highlight the instances associated with errors in products. All plots show increased RFI metric scores g on the interval [2,4] indicating an increased importance of the region within the feature space that is relevant for explaining error causes.

This visual representation, along with Table 2, aids in understanding the connections between features and the potential error, i.e., the error in the tail regions of the data distribution. In the following, we discuss how these aggregations reveal interesting regions.

Mean SHAP: This aggregation calculates the average SHAP value for a cluster across all instances of that cluster. SHAP values quantify the contribution of a feature to the prediction for an instance. Therefore, the Mean SHAP

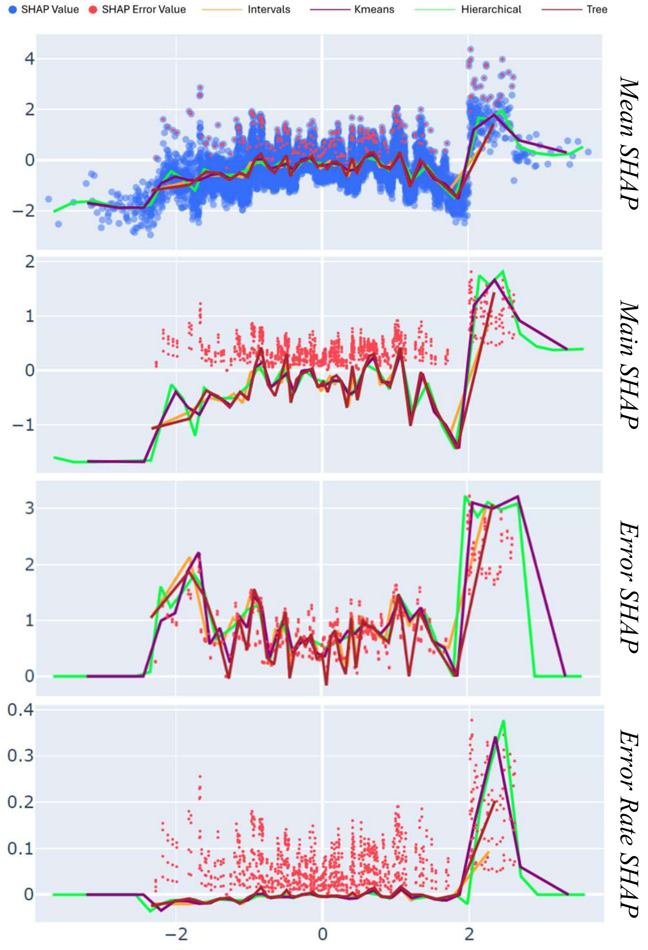

Figure 1. "Evaluation RFI Target feature" RFI aggregations over all partitions of the Target feature showing increased values for features values greater than two.

aggregation provides a measure of the average impact of the cluster on error predictions, highlighting clusters that have a strong relation to errors. Figure 1 Mean SHAP shows the increased average impact for a feature value greater than two.

Main SHAP: Main SHAP quantifies the average main effect within a specific cluster, isolating the effects from interactions with other features. This metric shed light on the standalone contribution of features to the model's output, revealing the intrinsic influence of the feature. This aggregation method is crucial for pinpointing features as explanations, facilitating which features are fundamental to predictions. The Main SHAP metric is instrumental in identifying key drivers by focusing on the direct impact of individual features. Figure 1 shows the Main SHAP plot.

Error Rate SHAP: Considers the error rate alongside with the region's impact. It is computed by multiplying the cluster mean SHAP with the error rate of the cluster. The error rate is the proportion of instances associated with errors relative to those that are not. This aggregation highlights impactful regions with frequent errors.

Error SHAP: This measure considerers SHAP values of all instances labeled as errors in a region. It offers a direct insight into which regions are responsible for errors, regardless of how often those errors occur. This highlights significant errors even in regions without strong associations with errors. SHAP plots for Error and Error Rate SHAP are also presented in Figure 1.

B. *Segment-Error*

The second scenario resembles the challenge of identifying production errors in a small value range, targeting production problem that affects the overall production process performance. The scenario follows the setup described earlier. However, the following features are considered as training data:

- **Target**: A uniformly distributed feature that causes a higher error rate of 10% in a small segment of its range.
- **Trap**: A uniformly distributed feature that causes a moderate error rate in five segments of its range in varied intensity of 0.2-1 % error rate.
- **Noise**: A uniformly continuous variable that accounts for 1% errors of the data as noise.

We argue that feature Target, as of its localized yet significant relations to errors, presents a more interesting case for error analysis than the Trap feature. Although the Trap feature encompasses more errors, it's less interesting as it again comprises a broader value range. The global importance measures, as reported in Table 3, show that the Trap feature ranks highest, while the Target feature, which we consider most important, ranks second. As anticipated, feature Noise, which was intended to be the least important, ranks last. These results show that global feature importance assigns higher importance to features that are less interesting for error analysis, thereby illustrating the limitations of global metrics for error analysis.

TABLE 3. " EVALUATION RESULT GFI" RANKING THE DECEPTIVE FEATURE "TRAP" AS IMPORTANT

GFI	Rank		
	1	*2*	*3*
Cover	Trap 144.75/12.76	Target 116.39/8.49	Noise 105.70/9.51
Gain	Trap 1.64/0.07	Target 1.59/0.06	Noise 1.53/0.06
Mean Abs. Shap	Trap 0.31/0.02	Target 0.22/0.02	Noise 0.16/0.02
Total Cover	Trap 93895.74/7866.96	Target 77933.30/7179.09	Noise 67254.76/6711.44
Total Gain	Trap 1066.32/41.3	Target 1063.8/54.0	Noise 975.96/59.6
Weight	Target 669.90/44.50	Trap 649.10/18.3	Noise 636.50/35.3

Notation: feature (mean/std)

Subsequently, the RFI is computed for every feature f using the proposed partitioning methods P to determine clusters $C \in P(f)$. The importance score $g(f, X, y, M)$ is than examined as the maximum value among all clusters C: $\max_{C \in P(f)} A(f, X, C)$. This illustration focuses on the proposed aggregations Slope SHAP and Z-Score SHAP for which the results of g in Table 3 are reported based on 10 repetitions, and the scatter plots with its RFI curves in Figure 2.

The RFI results in Table 4 shows that both metrics rank the more interesting feature Target as more important. The scatter plots in Figure 2 show the increased mean SHAP

TABLE 4. " EVALUATION RESULT RFI": RANKING THE INTERESTING FEATURE "TARGET" AS IMPORTANT

RFI		Rank		
A	*P*	*1*	*2*	*3*
Slope SHAP	Tree	Target 0.46/0.08	Trap 0.29/0.07	Noise 0.25/0.06
	Hierarch.	Target 0.43/0.05	Trap 0.27/0.07	Noise 0.23/0.05
	Interval	Target 0.44/0.05	Trap 0.27/0.07	Noise 0.23/0.06
	K-means	Target 0.42/0.06	Trap 0.27/0.08	Noise 0.22/0.06
Z-Score SHAP	Tree	Target 559/119.3	Trap 421/66.63	Noise 284/41.46
	Hierarch.	Target 486/73.44	Trap 377/52.87	Noise 245/59.44
	Interval	Target 470/34.6	Trap 348/40.74	Noise 230/26.65
	K-means	Target 479/59.99	Trap 351/40.95	Noise 233/42.2

Notation: feature (mean/std)

values in the interval [0.4,0.425]. In the following, we discuss how the other both aggregations reveal interesting regions.

Z-Score SHAP: The Z-Score SHAP metric quantifies the deviation of SHAP values within a specific region from the average contribution of a feature. It is measured in standard deviations from the mean impact of a feature across the entire dataset. It provides a measure of how anomalously a feature behaves in relation to its general impact. A high absolute Z-score indicates a significant deviation, pointing to features that have an unusually strong impact on predictions within regions compared to their average impact across all data. In this illustration, this can be observed over the interval [0.4,0.425] in the scatter plots in Figure 2.

Slope-SHAP: This metric is used to identifying points or regions where the contribution of the feature changes significantly. Slope SHAP quantifies the rate of change in the mean SHAP values by calculating the slope of these values using a rolling window over the partitioned feature value range. This aggregation considers the absolute values, thereby capturing the overall variability in the feature's impact. High Slope-SHAP values indicate features that exhibit pronounced shifts in their influence at specific thresholds, signaling potential transition points that could lead to errors in production. This method enables the pinpointing of errors that are caused by exceeding a threshold, which can be mitigated by setting or adjusting an alarm threshold in production. In this illustration, this can be observed over the intervals [0.3,0.4] and [0.4,0.5] in the scatter plots in Figure 2.

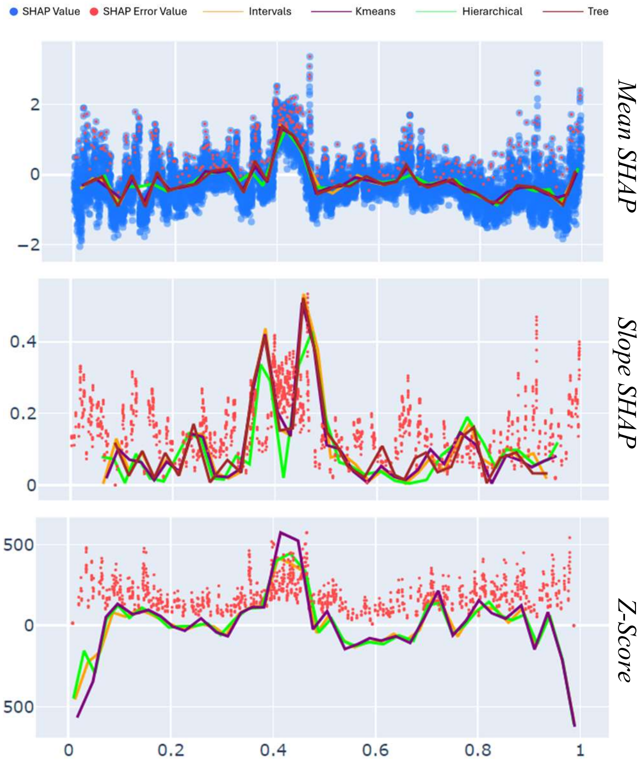

Figure 2. "Evaluation RFI Target feature" RFI aggregations over all partitions of the Target feature showing increased values for features values between 0.4 and 0.45.

These illustrations shows that the regional feature importance extend the insights gained from global feature importance measures in identifying the features that are most relevant for explaining the causes of errors in the synthetic data. The regional feature importance measures can capture the features that are more interesting for error analysis, while the global feature importance measures can be misleading or insufficient, as they take a global view and ignore rare but strong relations.

V. Real World Data Experiment

In this section, we evaluate the RFI measures on a real-world data set form the steel manufacturing domain [18]. The dataset contains 27 features and 52407 instances related to the quality of steel plates, and 158 binary labels indicating a product error, i.e., a production fault. We trained an XGBoost model which achieves a ROC AUC of 1 on the training data. Again, potential overfitting is not a concern for our experiments as they target feature importance and not prediction quality. This high performance ensures that the model can effectively capture the relationships within the data, making it suitable for evaluating the feature importance measures.

The RFIs are computed as each feature is divided in 20 partitions resulting in different cluster sizes depending on the feature. To compute the Slope SHAP aggregation a window size of five is used. In the following discussion, we compare the global and the regional feature importance rankings alongside with the RFI scatter plots.

The scatter plots show the curves of partitioning methods, where each break represents the mean of the cluster location over the feature value. For each cluster, the aggregations method is annotated on the right. As references the SHAP values of instances for error labels (red dots) are highlighted. However, only the Mean SHAP plot shows the real value of the error SHAP values. For all other scatter plots, the value is linearly scaled. Therefore, they show the right amplitude and position relative to each other, however, not the resulting SHAP values. These visualizations help in understanding how feature importance varies across different regions of the feature space, providing a nuanced view of feature contributions.

To identify the features that are most interesting for explaining the error causes and providing insights for quality management, we discuss the GFI and RFI of the top five rankings. The rankings are presented in the Table 5. The results show commonalities; however, there are also disagreements, which form the basis for the following discussion. Our analysis highlights the strengths and weaknesses of each method and provides a comprehensive understanding of feature importance in the context of manufacturing errors.

TABLE 5. "TOP FEATURE RANKINGS": STEEL DATA TOP 5 RANKINGS OF RFI AND GFI– HIGHLIGHTED TOP FIVE RANKINGS.

Aggregration *A*	Partitions *P*	Orientation_Index	Length_of_Conveyer	Edges_Y_Index	Outside_X_Index	Empty_Index	Minimum_of_Luminosity	Log_X_Index	Steel_Plate_Thickness	Y_Minimum	X_Minimum	Edges_Index	Luminosity_Index	Maximum_of_Luminosity	Square_Index	X_Maximum	X_Perimeter	
Z-Score	Tree	**4**	6	21	**3**	13	18	20	9	**1**	8	17	11	12	10	**5**	**2**	*RFI*
	Hierarch.	**2**	7	17	23	11	9	15	**4**	**1**	6	13	12	**5**	10	**3**	22	
	Interval	**3**	**5**	22	**4**	13	19	20	14	**1**	6	12	9	11	10	8	**2**	
	K-means	**2**	9	16	21	8	11	14	**4**	**1**	**3**	12	13	**5**	7	10	23	
Error Rate	Tree	**1**	**2**	14	**5**	8	**3**	17	6	10	7	12	13	15	**4**	9	16	
	Hierarch.	**2**	**5**	13	24	8	**1**	15	**3**	**4**	6	12	16	10	7	9	23	
	Interval	**1**	**2**	14	6	7	**3**	17	12	11	**5**	9	13	16	**4**	8	15	
	K-means	**1**	**4**	11	24	9	6	14	**3**	**2**	**5**	12	16	10	7	8	23	
Error SHAP	Tree	**1**	**2**	14	7	**5**	**3**	19	10	**4**	13	8	12	6	9	17	16	
	Hierarch.	**1**	**2**	11	15	7	**3**	17	6	**4**	10	8	13	**5**	9	14	21	
	Interval	**1**	**3**	14	**4**	7	**2**	19	10	6	12	8	11	**5**	9	18	17	
	K-means	**1**	**3**	11	14	**5**	**2**	17	7	**4**	10	8	13	6	9	16	21	
Mean SHAP	Tree	**1**	**2**	13	**4**	9	**3**	18	6	**5**	10	12	11	8	7	15	16	
	Hierarch.	**1**	**3**	13	27	7	**2**	15	6	**4**	9	10	12	**5**	8	11	25	
	Interval	**1**	**2**	13	**5**	8	**3**	17	10	**4**	9	11	12	6	7	15	16	
	K-means	**1**	**4**	11	27	9	**3**	14	**5**	**2**	7	10	12	6	8	15	25	
Main SHAP	Tree	**1**	**3**	12	**5**	7	**2**	16	8	**4**	11	9	10	6	14	17	19	
	Hierarch.	**1**	**3**	12	26	6	**2**	14	7	**4**	9	8	10	**5**	13	15	25	
	Interval	**1**	**3**	12	**5**	7	**2**	17	9	**4**	11	8	10	6	14	16	18	
	K-means	**1**	**3**	12	25	7	**2**	14	6	**4**	9	8	10	**5**	13	16	26	
Slope SHAPs	Tree	**1**	**2**	7	**4**	9	**3**	6	**5**	13	10	8	14	12	11	17	18	
	Hierarch.	**1**	**5**	**3**	15	6	**2**	**4**	7	11	16	8	12	9	10	13	19	
	Interval	**1**	**3**	**2**	**5**	9	**4**	8	6	14	7	11	13	10	12	18	17	
	K-means	**1**	**3**	**4**	13	**5**	**2**	6	7	9	16	11	14	10	8	15	21	

Weight	9	**3**	18	13	**1**	8	22	11	**4**	6	**2**	**5**	7	12	17	14	*GFI*
Gain	**2**	**3**	**1**	14	6	9	**4**	8	15	**5**	17	16	10	7	23	21	
Cover	**2**	8	**1**	**3**	14	6	**4**	**5**	10	9	19	16	7	11	18	21	
Total Gain	**1**	**2**	**4**	14	**3**	6	16	10	8	**5**	9	12	7	11	22	17	
Total Cover	**1**	**3**	**5**	**2**	7	6	15	8	**4**	10	11	12	9	13	20	17	
Abs. Mean SHAP	**1**	**4**	**2**	6	**5**	**3**	8	9	12	13	10	14	11	7	18	19	

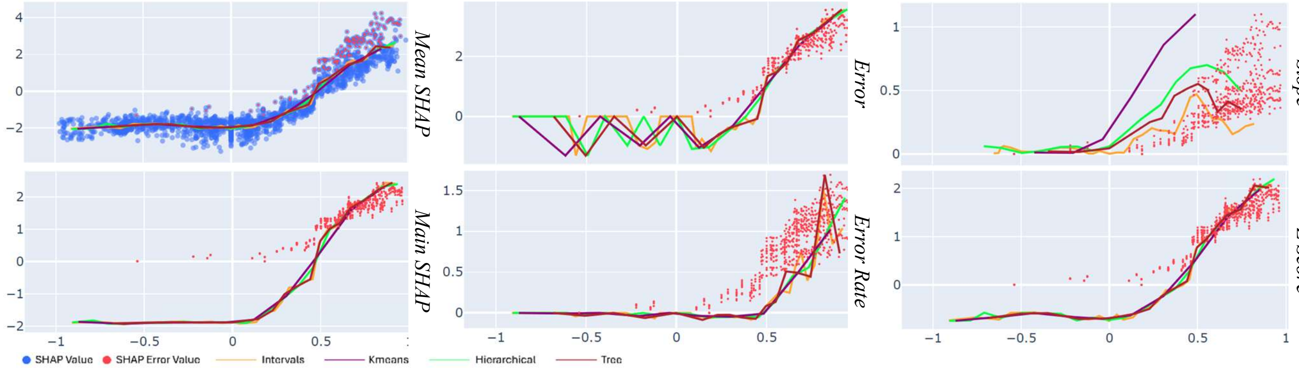

Figure 3. “Orientation_Index” The feature significantly influences error prediction within the range of [0.5, 1]. Main SHAP values indicates importance, Error SHAP indicates a strong association with error causes, and Error Rate SHAP underscores the feature’s relevance in regions with few data points but substantial error instances. Finaly, the Slope SHAP metric reveals a steep increase over the feature range [0, 0.5], where SHAP values exhibit a sharp incline.

A. Features with Global and Regional Importance

The following discussion presents the results about features which are important based on both GFI and RFI.

The GFI ranks Orientation_Index, as shown in Table 4, as the first and second most important feature, based on Total Gain, Total Cover, Gain, and Cover. This indicates that Orientation_Index on average impacts many data points in its splits which lead to substantial improvement in performance. In addition, Mean Abs. SHAP also places it on the first rank, highlighting the feature’s overall contribution to the prediction.

The RFI measures allow a more nuanced picture. Mean SHAP ranks Orientation_Index as first and second most important, showing its positive influence in the [0.5, 1] range, as shown in Figure 3. Main SHAP metrics provides an indication of how much the feature itself (rather than interactions with other features) can be considered as error explanation. The increased Main SHAP values in the interval [0.5,1] suggest that the feature is an interesting candidate for explaining errors in this region. Error SHAP quantifies the average impact of error instances. Error SHAP ranks it high across all partitioning methods, indicating a strong relation to error causes in the [0.5, 0.9] region. Error Rate SHAP ranks the feature as important on all partitioning methods. This aggregation considers the mean SHAP value and the error rate, focusing on regions with small numbers of data points with a significant amount of error instances. Error Rate SHAP shows increased values in the region of [0.5, 1], indicating separable errors for which the feature serves as a suitable explanation. The Slope SHAP metric exhibits a steep increase over the feature range of [0,0.5], where the feature shows a sharp increase of SHAP values. This indicates that the feature has a significant impact on the prediction of errors above the threshold of 0.5. Z-Score SHAP ranks the feature highly, indicating above-average impact. Overall, the RFI indicates that Orientation_Index is an interesting candidate to consider as explanation of error cause for the region above a feature value of 0.5.

GFI ranks Length_Of_Conveyer (Figure 4) as second to fourth, showing its frequent use and high impact. RFI also places it in the top five, with Mean SHAP showing increased values in [1690, 1700] and Main SHAP identifying it as the main driver. Error SHAP highlights strong relations to errors

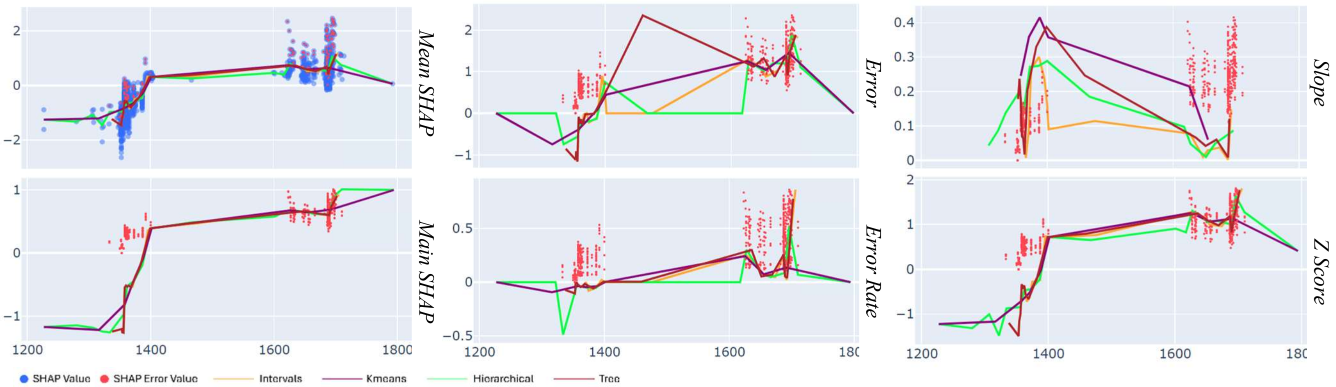

Figure 4. “Length_of_Conveyer”: Mean SHAP reveals increased SHAP values within the range of [1690, 1700] and Main SHAP values suggest that the feature itself is the primary driver in this region. Error SHAP emphasize a strong association with errors in the region [1600, 1700]. Additionally, Slope SHAP identifies a threshold, and Z-Score SHAP indicates an impact higher than on average above a feature value of 1400.

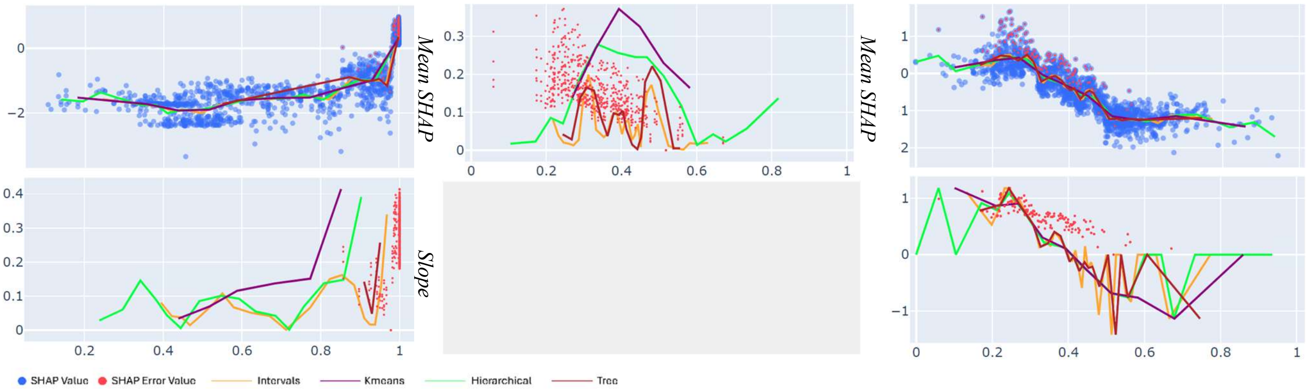

Figure 5. "Edges_Y_Index and Empty_Index": Edges_Y_Index (left) exhibits high importance in the Slope aggregation. The slope of this feature sharply increases in the region [0.9, 1]. Quality engineers can leverage this threshold to adjust alarms. Empty_Index (right) shows increased relations to errors in the interval [0.2, 0.3] indicated by Error SHAP.

in [1400, 1700], and Error Rate SHAP points to a critical error rate at 1700. Z-Score SHAP confirms its above-average impact above 1400, suggesting its critical role in product quality at high values.

The GFI measures rank Edges_Y_Index as the most important feature, highlighting its broad impact. Additionally, Mean Abs. SHAP also ranks the feature as the most important, indicating a substantial impact on the model's overall performance. The Slope SHAP aggregation ranks Edges_Y_Index high. Slope SHAP measures the change in mean SHAP values over the feature value range. Specifically, as shown in Figure 5, the slope is high in the region [0.9, 1], where Edges_Y_Index has a sharp increase of SHAP values. This presents a threshold where quality engineers could adjust production alarms to.

The GFI measure ranks Empty_Index as the first and the third most important feature, based on Weight and Total Gain. This shows that the feature is often used as split criteria, although it leads to improvements in prediction only in a few instances. The Mean Abs. SHAP ranks the feature fifth, indicating that it decently contributes to the overall model predictions. The RFI ranks Empty_Index (Figure 5) on Error SHAP and Slope SHAP as the fifth most important feature, according to the K-mean and the Tree partitioning method. The feature shows an increased contribution of errors in the interval [0.2,0.3] with decreasing effects up to 0.4 where the Slope SHAP also shows a threshold. Leveraging this threshold for production alarms could be valuable.

The GFI measures rank Outside_X_Index as the second and third most important feature, based on Total Cover and Cover. These indicate that the feature influences a large number of data points both in total and on average when used in the splits. The RFI, as shown in Figure 6, reveals that for error causes analysis, the interval [0, 0.01] is particularly interesting. The Mean SHAP in this region shows increased values, supported by the Error Rate metric, which shows an increased error rate in this region. Furthermore, the Main SHAP values are also increased, suggesting that Outside_X_Index itself is the primary driver of the effects. The Slope aggregation displaying the change of SHAP values shows a steep decrease, potentially indicating a threshold leading to actionable insights. Additionally, the Z-Score SHAP indicates above average performance of feature values <0.007. This shows that Outside_X_Index is important for explaining errors in production for small feature values. The GFI ranks Minimum_of_Luminosity as the fourth most

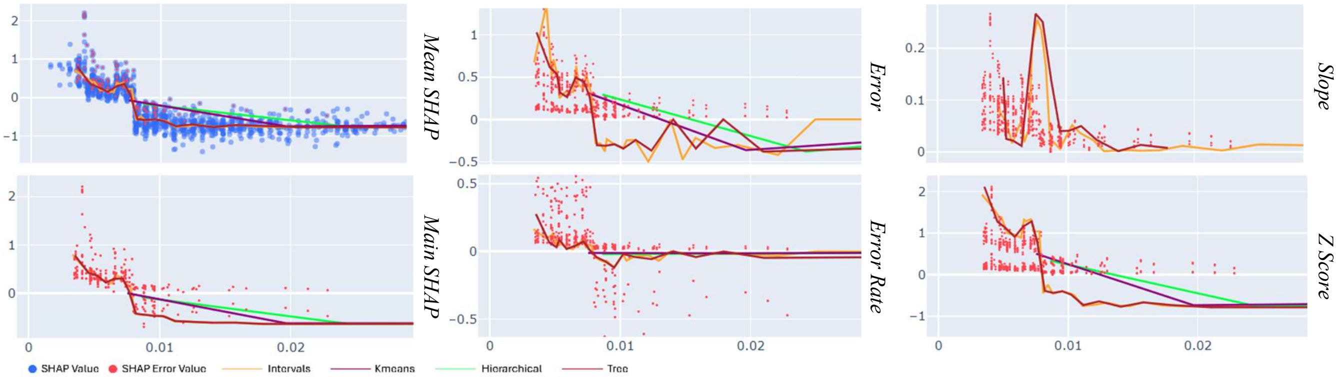

Figure 6. "Outside_X_Index": The interval [0, 0.01] is particularly interesting. Within this region, Mean SHAP values increase, supported by the Error Rate metric, which indicates an increased error rate. Moreover, Main SHAP values are higher, suggesting that Outside_X_Index itself drives the resulting effects. Slope SHAP displays a steep decrease, potentially indicating a threshold for actionable insights. Additionally, the Z-Score SHAP indicates above average performance of feature values < 0.0075.

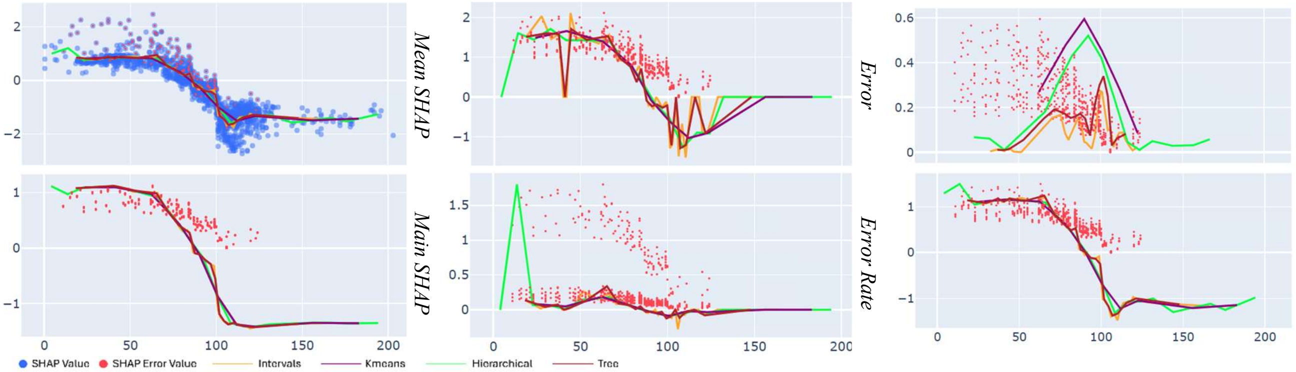

Figure 7. “Minimum_of_Luminosity”: An important feature for explaining error causes, especially in the region of [0, 75]. In this region, the feature has a positive impact on predicting errors, and not because of interactions. The feature shows a high error rate and a higher impact than on average in this region. The threshold of 75 could be a potential threshold for detecting errors.

important feature, based on Mean Abs. SHAP. This indicates a high impact over the entire feature value range for both error and non-error cases. The RFI ranks Minimum_of_Luminosity, as shown in Figure 7, second and third on Mean, second on Main, second and third for Error, second to fifth for Error Rate, and third and fourth rank for the Slope SHAP metrics. Specifically, Mean SHAP shows a positive influence on predicting errors in the region of [0, 75], highlighting its critical role for error analysis. In the same [0,75] region, the Main SHAP metric shows increased values, suggesting that the feature itself drives errors rather than interactions. The Error SHAP metric shows increased value for Minimum_of_Luminosity in the region of [0, 75], implying that the feature is more likely to cause errors in this region.

For feature values smaller than 25, the Error Rate SHAP shows increasing error rates, indicating a sparse region with high error ratio. The increased Z-score SHAP values for the region [0,75] emphasizes its impact in this region beyond the feature average. The Slope SHAP shows a sharp change of SHAP values which suggests a potential threshold for a feature value of 75. Overall, the RFI underscores that Minimum_of_Luminosity is a critical factor for the quality of products. It implies that the feature is more likely to serve as explanation in the region [0,75] with potential threshold of around 75 for identifying errors.

The GFI measures ranks Log_X_Index as one of the top five most important features, based on Gain and Cover. This indicates that the feature produces splits that affect a great number of data points and have a great impact on the prediction. For the RFI, as shown in Figure 8, Slope SHAP ranks the feature on fourth place on the hierarchical partitioning method and shows a threshold at a feature value of 1.25. The minor importance assigned from others RFIs points to the negative SHAP value above the feature value of 1.25, suggesting that feature is indicative of non-error instance above that threshold.

The GFI ranks Steel_Plate_Thickness (Figure 8) as the fifth most important feature, based on Cover. This indicates that the feature splits affect a large amount of data. However, the RFI reveals the relations to error causes. The error rate aggregation ranks the feature high for the hierarchical and k-means partitioning, showing that the error rate of Steel_Plate_Thickness is higher in the region of [100, 200]. The Z-Score SHAP ranks this feature on fourth, indicating a greater-than-on average impact in the region. The Slope

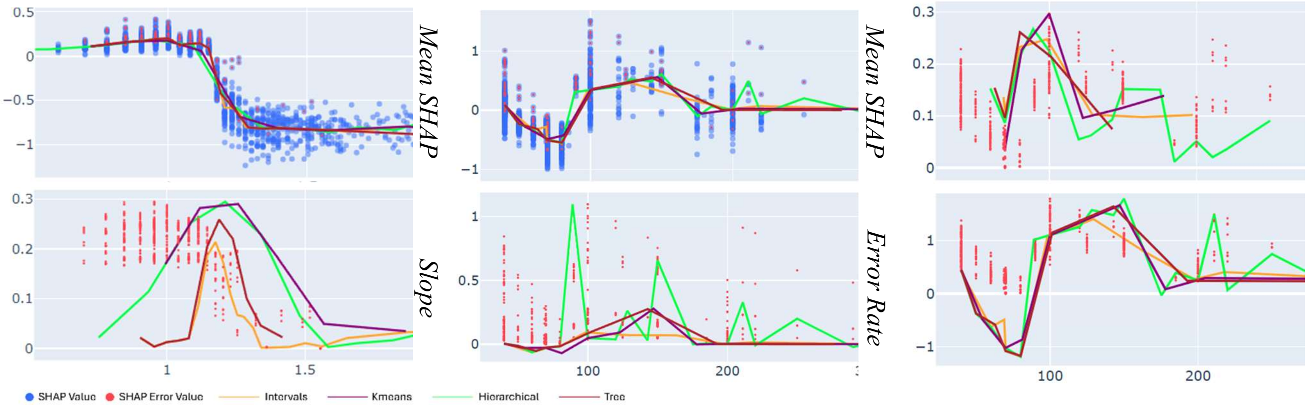

Figure 8. “Log_X_Index and Steel_Plate_Thickness”: Log_X_Index (left) is of minor interestingness for error explanation, except the threshold of around 1.25. Steel_Plate_Thickness (mid and right) exhibits and increased error rate and greater than average impact in the region of [100, 200], as indicated by Error Rate and Z-Score. The Slope SHAP also shows a threshold of around 100.

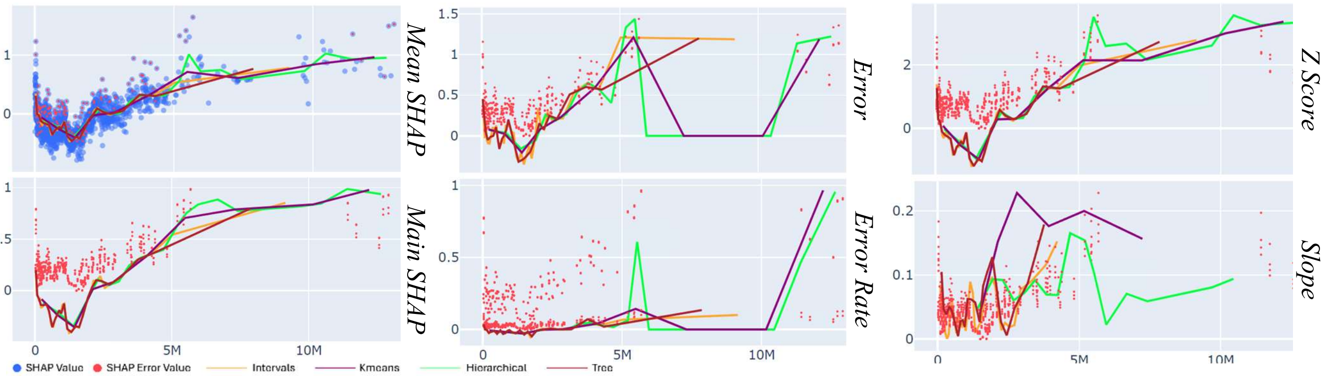

Figure 9. "Y_Minimum": The feature has an increasing impact on predicting errors in the region of [5M, 15M], and not because of interactions. It shows a high error rate and a higher than average impact in the region of >10M. Slope SHAP suggests some thresholds for <5M. Y_Minimum is a key factor for error analysis, especially in the regions of >5M.

metric indicates a threshold at a feature value of about 100. Consequently, Steel_Plate_Thickness is likely a critical factor in the range [100, 200], with increased error frequency and sufficient SHAP effects.

Feature Y_Minimum, as shown in Figure 9, is ranked as the fourth most important feature, based on Weight and Total Cover. This indicates a frequent use and many influenced data points, at least in some splits.

The Mean SHAP ranks Y_Minimum on second to fifth place, driven by positive effects in the region [5M, 15M]. This implies a critical influence on the product in this region. Main SHAP ranks the feature on fourth, showing increasing influence of the feature from [2M, 15M]. This indicates that the influence is caused by the feature itself. The Error SHAP metric ranks the feature in fourth and sixth place, showing increasing influence in the interval [2M, 7.5M] and above 10M. Error SHAP provides an indication of contribution to errors in products in these regions. The error rate SHAP metric ranks Y_Minimum high for the k-means and hierarchical partitioning. These partitioning methods focuses on separated regions, which also reveal patterns in sparse regions. The error rate is higher in the region of >10M, implying clearly separable product errors. Moreover, the Z-Score SHAP shows increasing values, indicating an impact greater than average. Therefore, Y_Minimum is interesting for error analysis and likely to explain the errors in the region of a feature value <5M.

The GFI ranks X_Minimum, as shown in Figure 10, as one of the five most important features, based on Gain and Total Gain. This suggests that using the feature as splitting criteria improves the predictive performance both in total and on average. The RFI ranks the feature as important on the Error Rate and the Z-Score metrics. Specifically, the feature shows an increased error rate at a feature value of about 1500, indicating separable errors with high frequency and a strong association with errors in that region. The Z-Score SHAP ranks the feature third for the K-means partition method, indicating an effect greater than average in this region. Consequently, X_Minimum is a promising candidate for explaining error on high feature values.

B. Features with Global Importance

In the following, we focus on the features that are ranked as important only GFI measures, but not by the RFI. The GFI ranks Edges_Index (Figure 11) as the second most important feature, based on Weight. This indicates that Edges_Index is frequently used in the splits of the model. However, the regional feature importance measures reveal that the feature is not particularly useful for explaining error causes. The SHAP values over the entire feature range have small effects. Moreover, the feature is often used as indicated by Weight, however, without effect on the predictive performance. The GFI rank Luminosity_Index (Figure 11) as the fifth important feature, based on Weight. This indicates that the feature is frequently used in the splits of the model. Luminosity_Index is not ranked as important from the RFI. The feature is placed on rank twelfth for Mean SHAP and tenth on the Main SHAP metrics, indicating some influence of the feature to the model

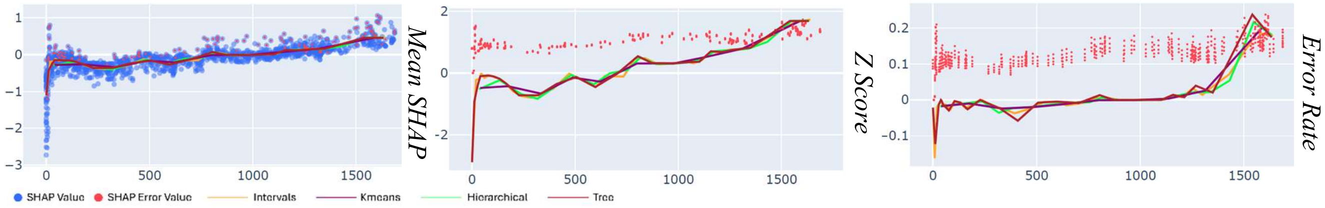

Figure 10. "X_Minimum, Luminosity_Index and Edges_Index": X_Minimum (left and mid) shows a high importance in the region around 1500, where the Error Rate and the Z-Score SHAP indicates a strong association with errors and a greater than average impact.

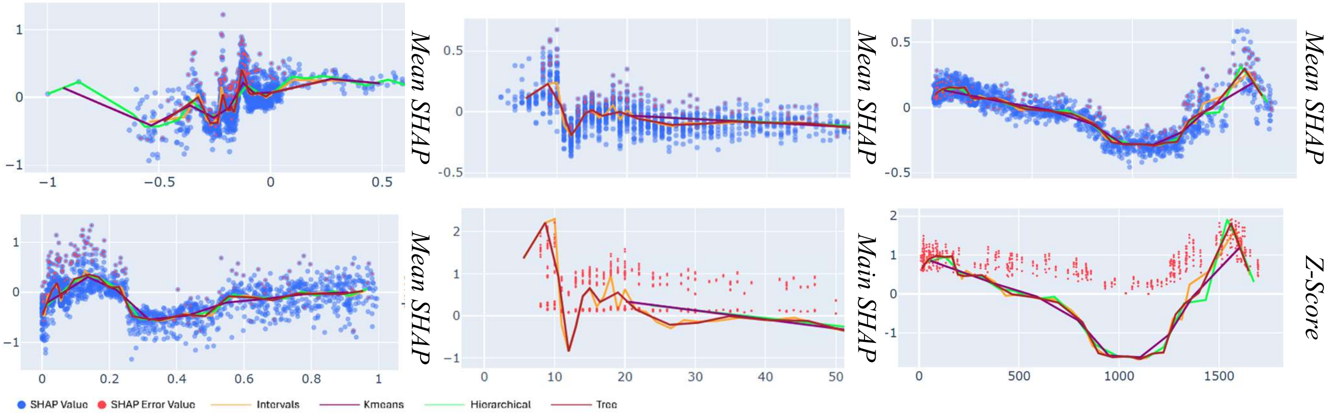

Figure 11. "X_Perimeter, Edges_Index, Luminosity_Index and X_Maximum": Edges_Index (top left) and Luminosity_Index (bottom left) are both important, based on Weight. The SHAP values of both features are small and do not show any clear patterns. Therefore, both feature are not very interesting. X_Perimeter (mid) shows increased main effects in the region [8, 10], indicating that the feature itself primarily causes the effect. X_Maximum (right) shows increased Z-Score SHAP values for feature values greater than 1500, indicating a higher than average impact of the feature in this region.

performance. However, due to its weak effects, this feature is not interesting for error analysis.

C. Features with Regional Importance Only

In the following, we focus on the features that are ranked in the top five by the RFI measures, but not by the GFI measures. These features are interesting because they reveal the limitations of the GFI measures in capturing the local patterns and behaviors of the features that are relevant for error analysis.

The GFI do not rank X_Perimeter, as shown in Figure 11, in the top five. The Z-Score SHAP on the tree and interval partitioning methods show increased effects in the interval [8,10]. However, Main SHAP ranks the feature with minor importance which shows that these effects are resulting from interactions with other features. The GFI do not rank X_Maximum (Figure 11) in the top five. The Z-Score SHAP on the tree and hierarchical partitioning methods show increased effects for feature values greater than 1500. Error Rate SHAP ranks the feature eighth and ninth, indicating some separable errors. However, Main SHAP ranks the feature with minor importance which shows that these effects also result from interactions.

The GFI do not rank Maximum_of_Luminosity, as shown in Figure 11, in the top five. However, the RFI rank the feature on the fifth most important. Error SHAP for the tree and the interval partition method indicate increased SHAP values of error instances on the interval [125, 200]. In the same region, the Main SHAP metrics also show that the feature is likely the primary driver of the effect. The increased Z-Score SHAP metric indicates an impact greater-than-average in the region. Therefore, Maximum_of_Luminosity is an interesting feature for error analysis.

The GFI does not rank Square_Index (Figure 12) in the top five. However, Error Rate SHAP ranks the feature fourth for both the tree and interval partitioning methods, indicating

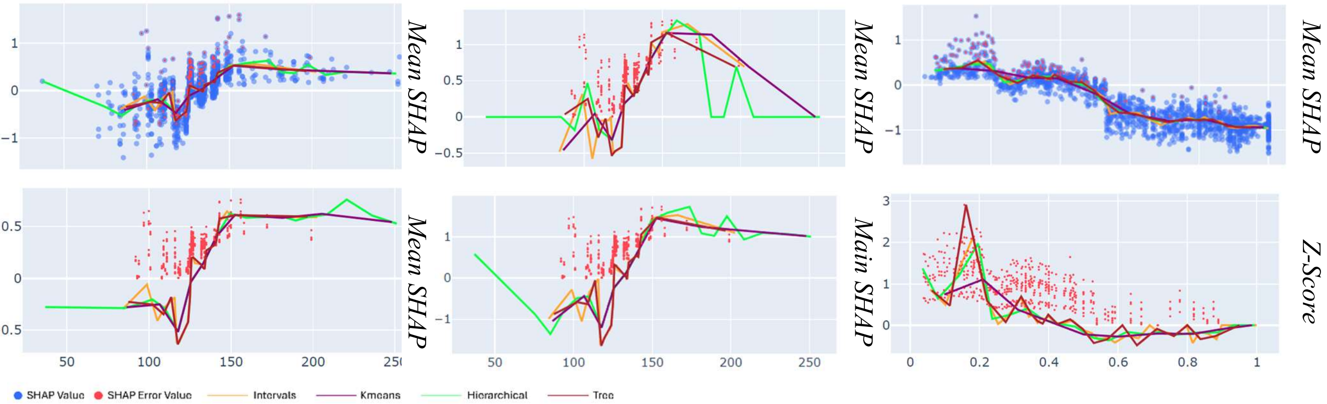

Figure 12. "Maximum_of_Luminosity and Square_Index": Maximum_of_Luminosity (left and mid) is a critical factor for the error analysis within the region [125, 200]. Error SHAP exhibits increased values, Main SHAP indicates that the feature itself influences the effect, with an impact greater than on average, as indicated by Z-Score SHAP. Square_Index (right) shows increased error SHAP in the region of [0.1, 0.2], which indicates a separated errors and identifies the feature as an interesting candidate for explaining errors in this region.

an increased error rate in the region of [0.1, 0.2]. This implies that the feature has separable regions with higher error frequency. Consequently, the feature is an interesting candidate for explaining these errors.

In summary, we evaluated the RFI on the dataset from the steel manufacturing domain using an XGBoost model. The RFI measures, applied across 20 partitions for each feature, provided insights that GFI measures could miss, particularly in identifying specific feature value intervals crucial for error prediction. This nuanced understanding of regional effects is vital for actionable insights in error analysis. The analysis revealed common assignments of importance across features comparing GFI and RFI. This shows that RFI captures the notion of importance as embedded by the GFI, which is global, or model performance-based. Moreover, the RFI revealed interesting features, and the provided interesting insights about patterns that the GFI missed. Understanding these regional nuances can lead to actionable insights for error analysis in manufacturing.

VI. Related Work

Root cause analysis in the production environment has been well studied [19] and several methods for model interpretability through XAI have been reported [20]. 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. In [21] saliency maps and class activation maps are extracted from a deep learning model. In [4] 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 [14]. 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 [13]. 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 [22]. In [23] Permutation Feature Importance is introduced. A global measure where the features are perturbed and the resulting performance loss of the model is taken as a measure of the feature's importance. Mehdiyev and Fettke [18] used Individual Conditional Expectation (ICE) [24] as the global importance. Another method possibly used are Partial Dependence Plots (PDP) [25]. 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 [26]. According to Lundberg [27], the Gini index is equivalent to the in XGBoost [17] implemented importance measure Gain, which *uses the average training loss reduction gained when using a feature for splitting*. Lundberg [27] 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 [28] 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 also LIME [12] could be considered. However, to compute explanations LIME uses sampling which is not restricted to solely interesting areas.

VII. Conclusion

In this paper, we introduced RFI 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, RFIs 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 proposed measures provided detailed insights on features – based on our experience [5] – are interesting; moreover, are partly missed by established methods. Therefore, we conclude that with the help of the proposed importance measures, quality managers get hints about interesting relations that are reflected in the prediction model to drive deeper analysis. Thus, quality managers benefit from adding the proposed importance measures to the pool of XAI methods and we thereby improve XAI for error prediction in manufacturing.

Subject to future work are questions about the integration of RFI in the machine learning pipeline. We assume that the measures are applied at the end of the pipeline, potentially after feature engineering and model optimization. However, the proposed measures may drive the analysis of features earlier in the pipeline as well. Additionally, future work may expand the range of partitioning and aggregation methods to enhance the detection of complex error patterns. Investigating the measures in larger datasets and diverse manufacturing settings could further validate and refine their applicability.

REFERENCES

[1] V. Göttisheim, H. Ziekow, U. Schreier and A. Gerling, "Shapley Values based Regional Feature Importance Measures Driving Error Analysis in Manufacturing", *DATA ANALYTICS 2022, The Eleventh International Conference on Data Analytics*, November, 13 - 17, 2022, Valencia, Spain, pp. 19-26, 2022.

[2] 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.

[3] 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.

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] M. Saarela and S. Jauhiaine, "Comparison of feature importance measures as explanations for classification models", *SN Appl. Sci. 3*, p. 272, 2021.

[16] G. Casalicchio, C. Molnar, and B. Bischl. "Visualizing the Feature Importance for Black Box Models". In: *Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science*, vol 11051. Springer, Cham, 2019.

[17] 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.

[18] M. Buscema, S. Terzi and W. Tastle, "Steel Plates Faults " *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science, 2010. [Online]. Available from: https://doi.org/10.24432/C5J88N.

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

[20] 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.

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

[22] 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.

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

[24] 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.

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

[26] T. Hastie, R. Tibshirani and J. Friedman, "Random forests", *The Elements of Statistical Learning*, Springer, pp. 587–604, 2009.

[27] S. M. Lundberg, "Interpretable Machine Learning with XGBoost", April, 2018. [Online] Available from: https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27, retrieved on 10/06/2024.

[28] 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