Data-analytics-based factory operation strategies for die-casting quality enhancement

Jun Kim1 · Ju Yeon Lee2

Received: 2 September 2021 / Accepted: 20 December 2021 / Published online: 10 January 2022 © The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
This paper proposes data-analytics-based factory operation strategies for the quality enhancement of die casting. We first define the four main problems of die casting that result in lower quality: [P1] gaps between the input and output casting parameter values, [P2] occurrence of preheat shots, [P3] lateness of defect distinction, and [P4] worker-experience-based casting parameter tuning. To address these four problems, we derived seven tasks that should be conducted during factory operation: [T1] implementation of exploratory data analysis (EDA) for investigating the trends and correlations between data, [T2] deduction of the optimal casting parameter output values for the production of fair-quality products, [T3] deduction of the upper and lower control limits for casting parameter input–output gap management, [T4] development of a preheat shot diagnosis algorithm, [T5] development of a defect prediction algorithm, [T6] development of a defect cause diagnosis algorithm, and [T7] development of a casting parameter tuning algorithm. The details of the proposed data-analytics-based factory operation strategies with regard to the casting parameter input and output data, data preprocessing, data analytics method used, and implementation are presented and discussed. Finally, a case study of a die-casting factory in South Korea that has adopted the proposed strategies is introduced.

Keywords Data analytics · Factory operation strategy · Quality enhancement · Die-casting

Abbreviations

\(i, i^*\) Variables that indicate the casting parameter data \((i = 1, 2, \ldots, I)\)

\(j\) Variables that indicate the sensor data \((j = I + 1, I + 2, \ldots, I + j)\)

\(V_i, V_j\) Observed output value of each casting parameter and sensor data

\(V_{\text{input}}\) Input value of each casting parameter data

\(t_{\text{constant}}\) Constant of the time period

\(V_{\text{avg}}, V_{\text{avg}}\) Average value of \(V_i\) and \(V_j\) observed during \(t_{\text{constant}}\) minutes

\(V_{\text{gap}}\) Average gap between the input and the output values of \(i\) during \(t_{\text{constant}}\) minutes

\(C_{\text{gap allowanace}}\) Constant indicating the allowance percentage

\(C_i, C_{\text{lower}}, C_{\text{upper}}\) Upper control limit of \(i\)

Lower control limit of \(i\)

Constant indicating the number of deviations allowed considering control limits

Constant indicating the number of deviations allowed considering defect causes

Coefficient of the regression model for each casting parameter and sensor data

1 Introduction

An industry that is not visible but is essential in producing the final product and that forms the basis of manufacturing competitiveness is called “root industry” [1]. The root industry provides the preprocessing technology for the manufacturing process of other industries, such as the automobile, shipbuilding, and information technology industries. The main task of the root industry is to process raw materials into parts of the final product, and quality management in the root industry is essential for improving the quality competitiveness of the final product. Some representative root industries are the die-casting, mold, plastic-working, heat treatment, and surface treatment industries. As most factories
in the root industry are small and old, problems regarding factory operation that can be solved with the application of data analytics are hard to be resolved. As smartification of the factories in other industries is already in progress, and as the interest in smart factories has been increasing recently, the need for data analytics application in the root industry is steadily being raised. This paper deals with a way of implementing data analytics in the root industry, focusing on die casting, a metal-casting process characterized by forcing molten metal into a mold cavity under high pressure.

Although the global casting production quantity is estimated to be 100 million tons per year and there are about 47,000 factories in the die-casting industry, 95% of the factories are small in scale, with less than 300 employees [2]. This has led to a lack of research and development capability of the most factories in the die-casting industry, which is connected to the chronic neglect of quality regarding problems resulting from the absence of data analytics. Outdated machines and poor factory environment have also served as barriers to the implementation of data-analytics-based problem-solving. The circumstances of the die-casting industry have forced the factories to operate based on the workers’ experience rather than on the use of the data analytics approach, and the situation has led to chronic quality-related problems. Here, we define four representative quality-related problems of factories in the die-casting industry that should be addressed by using data analytics in the factory operation: [P1] gaps between the casting parameter input and output values, [P2] occurrence of preheat shots, [P3] lateness of defect distinction, and [P4] worker-experience-based casting parameter tuning. The causes of these representative problems and the degree of data analytics advancement needed to address them are summarized in Fig. 1.

Let us examine each quality-related problem of die casting in detail. When a product is produced, the proper values of the key parameters of the die-casting machine (e.g., casting pressure, cylinder pressure, and injection velocity) should first be set in the machine. It would be ideal if the actual casting parameter output values were the same as the input values, but there exists a varied range of gaps between the output and input values in the field. Gaps usually occur due to the influence of the factory’s representative environmental factors, such as the factory temperature or humidity, and a significant gap between the input and output values of a casting parameter causes quality-related problems. We define this kind of mechanical problem affecting the casting quality as [P1].

When a die-casting machine is reactivated after the machine stops or when the casting parameter input values are adjusted while the machine is operating, preheat shots are likely to occur while the machine becomes to be ready. We define this as the [P2] problem. As the occurrence of preheat shots will render the resulting products not of fair quality, in which case they should be discarded, the points that are considered as occurrence of preheat shots during the die-casting process should be decided. This is usually decided, however, based on the worker’s experience, and thus often results in wastage, loss, or a product quality problem.

![Figure 1: Characteristics of the representative problems](image-url)
[P3] is directly related to product quality. A casted product is usually produced through several successive processes: casting, fitting, blasting, machining, and inspection. Although it would be ideal if the defects that occur in the casting process (e.g., short shots, pores, bubbles) were distinguishable right after the completion of the casting process, most defects can be distinguished only when all the succeeding processes have been completed. When a defective casted product flows into the processes following the casting process, wastage occurs, and the delivery lead time can be delayed if there are numerous defective casted products. The fact that the average time between the completion of the casting process and the quality inspection is about a week intensifies the problem, and the production plan is often changed to avoid delivery lead time delay.

[P4] is somewhat related to [P1] and [P3] in that casting parameter tuning should be conducted when a gap continuously appears between the casting parameter input and output values or when defects are successively detected. As most factories do not manage the numerical values of the gap between the casting parameter input and output values and do not use defect prediction algorithms, the casting parameter input values are adjusted only when defects that can be detected with the naked eyes consistently occur. How to tune the casting parameters is important; however, most die-casting factories are relying on the experience of their proficient workers to tune the casting parameters for machine input, but worker-experience-based casting parameter tuning does not always guarantee the best results and can even yield worse results with a higher defect occurrence rate.

In recent years, government support for advancing manufacturing processes has been provided to root industries, including the die-casting industry, and such government support has paved the way for the application of data analytics in the die-casting industry. According to the root industry technology trend report (KII (2017)), one of the biggest barriers to data-analytics-based factory management, the absence of data to be analyzed has started to be addressed through the construction of data collection infrastructures like the manufacturing execution system (MES) or point-of-production system (POP) in die-casting factories. However, no guideline has been developed to determine what collected data to analyze, how to analyze such, and even how the analysis results should be utilized in factory operation. This paper proposes data-analytics-driven operation management strategies for die-casting factories, focusing on addressing the aforementioned quality-related problems with real-case-based data analytics results obtained using real field data. The study thus sought to address an issue that is not merely theoretical but practical. As the technology transfer is expected to be fast because die-casting factories have similar collectable machine parameter and process data and encounter the same quality-related problems, the proposed strategies can be adopted by general factories in the die-casting industry. Thus, the main contribution of this paper is that it presents a practical guideline for data-analytics-based factory operation management to the factories in the die-casting industry.

The rest of this paper is organized as follows: Sect. 2 reviews the related studies. Section 3 discusses in detail the strategies for addressing each quality-related problem. Section 4 presents implementation cases of each strategy. Finally, Sect. 5 concludes the paper and discusses the recommendations for future studies.

2 Literature review

In this section, we present the previous studies related to data-analytics-driven approaches for supporting the manufacturing systems of die-casting factories. As to research regarding the die-casting process, studies of improving process efficiency and effectiveness in mechanical viewpoints (e.g. [3–8]) were actively conducted rather than data-driven viewpoints. The previous studies that applied data-driven methods on the die-casting process can be categorized into three types based on their subjects: (1) studies regarding the casting parameters; (2) studies regarding data analytics; and (3) studies regarding the manufacturing system of die-casting factories.

The previous studies on the casting parameters focused on revealing the correlations between the factory data or determining the optimal casting parameter values. Ratna and Prasad [9] tried to optimize the process parameters of a cold-chamber aluminum die-casting operation. Focusing on the pressure of the plunger and the temperature of the liquid aluminum, the researchers constructed an Ishikawa diagram and carried out an analysis of variance test to identify the impacts of the two parameters on the product quality. Finally, they developed an artificial neural network that derived the optimal casting parameter input values by utilizing the investigated impacts. Fitriana et al. [10] implemented failure mode and effect analysis and six-sigma-based diagnosis of a die-casting factory with five steps (i.e., measurement, analysis, improvement, and control) to improve the die-casting production process. Targeting a real die-casting factory in Indonesia, they tried to determine the appropriate standard operating procedures and parameter values by adopting the aforementioned methods. Morgado [11] concentrated on analyzing the monitoring data of die-casting factories in terms of the correlations between the process parameters and the factory’s environmental parameters. He integrated heterogeneous data sources of the die-casting process and preprocessed the noise of the data, after which he tried to visualize the impacts of the casting parameters on the quality of the die-casting process. Winkler et al. [12] tried to reveal the correlations between the casting
parameters (e.g., piston speed, piston acceleration, vacuum pressure, humidity, machine pressure) and the quality of the aluminum high-pressure die-casting process. Through numerical experiments, they concluded that the velocity of the piston and the temperature on the surface of the die significantly affect the quality of the die-casting product. Haghhighi et al. [13] developed a to-be-machined simulation-based method that automates digital fixture setups minimizing the scrap of the parts. By identifying all the possible locations of the parts in the machining fixture, the optimal location of the cast part setting on a machine tool is decided. Soban et al. [14] adopted the concept of visual analytics to determine which parameters cause high scrap rates in the die-casting process. They especially concentrated on the data regarding the acceleration phase and the second-phase velocity of the die-casting process. The results were utilized to identify and propose strategies to lower the scrap rate for die-casting factories. Kittur et al. [15] developed a back-propagation neural network algorithm to model the correlations between the parameters of the die-casting process (e.g., fast shot velocity, intensification pressure, phase change over point, holding time) and the quality-related parameters (e.g., surface roughness, hardness, and porosity). Through numerical experiments, the average absolute deviation of the proposed algorithm was revealed to be 7.27%. While other studies focused on the casting-related parameters, Liu et al. [16] concentrated on monitoring the energy data of die-casting machines. Indicators like energy per part and energy per action were developed to evaluate die-casting machines in terms of energy efficiency.

Most of the previous studies related to data analytics were conducted to predict or diagnose quality-related problems. Kim et al. [17] compared the performances of regression algorithms like linear regression, nonlinear regression, and tree-based regression, which predict the product quality of the die-casting process, through an experiment. Their experiment results showed that the tree-based regression algorithms outperformed the linear and nonlinear regression algorithms. Cashion et al. [18] developed a convolutional neural network that could classify the quality of the parts based on the image of the die used as a mold captured by a thermal camera. The accuracy of the algorithm was verified to be 90%. Gellrich et al. [19] proposed a quality prediction algorithm for the aluminum die-casting process using the visual analytics approach based on the feature selection technology. Their study, conducted in a die-casting factory producing aluminum knuckles, demonstrated the effectiveness of the algorithm that they developed. Kozlowski et al. [20] tried to enhance the quality of the high-pressure die-casting process by developing a data analytics model that predicts the fraction of faulty products using abnormal process parameter values. Weidrer et al. [21] mainly considered the temperature signal data extracted during a thermal manufacturing process. By decomposing the temperature signals using nonnegative matrix factorization, an unsupervised learning method, a simple signal can be converted into a physically meaningful data that can be utilized in data analytics for revealing a quality-related problem. Kim et al. [22] diagnosed status of the die-casting machine with a random forest algorithm utilizing casting parameters such as cylinder pressure, casting speed, cycle time, and spray time. Among the various status of the die-casting machine, they focused on to diagnose preheating status. Wu et al. [23] developed an online simulation method based on back propagation neural network to control and optimize speed of the injection system of die-casting machine. In order to control the injection speed in real time, they tried to reduce response time, overshoot, and steady-state error of the simulation method. Jianmin et al. [24] proposed a simulating analysis model to control temperature of die-casting mold. They analyzed how the mold temperature affects the quality of products, and built a simulation model with Magma software. Elser and Lechler [25] tried to automate die-casting process with numerical controls (NC). Among the processing variables of the die-casting process, they mainly focused on flow rate and drop height. They validated that the proposed automation algorithm can be utilized under the dedicated test environment. Hailin et al. [26] developed an intelligent control system to regulate vacuum pressure of the die-casting process. They utilized self-tuning proportional-integral-derivative algorithm to control the pressure of the die-casting process.

There have also been studies with systematic viewpoints that tried to construct practical data-analytics-driven systems for supporting factory administrators. Haokai and Peijie [27] constructed a remote monitoring system for the die-casting process that visualizes the machine status, operating parameters, injection data, and production information. They tried to enhance the quality of the die-casting process, starting from monitoring the data efficiently and effectively. Meisen et al. [28] proposed a framework for the semantic integration of monitoring data collected from different measurement systems and sensor networks. For a conceptual case study of the implementation of the proposed framework to high-pressure die-casting machines, they tried to aggregate the different process-parameter-related data (e.g., mold data, pump) in a hydraulic system. Lee et al. [29] proposed a big-data analytics platform especially suitable for supporting small and medium enterprises so they could realize an integrated data collection system environment between the legacy system and the data analytics platform, and so they could address quality-related issues by applying the data analytics models stored in the platform to die-casting factories. To verify the effectiveness of their proposed platform, they conducted a case study targeting a die-casting factory in South Korea. Zhao et al. [30] proposed data-analytics-driven die-casting smart-factory solutions concentrating on operation analysis.
and the decision-making support system. They also designed a three-layer cyber-physical system for die-casting factories. Park et al. [31] proposed an Internet of Things-based smart-factory architecture for small and medium die-casting companies that enables real-time monitoring of casting parameter data. They also proposed systematic methods of establishing a correlation between the casting parameters and the quality of production using data-analytics-based techniques. Vanli et al. [32] insisted that the die-casting process should achieve an integrated manufacturing system by adopting management strategies utilizing big data. They also ensured that easy access to the process data and data analysis for process optimization would enable the resolution of the die-casting process’s practical problems. Kim et al. [33] developed a data-analytics-based system for quality enhancement of the die-casting process in server-edge dualized structures. They tried to enhance the efficiency of implementing data analytics in the die-casting industry by adopting the dualized structure divided into an integrated server and an edge-computing device. Kim and Lee [34] proposed closed-loop data analytics system architecture for implementing data analytics in the environment of cyber-physical system. They also considered a server-edge dualized structure for developing the system architecture.

Although many studies have been conducted on data-analytics-driven approaches for supporting the manufacturing systems of die-casting factories, the previous studies mainly focused on the monitoring function or on generating simple data-analytics-based algorithms that cannot address operational issues. A crucial issue regarding data analytics in the field is how a factory can handle operational problems and can be efficiently and effectively operated by utilizing the data analytics results. Unlike the previous studies, this study focused on the operation strategies of die-casting factories dealing with the typical problems resulting in a lower-quality die-casting process. The main contribution of this paper is that it proposes a detailed guideline of data-analytics-driven factory operation strategies that can enhance the quality of the die-casting process and can be used to solve the problems occurring in the field. By defining the quality-related problems and the tasks that have to be accomplished to resolve them, we propose a systematic guideline for conducting data-analytics-driven factory operation.

3 Data-analytics-driven factory operation strategies

In this section, we describe our data-analytics-driven factory operation strategies for solving quality-related problems [P1]–[P4] in detail. In Sect. 3.1, the data that were utilized for establishing the proposed factory operation strategies are presented. In Sect. 3.2, the details of the strategy for addressing each problem are given. Finally, in Sect. 3.3, the proposed data-analytics-based factory operation strategies are summarized.

3.1 Data utilized for the proposed data-analytics-driven factory operation strategies

As this paper targets die-casting factories and presents data-analytics-driven operation management strategies for solving their quality-related problems, die-casting factories should first be understood, and how the data that were utilized for the study were chosen should be explained. Die-casting factories carry out the typical sequential production processes of casting, fitting, blasting, machining, and inspection. Raw materials are melted and forced into a mold by high pressure, and the approximate shape of the product is formed. Then, the impurities and residue are detached from the product, and the rough surface is made smooth in the fitting and blasting processes. In the machining process, the delicate shape of the product is formed with a computer numerical control machine. In the final inspection process, the quality of the finished product is investigated, and the product is released. In each stage of the production process, a number of data are generated and can be collected. As this study focused on the quality-related problems concerning the casting process in particular among all the production processes, we discuss herein the data generated when the casting process is conducted.

The casting-related data that can be utilized as independent variables of data analytics can be categorized into two groups: (1) casting parameter data and (2) sensor data. When a die-casting machine casts a product, the values of the casting parameters (e.g., casting pressure, injection velocity, physical strength) generated during the product casting are measured and provided to the users through a built-in data interface module. Although the types of casting parameter data that can be provided by the data interface module differ depending on the type and age of the die-casting machine, the primary casting parameters are managed by the data interface modules of all die-casting machines. To collect the other data that cannot be provided by the machine, sensors can be attached to the die-casting machine and the peripheral machines (devices). Data like the temperature and pressure of the molds, the temperature of the heating furnace and coolant (which assist the die-casting machine), and the temperature and humidity of the factory can be collected by the sensors. The sensor data can be utilized as the important independent variables as well as the casting parameter data. This study assumed that the die-casting factories that would apply the proposed data-analytics-driven factory operation strategies already have existing infrastructure for collecting the casting parameter and sensor data.
The dependent variables that were utilized for the proposed data-analytics-driven factory operation strategies were selected considering the problems defined earlier. The problems that require dependent variables for constructing data analytics models that can diagnose and predict the statuses of the casting process and product are [P2] and [P3]. [P2] needs outcome data on whether a preheat shot was produced or not, and [P3] needs quality inspection data indicating whether a product is defective or not. This means that the consequence data of the preheat shot and the product quality have to be secured and matched with the values of the relevant independent variables to apply the data-analytics-driven factory operation strategies. This study also assumed that die-casting factories have already collected the values of the dependent variables and have already matched these with the values of the independent variables for a certain period before carrying out the proposed strategies.

3.2 Details of the proposed factory operation strategies for solving each quality-related problem

In this section, we present the details of the proposed data-analytics-driven factory operation strategies for solving each of the quality-related problems defined earlier. Crucial information on the following is needed to understand the proposed strategies and to apply them to die-casting factories: (1) the input and output data; (2) data preprocessing; (3) the data analytics method used; and (4) implementation of the strategy. For each quality-related problem, the proposed factory operation strategy that can address it will be described in terms of these four categories of information.

3.2.1 Proposed factory operation strategy for solving [P1]

The purpose of the proposed factory operation strategy for solving [P1] is to understand the relationship between the casting parameter and sensor data and to derive some insights on managing the gaps between the casting parameter input and output data. In the strategy, statistical analysis rather than advanced machine-learning-based technologies is useful and should be conducted. By understanding the factory data through statistical analysis, the proper casting parameter output values for the production of fair-quality products and the gap allowance can be derived, along with insights like when a gap occurs often and which data are relevant. The main tasks for the strategy are as follows: [T1] implementation of exploratory data analysis (EDA) for investigating the trends and correlations between the data; [T2] deduction of the optimal casting parameter output values for the production of fair-quality products; and [T3] deduction of the upper and lower control limits for casting parameter input–output gap management.

3.2.1.1 Input and output data In die-casting factories, the casting parameter input values are usually constantly changed while the sensor data input values are usually fixed. To illustrate this, as the sensor data (e.g., temperature and humidity of die-casting factories) are not controllable in reality, the sensor data of the peripheral machines (e.g., temperature of the coolant and heating furnace) are usually utilized along with the relatively constant output values. The types of input data needed for the proposed data-analytics-driven strategy for solving [P1] should be decided considering the environment. The casting parameter input and output values and the sensor data output values are basically needed, and quality classification of each shot is also needed as the input data of the strategy. The input data should be managed with a time stamp for integration with relevant data. An issue that needs to be addressed is that data on the temperature and humidity of the factory should be gathered in multiple spots of the factory. The factory temperature and humidity are usually revealed as crucial data affecting the casting quality, so the data should be managed well as the input data of the strategy. The temperature and humidity of the factory, however, are measured differently according to the locations of the sensors in the factory, such as a section near the factory exit and a section in the middle of the factory. The factory should be properly divided into sections considering the arrangement of the exits, windows, and machines, and the data of each section have to be considered the input data of the strategy.

From the implementation of the proposed data-analytics-driven factory operation strategy for solving [P1] using the casting parameter input data, the following data was found to be useful for enhancing the quality of die-casting factories’ products: the statistical values of each casting parameter and sensor data, the correlation coefficients between the data, the optimal output value of each casting parameter, and the upper and lower control limits of each casting parameter. The statistical values and correlation coefficients can be used as quantitative data-analytics-driven factors to understand the die-casting factory, and the optimal casting parameter output values and control limits can be utilized as factors affecting the closing of the gaps between the casting parameter input and output values. The input and output lists of tasks for [P1] are summarized in Table 1.

3.2.1.2 Data preprocessing What has to be done first in the data preprocessing process for the proposed data-analytics-driven factory operation strategies is to integrate the casting parameter input data and generate datasets to be utilized for statistical analysis. The key parameters that can integrate the relevant casting parameter input and output data, sensor data, and quality classification are the time stamp and the shot ID. The casting parameter data and quality classification can be matched with the shot ID, and the cast-
Table 1 Input and outputs list of tasks in [P1]

| Task | Input | Data type       | Output | Data type      |
|------|-------|-----------------|--------|----------------|
| [T1] | Output values of casting parameter | Numerical | Time-series graph or boxplot | Graphic |
|      | Output values of sensor data   | Numerical | Statistical value | Numerical |
|      | Time stamp                   | Time     | Correlation coefficient | Numerical |
| [T2] | Output values of casting parameter | Numerical | Optimal output values of casting parameter | Numerical |
|      | Quality of product           | Categorical |               |             |
|      | Time stamp                   | Time     |               |             |
| [T3] | Output values of casting parameter | Numerical | Upper control limit of input–output gap | Numerical |
|      | Input values of casting parameter | Numerical |               |             |
|      | Quality of product           | Categorical |               |             |
|      | Time stamp                   | Time     | Lower control limit of input–output gap | Numerical |

Interpolation parameter and sensor data can be matched with the time stamp. As the casting parameter and sensor data often have different sample rate, data interpolation should be done when they are integrated with the time stamp. For example, let us assume that casting parameter and sensor data are collected every 30 s and 1 min, respectively. In this case, every other casting parameter data cannot have paired sensor data. The sensor data utilized as the input data of the strategy are observed to be continuous time-series data, so we can infer that the interpolated value (e.g., the average value of the previous and next sensor data corresponding to the target casting parameter) can be matched with the casting parameter value.

As real field data have numerous noises and missing values, the integrated input data have to be filtered through the data preprocessing processes. Recall that one of the primary purposes of the proposed strategy for solving [P1] is to understand the generalized environment of the factory. In other words, the data that need to be examined are the integrated input data of the strategy trimmed through outlier elimination and missing-value treatment, which make the data representative of the factory’s generalized circumstances. When outlier elimination is conducted, outliers have to be detected for each subset of the strategy’s input data classified by quality (e.g., fair-quality and defective types). The fact that the data distribution is observed to differ according to the quality of the product can result in the neglect of the data characteristics if the classified subset of the data is not considered. There is a possibility that some trends of the casting parameter or sensor data observed in a certain quality-related problem will be treated as outliers. For each classified casting parameter and sensor data, outliers can be determined with the interquartile range (IQR) of the data, which can be calculated as follows: (3rd quartile of the data) − (1st quartile of the data) = 1.5 × IQR. Data lower than (1st quartile − (1.5 × IQR)) or higher than (3rd quartile + 1.5 × IQR) can be investigated as outliers, and the outliers should be eliminated. Although there are numerous interpolation-based methods for treating missing values, interpolation-based missing-value treatment may lead to a nonconformity problem in the comprehension of data trends. Thus, we recommend that the missing values and outliers be eliminated.

3.2.1.3 Data analytics method In this subsection, we will present how data analytics is conducted in the proposed strategy for solving [P1] in relation to the three tasks that we defined. An issue that we have to discuss prior to presenting the details of the data analytics method is how to group the input data of the strategy to conduct statistical analysis. As the distribution and trend of the casting parameter and sensor data have deep correlations with time series, statistical analysis should also be conducted after grouping the data based on the time-series-based criteria. Moreover, although the EDA results sometimes cannot show certain data trends when EDA is conducted with a single-unit data value, the results of EDA utilizing statistical values like the average value or standard deviation of multiple data values may provide some meaningful insights. Therefore, we need to conduct statistical analysis utilizing data grouped according to the time-series-based criteria, such as days, week, sample rate, and seasons, or a certain number of successive data. For the first defined task (i.e., the implementation of EDA for investigating the trends and correlations between the data), the most popular and the simplest way of determining the distribution and tendency of the data is to visualize the data. For the visualization methods, we can use a time-series graph and a box plot. By observing the data with the time-series graph, the statistical information of several cases can be determined. For example, a case in which a certain parameter shows a periodic increase or decrease according to the time can be identified, along with how a certain parameter increases or decreases when the machine status is changed. With the box plot, approximate information regarding the distribution of the data can be obtained.
As the visualization methods of EDA present only approximate information that can be detected with the naked eyes, more specific statistical information should be obtained from each data by calculating the statistical values of each data, such as the average, maximum, minimum, quartiles, and standard deviation. When we analyze the data grouped in accordance with each quality classification, we can derive specific characteristics of the data distribution for fair quality and for each defective category as well as the time-relevant characteristics of the data, using time-series-based grouped data. Some insights (e.g., a certain casting parameter) can be obtained from the statistical values. For example, the statistical values are observed to be higher when a certain defect occurs compared with fair-quality products, or a certain casting parameter is observed to be higher during a specific time period compared with the other time periods. Such insights can help factory administrators understand the factory with the use of quantitative data. The gaps between the casting parameter input and output values can also be analyzed in a similar way.

The correlations between the casting parameter and sensor data are also important statistical data that can be utilized for understanding a factory quantitatively. By calculating correlation coefficient using all the data and the time-series-based grouped data, how the casting parameter input and output values can also be analyzed in a similar way.

With regard to the second main task of the proposed strategy for solving [P1] (deduction of the optimal casting parameter output values for the production of fair-quality products), the optimal casting parameter output values for the production of fair-quality products can be simply derived from the EDA results. The calculated statistical values of fair-quality products, such as the average or median values of each casting parameter, can be regarded as the optimal casting parameter output values. Whether the seasonal factors affect the optimal casting parameter values should be additionally considered. We have to check if the statistical values have differences in statistical significance according to the groups of data divided by time period. When the differences are statistically significant, the optimal casting parameter output values should be derived, along with the different time periods, such as day and night, hours, months, and seasons.

Based on the derived optimal casting parameter output values, the upper and lower control limits of each casting parameter can be derived using IQR. Let us assume that $x_{ij}^*$ represents the optimal output value of casting parameter $i$ of product $j$, and that $x_{ij}^{fair}$ indicates the observed values of the fair-quality products’ casting parameter $i$ of product $j$. The upper and lower control limits of each casting parameter of product $j$ can be derived as follows: $Upper \ limit = x_{ij}^* + c \times 3rd \ quartile \ of \ (x_{ij}^{fair} - x_{ij}^*)$, $Lower \ limit = x_{ij}^* + c \times 1st \ quartile \ of \ (x_{ij}^{fair} - x_{ij}^*)$, where $c$ is a constant. Note that if there is a difference in the optimal output values of specific casting parameter or sensor data by time period, the lower and upper limits should also be managed by time period.

### 3.2.1.4 Implementation

The results of the proposed strategy for solving [P1] contain quantitative information that can help factory administrators understand the factory data and obtain insights that can support their decision making on how to address [P1]. However, the effects of the proposed strategy with the purpose of managing gaps between the casting parameter input and output values can eventually be maximized when a monitoring system of casting parameter and sensor data is developed using quantitative information. The optimal output values of each casting parameter and sensor data can be set as the standard guidelines for real-time data management, and the gap between the real-time and guideline data can be calculated and monitored in real time. If the gap is monitored continuously and an alarm occurs when the real-time data deviate from the lower or upper control limits derived from the proposed strategy, [P1] can be effectively addressed. The insights obtained regarding the correlations between the data inferred from carrying out the proposed strategy can also be utilized more efficiently when a monitoring system is established.

It should also be noted that when the statistical values of the data are revealed to be significantly different by time period (e.g., shift, season, month), we should manage the data and adopt all the proposed strategies for solving [P1]–[P4] separately according to the statistically significant time periods.

### 3.2.2 Proposed factory operation strategy for solving [P2]

The purpose of the proposed data-analytics-driven factory operation strategy for solving [P2] is to determine whether the die-casting machine is currently in a normal or preheat condition. Unlike the proposed strategy for solving [P1], the proposed strategy for solving [P2] needs a more advanced data analytics method to diagnose the status of the die-casting machine, such as machine learning methods that can comprehensively consider the variances of the casting parameters. The main task of the strategy is to develop a preheat shot diagnosis algorithm ([T4]).
3.2.2.1 Input and output data As the status of the die-casting machine is determined by the casting parameter data directly related to the machine, the values of the casting parameter data should be utilized as input data of the proposed strategy for solving [P2]. Sensor data that can affect the casting parameter data should also be considered input data. Additionally, the labels of the machine status that indicate whether the condition of the machine is normal or preheat have to be considered input data for utilization as dependent variables of the developed algorithm. An important issue regarding the input data of the proposed strategy is the classification problem of the labels. To develop a preheat shot diagnosis algorithm, the exact status of the die-casting machine should be collected for a certain period of time. However, in die-casting factories, factory administrators usually do not collect the exact machine status data indicating whether a shot is a preheat shot or not because doing so requires much effort. Even so, as the exact labels have to be collected to solve [P2], time and effort must be invested for such. The output of the strategy is a machine-learning-based algorithm for diagnosing preheat shots and classification of the casted shot as normal or preheat. The input and output lists of tasks in [P2] are shown in Table 2.

3.2.2.2 Data preprocessing The first stages of data preprocessing (i.e., integration of the input data, outlier elimination, and missing-value treatment) can be conducted in a way similar to the proposed strategy for solving [P1]. The discriminated data preprocessing in the strategy concerns the generation of training datasets utilized in the development of machine-learning-based algorithms. How training datasets are organized affects the performance of such algorithms. With regard to the organization of the training datasets for the strategy, there were two main issues: (1) the independent variables of the algorithm that was to be developed (casting parameter data) could have subordinate relations with each other and (2) the number of dependent variables (labels of the machine status) is often imbalanced. Let us examine these two issues and how we can handle them using data preprocessing.

The accuracy of the machine-learning-based data analytics model can vary according to the data sets utilized. To illustrate this point, considering that a large number of data columns does not always yield the best results, for example, when multiple independent variables with high correlations are simultaneously considered a training dataset of the machine-learning-based algorithm, an overfitting problem can occur due to the redundancy of the data features. There are numerous data that can be collected from a die-casting factory, and only the data that can well represent the features of the factory should be selected. A way of selecting best-fitting data columns is to conduct principal component analysis (PCA), an eigenvector-based multivariate analysis method. PCA is an orthogonal linear transformation method that transforms data into a new coordinate system sequencing the scalar projection of the data on the volume of variance. PCA also provides a principal component score (weight of coefficient vector) for each vector of data, and a high score (weight of coefficient vector) means that the data can represent the domain well. The ratio of the weight of coefficient vector to the total weight signifies the degree to which the corresponding data can represent the domain. When all the weights of coefficient vector are arranged in descending order, the dataset corresponding to the over 0.8 or 0.9 weight of the accumulated proportion from the beginning is selected as the representative dataset.

Of course, there are diverse methods to determine the relative significance of manufacturing process parameters other than PCA; however, PCA is recommended with the following reasons. First, it can effectively reduce dimensions of the data in the direction of maximize deviation of data characteristic. The objective of PCA is to select combination of data that can cover the maximum range of data features without redundancy. As mentioned earlier, it is needed to eliminate redundancy of data features caused by considering multiple variables that have high correlation with each other. PCA can efficiently eliminate the redundancy of data and reduce over-fitting problem. Second, unlike the other methods, PCA provides score of each variable that stands for the relative significance in terms of covering data features without redundancy. The scores can efficiently support the decision makers to select the proper processing variables. Lastly, it is easy to use since almost all the commercial statistical analytics tools provide PCA function. Since this study is conducted to support efficient and effective spread of data-driven operation strategies to die-casting industry, we recommend PCA that can be easily utilized.

In the die-casting industry, there is an extreme preheat-normal shot occurrence imbalance. When a data-analytics-based algorithm for diagnosing preheat shots is developed

| Task                  | Input                        | Data type        | Output                        | Data type        |
|-----------------------|------------------------------|------------------|-------------------------------|------------------|
| [T4]                  | Output values of casting parameter | Numerical        | Diagnosed machine status (Preheat or Normal) | Categorical      |
|                       | Output values of sensor data  | Numerical        |                               |                  |
|                       | Machine status (for training) | Categorical      |                               |                  |
|                       | Time stamp                   | Time             |                               |                  |
utilizing imbalanced data, there is a high probability that the base rate fallacy (i.e., the tendency to ignore the data pertaining to a small number of cases and to focus on the base rate data) will occur. For example, in the case of the die-casting industry, as the number of normal-shot data is generally much bigger than the number of preheat shot data, an algorithm that diagnoses all the shots as a normal shot can have a higher classification accuracy than other algorithms. To avoid committing the base rate fallacy, the data imbalance should be resolved by conducting oversampling, such as the synthetic minority oversampling technique (SMOTE).

3.2.2.3 Data analytics method As the purpose of this study was to propose general data-analytics-driven factory operation strategies rather than to propose a brand-new data analytics method, we focus here on how we can effectively utilize the previously developed data analytics methods. As there are several previously developed machine learning methods, such as decision tree, random forest, support vector machine (SVM), neural network, AdaBoost, and XGBoost, and as the performance of the data-analytics-based algorithm depends on the machine learning method used, selecting the proper machine learning method is an important issue. A simple and clear way of doing this is to compare the performances of all the data-analytics-based algorithms, such as through the mean absolute percentage error (MAPE). For evaluating the performance of each data-analytics-based algorithm, k-fold cross-validation is recommended. k-fold cross-validation splits data in k sets, and a set is selected as the test set while the other k − 1 sets are combined into the corresponding training set. This means that the performance of each algorithm can be evaluated with k numerical experiments. A machine-learning-based algorithm that shows the best performance can be selected as the suitable data-analytics-based algorithm.

The next step we should carry out after we finalize and generate a data-analytics-based algorithm is to calculate the feature importance of each of the data utilized for data analytics, which indicates the relative impact of such data on the algorithm performance. When the developed algorithm is used in the field, the algorithm performance evaluated in the algorithm validation steps cannot be guaranteed. Unlike the preprocessed training data, real field data have noises that can affect the performance of the data-analytics-based algorithm. Moreover, due to the great volume of field data, the computational time required for generating the algorithm can be too long. These problems can be solved by considering the concept of feature importance. The feature importance of each column data can be calculated as the relative extent to which it is reduced by ignoring the column data when the data analytics model is generated. The steps for deriving the feature importance are as follows: (1) develop data analytics models utilizing the selected machine learning method, and exclude each one of the data from one model; (2) calculate the gaps in accuracy between the original data analytics model and each model from which one of the data was excluded; and (3) normalize the gaps. The normalized gap becomes the feature importance of each excluded data. When the data analytics model generated using only the data with a high feature importance value is applied to the field, the algorithm’s validation performance can decrease, but its field implementation performance may increase. The features to be finally adopted should be selected based on the field application test results.

The last step is to set criteria for the update of the generated data-analytics-based algorithm. With the passage of time and as the factory environment changes, the field accuracy of the data-analytics-based algorithm can decrease. To maintain the accuracy of the data-analytics-based algorithm, the algorithm should be continuously updated with the most recent data. The criteria for algorithm update can be set with two standards: time and field accuracy. When a certain time period flows or when the field accuracy of the data-analytics-based algorithm becomes lower than a certain level, we can update the algorithm. However, as the updated algorithm does not guarantee better performance than the original algorithm, we should compare the performances of the updated and original algorithms and adopt the one with a better performance.

3.2.2.4 Implementation When a shot is casted, the real-time casting parameter data relevant to the shot are integrated in a record and should be entered in the developed preheat shot diagnosis algorithm. When the machine status at the time that the shot is casted is determined as preheat, the shot should be discarded from the lot. For the implementation of the proposed strategy, an infrastructure that enables real-time diagnosis of the die-casting machine status and that enhances the quality of die casting by discarding the shots that are not of fair quality is needed.

3.2.3 Proposed factory operation strategy for solving [P3]

The purpose of the proposed data-analytics-driven factory operation strategy for solving [P3] is to predict whether a casted product will be of fair quality or will become defective right after a shot is casted. Moreover, by diagnosing the causes of the defects created when a shot is predicted to have a defect, the said strategy can assist the factory administrator in making a quality-related decision. There are two main tasks for the strategy: [T5] development of a defect prediction algorithm and [T6] development of a defect cause diagnosis algorithm. As the stages of data preprocessing and developing a defect prediction algorithm are similar to those of the proposed strategy for solving [P2], we will skip
the said stages here and we will focus on the differentiated contents.

### 3.2.3.1 Input and output data

For the input data of the strategy, the casting parameter data, sensor data, and quality classification for each shot are utilized when the target factory for the strategy has already constructed an infrastructure for tracking each shot over all the processes of the die-casting factory. To develop a defect prediction algorithm that predicts the quality of each single shot, the quality information of each shot should be mapped with the corresponding casting parameter and sensor data. This means that the ID of each shot has to be granted so that the data regarding a shot will be trackable when the shot is realized to be defective. However, there are actually few die-casting factories that have infrastructure enabling data tracking in product unit. For the factories that cannot develop a defect prediction algorithm in shot unit because they do not have a data collection infrastructure, we propose that a defect prediction algorithm be developed in lot unit. Lot information, including the production time, production quantity, and number of defects, is managed in most die-casting factories. By using the lot information and time stamps, we can match a lot with the set of casting parameter and sensor data collected during the production of the products belonging to the corresponding lot. We can then develop a defect prediction algorithm that can forecast the number of defects for each lot. The results of the defect prediction algorithm in lot unit are also meaningful because the defects of casted products often occur consecutively, and as such, defective casted products are likely to belong to the same lot. As a lot includes a certain number of shots, single values of the casting parameter and sensor data of each shot cannot be utilized as independent variables. Instead, what can be utilized as independent variables are the calculated statistical values of each casting parameter and sensor data of the shots included in the same lot (e.g., average, minimum, maximum, skewness, standard deviation, increasing velocity, decreasing velocity). The shapes of the dependent variables (labels) can also be somewhat different from those in the algorithm in product unit, which considers the quality classification of each shot as a dependent variable. There are several ways of setting the dependent variables of the defect prediction algorithm in lot unit. The number of each quality classification, the ratio of each quality classification to the total product, and the level of defective shots contained in the lot can be regarded as the dependent variables. Among the candidates, we recommend that the level of defective shots contained in the lot be considered a dependent variable because the machine-learning-based algorithm can show better performance when the algorithm classifies data as a class than predicts a certain numerical value. When the classification levels of lots according to the proportion of defective shots among all the shots are 0–10%, 11–20%, 21–30%, and over 30%, for example, the developed algorithm becomes a classification algorithm, and its accuracy may be better than that of the other cases. The input and output lists for each task in [P3] are illustrated in Table 3.

### 3.2.3.2 Data analytics method

As the data analytics method for the development of a defect prediction algorithm is similar to the strategy for solving [P3], as mentioned earlier in this subsection, we present herein the details of the data analytics method for the development of a defect cause diagnosis algorithm. To enhance the quality of die casting, the defect causes should be determined because the defects of die casting tend to continuously occur in factories as a result of the casting process until the machine is adjusted right after the first defect occurrence. We propose two ways of developing a defect cause diagnosis algorithm depending on the machine learning method used.

When tree-based data analytics models are utilized, we can infer defect causes from the branch points of the trees. When decision-tree-based ensemble models utilizing boosting or bagging are developed, multiple trees that consider randomly or statistically selected data features

---

**Table 3** Input and output lists of tasks in [P3]

| Task | Input                                                                 | Data type  | Output                      | Data type |
|------|-----------------------------------------------------------------------|------------|-----------------------------|-----------|
| [T5] | Output values of casting parameter                                    | Numerical  | Predicted quality of product| Categorical|
|      | Output values of sensor data                                          | Numerical  |                             |           |
|      | Quality of product (for training)                                     | Categorical|                             |           |
|      | Time stamp                                                             | Time       |                             |           |
| [T6] | Quality of product                                                    | Categorical| Cause of defect             | Categorical|
|      | Output values of casting parameter                                    | Numerical  |                             |           |
|      | Input values of casting parameter                                     | Numerical  |                             |           |
|      | Output values of sensor data                                          | Numerical  |                             |           |
|      | Input values of sensor data                                           | Numerical  |                             |           |
|      | Upper control limit of input–output gap                               | Numerical  |                             |           |
|      | Lower control limit of input–output gap                               | Numerical  |                             |           |
are generated, and an integrated model is constructed. Whether a shot is defective or not is determined by the principle of majority rule reflecting the prediction results of each tree. When a shot is predicted as a defect, we can derive specific data conditions that make the model decide that the shot is defective from a set of majority trees. We extract the conditions of the lowest branches of trees included in the majority set and consider the intersection conditions for the lower-bound conditions as the causes of the defect. For example, when the lowest branches are observed to be \( x_1 \leq 20, x_4 \leq 15, x_2 \geq 20, \) and \( x_3 \geq 30, x_1 \leq 15 \) and \( x_2 \geq 30 \) become the defect causes.

When non-tree-based machine-learning-based algorithms are considered, we can utilize the upper and lower control limits of each casting parameter and sensor data derived in the proposed strategy for solving [P1] for deriving the defect causes. Whenever a defect is predicted, we can compare the corresponding values of the casting parameter and sensor data with the control limits and regard the deviated condition as a defect cause. The proposed algorithm is summarized in Table 4.

### 3.2.3.3 Implementation

When every shot is casted, a record consisting of the real-time values of the casting parameter and sensor data utilized as independent variables of the defect prediction algorithm should be generated, and then, whether the shot is defective or not should be predicted right after the shot is casted, by implementing the defect prediction algorithm. When a shot is predicted as defective or when a lot is predicted to include more than a certain percentage of defective shots, the shot or lot have to be discarded. When a defect is detected, its causes should be diagnosed and relayed to the factory administrators.

### 3.2.4 Proposed factory operation strategy for solving [P4]

The purpose of the proposed data-analytics-driven factory operation strategy for dealing with [P4] is to derive quantitative data-analytics-based guidelines for the casting parameter tuning process. As a casting parameter can be affected by the other casting parameter data and by the sensor data, the extent of the adjustment of a casting parameter input value needed to change the output value to the intended value can differ according to the factor’s circumstances. Therefore, we should determine the extent of the casting parameter input value adjustment considering its correlations with the other data. The main task of the strategy is to develop a casting parameter tuning algorithm considering the characteristics of the problem ([T7]).

#### 3.2.4.1 Input and output data

To implement the proposed strategy for solving [P4], we need information regarding when we will tune the casting parameter input value and how we will do this. For the former information, we need the upper and lower control limits of each casting parameter data and the defect causes, which are the output data of the implemented strategies for solving [P1] and [P3], respectively, as the input data of the strategy. When the values of a specific casting parameter are observed to have deviated from the limits or when similar defect causes are continuously detected, the factory administrators can make a decision to tune the casting parameter input value. The input data of the strategy that can be utilized as independent variables are the input value of each target casting parameter data and the corresponding output values of the other casting parameter data and of the sensor data. For an output value of each casting parameter, the input value of the casting parameter and the output values of the other data should be integrated, and the integrated data should be preprocessed. As the preprocessing procedures are similar to those in the

| Table 4 Defect cause diagnosis algorithm |
|-----------------------------------------|
| >> Start the algorithm when a shot is predicted as defective |
| >> If the defect prediction algorithm is a tree-based model; |
| >> search the trees included in the majority set that classify a shot as defective; |
| >> bring the lowest branches of each tree; |
| >> for the lower-bound conditions, get the intersection conditions of each corresponding data; |
| >> for the upper-bound conditions, get the union conditions of each corresponding data; and |
| >> diagnose the defect causes as the derived conditions; |
| >> Otherwise, |
| >> bring the values of each casting parameter and sensor data corresponding to a shot that is predicted as defective; |
| >> compare the values with the upper and lower control limits derived from the strategy for solving [P1]; |
| >> find the values deviating from the control limits; and |
| >> diagnose the defect causes as the corresponding control limits of the data; |
other proposed strategies in terms of outlier elimination and missing-value treatment, we will no longer discuss the data preprocessing part here. The output value of the target data is regarded as a dependent variable. As the output of the strategy, regression equations that can infer how the output value of a certain casting parameter data will be changed as the input value is changed are derived. Moreover, an expert system that determines when to tune the casting parameter input value and the extent of adjustment needed is derived. Input and output lists of tasks in [P4] are shown in Table 5.

3.2.4.2 Data analytics method

The data analytics method can be divided into two components: that for finding the triggers of casting parameter tuning and that for deciding the extent of input adjustment needed for a target casting parameter data. For the former component, we should develop an expert system that can detect the criteria for activating the casting parameter tuning algorithm. The triggers of the casting parameter tuning process can be defined as three cases: (1) when the gap between the input and output values is continuously observed to be high; (2) when the real-time values of the casting parameter or sensor data continuously deviate from the control limits; and (3) when the number of certain defect causes is predicted in a certain time period. For the first case, we have to set a specific time period during which a gap between the input and output values will be continuously observed, as well as the extent of gap allowance. For example, when we set the time period as 10 min and the gap allowance as 5%, the average gap between the casting parameter input and output data for every 10-min period should be calculated, and whenever the average gap deviates 5%, the system has to detect the circumstance. The time period and gap allowance can be set differently as casting parameters. In the second case, we can set the deviation allowance number and standard time period. For example, in a case where we set the deviation allowance number as 5 and the time period as 10 min, the system detected a circumstance in which the real-time values of a certain casting parameter deviated more than 5 times in the recent 10 min as a trigger of casting parameter tuning. The criteria for the third case are somewhat similar to those for the second case. The deviation allowance number of a certain defect cause’s occurrence and the time period should be set.

When a trigger is detected, we should determine the intended change of the target casting parameter data’s output value. When the first trigger is activated, the average gap can be the intended extent of adjustment, and the average extent of deviation can be the extent of intended change for the second case. For the third case, the average gap between the real-time data and the defect cause diagnosis criteria can be determined as the intended adjustment value.

The next step is to develop a data analytics model for each casting parameter data that derives the input value of the target casting parameter for adjusting the intended output values. Although there are many methods that can derive the input value for casting parameter tuning, we recommend the multivariate linear regression method, which can intuitionistically present the relationship between the input and output values of the target casting parameter data with mathematical models reflecting the correlations between the target casting parameter and other data. To modify the input value of the casting parameter as a way of changing the output value to the intended amount, and to develop the function, we need clear guidelines, such as mathematical equations. Moreover, the fact that the key casting parameter and sensor data are those of the velocity, pressure, and temperature, which are revealed to be proportional to one another, makes the multivariate linear regression method suitable for the strategy. For a target casting parameter data that we want to use to derive a regression model, let us assume that there are \( n \) instances of \( m \) relevant independent variables \( x_i (i = 1, \ldots, m) \), and that \( x_{m+1} \) indicates the input value of the casting parameter data. In the case in which \( H \) indicates the function of deriving the output value of the target casting parameter, the purpose of developing a multivariate linear regression model is to derive an equation from

| Task | Input | Data type | Output | Data type |
|------|-------|-----------|--------|-----------|
| [T7] | Upper control limit of casting parameter | Numerical | Time to modify input values of casting parameter or sensor data | Time |
|      | Lower control limit of casting parameter | Numerical |        |           |
|      | Upper control limit of sensor data | Numerical |        |           |
|      | Lower control limit of sensor data | Numerical |        |           |
|      | Output value of casting parameter | Numerical | Modified input value of casting parameter | Numerical |
|      | Input value of casting parameter | Numerical |        |           |
|      | Output value of sensor data | Numerical |        |           |
|      | Input value of sensor data | Numerical | Modified input value of sensor data | Numerical |
|      | Quality of product | Categorical |        |           |
|      | Defect cause | Categorical |        |           |
\[ H(x_1, \ldots, x_{m+1}) = w_1x_1 + w_2x_2 + \cdots + w_{m+1}x_{m+1} + c, \] where \( w_i \) is the coefficient weight and \( c \) is a constant. The cost function of the regression model is calculated as

\[ \text{cost}(W, b) = \frac{1}{n} \sum_{i=1}^{n} (H(x_1, \ldots, x_{m+1}) - y_j)^2, \]

where \( x_i \) and \( y_j \) are the value of each independent variable and the value of the dependent variable for instance \( j \). The values of \( w_i \) for \( i = 1, \ldots, m + 1 \) and \( c \) are derived by minimizing the cost function with the gradient descent algorithm.

The input value of the target casting parameter can be derived through the backward calculation of the derived equation.

Selecting proper independent variables relevant to a target casting parameter data is an important issue when a regression model is generated. We can consider all the casting parameter and sensor data or consider some parts of the dataset with high correlations with the target casting parameter data. As the accuracy of the regression model can vary according to the dataset utilized, we should compare the accuracy levels of several cases, such as by considering all the data and considering only the data that has higher correlation coefficient values than a certain extent. The regression model with the highest accuracy level will be selected as the equation to be utilized for the casting parameter input value tuning algorithm.

Whenever a trigger of the casting parameter tuning process is detected and the intended extent of output value change is derived, the input value that should be adjusted is determined using the developed regression models. By substituting \( x_i (i = 1, \ldots, m) \) with the average value of the corresponding data recently observed during a certain time period and \( H(x_1, \ldots, x_{m+1}) \) with the intended output value of the target casting parameter data, we can calculate the input value of the target casting parameter data. It was noted, however, that when an input value of a target casting parameter is adjusted, the other casting parameter data related to the target casting parameter data are also affected. Hence, the input values of the related casting parameters included in the regression model of the target casting parameter should also be adjusted, along with the input values of the target casting parameter, maintaining the same extent of adjustment of the output values. The expected output value of the target casting parameter data should be reflected in the stage.

3.2.4.3 Implementation To implement the results of the proposed factory operation strategy for solving [P4] in die-casting factories, the consecutive stages of the casting parameter tuning process, from the detection of process triggers to the adjustment of the input value of the target casting parameter data, should be integrated into a complete system. Table 6 shows the detailed stages of the system, including the function of trigger detection, the determination of the intended change of the target casting parameter data’s output value, and the extent of the target casting parameter data’s input value adjustment.

3.3 Summary of the proposed factory operation strategies

In this section, we summarize the four proposed data-analytics-driven factory operation strategies for solving [P1]–[P4] and present systematic viewpoints of implementing the strategies in an integrated way. For each strategy, we derived the detailed tasks for solving each quality-related problem, as follows: [T1] implementation of EDA for investigating the trends and correlations between data, [T2] deduction of the optimal casting parameter output values for the production of fair-quality products, [T3] deduction of the upper and lower control limits for casting parameter input–output gap management, [T4] development of a preheat shot diagnosis algorithm, [T5] development of a defect prediction algorithm, [T6] development of a defect cause diagnosis algorithm, and [T7] development of a casting parameter tuning algorithm. The algorithms that we derived are summarized in Fig. 2.

Table 6: Casting parameter tuning algorithm

| Step | Description |
|------|-------------|
| 1    | Calculate the average output value of each casting parameter data \( i \) and sensor data \( j \) for the recent \( t_{\text{constant}} \) minutes, and set the value as \( V_i^{\text{avg}} \) and \( V_j^{\text{avg}} \), respectively. |
| 2    | Calculate the average value of \( (V_i - V_i^{\text{input}}) \) for each casting parameter data \( i \) for the recent \( t_{\text{constant}} \) minutes, and set the value as \( V_i^{\text{prop}} \) |
| 3    | If \( \frac{|V_i^{\text{prop}}|}{V_i^{\text{avg}}} \times 100 \geq C_{\text{allowance}}^{\text{prop}} \) for any \( i \), update set \( A \) with \( i \) that satisfies the condition, and \( V_i^{\text{change}} \) as \( V_i^{\text{prop}} \). Then go to Step 6 |
| 4    | If \( n(\{ V_i | V_i > C_i^{\text{upper}} \}) \cup \{ V_i | V_i < C_i^{\text{lower}} \}) > C_{\text{allowance}}^{\text{prop}} \) for any \( i \) during the recent \( t_{\text{constant}} \) minutes, update set \( A \) with \( i \) that satisfies the condition, and \( V_i^{\text{change}} \) as the average value of deviation \( (V_i - C_i^{\text{upper}}) \) or \( (V_i - C_i^{\text{lower}}) \). Then go to Step 6 |
| 5    | If the number of certain defect cause criteria during the recent \( t_{\text{constant}} \) minutes for any \( i \) is observed to be more than \( C_{\text{criteria}}^{\text{allowance}} \), update set \( A \) with \( i \) that satisfies the condition, and \( V_i^{\text{change}} \) as the average value of deviation. Then go to Step 6 |
| 6    | For each regression equation of \( i^* \in A, V_i^{\text{change}} = \sum_{j \neq i^*} w_{ij} V_j + \sum_{j \neq i^*} w_{ij} V_j^{\text{input}} + w_{ij} V_j^{\text{prop}} + c, \) substitute \( V_i \) with \( V_i^{\text{avg}} + V_i^{\text{prop}}, V_j \) with \( V_j^{\text{avg}} \), and \( V_j \) with \( V_j^{\text{change}} \) |
| 7    | Calculate \( V_i^{\text{input}} \) for each \( i^* \in A \) and adjust the input value as \( V_i^{\text{input}} \) |
| 8    | For each \( i^* \in A \), calculate \( V_i^{\text{input}} \) in the same manner as in Step 6, and adjust the input value as \( V_i^{\text{input}} \) |
Recall that the tasks for solving each quality-related problem that we defined are related with one another, and thus, the precedence relationships of the tasks should be considered when the data-analytics-driven operation strategies are implemented. Figure 3 presents the sequence of implementation of the tasks of the proposed data-analytics-driven factory operation strategies. Before implementing the tasks, the data that will be utilized for the said strategies should be collected and integrated. Then, [T1] should be conducted, and the statistical information essential for data analytics, such as the correlation between the data, should be deduced. The results of [T1] are utilized in the following stages and for deriving useful information for the quality enhancement of die casting. The next stages consist of two parallel sequential steps: one consisting of [T2] and [T3] and dealing with [P1] and the other consisting of [T4], [T5], and [T6] and dealing with [P2] and [P3]. Between the two sequential steps, the former is recommended to be conducted first as the results of [T3] can be used when [T6] is conducted. After [T2] is conducted, [T3] is carried out using the optimal casting parameter output value deduced from [T2]. For the latter sequential steps, [T4] should be conducted prior to [T5] because preheat shots are regarded as either defective or of fair quality and are managed separately. This means that preheat shots should be eliminated and not considered when [T5] is conducted. When a defect prediction algorithm is developed, [T6], which infers the specific causes of the defect, is carried out using the developed defect prediction algorithm or the upper and lower control limits of the casting parameters. After the two parallel stages are completed, [T7] is carried out, triggered from the results of [T3] and [T6].

4 Case study

In this section, we present a case of implementation of the proposed data-analytics-driven factory operation strategies to a real die-casting factory in South Korea. We first briefly introduce the target factory in Sect. 4.1 and then describe the results of the application of the strategies to the target factory in Sect. 4.2.

4.1 Target factory

For the target factory of our case study, we selected a typical die-casting factory in South Korea that carries out the general processes of die casting: casting, fitting, blasting, machining, and inspection. The target factory’s average production cycle time (from casting to inspection) is about a week, and the factory produces 10 million products every year. All the quality-related problems defined herein existed prominently in the target factory, so it was
properly to implement the proposed data-analytics-driven factory operation strategies in the target factory, and to verify their effectiveness with such an application case. In the target factory, about 20 types of defects are managed. Among these, we selected the two casting-process-related defects of short shot and pores, which commonly occur and are generally tracked in die-casting factories. In short shot, a product is not fully casted. As for pores, when air or impurities permeate a product during the casting process, pores appear in the casted product. Both defects can be easily detected by inspecting appearance and surface of the products with the naked eye.

In the target factory, 14 Toyo die-casting machines are operated, and infrastructure for collecting casting parameter and sensor data had been constructed when the study began. The casting parameter and sensor data collected from the target factory are shown in Table 7. The count number of a shot, the time stamp, the casting parameter values at the time of casting of a shot, and the cycle time were utilized. Injection velocity is the average speed at which molten metal is forced into a mold, and high speed velocity is the highest speed of casting a shot. Pressure-related parameters like the physical strength of the casting machine and the biscuit thickness, cylinder pressure, casting pressure, and casting

| Data category      | Data list                      |
|--------------------|-------------------------------|
| Casting parameter data | Shot number, Injection velocity, High speed velocity |
|                    | Physical strength, Biscuit thickness, Cylinder pressure |
|                    | Casting pressure, Pressure increase time, Spray time 1 |
|                    | Spray time 2, Spray time 3, Cycle time |
| Sensor data        | Machine, Heating furnace temperature, Coolant temperature, Air pressure |
|                    | Factory, Factory temperature, Factory humidity |

Table 7 Target factory data used in this study
pressure increase time were also utilized. After the injected molten metal is pressed with the cylinder in a mold, vapor and air steam are sprayed onto the shot from the nozzles of the casting machine. Although the number of nozzles depends on the type of machine, we considered the data with three spray nozzles in this study. Among the sensor-related data, we selected the parameters that are important to manage and that can be collected with less effort, without having to construct a data collection infrastructure. We utilized the heating furnace temperature, coolant temperature, and air pressure as the casting-machine-related sensor data, and the factory temperature and humidity as the factory-related sensor data. Of course, more varied types of data can be utilized for data analytics, but we focused on the general data that most die-casting factories manage and that can be collected with relatively less effort in our case study.

4.2 Results of the implementation of the proposed factory operation strategies

4.2.1 Implementation of the proposed strategy for solving [P1]

As the volume of EDA results is too large to present in this paper, we will present only the general EDA results. For the collected casting parameter and sensor data, we first tried to derive some insights from statistical values like the average, standard deviation, minimum, and maximum values of each group of data divided by the time periods. The statistical values were observed to be different according to the shift (day and night) and the season (spring, from March to May; summer, from June to August; autumn, from September to November; and winter, from December to February), so we carried out the proposed data-analytics-driven factory operation strategies for solving [P1]–[P4] for each time division. The case presented in this section is for the day shift of the spring season. Compared with the statistical values of the fair-quality products, the cylinder pressure and casting pressure of the short-shot products tended to be lower whereas the physical strength was observed to be generally higher. Whenever the pore defect type occurred, the injection velocity and high speed velocity were higher than those of the fair-quality products. Among the sensor data, the factory’s high temperature and low humidity were observed when defects occurred in general. Through the statistical values of the casting parameters and sensor data, identifying the tendencies of the main factors that affect the occurrence of each defect and deciding the main factors that should be managed were enabled. Based on the statistical values that we calculated, we set the average value of each casting parameter data for fair-quality products as the optimal output casting parameter, and derived the lower and upper control limits of each data using IQR.

We also conducted statistical analysis of the gap between the input and output values of the casting parameter data. We calculated the average absolute gap, standard deviation of the gap, maximum gap, and minimum gap between the input and output values of the casting parameter data, and the results are presented in Table 8. The average gap and standard deviation of the gap between the input and output values of the cylinder pressure, physical strength, and casting pressure were observed to be high, whereas those of the other parameters were observed to be relatively stable. Based on the results shown in Table 8, we concluded that we should manage mainly the gaps between the input and output values of the cylinder pressure, physical strength, and casting pressure.

Table 9 presents some representative correlations between the target factory’s casting parameter and sensor data. The red and blue colors indicate positive and negative correlations, respectively, and the deeper the color, the stronger the correlation. It can be seen that cylinder pressure has strong positive correlations with casting pressure and spray time 1 and has a negative correlation with heating furnace temperature. The fact that factory temperature affects cylinder pressure, casting pressure, spray time 1, and heating furnace temperature can also be seen in Table 9. It can thus be said

| Statistical value | Injection velocity | High speed velocity | Cylinder pressure | Biscuit thickness | Physical strength |
|-------------------|--------------------|---------------------|------------------|------------------|------------------|
| Average           | 0.002              | 0.005               | 1.752            | 0.006            | 12.678           |
| Std               | 0.012              | 0.0214              | 2.441            | 0.019            | 7.845            |
| Min               | -0.071             | -0.079              | -3.912           | -3.245           | -40.678          |
| Max               | 0.073              | 0.041               | 4.147            | 2.175            | 51.221           |

| Statistical value | Pressure increase time | Casting pressure | Spray time 1 | Spray time 2 | Spray time 3 |
|-------------------|------------------------|------------------|-------------|-------------|-------------|
| Average           | 0.002                  | 7.412            | 0.897       | 0.018       | 0.012       |
| Std               | 0.008                  | 6.576            | 3.427       | 0.214       | 0.027       |
| Min               | -0.781                 | -31.547          | -2.174      | -1.224      | -0.487      |
| Max               | 0.671                  | 51.133           | 2.397       | 0.576       | 0.553       |
that the highlighted pairs of data should be considered and managed together when factory administrators try to make decisions based on the data.

By implementing the proposed strategy for solving [P1], efficient monitoring of casting parameters and sensor data was enabled for the target factory. Whenever the data gets out of upper or lower control limit, the monitoring system alarms workers, so that it becomes possible to maintain the casting parameters and sensor data within the proper range. As a result, overall defect rate of the target factory decreases from 7.84% to 5.23%.

### Table 9 Correlations between the casting parameter and sensor data

|                      | Injection velocity | High speed velocity | Cylinder pressure | Casting pressure | Spray time 1 | Heating furnace temperature | Factory temperature | Factory humidity |
|----------------------|--------------------|---------------------|------------------|-----------------|--------------|----------------------------|---------------------|----------------
| Injection velocity   | 1.0000             | -0.3965             | 0.3935           | 0.3938          | 0.1425       | -0.2996                    | -0.1593             | -0.1587        |
| High speed velocity  | -0.3965            | 1.0000              | 0.2652           | 0.2650          | 0.0234       | 0.1024                     | 0.0356              | 0.0944         |
| Cylinder pressure    | 0.3935             | 0.2652              | 1.0000           | 0.9995          | 0.6514       | -0.4719                    | -0.3056             | -0.1721        |
| Casting pressure     | 0.3938             | 0.2650              | 0.9995           | 1.0000          | 0.6515       | -0.4719                    | -0.3055             | -0.1723        |
| Spray time 1         | 0.1425             | 0.0234              | 0.6514           | 0.6515          | 1.0000       | -0.4681                    | -0.3609             | -0.0546        |
| Heating furnace temp | -0.2996            | 0.1024              | -0.4719          | -0.4719         | -0.4681      | 1.0000                     | 0.6051              | 0.0700         |
| Factory temperature  | -0.1593            | 0.0356              | -0.3056          | -0.3055         | -0.3609      | 0.6051                     | 1.0000              | -0.5909        |
| Factory humidity     | -0.1587            | 0.0944              | -0.1721          | -0.1723         | -0.0546      | 0.0700                     | -0.5909             | 1.0000         |

*Red – positive correlation; blue – negative correlation

### 4.2.2 Implementation of the proposed strategy for solving [P2]

Before developing a preheat shot diagnosis algorithm, we conducted two preprocessing processes: PCA and SMOTE. For carrying out PCA, we utilized the casting parameter and sensor data of the target factory and selected the set of data that we would utilize for developing a data-analytics-based algorithm. Table 10 shows the proportion of the weight of the coefficient vector of each casting parameter and sensor data. We selected factory humidity, factory temperature,

|                      | Factory humidity | Factory temperature | Heating furnace temperature | Casting pressure | Physical strength | High speed velocity | Cylinder pressure | Injection velocity |
|----------------------|------------------|---------------------|----------------------------|-----------------|------------------|--------------------|-------------------|-------------------|
| Weight of coefficient vector | 0.1602           | 0.1245              | 0.1131                     | 0.1112          | 0.1107           | 0.0945             | 0.0737            | 0.0518            |
| Accumulated proportion | 0.1602           | 0.2847              | 0.3978                     | 0.509           | 0.6197           | 0.7142             | 0.7879            | 0.8397            |
| Coolant temperature  | 0.0321           | 0.0301              | 0.0291                     | 0.0218          | 0.0187           | 0.0113             | 0.0091            | 0.0081            |
| Air pressure         | 0.8718           | 0.9019              | 0.931                      | 0.9528          | 0.9715           | 0.9828             | 0.9919            | 1                 |
| Spray time 1         |                  |                     | 0.931                      | 0.9528          | 0.9715           | 0.9828             | 0.9919            | 1                 |
| Spray time 2         |                  |                     |                            |                 |                  |                    |                   |                   |
| Pressure increase time |                 |                     |                            |                 |                  |                    |                   |                   |
| Cycle time           |                  |                     |                            |                 |                  |                    |                   |                   |
heating furnace temperature, casting pressure, physical strength, high speed velocity, cylinder pressure, injection velocity, coolant temperature, and air pressure, whose corresponding accumulated weight proportions were over 0.9, as the main factors of the data analytics model that we developed. In the case of the target factory, the initial ratios of the number of normal shots to the number of preheat shots were 9.9:0.1, so we conducted SMOTE with $k=5$ until the ratio became 6.5:3.5. The final ratio was determined through repetitive numerical experiments as the value corresponding to the developed algorithm’s best performance, avoiding the overfitting issue.

After carrying out data preprocessing, we developed preheat shot diagnosis algorithms using the random forest, SVM, neural network, AdaBoost, and XGBoost methods, and conducted fivefold cross-validation of such algorithms. The performance of each machine-learning-based algorithm is presented in Table 11. As the MAPE of the XGBoost-based preheat shot diagnosis algorithm outperformed those of the other algorithms, we adopted the XGBoost-based algorithm. We also figured out that the accuracy of the preheat shot diagnosis algorithm increases when SMOTE is applied.

For the further analysis, we derived detailed accuracies of the XGBoost algorithm trained with each data set comprised of monthly data as shown in Table 12. The accuracy tended to decrease when temperature of factory extremely increases or decreases in summer or winter. Particularly, the accuracy of the preheat diagnosis algorithm was the lowest with 85.3% and 87.2% for each case of not applying SMOTE and applying SMOTE, respectively, in August. It means that as the environment of the factory becomes extreme condition, unexpected variables that affect the machine status occur, and thus, performance of the developed algorithm decreases. Moreover, performance gap between the cases that applied SMOTE and not was also larger when the temperature of the factory was low or high in winter and summer. The fact that temperature and humidity of the factory affect the other casting parameter and sensor data as shown in Sect. 4.2.1 can also explain the results.

For the developed XGBoost-based algorithm, we also calculated the feature importance. As shown in Table 13, the feature importance of coolant temperature, factory humidity, injection velocity, and high speed velocity was relatively lower than that of the other data. As the feature importance indicates the relative impact of each data on the developed algorithm, we compared the performances of the data analytics models considering all the data and considering the top 6 data. As shown in Table 14, the validation accuracy of the preheat shot diagnosis algorithm considering only the top 6 data was slightly lower than that of the original algorithm, but it still showed good performance. Thus, we applied both algorithms to the field and compared their field accuracy levels. Based on the results shown in Table 14, the field accuracy of the algorithm that considered only the top 6 data was generally higher than that of the algorithm that considered all the data. This shows that when the developed data-analytics-based algorithm is applied in the field, several highly influential variables investigated by feature importance should be considered.

By implementing the proposed strategy for solving [P2], the target factory could distinguish preheat shot more precisely. Before adopting the strategy, the target factory eliminated 10 shots of the product as preheat shots whenever reoperating the die-casting machine. However, it was realized that preheat shots occur less than 10 shots in summer and more than 10 shots in winter by conducting the case study. By adopting the preheat shot diagnosis algorithm, unnecessary elimination of normal products decreased in summer. Moreover, by eliminating preheat shots previously diagnosed as the normal, overall quality of the target factory could be enhanced.

| Table 11 | Accuracy levels of the developed preheat shot diagnosis algorithms |
|----------|-------------------------------------------------------------|
| Data model | Accuracy (%) | Without SMOTE application | With SMOTE application |
| Decision tree | 71.4 | 76.9 |
| Random forest | 89.5 | 90.1 |
| Neural network | 83.6 | 85.2 |
| Support vector machine | 86.1 | 87.9 |
| AdaBoost | 88.7 | 90.1 |
| XGBoost | 91.4 | 92.5 |

| Table 12 | Accuracy of the preheat shot diagnosis algorithm for each data set |
|----------|-----------------------------------------------------------------|
| Data Set | Accuracy (%) | Without SMOTE application | With SMOTE application |
| Jan | 86.2 | 88.2 |
| Feb | 87.9 | 90.1 |
| Mar | 93.2 | 93.9 |
| Apr | 95.3 | 96.0 |
| May | 94.9 | 95.7 |
| Jun | 92.7 | 93.1 |
| Jul | 87.1 | 88.1 |
| Aug | 85.3 | 87.2 |
| Sep | 91.7 | 92.3 |
| Oct | 94.8 | 95.7 |
| Nov | 95.6 | 96.4 |
| Dec | 91.8 | 92.7 |
Before we developed a defect prediction algorithm, we also conducted the data preprocessing processes of PCA and SMOTE, as with the strategy for solving [P2]. The PCA result obtained is the same as that of the PCA conducted in the strategy for solving [P2] because the same independent variables consisting of casting parameter and sensor data were considered. As it is possible for the target factory to trace the defects in a product, we developed a defect prediction algorithm that predicts the quality classification of each casted shot. Utilizing the 10 variables of cylinder pressure, air pressure, physical strength, casting pressure, heating furnace temperature, factory temperature, coolant temperature, factory humidity, injection velocity, and high speed velocity selected from the PCA results, we generated data-analytics-based algorithms for predicting defects with various machine learning methods (i.e., decision tree, random forest, neural network, SVM, AdaBoost, and XGBoost). Table 15 shows the performance of each machine learning method under fivefold cross-validation. Among the methods, the accuracy of XGBoost was observed to be the highest, so we adopted the XGBoost method as the main machine learning method of our defect prediction algorithm.

Moreover, based on the results shown in Table 15, we determined that the accuracy of the data-analytics-based algorithms when SMOTE was applied was higher than that when SMOTE was not applied. Based on the comparison of the accuracy levels of the algorithms when SMOTE was applied and not applied in both the preheat shot diagnosis and defect prediction algorithms, it can be said that the application of SMOTE can elevate the performance of the data-analytics-based algorithms in die-casting factories.

Table 16 shows the detailed accuracy of the XGBoost-based defect prediction algorithm for each monthly data set. The accuracy of the defect prediction algorithm was the worst in August and was the best in April. Moreover, the accuracy was relatively low when it was winter or summer. This is because extreme temperature and humidity of the factory affects the defect, so that unexpected variables that cause the defects occur.

Table 17 presents the feature importance value of each data for the XGBoost-based defect prediction algorithm, and Table 18 shows a performance comparison for a case that considers parts of the dataset that has high feature importance and a case that considers the whole dataset. As can be seen in Table 17, unlike in the case of developing a preheat diagnosis algorithm, the feature importance values of all the data seem similar. In this case, considering defect prediction algorithms developed using parts of the dataset with the highest feature importance may lower the performance because all the data are equally important. The results presented in Table 18 show that when only the top 5 data are considered, both the validation accuracy and field accuracy of the defect prediction algorithm significantly decrease compared with the original algorithm. These results demonstrate that it is useless to apply the defect prediction algorithm considering only the data with high feature importance.

As we adopted the XGBoost-based algorithm, a tree-based algorithm, for the target factory, we could derive the defect causes using the branch points of the trees included in the majority set that classifies a shot as defective. Figure 4 shows an example of the trees in the majority set that predicts the

| Table 13 Feature importance values of the preheat shot diagnosis algorithm |
|-----------------------------|-------------------------------|
| Data                        | Cylinder pressure | Air pressure | Physical strength | Casting pressure | Heating furnace temperature |
| Feature importance Data     | 0.217             | 0.198        | 0.187             | 0.161           | 0.121                       |
| Feature importance          | 0.087             | 0.012        | 0.009             | 0.006           | 0.002                       |

| Table 14 Accuracy of the preheat diagnosis algorithm utilizing selective data |
|-----------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| Data model                  | Validation accuracy (%)       | Field accuracy (%)           |                               |                               |
|                             | Considering all the data       | Considering the top 6 data    | Considering all the data       | Considering the top 6 data    |
| Decision tree               | 76.9                          | 74.8                         | 69.7                          | 71.1                          |
| Random forest               | 90.1                          | 88.3                         | 84.4                          | 88.9                          |
| Neural network              | 85.2                          | 82.1                         | 71.3                          | 77.1                          |
| Support vector machine      | 87.9                          | 83.9                         | 83.9                          | 82.4                          |
| AdaBoost                    | 90.1                          | 87.4                         | 86.5                          | 88.7                          |
| XGBoost                     | 92.5                          | 90.7                         | 87.3                          | 92.1                          |
quality of the target factory’s die casting. When a defect was predicted, we extracted the trees that we needed and investigated the lowest branches of the trees. Then we treated the intersection conditions for each casting parameter and sensor data of the branches as the lower-bound conditions, and the union conditions as the upper-bound conditions.

By implementing the proposed strategy for solving [P3], defective product could be efficiently eliminated right after the casting process was conducted. It removed the entrance of defective WIP to the post processes, so that the productivity of the target factory increased. Compared with before the strategy was applied to the target factory, productivity increased about 1.8%. Moreover, overall defect rate detected in the final inspection process decreased from 7.84% to 5.23%.

4.2.4 Implementation of the proposed strategy for solving [P4]

To implement our proposed data-analytics-based factory operation strategy for solving [P4], we first developed regression models of each casting parameter data, except for the shot number, which indicates the ID of each shot. As mentioned in Sect. 3.2.4, the independent variables used can affect the performance of the regression model, so we developed regression models using all the casting parameter and sensor data and the datasets consisting of the data with over 0.2, 0.3, 0.4, and 0.5 correlations with the target casting parameter data. The accuracy levels of the regression models for each case were observed to be 0.71, 0.78, 0.87, 0.71, and 0.62, respectively, so we utilized the regression models developed with the casting parameter and sensor data whose correlations with the target casting parameter data were over 0.3. Then, to implement the casting parameter tuning algorithm illustrated in Sect. 3.2.4, we set the values of the constants needed for executing the algorithm as shown below.

\[ t_{\text{constant}} = 10 \text{ minutes} \]

\[ C_{\text{gap allowance}} = 7\% \]

\[ C_{\text{limit allowance}} = 10 \text{ times} \]

\[ C_{\text{criteria allowance}} = 10 \text{ times} \]

We also utilized the data regarding the control limits and the defect cause criteria with the results of the proposed strategies for solving [P1] and [P3], respectively.

By implementing the proposed strategy for solving [P3], the target factory was enabled to adjust input values of casting parameter and sensor data properly according to the factory environment. Compared with the past that the workers adjusted the input values based on their experience, more accurate adjustment was enabled, and thus, the overall quality of the target factory could be elevated.

### 4.2.5 Implementation of the integrated system

As the proposed strategies for each of the quality-related problems defined herein are to be implemented in die-casting factories, the results of the implementation of each strategy (e.g., control limit values, data-analytics-based algorithms, and an expert system) were derived. This section presents how the results of each strategy’s implementation can be applied and implemented as a consolidated architecture in the field. The results can be integrated as a system for implementing the data analytics results to the field, as shown in Fig. 5. When real-time data are collected, the collected data are first pre-processed into a form of structure that is suitable for utilization in the implementation system. Using the preprocessed data, upper and lower control limits are continuously derived and updated. The real-time data are monitored referring to the derived control limits, and when the real-time data deviate
from the control limits, an alarm is sounded. Every time a shot is casted, the preheat shot diagnosis algorithm is implemented, and when a shot is diagnosed as a preheat shot, it is discarded. When a shot is diagnosed as a normal shot, the defect prediction algorithm is implemented. When the shot is predicted as defective, it is discarded, and the defect causes are diagnosed. Then, the factory administrators are notified of the diagnosed defect causes. When an alarm is sounded as a real-time data deviates from the control limits or as a defect occurs, an expert system decides whether to adjust the input value of the casting parameter that resulted in the problem or not to. When the system decides to adjust the input value of the casting parameter, the extent of adjustment of the output value is determined according to the factory’s circumstances, and the input values of the target and related casting parameters are calculated using the regression equations. Finally, the input values of the corresponding casting parameters are adjusted.

4.3 Summary of the proposed strategies

A summary of the strategies proposed for addressing each of the quality-related problems defined herein is presented in Table 19. As data analytics needs a certain degree of factory infrastructure related to data collection and system control, it may be hard for all die-casting factories to fully follow the guidelines of the proposed data-analytics-driven factory operation strategies. We thus recommend that die-casting factories conduct the strategies by stage. For the first stage, the strategy for solving [P1] can be implemented. Although the real-time data monitoring system we propose needs a real-time data collection and visualization infrastructure, the other tasks of the strategy can be carried out only by collecting data. As the strategy needs statistical analysis with a relatively low extent of data analytics advancement, factories that do not have experience in data analytics can follow the strategy more easily than the other strategies. Using the accumulated casting parameter and sensor input and output data, the optimal parameter value, the upper and lower control limits of each data, the correlations between the data, and the statistical values of the data can be derived with statistical methods. The results can help the factories understand their own operations based on the collected data, and can enable them to determine the proper casting parameter or sensor data settings. Moreover, the factory’s statistical information can provide some insights regarding the proper factory operation, such as when the factory temperature is high, the casting pressure input value should be set lower than the expected value. It is true that some workers have some knowledge of the intricacies of die-casting factory operation due to their extensive relevant experience, but organizing and verifying their knowledge with the specific data obtained is also important. When the related knowledge is documented, it can also support the workers with lower experience. If a real-time automated data collection infrastructure is constructed, a real-time data monitoring system containing the factory’s statistical information can be developed, and the system can enhance the factory’s product quality.

The next stages are strategies for solving [P2] and [P3]. The strategies focus on solving the problem by utilizing the collected data. Using the casting parameter data, sensor data, and classification of each product, a preheat shot diagnosis algorithm, a defect prediction algorithm, and a defect cause

| Table 17 Feature importance values of the defect prediction algorithm |
|------------------------|---------------------|---------------------|---------------------|---------------------|
| Data                   | Cylinder pressure   | Air pressure        | Physical strength   | Casting pressure    | Heating furnace     |
| Feature importance     | 0.115               | 0.112               | 0.107               | 0.103               | 0.098               |
| Data                   | Factory temperature | Coolant temperature | Factory humidity    | Injection velocity  | High speed velocity |
| Feature importance     | 0.096               | 0.094               | 0.092               | 0.092               | 0.091               |

| Table 18 Accuracy levels of the defect prediction algorithm utilizing selective data |
|------------------------|---------------------|---------------------|---------------------|---------------------|
| Data model             | Validation accuracy (%) | Field accuracy (%) |
|                        | Considering all the data | Considering the top 6 data | Considering all the data | Considering the top 6 data |
| Decision tree          | 74.7                 | 49.8                | 67.3                 | 45.2                |
| Random forest          | 89.4                 | 52.4                | 77.2                 | 47.6                |
| Neural network         | 82.1                 | 44.3                | 76.3                 | 43.1                |
| Support vector machine | 84.4                 | 47.9                | 77.5                 | 42.7                |
| AdaBoost               | 92.3                 | 51.6                | 85.2                 | 49.4                |
| XGBoost                | 96.3                 | 54.8                | 86.1                 | 50.6                |
diagnosis algorithm were developed. These strategies need more advanced data analytics skill and infrastructure than the first strategy. The purposes of the strategies are diagnosing the machine status and predicting the quality of the casted products in real time. Hence, a real-time data collection infrastructure and a data analytics and implementation infrastructure should be constructed. For the data preprocessing method, outlier elimination, missing-value treatment, PCA, and SMOTE can be utilized. SMOTE is an important data preprocessing method because the number of preheat shot and defect data is much lower than the number of normal data. The adoption of SMOTE can increase the performance of the proposed data-analytics-based algorithms. For the data analytics method that was used, XGBoost showed the best performance based on

---

**Fig. 4** Example of the trees included in the majority set

**Fig. 5** System architecture of the system implementation
the numerical experiment results that we obtained. When the proposed strategies are conducted in die-casting factories, the decision-making processes regarding the direct quality-related problem of the product, which are necessary in factory operation, can become more accurate.

The last stage is the strategy for solving [P4]. The strategy utilizes the information derived from the other strategies and infers the knowledge needed for tuning the casting parameters from such information. Using the casting parameter and sensor data, regression models that can infer the expected values of each data according to the factory circumstances are derived. When a casting parameter tuning trigger is detected, the extent of casting parameter adjustment needed is determined, and the input value of the corresponding data is changed. The realization of the last strategy to construct an automated real-time casting parameter tuning system was the final goal of this study. This strategy is more difficult to implement than the other strategies as it needs high data analytics advancement and complete pre-construction of the other strategies.

5 Conclusions
This paper proposed data-analytics-based factory operation strategies for quality enhancement of die casting. We first defined the four main problems of die-casting factories that result in a decrease in quality: [P1] a gap between the casting parameter input and output values, [P2] occurrence of preheat shots, [P3] lateness of defect distinction, and [P4] worker-experience-based casting parameter tuning. Then, for each strategy, we derived the following seven tasks: [T1] implementation of EDA for investigating the trends and correlations between the data, [T2] deduction of the optimal casting parameter output values for the production of fair-quality products, [T3] deduction of the upper and lower control limits for casting parameter input–output gap management, [T4] development of a preheat shot diagnosis algorithm, [T5] development of a defect prediction algorithm, [T6] development of a defect cause diagnosis algorithm, and [T7] development of a casting parameter tuning algorithm. The detailed contents of the proposed data-analytics-driven factory operation strategies in terms of input/output data, data preprocessing, data analytics method used, and implementation are presented. We also conducted a case study of a die-casting factory in South Korea that has adopted the proposed strategies.

The main contribution of this work is that it provides guidelines for data-analytics-based factory operation strategies for quality enhancement that can be utilized by general factories in the die-casting industry and presents a case involving a real die-casting factory’s implementation of such strategies. Following the provided guidelines, factories can address their quality-related issues. Of course, there are methods other than those introduced herein that can be utilized, but applying other specific methods of data...
preprocessing or machine learning is simply a matter of extending the basic structure of the strategies we introduced herein. For the future works, we will try to consider the concept of transfer learning, a research problem in machine learning that focuses on storing the knowledge gained while solving one problem and applying it to a different but related problem. When there are already enough die-casting factory data analytics cases, we can use the transfer learning concept in developing data-analytics-based algorithms by referring to such cases. The application of the transfer learning concept can make it easier for die-casting factories to implement the data-analytics-based factory operation strategies that we proposed.

Author contribution Jun Kim collected and analyzed quality relevant data of die-casting processes, established factory operation strategies, designed the structure of this paper, and wrote the manuscript. Ju Yeon Lee supervised the entire process of data analysis, established factory operation strategies, examined the structure of the paper, and made suggestions on the details of the paper.

Funding This work was supported by the Ministry of Economy and Finance under the Manufacturing Innovation Support Project (PJH21010) and the Ministry of Trade, Industry & Energy (MOTIE, Korea) under the Industrial Technology Innovation Program (20003661).

Data availability The datasets used or analyzed during the current study are available from the corresponding author on reasonable request. However, data belonging to the company cannot be provided even if requested.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Park BC (2013) On the policy for strengthening the global competitiveness of the fundamental manufacturing industry. J Region Stud 21(4):1–19
2. Committee on Human Resources Development in Root Industry (2018) Report on survey and analysis of manpower status in root industry
3. Scampone G, Pirovano R, Mascetti S, Timelli G (2021) Experimental and numerical investigations of oxide-related defects in Al alloy gravity die castings. Int J Adv Manuf Technol 1–16
4. Singh R, Madan J (2019) A computer-aided system for multi-gate gating-system design for die-casting dies. Int J Adv Manuf Technol 101(5):1793–1806
5. Dong Y, Su F, Zhu G, Yang G (2019) Study and implementation of a variable parameter modeling for die casting mold generation. Int J Adv Manuf Technol 105(7):3543–3561
6. Zhao X, Meng D, Zhang J, Han Q (2020) The effect of heat treatment on die casting aluminum to apply self-pierce riveting. Int J of Adv Manuf Technol 109(9):2409–2419
7. Andrade-Gutiérrez ES, Carranza-Bernal SY, Hernandez-Sandoval J, Gonzalez-Villarreal AJ, Berber-Solano TP (2018) Optimization in a flexible die-casting engine-head plant via discrete event simulation. Int J Adv Manuf Technol 95(9–12):4459–4468
8. Cao L, Liao D, Sun F, Chen T (2018) Numerical simulation of cold-lap defects during casting filling process. Int J Adv Manuf Technol 97:2419–2430
9. Ratna S, Prasad D (2014) Application of artificial neural network for optimization of cold chamber aluminium die casting. Int J Innov Res Adv Eng 1(5):43–47
10. Fitiriana R, Saragih J, Sarasaty S (2014) Implementation of six sigma and data mining to improve die casting production process at PT. AB. Proc 7th Int Seminar Industr Eng Manage 64–71
11. Morgado JPGB (2015) Knowledge elicitation by merging heterogeneous data sources in a die-casting process. Univ Coimbra
12. Winkler M, Kallien L, Feyertag T (2015) Correlation between process parameters and quality characteristics in aluminum high pressure die casting. NADCA Die Casting Congress & Exposition
13. Haghhighi F, Ramnath S, Kalish N, Shah JV, Shah JJ, Davidson JK (2016) Method for automating digital fixture-setups that are optimal for machining castings to minimize scrap. J Manuf Syst 40:15–24
14. Soban D, Thornhill D, Salunkhe S, Long A (2016) Visual analytics as an enabler for manufacturing process decision-making. Procedia CIRP 56:209–214
15. Kuttir JK, Patel GM, Parappagoudar MB (2016) Modeling of pressure die casting process: an artificial intelligence approach. Int J Metal Cast 10(1):70–87
16. Liu W, Tang R, Peng T (2018) An IoT-enabled approach for energy monitoring and analysis of die casting machines. Procedia CIRP 69:656–661
17. Kim A, Oh K, Park H, Jung JY (2017) Comparison of quality prediction algorithms in manufacturing process. ICIC Exp Lett 11(6):1127–1132
18. Cashion K, Powar N, De Neff R (2018) Kress R (2018) Part quality assessment using convolution neural networks in high pressure die casting. Electron Imag 9:2771–2776
19. Gellrich S, Beganovic T, Mattheus A, Herrmann C, Thiede S (2019) Feature selection based on visual analytics for quality prediction in aluminium die casting. 2019 IEEE 17th Int Conf Industr Inform (INDIN) 66–72
20. Kozlowski J, Jakimiuk M, Rogalewicz M, Sika R, Hajkowski J (2019) Analysis and control of high-pressure die-casting process parameters with use of data mining tools. Int Sci Tech Conf Manuf 253–267
21. Weiderer P, Tomé AM, Lang EW (2020) A NMF-based extraction of physically meaningful components from sensory data of metal casting processes. J Manuf Syst 54:62–73
22. Kim JS, Kim J, Lee JY (2020) Die-casting defect prediction and diagnosis system using process condition data. Proc Manuf 51:359–364
23. Wu Z, Wu Y, Zhang E (2020) Study on die casting speed control based on neural network. In: 2020 International Conference on Communications, Information System and Computer Engineering (CISCE) 407–410
24. Jianmin TANG, Laijie WU, Haidun LIU, Bo LI, Dejun ZENG (2021) Quality control study on high pressure casting based on magma and ultra-red simulation. J Phys: Conf Series 1986(1):012130
25. Elser A, Lechler A (2021) Implementation of a model based numerical control for the gravity die casting process. In 2021 22nd IEEE International Conference on Industrial Technology (ICIT) 1:1094–1099
26. Hailin L, Jinhong W, Yong L (2021) Intelligent control system design for high vacuum die casting process. In 2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) 300–303
27. Haokai H, Peijie L (2016) Design of remote monitor system for die casting units. 2016 2nd International Conference on Control Science and Systems Engineering (ICCSSE) 252–255
28. Meisen T, Rix M, Hoffmann M, Schilberg D, Jeschke S (2016) A framework for semantic integration and analysis of measurement data in modern industrial machinery. Autom Commun Cybernet Sci Eng 2015(2016):893–905
29. Lee JY, Yoon JS, Kim BH (2017) A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: An empirical case study of a die casting factory. Int J Precision Eng Manuf 18(10):1353–1361
30. Zhao Y, Qian F, Gao Y (2018) Data driven die casting smart factory solution. Recent Adv Intell Manuf 2018:13–21
31. Park S, Kim C, Youn S (2019) Establishment of an IoT-based smart factory and data analysis model for the quality management of SMEs die-casting companies in Korea. Int J Distrib Sens Netw 15(10):1–21
32. Vanli AS, Akdogan A, Durakbasa MN (2019) Tools of industry 4.0 on die casting production systems. Proc Int Symp Prod Res 2019:328–334
33. Kim J, Kang HS, Lee JY (2020) Development of intelligence data analytics system for quality enhancement of die-casting process. J Kor Soc Precision Eng 37(4):247–254
34. Kim J, Lee JY (2021) Server-edge dualized closed-loop data analytics system for cyber-physical system application. Robot Comput Integr Manuf 67:102040

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.