Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing

Julian Senoner, Torbjørn Netland, Stefan Feuerriegel

Department of Management, Technology, and Economics (D-MTEC), ETH Zurich, Switzerland

jsenoner@ethz.ch, tnetland@ethz.ch, sfeuerriegel@ethz.ch

We develop a data-driven decision model to improve process quality in manufacturing. A challenge for traditional methods in quality management is to handle high-dimensional and nonlinear manufacturing data. We address this challenge by adapting explainable artificial intelligence to the context of quality management. Specifically, we propose the use of nonlinear modeling with SHAP values to infer how a set of production parameters and the process quality of a manufacturing system are related. Thereby, we contribute a measure of process importance based on which manufacturers can prioritize processes for quality improvement. Grounded in quality management theory, our decision model selects improvement actions that target the sources of quality variation. The decision model is validated in a real-world application at a leading manufacturer of high-power semiconductors. Seeking to improve production yield, we apply our decision model to select improvement actions for a transistor chip product. We then conduct a field experiment to confirm the effectiveness of the improvement actions. Compared to the average yield in our sample, the experiment returns a reduction in yield loss by 21.7%. Furthermore, we report on results from a post-experimental roll-out of the decision model, which also resulted in significant yield improvements. We demonstrate the operational value of explainable artificial intelligence by showing that critical drivers of process quality can go undiscovered by the use of traditional methods.

Key words: manufacturing, quality management, artificial intelligence, SHAP value method

1. Introduction

Manufacturing processes do not always generate outcomes that meet the desired quality specifications. Poor quality creates unnecessary scrap and rework, thereby having a substantial negative impact on financial performance. Moreover, it can lower schedule adherence, increase inventory levels, and make other improvement opportunities less apparent (Ittner 1994). The American Society for Quality estimates that poor quality generates 10% to 15% of the operating expenses in manufacturing companies.[1] Motivated by such figures, manufacturers continuously seek to improve their

---

[1] American Society for Quality. *Cost of Quality (COQ)*. URL: `https://asq.org/quality-resources/cost-of-quality`, last accessed on June 21, 2021.

process performance. To that end, quality management theory suggests to identify and eliminate sources of *quality variation* (Taguchi 1986, Schmenner and Swink 1998, Zantek et al. 2002, Field and Sinha 2005, Hopp and Spearman 2011).

Quality improvement has long been supported by statistical methods (e. g., Shewhart 1926). Existing approaches for identifying sources of quality variation focus on linear associations (e. g., Zantek et al. 2002). However, modern manufacturing settings are characterized by high-dimensional data (Kusiak 2017), which frequently involve nonlinear relationships. For instance, in semiconductor fabrication, manufacturers can collect several thousand interrelated measurements for each individual product unit. When neglecting nonlinearities under high-dimensional conditions, manufacturers may not identify important drivers of process quality. Consequently, there is a need for methods that better accommodate nonlinearities in manufacturing data.

This paper addresses the above shortcoming by developing a data-driven decision model for improving process quality in manufacturing. The decision model follows two steps: first, prioritizing processes for quality improvement and, subsequently, selecting suitable improvement actions. For this purpose, we adapt the Shapley additive explanations (SHAP) value method (Lundberg and Lee 2017) from the field of explainable artificial intelligence (AI) to the context of quality management. By combining SHAP values and nonlinear modeling, we provide a novel measure of process importance that is grounded in quality management theory. Our measure of process importance estimates to what extent the production parameters from a given process contribute to variation in overall process quality. This supports the effective allocation of improvement efforts.

We validate our decision model in a real-world application at Hitachi ABB Power Grids (hereafter referred to as Hitachi ABB), a leading manufacturer of high-power semiconductors. The semiconductor industry is particularly suitable for our research because it is typically subject to high-dimensional data and costly yield losses for which the underlying reasons are often hard to identify. Using historical manufacturing data of a transistor chip product, we apply the decision model to select improvement actions. We confirm the effectiveness of the selected improvement actions in a field experiment. The field experiment shows that, compared to the average yield in our sample, the improvement actions reduce yield loss by 21.7%. After the experimental validation, Hitachi ABB integrated our decision model into their quality management. We report on results from a post-experimental roll-out to a different transistor chip product, which led to a reduction in yield loss of 51.3%.

This paper makes two main contributions. First, we contribute to the practice and literature of quality management by proposing a data-driven decision model that is designed to handle both high-dimensional and nonlinear manufacturing data. For this, we provide a measure of process importance based on which manufacturers can prioritize processes for quality improvement. Second,

we contribute to the research on empirical operations management (cf. Terwiesch et al. 2019) by providing experimental evidence that explainable AI is effective for quality improvement. Overall, our work adds to the emerging literature on data-driven decision making in operations management (cf. Mišić and Perakis 2020, Olsen and Tomlin 2020, Bastani et al. 2021).

The remainder of this paper is structured as follows. In Section 2, we review the literature on quality management and reveal the scarcity of methods designed for nonlinear manufacturing data. In Section 3, we develop a decision model for improving process quality via explainable AI. In Section 4, we apply the decision model in the transistor chip production of Hitachi ABB, and in Section 5, we conduct a field experiment. In Section 6, we implement the decision model into a different production setting at Hitachi ABB. In Section 7, we perform robustness checks and compare our decision model against existing approaches in quality modeling. In Section 8, we discuss implications for quality management, and, in Section 9, we conclude.

2. Related literature

We draw upon two streams of research: quality modeling and explainable AI.

2.1. Quality modeling

Managing process quality has a long history in manufacturing. In the 1920s, statistician Walter A. Shewhart at the Bell Laboratories of Western Electric proposed to analyze data collected from manufacturing processes by means of statistical techniques (Shewhart 1926). Shewhart showed that controlling the manufacturing processes was key to ensuring the quality of the system output. To meet this objective, quality management theory suggests that manufacturers should identify and eliminate the sources of variation; only focusing on the expected outcome of a process or the adherence to tolerance specifications is insufficient (Taguchi 1986, Taguchi and Clausing 1990).

Quality modeling aims at capturing the relationships between a set of production parameters and process quality through analytical methods. The process quality of a manufacturing system is often measured by its output—the products produced. For example, it can be described through physical product measurements (Zantek et al. 2002), predefined quality levels (Chien et al. 2007), or yield (Wu and Zhang 2010). In other cases, process quality can be measured by performance metrics, such as service levels, throughput times, equipment effectiveness, or energy consumption. Production parameters can, for instance, refer to process features (Tsai 2012), material routings (Chen et al. 2005), or (intermediate) product properties (Zantek et al. 2002). Typically, these production parameters belong to different processes. Understanding how various production parameters and process quality relate is crucial for improving the performance of a manufacturing system.

A common approach in quality modeling is to learn a functional relationship between process quality and the observed production parameters. The learned representation can be viewed as

a meta-model, in which the underlying physical mechanisms within a manufacturing system are reproduced (cf. Yu and Popplewell 1994). Examples in the literature include association rule mining (e. g., Chen et al. 2005) and decision trees (e. g., Chien et al. 2007). By interpreting the functional relationships, one can inform the choice of potential improvement actions, but these do not directly target the sources of quality variation. Therefore, the most effective improvement actions may not be selected.

Aligned with quality management theory, there are approaches that first aim at identifying sources of variation and, subsequently, allocate resources to improve process quality. An example is provided by Zantek et al. (2002). The authors model process quality as a linear combination of production parameters and quality variables measured at intermediate inspections throughout a manufacturing system. The sources of variation are then identified by decomposing the variance of each quality variable into components. The work by Zantek et al. (2002) makes two assumptions that limit its applicability in practice. First, it assumes that the underlying relationships are linear; therefore, potential nonlinearities are neglected. Second, it assumes that quality variables are measured at intermediate inspections in-between processes. However, due to physical dependencies, process quality can often only be measured at the final stage of a manufacturing system (e. g., yield). We address these shortcomings by adapting explainable AI to the context of quality management.

2.2. Explainable AI

Nonlinear modeling is often used to capture complex relationships in high-dimensional operational settings (e. g., Cui et al. 2018). However, the underlying decision rules of nonlinear models are not always self-explanatory. As a remedy, various approaches have been proposed to understand how inferences from nonlinear models are formed (Guidotti et al. 2018). Approaches that leverage post hoc explanations of model predictions can be subsumed under the term "explainable AI." Two concepts in explainable AI are relevant to our work: feature importance and feature attribution.

Feature importance measures the extent to which a feature is responsible for forming a functional logic. A common approach is to compare the performance of a predictive model against one without a specific feature of interest while considering all possible nonlinear relationships (i. e., ablation study). A similar approach is to measure the difference in prediction performance when a feature is permuted randomly so that its influence is omitted (Breiman 2001). There are also feature importance measures tailored to specific models including decision trees (Breiman et al. 1984) and support vector machines (Guyon et al. 2002).

Feature attribution infers the marginal effect of a feature on a model prediction. Specifically, it estimates how the model prediction changes with a feature of interest. For instance, in linear

modeling, the marginal effect of a feature is quantified by the coefficients. In nonlinear modeling, feature attribution can be estimated via partial dependence plots (Friedman 2001) or locally interpretable model-agnostic explanations (LIME for short; Ribeiro et al. 2016).

A combination of both feature importance and feature attribution is given by the SHAP value method (Lundberg and Lee 2017). The SHAP value method infers the underlying decision rules of predictive models by decomposing a prediction into the contribution (called the "SHAP value") of each feature. A summary of the SHAP value method is provided in Appendix A. Instead of explaining model predictions, we leverage SHAP values to infer how various production parameters and the process quality of a manufacturing system are related.

3. Decision model

In this section, we develop a data-driven decision model to improve process quality in manufacturing. We define a formal manufacturing setting, give our problem description, and provide the model specification.

3.1. Manufacturing setting

We consider a manufacturing system with sequential processes (see Figure 1).[2] Each process is specified by production parameters that potentially influence the performance of the manufacturing system. The overall performance outcome is measured by a process quality variable (e. g., yield), which is observed at the final stage of the manufacturing system.
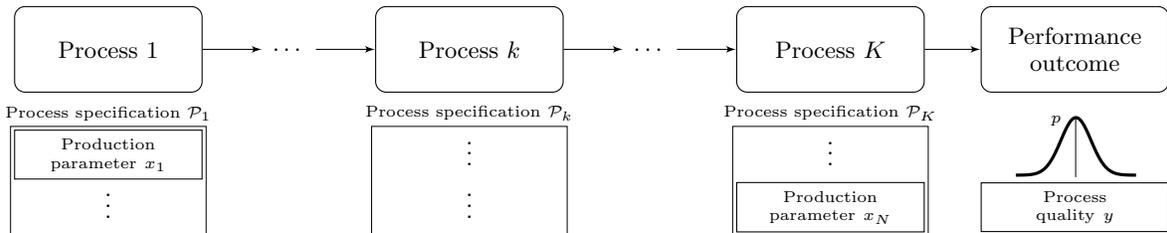

Figure 1   Manufacturing system

The manufacturing system generates data on production parameters x and process quality outcomes y. Overall, there are $j = 1, \ldots, N$ production parameters and $i = 1, \ldots, M$ observations (e. g., product units). For notation, we use superscript indices when referring to observations and subscript indices when referring to production parameters. Formally, let $y^{(i)} \in \mathbb{R}$ refer to the measured process quality of the i-th observation. Further, let $x^{(i)} \in \mathbb{R}^N$ denote the observed production parameters of the i-th observation with production parameters $j = 1, \ldots, N$. Analogously, let $x_j \in \mathbb{R}^M$ denote

---

[2] Our approach can also be extended to manufacturing systems where sub-components are processed in parallel. For this, the sub-components must be considered as one product observation.

the j-th production parameter across observations $i = 1, \ldots, M$. The single value of a production parameter is given by $x_j^{(i)}$.

The production parameters are captured at different processes $k = 1, \ldots, K$. Here, let the process specification $\mathcal{P}_k \subseteq \{1, \ldots, N\}$ define which specific production parameters belong to a certain process k. Each production parameter is associated with exactly one process; that is, $\mathcal{P}_{k'} \cap \mathcal{P}_{k''} = \varnothing$ for $k' \neq k''$ and $\bigcup_k \mathcal{P}_k = \{1, \ldots, N\}$.

3.2. Problem description

Our objective is to allocate improvement actions to the processes associated with a large influence on the system performance (i. e., the overall process quality outcome). To achieve this, quality management theory suggests targeting the sources of variation (Taguchi 1986, Schmenner and Swink 1998, Zantek et al. 2002, Hopp and Spearman 2011). In practice, any manufacturing system generates outcomes that are subject to varying quality (see distribution p in Figure 1). Without loss of generality, we assume that high process quality is preferable. The spread around the process quality mean is a proxy for the improvement potential of the manufacturing system. Essentially, when there is variation in process quality, there exist opportunities to learn from better (right tail of the distribution) and avoid worse outcomes (left tail of the distribution). Thereby, manufacturers can shift future outcomes from the negative tail to the positive tail of the distribution. Therefore, by targeting processes responsible for variation in overall process quality, manufacturers can implicitly improve the overall process quality mean.
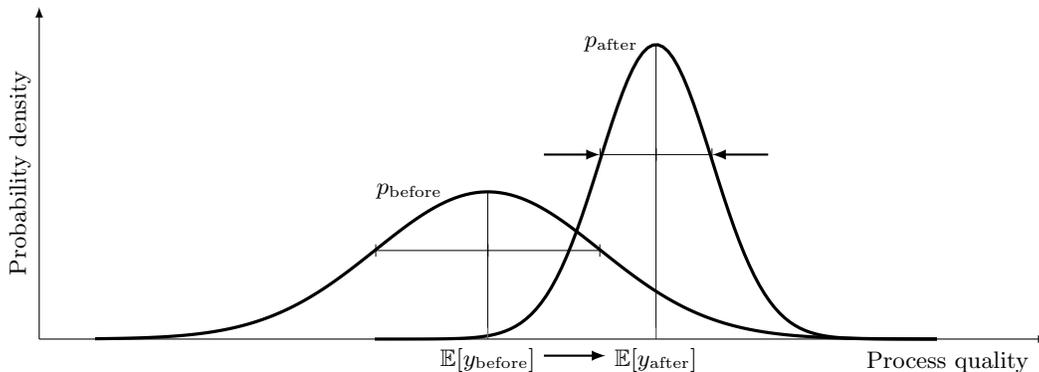

Figure 2   Illustrative example of quality improvement

Consider Figure 2, which compares the process quality of a manufacturing system before and after implementing improvement actions. To improve process quality, one first identifies the underlying processes that are associated with the largest contribution to variation around the overall process quality mean. For these processes, avoiding outcomes that are associated with below-average process quality (i. e., left tail of distribution p_{before}) leads to two coupled effects: As evident

from distribution p_{after}, it reduces the inherent variation in process quality and improves the mean process quality simultaneously (i. e., shifts the distribution to the right and makes it narrower). As a result, the manufacturing system becomes more robust and generates better quality outcomes.

Supposedly, one could also identify actions that improve the mean process quality "directly." However, this approach does not work in quality management, because optimizing against the mean would require modeling production parameters and possible improvement actions jointly. For this, one would have to define all possible improvement actions for every process in a manufacturing system—even those that will eventually not be prioritized for improvement. This is highly impractical and expensive due to myriad possible actions that have to be compiled by domain experts. Therefore, manufacturers instead choose to first identify the processes with the largest improvement potential and then only determine improvement actions for those processes.

The implementation of improvement actions depends on the characteristics of the manufacturing processes. For some production parameters, it may be possible to manipulate the absolute parameter values directly. For example, if the temperature in a certain process is associated with an influence on process quality, an improvement action can be to adjust the temperature levels. However, in some processes, production parameters cannot be modified directly. In this particular case, there can be global improvement actions that affect all production parameters from the same process. For instance, if production parameters depend on the production equipment used, a common solution is to change the material routing through processes.

Following the above rationale, decisions about quality improvement are generally informed by a two-step approach (cf. Zantek et al. 2002). In the first step, manufacturers aim at prioritizing the processes associated with the largest contribution to variation in the overall process quality. In the second step, for the prioritized processes, manufacturers determine possible improvement actions and select the ones that promise the largest quality improvement. In theory, each process that contributes to variation in process quality should be targeted with improvement actions. However, in practice, there are limited resources that can be allocated, thus restricting improvement actions to a maximum number of p_{max} processes.

This paper formalizes the above two-step approach for quality improvement into a data-driven decision model. Because processes are potentially interdependent, the decision model must account for nonlinear interdependencies. For example, such nonlinearities can appear if a production parameter $x_{j'}$ from a process k' interacts with a production parameter $x_{j''}$ from another process k''. This motivates a nonlinear modeling approach, as introduced in the following.

3.3. Model specification

We now specify our two-step decision model for improving process quality in manufacturing (Figure 3). The initial input for the decision model is given by historical manufacturing data

$\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$. Based on this input, we learn a nonlinear meta-model f that reproduces the associations between production parameters $x^{(i)}$ and process quality $y^{(i)}$. In Step 1, the decision model utilizes the process specifications $\mathcal{P}_k$ and feature attributions $\phi_j^{(i)}$ to return a process prioritization p^*. For this, we develop a measure of process importance that estimates the extent to which the production parameters from a given process are associated with variation in overall process quality. For each prioritized process, decision makers must compile a set of candidate actions $\mathcal{A}_k$. In Step 2, the decision model uses feature attributions $\phi_j^{(i)}$ to select those improvement actions $a_k^* \in \mathcal{A}_k$ with the largest estimated effect on overall process quality.
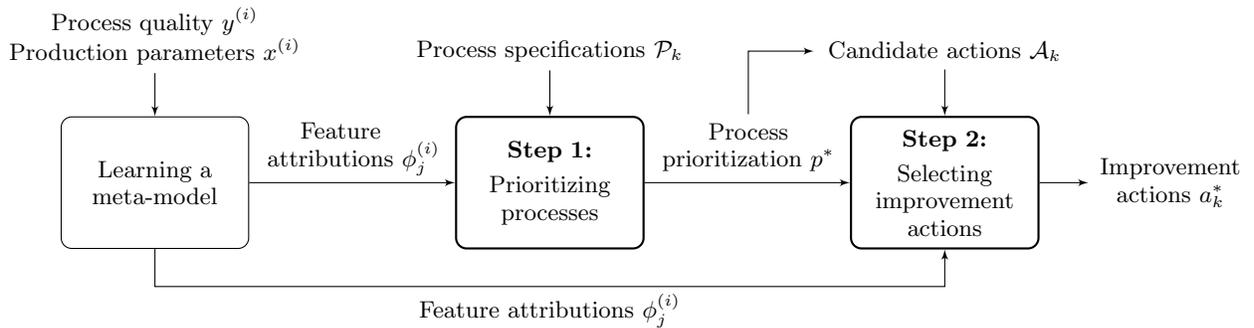
Notes: The figure shows our decision model for improving process quality in manufacturing. The decision model (formalized in sections 3.3.1 to 3.3.3) is empirically validated at Hitachi ABB (sections 4 to 7). Section 4 describes the empirical context (sections 4.1 to 4.3), provides the implementation details of the meta-model (Section 4.4), and reports the numerical results of Step 1 (Section 4.5.1) and Step 2 (Section 4.5.2). The improvement actions, which are returned by the decision model, are experimentally validated in Section 5. Section 6 provides evidence from a post-experimental roll-out where we apply the decision model to a different production setting at Hitachi ABB. Section 7 reports robustness checks in which we implement different meta-models (Section 7.1) and compare the decision model against linear methods (Section 7.2) and a decision tree heuristic (Section 7.3). Additional numerical examples are included in the Online Supplements.

Figure 3   Decision model

3.3.1.   Learning a meta-model. The basis for the decision model is a meta-model $f : \mathbb{R}^N \to \mathbb{R}$ that is estimated based on past observations of production parameters and process quality outcomes. This can be an arbitrary predictive model f that can emulate high-dimensional and nonlinear relationships (e. g., tree ensembles, deep neural networks). Model f is estimated with the objective of minimizing the error between the true and estimated process quality; that is,

$$\min_{f} \mathbb{E}[\ell(y, f(x))], \tag{1}$$

where ℓ is a convex loss function (e. g., mean squared error). Provided f is well specified, we obtain a meta-model of the physical processes that allows explaining how various production parameters and process quality are related.

We infer the underlying relations in a manufacturing system via the SHAP value method (Lundberg and Lee 2017, Lundberg et al. 2020). Specifically, we use SHAP values (for details, see

Appendix A) to provide explanations of how the estimated process quality changes when the effect of a production parameter is omitted. For this, the SHAP value method explains model f locally at each observation i. The explanation is formally given by additive feature attributions that sum to the output of the meta-model; that is, $f(x^{(i)}) = \phi_0 + \sum_{j=1}^{N} \phi_j^{(i)}$, where $\phi_0 = \mathbb{E}[f(x)]$ and $\phi_j^{(i)} \in \mathbb{R}$ corresponds to the SHAP value of the observed production parameter $x_j^{(i)}$. In our setting, a SHAP value gives the estimated deviation from the expected process quality $\mathbb{E}[f(x)]$ that can be attributed to an observed production parameter $x_j^{(i)}$. Here, negative SHAP values indicate a decrease in the estimated process quality, whereas positive SHAP values indicate an increase in the estimated process quality. The larger the absolute SHAP value, the larger the estimated change in process quality. The computation of SHAP values is repeated for all observations, thereby returning feature attributions $\phi_j^{(i)}$ for $i = 1, \ldots, M$ and $j = 1, \ldots, N$.

REMARK 1. *The SHAP value method is the only additive feature attribution method that satisfies missingness, consistency, and local accuracy (Theorem 1 in Lundberg and Lee 2017).*

SHAP values guarantee three properties: (1) missingness, (2) consistency, and (3) local accuracy (see Remark 1). In our setting, missingness assures that absent production parameters have no feature attribution. Consistency is required to make meaningful comparisons of feature attributions across production parameters. Local accuracy ensures that feature attributions sum to the model output; that is, $f(x^{(i)}) = \phi_0 + \sum_{j=1}^{N} \phi_j^{(i)}$, and thus give an estimate of changes in process quality.

3.3.2. Step 1: Prioritizing processes. Step 1 of the decision model determines the processes that should be prioritized for improvement actions. Here, we are interested in identifying the processes associated with a large estimated contribution to variation in overall process quality. For this, we make use of the previously computed feature attributions $\phi_j^{(i)}$, which quantify the estimated deviation from the expected process quality attributed to each production parameter at the observation level (i.e., local understanding). However, prioritizing processes for quality improvement requires an importance measure at the process level (i.e., global understanding). To achieve this, we develop our measure of process importance. We proceed by aggregating the absolute feature attributions from the observation level onto the production parameter level and then from the production parameter level onto the process level.

We first aggregate the absolute feature attributions at the production parameter level as a measure of feature importance. Here, a large feature importance points toward production parameters with a large estimated contribution to variation in the overall process quality. Later, we refer to these production parameters as quality drivers. Formally, the decision model computes the mean absolute feature attributions (i.e., feature importance) across all observations via

$$\Phi_j = \frac{1}{M} \sum_{i=1}^{M} |\phi_j^{(i)}| \ \text{ for all } \ j = 1, \ldots, N. \tag{2}$$

We then aggregate the mean absolute feature attributions at the process level. Because contrary effects of production parameters within a process should not cancel out, we aggregate the absolute values. This is given by our definition of process importance, which quantifies the estimated contribution to variation in overall process quality that can be attributed to the production parameters from a given process under the meta-model f.

DEFINITION 1 (PROCESS IMPORTANCE). *The process importance of process k is given by*

$$\Theta_k = \sum_{j \in \mathcal{P}_k} \Phi_j, \tag{3}$$

where Φ_j is the mean absolute feature attribution of the j-th production parameter.

Our measure of process importance has two desirable characteristics. First, due to the local accuracy property of SHAP values, it provides a meaningful estimate of the contribution to variation in overall process quality. Therefore, it is aligned with quality management theory, which suggests targeting sources of quality variation. Second, it accounts for interaction effects because feature attributions are computed over all possible subsets of production parameters. This is different from ablation studies where the importance of two highly informative and perfectly correlated features would be underestimated.

Using the process importance from Definition 1, we identify the processes with the largest estimated contribution to variation in overall process quality via

$$p^* \in \underset{p \in \{0,1\}^K}{\arg\max} \sum_{k=1}^{K} p_k \, \Theta_k \quad \text{s.t.} \sum_{k=1}^{K} p_k \le p_{\max}. \tag{4}$$

Here, p^* is a K-dimensional binary vector that specifies which processes should be prioritized for improvement actions. The additional constraint limits the number of processes that a manufacturer prioritizes for improvement actions to $p_{\max}$. While this modeling step has determined which processes should be prioritized, the next step selects actions for quality improvement.

3.3.3. Step 2: Selecting improvement actions. Step 2 of the decision model selects improvement actions for the prioritized processes. For every prioritized process k, decision makers must compile a set of possible candidate actions $\mathcal{A}_k = \{a_k^1, \ldots, a_k^{Q_k}\}$, where Q_k defines the number of candidate actions for process k. The availability of actions depends on the degrees of freedom modifiable within a prioritized process. Continuous production parameters (e. g., temperature) can be addressed by discretization. The choice of candidate actions can be informed by analyzing the relationships between the observed production parameters and the feature attributions (i. e., $\{(x^{(i)}, \phi^{(i)})\}_{i=1}^{M}$) or by leveraging domain expertise. There may also be candidate actions that affect more than one production parameter from a process. For example, one can choose to prioritize

certain machines over other machines from the same process. In this case, the possible candidate actions are simply the set of machines that can be used to carry out that process.

Due to complexities and interdependencies in manufacturing systems, improvement actions have to be tailored to specific processes to be most effective. We, therefore, select improvement actions at the process level. For each prioritized process, the objective is to select the action $a_k^* \in \mathcal{A}_k$ for which the estimated effect on process quality (i.e., $\mathbb{E}[f(x) \mid a_k^*]$) is maximal.

To quantify the marginal effects of all possible candidate actions on the estimated process quality, we draw upon the local accuracy property of SHAP values. Specifically, we aggregate the process-level feature attributions for each action. Note that this requires that actions have been observed previously. Let $\mathcal{I}_k^q \subseteq \{1, \ldots, M\}$ define which in-sample observations i have received treatment in the form of an action a_k^q. Then, the mean feature attribution at the action level is computed via

$$\Psi_k^q = \frac{1}{|\mathcal{I}_k^q|} \sum_{i \in \mathcal{I}_k^q} \sum_{j \in \mathcal{P}_k} \phi_j^{(i)} \quad \text{for all} \quad q = 1, \ldots, Q_k \quad \text{in prioritized process } k. \tag{5}$$

This quantifies the average change in the estimated process quality that can be attributed to an action. Positive values of Ψ_k^q are associated with an increase in the estimated process quality, whereas negative values are associated with a decrease in the estimated process quality. Therefore, the action with the largest mean feature attribution gives the largest estimated quality improvement under meta-model f. Moreover, the action accounts for nonlinearities because the feature attributions have previously been computed over all possible subsets of production parameters.

The decision model selects an improvement action via

$$a_k^* \in \arg\max_q \Psi_k^q. \tag{6}$$

The procedure can be repeated to select actions a_k^* for all prioritized processes.

4. Empirical application

We now validate the proposed decision model in the semiconductor industry. Semiconductor manufacturing generally involves several hundred processes that are interrelated and often take months to complete.[3] Owing to the high complexity of the fabrication procedure, identifying quality drivers is challenging. Therefore, semiconductor manufacturers frequently face considerable yield losses that substantially affect their financial performance. Manufacturers are usually cautious about reporting their yield, but estimates range between 85% and 95%.[4] Against this background, improving process quality promises to have a major economic impact.

---

[3] Hitachi. *Semiconductor - Metrology and Inspection.* URL: `https://www.hitachi-hightech.com/global/products/device/semiconductor/metrology-inspection.html`, last accessed on June 21, 2021.

[4] McKinsey & Company (2017). *Reimagining fabs: Advanced analytics in semiconductor manufacturing.* URL: `https://www.mckinsey.com/industries/semiconductors/our-insights/reimagining-fabs-advanced-analytics-in-semiconductor-manufacturing`, last accessed on June 21, 2021.

4.1. Empirical setting

Our research is carried out at Hitachi ABB in Lenzburg, Switzerland. Hitachi ABB is a leading manufacturer of transistors, thyristors, and diodes for applications in power transmission, transportation, electrical drives, and renewable energy. The objective of our research is first to prioritize processes for quality improvement and, subsequently, to select suitable improvement actions. For this purpose, we receive historical manufacturing data from a transistor chip product. Despite high capacity utilization and considerable fabrication costs, Hitachi ABB agreed to experimentally validate the two most promising improvement actions in their transistor chip production.

The transistor chip production at Hitachi ABB consists of 200 processes that are carried out in a low-vibration and temperature-constant clean room. The raw material input for the transistor chips is a thin silicon slice called a "wafer." The production is organized based on the batch principle, where 25 wafers form an entity (see Figure 4). Wafers in the same production batch are processed together and do not move to the next process until all wafers are finalized. During fabrication, the electrical properties of the wafers are modified by introducing impurities into the silicon crystal. Once a production batch has been completed, each wafer is cut into 62 rectangular transistor chips that form the final product. As part of the quality testing, which corresponds to the last stage in the fabrication procedure, each transistor chip is exposed to realistic field conditions to measure its quality via electrical response variables, such as currents and voltages. If a transistor chip does not meet the required quality specifications, it is scrapped at a cost.
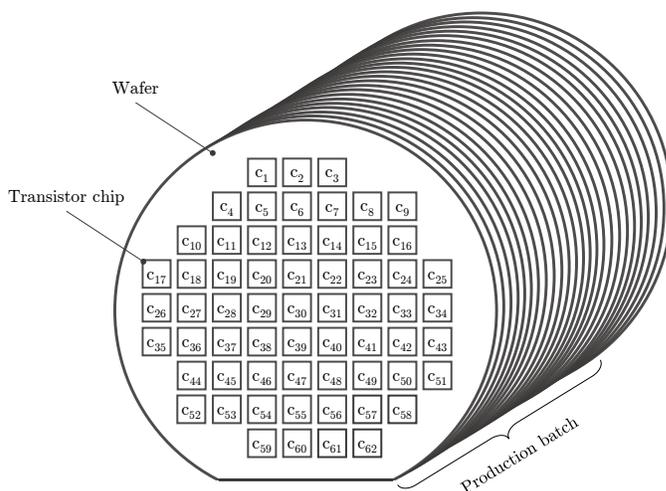

Figure 4   Production output

The main process types are impurity doping, oxidation, photolithography, etching, and metallization, as described in the following. (1) Impurity doping is applied to alter the conductivity of

the wafer and is performed by diffusion and implantation processes. (2) Oxidation processes are required to grow thin layers of silicon dioxide on top of the wafer surface. These films serve as insulators and are essential to specify the implantation regions during impurity doping. (3) Photolitography exposes the wafer to UV light to transfer the geometrical pattern of the semiconductor devices to an underlying layer. (4) Etching processes are applied to selectively remove material from the wafer surface. (5) Metallization is utilized to deposit conductive layers on both the front and back of the wafer to interconnect the electronic circuits of the devices. Each process type is executed several times with many interdependent production parameters. All production parameters potentially influence process quality. Additionally, there may be critical combinations of production parameters that can trigger undesired interaction effects. For example, a machine from a given process may only induce quality issues if it is used in combination with another machine from a different process. Therefore, identifying quality drivers in this setting requires methods that can handle nonlinearities.

4.2. Operational data

Hitachi ABB provided us with historical data on $M = 1{,}197$ production batches (approximately 1.8 million transistor chips) produced prior to April 2019. The fabrication conditions of each production batch are described by $N = 3{,}614$ production parameters from $K = 200$ different processes. Hitachi ABB also provided us with process specifications $\mathcal{P}_k$ that define which production parameters belong to a certain process. In addition, each production batch i is associated with a quality variable (from the electrical testing) given by the yield $\mu^{(i)} \in \mathbb{R}$. The yield for a given production batch is defined as the ratio between the number of transistor chips that met the required quality specifications and the number of chips inspected. The company protected confidential information by scaling the yield variable between 0 and 100; that is, $y^{(i)} = 100 \times \frac{\mu^{(i)} - \mu_{\min}}{\mu_{\max} - \mu_{\min}}$. This normalization maintains the distributional pattern and still allows us to later report the improvements actually achieved.

4.3. Descriptive statistics

The distribution of the normalized yield in our sample is shown in Figure 5. The average normalized yield is 82.1 (standard deviation of 12.3). Approximately 50% of the production batches have a normalized yield above 85.5. According to Hitachi ABB, production batches with a normalized yield below 65 can be considered "low performers." In our dataset, this corresponds to approximately 10% of the production batches. The aim of our decision model is to select improvement actions that address the long tail of the distribution.
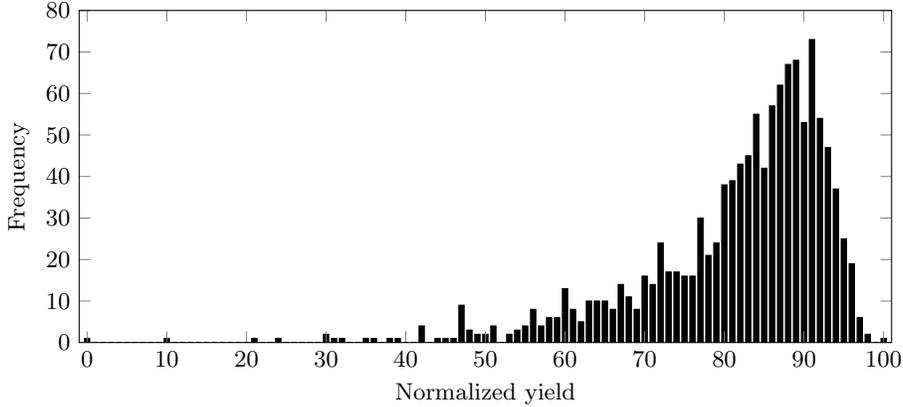
Figure 5 **Histogram of normalized yield across production batches**

Table 1 lists exemplary production parameters captured in the fabrication processes. For confidentiality reasons, we later only refer to the anonymized production parameters (i. e., x_j). In general, we distinguish production parameters at the process and product levels. Process parameters describe machine-related properties (e. g., the average pressure measured in a machine), whereas product parameters relate to physical product characteristics during fabrication (e. g., the number of particles counted in a production batch). All production parameters are measured before quality testing and are thus potential quality drivers.

Table 1 **Exemplary production parameters**

Description	Production parameters	Unit	Level
Average pressure	x_{282}, x_{345}, x_{649}, x_{712}, x_{1213}	mTorr	Process
Average gas flow	x_{294}, x_{353}, x_{507}, x_{661}, x_{724}, x_{784}, x_{1166}, x_{1194}, x_{1225}, x_{3121}	sccm	Process
Average scan speed	x_{308}, x_{367}, x_{515}, x_{675}, x_{738}, x_{792}, x_{1174}, x_{1202}, x_{1239}, x_{3129}	$\mathrm{mm\,s^{-1}}$	Process
...	...	...	...
Layer thickness	x_{45}, x_{46}, x_{47}, x_{3369}, x_{3370}, x_{3371}, x_{3594}, x_{3595}, x_{3596}	Å	Product
Particle count	x_{170}, x_{171}, x_{172}, x_{252}, x_{253}	—	Product
Scratch count	x_{197}, x_{198}, x_{199}, x_{272}, x_{273}	—	Product
...	...	...	...

We also assess the Pearson correlation coefficients between the non-constant production parameters and the normalized yield (Figure 6). The absolute Pearson correlation coefficient exceeds the value of 0.25 only for one of the 3,614 production parameters. This corresponds to production parameter x_{302} from process 25, for which the correlation coefficient amounts to 0.54. The low correlation between the production parameters and the normalized yield can have two reasons: either the production parameters are not quality drivers at all, or there are nonlinear relationships

not captured by the correlation coefficients. The latter may, for example, occur if different quality drivers are interrelated across production processes. This motivates a nonlinear modeling approach to identify quality drivers. As demonstrated later, our decision model locates quality drivers beyond production parameter x_{302} for which we also confirm a critical influence on the normalized yield.
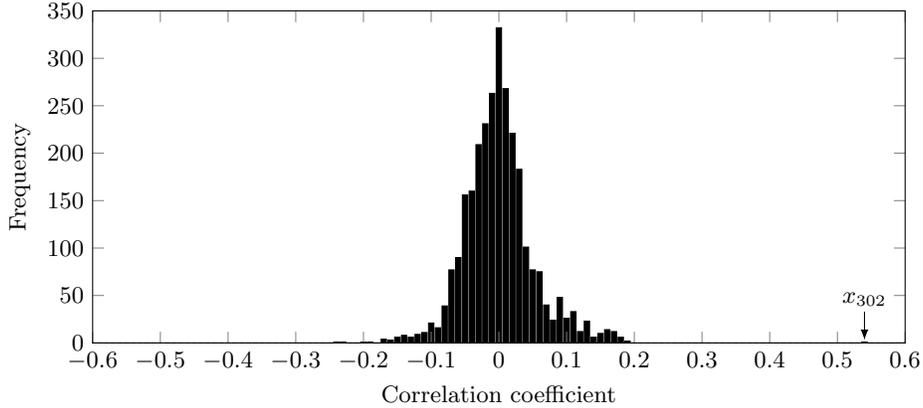

Figure 6   **Histogram of correlation coefficients between production parameters and normalized yield**

4.4.   Implementation details of the meta-model

The meta-model is estimated based on all production parameters and the normalized yield using gradient boosting with decision trees (Ke et al. 2017). Gradient boosting belongs to the category of tree ensemble algorithms, which are known for performing well on complex datasets and have already been applied in other operational applications (e. g., Cui et al. 2018, Sun et al. 2020). We utilize common procedures (cf. Hastie et al. 2009) and split our data into a training set (80% of the data) for estimating parameters and a holdout set (remaining 20%) for evaluating modeling performance. The training set contains 957 production batches, and the holdout set contains 240 production batches. The meta-model is trained and tuned only based on the training set (grid search with cross-validation for hyperparameter tuning; see Appendix C). A comparison with alternative meta-models is provided in Appendix B.

We compute the feature attributions of all production parameters with the tree implementation of the SHAP value method (see Lundberg et al. 2020, for details). This implementation utilizes the structure of tree-based models to compute SHAP values with high efficiency. On conventional office hardware (Intel Core i7-8550U processor with 1.8 GHz), the computation of feature attributions for the entire training set takes around one second.

4.5.   Numerical results

In the following, we report the results from our empirical application. To identify improvement potentials, we first locate the processes associated with the largest estimated contribution to variation in the normalized yield (Step 1). Recall that, due to high capacity utilization and fabrication

costs, Hitachi ABB agreed to only validate two improvement actions experimentally. Therefore, we set $p_{\max} = 2$ in the decision model, thereby prioritizing two processes. For these processes, we compile a set of candidate actions $\mathcal{A}_k$ together with the process engineers at the company. Then, the decision model selects the actions that promise the largest estimated yield improvement (Step 2).

4.5.1.   Step 1: Prioritizing processes. The decision model now determines the two processes associated with the largest estimated contribution to variation in the normalized yield. First, the mean absolute feature attribution Φ_j is returned for each production parameter. Around 99% of the production parameters have a mean absolute feature attribution $\Phi_j < 0.1$ and are thus hardly relevant in describing variation in the normalized yield. The top 10 quality drivers in the transistor chip production are listed in Table 2. Overall, six of the top 10 production parameters stem from the same implantation process (process 25). This includes production parameter x_{302}, for which we previously determined a high correlation with the normalized yield (cf. Section 4.3; note that we here report the correlation in the training set). Therefore, it is not surprising that our decision model lists production parameter x_{302} among the top 10 quality drivers. In contrast, the other quality drivers have a small correlation with the normalized yield. This suggests the presence of nonlinear relationships not captured by linear correlation.

Table 2   Top 10 quality drivers ranked by mean absolute feature attribution

Rank	Production parameter	Process	Process type	Correlation coefficient	Mean absolute feature attribution Φ_j
1	Parameter x_{302}	Process 25	Implantation	0.52	0.74
2	Parameter x_{313}	Process 25	Implantation	0.02	0.58
3	Parameter x_{2357}	Process 166	Etching	0.09	0.53
4	Parameter x_{889}	Process 83	Photolithography	0.02	0.40
5	Parameter x_{1854}	Process 145	Etching	0.19	0.39
6	Parameter x_{295}	Process 25	Implantation	0.09	0.38
7	Parameter x_{279}	Process 25	Implantation	0.01	0.34
8	Parameter x_{405}	Process 36	Photolithography	-0.11	0.33
9	Parameter x_{280}	Process 25	Implantation	-0.07	0.30
10	Parameter x_{340}	Process 25	Implantation	0.01	0.29
...	...	...	...	...	...

Notes: The correlation between the production parameters and the normalized yield is reported based on the Pearson correlation coefficient in the training set.

The decision model computes the process importance Θ_k (i.e., mean absolute feature attribution at the process level) and then returns the two processes that should be prioritized for improvement actions. The output suggests that process 25 and process 166 are associated with the largest estimated contribution to variation in the normalized yield. The distribution of process importance

values for all 200 processes is shown in Figure 7. The decision model indicates that, for 159 processes, the attributed influence on the normalized yield is negligible ($\Theta_k < 0.1$). In contrast, the two prioritized processes explain a comparably large portion of the variation in the normalized yield.
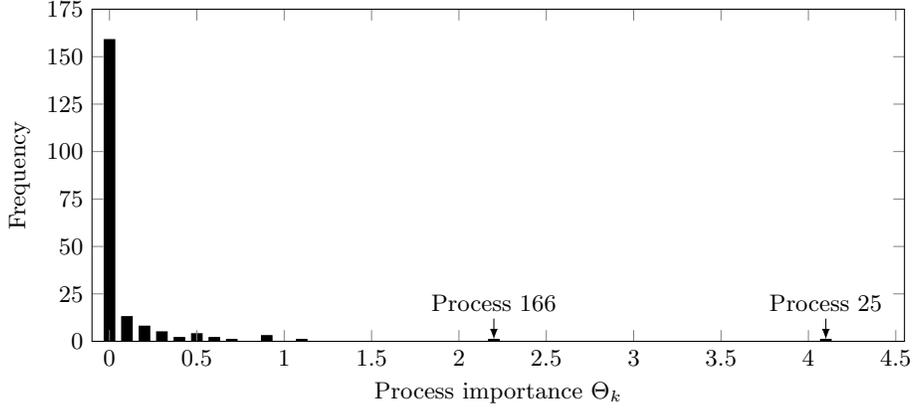

Figure 7   Histogram of process importance values

Table 3 provides more detailed information about the two prioritized processes. Process 25 is an implantation process, comprising 63 measured production parameters with a process importance of 4.18. Process 166 is an etching process, comprising 173 measured production parameters with a process importance of 2.30. Both processes can be carried out by different machines (i. e., implantation process 25 by machines QI613, QI614, or QI615 and etching process 166 by machines QP211, QP212, QP232, or QP233). Next, we will compile a set of candidate actions and select the actions that promise the largest estimated yield improvement.

Table 3   Overview of prioritized processes

	Process 25	Process 166
Process type	Implantation	Etching
Associated production parameters	$x_{278}, \ldots, x_{340}$	$x_{2197}, \ldots, x_{2369}$
Associated machines	QI613, QI614, QI615	QP211, QP212, QP232, QP233
Process importance Θ_k	4.18	2.30

4.5.2.   Step 2: Selecting improvement actions. Possible improvement actions depend on whether a production parameter can be changed. In our setting, the production parameters in process 25 and process 166 are measurements that cannot be directly manipulated. The process engineers at Hitachi ABB suggested that the measured production parameters can depend on the production equipment used, and, therefore, possible improvement actions involve a change in the material routing. This is achieved by altering the corresponding prioritization of machines so that

one machine is preferred over others from the same process. Accordingly, the set of candidate actions $\mathcal{A}_k$ contains the available production equipment for both prioritized processes, respectively. Next, the decision model will assess whether there are performance differences between the production equipment used.

The decision model now selects improvement actions in the form of machine prioritizations. The candidate actions in both processes are given by the machines that can be selected. In process 25, the set of candidate actions is defined as $\mathcal{A}_{25} = \{a_{25}^1, a_{25}^2, a_{25}^3\}$ with machine QI613 (a_{25}^1), machine QI614 (a_{25}^2), and machine QI615 (a_{25}^3). In process 166, the set of candidate actions is defined as $\mathcal{A}_{166} = \{a_{166}^1, a_{166}^2, a_{166}^3, a_{166}^4\}$ with machine QP211 (a_{166}^1), machine QP212 (a_{166}^2), machine QP232 (a_{166}^3), and machine QP233 (a_{166}^4). For each action, the decision model returns the mean feature attribution, which quantifies the estimated average influence on the normalized yield (Table 4). Here, positive values refer to actions that have a positive association with the normalized yield, whereas negative values refer to a negative association with the normalized yield. The decision model estimates that action a_{25}^3 and action a_{166}^2 are associated with a positive influence on the normalized yield and are thus selected as improvement actions. For both prioritized processes, a Kruskal–Wallis test confirms that the feature attributions of the candidate actions differ at a statistically significant level ($p < 0.001$). This provides strong evidence of machine-related performance differences.

Table 4   Mean feature attributions per action

Candidate action	Process 25			Process 166			
	a_{25}^1	a_{25}^2	a_{25}^3	a_{166}^1	a_{166}^2	a_{166}^3	a_{166}^4
Associated machine	QI613	QI614	QI615	QP211	QP212	QP232	QP233
Mean feature attribution Ψ_k^q	−13.40	−11.42	2.40	−0.86	1.57	−12.54	−13.48
Action selected	No	No	Yes	No	Yes	No	No
H-statistic	374.11***			694.55***			

Significance level: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$
Notes: The H-statistic is reported based on a Kruskal–Wallis test.

As improvement actions, the decision model suggests selecting machine QI615 (a_{25}^3) in process 25 and machine QP212 (a_{166}^2) in process 166 improves the normalized yield. Figure 8 shows the suggested material routing through the two prioritized processes. Hitachi ABB confirmed that the selected improvement actions do not introduce any new capacity constraints in their transistor chip production. The remaining machines can be used to perform other implantation and etching processes, which, according to our decision model, have no estimated influence on the normalized yield. In the following, we validate the two improvement actions by conducting a field experiment.
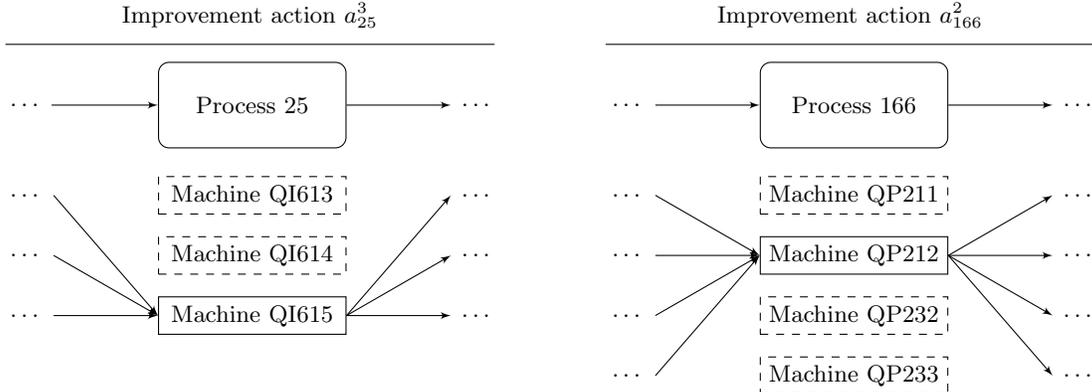
Improvement action a^3_{25} Improvement action a^2_{166}

Figure 8   **Suggested material routing through prioritized processes**

5. Experimental validation of selected improvement actions

The effectiveness of the two selected improvement actions is validated as follows. First, we determine the projected treatment effect by statistically analyzing historical production batches in the holdout set. Then, we conduct a field experiment at Hitachi ABB to demonstrate that the projected treatment effect is achieved after implementation.

5.1. Projected treatment effect

To avoid overfitting, the projected treatment effect for the two selected improvement actions must be based on observations that were not included in the estimation of meta-model f. For this, we consider the 240 production batches in the holdout set. Figure 9 compares the normalized yield of the production batches that received treatment in the form of action a^3_{25} and action a^2_{166} (i.e., processed by machine QI615 in process 25 and machine QP212 in process 166) with the production batches processed differently. The plot suggests that the production batches that were processed with action a^3_{25} and action a^2_{166} are associated with a considerably larger normalized yield. The projected treatment effect is given by the mean difference in normalized yield, amounting to 22.7 units. A Welch's t-test confirms that this difference is statistically significant ($p < 0.001$).
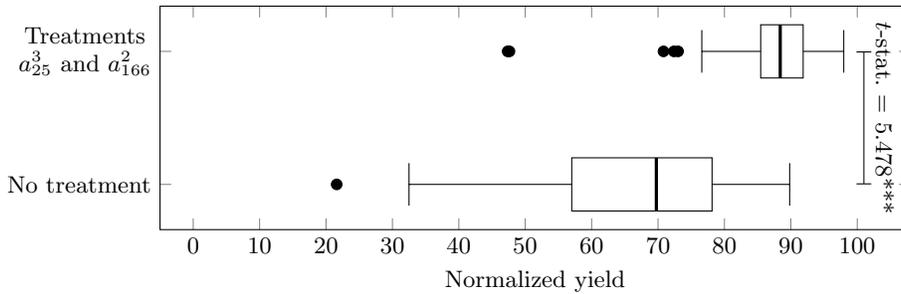
Significance level: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$
Notes: The box plot shows the normalized yield for those observations in the holdout set that have been treated with actions a^3_{25} and a^2_{166} (i.e., processed by machines QI615 and QP212) against those that have not. The whisker length is given by the 1.5 interquartile range, and the 50% quantile is highlighted as a bold line. We report the mean difference in normalized yield and the statistical significance based on Welch's t-test.

Figure 9   **Projected treatment effect**

Although the results are statistically significant, an analysis of historical observations is not sufficient to demonstrate that the selected improvement actions will also be effective in the future. Moreover, the fabrication conditions in all other processes of the above analysis were not held equal and, therefore, the projected treatment effect only gives an approximation. We addressed this by conducting a field experiment.

5.2. Design of field experiment

Our experimental validation followed the best practice in semiconductor manufacturing. For this, Hitachi ABB produced a completely new production batch of transistor chips (i.e., not included in the original dataset), where the routing of materials through the fabrication processes was altered. More specifically, the field experiment was carried out based on a 2×2 factorial design, such that the selected improvement actions were tested at both prioritized processes 25 and 166. The underlying principle of this design is to subdivide the experimental population (the production batch) into four groups such that two different improvement actions (machine prioritizations) can be tested at two different levels (processes). Overall, one group is subject to both improvement actions, one group to none, and two groups to only one.

Hitachi ABB has previously used 2×2 factorial experiments to test various hypotheses about quality issues. The production batch used in the experiment consisted of 24 wafers that were split into four equal groups with six wafers each (i.e., 372 transistor chips per group). Except for the prioritized processes 25 and 166, in which we tested the selected improvement actions, all wafers were processed under the exact same conditions. This ruled out confounders that could have influenced the results of the experiment. We report the difference in normalized yield between all four groups to establish the effectiveness of the selected improvement actions. Note that, to ensure comparability, the yield normalization in the field experiment was identical to that used previously.

5.3. Field experiment

The experiment was performed as follows. Group 1 served as the control group; Group 2 received a treatment in the form of improvement action a_{25}^3; Group 3 received improvement action a_{166}^2; and Group 4 received both improvement actions a_{25}^3 and a_{166}^2. Table 5 lists the machine prioritization.

The results of the experiment confirm that the treated groups outperform the control group in terms of yield performance (Table 5). The wafers in the control group (i.e., Group 1 with no improvement action) have a normalized yield of 38.2. In contrast, the wafers processed with treatment a_{25}^3 (Group 2) have a normalized yield of 50.6. For the wafers that received treatment a_{166}^2 (Group 3), we record a normalized yield of 85.7. The largest effect can be observed for the wafers that received both treatments a_{25}^3 and a_{166}^2 (Group 4), resulting in a normalized yield of 86.0. Overall, this improved the normalized yield over the control group by 47.8 units. This result

is consistent with the projected treatment effect (i.e., both groups are within 90% of empirical distribution).

Table 5   Experimental results

	Control group	Treatment groups		
	Group 1	Group 2	Group 3	Group 4
Treatment	No treatment	a_{25}^3	a_{166}^2	a_{25}^3 and a_{166}^2
Machine in process 25	QI613	QI615	QI613	QI615
Machine in process 166	QP233	QP233	QP212	QP212
Normalized yield	38.2	50.6	85.7	86.0
Absolute improvement over control group	—	+12.4	+47.5	+47.8

Notes: The table presents the results from a 2×2 factorial experiment at Hitachi ABB. We investigate the yield influence of different process routings to validate the selected improvement actions. Except for the prioritized processes 25 and 166, all wafers are processed by identical machines. In process 25 and process 166, we split the production batch into four equal groups such that one group receives both treatments, one group receives no treatment, and two groups receive one of the treatments. We report the resulting normalized yield for each group to show the effectiveness of the improvement actions a_{25}^3 and a_{166}^2.

5.4.   Interpretation of experimental results

Comparing the control group (Group 1) and the group with both treatments (Group 4) returns an improvement in the normalized yield by 47.8 units. We acknowledge that this yield improvement is relatively large considering the projected treatment effect. The reason is that the experiment compares the machines associated with the worst performance against the machines associated with the best performance (cf. Table 4). Nevertheless, a selection bias can be ruled out because Hitachi ABB did not select among production batches for the experiment. Instead, they dedicated a random batch of wafers to be treated experimentally. Hitachi ABB further ensured that the experimental batch was representative by comparing the average non-normalized yield of our sample (i.e., $\frac{1}{M} \sum_{i=1}^{M} \mu^{(i)}$) against the yield of Group 4.[5] The company informed us that both improvement actions resulted in a 21.7% reduction in yield loss.

The experiment confirms that both improvement actions have a positive effect on the normalized yield. We note that the process importance of process 25 is larger than that of process 166. Nonetheless, the experimental results suggest the yield improvement due to action a_{166}^2 is larger compared to that due to action a_{25}^3. This can be explained as follows. The process importance values (cf. Table 3) are computed across all observations in the training set. Here, the process associated with the largest estimated contribution to variation in the normalized yield is not necessarily the process that, for a single production batch, relates to the largest yield gain. Specifically, we observed that,

---

[5] Note that the yield variable has been normalized to preserve the confidentiality of the absolute figures. The normalization in the experiment is equivalent to the one used previously.

in our sample, the worst-performing machines in process 25 (i. e., QI613 and QI614) have been used relatively more often than the worst-performing machines in process 166 (i. e., QP232 and QP233). Therefore, the decision model attributes more importance to process 25. This is a strength of the decision model because it not only takes into account marginal contributions, but also how often certain actions have been used. Further note that, once transistor chips are damaged, they cannot be restored, because the effects for both improvement actions are not additive.

The research team presented the results to senior managers and production experts at Hitachi ABB. A discussion among the domain experts led to suggestions regarding the sources of machine-related performance differences. In implantation process 25, the worst-performing machines, QI613 and QI614, are prone to particles that induce chip failures during processing. In contrast, machine QI615 does not engender the same amount of contamination. In etching process 166, there is a specific machine type whose loading mechanism scratches the wafer surface, thereby damaging a subpopulation of chips at four wafer positions. Machine QP212 is not affected by this mechanical problem and has thus a positive association with the normalized yield. Overall, the improvement actions are effective, as the worst-performing machines can be used to perform other processes where the influence on the normalized yield is not critical.

6.   Evidence from post-experimental roll-out

Motivated by the results of the field experiment, Hitachi ABB decided to integrate our decision model into their quality management. Their first post-experimental roll-out targeted a transistor chip product that is different from the one used in our initial field research (sections 4 and 5). The company chose this product because it had been subject to a comparably low yield for which the underlying mechanisms were not understood. Together with the process engineers, we implemented our decision model using product and process measurements from the new setting. The decision model identified two etching processes associated with a large estimated contribution to variation in the yield. The process engineers used this new insight to implement two improvement actions targeting the negative yield associations in both prioritized etching processes.

We report the quality improvements that were achieved in the running operations of Hitachi ABB. Figure 10 compares the yield distributions of the production volumes produced before and after the implementation of the two improvement actions (corresponding to February–June 2020 and July–August 2020, respectively). Note that semiconductor manufacturers like Hitachi ABB produce thousands of transistor chips daily. By comparing the change in the distributions, we find a reduced yield variation and improved yield mean, which is in line with the theoretical background of this work (cf. Figure 2). Overall, the yield loss of the considered transistor chip product was reduced by 51.3% ($p < 0.001$). The improvement remained consistent throughout the entire study period.
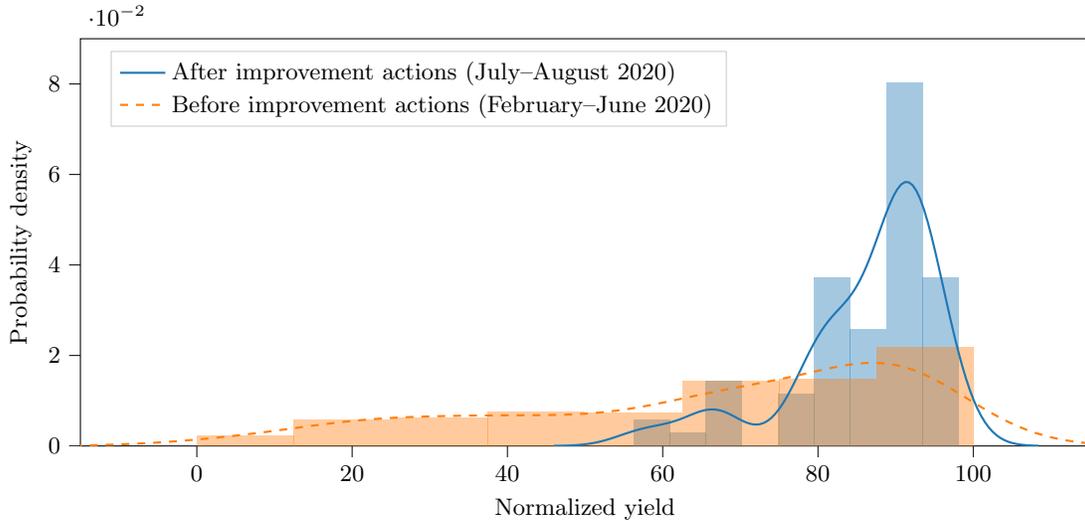
Figure 10   Comparison of normalized yield before and after improvement actions

Evidently, the two selected improvement actions were associated with significant yield improvements. Yet the underlying quality drivers were not identified previously. On the one hand, Hitachi ABB confirmed that the two identified processes were not detected with the company's standard quality management tools (based on correlation analysis and expert knowledge). On the other hand, alternative methods could not provide improvement actions. Specifically, we implemented a lasso with additional quadratic terms and interaction terms between production parameters to consider potential nonlinearities. The lasso set all coefficients to zero and, in this case, incorrectly suggested that no production parameter is associated with the normalized yield. This again highlights the operational value of using explainable AI for quality improvement.

7.   Robustness checks

In this section, we perform several robustness checks related to our empirical application (Section 4). We confirm that using an alternative meta-model results in consistent findings and compare our decision model against linear methods and a decision tree heuristic.

7.1.   Selection of improvement actions with alternative meta-model

We implement a random forest as a meta-model to evaluate whether we arrive at the same conclusions as before. The random forest is chosen because its modeling performance is on par with gradient boosting (see Appendix B). We first compare the rankings of process importance values Θ_k determined based on the gradient boosting and random forest meta-models. We find that the rankings are nearly identical (Spearman rank-order correlation coefficient of 0.93 with $p < 0.001$). The decision model based on the random forest also associates implantation process 25 and etching process 166 with the largest estimated contribution to variation in the normalized yield. In

addition, it suggests that machine QI615 in process 25 and machine QP212 in process 166 should be selected as improvement actions. Therefore, both Step 1 and Step 2 of the decision model lead to consistent results regardless of the two implemented meta-models.

7.2. Comparison against linear methods

An example of linear methods in quality modeling is provided by Zantek et al. (2002). Their method models process quality as a linear combination of production parameters and quality variables to allocate improvement resources. However, the work by Zantek et al. (2002) assumes that quality variables are measured at intermediate inspections throughout a manufacturing system. This is not applicable in our setting where the quality variables (i.e., yield) are only measured at the end of the transistor chip production.

An alternative linear approach is to model process quality only based on production parameters without quality variables from intermediate inspections. We implement this approach with a lasso to evaluate whether the model coefficients point toward quality drivers. We find that the only non-zero coefficient is related to production parameter x_{302} from implantation process 25. Recall that this production parameter has a strong linear correlation with the normalized yield; therefore, this finding is not surprising. In contrast, all coefficients related to etching process 166 are set to zero, thereby indicating no influence on the normalized yield. This contradicts the outcomes of the field experiment, which confirmed a large yield influence for process 166. We also implemented a lasso with quadratic terms and interaction terms, which did not provide additional findings.

7.3. Comparison against decision tree heuristic

A nonlinear approach for quality modeling is to interpret the functional relationships of a decision tree. The underlying idea is to manually analyze actions that are frequent in tree leaves with large predicted yield values. As such, this approach does not necessarily target the sources of variation as would be required for prioritizing processes. We implement an optimal decision tree (Bertsimas and Dunn 2017) and compare it against our decision model. We find implausible results for process 166 when analyzing the functional relationships. Specifically, the tree splits suggest that machine QP233 has no negative association with the normalized yield. This is at odds with the field experiment, which confirmed that machine QP233 is associated with substantial yield loss.

8. Discussion

In this section, we discuss contributions, limitations, and practical implications.

8.1. Contributions to the literature

This paper proposes a data-driven decision model for improving process quality in manufacturing. The decision model has three properties that address the limitations of existing approaches in

quality modeling. First, it is closely aligned with quality management theory by first targeting the sources of variation and, subsequently, selecting actions for quality improvement. Second, it is designed to handle manufacturing data that are both high-dimensional and nonlinear. Third, it contributes a measure of process importance based on which manufacturers can prioritize processes for quality improvement without requiring access to quality variables from intermediate inspections.

Our measure of process importance supports the effective allocation of improvement efforts, even when manufacturing data is subject to nonlinearities. In our real-world application at Hitachi ABB, we provide evidence from two independent interventions where traditional methods for quality management did not provide sufficient insights. In contrast, the proposed decision model revealed critical relationships, which eventually led to significant quality improvements. In addition, we provide a simulation (see Online Supplements) that confirms our measure of process importance is effective in locating quality drivers under nonlinearities. Overall, our field research demonstrates the operational value of explainable AI.

8.2. Limitations

A limitation of the proposed decision model is that it makes inferences based on correlations. Recent work has established that assuming causation in explainable AI can lead to misleading interpretations and thus poor decisions (Bastani et al. 2018). Moreover, it has been shown that post hoc explanation methods can be fooled via adversarial attacks (Slack et al. 2020). While this evidently has not been the case in our empirical setting, it highlights the importance of involving domain experts in the development of candidate actions once processes have been prioritized for quality improvement. It is unlikely that manufacturers are willing to implement improvement actions based on non-explainable model outputs. Although our decision model by itself cannot guarantee causality, it provides suggestive input to process engineers, who can then identify potential causal pathways that would be consistent with the inferences from the model. Finally, determining whether the selected actions actually lead to quality improvements requires experimental validation.

Another limitation of the decision model is that it can only select improvement actions based on past observations of production processes. Consequently, the decision model is incapable of identifying quality drivers that have no variation or have not been observed previously. This is a common assumption, as well as limitation, of quality modeling methods (e. g., Zantek et al. 2002). Consider the example of a baking process that is operated at a constant out-of-range temperature and thus consistently produces quality defects. In this case, the decision model would not be able to point toward temperature as a potential root cause because suitable temperature levels have not been observed. To overcome this limitation, manufacturers commonly use classic quality management methods, such as Design of Experiments (Fisher 1935), which introduce controlled variation into a process.

8.3. Practical implications

Our decision model can be efficiently adopted into quality management practice. Thereby, manufacturers can generate new insights from available data, which are often not analyzed effectively (Kusiak 2017, Corbett 2018). We make no specific assumption about which meta-model is used. This enables the straightforward use of established models in the operations management literature (e. g., tree ensembles; Cui et al. 2018, Sun et al. 2020, Bastani et al. 2021). As part of the robustness checks, we showed that using an alternative meta-model (with similar modeling performance) results in identical improvement actions. In addition, the decision model is specified generically so that it only requires lightweight input in the form of production parameters, a measured process quality variable, and process specifications.

Our field research was carried out in the semiconductor industry, which has several favorable conditions for using explainable AI. First, semiconductor manufacturing is highly automated, which eases the capture and system coverage of data. Second, fabrication steps are clearly defined, which allows tracking each product to a distinct process. Third, semiconductor manufacturers face costly yield losses, which motivates the investments in quality improvement. Notwithstanding the benefits of our research setting, our decision model generalizes to manufacturing settings beyond semiconductor fabrication. The biggest hurdle is the representation of data covering all relevant processes and production parameters. When important production parameters are omitted, there is a risk that quality drivers may go unnoticed. Other industries likely to have favorable conditions include pharmaceuticals, petrochemicals, and automated production lines for fast-moving consumer goods or printed circuit boards. The decision model likely performs worse in labor-intensive manufacturing because manual processes are often quality drivers but challenging to capture digitally. It is also problematic if quality issues are caused by unnoticed supplier performance, such as inferior material properties. With the ongoing digitization of manufacturing, we expect that the challenges of data representability will be reduced in the future.

Our approach to quality improvement is relevant to other application areas in management. In marketing, for example, it could be used to understand drivers of customer churn and target at-risk individuals with suitable incentives. In supply chain management, it could be used to assess influential variables of supplier risk. In healthcare management, it could be used to improve hospital operations responsible for between-patient variations in readmission risk. We leave these promising opportunities for future research.

9. Conclusion

In this paper, we proposed a data-driven decision model to improve process quality in manufacturing. The decision model first prioritizes processes for quality improvement and then selects

improvement actions. As a particular benefit, the decision model is designed to handle nonlinear manufacturing data. This is achieved by a novel combination of nonlinear modeling and SHAP values from the field of explainable AI. We demonstrated the effectiveness of our approach in a field experiment at Hitachi ABB. After the field experiment, the company adopted the decision model and achieved significant quality improvements. The generic design of our decision model allows for widespread applicability in manufacturing settings with high data coverage. Based on our work, we see promising opportunities for explainable AI in management science.

Acknowledgments

References

Bastani H, Bastani O, Kim C (2018) Interpreting predictive models for human-in-the-loop analytics, URL https://hamsabastani.github.io/interp.pdf.

Bastani H, Zhang D, Zhang H (2021) Applied machine learning in operations management. Babich V, Birge J, Hilary G, eds., *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management, Forthcoming (Springer Nature).

Bertsimas D, Dunn J (2017) Optimal classification trees. *Machine Learning* 106(7):1039–1082.

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees* (Belmont, CA: Wadsworth).

Chen WC, Tseng SS, Wang CY (2005) A novel manufacturing defect detection method using association rule mining techniques. *Expert Systems with Applications* 29(4):807–815.

Chien CF, Wang WC, Cheng JC (2007) Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications* 33(1):192–198.

Corbett CJ (2018) How sustainable is big data? *Production and Operations Management* 27(9):1685–1695.

Cui R, Gallino S, Moreno A, Zhang DJ (2018) The operational value of social media information. *Production and Operations Management* 27(10):1749–1769.

Field JM, Sinha KK (2005) Applying process knowledge for yield variation reduction: A longitudinal field study. *Decision Sciences* 36(1):159–186.

Fisher RA (1935) *The Design of Experiments* (Edinburgh and London: Oliver and Boyd).

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5):1189–1232.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. *ACM Computing Surveys* 51(5):1–42.

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1/3):389–422.

Hastie T, Tibshirani R, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York, NY: Springer), 2nd edition.

Hopp WJ, Spearman ML (2011) *Factory Physics* (Long Grove, IL: Waveland Press), 3rd edition.

Ittner CD (1994) An examination of the indirect productivity gains from quality improvement. *Production and Operations Management* 3(3):153–170.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)* .

Kusiak A (2017) Smart manufacturing must embrace big data. *Nature* 544(7648):23–25.

Lundberg S, Lee SI (2017) A unified approach to interpreting model predictions. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)* .

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1):56–67.

Mišić VV, Perakis G (2020) Data analytics in operations management: A review. *Manufacturing & Service Operations Management* 22(1):158–169.

Olsen TL, Tomlin B (2020) Industry 4.0: Opportunities and challenges for operations management. *Manufacturing & Service Operations Management* 22(1):113–122.

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: Explaining the predictions of any classifier. Krishnapuram B, Shah M, Smola A, Aggarwal C, Shen D, Rastogi R, eds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*.

Schmenner RW, Swink ML (1998) On theory in operations management. *Journal of Operations Management* 17(1):97–113.

Shapley LS (1953) A value for n-person games. Kuhn HW, Tucker AW, eds., *Contributions to the Theory of Games*, 307–318, Annals of Mathematics Studies (Princeton, NJ: Princeton University Press).

Shewhart WA (1926) Quality control charts. *Bell System Technical Journal* 5(1):593–603.

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 180–186.

Sun J, Zhang D, Hu H, Van Mieghem JA (2020) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* .

Taguchi G (1986) *Introduction to Quality Engineering: Designing Quality into Products and Processes* (Tokyo: Asian Productivity Organization).

Taguchi G, Clausing D (1990) Robust quality. *Harvard Business Review* 68(1):65–75.

Terwiesch C, Olivares M, Staats BR, Gaur V (2019) A review of empirical operations management over the last two decades. *Manufacturing & Service Operations Management* 22(4):656–668.

Tsai TN (2012) Development of a soldering quality classifier system using a hybrid data mining approach. *Expert Systems with Applications* 39(5):5727–5738.

Wu L, Zhang J (2010) Fuzzy neural network based yield prediction model for semiconductor manufacturing system. *International Journal of Production Research* 48(11):3225–3243.

Yu B, Popplewell K (1994) Metamodels in manufacturing: A review. *International Journal of Production Research* 32(4):787–796.

Zantek PF, Wright GP, Plante RD (2002) Process and product improvement in manufacturing systems with correlated stages. *Management Science* 48(5):591–606.

Appendix

A. SHAP value method

The SHAP value method (Lundberg and Lee 2017, Lundberg et al. 2020) infers the underlying decision rules of predictive models by decomposing a prediction $f(x)$ into the contribution (called the "SHAP value") of each feature j. For this, the SHAP value method combines local model explanations and game theory (Shapley 1953). Intuitively, the SHAP value method can be viewed as a cooperative game (i. e., reproducing a prediction) where the payoff (i. e., the prediction) must be allocated fairly among individual players (i. e., the features values) based on their contribution.

SHAP values are computed at the observation level; that is, each value in a feature vector x receives its own SHAP value. A particular benefit of SHAP values is that they can be interpreted both locally (at the observation level) and globally (at the model level). Based on this, both feature attribution and feature importance can be quantified: Feature attribution is directly determined via the SHAP values, whereas feature importance is determined by averaging the absolute SHAP values across observations (Lundberg et al. 2020).

The SHAP value method (Lundberg and Lee 2017, Lundberg et al. 2020) extends the traditional Shapley value definition from game theory to its use in predictive models; that is,

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! \, (N - |S| - 1)!}{N!} \left[f_x(S \cup \{j\}) - f_x(S) \right], \tag{A1}$$

where F defines the set of all features, N defines the total number of features, and S defines all the possible subsets $F \setminus \{j\}$. Here, $f_x(S) = \mathbb{E}[f(x) \,|\, x_S]$ denotes the expected value of the predictive model conditioned on a subset S of input features (as defined in Lundberg et al. 2020). Consequently, $f_x(S \cup \{j\}) - f_x(S)$ estimates the marginal contribution of adding the feature j to the feature set S. The effect of adding the feature depends also on other features in the model and is thus computed for all possible subsets S. Therefore, the feature attribution ϕ_j quantifies the weighted average of all marginal contributions of the feature j to the model prediction $f(x)$, while considering nonlinear relationships (cf. Lundberg and Lee 2017).

Estimating feature attributions via Equation (A1) requires 2^N estimations and is thus computationally expensive. As a remedy, Lundberg and Lee (2017) developed model-specific methods for computing SHAP values in a computationally efficient manner, such as for deep neural networks (Lundberg and Lee 2017), kernel-based models (Lundberg and Lee 2017), and tree-based models (Lundberg et al. 2020).

SHAP values have three desirable properties (Lundberg and Lee 2017, Lundberg et al. 2020): (1) local accuracy, (2) missingness, and (3) consistency. Local accuracy states that the sum of all feature attributions equals the prediction; that is, $f(x) = \phi_0(f) + \sum_{j=1}^{N} \phi_j(f, x)$, where $\phi_0(f)$ is the expected model output. Missingness states that absent features have no attribution; that is, if $f_x(S \cup \{j\}) = f_x(S)$ for all subsets S in the power set of F, then $\phi_j(f, x) = 0$. Consistency states that increasing the impact of a feature on the model does not decrease the attribution of that feature; that is, for any models f and f', if $f'_x(S) - f'_x(S \setminus \{j\}) \geq f_x(S) - f_x(S \setminus \{j\})$ for all subsets S in the power set of F, then $\phi_j(f', x) \geq \phi_j(f, x)$. SHAP values are the only additive feature attribution method that satisfies these three properties (Theorem 1 in Lundberg and Lee 2017).

B.   Comparison of meta-models

We compare the modeling performance of gradient boosting with eight alternative meta-models f. We follow previous literature (e. g., Cui et al. 2018) and implement both linear and nonlinear models. We also introduce a naïve baseline, which is defined as the in-sample mean of the normalized yield. Each model is subject to hyperparameter tuning (see Appendix C). The modeling performance is measured via the deviation between the model output $f(x^{(i)})$ and the true normalized yield $y^{(i)}$ in the holdout set $\mathcal{H}$. Here, we draw upon the root mean squared error (RMSE) given by $\sqrt{\frac{1}{M_\mathcal{H}} \sum_{i=1}^{M_\mathcal{H}} (y^{(i)} - f(x^{(i)}))^2}$ and the mean absolute error (MAE) given by $\frac{1}{M_\mathcal{H}} \sum_{i=1}^{M_\mathcal{H}} |y^{(i)} - f(x^{(i)})|$. The advantage of using the MAE to compare different model specifications is its interpretability, as it directly transfers to the normalized yield. Additionally, both the MAE and the mean absolute feature attribution are measured as mean absolute deviations.

Table A1   **Comparison against alternative meta-models**

Model	MAE	RMSE	Statistical comparison (t-statistic)		
			In-sample mean	Lasso	Gradient boosting
Naïve baseline					
In-sample mean	9.103	12.404	—	−2.179*	−4.128***
Linear models					
Linear regression	9.082	12.414	0.027	−2.145*	−4.088***
Ridge regression	8.747	12.111	0.463	−1.692*	−3.629***
Lasso	7.528	10.513	2.179*	—	−2.016*
Elastic net	7.528	10.513	2.179*	0.000	−2.016*
Nonlinear models					
Support vector regression	7.480	10.961	2.158*	0.069	−1.845*
Deep neural network	6.740	9.756	3.325***	1.198	−0.801
Optimal decision tree	6.718	9.599	3.394***	1.247	−0.779
Random forest	6.250	9.022	4.143***	2.015*	−0.021
Gradient boosting	6.237	9.114	4.128***	2.016*	—

Significance level: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$
Notes: The table compares the out-of-sample model performance for nine different meta-models and a naïve baseline. The linear regression makes additional use of recursive feature elimination to prevent overfitting. The optimal decision tree is based on Bertsimas and Dunn (2017). We perform t-tests on the absolute prediction errors to show that the model performances differ at a statistically significant level. The t-statistics and significance levels (one-sided) are reported based on Welch's t-test.

Table A1 presents the out-of-sample modeling performance. The estimation results confirm that gradient boosting achieves the best performance among all models (MAE of 6.237). Among the nonlinear models, the random forest appears on par with gradient boosting. Overall, gradient boosting outperforms all linear models at a statistically significant level. The improvement over the best linear models (i. e., lasso and elastic net) amounts to 17.1%. The results suggest that gradient boosting is superior in modeling the underlying physical processes, thus making its choice particularly suitable for our setting.

C. Hyperparameter tuning

Table A2 lists the hyperparameters evaluated based on a grid search with 5-fold cross-validation on the training set. Note that deep neural networks are usually used along long rather than wide data. Hence, additional effort was needed to achieve a favorable performance in our high-dimensional setting. We considered different model architectures, including layers with regularization and convolutional layers. The best-performing architecture draws upon kernel regularization to prevent overfitting.

Table A2   Grid search for hyperparameter tuning

Model	Tuning parameters	Tuning range
Linear regression	Number of features (recursive feature elimination)	1, 2, **3**, ..., 3614
Ridge regression	Regularization strength α	0.01, 0.1, 1, 10, 100, 500, 1000, 2000, **5000**
Lasso	Regularization strength α	0.01, 0.1, **1**, 10, 100
Elastic net	Regularization strength α	0.01, 0.1, **1**, 10, 100
	Regularization ratio	0, 0.25, 0.5, 0.75, **1**
Deep neural network	Number of hidden layers	**2**, 5
	Number of neurons in 1st hidden layer	**100**
	Number of neurons in other hidden layers	**10**
	Learning rate	**0.00001**
	Epochs	5000, **10000**
	Batch size	**64**
	Kernel regularization strength α	0, **1**
	Dropout rate in 1st hidden layer	**0**, 0.5
	Dropout rate in other hidden layers	**0**
Support vector regression	Kernel function	**Radial**
	Cost parameter C	0.01, 0.1, 1, **10**, 100
	γ-Parameter	0.001, **0.01**, 0.1, 1
Optimal decision tree	Maximum tree depth	1, 2, **3**, 4, 5
Random forest	Number of trees	50, 100, **400**
	Maximum tree depth	None, 2, 5, 10, **50**
	Maximum features considered for split	**Auto**, 1, 3, 5, 10
Gradient boosting	Number of trees	50, 100, 400, **1200**
	Maximum tree depth	**3**
	Learning rate	**0.01**
	Regularization strength α	**0**, 0.1, 1, 4
	Regularization strength λ	0, 0.1, 1, **4**
	Bagging fraction	**0.2**, 0.6, 1
	Feature fraction	**0.2**, 0.6, 1

Notes: The selected hyperparameters are highlighted in bold.

As an additional data representation, the lagged normalized yield was included as a predictor, but modeling performance still appeared on par. Moreover, the decision model selected the exact same actions.

D. Analysis of feature attribution

Figure A1 shows the SHAP summary plot (Lundberg et al. 2020) for the top 10 quality drivers.
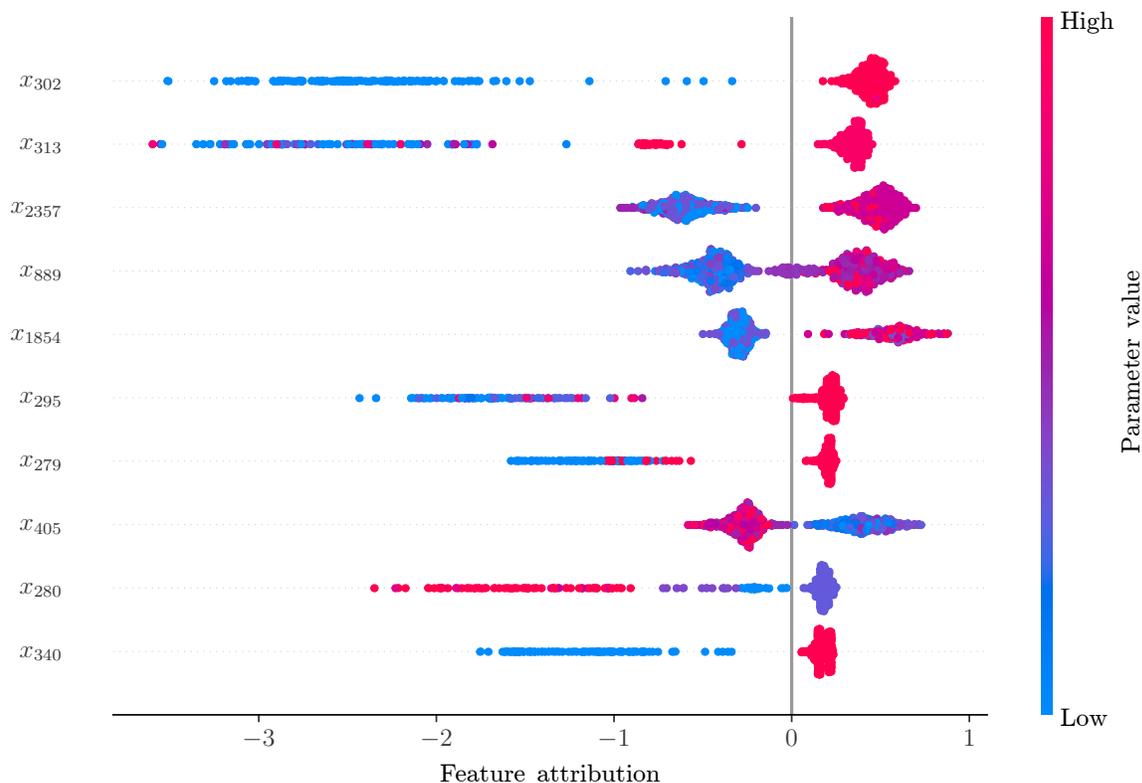

Figure A1   **Feature attribution for top 10 quality drivers**

Figure A2 compares the mean absolute feature attributions and the Pearson correlation coefficients between the production parameters and the normalized yield. The figure shows that, except for production parameter x_{302}, quality drivers tend to have a low correlation with the normalized yield.
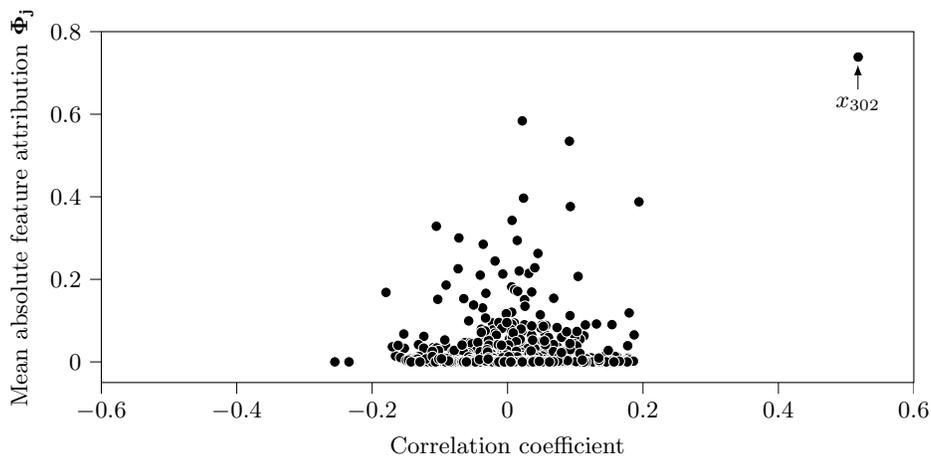

Figure A2   **Comparison of correlation coefficients and mean absolute feature attribution**