WARSAW UNIVERSITY OF TECHNOLOGY Index 351733
FACULTY OF CIVIL ENGINEERING
COMMITTEE FOR CIVIL AND WATER ENGINEERING
POLISH ACADEMY OF SCIENCES ISSN 1230-2945

DOI: 10.24425/ace.2024.149854

ARCHIVES OF CIVIL ENGINEERING

Vol. LXX ISSUE 2 2024

pp. 119 –134

© 2024. Justyna Dzięcioł, Wojciech Sas.

This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (CC BY-NC-ND 4.0, https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited, the use is non-commercial, and no modifications or adaptations are made.

Research paper

Estimation of the coefficient of permeability as an example of the application of the Random Forest algorithm in Civil Engineering

Justyna Dzięcioł[1], Wojciech Sas[2]

Abstract: A new world record for crude steel production was recorded in 2021, which increased by 3.8% over 2020. This also affected the amount of slag produced with this production. Total waste from industrial and construction production throughout the European Union accounts for as much as 48%. Therefore, waste management should provide for the recovery of as many resources as possible. European Union strategies in line with the circular economy objectives focus on ensuring policy coherence in the areas of climate, energy efficiency, construction and demolition waste management and resource efficiency. Slags are a material of interest to researchers in terms of their use in construction. Slags, on the one hand, are materials that are becoming better understood on the other hand, we are making sure of the heterogeneity of these materials. The characteristics of physical properties of slags are influenced by many factors, including the furnace split in which they are produced. This prompts the search for tools to help determine the parameters of slags based on already available data. The study aimed to verify the hypothesis that it is possible to determine the parameter of the filtration coefficient, relevant to applications in earth structures using the machine learning algorithm – Random Forest. In the study, two types of material were analysed: blast furnace slag and furnace slag. The results of the analysis yielded a high coefficient of determination (R^2) – 0.84–0.92. This leads us to believe that the algorithm may prove useful in determining filtration parameters in slags.

Keywords: circular construction, slag, Machine Learning, Random Forest, coefficient of permeability

[1]MSc, Eng., Warsaw University of Life Sciences – SGGW, Institute of Civil Engineering, 159 Nowoursynowska, 02-776 Warsaw, Poland, e-mail: justyna_dzieciol@sggw.edu.pl, ORCID: 0000-0002-2436-9748
[2]Prof., DSc., PhD., Eng., Warsaw University of Life Sciences – SGGW, Water Centre SGGW, 159 Nowoursynowska, 02-776 Warsaw, Poland, e-mail: wojciech_sas@sggw.edu.pl, ORCID: 0000-0002-5488-3297

1. Introduction

The efficient application of resources has become one of the key challenges for a long-term sustainable construction industry. Implementing a circular approach by valorizing waste and transforming it into eco-efficient building materials is an effective path for managing waste previously deposited in landfills. Meanwhile, this attitude offsets the depletion of natural resources and has a positive impact on environmental issues [1–4]. Generally, in European Union countries, slag is used as a by-product or is end-of-waste. If it is considered a waste for the smelter where it is produced it complies with the end-of-waste definition. This allows slag to be subjected to recovery, including recycling in the construction industry, for example. The different types of slags depending on their source (blast furnaces, thermal power stations, etc.) affect the possibility of their subsequent use in the construction sector, this is of course due to differences in the physical and chemical properties of these materials. A good understanding of the parameters of individual anthropogenic materials supports the design of possible solutions for their application in the construction sector [5–7].

One of the predictive solutions that may prove helpful in the analysis of parameter prediction based on previously available data may be Machine Learning algorithms. In recent decades, these techniques have been continuously developing and generating new algorithms with increasingly versatile applications. Most of the algorithms developed in recent years can be used for both categorization and regression prediction. Categorization algorithms (also known as classification algorithms) are used to classify data into predefined categories or classes. The goal is to train a Machine Learning model that can accurately predict the class of new, unseen data points based on the features or attributes of the data. Regression algorithms, on the other hand, are used to predict a continuous numerical value, such as a price, a temperature, or a stock price, based on a set of input features or variables. The goal is to train a machine learning model that can accurately predict the value of the target variable for new, unseen data points. Algorithms such as Artificial Neural Networks (ANN) are well recognized and have already found application in Civil Engineering [8–13]. Others like Random Forest, which will be analyzed in this article, is a newer algorithm whose effectiveness in predicting Civil Engineering phenomena has not been well documented.

The purpose of this paper will be to present the applicability of the Random Forest algorithm for determining the filtration coefficient in Blast Furnace Slag. Using validation methods, error analysis (MSR, RMSE and R^2) and SHAP analysis, were used statistically to evaluate the accuracy of the model.

1.1. Random Forest

The Random Forest technique is a relatively new Machine Learning method described by Breiman in 2001 [14]. The concept is based on bagging (“bootstrap aggregation”) [15], which is one way to reduce the variance of an estimate after averaging multiple estimates. Training M different trees on different subsets of data, chosen randomly with replacement,

will give the following relationship, Eq. (1.1):

$$f(x) = \sum_{m=1}^{M} \frac{1}{M} f_m(x) \tag{1.1}$$

where: f_m – m this tree.

Models derived from the Random Forest Algorithm often have very high predictive accuracy [16] and have also been widely used in many applications.

The Random Forest algorithm described by Breiman [14] uses two important elements to reduce generalization error. When creating individual trees, only a random subset of features is considered for each split, and each tree is given a randomly drawn subset of observations to train. Typically, there is an approach of bootstrap aggregation or bagging [17]. Random Forest models were comprehensively described by Louppe [18], and theoretical results on the hyperparameters of Random Forest models were summarized by Scornet [19]. Parameters for the decision trees themselves are often included in Random Forest tuning [20–22].

The reference implementation studied in this article is from the R language, the packages used are: Random Forest [23], Caret [24, 25]. The method of operation of the algorithm is presented in Fig. 1.

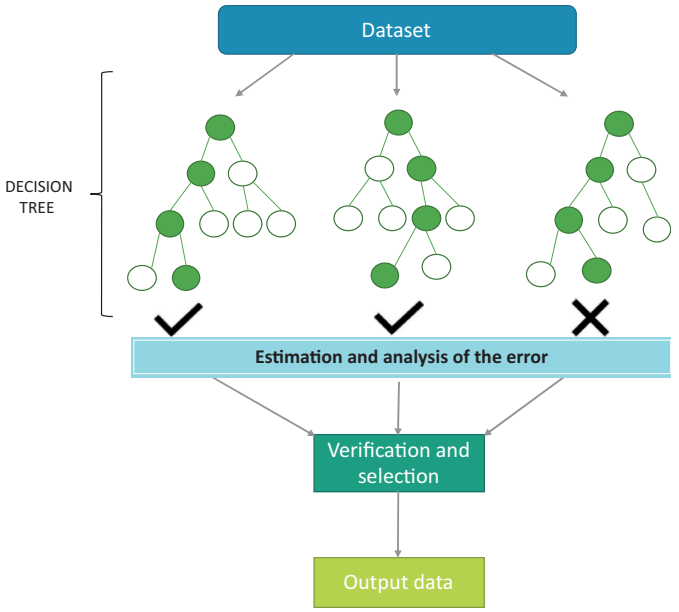

Fig. 1. Diagram of the functioning of the Random Forest algorithm

Random forest is a method of averaging multiple deep decision trees, trained on different parts of the same training set, to overcome the problem of over-fitting a single decision tree [14, 26, 27].

Among the biggest advantages of Random Forest is that it can be used to solve both problems and can be used in regression as well as classification. It can handle a large number of data sets with high dimensionality. Including datasets with more than a few thousand variables. Random Forest is very successful in identifying significant variables. Modeling with the Random Forest algorithm is based on bootstrap sampling samples processing input data with replacement. The algorithm can effectively estimate missing data and easily maintains accuracy, even when given a large amount of data [28, 29]. Disadvantages of the Random Forest algorithm include the occurrence of over-fitting the model to the observation data if the data is overly noisy. Overfitting means that the model fits the training dataset well, but not the test dataset. Random Forest is an algorithm that is difficult to interpret in terms of controlling its performance, it presents a black-box approach. This makes it challenging to interpret because it is only possible to check random seeds and various parameters. The paper attempts to interpret this algorithm using the SHAP technique.

2. Material and methodology

2.1. Material

The constant head method was used to test permeability characteristics for Blast Furnace Slag. The method is characterized by simplicity and unchanging test conditions, and the constant head method alone is one of the most reliable techniques for measuring permeability in non-cohesive soil [30–32]. The coefficient of permeability study used an aggregate of Blast Furnace Slag tested samples were from several parties of the material. Basic data on the grain size ranges of the tested samples are presented in Table 1. and data on physical parameters are included in Table 2.

Table 1. Grain size curve of tested materials of Blast Furnace Slag

Size particle	d_5 [mm]	d_{10} [mm]	d_{17} [mm]	d_{20} [mm]	d_{30} [mm]	d_{50} [mm]	d_{60} [mm]	d_{90} [mm]
Min.	0.20	0.60	1.70	2.50	7.80	9.50	10.50	15.00
Mean	0.45	0.83	2.77	4.40	8.70	10.53	11.67	16.67
Max.	0.60	1.00	4.00	7.50	9.30	11.10	12.50	19.00

Part of the samples were compacted with a Proctor normal energy [33] of 0.59 [J/cm^3], while others while no additional compaction energy was applied to the rest of the samples. The minimum, maximum and average values contained in Tables 1 and 2 were determined from data derived from an observation database created on the basis of laboratory tests carried out – filtration coefficient testing, granulometric testing and other physical features.

Table 2. Physical parameters of Blast Furnace Slag

	Volumetric density	Porosity	Index porosity	Homogeneity index – Cu	Grain size curvature index – Cc	Hydraulic gradient
Min.	0.80	0.56	1.29	12.00	7.21	0.08
Mean	0.88	0.62	1.69	14.57	8.08	0.55
Max.	1.03	0.66	1.93	17.50	9.66	1.02

2.2. Algorithm application methodology

A schematic of the process and methodology of the analysis conducted is presented in Fig. 2. The data were collected and cleaned before the analysis began. In the learning and testing process, the data was divided into a training sample – of 70% and a test sample – of 30%. Validation of the model was performed using the 10-fold Cross Validation method [34, 35]. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited sample of data.

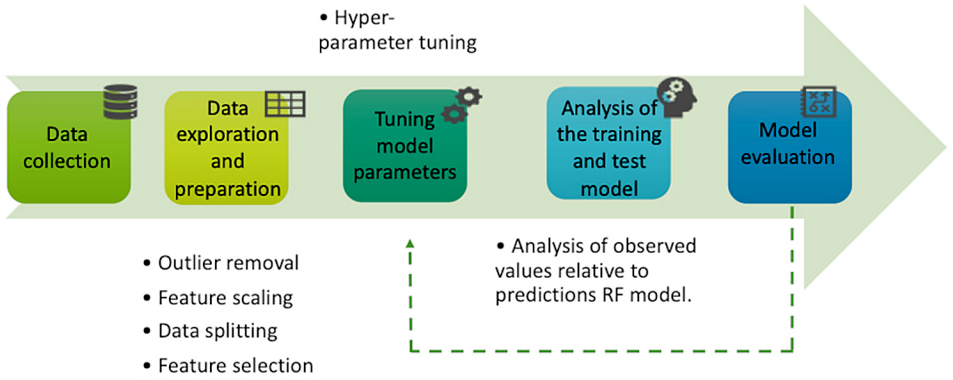

Fig. 2. General workflow with the Random Forest algorithm

The 10-fold Cross Validation method procedure has one parameter k, which refers to the number of groups into which the data sample is to be divided. Cross-validation is mainly used in Machine Learning to estimate a model's ability to predict data. This is a popular method, simple to understand and generally gives a less biased or less optimistic estimate of the model's ability than other methods. This allows the model obtained in the learning and testing process to be more reliable. The results of k-fold cross-validation were summarized using error analysis. For each model, the following were estimated [12]:

– Root Mean Square Error (RMSE), Eq. (2.1):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y})^2} \tag{2.1}$$

where: n – is the number of observations, $\hat{y}_i$ – the estimated value, y_i – the observed value.

– Mean Squared Residuals (MSR), Eq. (2.2):

$$\text{MSR} = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \tag{2.2}$$

where: n – is the number of observations, $\hat{y}_i$ – the estimated value, y_i – the observed value.

– Coefficient of determination (R^2), Eq. (2.3):

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} \tag{2.3}$$

where: $\overline{y}$ – is the mean values, $\hat{y}_i$ – the estimated value, y_i – the observed value.

The SHAP analysis was also performed.

3. Results

The Random Forest algorithm uses a bootstrap aggregation, where a single tree is built on a random sample of the dataset (about 70% of the data). The Random Forest algorithm uses bootstrap aggregation, where a single tree is built on a random sample of the dataset (about 70% of the data). The remaining observations are referred to as out-of-bag (OOB) and used as a method of assessing the quality of the Random Forest model, which involves using some of the training data samples that were not used in a given decision tree to assess the prediction of that tree [36, 37]. This activity is repeated many times and the results are averaged. Each tree is developed rather than reduced based on error measures, and this means that the variance of each of these individual trees is high. The variance can be reduced by averaging the results without increasing the bias. An example of an observation-based decision t]ree is shown in Fig. 3.

Another important parameter of Random Forest is that in parallel with the random sample of data (bagging), it also takes a random sample of input features at each split. The model using the R language uses the 'random Forest' package, which was used for the default random number of predictors that are sampled, for the regression it is the total number of predictors divided by three. The number of predictors that the algorithm randomly selects at each split can be changed through the model-tuning process.

Machine Learning techniques are mostly based on one or more parameters that need to be set before the learning process begins, these are named hyperparameters. The choice of values for each hyperparameter has a significant impact on the performance of any

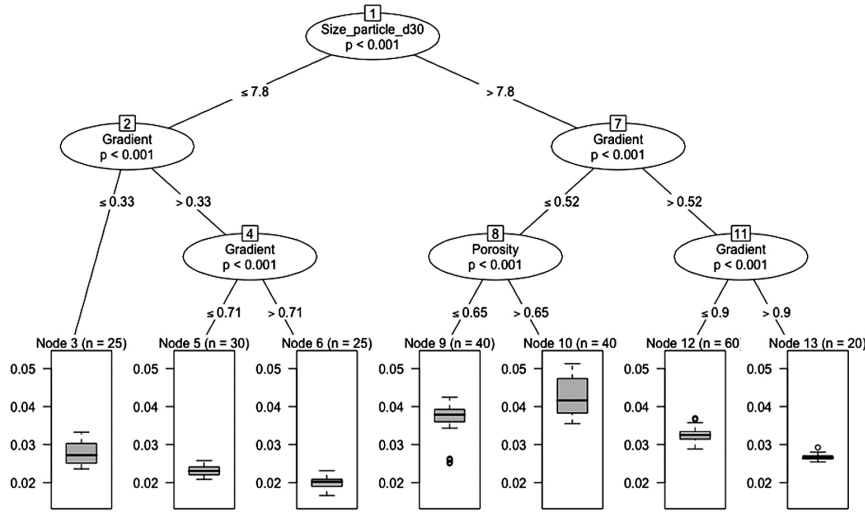

Fig. 3. Sample decision tree based on observation

particular model. Therefore, it is very important to identify and set the appropriate values for the model's hyperparameters before the building process. In machine learning, the process of optimal hyperparameters for a model is called tuning. The process of tuning hyperparameters for models performed manually is tedious and time-consuming, so it is worth using solutions that shorten this procedure. In the case of the R programming language, this is the Caret package [23]. It provides a set of tools for performing parameter searches using a grid graph and is used for preliminary comparative analysis.

There are three main hyperparameters to adjust for the Random Forest algorithm:

– mtry – the number of predictors (integer) that will be randomly sampled at each split when creating tree models,
– trees – the number of trees (integer) included in the ensemble,
– min_n – the minimum number of data points at a node (integer) that are required for a node to be split further.

Graphical results of this analysis are presented in Fig. 4.

Based on the compilation of the relevant parameters affecting the algorithm and analyzed from this perspective, the optimization of the hyperparameters of the algorithms. Fig. 5 presents the tuning of hyperparameters based on errors with different observations.

Based on parameter tuning (Figs. 4 and 5), the parameter mtry was estimated to be 9 and min_n equal to 8, all of which were used to tune the Random Forest algorithm. The error results generated for the individual learning, testing and validation trials are presented in Table 3. Analysis of the MSR, RMSE and R^2 errors indicates a high prediction explanation coefficient for the Random Forest model tuned with hyperparameters.

The analysis of the prediction results compared with the coefficient of filtration observations resulting from the tests are presented in Figure 6. The prediction was performed using 10-fold validation sampling. The results of the error analysis and matching are presented

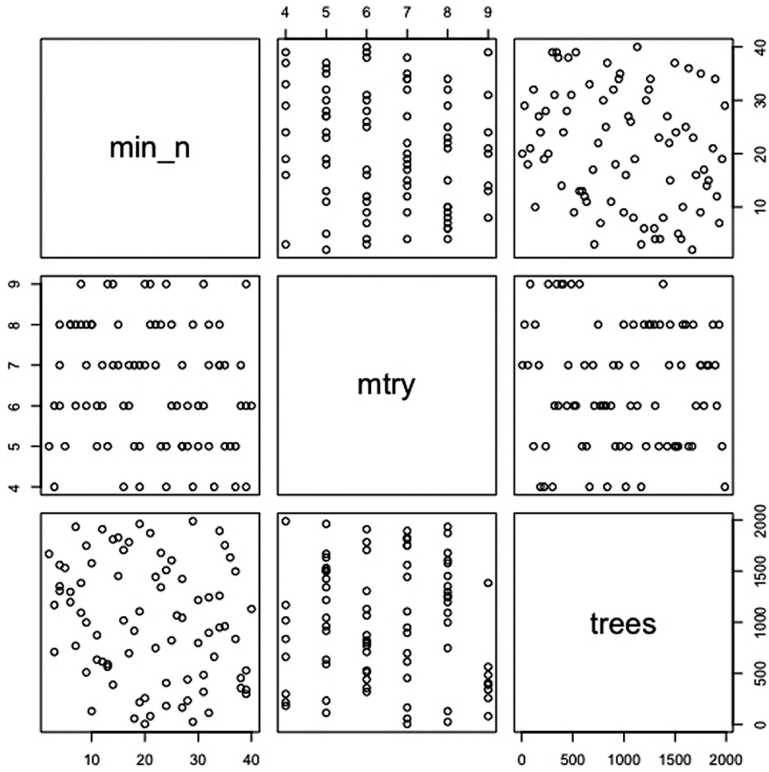

Fig. 4. Graphical analysis of preliminary tuning results

Table 3. Analysis of prediction errors and fits for different data sets

Metric	Train set	Test set	Validation set
MSR	$4.96 \cdot 10^{-6}$	$9.43 \cdot 10^{-6}$	$5.73 \cdot 10^{-6}$
RMSE	0.0020	0.0031	0.0023
R^2	0.924	0.8359	0.9089

above in Fig. 6, the results correlate with those obtained for the earlier trials included in Table 3. As early as 1992, Vukovic and Soro [38] noted that applying different empirical equations to the same porous medium material can provide different hydraulic conductivity values. The results obtained can differ by up to several tens of percent. In this context, it is important to search for methods with higher reliability of prediction results.

The most relevant parameters (Fig. 7) of the physical properties and grain size of the material were compiled based on the Random Forest model developed, the parameters with the highest impact affecting model building were gradient, compaction energy and volumetric density.

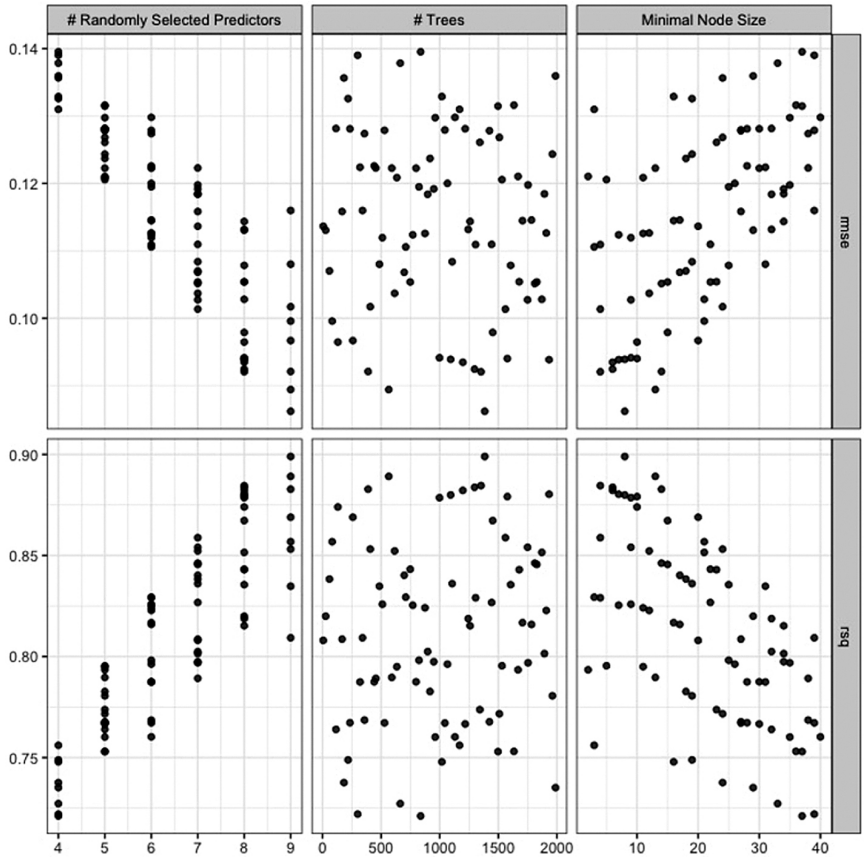

Fig. 5. Graphical analysis of tuning results

In the analysis of the interpretation and explanation of the prediction formed based on the random forest algorithm was used – Shapley Additive explanations (SHAP). It is mainly based on Shapley values, a method of cooperative game theory [39–41]. The model, developed by Lloyd Shapley, is a unique method of assigning rewards from a cooperative game [41]. The game is a Machine Learning model, the parameter values are the players in the game, and the predicted filter coefficient is the outcome of the game. The Shapley value gives a unique solution, fairly assigning the contribution of each player (parameter) to the game score. The Shapley value determines the validity of feature i in Eq. (3.1) below:

$$\emptyset_i\,(val) = \frac{1}{N!} \sum_{S \subseteq \{x_1,\ldots,x_N\}\setminus\{x_i\}} |S|!\,(|N| - |S| - 1)![val\,(S \cup \{x_i\}) - val\,(S)] \tag{3.1}$$

where: S – is a subset of the features in the model, $val(S)$ corresponds to the model output for S, N – is the total number of features, x – the feature value for the sample to be explained, that is $x \triangleq \{x_1, \ldots, x_N\} \in R^N$.

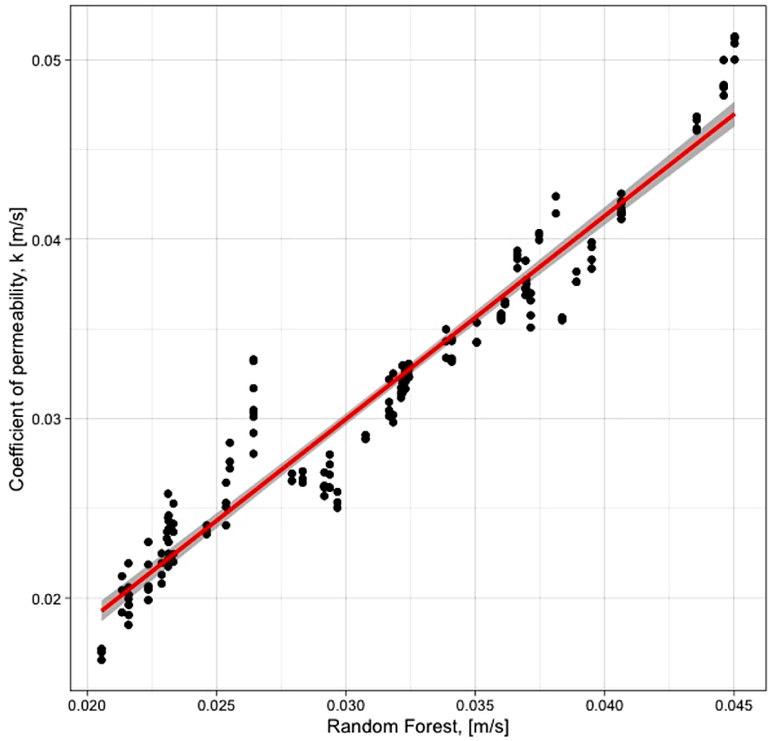

Fig. 6. Comparison of modeling and observation results

The relation $val(S \cup \{x_i\}) - val(S)$, provides the marginal contribution of x_i. These marginal contributions are assigned weights, on the various ways the subset could have been formed before adding x_i: $|S|!$ and then adding x_i: $(|N| - |S| - 1)!$. Add up all the possible sets of S, and then calculate the average of $\frac{1}{N!}$.

Shapley Additive Explanations (SHAP) allows local interpretation of the prediction by calculating each validity score of physical property characteristics and grain size characteristics for each prediction sample. SHAP is also used to derive an accurate global interpretation of the model, giving rise to its high representativeness as a post-hoc IML method (Fig. 8).

The Shapley value is a unique solution because it satisfies the axioms of symmetry, imitation and additivity [39, 41–43]. Symmetry implies that if the relative marginal contribution of the parameters in question is the same, then the Shapley value assigned to each feature value will also be the same. Imitativeness implies that if the value of a given parameter does not affect the model, then the assigned Shapley value will be zero. Finally, additivity means that the sum of the Shapley values of all feature values in the model is equal to the model output [40, 44]. As such, the total contribution of all parameter values will be equal to the impact of all feature values on the model output minus the impact in the absence of the parameter value.

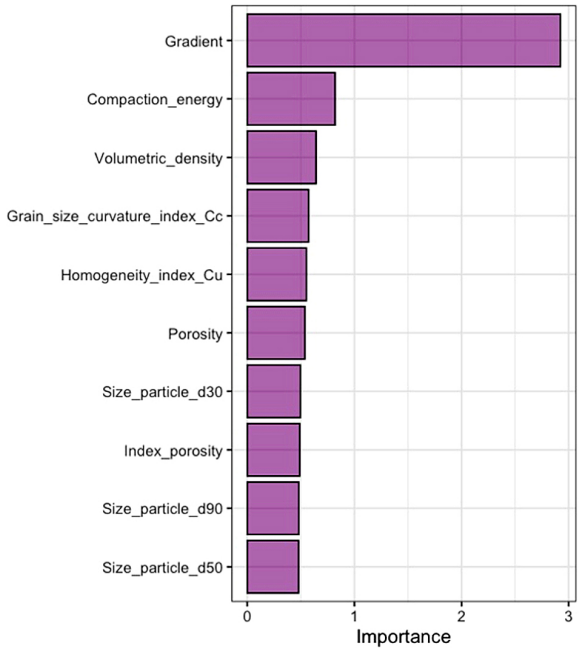

Fig. 7. Random Forest variable importance

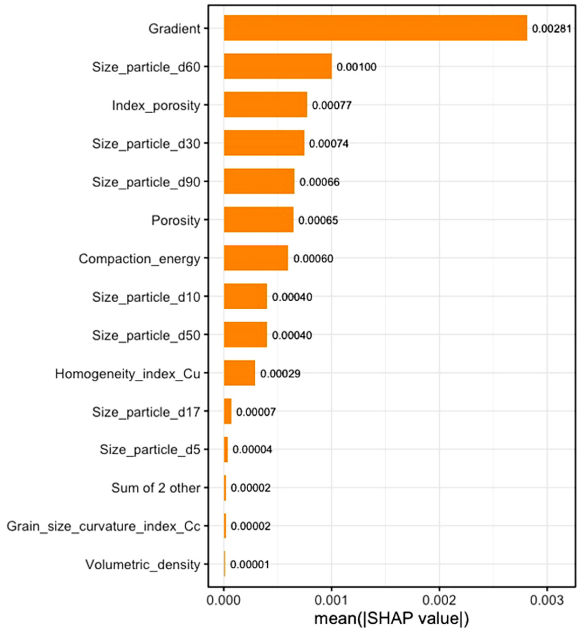

Fig. 8. Graphical analysis of results for averages of SHAP values

A graph of the prediction results of the coefficient of filtration, which is a global interpretation of the model, is shown in Fig. 8, where the average SHAP values, that is, the effect of the average feature on the output of the model, are presented. Gradient and size particle d_{60} turned out to be the most significant rank of importance of the feature parameters. Due to the local representativeness properties of SHAP, it can provide both a global importance score and an explanation of individual predictions in the SHAP summary chart, allowing for a richer visual summary, as shown in Fig. 9. In the chart, the parameters are arranged in descending order based on relative importance $\sum_{j=1}^{N} \left|\emptyset_i^{(j)}\right|$, where $\emptyset_i$ is the Shapley value of the parameter i, j is the sample, and N is the total number of samples. Each dot in the summary plot represents a sample versus its effect on the model output $\emptyset(j)$. The color of each sample represents the relative value of the parameters, i from low to high [42, 45]. Fig. 9 shows the most important parameter in the summary – the gradient. The horizontal spread of SHAP values indicates the change in parameter values. The greater the spread, the greater the change in the parameter and thus the greater the importance of the covariate. Covariates are ranked from most to least influential by their average absolute SHAP value.

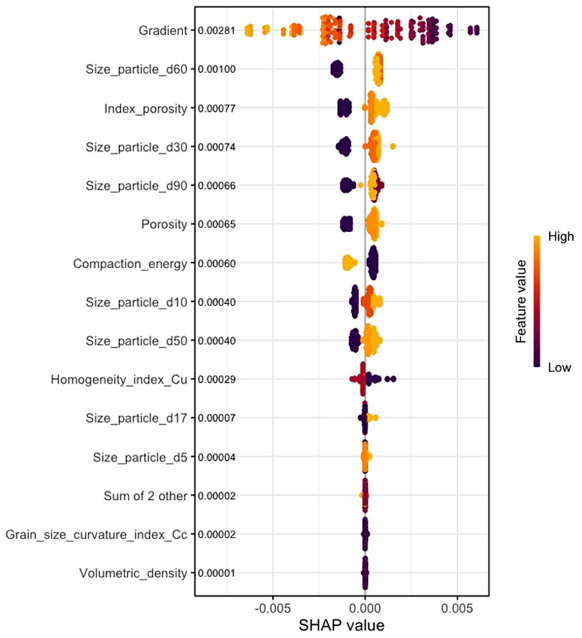

Fig. 9. SHAP value bee graph

The SHAP value in Fig. 10 for each parameter from the point of view of individual values is shown in the relevant bar. The vertical dashed line indicates the expected SHAP value. The SHAP values add up to the final model prediction. We can also look at the SHAP values from the point of view of the coefficient of filtration (Fig. 10). We can see the exact change in the filter coefficient resulting from the inclusion of each variable.

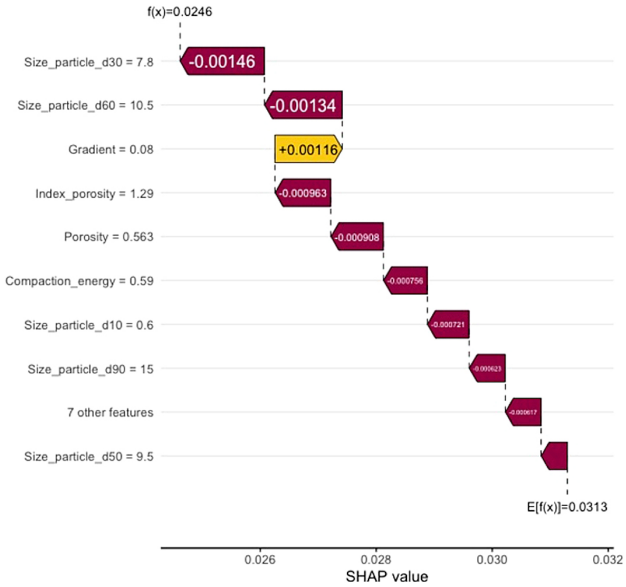

Fig. 10. Waterfall graph of SHAP value results

4. Summary

The research was aimed at expanding the knowledge of the possibility of using the Random Forest algorithm to determine the coefficient of filtration in anthropogenic aggregate – Blast Furnace Slag. Random Forest is a complex model that consists of many individual trees, and the final prediction is made by considering the output of each tree. To interpret the algorithm's prediction results well, auxiliary methods were used the SHAP. Since direct insight into the performance of the model is difficult, it is important to assist with methods to better understand the relevant parameters affecting the model. The obtained results of matching the results of the algorithm with the observations are promising (R^2: 0.84–0.92), and the resulting errors are relatively low: RMSE in the range of 0.002 to 0.0031 and MSR $4.96 \cdot 10^{-6}$ – $9.43 \cdot 10^{-6}$. It is worth further research and analysis on expanded datasets considering more types of material. Verifying the distribution of the resulting parameters using the sample tree and comparing them with the results of the relevant parameters affecting the model can help identify the parameters that are important for obtaining good-quality predictive estimates. SHAP assists in clarifying these relationships, but this does not mean that the results are easily interpretable. Machine learning algorithms remain black box models, and some of the relationships identified by the model do not guarantee that the results are interpretable based on previously known material property relationships. The combination of Random Forest and SHAP algorithms provides in-depth insight into the relationships between variables and parameters affecting the filter coefficient. This allowed us to identify the physical parameters with the greatest impact on the model, these included gradient, size particle d_{60}, index of porosity and size particle d_{30}. As previously mentioned,

further research and analysis using other algorithms as well are needed to reliably determine the usefulness of the Random Forest algorithm for predicting the filtration coefficient for anthropogenic materials.

References

[1] J. Dzięcioł and M. Radziemska, "Blast furnace slag, post-industrial waste or valuable building materials with remediation potential?", *Minerals*, vol. 12, no. 4, art. no. 478, 2022, doi: 10.3390/min12040478.

[2] G.C. Ulubeyli and R. Artir, "Sustainability for blast furnace slag: use of some construction wastes", *Procedia Social and Behavioral Sciences*, vol. 195, pp. 2191–2198, 2015, doi:10.1016/J.SBSPRO.2015.06.297.

[3] R. Trach, M. Poł ński, and P. Hrytsiuk, "Decision making in choosing a network organizational structure in integrated construction projects", *Archives of Civil Engineering*, vol. 67, no. 2, pp. 195–208, 2021, doi: 10.24425/ACE.2021.137163.

[4] J. Witkowska-Dobrev, et al., "Effect of sewage on compressive strength and geometric texture of the surface of concrete elements", *Structural Concrete*, vol. 24, no. 1, pp. 468–484, 2023, doi: 10.1002/suco.202200467.

[5] T. He, Z. Li, S. Zhao, X. Zhao, and X. Qu, "Study on the particle morphology, powder characteristics and hydration activity of blast furnace slag prepared by different grinding methods", *Construction and Building Materials*, vol. 270, art. no. 121445, 2021, doi: 10.1016/J.CONBUILDMAT.2020.121445.

[6] M. Radziemska, et al., "Recycling of blast furnace and coal slags in aided phytostabilisation of soils highly polluted with heavy metals", *Energies (Basel)*, vol. 14, no. 14, 2021, doi: 10.3390/en14144300.

[7] M. Valcuende, F. Benito, C. Parra, and I. Miñano, "Shrinkage of self-compacting concrete made with blast furnace slag as fine aggregate", *Construction and Building Materials*, vol. 76, pp. 1–9, 2015, doi: 10.1016/j.conbuildmat.2014.11.029.

[8] N.D. Lagaros, "Artificial neural networks applied in civil engineering", *Applied Sciences*, vol. 13, no. 2, art. no. 1131, 2023, doi: 10.3390/APP13021131.

[9] I. Flood and N. Kartam, "Neural networks in civil engineering. I: principles and understanding", *Journal of Computing in Civil Engineering*, vol. 8, no. 2, pp. 131–148, 1994, doi: 10.1061/(ASCE)0887-3801(1994)8:2(131).

[10] I. Flood, "Towards the next generation of artificial neural networks for civil engineering", *Advanced Engineering Informatics*, vol. 22, no. 1, pp. 4–14, 2008, doi: 10.1016/J.AEI.2007.07.001.

[11] R. Trach, Y. Trach, and M. Lendo-Siwicka, "Using ANN to predict the impact of communication factors on the rework cost in construction projects", *Energies*, vol. 14, no. 14, art. no. 4376, 2021, doi: 10.3390/EN14144376.

[12] J. Dzięcioł and W. Sas, "Perspective on the application of machine learning algorithms for flow parameter estimation in recycled concrete aggregate", *Materials*, vol. 16, no. 4, art. no. 1500, 2023, doi: 10.3390/MA16041500.

[13] M.F. Hasan, O. Hammody, and K.S. Albayati, "Estimate final cost of roads using support vector machine", *Archives of Civil Engineering*, vol. 68, no. 4, pp. 669–682, 2022, doi: 10.24425/ace.2022.143061.

[14] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] L. Breiman, "Bagging predictors", *Machine Learning*, vol. 26, no. 2, pp. 123–140, 1996.

[16] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms", *ACM International Conference Proceeding Series*, vol. 148, pp. 161–168, 2006, doi: 10.1145/1143844.1143865.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning with applications in R*. Springer, 2013.

[18] G. Louppe, "Understanding Random Forests: from theory to practice", PhD thesis, University of Liege, Belgium, 2014, doi: 10.48550/arxiv.1407.7502.

[19] E. Scornet, "Tuning parameters in random forests", *ESAIM Proceedings and Surv*, vol. 60, pp. 144–162, 2018, doi: 10.1051/proc/201760144.

[20] M.N. Wright, S. Wager, and P. Probst, "Ranger: a fast implementation of Random Forests", 2022.

[21] S.J. Wright and B. Recht, *Optimization for data analysis*. Cambridge University Press, 2022, doi: 10.1017/9781009004282.

[22] M.N. Wright and A. Ziegler, "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R", *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/JSS.V077.I01.

[23] A. Liaw and M. Wiener, "Classification and Regression by randomForest", *R News*, vol. 2, no. 3, 2002.

[24] M. Kuhn, *The caret Package*. 2019. http://jmlr.org/papers/v18/17-269.html

[25] M. Kuhn, "Building Predictive Models in R Using the caret Package", *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008, doi: 10.18637/JSS.V028.I05.

[26] R. Genuer, J.-M. Poggi, and C. Tuleau, "Random Forests: some methodological insights", Nov. 2008, doi: 10.48550/arxiv.0811.3619.

[27] R. Genuer and J.-M. Poggi, *Introduction to Random Forests with R.* Cham: Springer International Publishing, 2020, pp. 1–8. doi: 10.1007/978-3-030-56485-8_1.

[28] A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: AnEnsemble Classifier", in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 26. Springer Science and Business Media Deutschland GmbH, 2019, pp. 758–763, doi: 10.1007/978-3-030-03146-6_86.

[29] P. Probst and A.-L. Boulesteix, "To Tune or Not to Tune the Number of Trees in Random Forest", *Journal of Machine Learning Research*, vol. 18, pp. 1–18, 2018. http://jmlr.org/papers/v18/17-269.html

[30] W. Sas, J. Dzięcioł , and A. Gł uchowski, "Estimation of recycled concrete aggregate's water permeability coefficient as earth construction material with the application of an analytical method", *Materials*, vol. 12, no. 18, art. no. 2920, 2019, doi: 10.3390/ma12182920.

[31] W. Sas and J. Dzięcioł , "Determination of the filtration rate for anthropogenic soil from the recycled concrete aggregate by analytical methods", *Scientific Review Engineering and Environmental Sciences*, vol. 27, no. 2. pp. 236–248, 2018, doi: 10.22630/PNIKS.2018.27.2.23.

[32] W. Sas, et al., "Geotechnical and environmental assessment of Blast Furnace Slag for engineering applications", *Materials*, vol. 14, no. 20, art. no. 6029, 2021, doi: 10.3390/ma14206029.

[33] PN-EN 13286-2:2007 Unbound and hydraulic binder mixtures: Part 2: Methods for determining density in relation to water content. Proctor compaction.

[34] M.W. Browne, "Cross-Validation methods", *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 108–132, 2000, doi: 10.1006/JMPS.1999.1279.

[35] T. Fushiki, "Estimation of prediction error by using K-foldcross-validation", *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, 2011.

[36] S. Janitza and R. Hornung, "On the overestimation of random forest's out-of-bag error", *PLoS One*, vol. 13, no. 8, art. no. e0201904, 2018, doi:10.1371/JOURNAL.PONE.0201904.

[37] G. Martínez-Muñoz and A. Suárez, "Out-of-bag estimation of the optimal sample size in bagging", *Pattern Recognition*, vol. 43, no. 1, pp. 143–152, 2010, doi: 10.1016/J.PATCOG.2009.05.010.

[38] M. Vukovic and A. Soro, *Determination of hydraulic conductivity of porous media from grain-size composition*. Littleton: Water Resources Publications, 1992.

[39] S.M. Lundberg and S.I. Lee, "A unified approach to interpreting model predictions", in *Neural Information Processing Systems*, vol. 2017. 2017, pp. 4766–4775, doi: 10.48550/arxiv.1705.07874.

[40] L. Merrick and A. Taly, "The Explanation Game: Explaining Machine Learning Models Using Shapley Values", in *Machine Learning and Knowledge Extraction. Lecture Notes in Computer Science,* vol. 12279. Springer, 2020, pp. 17–38, doi: 10.1007/978-3-030-57321-8_2.

[41] L.S. Shapley, "A Value for n-Person Games", in *Contributions to the Theory of Games (AM-28),* vol. 2. Princeton University Press, 1953, pp. 307–318.

[42] B. Rozemberczki, et al., "The Shapley Value in Machine Learning", in *Proceedings of the 31st International Joint Conference on Artifical Intelligence, IJCAI22*. International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5572–5579, doi: 10.24963/ijcai.2022/778.

[43] M. Anjum, K. Khan, W. Ahmad, A. Ahmad, M.N. Amin, and A. Nafees, "New SHapley Additive ExPlanations (SHAP) Approach to Evaluate the Raw Materials Interactions of Steel-Fiber-Reinforced Concrete", *Materials*, vol. 15, no. 18, art. no. 6261, 2022, doi: 10.3390/ma15186261.

[44] I.U. Ekanayake, D.P.P. Meddage, and U. Rathnayake, "A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)", *Case Studies in Construction Materials*, vol. 16, 2022, doi: 10.1016/j.cscm.2022.e01059.

[45] H. Errousso, E.A. Abdellaoui Alaoui, S. Benhadou, and H. Medromi, "Exploring how independent variables influence parking occupancy prediction: toward a model results explanation with SHAP values", *Progress in Artificial Intelligence*, vol. 11, no. 4, pp. 367–396, 2022, doi: 10.1007/S13748-022-00291-5.

Szacowanie współczynnika filtracji jako przykład zastosowania algorytmu Random Forest w budownictwie lądowym

Słowa kluczowe: budownictwo cyrkularne, żużel, uczenie maszynowe, las losowy, współczynnik przepuszczalności.

Streszczenie:

Nowy światowy rekord produkcji stali surowej odnotowano w 2021 r., wzrosła wynosił 3,8% w stosunku do roku 2020. Wpłynęło to jednocześnie na ilość wytwarzanego wraz z tą produkcją żużla. Cał kowitej ilości odpadów pochodzące z produkcji przemysłowej i budowlanej w całej Unii Europejskiej stanowią aż 48%. Dlatego, gospodarka odpadami powinna zapewniać odzysk jak największej ilości zasobów. Strategie Unii Europejskiej zgodne z celami gospodarki cyrkularnej koncentrują się na zapewnieniu spójność polityki w obszarach: klimatu, efektywności energetycznej, gospodarowania odpadami z budowy i rozbiórki oraz zasobooszczędności. Żużle są materiałem, który interesuje badaczy pod kątem ich zastosowania w budownictwie. Żużle z jednej strony są materiałami coraz lepiej poznanymi z drugiej upewniamy się o niejednorodności tych materiałów. Na charakterystykę właściwości fizycznych żużli wpływa wiele czynników m.in. rozdaj pieca w jakim powstają. Skłania to do poszukiwania narzędzi pomocnych w wyznaczaniu parametrów żużli, jaki materiałów antropogenicznych. Celem badań było zweryfikowanie hipotezy, że możliwe jest wyznaczanie parametru współczynnika filtracji, istotnego do zastosowań w konstrukcjach ziemnych z wykorzystaniem algorytmu uczenia maszynowego – Random Forest. W pracy przeanalizowano dwa rodzaje materiału: żużel wielkopiecowy oraz paleniskowy. Wyniki analizy pozwoliły na otrzymanie wysokiego współczynnika determinacji (R^2) – 0.84–0.92. Pozwala to sądzić, że algorytm ten może okazać się użyteczny przy wyznaczaniu parametrów filtracyjnych w żużlach.

Received: 2023-05-07, Revised: 2023-06-30