

Enhancing Product Quality in High-Variant Manufacturing: Combining Physics-Based Simulations and Data Science for Target Variable Estimation in an IoT- and Machine Learning-Driven Context

[Manuela Larissa Schreyer](#)*, Alexander Gerber, Steffen Neubert, [Peter Simon](#)

Posted Date: 4 December 2025

doi: 10.20944/preprints202512.0479.v1

Keywords: process optimization; machine learning; heterogeneous production processes; data science; target variables; reduced/limited dataset



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Enhancing Product Quality in High-Variant Manufacturing: Combining Physics-Based Simulations and Data Science for Target Variable Estimation in an IoT- and Machine Learning-Driven Context

Manuela Larissa Schreyer ^{1,*}, Alexander Gerber ², Steffen Neubert ² and Peter Simon ³

¹ Austria Metall GmbH, Lamprechtshausenerstr. 61, 5282 Ranshofen, Austria

² AMAG Casting GmbH, Lamprechtshausenerstr. 61, 5282 Ranshofen, Austria

³ AMAG Rolling GmbH, Lamprechtshausenerstr. 61, 5282 Ranshofen, Austria

* Correspondence: manuela.schreyer@amag.at

Abstract

Due to growing demands for quality, sustainability, and digitalization, data science and artificial intelligence are gaining importance across industries. The extensive product range in many sectors often poses considerable challenges. For example, machine learning (ML) models may struggle with limited data per product type. This paper presents a method that combines data science and physics-based simulations for target variable estimation, with the aim of creating a comparable database for effective statistical analysis and ML models — based on a use case from the aluminum production industry. A key advantage of this approach is that it can effectively model even production variants with very low quantities. The following discussion will present how this method can be used to enhance production processes, specifically to identify parameters that directly influence product quality. Furthermore, the work explores the potential for precisely controlling these parameters using ML models and discusses some major challenges. The article demonstrates that integrating data science, technological knowledge and physics-based simulations is an effective methodology for estimating target variables, facilitating precise modelling even for production variants with minimal quantities and improving production processes.

Keywords: process optimization; machine learning; heterogeneous production processes; data science; target variables; reduced/limited dataset

1. Introduction

The fields of data science and artificial intelligence are assuming an increasingly important role in a multitude of industries. This is primarily attributable to an increased demand for higher quality and sustainability, in conjunction with the objective of enhancing efficiency and competitiveness [1–3]. The capacity to analyze large amounts of data enables automated and more precise decision-making. Furthermore, advancements in digitalization are instrumental in facilitating the integration of data-driven models.

Data-driven processes enable the efficient and precise handling of a wide variety of issues. The present article focuses on the optimization of production processes, with a particular emphasis on reducing waste and improving product quality. This paper analyses a selection of key challenges that can arise in the field of data science, as observed in many industrial sectors. It demonstrates possible solutions in the context of a specific use case in the aluminum processing industry, with a particular focus on aluminum ingot casting at AMAG.

The large quantity and complexity of production data necessitate the implementation of efficient techniques to derive insights that may facilitate the further optimization of production processes. In [4], for instance, a 7-step problem-solving strategy for the sustainable implementation of big data projects is presented. As is the case in numerous industrial sectors including, the product range is particularly extensive. In industry, this means that many different product variants are manufactured. For instance, the manufacturing of products of various sizes or material compositions. Therefore, the data is very heterogeneous, and some product variants are produced only rarely, such as custom-made products. Conversely, other product variants are manufactured in large quantities. However, a model should be capable of covering all cases equally well, including individual cases with low production volumes, as these are often associated with higher margins. This and other factors can result in significant challenges when conducting statistical analysis and building machine learning models [4].

In many cases, a substantial database is available for use in this context. However, the individual samples in the database are frequently not comparable due to the extensive product range and the resulting very different production histories. The wide range of products and the large variety of production settings pose a considerable challenge for statistical analyses and machine learning due to the extremely heterogeneous data, which leads to poor model generalization [5,6]. Conventional models produce distorted or inaccurate results due to this variability, as they do not consider the differences between settings [5]. There are various approaches to overcoming this challenge, such as clustering or data transformation methods, which help to extend analysis and models to infrequently produced variants [7].

The primary section of the subsequent article presents an additional approach, in which a data transformation method is introduced that uses data science methods in combination with finite element method (FEM) simulations to create a comparable database for effective statistical analysis and machine learning (ML) models based on a use case. A key advantage is that production variants, even those in very low quantities, can be modelled effectively and used for data analysis.

In addition, the following discussion will present the usage of this method for more efficient data analysis, with the objective of optimizing product quality and reducing waste. In the course of this study, a comparison will be made between data analyses based on the introduced transformation-method and non-transformed data. The objective of this comparison is to demonstrate the efficacy of the method employed. The subsequent section elucidates the manner in which the insights obtained – specifically, insights into parameters that exert a direct influence on product quality – contribute to the enhancement of efficiency. Furthermore, it delves into the potential for precise control of these parameters through the utilization of machine learning methods, as well as the associated requirements and challenges.

2. Methodological Framework – Calculating a Target Value

2.1. Use Case: Application Example and Description of the Data

This application example deals with the production of plates for the aerospace industry, in particular with the casting of aluminum ingots, which are then subsequently rolled into plates from which aerospace components are finally milled. A schematic representation of the production process is shown in Figure 1.

During the casting of the ingots, it is possible that non-metallic inclusions may be formed in the material, either as a result of the melting and casting process itself or as a consequence of the used input material. Such non-metallic inclusions may result in the rejection of the final product. This is determined during laboratory testing after the ingots have been rolled into plates [2,3,7,8]. Depending on the intended use of the end product, the final product must fulfil strict quality requirements, which are ensured by appropriate quality controls [9]. Among other things, the plates are subjected to ultrasonic (US) testing to detect non-metallic inclusions. If a plate contains too many or too large indications, the entire plate is rejected.

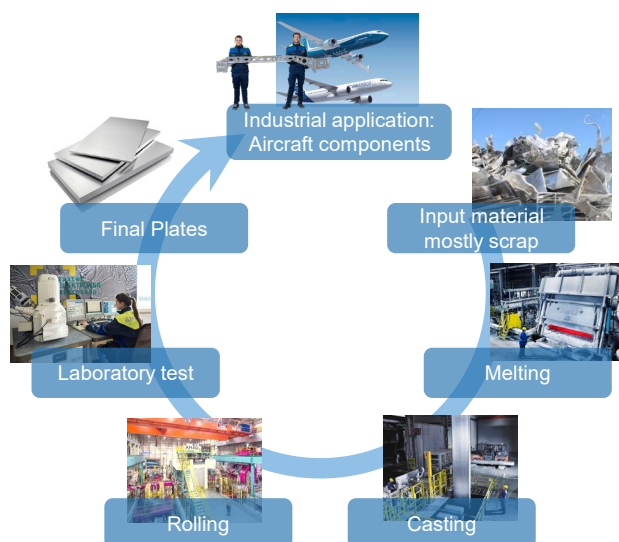


Figure 1. Schematic representation of the production process.

It is important to note that the origin of the non-metallic inclusions can be attributed to the casting process. But for technical reasons, the US-test can only be carried out on the final plates, rather than immediately after the casting process on the cast ingot. Nevertheless, this implies that numerous production processes (e.g. homogenisation, rolling, heat treatment, stretching, cutting, etc.) occur between the initial creation and subsequent detection, which can span several weeks [3].

The primary aim is to reduce the rejection rate (percentage of products that are rejected due to quality issues), the plant utilization rate and, consequently, the delivery time. One approach to achieving this is to identify potential influencing factors in the formation of non-metallic inclusions during the casting process using statistical analysis and machine learning models. Some crucial considerations that must be addressed are highlighted in sections 2.2.1, 2.2.2 and 3.1.

As will be demonstrated in the following paragraphs, this example serves also to illustrate the significant impact which the rolling production process can have on the rejection rate with respect to casting related non-metallic inclusions. This, in turn, has implications for statistical analyses and machine learning models, as already mentioned. Subsequently, methods from statistics are presented that can take this aspect into account in order to generate a target variable for statistical analyses and machine learning models that is independent of the subsequent production process. The issue of multi-step production manifests itself in a variety of industrial sectors. In the field of statistics, there exists a multitude of methodologies for the calculation of a target variable that is independent of 'intermediate steps' [7,10,11]. The present paper proposes one such methodology and discusses its advantages over alternative approaches.

In the field of statistics, the term "target variable" is used to describe the variable whose prediction or explanation is being sought as part of an analysis or prediction model. It represents the result or answers that can be predicted by analyzing the relationship to other explanatory variables (predictors). In this application example, the target variable should describe the non-metallic inclusions created by the casting process (in terms of size and number), regardless of all subsequent processes.

During the US-test, the precise position (thickness, width and length coordinates) of the defects and their dimensions are identified. These coordinates are then calculated, in a first step, back to the ingot, as illustrated in Figure 2. For more Details see [3,8].

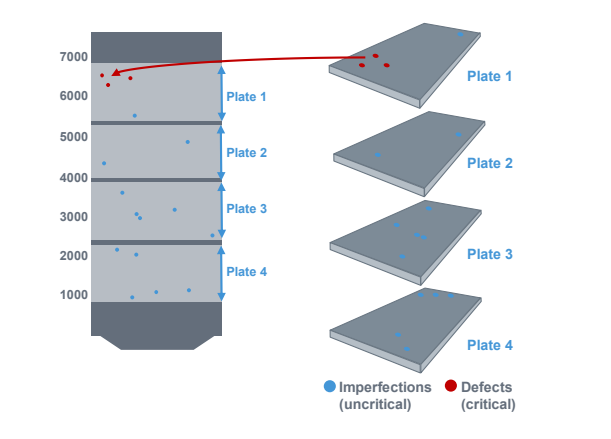


Figure 2. Recalculation of the US-indications from the plates to the original ingot.

It is imperative to implement a seamless digital recording system for all sawing, milling and cutting operations throughout the production process. This will facilitate the calculation of the defect position back to the casting ingot. Even temperature differences in the material during production must be taken into account in order to account for density variations in the material. Any inconsistencies in the data acquisition process will result in significant errors in the calculated coordinates on the ingot, due to the considerable dimensional change.

2.2. Industrial Challenge and Specific Implementation Scenario

2.2.1. General Challenge

Multi-stage production processes are established in a large number of industrial sectors. A recurrent phenomenon is the occurrence of errors in the production process, which are only identified at the end of the process – for example, during the final quality check [4]. This complicates the identification of causal relationships between specific production parameters and the detected defect, as relevant data must be linked across several process steps in order to precisely determine the error position in the product at the moment of occurrence. This, in turn, is required to make the analysis of the associated process parameters possible. Furthermore, intermediate steps in production between the occurrence and detection of the error can influence the final error pattern [4]. As demonstrated in the example provided, intermediate steps have the capacity to influence the size of the error.

The determination of a valid target variable and the development of robust analysis models are therefore challenging tasks. In many cases, the original process step that led to the error is known, which simplifies the analysis – but the analysis remains challenging, nonetheless. One such example is that of non-metallic inclusions during the casting process (see section 2.1). [2,3,7,8]

First, it is necessary to identify the intermediate production steps that could influence the occurrence or detection of defects and determine the nature of this influence. In most industries, there is a wide range of product variants, which can manifest as differences in size, shape, material or surface. Furthermore, there are often various facilities for producing the products in question. Intermediate production steps that could influence defect detection often become apparent when analyzing reject rates or error characteristics between different variants [4,7].

It is usually not sufficient to focus on specific variants in analyses because the production volumes for individual variants are often relatively low, particularly for high-quality specialty products. Therefore, the development of methods for transforming data from different variants to ensure comparability and enable cross-variant analysis is essential. It should be noted that data-driven statistical methods only lead to satisfactory results if there is a sufficiently large sample size of variants. In the following section, a method using FEM simulations will be presented, which offers advantages over statistical methods, particularly for smaller sample sizes.

2.2.2. Specific Application - Dependence of the Reject Rate on the Final Plate Thickness

AMAG’s product range is comprehensive and includes a wide variety of plate formats, alloys and ingot formats. In addition, a considerable number of different systems are available for production. In the following sections, one specific casting-system, alloy and ingot format will be the focus of the discussion. However, the method described can be applied to other casting-systems, alloys and ingot formats.

As already mentioned, the aim in this application example is to use statistical methods to find possible influencing variables on the formation of non-metallic inclusions during casting (measured via ultrasonic tests at the final plates). One of the biggest challenges is the strong dependence of the size of ultrasonic indications, measured on the final plates, on the rolling process, which becomes visible in the dependence on the final plate format.

Figure 3 illustrates the rejection rate, based on US indications, for the selected data set (one particular system, alloy and ingot format). The x-axis depicts the final plate thickness, while the y-axis represents the mean rejection rate for each final plate thickness. In Figure 3, the plate thickness was divided into nine groups (G1 to G8). Group G1 contains the thinnest plates, with the plate thickness increasing with each successive group number (G2, G3, etc.). The figure shows that the rejection rate (based on US indications) is highly dependent on the final plate thickness of the product, although only one specific alloy, one casting format and one casting plant are taken into account.

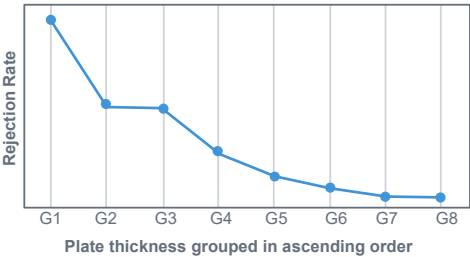


Figure 3. Dependence of the reject rate on the final plate thickness (one specific casting-system, alloy and ingot format).

This demonstrates that there is no comparability with regard to US scrap between products with the same alloy and ingot format from the same casting-plant, but with different final plate thicknesses. It can also be shown that there are huge differences in the scrap-rate between different alloys and plants. Nevertheless, this issue will not be addressed in any further detail in this article. This article only concentrates on only one specific casting-plant, one alloy and one ingot-format.

The aspect of the rejection rate’s high dependence on the final plate thickness raises the question of what implications this has for statistical analyses or machine learning approaches. Due to the fact that the products examined are high-quality specialty products for the aerospace industry, production volumes are relatively low compared to other products. Consequently, it is not feasible to concentrate on individual final plate formats in the analysis to overcome this issue. In principle, the aim is to create the most exhaustive database possible, in which the products can be compared. The utilization of data transformation methods serves this purpose, with a detailed description provided below.

However, it is first necessary to identify the cause of this behavior. It is already known that the dependence on the final plate format is generated by several different factors. One cause is, for example, the dependence of the US-untested area in thickness-direction on the final plate thickness. A detailed description of this behavior can be found in [8,12], which also presents a solution approach for accounting for this effect using statistical methods when calculating a target variable. However, this aspect is not considered further here. In the following, only the thickness range, which is always US-tested for all plate formats, is considered in order to be able to ignore this aspect and focus on other issues. For further consideration of this aspect, the explanations in [8,12] can then be followed.

Another cause is the presence of varying rolling forces, dependent on the final plate format, when the cast ingots are processed into plates. This topic is discussed in more detail below, and a methodology for calculating the target variable is presented.

Figure 4 shows that the thickness of the final plate has a significant impact on the size of indications detected during the US-test. An increase in plate thickness results in variations in rolling pressures, which in turn lead to different indication sizes. Figure 4 illustrates this relationship by displaying box plots grouped by plate thickness. The labels G1 and G8 represent the thinnest and thickest plates, respectively. The indications are classified as ‘Centre’ (in the middle of the ingot in thickness direction) or ‘Outer’ (at the top and bottom of the ingot in thickness direction) defects. The results demonstrate that as plate thickness increases, the sum of indication areas per ingot decreases. The reason for this behavior is that the indication size decreases with increasing plate thickness due to different rolling pressures [8]. In addition, the greater the final plate thickness, the more indications fall below the detection limit for ultrasonic testing.

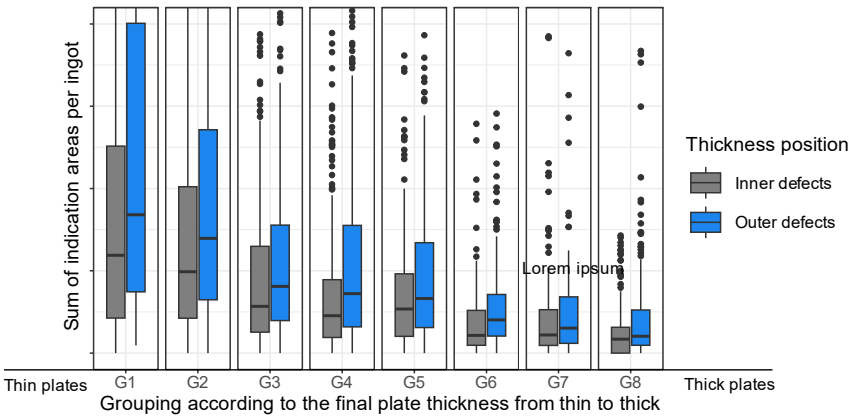


Figure 4. Effects of plate thickness on indication size.

As a direct consequence of this observation, it is essential to consider differences in rolling pressures when calculating the target variable. This ensures that the target variable accurately reflects casting-related non-metallic influences, regardless of any rolling effects.

The target variable, which is independent of rolling parameters, is a key requirement for the statistical analysis of potential influencing factors on casting quality and machine learning algorithms. It also provides foundry technologists with a valuable tool for evaluating the current casting quality.

2.3. Solution Approach and Use Case Implementation

2.3.1. General Approach

The present study aims to address the question of how adequate target values can be defined in multi-stage industrial processes. Multi-stage industrial processes are to be found in almost all areas of industrial manufacturing and processing [4]. These processes are characterized by a series of interconnected steps, frequently involving a degree of interdependence, with the objective of yielding a final product from the initial raw materials. In the following, the challenges and potential solutions are discussed, as this issue manifests itself in a variety of industrial areas.

The utilization of the pure reject rate as a target value necessitates critical examination. In any case, a thorough examination is necessary to ascertain whether there are hidden correlations with process steps that are not taken into account in the data analysis [7]. In such cases, the utilization of the reject rate has the potential to introduce inconsistencies. There are a variety of approaches that may be adopted in order to enhance both the informativeness of the target variable and the robustness of the analysis. A comprehensive overview is provided in [7].

The presence of hidden correlations can pose a significant challenge, as they reveal statistical relationships between the target variable and external variables that should not be regarded as influencing factors. An illustration of this is the thickness of the plates, which are not known at the time of casting. Nevertheless, as demonstrated previously, it exerts an influence on the rejection rate due to non-metallic indications. A range of statistical methods are available that are suitable for neutralizing such distortions. Common methods include grouping by confounding variables, variance stabilization and transformations to a reference distribution [7,10,11]. These methodologies facilitate the establishment of a comparable target variable; however, they necessitate sufficiently large samples. In a significant number of industrial applications, the issue that arises is that a sufficiently large sample is not available for each production variant. The production of custom-made products and the high variety of variants to meet the requirements of a large number of potential customers results in a high heterogeneity of data, with individual groups often being underrepresented. In the cases previously mentioned, data-driven approaches frequently reach their limits. The subsequent section presents a case study exemplifying the effective integration of FEM simulations and data-driven statistical methodologies to adequately address the issue of limited sample sizes.

The fundamental question pertains to the systematic utilization of FEM simulations for the realistic and comparable calculation of defect patterns, such as non-metallic inclusions, in multi-stage processes – despite the complex interactions between the individual manufacturing steps. In many cases, the defect pattern (e.g. inclusions, distortion, cracks, dimensional deviations, etc.) that ultimately leads to scrap is caused by a single process step. However, the final extent of the defect is influenced by one or even several other process steps in combination [13]. For instance, it is possible that existing defects (inclusions, cracks or shape deviations) may be further worsened under certain conditions. The presence of overlapping effects complicates the process of identifying the precise cause of the defect. The approach described aims to disaggregate these influences and achieve an objective, comparable assessment of the defect patterns.

In the initial stage, the manufacturing steps relevant to the defect pattern (e.g. forming, heat treatment and machining) are simulated using an FEM model [14].

In order to identify the relevant manufacturing steps, a deep technical understanding is required. The objective of this approach is to obtain quantitative data regarding the extent to which this step impacts the subsequent defect pattern.

In the subsequent step, the defect pattern is ‘normalized’ using the FEM results. In this context, the term ‘normalization’ is to be understood as follows: The real defect pattern is corrected mathematically or statistically, considering the influencing factors whose effects are described by the FEM simulations.

The result shows a corrected error pattern that is independent of the influences of the individual ‘intermediate’ process steps. The corrected error values (and thus also the reject numbers) are adjusted for process conditions such as temperature, product format etc. Using this method, it is possible to compare different production variants and evaluate, analyze or predict reject rates for all products in a uniform manner. This results in an expansion of the database, which in turn enables the development of analyses and prediction models.

The method can be summarized as follows: Each relevant production step is simulated individually to determine its individual relevance to the error pattern. Based on the simulation results, the influence of this contribution is mathematically extracted from the measured error patterns. This results in a standardized and comparable error measure, regardless of the extent to which, for example, forming or heat treatment were actually carried out. This approach allows error patterns to be calculated correctly and used for data analysis [7].

2.3.2. Implementation in the Industrial Use Case

In the use case delineated in section 2.2.2, the question arises about how the aforementioned dependency on rolling parameters should be taken into account when calculating the target variable.

For this purpose, the degree of deformation was calculated for some individual final plate thicknesses using FEM simulations, depending on the final plate thickness [14].

This simulation for an individual final plate thicknesses was then extended to all possible plate thicknesses by calculating an estimate using regression for all possible plate thicknesses, resulting in a function for the forming factor as a curve over the ingot thickness, which depends on the final plate thickness. In more detail, firstly, a polynomial regression model is separately created for each final plate thickness in order to model the degree of deformation over the ingot thickness position. Subsequently, an additional polynomial regression model is formulated for each defect position, which delineates the degree of deformation as a function of the final plate thickness. The resulting models are then combined, and a function is defined that calculates the degree of deformation for any combination of final thickness and position in the ingot.

Figure 5 shows the curves of the forming factor as a function of the ingot thickness dependent on the final plate thickness [15–17]. A thick final plate is characterized by a dark blue curve and a thin final plate by a light blue curve.

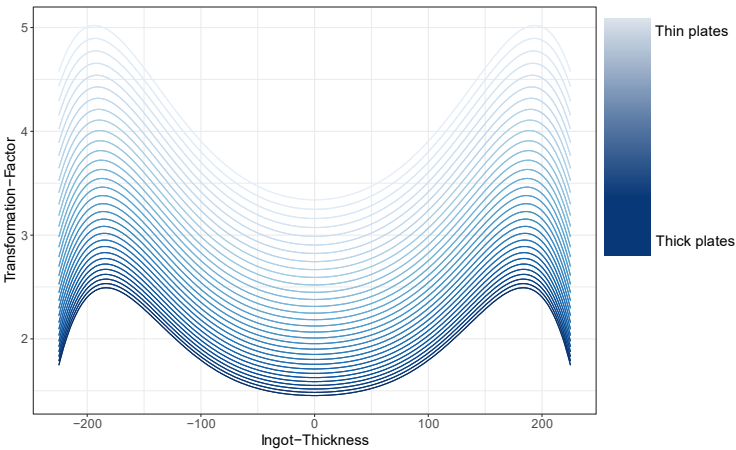


Figure 5. Transformation-Factor for different final plate thickness based on FEM simulations.

From this function, which describes the transformation-factor over the ingot thickness, an inverse function can now be calculated. This allows the defect size measured on the final plate (via US-test) to be transformed, thus determining the actual defect size of the original defect in the ingot.

As previously stated, an increase in the final plate thickness will result in a greater number of US-indications falling below the detection limit for ultrasonic testing. To take this into account, the detection limits are transformed from the final plate to the ingot in the same way. All indications that fall below the detection limit of the maximum plate thickness contained in the database are then removed, because these indications would not be detected if a plate with the maximum possible plate thickness had been produced from the ingot under consideration. This ensures comparability between the different plate thicknesses, and the target variable can be calculated based on the remaining indications.

Figure 6 illustrates now the transformed and reduced sum of indication areas per ingot by displaying box plots grouped by plate thickness. It is now evident that there is no longer a dependency on the final plate thickness. Consequently, the influence of deformation during rolling has been eliminated, and a comparability target variable between the different plate thicknesses is available for analysis.

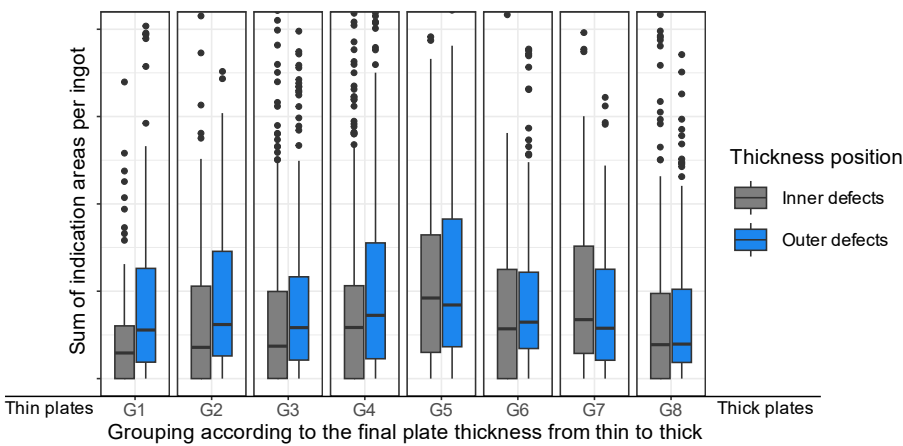


Figure 6. Transformed and reduced sum of indication areas per ingot (for thickness range, which is always US-tested for all plate formats).

3. Results

In the context of machine learning models and statistical data analysis for possible influences on product quality, the use of a comparable target variable is essential for the identification of significant and reliable patterns. Inconsistent definition or scaling of the target variable can lead to inaccuracy. This has a negative impact on performance and interpretability of results. Ensuring comparability of the target variable within the dataset used for machine learning models and statistical data analysis is therefore essential, as it enables the learning of relationships that are valid across the dataset. Consequently, the results become reproducible, and the model retains stability when applied to similar data sets. As part of data preparation, it is therefore essential to ensure a standardized definition of the target variables [7].

This section delineates the manner in which the target variable can be utilized to analyze the influencing factors in the casting process and how to control the identified influencing parameters using machine learning (ML) models. It also discusses the advantages of this method of calculating a target variable in comparison to other methods (for details see [7]). Firstly, the concluding definition of the target variable for the use case described in section 2.1 is presented.

3.1. Use Case: Calculation of a Quality Measure Per Batch

The target variable is calculated with consideration for the entire process chain, from the casting of the ingots to the final ultrasonic testing of the plates. The following aspects were considered:

1. Influence of defect position and final plate thickness: The position of the defects in thickness direction and the final plate thickness have been demonstrated to influence the defect size. The modelling was conducted utilizing FEM simulations and statistical methodologies. This modelling was integrated into the re-calculation of the defect size from the final plates to the original ingot as described in section 2.3.2.
2. Estimation of the defect area in the US-untested area: It should be noted that certain areas of the plates, due to technical constraints, could not be included in the testing process. Depending on the plate thickness, this affects 4% to 24% of the ingot (see [8,12]). For the US-untested areas, the defect area was estimated using an ML model – depending on the plate thickness and the defects found in the tested areas. Additional information is presented in [8].
3. Target variable per ingot and batch: The calculated defect-areas result in a target variable per ingot (sum of defect areas per US-tested ingot weight). The target variable per batch is calculated as the median of the ingot target variables – excluding outliers that were previously identified and excluded by Monte Carlo simulations. Further details are available in [7].

4. Section-by-section quality assessment: The target variable was also calculated for individual sections of the ingot in order to detect changes in quality during the casting process. See [3] for details.

3.2. Analysis of Influencing Parameters

3.2.1. Use Case: Analysis Result

It is well established that temperatures of all kinds frequently assume a pivotal role in production, particularly during the casting process. During the casting process, hundreds of different signals, including numerous temperature signals, are recorded. Feature selection methods were used to evaluate the relevance of this data with regard to the defined target variable [3]. This made it possible to identify the most meaningful signals and prioritize them for further analysis.

The subsequent section will focus on a specific signal. For reasons of confidentiality, the exact name of the signal cannot be disclosed. Consequently, the signal will be designated Signal ‘TS’ in the ensuing discussion.

The Signal ‘TS’ is recorded throughout the entire casting process, which lasts approximately two hours. This continuous recording enables a detailed analysis of the signal curves over the entire duration of the process. Data-driven investigations have identified a specific parameter within this signal ‘TS’ that has a significant influence on the final product quality [7,18,19]. This correlation is clearly evident in the calculated quality measure per batch, which serves as the target variable, as shown in Figure 7. The evaluation proves that a certain pattern in the Signal ‘TS’ can be attributed to quality deviations. This pattern is called Parameter P_{TS} (Parameter of TS) in the following.

Figure 7 shows a comparison of the Parameter P_{TS} between good- and poor-quality products using box plots (for more details see [7,18,19]). The rejection rate is shown on the left-hand side. No significant difference between the two quality groups can be seen here, which is confirmed by the high p-value of 0.86 [20]. The right-hand side of the figure shows the calculated target variable (as described in the previous sections), which shows a significant difference with a p-value close to zero. This finding underscores the conclusion that the correlation between the Parameter P_{TS} and product quality cannot be adequately represented by the rejection rate. Only by using the calculated quality measure per batch does this correlation become evident. The reason for this is that the rejection rate still contains ‘hidden correlations’ with the final plate thickness, which distort the analysis result [7].

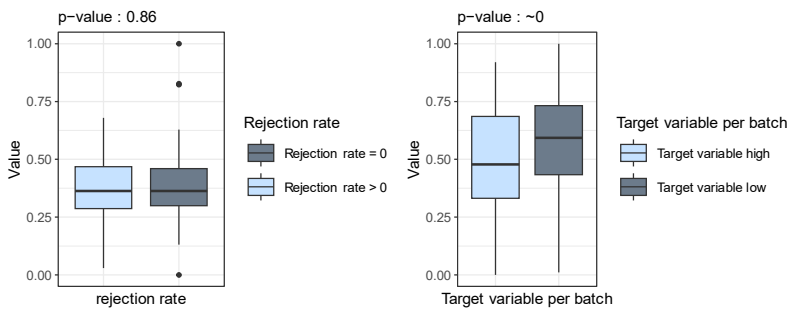


Figure 7. Boxplot of normalized Parameter P_{TS} form Signal ‘TS’ for good- and poor-quality products classified according to rejection rate (left side) and low / high Quality measure per batch (right side).

3.2.2. General Approach

In this section, the rationale behind the meticulous definition of the target variable is examined. Section 3.2.1 provides answers to these questions using an example for illustration.

The first step in a data-driven analysis is to identify the relevant factors influencing the calculation of the target variable. It is of crucial importance here to understand and correct process-related influences, also known as ‘hidden correlations’ (see section 2.2.1). The methodology for

calculating the target variable described in section 2.3 enables the identification of significant factors influencing product quality, which would not be evident when using alternative approaches, as shown in the example in section 3.2.1.

In addition to the method presented in sections 2.3 for calculating a valuable target value using FEM simulations, there are numerous other approaches to data transformation that aim to correct hidden correlations. In [7], several methods are presented and applied to the example presented in section 3.2.1. It was found that the transformations can improve the target variable. However, the calculation of the quality measure per batch described in section 3.1 delivers the best results. Although it is the most complex method, it is also the most precise, especially for small sample sizes. The other transformation methods described in [7], on the other hand, only work reliably with large sample sizes and are therefore limited in their application, especially in industrial environments.

The relevance of this approach is evident for the following reason: The targeted analysis and modelling of quality-relevant parameters (as described in the following section) provides valuable insights that contribute to the optimization of the production processes, particularly with regard to reducing scrap and ensuring consistent product quality.

3.3. Application - Controlling Influencing Parameters Using ML Models

3.3.1. General Approach

Machine learning models are becoming increasingly relevant in industrial manufacturing. The aim is to ensure product quality, reduce the rejection rate and use resources more efficiently. A key advantage of ML models is their ability to control critical parameters in a targeted manner. The full potential of this data-driven approach is particularly evident in automated and dynamic production environments [4,13,18,19].

A concrete example is the parameter P_{TS} whose influence on product quality can be proven to be significant as shown in the previous section 3.2. Deviations from the optimal range usually lead to quality deviations and increased scrap. To solve this problem, a machine learning model is being developed with the aim of controlling the parameter in a targeted manner. This method is designed to stabilize the signal value and adapt it to the given conditions. The model uses historical process data and relevant influencing factors to make accurate predictions.

Essentially, two issues can be identified that need to be examined more closely. The procedure described below can be applied to general industrial processes in which products are manufactured in batches, the conditions prior to the start of the process differ from those during the production process and signal values are continuously recorded throughout the entire production process.

The first question relates to predicting a signal value based on information that is available at the start of the process (example: start of casting). The aim is to use suitable modelling to enable the most accurate possible estimation of the expected signal value at the start of the production process, taking into account external factors (in the example: factors such as flow times in the melting channel system from the melting furnace to the mold, times in melt cleaning systems or ambient conditions).

Question 1: Predicting the signal value at the start of the process

- Data basis for the model:

The utilization of historical process data for batches that have already been produced in recent years is a key component. In this context, it is essential to ensure that the production process has not undergone significant changes during the period under consideration. Furthermore, specific external factors at the time of process initiation (e.g. ambient temperature, raw material properties) are utilized.

- Model:

The forecast is derived from the observed start conditions and historical patterns. The prediction is made for a single value per batch or process start. In such circumstances, static models such as regression or classification methods can be employed. Consequently, a conventional time series analysis is not required in this instance.

- Data preparation effort:

Preliminary data preparation and validation is a necessity. In comparison with question 2, the computing time and streaming requirements are lower.

- Model evaluation:

The accuracy of the signal value prediction at the initiation of the process, and the stability under various start conditions, are of paramount importance in this context. In the process of feature selection, it is firstly imperative to allocate particular attention to the identification of potential parameters that can be utilized to regulate the signal value at the start of the process. It is important to note that not all parameters can be controlled in a targeted manner (e.g. waiting times between two events). Consequently, it is crucially important in this context to keep these parameters as stable as possible across different start conditions.

The second question addresses continuous prediction during the ongoing production process (for example [4,21]). The aim of the model is to respond dynamically to changes during the production process and in the final step to intervene in a controlling manner if necessary.

Question 2: Real-time prediction during the process

- Data basis for the model:

Firstly, the model utilizes historical process data for batches that have been produced in recent years. Secondly, continuous sensor data is used during ongoing operation. This constitutes a fundamental distinction from question 1. At this point, it is important to note that external factors may be subject to change (e.g. fluctuations in metal flow, external conditions, etc.) during the production process.

- Model:

In this model, chronologically sequential process and sensor data play a pivotal role. It is imperative that alterations over time, such as trends or deviations, are detected and incorporated in real time. For this purpose, time series models can be utilized.

- Data preparation effort:

In contrast to question 1, continuous availability of data throughout the entire forecast period is required (e.g. IoT data). Furthermore, it is imperative to undertake real-time verification of data quality and model performance.

- Model evaluation:

In addition to the challenges outlined in question 1, there is also the issue of forecast quality under time pressure, as well as the response speed of the model. This is undoubtedly also pertinent to question 1, but is likely to be an even more significant challenge in question 2, given that parameter settings frequently only impact the parameter to be controlled following a specified time delay. This is described in more detail in 3.3.3.

In addition, it is necessary to discuss how the prediction models can be used productively in the IoT environment. Further challenges are outlined in section 3.3.3.

3.3.2. Use case

The implementation of this ML models involves the evaluation of historical data, with consideration given to data from batches since 2018 (comprising various alloys). It should be noted that the production specifications vary depending on the alloy.

With regard to question 1, an initial prediction model was developed based on a random forest regressor. The aim was to improve existing forecasts and create a robust basis for further analysis.

Figure 8 on the left side shows how the parameter has been controlled to date without using an ML approach. The Y-axis shows the difference between the input parameter, which is specified as the target value, and the actual value. It is very clear that the actual value is almost always below the target value.

Figure 8 (right side) shows the first prototype of an ML approach (random forest) for better control of the parameter [15–17,22]. It can be seen that the prediction deviates much less from the

specified value and that there is no longer any overestimation or underestimation. The sample size without missing/invalid values is approximately 1.200 batches.

The following key figures were used to evaluate the model quality [15]: The root mean squared error (RMSE = 2.56) is a quantitative key figure that describes the average deviation of the forecasts from the actual value in the original units of the target variable. A root mean square error (RMSE) of 2.56 therefore means that the forecasts fluctuate by this amount on average.

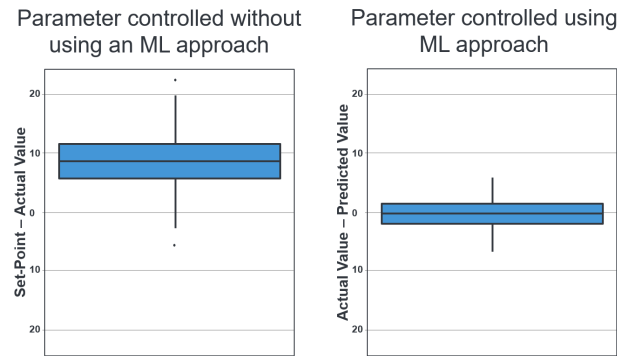


Figure 8. Left: Parameter controlled to date without using an ML approach; Right: Parameter controlled using an ML approach.

According to the model $R^2 = 0.65$, the model explains a variance of around 65% in the target variable. It was found that it is possible to map a significant proportion of the dispersion in the data. However, there is still room for improvement.

The mean absolute error (MAE = 2.07) refers to the mean absolute deviation of the forecasts from the actual value. The calculated forecast deviation of 2.07 is lower than that of the RMSE of 3.01, which indicates that the ‘typical’ errors are closer to the larger errors than previously assumed.

The model represents a significant improvement over the previous control system. With $R^2 = 0.65$ and an MAE of 2.07, the developed model already delivers very solid prediction quality for a first prototype and provides a robust basis for further development.

After an initial model comparison (Random Forest, XGBoost, linear models [15,22–24]) were carried out and no optimization could be achieved compared to the current Random Forest model, the focus of further work will now be on expanding and improving the features.

Therefore, the next step is feature-engineering. The aim is to increase the information in the input data in order to provide the model with additional parameters. The focus should be on the use of domain knowledge. The derivation of variables that reflect specific process- or material-knowledge is an essential aspect of this step.

After implementing the feature expansion, a comprehensive validation of the model quality is planned, including cross-validation and residual analysis.

This further work aims to use an extended feature base to evaluate whether it is possible to optimize the existing random forest model. In this context, hyperparameter tuning will be used, which includes various methods such as grid search, random search and Bayesian optimization [25]. A recent review of alternative model approaches will also be conducted.

3.3.3. Further Work & Challenges in General

In addition to further model development, operational feasibility must also be examined. This includes, in particular, the availability and data-quality of the input parameters for the model at the time the model makes a forecast. Incorrect data can lead to wrong predictions. It is therefore imperative to verify the data for plausibility and consistency, and to correct any discrepancies [26]. Furthermore, a comparison of the model types in terms of their performance and interpretability is necessary. The decision in favor of an explainable model depends on the priorities. An explainable

model is preferred when safety and traceability are more important than marginal performance gains.

In industrial applications, explainable models are preferred because transparency and traceability are crucial for process reliability and compliance. The use of black-box models is only justified in rare cases by minimal performance gains, particularly due to security and acceptance issues. In this case, risks must be compensated for by additional measures. In [26] and [27] for example, a perspective on this is presented in the automotive, chemical and process industries.

Addressing question 2 poses even greater challenges in the field of data science, which will be discussed below.

In the second question, the problem could be described as time-variable, condition-dependent lags and interval-based areas of effect. This is explained in the following, as it is found in various industrial areas. In [28], for example, a method is described for identifying delayed effects and cause-and-effect relationships in industrial sensor networks.

For each product or batch produced, several signals are recorded throughout the entire production period under specific product specifications. A major problem is that the signals can be shifted in time relative to each other, depending on the position of the sensors. The time delay (lag) is not constant. The dependence of the phenomenon manifests itself in two ways. On the one hand, there is a dependence on static factors, such as product specifications. On the other hand, there is a dependency on dynamic factors such as production speed or specific process fluctuations. The aim is to synchronize the signals in order to ensure comparability for predictions and further analysis.

Figure 9 shows a schematic example of two unsynchronized raw signals with a time shift of 0.5 seconds (first plot), and the synchronized signals without a time shift below for comparison [15].

In addition, it must be taken into account that some signals are influenced not only by a single point in time, but also by behavior in a preceding time interval. Here is an example: the value of a signal at time x depends not only on another signal at time $x - y$, but also on the behavior of this signal in then interval $(x - y_1, x - y_2)$. Therefore, the area of effect is defined by a time window, rather than a single point in time.

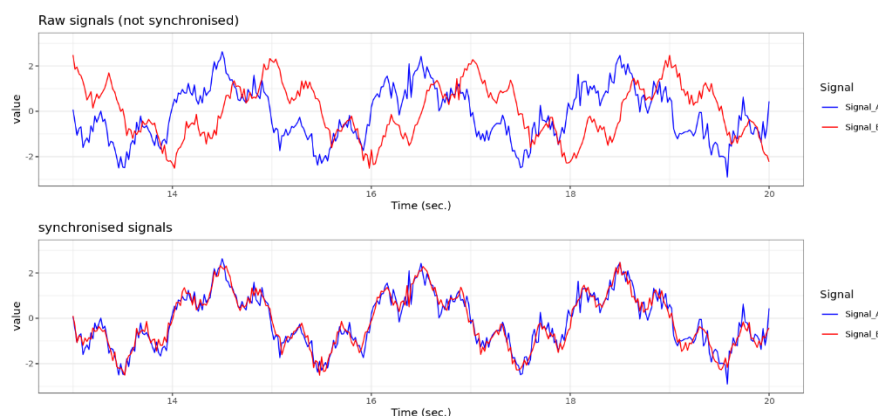


Figure 9. First Plot: schematic example for two unsynchronized raw signals with time-shift of 0.5 seconds; Second Plot: synchronized signals.

Future research will explore this issues in more detail. The aim is to develop a real-time forecasting model to predict the signal during production. With regard to the problem described above and the resulting questions, the following approaches to a solution are conceivable:

At the beginning of the research project, a baseline analysis should be carried out. For an initial assessment, this analysis may use methods such as cross-correlation and classic Dynamic Time Warping (DTW). These are time series analysis methods that enable the comparison of two time series [27]. The synchronized signals obtained in this way can then be used to train a prediction model for the target variables. The aim is to develop a model that considers time delays between signals and

their areas of effect, based on product specifications and process factors (basics of time series analysis with R, see for example [29])

3.3.4. Summary & General Vision for the Future

A key advantage of machine learning models is their ability to control critical parameters in a targeted manner. This data-driven approach unleashes its full potential, particularly in automated and dynamic production environments.

A concrete example, which has been presented above, is the targeted control of a parameter that has been proven to have a significant influence on product quality (see section 3.2). Deviations from the optimal range regularly lead to quality defects and increased scrap. To solve this problem, a machine learning model was developed that stabilizes the parameter and adapts it to the respective production conditions. It uses historical process data and relevant influencing factors to predict the value precisely.

As shown in section 3.3.2, a prototype for predicting the start value of a relevant parameter has already been developed. This serves as the basis for the targeted control of the parameter at the start of the process. In the next steps, the model will be further developed to enable real-time prediction of the signal value throughout the entire casting process.

Such a model could provide plant operators with specific control recommendations to keep the signal value within a defined range. This would be a significant step towards intelligent process control. A future vision could also envisage such models being integrated directly into the plant control system and intervening automatically.

Implementing this vision would increase process reliability and product quality and significantly improve the efficiency and responsiveness of industrial plants. The use of ML models therefore represents a strategically sensible and technologically forward-looking approach for modern production facilities [26,27].

3.4. Financial Benefit

The targeted control and analysis of process parameters is a pivotal aspect of data-driven production optimization at AMAG. A substantial number of projects (one example is presented in section 2.1 and 3.2, with further examples available in [18,19,30]) demonstrate that the utilization of data science, machine learning and interactive data analysis tools, in particular, yields the following advantages for the company:

The systematic analysis of process data facilitated the identification of patterns indicative of quality deviations. As demonstrated in section 3.2, the efficacy of utilizing data science to minimize scrap can be effectively illustrated. In this use-case, a 70% reduction in scrap was achieved [18,19].

The example illustrates that the targeted control of identified quality-relevant parameters can lead to a reduction in scrap, optimized plant utilization and shorter delivery times (see example in section 3.3.2).

The synergy that has been cultivated among the production, technology, IT, data engineering and analysis departments has been found to be remarkably efficacious.

Digitalization throughout the entire process chain facilitates the monitoring of machines and production processes [18,31]. The utilization of automated quality controls, leveraging artificial intelligence (AI) and predictive maintenance methodologies, is poised to play a pivotal role in the future by reducing machine downtime and ensuring product quality. Digital product tracking also facilitates complete traceability of errors [31,32]

These measures form the basis for innovation and competitiveness. The identification of quality-relevant parameters using data-driven approaches, and the subsequent specific control of these process parameters, is fundamental to the establishment of forward-looking, data-driven quality assurance. This is a pivotal factor in determining a company's future competitiveness in the global market.

4. Discussion

This example describes an approach that can be used in various industrial processes, particularly in multi-stage production processes where complex interrelationships exist and traditional methods are ineffective.

First, a precise target variable must be defined. This target variable acts as a quality indicator and forms the basis for any data-driven optimization. It is crucial that this target variable accurately describes product quality and can be compared across different production conditions [7].

Multi-stage industrial processes consist of coordinated production steps that depend on each other, which can result in so-called 'hidden connections' [7]. These are correlations between final quality parameters and external factors that should not actually have any influence. One example from aluminum production is the final plate thickness of the rolled ingot, which remains unknown at the time of ingot casting yet still influences the rejection rate of casting related defects. This can interfere with correct analysis and require special methods to neutralize it. The results presented here indicate that combining data science methods with physics-based FEM simulations can significantly improve the definition of target variables for statistical analyses and ML models. Compared to classic, purely data-driven approaches, the presented method enables robust, comparable target variables, even with small sample sizes. This is particularly relevant in industrial practice, where many specialized products are manufactured in small quantities, and where classic statistical methods often reach their limits.

The second step is to identify and analyze the factors influencing the calculated target variable. This analysis can be performed using machine learning approaches, such as feature importance from ML models, or by applying domain knowledge, such as material and process properties. The aim of this step is to identify relevant production parameters to be specifically controlled, taking into account given boundary conditions such as initial conditions or dynamic process parameters such as flow velocity or temperature changes.

The results of the investigation confirm the working hypothesis that considering and correcting process-related influences (as in the example the rolling process) is essential for correct data analysis. Previous studies have highlighted the issue of 'hidden correlations' in production data [7]. The method developed in this context addresses this challenge by using FEM simulation and data transformations to generate a target variable that is independent of subsequent process steps. This enables optimized comparability of the target variable between different product variants.

A key finding of the study is that applying this method enables significant factors influencing product quality to be identified, that would not be evident using alternative approaches. This can be illustrated by considering the casting signal 'TS' (Example in section 3.2.2), the influence of which on quality only becomes apparent through the transformation of the target variable using FEM simulation. The study emphasizes the importance of defining the target variable precisely for the successful application of statistical analyses and machine learning models in production contexts. Section 3.4 also describes the financial benefit that can be achieved by specifically controlling the signal 'TS', which has been identified as relevant to the quality of the final product.

The third step is where the real strength of machine learning comes into play. The results indicate that the Signal 'TS' (Example in section 3.2.2) significantly influences the quality of the final product. This has far-reaching implications: being able to control critical process parameters specifically opens up new avenues for intelligent, data-driven process control. Even in the prototype phase, the ML model presented in this paper demonstrates considerable optimization compared to current control methods. In future, integrating such models into plant control systems could enable automatic, adaptive process optimization.

In addition, efficient data processing is essential. To this end, cloud storage as a scalable storage option (e.g. AWS, Azure) and advanced ETL (Extract-Transform-Load) processes must be implemented. These enable the extraction, transformation and conversion of data from various sources into a standardized schema [33]. The use of high-performance computers or cloud-based platforms [34] is crucial for building an AI infrastructure.

Further steps also include the deployment of machine learning models within the domain of IoT environments, which gives rise to a number of technical and organizational challenges [26,35,36]. A pivotal factor in this regard pertains to the deployment location, namely whether it is situated on the edge, in close proximity to the machine, facilitating low-latency responses, or if it is hosted in the cloud, which offers augmented computational resources but may result in latency [35,36].

In order to guarantee accessibility for end-users, it is necessary to incorporate suitable services for model querying [36]. In addition, AI tools and analysis platforms are essential, such as Business Intelligence (BI) tools, machine learning platforms and AI services [37–39].

Furthermore, the frequency of data transmission, latency, and required reaction times must be carefully analyzed, particularly in the context of real-time prediction scenarios. It is imperative that models deliver results within seconds; therefore, efficient buffering strategies must be implemented to prevent delays [36].

Model robustness is equally critical, requiring continuous monitoring of sensor data quality to detect anomalies such as outliers. It is imperative to meticulously monitor the performance of models in order to identify any deviations or discrepancies in predictions. Moreover, it is essential to incorporate mechanisms for automated or manual retraining of models in response to changes in process conditions. Finally, the model's ability to handle measurement inaccuracies and data gaps is essential for ensuring reliability in real-world production settings (big data challenges in manufacturing see also for example [4,13,26,27]).

Future work must address all these challenges in order to enable the use of the ML model, the first prototype of which was presented in this paper, in a production environment.

In summary, this work demonstrates that the presented methodology can be applied to other industrial processes where complex relationships and small data sets make classical statistical analysis challenging. The combination of simulation, technological knowledge, data science and machine learning is a promising approach for enabling robust analyses and predictions, even with heterogeneous and small datasets.

Several areas of research are emerging for the future:

- Extension of the method for calculating a target variable using FEM simulations to other alloys, formats, and production lines.
- Development of analysis methods for small, non-normally distributed datasets and imbalanced datasets.
- Development of real-time prediction models for process signals with time-variable, condition-dependent lags and interval-based areas of effect.
- Integration of real-time prediction models for productive use in IoT and cloud environments.

The present study provides a comprehensive overview of the potential benefits that can be realized through the integration of simulation, technological expertise and data-driven analysis in the context of industrial production. The findings of this study are of particular pertinence not only to the aluminum industry, but also to other industries characterized by complex production processes.

Author Contributions: Conceptualization, M.L.S.; methodology, M.L.S, A.G, S.N and P.S.; software, M.L.S. and P.S.; validation, M.L.S, A.G and S.N.; formal analysis, M.L.S.; investigation, M.L.S.; resources, M.L.S.; data curation, M.L.S.; writing—original draft preparation, M.L.S.; writing—review and editing, M.L.S, A.G, S.N and P.S.; visualization, M.L.S.; supervision, M.L.S.; project administration, M.L.S.; funding acquisition, M.L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study and to allow for the commercialization of research findings.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AMAG	AMAG Austria Metall AG
AWS	Amazon Web Services
BI	Business Intelligence
DTW	Dynamic Time Warping
ETL	Extract, Transform, Load
FEM	Finite Element Method
IoT	Internet of Things
MAE	Mean Absolute Error
ML	Machine Learning
RMSE	Root Mean Square Error
TS	Time Series
US	Ultrasonic

References

1. Lee, J.; Kao, H.-A.; Yang, S. Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia CIRP* **2014**, *16*, 3–8.
2. Schreyer, M.; Gerber, A.; Neubert, S. Data Analysis of Production Data for Continuous Casting of Aluminium Rolling Ingots. *Thermec*, 2023.
3. Tschimpke, M. J.; Trutschnig, W.; Schreyer, M.; Neubert, S.; Gerber, A. Statistical methods for prediction of the quality of aluminium ingots. *Thermec*, 2023.
4. Escobar, C.; MCGovern, M.; Morales-Menendez, R. Quality 4.0: a review of big data challenges in manufacturing. *J. Intell. Manuf.* **2021**, *32*, 2319-2334.
5. Lee, J.; Jang, J.; Tang, Q.; Jung, H. Recipe Based Anomaly Detection with Adaptable Learning: Implications on Sustainable Smart Manufacturing. *Sensors* **2025**, *25*(5), 1457.
6. Wilhelm, Y.; Schreier, U.; Reimann, P.; Mitschang, B.; Ziekow, H. Data Science approaches to quality control in manufacturing: A review of problems, challenges and architecture. In *Service-Oriented Computing: 14th Symposium and Summer School on Service-Oriented Computing, SummerSOC 2020*; Springer: Cham, Switzerland, 2020; pp. 45–65.
7. Schreyer, M. Key Concepts and Challenges in Industrial Processes with a Focus on Defining Target Variables in Multi-Stage Production and Batch Data. In *Data Mining - Foundational Concepts and Cutting-Edge Advancements*, Tangina Sultana; Publisher: IntechOpen, London, 2025.
8. Schreyer, M.; Tschimpke, M.; Gerber, A.; Neubert, S.; Trutschnig, W. Applied Statistics in Industry: Defining an Appropriate Target Variable and Analysing Factors Affecting Aluminium Ingot Quality. In *Combining, Modelling and Analyzing Imprecision, Randomness and Dependence*; Ansari, J., et al., Eds.; SMPS 2024; Vol. 1458, pp. 457–465.
9. Prillhofer, B.; Antrekowitsch, H.; Bottcher, H.; En-Right, P. Nonmetallic inclusions in the secondary aluminium industry for the production of aerospace alloys. In *Light Metals–Warrendale Proceedings*, TMS 2008, 603.
10. Montgomery, D. C. *Statistical Quality Control: A Modern Introduction*, 7th ed.; Wiley: Hoboken, NJ, USA, 2013.
11. Faraway, J. J. *Linear Models with R*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2015.
12. Schreyer, M.; Tschimpke, M.; Neubert, S.; Gerber, A.; Trutschnig, W. Statistical methods for measuring and predicting the quality of aluminum slabs. *Light Metal Age* **2023**, 46–53.
13. Filz, M.-A.; Gellrich, S.; Herrmann, C.; Thiede, S. Data-driven Analysis of Product State Propagation in Manufacturing Systems Using Visual Analytics and Machine Learning. *Procedia CIRP* **2020**, *93*, 449-454.

14. Simon, P.; Falkinger, G.; Scheiblhofer S. Hot Rolling Simulation of Aluminium Alloys using LS Dyna. *11th European LS-DYNA Conference*, 2017.
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
16. Wickham, H. *Elegant Graphics for Data Analysis*; Springer-Verlag: New York, NY, USA, 2016. Available online: <https://ggplot2.tidyverse.org> (accessed on 3 Nov 2025).
17. Wickham, H.; et al. *dplyr: A Grammar of Data Manipulation (Version 1.1.3)*, **2023**. Available online: <https://dplyr.tidyverse.org> (accessed on 3 Nov 2025).
18. Schreyer, M.; Gerber, A.; Neubert, S.; Tschimpke, M.; Trutschnig, W. Interactive Visual Data Analysis for Quality Optimization in Aluminum Casting. Manuscript planned for submission to *VIS 2026: Visualization & Visual Analytics*.
19. Schreyer, M.; Gerber, A.; Neubert, S. Process & product optimization using interactive visual data analysis tools. *ALUMINIUM 2025*, *accepted*.
20. Burchett, W. W.; Ellis, A. R.; Harrar, S. W.; Bathke, A. C. Nonparametric Inference for Multivariate Data: The R Package nrmv. *J. Stat. Softw.* **2017**, 1–18.
21. Lu, Z.; Ren, N.; Xu, X.; Li, J.; Panwisawas, C.; Xia, M.; Dong, H. B.; Tsang, E.; Li, J. Real-time prediction and adaptive adjustment of continuous casting based on deep learning. *Communications Engineering* **2023**, 2, 34.
22. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, 2(3), 18–22.
23. Chen T.; He T. *xgboost: Extreme Gradient Boosting*, **2023**. Available online: <https://CRAN.R-project.org/package=xgboost>
24. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, 28(5), 1–26. Available online: <https://cran.r-project.org/package=caret>
25. Bartz, E.; Bartz-Beielstein, T.; Zaefferer, M.; Mersmann, O., Eds. *Hyperparameter Tuning for Machine and Deep Learning with R – A Practical Guide*; Springer, Singapore, 2023.
26. Matamoros, O.; Takeo, G.; Moreno, J.; Ceballos Chavez, B. A. Artificial Intelligence for Quality Defects in the Automotive Industry: A Systemic Review. *Sensors* **2025**, 25, 1288.
27. Mowbray, M.; Vallerio, M.; Perez-Galvan, C.; Zhang, D.; del Rio-Chanona, E.; Navarro-Brull, F. Industrial data science – a review of machine learning applications for chemical and process industries. *Reaction Chemistry & Engineering* **2022**, 7.
28. Chen, R.; Liang, S.; Wang, J.; Yao, Y.; Su, J.-R.; Liu, L.-L. Lag-Specific Transfer Entropy for Root Cause Diagnosis and Delay Estimation in Industrial Sensor Networks. *Sensors* **2025**, 25, 3980.
29. Krispin, R. *Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R*; Packt Publishing: Birmingham, UK, 2019.
30. Neubert, S.; Schreyer, M.; Gerber, A.; Pucher, P. Umsetzung von Big-Data-Analysen zur Verbesserung der Walzbarrenqualität. *ALUMINIUM* **2023**, 10, 35-38.
31. Haidenthaler, A.; Schreyer, M.; Pfeiffer, P.; et al. Insights into a cold rolling mill: a survey through length-related process data, strip splitting, and multiple passes. *Int. J. Adv. Manuf. Technol.* **2025**, 138, 2395–2411.
32. AMAG Austria Metall AG. *AluReport 2025 Ausgabe 2*, 2025. Available online: <https://www.amag-al4u.com/media/publikationen> (accessed on 3 Nov 2025).
33. Elmasri, R.; Navathe, S. B. *Fundamentals of Database Systems*, 7th ed.; Pearson: Boston, MA, USA, 2015.
34. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A. D.; Katz, R.; Konwinski, A.; et al. A view of cloud computing. *Communications of the ACM* **2010**, 53(4), 50–58.
35. Andriulo, F.; Fiore, M.; Mongiello, M.; Traversa, E.; Zizzo, V. Edge Computing and Cloud Computing for Internet of Things: A Review. *Informatics* **2024**, 11, 71.
36. Ficili, I.; Giacobbe, M.; Tricomi, G.; Puliafito, A. From Sensors to Data Intelligence: Leveraging IoT, Cloud, and Edge Computing with AI. *Sensors* **2025**, 25, 1763.
37. Microsoft. *Azure Machine Learning Documentation*, 2023. Available online: <https://docs.microsoft.com/en-us/azure/machine-learning/> (accessed on 3 Nov 2025).

38. Microsoft. *Power BI Documentation*, 2023. Available online: <https://docs.microsoft.com/en-us/power-bi/> (accessed on 3 Nov 2025).
39. Microsoft Corporation. *Microsoft Power BI [Software]*, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.