COMPUTER METHODS IN MATERIALS SCIENCE

2023, vol. 23, no. 2, 45–56
https://doi.org/10.7494/cmms.2023.2.0809

Machine learning methods for diagnosing the causes of die-casting defects

Alicja Okuniewska*, Marcin Perzyk, Jacek Kozłowski

Warsaw University of Technology, Faculty of Mechanical and Industrial Engineering, Institute of Manufacturing Technologies, ul. Narbutta 85, 02-524 Warsaw, Poland.

Abstract

The research was focused on analyzing the causes of high-pressure die-casting defects, more specifically on casting leakage, which is considered perhaps the most important and common defect. The real data used for modelling was obtained from a high-pressure die-casting foundry that manufactures aluminum cylinder blocks for the world's leading automotive brands. This paper compares and summarizes the results of applying advanced modelling using artificial neural networks, regression trees, and support vector machines methods to select artificial neural networks as the most effective method to perform a multi-dimensional optimization of process parameters to diagnose the causes of die-casting defects and to indicate the future research scope in this area. The developed system enables the prediction of the level of defects in castings with satisfactory accuracy and is therefore a highly relevant reference for process engineers of high-pressure foundries. This article indicates exactly which process parameters significantly influence the formation of a defect in a casting.

Keywords: fault diagnosis, machine learning tools, neural network, classification trees, support vector machine

1. Introduction

Metal casting, which is a part of the metallurgy sector, is one of the most popular manufacturing processes (Miłek, 2017). The largest application segment for castings is the automotive industry (Grand View Research, 2019). It is expected that due to the continuous work on using aluminum for weight reduction of vehicles to increase their energy efficiency, the demand for castings from the automotive sector will increase continuously (Grand View Research, 2019). In addition, aluminum castings are becoming increasingly popular due to environmental trends and, more specifically, through their recyclability. The creation and use of recyclable materials reduce the environmental impact of an activity. Therefore, special attention should be paid to the development of the foundry industry. This development should be based on state-of-the-art technologies, using smart specialization to increase the competitiveness and innovativeness of the sector (Grand View Research, 2019), in order to meet the growing demand and requirements of customers and care about the environment.

Referring to the foundations of the Industry 4.0 concept, it can be concluded that data today is one of the most valuable raw materials in the industry. The main goal is to extract information, knowledge, and wisdom from the acquired process data (Grzegorzewski & Kochański, 2019a). Ultimately, the Smart Factories (Xu et al., 2018) of the future will integrate

*Corresponding author: alicja.okuniewska.dokt@pw.edu.pl
ORCID ID's: 0000-0001-9635-8557 (A. Okuniewska), 0000-0003-4901-7746 (M. Perzyk), 0000-0002-5297-1509 (J. Kozłowski)

the physical and digital worlds and be able to effectively and independently (Wang et al., 2015), control the process in order to diagnose the causes of product defects. Referring to the level of sophistication of foundry processes and their ability to collect process data, according to the concept of Industry 4.0, it is important to mention that the die-casting process presents one of the highest levels (Perzyk et al., 2019). This also refers to the definition of "Quality 4.0" (Jacob, 2017; Raluca, 2021), which supports the identification of the relations between the quality management of manufactured products and the concept of Industry 4.0. It seeks the overall digitalization of the quality management process in manufacturing companies through artificial intelligence (AI) techniques, which are machine learning (ML) methods (Bowers & Pickerel, 2019). The usefulness of using machine learning techniques has been highlighted (see Khan et al., 2022), among other reasons, because of the need to use the right software architecture to process large data sets.

The topic of analyzing casting defects in high-pressure die casting is highly significant and has attracted the attention of many researchers. High-pressure die casting is a process that requires the non-destructive testing of critical components using modern technology and advanced data analysis methods. Excellent examples are the Pareto analyses which have been conducted, in which critical areas were identified by performing a case study and prioritizing casting defects, for which a causal analysis was then performed (Bharambe et al., 2023; Govindarao et al., 2022) or an analysis using ANOVA and Taguchi analysis (Tariq et al., 2021). Analyses using artificial intelligence methods such as image analysis (Parlak & Emel, 2023) or artificial neural networks (De-Jian & Young-Peng, 2021), which are very relevant in modern digital reality, are also conducted. Analyses of the causes of defects in die castings using machine learning models have not been published before and are the focus of this article.

It may be caused by the fact that there is a constant desire to improve productivity and profitability in order to increase the competitiveness of enterprises. In countries where labor and energy costs are high, it is possible to work on reducing them through the application of modern technologies. In particular, the trend is towards increasing industrial productivity by completely reducing waste (Seit, 2018), including defective products. The foundry industry therefore, suffers from the production of products with defects, and from a large number of process parameters, some of which can affect the formation of defects in the product (Patil & Inamdar, 2014). The final cost of a given product is determined by the cost of its manufacture, and it depends, in particular, on the used materials and processing costs.

The production of a given casting of a quality suitable for the customer may require additional quality verifications and quality audits, which could be reduced by applying an appropriate methodology allowing to predict the occurrence of a defect in the casting. Each defect that arises necessitates disposal of a defective product, recovery of used materials, or repair using new additional materials, extra working hours of operators and devices. Moreover, defects in products increase the total time of delivery to the customer, as it requires additional production than ordered by the customer. It is important to mention that a reduction of defects occurrence even by 1%, can ensure annual savings of several million PLN for medium-sized foundries (Falęcki, 1997).

Therefore, the proper diagnosis of product defects and their causes is an extremely important issue and arouses high interest among scientists, technologists, and experts of production companies. In the process under consideration, a characteristic feature is that the production costs (strictly the casting process) can be comparable to the costs of finishing and quality control. Therefore, it is important to detect product defects at an early stage in the process, thus saving up to 50% of the total manufacturing costs. Consequently, the main goal of the research is to diagnose the causes of manufacturing defects, more specifically on die-casting leakage, based on advanced data modeling using artificial neural networks, regression trees, and support vector machines methods, to compare and find the most effective method and apply it at an intermediate stage after the production of the casting and before the finishing phase.

2. Experimental procedures

High-pressure die casting (HPDC) is a highly precise and very expensive manufacturing process, described in detail in (Kozłowski et al., 2019). Therefore, the main objective is to reduce the amount of waste produced during the casting process (Sabau, 2006). This is very significant and still difficult to achieve, because of the above-mentioned complexity of metallurgical processes and the continuous lack of complete understanding of them. In order to improve the quality of castings by learning the complex relationships between the quality of the final product defined by the value of the output parameter and the parameters of the casting process, a systematic data analysis strategy has been developed using machine learning (ML) techniques.

Discovering relationships between process parameters derived from different stages of the casting process is highly complex and currently almost impossible.

For advanced data-driven (soft) modelling, production data were collected from one of the leading high-pressure die-casting foundries producing castings for the automotive industry, these data form the basis of the proposed research methodology. The cooperating foundry collects over 60,000 new datasets with new observations per month in its databases. The actual data used for the research contained over 10,000 observations with 59 parameters, characterizing the manufactured castings. The first step of machine learning implementation, is called data acquisition (Chen S. & Kaufmann, 2022), so the first selection of the range of variables and the sampling period was made using engineering, expert knowledge by the process engineers. From all the relevant process parameters, the parameters named as input variables affecting the output variable defined as leakage, detected during the high-pressure test, were selected. Leakage is expressed in cubic centimeters and indicates the volume of compressed air that has leaked through the casting, providing a measure of the tightness of the casting, and indicating its quality. There are certain ranges of leakage allowed by the recipients, further ranges classify the casting for repair, or scrap. Generally, the higher the leakage value, the lower the quality of the casting produced.

Each manufacturing data must go through data pre-processing stages to transform raw data into an understandable format (Agarwal, 2015) and to handle its imperfections to organize it for further data processing procedures. The mentioned stage is very important because when faulty data is applied to models, we can get very erroneous conclusions that can affect further decision-making processes. Therefore, it is believed that up to 80% (Ruiz, 2017) of the time spent on data analysis, is dedicated to data preprocessing, in order to finally obtain correct conclusions that enable proper future predictions. Performing data preprocessing involves five main tasks: data cleaning, data integration, data transformation, data reduction, and data discretization (Grzegorzewski & Kochański, 2019b).

The process of preparing the data for further analysis is described in detail in the publication (Okuniewska et al., 2021). Five datasets were created from the full dataset, based on the flow charts of the variables. In each of the datasets, different ranges of the dependent variable and different numbers of observations were included. The first and third datasets created were relatively large datasets containing more than 10,000 observations, where the first dataset included all data, while the third included observations representing a distribution close to a normal distribution of the dependent variable. The other three data sets, meaning the second, fourth and fifth, were relatively small, as they contained a maximum of 140 observations. They were created to test the cause of the occurrence of high values of the dependent variable on an undesired level.

In the next step, optimization of the number of independent variables should be applied, limiting them to those that are significant in terms of their impact on the dependent variable. This is referred to a significance analysis and can be performed using statistical methods such as the Kruskal–Wallis test or ANOVA analysis of variance (Perzyk et al., 2008). The computational analysis resulted in two parametric statistics: p for both tests and additionally the H statistics for the Kruskal–Wallis test, and F for the ANOVA test (Okuniewska et al., 2021). After applying the aforementioned statistical methods, optimization of the number of independent variables can be carried out. The optimization is possible by analyzing the value of the obtained p-statistic. All variables having a p-value less than 0.05 were selected for further advanced data-driven modelling. Finally, for each of the five datasets, three variable selection criteria (K-W – based on the Kruskal–Wallis test results, reversed K-W – based on the reversed Kruskal–Wallis test results, and ANOVA – based on the ANOVA classical and reversed) were added (Okuniewska et al., 2021).

Consequently, 15 datasets were used for further modelling, including 6 large datasets (first and third data sets organized according to the K-W, reversed K-W, and ANOVA criterion) and 9 small datasets (second, fourth and fifth data sets organized according to the K-W, reversed K-W, and ANOVA criterion) with different numbers of process parameters, quantifies in Figure 1. The effectiveness of the methods used for data preprocessing can be confirmed by the illustrated number of variables that were selected for further analyses, the best result was obtained for the small dataset (fifth) where the number of the independent variable was optimized by 81% in comparison to the original number (Fig. 1) (Okuniewska et al., 2021).

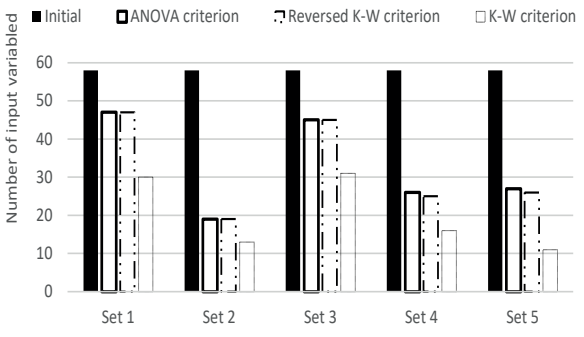

Fig. 1. Comparison of the number of independent variables optimised by the statistical methods (Okuniewska et al., 2021)

The focus then turns to an important step in building machine learning models, involving regression problems handled by machine learning methods. For this purpose, the features of the different methods were considered (Tseng et al., 2004). The methods most commonly used among researchers for product defect extraction are support vector machines (SVM), decision trees (DT), artificial neural networks (ANN), and Bayesian networks (BN) (Bártová et al., 2021). Bayesian networks method was tested on the foundry data in (Sata & Ravi, 2017; StatSoft, 2011a; Thomas et al., 2004).

In this methodology, the advanced modelling was initiated with applying the ANN method, which demonstrates considerable ability to represent complex hidden relationships between parameters of the manufacturing process. This method allows complex and complicated non-linear problems to be solved based on the data. Artificial neural networks can even make inferences from incomplete data containing noise. They have the ability to process information in parallel as their computational power allows them to perform more than one task at the same time. After the learning process, the networks are capable of effective prediction and generalization. This method is suited to working with large datasets as it stores data information throughout the network rather than in an additional database (Mijwel, 2018).

For the further and more precise determination of the relevant parameters, the decision tree (DT) method has been used due to its radically different nature and prevalence in industrial and other applications (Chen R.-S. et al., 2005; Hur et al., 2006; Rokach & Maimon, 2006). The use of DT has advantages such as speed of computation and facilitated interpretation of results due to their graphical tree representation (Perzyk & Soroczyński, 2019). This solution as rule-based is more transparent in comparison to black-box solutions such as artificial neural networks. Decision trees are characterized by a procedure for selecting relevant features from the examined data set in order to find the best division of the data into parts with maximum uniformity. This process is then repeated for each of the resulting data fragments (Timofeev, 2004). Often, the decision trees developed are very complex, as they contain many levels and a large number of variables. This method deals perfectly with outliers, it is even able to separate them in one of the nodes, the tree structure also does not change with respect to monotonic transformations of the independent variables, so any variable can be replaced by its logarithm or square root and the tree structure will not change (Timofeev, 2004).

Additionally, due to current trends and the ability to work with noisy data, it was decided to test the modelling using the support vector machine method. In recent years, SVM, as one of the most effective machine learning techniques (Karimi et al., 2019), has attracted the attention of many researchers involved in modeling manufacturing data to improve product quality, support decision-making, and implement process diagnostics (Esmaeilian et al., 2016; Köksal et al., 2011; Del Vecchio et al., 2019). This method has many advantages, including the ability to effectively consider data having variables of continuous type and variables of categorical type, with non-linear relations and not having normal or close-to-normal distribution. The method can work perfectly with noisy data, complex data sets, and even with numerous outliers. It avoids model overfitting and provides performance at the expected level (Vapnik, 2000). These features of the support vector machine method can be very useful for modelling foundry data, due to their specificity, that is, imperfect quality, complexity, variety of types of variable distributions, the occurrence of correlations between different process parameters, and lack of balance in the representation of values (Okuniewska et al., 2021).

Finally, a multidimensional optimization of the process parameters was carried out to illustrate what exactly influences the formation of a defect in a product, in this case a casting. More specifically, it was determined which process parameter values influence the production of defective castings.

3. Results and discussion

3.1. Artificial neural networks

In order to diagnose the hidden dependencies, which are highly complex, the ANN method was used (Okuniewska, 2020). The application and results of ANN were described in detail in the connected article (Okuniewska et al., 2021).

In the analysis, a multilayer perceptron (MLP) was applied, creating a unidirectional network with an input layer, an output layer and one hidden layer. Learning of the multi-layer perceptron was performed by presenting datasets divided in different proportions into a learning, testing and validation set. The datasets contained sets of inputs for successive learning observations and corresponding examples of outputs with which the modelled artificial neural network should respond.

During the research, it was very important to specify a test set equal to 0%, 10%, 15% or 20%. However, for real data from foundries, the creation of test and validation sets (completely independent) is often omitted in the process of learning the network, as real data may have deficiencies and be unbalanced, which

affects the difficulty of selecting such sets. However, in the presented study, the quality of the network was also checked for models having test and validation sets.

Next, the architecture of the network was determined, i.e., the number of layers, the number of neurons in the layers. During this step, it had to be kept in mind that determining the right number of hidden neurons and hidden layers is a kind of challenge for the SSN designer. In the present analysis, the numbers of neurons were kept small to avoid overfitting the model and ranged from 7 to 23 for large datasets (i.e., first and third) and 2 to 5 for small datasets (i.e., second, fourth and fifth). It was concluded that there was no reason to create more than one hidden layer, as this would not increase the quality of the result and would only complicate the neural model. The final step was learning the network, during which a tangent (tan) activation function was used in the hidden layer and tangent-linear (tan-lin) functions in the output layer.

The model performance was checked by the root mean square error (RMSE) result, calculated from the difference between predicted and actual leakage values.

The most promising and best results were obtained for modeling big data sets, where RMSE was equal to 0.84, and was obtained during the test without testing subset, for the model built with tangent-linear (tan-lin) activation function and with 23 neurons at the hidden layer (Fig. 2).

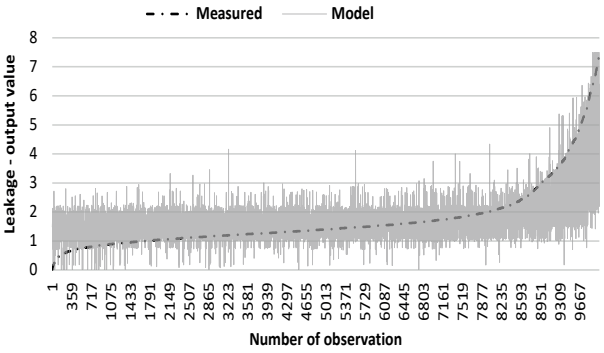

Fig. 2. Model and measured results comparison for the third K-W criterion set

An ablation study was not used, as were some other known methods for determining the significance of input variables from advanced target models (see e.g., Perzyk et al., 2008). Instead, an approach based on output optimization was tested, which, in principle, makes it possible to capture the simultaneous influence of multiple, not preselected variables, which was important because of possible hidden connections between such variables.

On the other hand, the modeling of small data sets gave the best RMSE result of 0.9, for the test without learning stop, the model contained 5 neurons, built with tangent activation function in the hidden layer and linear activation function at the output (tanh-lin) (Fig. 3).

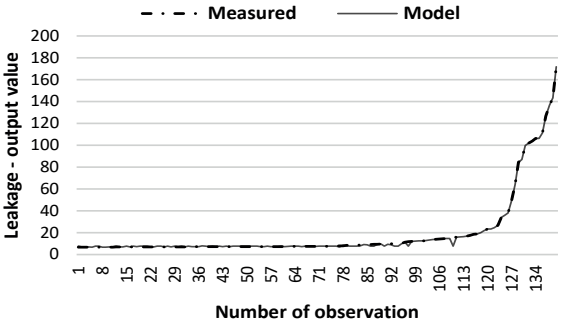

Fig. 3. Model and measured results comparison for the fourth ANOVA criterion set

After the application of advanced modeling with ANN, the multidimensional optimization of the process parameters was performed. This research was described in (Okuniewska et al., 2021). The results obtained showed that, in most cases, the optimization was not able to illustrate the exact cause of the casting defect formation. However in several cases, it was possible (Okuniewska et al., 2021).

Generally, it was found that extracting information from a data-driven model such as a neural model was crucial. Optimization methods were tried, but because of the problem's complexity, meaning the multidimensionality and the probability of local extremes, only in a few cases repeatability of results was obtained. It was concluded that there is a need to develop other methods for analyzing soft models, filtering out those obscuring dependencies, therefore the methods of decision trees and support vector machines were applied.

3.2. Regression trees

During the creation of a modelling plan using decision trees, more specifically classification and regression trees (C&RT) should be taken into consideration when determining the criteria for assessing prediction fidelity, selecting splits, determining when to stop the split generation process, and selecting the "right-sized" tree is the crucial point of the obtained results so of the prediction quality. The divisions can be continued until a perfect classification is obtained.

However, this does not make much sense as the resulting tree structure would be very complex, and such a model would most likely not give good predictions of new observations (StatSoft, 2011a). Therefore, a split stopping rule based on setting the value of the minimum number of cases in a given node was used. In the plan of computations, the effect of the value of the minimum number of cases in a given node on the quality of the prediction was checked. Generally, the higher the number of cases in a given node, the worse the prediction quality (Fig. 4).

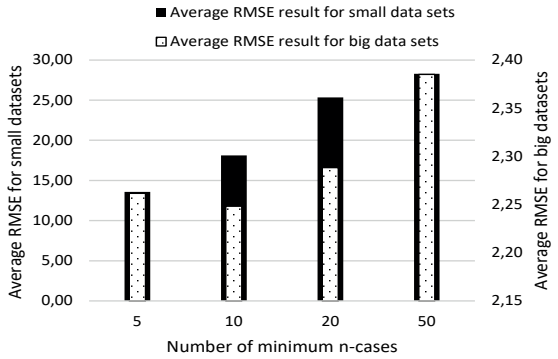

Fig. 4. The average value of RMSE with regard to the set number of minimum *n*-cases for big and small data sets modelled without V-fold cross-validation

The research also tested the impact of modelling with or without V-fold cross-validation. This method is useful when there is no test sample, and the learning sample is too small to form a separate test sample. The set value of the V-fold cross-validation parameter informs about the number of subsamples, which will be randomly created from the learning sample (StatSoft, 2011a). In the case of the data used for the study, the application of V-fold cross-validation resulted in obtaining outcomes equal to the average value of all observations of the dependent variable. Thus, regression trees with V-fold cross-validation are not able to learn or predict changes in the value of the dependent variable, which is why they were not used in further stages of the study.

The modelling was performed on data from the above described fifteen datasets tested for the combination of a minimum number of *n*-cases (5, 10, 20 and 50), without V-fold cross-validation value. The lowest RMSE value was obtained for the large dataset set according to the reversed K-W and ANOVA criterion and a minimum 50 number of cases (Fig. 5). On the other hand, the lowest RMSE value for the small data set was 7.1, in the K-W criterion data set and minimum 5 number of cases (Fig. 6).

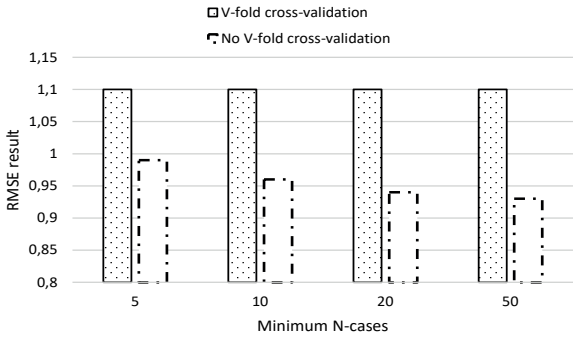

Fig. 5. Third (ANOVA and reversed K-W criteria) set modelling results, for four minimum *n*-cases with and without V-fold cross-validation

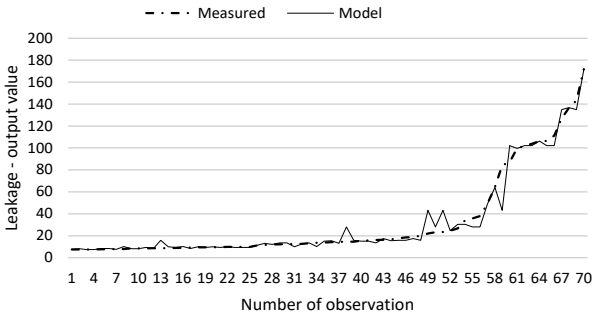

Fig. 6. Second set modeling results in the K-W criterion, with 5 minimum *n*-cases, without V-fold cross-validation

3.3. Support vector machines

In order to effectively deal with the problems and improve the quality due to current trends, the modeling method known as SVM was applied. This method at one point became competitive with the artificial neural network (ANN) method. Regression models were developed for this by using 8 different data sampling methods for each of the five datasets according to three criteria and therefore different numbers of predictor variables. The method constructs non-linear decision boundaries based on the performed regression. It is a concept of the decision boundary, which is divided by the construction of boundaries separating objects of different class membership. It is characterized by relatively high flexibility in the approach to regression problems, due to the nature of the feature space – predictors, in which the boundaries are built. The optimal separating hyperplane is built in an iterative learning algorithm, minimizing a certain error function.

The model is configured by adjusting the type of error function, i.e., type one, i.e., epsilon-SVN regression,

and type two, i.e., ni-SVM regression, and selecting the kernel function: linear, polynomial, radial basis function (RBF) and sigmoidal. In the regression, the dependence of the functional dependent variable a, on the set of independent variables b, is in the deterministic type (*f*), calculated with some addition of random noise, according to Equation (1) below:

$$a = f(b) + noise \tag{1}$$

Thus, the main task of the applied method is to find the form of the function *f*, which should give the best possible value of the dependent variable for new cases, which the model has not presented before. The model is learned by sequential minimization of the error function, using two methods SVM type 1 according to Equation (2), and SVM type 2, computed according to Equation (3), where for a constant C called capacity, w is a vector of coefficients, ξ_i, ξ_i^1 are the parameters of the overlapping cases, and the index *i* numbers of *N* learning cases (StatSoft, 2011b).

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi_i^1 \tag{2}$$

$$\frac{1}{2}w^T w - C(v\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^1) \tag{3}$$

The most commonly used kernel function in SVM is the RBF function, due to its limited range in the *b*-variable field (StatSoft, 2011). In the present study, the results of applying all four kernel types were checked and compared (Fig. 7). The lowest RMSE value of 1.1 was obtained in a large data set using the polynomial kernel function with SVM type 1, while in a small data set the best result of 29.2 was obtained using the RBF kernel function with SVM type 2 (Fig. 8). During the research, difficult applications of SVM models were noted.

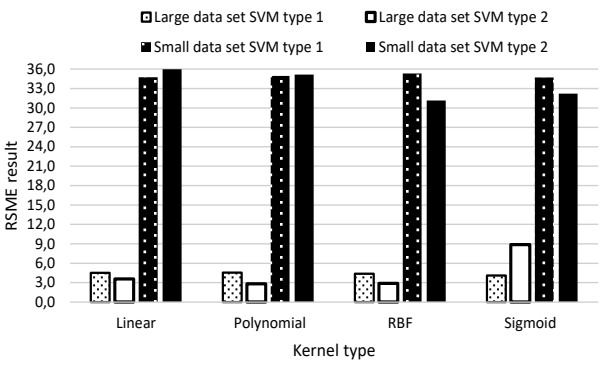

Fig. 7. Comparison of the average RMSE results for four kernel types for two SVM types

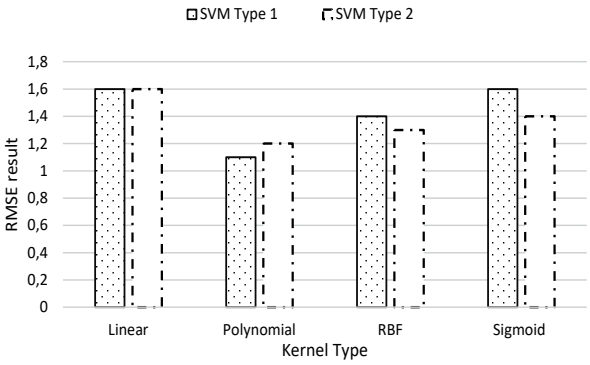

Fig. 8. Third set in the K-W criterion average RMSE results for four kernel types for two SVM types

3.4. Model testing for multidimensional optimization of process parameters

Multidimensional optimization of process parameters was possible by selecting the most effective method for advanced modelling based on large data sets. Accordingly, the research was initiated by repeating the modelling with the artificial neural network method included both the absolute best models (but with a lack of generalizability (obtained without testing set) as well as models characterized by a higher value of mean square error of prediction but having the ability to generalization (containing test sets) with simultaneous notation of the weights and programming of the model response (Okuniewska et al., 2021).

The information from the neural model can be extracted in various ways, preferably by appropriately scheduled querying of the network (Perzyk et al., 2008). This way, called the 'pedagogical' way, consists of treating the model as a black box, using a suitably designed network interrogation procedure to obtain the information sought (Perzyk et al., 2008). There is also a second way, called 'decomposition', which involves analyzing the weights of the artificial neural networks created, or more generally the individual parameters of the model (Perzyk et al., 2008). However, the approach based on analyzing the weights of the networks has proven to be insufficient (Garson, 1991; Perzyk et al., 2003). This is because each network learning process generates different weights, which are the source of significant differences in significance coefficient values (Perzyk et al., 2008). For this reason, the strategy developed here took the first approach – the 'pedagogical' approach – and included a multivariate optimization of all process parameters for maximum and minimum defect (leakage) values. The idea behind such

an approach was to assume that, under conditions of possible changes in all process parameters, occurring randomly, the repetitive values favoring and preventing the defect will take on those of them that actually have a significant role. The results of such optimization should also determine the direction of the induced changes. Through this analysis, it was possible to further optimize the process parameters for the maximum value of the leakage, meaning a casting with a defect, and a minimum leakage, so a casting without a defect. The optimalization was performed by using the MS Excel program's Solver add-in.

The results of the analysis for the five data sets according to the three criteria indicate that, in most cases, multidimensional process parameter optimization is unable to visualize what exactly influences the formation of a defect in a product, in this case a casting. More precisely, it is not always able to determine which process parameter values influence the production of defective castings.

Nevertheless, it has been possible to obtain such an answer in a few important cases (Figs. 9–13). It has been indicated that increased values of the parameters: 'multiplication delay,' 'urban water temperature' and 'alloy dosing time 2,' as well as decreased values of the parameters: 'blowing time' and 'cooling circuit flow 14,' influence the achievement of higher leakage values and thus the formation of a defect in the casting.

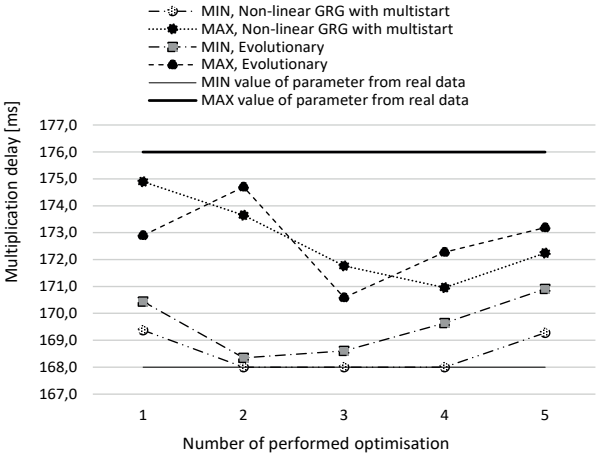

Fig. 9. Results of multidimensional optimization of first set, according to K-W criterion, for a network with 100% values in the learning set, 19 neurons in the hidden layer and a tanh activation function at the output for the 'multiplication delay'

This parameter (Fig. 9) is a key stage in the process that aims to reduce the shrinkage porosity of the castings produced, through the forced feeding of the liquid alloy into the solidifying casting. This parameter defines the point at which this phase of the process begins, and its significance obtained from modelling is not surprising. It is possible that in the case of forced metal feed starting too late, it may be inefficient due to the large fraction of solidified metal in the casting.

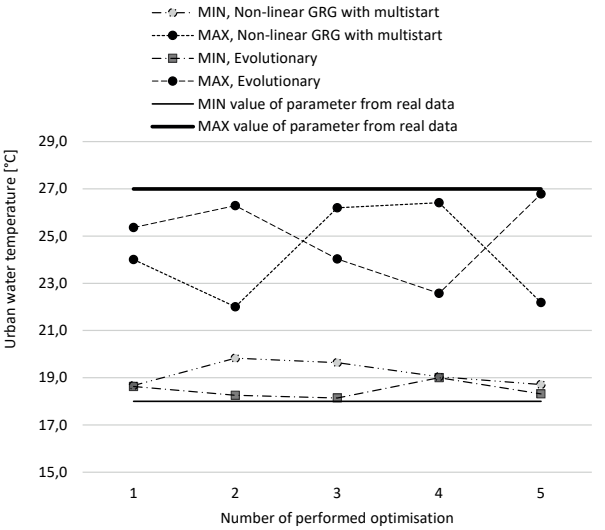

Fig. 10. Results of multidimensional optimization of first set, according to K-W criterion, for a network with 100% values in the learning set, 22 neurons in the hidden layer and a tanh activation function at the output for the 'urban water temperature'

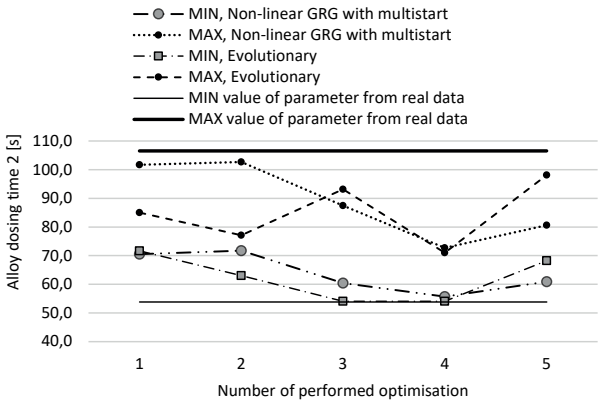

Fig. 11. Results of multidimensional optimization of first set, according to reversed K-W and ANOVA criterion, for a network with 100% values in the learning set, 22 neurons in the hidden layer and a tanh activation function at the output for the 'alloy dosing time 2'

The next results (Fig. 10) obtained for this parameter may be related to the imperfection of the mould temperature stabilization system, whose role in the formation of porosity is quite evident. The impact mechanism for the third parameter (Fig. 11) is not clarified. Perhaps a significant role is played here by correlations with other process parameters.

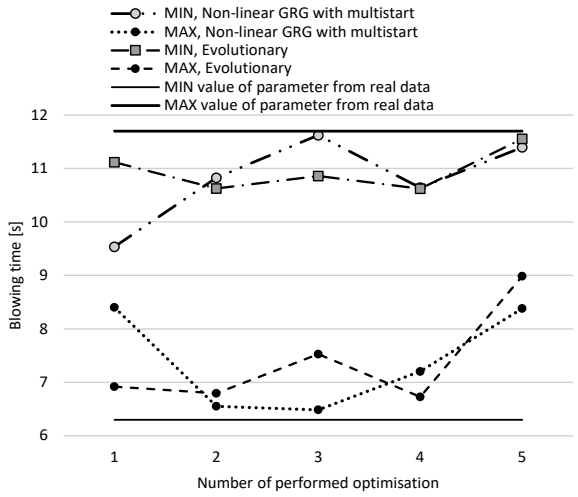

Fig. 12. Results of multidimensional optimization of third set, according to reversed K-W and ANOVA criterion, for a network with 100% values in the learning set, 21 neurons in the hidden layer and a tanh activation function at the output for the 'blowing time'

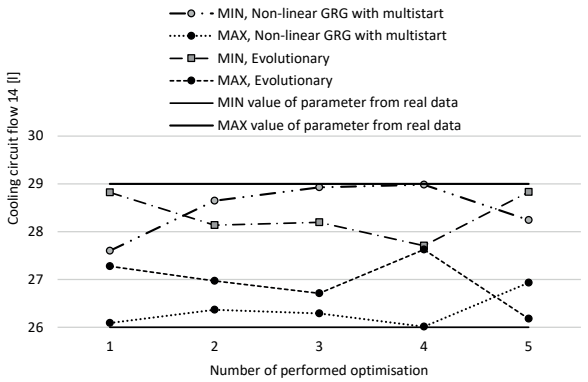

Fig. 13. Results of multidimensional optimization of the fourth set, according to K-W criterion, for a network with 100% values in the learning set, 4 neurons in the hidden layer and a tanh activation function at the output for the 'cooling circuit flow 14'

Interpretation of the impact of the fourth parameter (Fig. 12) is not simple, including for foundry technical staff, and would require deeper analyses and additional studies. The fifth parameter (Fig. 13) detected seems quite natural, as a reduction in the intensity of water flow in certain channels, resulting in a reduction in the local cooling intensity of the casting, may have the effect of increasing the solidification time and concentrating the porosity of the casting in that location leading to leakage.

Although it has been possible to obtain promising results in some cases, whereby it is possible to select the relevant variables and their specific values as being relevant to the process under study, indicating that they must be treated as critical by the foundry staff, it has not been possible in every case to obtain such information from the models created.

4. Conclusions

The formation of defects in castings often seems to be random, and the causes are frequently unknown. Data coming from many stages of the casting processes are very complex and perfectly fit the purpose of applying machine learning tools dedicated to large and complex data sets. The applied tools have been tested to work with data of imperfect quality, complexity, variety of types of variable distributions, the occurrence of correlations between different process parameters, and the lack of balance in the representation of values (Okuniewska et al., 2021).

The results of the analysis based on the multidimensional optimization of input variables showed that specific parameter values favor higher values of leakage and consequently defect the formation of the product. It was indicated that increased values of the parameters: multiplication delay, urban water temperature and alloy dosing time 2, as well as decreased values of the parameters: blowing time and cooling circuit flow 14, influence the achievement of higher leakage values and thus the formation of a defect in the casting. Although it has been possible to obtain promising results in some cases, whereby it is possible to select the relevant variables and their specific values as relevant to the process under study, indicating that they must be treated as critical by the foundry staff, it has not been possible in every case to obtain such information from the models created.

Summarizing the results obtained from three types of machine learning methods tested for the same problem, on the same data sets, and presented in the current article, the following can be concluded: the smallest RMSE means the best fit of the model to the data and the best prediction quality. The lowest RMSE and the best model fitting were obtained using the three methods for the large data set (which contained over 10,000 observations), so more data means more accurate estimates and better-quality prediction of the value of the dependent variable. The best results using each method were obtained in the large dataset (third with reversed K-W and ANOVA criteria). This results in the mentioned set showing significant variability in the values of the dependent variable and the presence of elevated values indicative of a defect in the product, the distribution of the sorted values of the dependent variable was close to a normal distribution. The most accurate values were obtained from the Artificial Neural Network method, while Regression Trees appeared to be slightly less precise, especially for the small data sets (described in section 2). Generally, in sets with small amounts of noisy data, i.e., in small, generally unbalanced data

sets, the best results in identifying the most significant variables were also obtained with the artificial neural networks method (Tab. 1). The best results using each method were obtained for the second and fourth small datasets determined by the reversed K-W and ANOVA criteria. This set also showed significant variability in the values of the dependent variable and the presence of elevated values indicative of a defect in the product. The values found from SVM, using the same research methodology essentially reflect the expected tendencies, however, their values are less accurate than those obtained from ANN and DT.

Table 1. Comparison of the RMSE result between three machine learning methods

Machine Learning Method	Size of the data set	Coefficient: Root Mean Square Error (RMSE)
Artificial Neural Networks (ANN)	small	0.90
	big	0.86
Classification and Regression Tress (C&RT)	small	7.10
	big	0.93
Support Vector Machine (SVM)	small	29.20
	big	1.10

The study identified general problems associated with data-driven modelling, namely the inherent randomness of neural network models (having different weights), the difficult application of SVM models, and the limited performance of models developed using the decision tree method. Previous research (Okuniewska et al., 2021; StatSoft, 2011a) has shown that the same problem applies to modelling using a Bayesian network, and problems with optimization as a potential approach useful in model inference leading to the identification of the root causes of process errors. The randomness of the optimization results is very likely as local minima are typical. Therefore, reasoning based on neural models is limited and can be used as a source of rather general indications regarding the recommended parameters of the production process. The methodology of using this type of information in making decisions in specific situations was presented (for a similar metallurgical process) in the recently published work by Perzyk et al. (2022).

Despite the problems detected, the results obtained in the case of modelling using the artificial neural networks method seem to be promising and supply the motivation for further research. The results indicate that predicting the defect level in castings can be done with satisfactory accuracy and therefore they can be a very significant benchmark for high-pressure foundries. However, there is still a need to develop the data preparation methodology to ensure that cases of insufficient quality are adequately represented. It is planned to extend the research to include classification models and, if satisfactory prediction results are obtained, it will be possible to extend the research towards the creation of models that predict quality with a set time in advance.

References

Agarwal, V. (2015). Research on data preprocessing and categorization technique for smartphone review analysis. *International Journal of Computer Applications*, *131*(4), 30–36. https://doi.org/10.5120/ijca2015907309.

Bártová, B., Bína, V., & Váchová, L. (2022). A PRISMA-driven systematic review of data mining methods used for defects detection and classification in the manufacturing industry. *Production*, *32*, e20210097. https://doi.org/10.1590/0103-6513.20210097.

Bharambe Ch., Jaybhaye, M.D., Dalmiya, A., Daund, C., & Shinde, D. (2023). Analyzing casting defects in high-pressure die casting industrial case study. *Materials Today: Proceedings*, *72*(3), 1079–1083. https://doi.org/10.1016/j.matpr.2022.09.166.

Bowers, K., & Pickerel, T.V. (2019). Vox Populi 4.0: big data tools zoom in on the voice of the customer. *Quality Progress*, *52*(3), 32–39.

Chen R.-S., Wu R.-C., & Chang C.-C. (2005). Using data mining technology to design an intelligent CIM system for IC manufacturing. In *Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks*. IEEE. https://doi.org/10.1109/SNPD-SAWN.2005.78.

Chen, S., & Kaufmann, T. (2022). Development of Data-Driven Machine Learning models for prediction of casting surface defects. *Metals*, *12*(1), 1. https://doi.org/10.3390/met12010001.

De-Jian, X., & Young-Peng, Y. (2021). A Neural Network Based Defect Prediction Approach for Virtual High Pressure Die Casting. *Journal of Physics: Conference Series*, *1948*, 012019. https://doi.org/10.1088/1742-6596/1948/1/012019.

Del Vecchio, C., Fenu, G., Pellegrino, F.A., Di Foggia, M., Quatrale, M., Benincasa, L., Iannuzzi, S., Acernese, A., Correra, P., & Glielmo, L. (2019). Support Vector Representation Machine for superalloy investment casting optimization. *Applied Mathematical Modelling*, *72*, 324–336. https://doi.org/10.1016/j.apm.2019.02.033.

Esmaeilian, S., Behdad, B., & Wang (2016). The evolution and future of manufacturing: A review. *Journal of Manufacturing Systems*, *39*, 79–100. https://doi.org/10.1016/j.jmsy.2016.03.001.

Falęcki, Z. (1997). *Analiza wad odlewów*. Wydawnictwa AGH.

Garson, D. (1991). Implementing neural network connection weights. *AI Expert*, *6*(4), 45–51.

Govindarao, R., Eshwara, K., & Srinivasa Rao, P. (2022). "Defect analysis and remedies in the High-Pressure Diecasting Process with ADC-12 Alloy" – A Technical review. *American Journal of Multidisciplinary Research & Development (AJMRD)*, *04*(07), 1–8. https://www.ajmrd.com/wp-content/uploads/2022/07/A470108.pdf.

Grand View Research (2019). Metal Casting Market Size, Share & Trends Analysis Report by Material (Aluminum, Iron, Steel), by Application (Automotive & Transportation, Building & Construction, Industrial), by Region, and Segment Forecasts, 2020–2025. https://www.grandviewresearch.com/industry-analysis/metal-casting-market.

Grzegorzewski, P., & Kochański, A. (2019a). From data to reasoning. In P. Grzegorzewski, A. Kochanski, J. Kacprzyk (Eds.), *Soft Modeling in Industrial Manufacturing* (pp. 15–25). Springer Cham. https://doi.org/10.1007/978-3-030-03201-2_2.

Grzegorzewski, P., & Kochański, A. (2019b). Data preprocessing in industrial manufacturing. In P. Grzegorzewski, A. Kochanski, J. Kacprzyk (Eds.), *Soft Modeling in Industrial Manufacturing* (pp. 27–41). Springer Cham. https://doi.org/10.1007/978-3-030-03201-2_3.

Hur, J., Lee, H., & Baek, J.-G. (2006). An intelligent manufacturing process diagnosis system using hybrid data mining. In P. Perner (Ed.), *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, 6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14–15, 2006, Proceedings* (pp. 561–575). Springer Berlin, Heidelberg. https://doi.org/10.1007/11790853_44.

Jacob, D. (2017). *Quality 4.0 impact and strategy handbook. Getting digitally connected quality management*. Retrieved May 24, 2021, from http://generisgp.com/2018/02/15/the-quality-4-0-impact-and-strategy-handbook/.

Karimi, F., Sultana, S., Shirzadi Babakan, A., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*, *75*, 61–75. https://doi.org/10.1016/j.compenvurbsys.2019.01.001.

Khan, W., Kumar, T., Cheng, Z., Raj, K., Roy, A.M., & Luo, B. (2022). SQL and NoSQL databases software architectures performance analysis and assessments – A systematic literature review. *Big Data and Cognitive Computing*, *7*(2), 97. https://doi.org/10.3390/bdcc7020097.

Köksal, İ., Batmaz, M.C., & Testik, M. (2011). A review of data mining applications for quality improvement in the manufacturing industry. *Expert Systems with Applications*, *38*(10), 13448–13467.

Kozłowski, 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. In A. Hamrol, A. Kujawińska, M. Barraza (Eds.), *Advances in Manufacturing II* (Vol. 2: *Production Engineering and Management*, pp. 253–267). Springer Cham. https://doi.org/10.1007/978-3-030-18789-7_22.

Mijwel, M.M. (2018). *Artificial Neural Networks advantages and disadvantages*. Research Gate. https://www.researchgate.net/profile/Maad-Mijwil/publication/323665827_Artificial_Neural_Networks_Advantages_and_Disadvantages/links/5aa2c01faca272d448b5a23d/Artificial-Neural-Networks-Advantages-and-Disadvantages.pdf.

Miłek, D. (2017). Development of the foundry industry in Poland. In *Metal 2017. 26th International Conference on Metallurgy and Materials May 24th–26th 2017 / Hotel Voronez I, Brno, Czech Republic, EU* (pp. 2250–2256). https://www.confer.cz/metal/2017/read/1718-the-development-of-the-foundry-industry-in-poland.pdf.

Okuniewska, A. (2020). Methods review of advanced data analysis tools, in process control and diagnostics. In K. Piech (Red.), *Zagadnienia aktualnie poruszane przez młodych naukowców. 17* (pp. 95–98). Creativetime.

Okuniewska, A., Perzyk, M.A., & Kozłowski, J. (2021). Methodology for diagnosing the cause of die-casting defects, based on advanced big data modelling. *Archives of Foundry Engineering*, *4*, 103–109. https://doi.org/10.24425/afe.2021.138687.

Parlak, I.E., & Emel, E. (2023). Deep learning-based detection of aluminum casting defects and their types. *Engineering Applications of Artificial Intelligence*, *118*, 105636. https://doi.org/10.1016/j.engappai.2022.105636.

Patil, G.G., & Inamdar, K.H. (2014). Prediction of casting defects through artificial neural network. *International Journal of Science, Engineering and Technology*, *02*(06), 298–314.

Perzyk, M., & Soroczynski, A. (2019). Assessment of selected tools used for knowledge extraction in industrial manufacturing. In P. Grzegorzewski, A. Kochanski, J. Kacprzyk (Eds.), *Soft Modeling in Industrial Manufacturing* (pp. 75–88). Springer Cham. https://doi.org/10.1007/978-3-030-03201-2_5.

Perzyk, M., Kochański, A., & Kozlowski, J. (2003). Istotność względna sygnałów wejściowych sieci neuronowej. *Informatyka w Technologii Materiałów*, *3*(3–4), 125–132.

Perzyk, M., Biernacki, R., & Kozlowski, J. (2008). Data mining in manufacturing: Significance analysis of process parameters. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *222*(11), 1503–1516. https://doi.org/10.1243/09544054JEM1182.

Perzyk, M., Dybowski, B., & Kozłowski, J. (2019). Introducing advanced data analytics in perspective of Industry 4.0. in die casting foundry. *Archives of Foundry Engineering*, *19*(1), 53–57. https://doi.org/10.24425/afe.2018.125191.

Perzyk, M., Kochański, A., & Kozłowski, J. (2022). Fundamentals of a recommendation system for the aluminum extrusion process based on data-driven modeling. *Computer Methods in Materials Science*, *22*(4), 173–188. https://doi.org/10.7494/cmms.2022.4.0782.

Raluca, D. (2021). Knowledge management systems in Quality 4.0. *MATEC Web of Conferences*, *342*, 09003. https://doi.org/10.1051/matecconf/202134209003.

Rokach, L., & Maimon, O. (2006). Data mining for improving the quality of manufacturing: a feature set decomposition approach. *Journal of Intelligent Manufacturing*, *17*(3), 285–299. https://doi.org/10.1007/s10845-005-0005-x.

Ruiz A. (2017). *Breaking the 80/20 rule: How data catalogs transform data scientists' productivity*. IBM Cloud. https://medium.com/@armand_ruiz/breaking-the-80-20-rule-how-data-catalogs-transform-data-scientists-productivity-7759a23a8893

Sabau A.S. (2006). Alloy shrinkage factors for the investment casting process. *Metallurgical and Materials Transactions*, *37B*(1), 131–140. https://doi.org/10.1007/s11663-006-0092-x.

Sata, A., & Ravi, B. (2017). Bayesian inference-based investment-casting defect analysis system for industrial application. *The International Journal of Advanced Manufacturing Technology*, *90*(9–12), 3301–3315. https://doi.org/10.1007/s00170-016-9614-0.

Seit J. (2018). *Trends and challenges: the die-casting industry on the road to the future, future prospects, spotlight metal the network for metal casting*. Retrieved 10 September, 2021, from https://www.spotlightmetal.com/trends-and-challenges-the-die-casting-industry-on-the-road-to-the-future-a-676717/.

StatSoft (2011a). *Drzewa klasyfikacyjne i regresyjne (C&RT)*. Retrieved 10 November, 2022, from https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstcart.html.

StatSoft (2011b). *Metoda wektorów nośnych*. Retrieved 10 November, 2022, from https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstsvm.html.

Tariq, S., Tariq, A., Masud, M., Rehman, Z. (2021). Minimizing the casting defects in high-pressure die casting using Taguchi analysis. *Scientia Iranica, International Journal of Science & Technology*, *29*(1), 53–69. https://doi.org/10.24200/sci.2021.56545.4779.

Thomas, P., Suhner, M.-C., Meutelet, B., & Brachotte, G. (2004). Quality monitoring of high-pressure die casting process based on Bayesian and neural networks. *IFAC Proceedings Volumes*, *37*(15), 299–304. https://doi.org/10.1016/S1474-6670(17)31040-6.

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications* [Master Thesis]. CASE Center of Applied Statistics and Economics, Humboldt University, Berlin.

Tseng, T.-L., Jothishankar, M.C., Wu, T., Xing, G., & Jiang, F. (2004). Applying data mining approaches for defect diagnosis in manufacturing industry. In *IIE Annual Conference and Exhibition 2004 – Houston, TX, United States. Duration: May 15 2004 → May 19 2004* (pp. 1441–1447).

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer New York, NY.

Wang, L., Törngren, M., & Onori, M. (2015). Current status and advancement of cyber-physical systems in manufacturing. *Journal of Manufacturing Systems*, *37*(2), 517–527. https://doi.org/10.1016/j.jmsy.2015.04.008.

Xu, L.D., Xu, E.L., & Li, L. (2018). Industry 4.0: state of the art and future trends. *International Journal of Production Research*, *56*(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806.