Artificial Intelligence Algorithm for Tailing Storage Facility Soil Classification Based on CPT Measurements

Natalia Duda-Mróz, Sergii Anufriiev, Wioletta Koperska, Paweł Stefaniak
KGHM Cuprum Research and Development Centre Ltd., gen. W. Sikorskiego 2-8, 53-659 Wroclaw, Poland

Paweł Stefanek
KGHM Polska Miedź S.A., M. Skłodowskiej-Curie 48, 59-301 Lubin, Poland

ABSTRACT
Due to the high environmental risks and negative impact of a failure, tailings storage facilities (TSFs) need constant monitoring. Advanced mathematical models have been developed in the past to predict the behavior of TSFs and raise alerts if needed. To be precise and reliable, such models need a spatial distribution of soil types within the dam as an input. Getting this data from laboratory measurements is time and cost-consuming. In this article, we propose an ANN-powered algorithm, which allows us to accurately estimate the soil distribution based on a cone penetration test (CPT).

INTRODUCTION
Tailings Storage Facility (TSF) belongs to the group of large-size geotechnical objects consisting of earth embankments intended for the cost effective storage of post-flotation waste and water. In this article, we focus on the active Żelazny Most facility operating as part of the activities of KGHM's underground copper ore mines in southwest Poland (Figure 1). A typical TSF stores fine residuals from mineral processing and is usually surrounded by tailings dams. The disposed materials of a TSF are usually the materials left over from the process of separating the non-economic fraction of the ore from the valuable fraction. In an ore processing plant, the ore undergoes crushing and grinding processes. As a result, fine material is obtained, which makes it possible to obtain valuable raw materials in the flotation process. We can distinguish 3 main methods of expanding the TSF known in practice: upstream, centerline, and downstream as shown in Figure 1b. The upstream method is the most common and, at the same time, the most dangerous, so the continuous monitoring of TSFs is even more important.

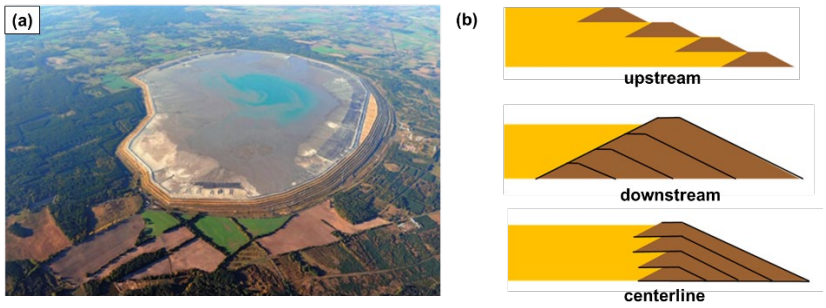

Figure 1. a) Zelazny Most TSF located in SW Poland, b) TSF construction methods: upstream, downstream, centerline.

The development of a TSF (design, expansion plan) requires knowledge of geotechnical properties, including grain size distribution, density, and mechanical and hydrogeological properties. For this purpose, geotechnical tests are used, including laboratory and field tests. Laboratory tests are very precise. They enable obtaining many geotechnical parameters. What's more, they are carried out under controlled research conditions. On the other hand, they are very expensive, which is the main reason for their limitated availability. The aim is therefore to intensify monitoring and field tests, which are very easy to conduct. Substantially, one field test provides a complete examination of the soil profile. Unfortunately, it is not possible to directly estimate geotechnical parameters on their basis. In the literature and practice, there are known applications that, based on correlations with laboratory tests, allow for a thorough examination of the structure of a geotechnical object. However, most of them concern natural soils. In the case of the Zelazny Most facility, we have dealings with tailings of anthropogenic origin. The tailings are made of waste generated at different times by 3 mines and the mined ore itself is highly inhomogeneous.

In this article, we focus on the application of various methods of classifying CPT field test data to determine the sands-to-fines ratio (SFR) parameter necessary to assess dam stability. The research material and preliminary methodology were developed as part of the IlluMINEation project (see the project website). At that time, several machine learning methods were used in the classification task (Koperska 2022). Research was also continued as part of the SEC4TD project (see the project website), where, due to the large variety of tested samples, it was decided to use deep learning methods.

PROPOSED APPROACH

The CPT test has a fixed standard process. An electric piezocone is pressed into the subsoil (see Figure 2), and at every 2 cm of depth, it measures the following parameters:

- measured cone resistance q_c,
- sleeve friction resistance f_s, and
- dynamic pore water pressure u_2.

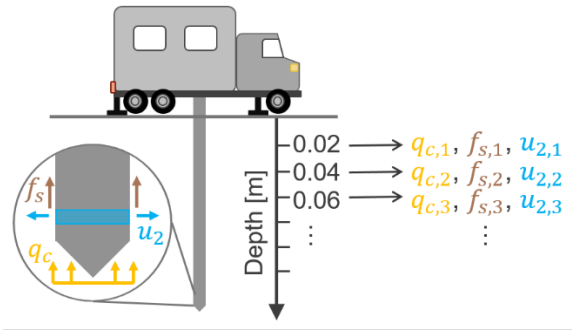

Figure 2. The diagram showing the course of the CPT test

An example of laboratory research is grain size analysis, which is used to derive the particle size distribution of soils. Samples are taken from different elevations, and their height may vary. According to ASTM standards, the test consists of preparing a

soil sample and passing it through the sieves, arranged in the following order: the coarsest one on top and the finer sieves below. Based on measured results, the grain size distribution is specified, which is a plot of soil particle diameter versus the percentage of the dry sample by weight that is smaller than that diameter. In accordence with ASTM standard, the soil classification by size consists of five types presented in Table 1.

Table 1. The five groups classified by grain size [mm] for ASTM standard.

Class	***Clay***	***Silt***	***Sand***	***Gravel***	***Cobble***
Range of grain diameter	≤ 0.005	$(0.005, 0.075\rangle$	$(0.075, 4.75\rangle$	$(4.75, 75\rangle$	> 75

After finding the grain size distribution, the coefficient of uniformity C_u and the coefficient of curvature C_c can be determined from the following formulas:

$$C_u = \frac{D_{60}}{D_{10}},$$

$$C_c = \frac{D_{30}^2}{D_{10} D_{60}},$$

where D_x is the grain size that corresponds to x percent passing.

The approach developed in the work consists of developing a flotation tailings classifier based on CPT measurements. The main objective is to estimate the SFR parameter in a more cost-effective and faster way compared to laboratory tests. The methods of classifying natural soils known in the literature use two parameters from the CPT test and specific classification thresholds. For example, methods using partition curves on two-dimensional plots can be found in Douglas 1981 and Robertson 1990. In Bhattacharya 2006, the authors present examples of the use of machine learning methods such as decision trees, ANN, or SVM. An interesting example of the general regression neural network application is presented in Kurup 2006. However, there is no guide on how to do this for tailings or other anthropological grounds. For this reason, continuing the study started in the IlluMINEation project, we are developing this research thread in order to achieve greater prediction accuracy for such difficult and complex geotechnical objects.

DATA PREPARATION

Based on grain-size distributions (see Figure 3) that were found in the laboratory and the ASTM norm, it is possible to define SFR as the proportion between coarse-grained and fine-grained particles:

$$SFR = \frac{Sand\ [\%] + Gravel\ [\%] + Cobble\ [\%]}{Silt\ [\%] + Clay\ [\%]}.$$

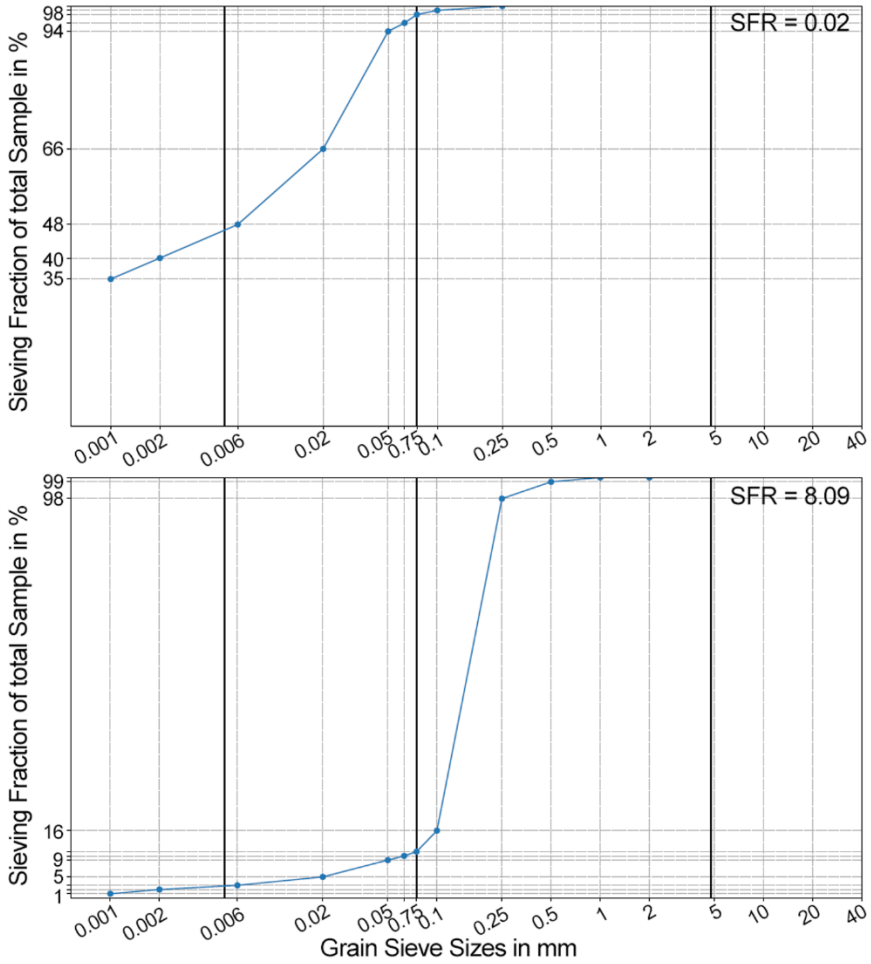

Figure 3. Examples of grain size distributions. Upper panel: low SFR value, lower panel: high SFR value.

Tailings are the partial product generated by crushing, milling, classifying, flotation, thickening, filtration, and drying processes. As reported by Stefanek 2017, tailings have the form of a slurry, in which solids represent 7 to 9% by volume, so SFR for them will have smaller values rather than greater ones, which is confirmed by the histogram presented in Figure 4 below.

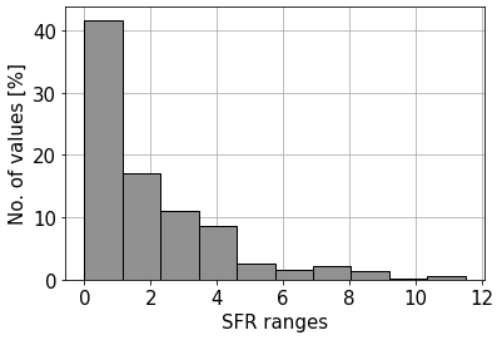

Figure 4. Histogram of SFR values.

The data contains more than 40% values with SFR less than 1.15, and other possible ranges are significantly lower. Therefore, the dataset is imbalanced, so the

Synthetic Minority Oversampling Technique (SMOTE) was used to increase the number of rest values. For that purpose, it was necessary to classify SFR, which is why the values were assigned to three groups. As a method result, the number of higher SFR samples is increased.

In the case of CPT measurements (see Figure 5), only parts that correspond to the soil samples that were taken for the grain size analysis are useful to the model. There are some cases where a small number of samples were taken, as shown in Figure 5A, so only a few readings of CPT are informative for the model. On the other hand, there are also CPT tests that have more corresponding parts, as shown in Figure 5B. Therefore, it is necessary to apply pre-processing to the whole probe for each test. Next, validated values that correspond to probes taken to the laboratory have to be aggregated into statistics.

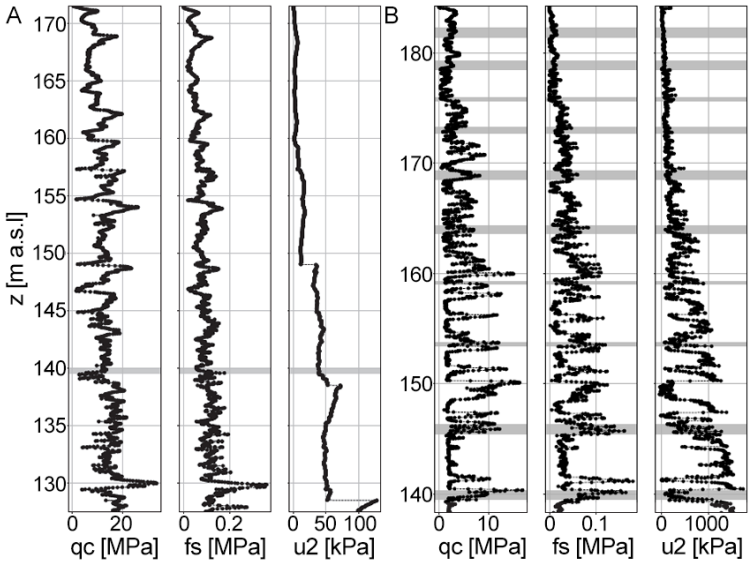

Figure 5. Examples of CPT tests measurements where A one sample B numerous samples (shadowed areas) were taken for Grain Size Analysis.

Hence, the model-independent variables are the following statistics: mean, standard deviation, minimum, maximum, and quantiles (0.05, 0.25, 0.50, 0.75, 0.95) each from q_c, f_s, u_2 measurements, and the dependent variable is SFR.

METHODOLOGY

The goal of this article is to test different machine learning approaches for the definition of SFR parameters based on the CPT measurements, with a special focus on artificial neural networks. First, we apply to our data three classic machine learning techniques: linear regression, random forest, and gradient boosting. The performance of these algorithms is then used as a baseline for the evaluation of the neural networks performance.

We have limited ourselves to a relatively simple architecture, which includes only dense layers and dropout layers, which allows us to reduce the overfitting during model training (Srivastava 2014). Even in such a simple case, selecting the proper parameters of the model can be challenging, thus, in this article, we run a Monte Carlo

style optimization of network parameters. We test one hundred structures randomly selected from a pre-defined range of parameters, listed in Table 2.

Table 2. Values of neural network parameters used during the optimization and their selected values, that assured the best performance

Parameter	Value range	Best value
Total number of dense layers	[2, 30]	8
Number of neurons in the first layer	[8, 300]	239
Total number of dropout layers	[0, 3]	1
Dropout value	[0.1, 0.5]	0.1892

At the beginning of the network, dense layers were alternated with dropout layers, until the maximum number of dropout layers was reached. For each consecutive new layer, the number of neurons equals the number of neurons in the previous layer times a random multiplier from the range [0.25, 1.25]. Such a range of multipliers allowed in general to decrease the sizes of consecutive dense layers, although some local increases are possible. Finally, some parameters were constant during optimization, such as: the optimizer (Adam), learning rate (0.001), activation function (relu), and loss function (mean squared error). Each out of one hundred randomly generated structure have been trained for 10 times in order to select the best performing model.

As was mentioned before, to increase the number of higher SFR values, the SMOTE method was applied. Oversampling by creating 'synthetic' examples was proposed by Chawla 2002. This implementation uses five randomly chosen nearest neighbors and generates one 'synthetic' sample in the direction of each of them. That causes the classifier to create larger and less specific decision regions. Blagus 2013, investigated the performance of SMOTE on high-dimensional data and concluded that it is beneficial for k-NN classifiers if the number of variables is reduced by performing some type of variable selection. This method introduces a correlation between some samples, but not between variables.

Furthermore, it is possible to classify tailings based on SFR values (see Table 2) to the cohesive ground (class I-V) and non-cohesive (class VI):

Table 3. SFR values separating tailings groups.

Class	***I***	***II***	***III***	***IV***	***V***	***VI***
Range of SFR	≤ 0.001	$(0.001, 0.6\rangle$	$(0.6, 1.5\rangle$	$(1.5, 2.5\rangle$	$(2.5, 7.4\rangle$	> 7.4

Source: Koperska et al. 2022.

RESULTS

Results from the proposed neural network model were compared with such basic models as Linear Regression, Random Forest, and Gradient Boosting Method. Before the probe was oversampled, the results from basic models (see Figure 6 and Table 3) nor from the Neural Network model were not satisfying.

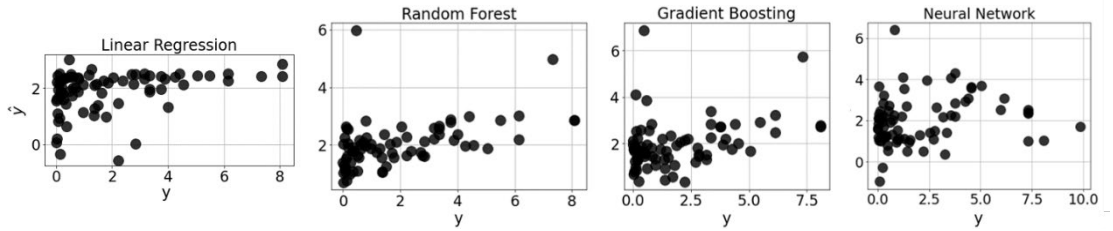

Figure 6. Actual versus predicted SFR values for the basic methods.

After the SMOTE method was used and the training set was class-balanced, as suggested in Hulse 2007, both the basic methods and neural network performance improved (see Figure 7 and Table 3). Regardless of the use of the oversampling method, the linear regression had the worst outcome. On the other hand, the greatest improvement was observed for the Gradient Boosting and Neural Network models, for which r-square increased the most. Moreover, the best results for the validation set were obtained from the model proposed by the authors.

Table 4. R-square comparison for training and validation sets and different models with and without the previously used SMOTE method.

		R-square for models			
Was the SMOTE method used?	***Set***	***Linear Regression***	***Random Forest***	***Gradient Boosting***	***Neural Network***
No	Training	0.1012	0.8715	0.8789	0.9064
	Validation	0.101	0.2111	0.0845	0.158
Yes	Training	0.3362	0.9691	0.9123	0.9453
	Validation	0.1577	0.726	0.7472	0.7688

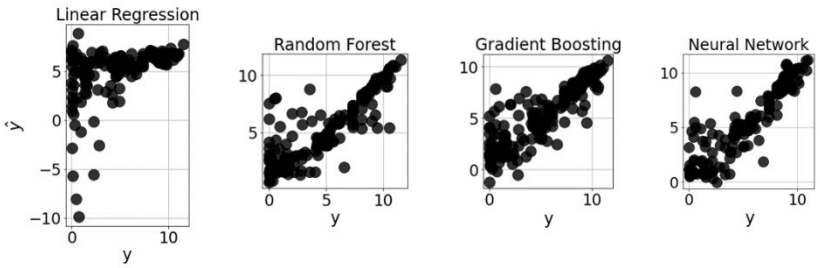

Figure 7. Actual versus predicted SFR values for the basic methods, after SMOTE method was used.

Figure 8 shows the analysis of the number of layers in the proposed NN model. The histogram is right-skewed (the median is equal to eight, which is less than the mean, that is approximately nine). The models were put into three similar size groups. The first group was for the models with 2 to 6 layers; it was noticed that the r-square was the most varied there. The other two groups were more concentrated. Despite these differences, each group has a median significantly above 0.5.

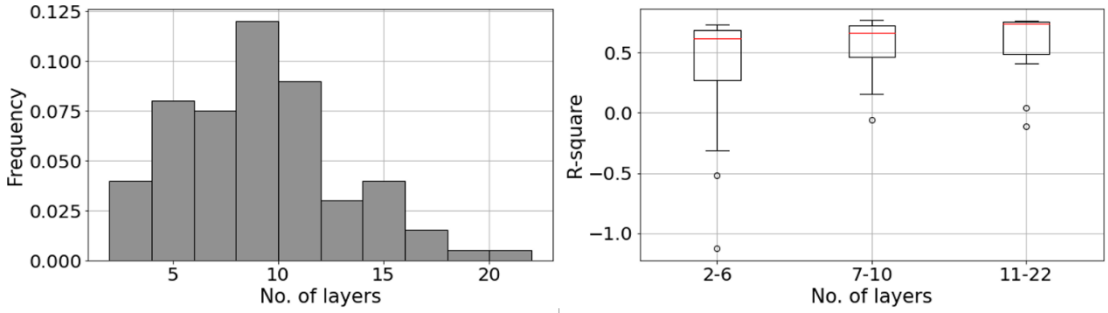

Figure 8. Left panel: histogram of the number of layers in the Neural Network model (NN). Right panel: Box-plots of R-square values for the specified ranges of the number of layers in the NN model.

Figure 9 shows the analysis of the size of 1st layer in the NN model. The histogram is uniform-shaped, so selected ranges of sizes are of similar length. The first box plot has the highest IQR value, and its median r-square is below 0.5. In the other cases, whole boxes are above that threshold, and the medians of the coefficient of determination are close to 3rd quartiles.

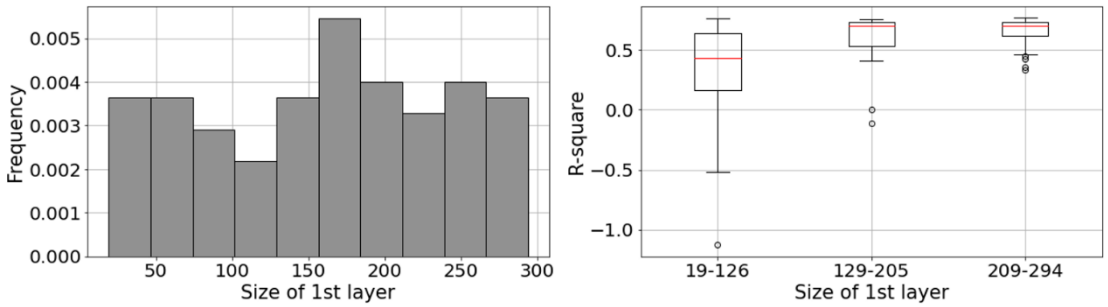

Figure 9. Left panel: histogram of the size of 1st layer in the NN model. Right panel: Box-plots of R-square values for the specified ranges of the size of 1st layers in the NN model.

CONCLUSION

In this article, a machine learning approach for the estimation of the SFR parameter for tailings based on cone penetration test (CPT) data is proposed. First, the process of data preparation and matching of CPT data and laboratory measurements is described. Then several conventional machine learning methods, such as linear regression, random forest, and gradient boosting, are applied to the data. Finally, an artificial neural network model is used. In order to select the optimal structure of the network, a Monte-Carlo optimization procedure is proposed.

Initial results obtained by all models have shown very low efficiency on the validation dataset compared to the training data, which may suggest that the models were overfitted. It may be caused by an insufficient amount of data. Besides that, the distribution of SFR values in the training sample was very right-skewed. In order to tackle these issues, the synthetic minority oversampling technique (SMOTE) has been applied to the data. It allowed for an increase in the overall number of samples, especially the number of data points with higher SFR values.

After the application of SMOTE, the results have improved significantly. All techniques besides linear regression have shown descent performance ($R^2 > 0.7$), with the neural network model working the best ($R^2 = 0.7688$). Comparison between different network structure parameters has shown that for the described problem, more complex structures performed better. Both the increase in network depth (i.e., the number of dense layers) and width (i.e., the number of neurons in the first layer) resulted in an increase in model performance.

This leads us to the conclusion that the proposed method can be further improved by applying more complex neural network architectures, such as convolutional neural networks. On the other hand, in this case, overfitting risks emerge, so for such research, more data may be necessary. We are going to study the potential of this approach in our future research.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 869379 (project IlluMINEation) and from EIT RawMaterials GmbH under Framework Partnership Agreement No 21123 (project Sec4TD).

REFERENCES

Bhattacharya, B., Solomatine, D.P., *Machine learning in soil classification*, in Neural networks, 19(2), 2006, pp. 186-195186–195

Blagus, R., Lusa, L. *SMOTE for high-dimensional class-imbalanced data*. BMC Bioinformatics 14, 106 (2013).

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., *SMOTE: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 321-357, 2002.

Douglas, B.J. and Olsen, R.S. (1981) *Soil Classification Using Electric Cone Penetrometer*. Proceedings of Conference on Cone Penetration Testing and Experience, St. Louis, 26-30 October 1981, 209-227.

Hulse JV, Khoshgoftaar TM, Napolitano A: *Experimental perspectives on learning from imbalanced data. Proceedings of the 24th international conference on Machine learning*. 2007, Corvallis, Oregon: Oregon State University, 935-942.

IlluMINEation project website www.illumineation-h2020.eu

Koperska, W., Stachowiak, M. Duda-Mróz, N. et al. *The Tailings Storage Facility (TSF) stability monitoring system using advanced big data analytics on the example of the Żelazny Most Facility*, Archives of Civil Engineering, 68(2).

Kurup, P. U., and Griffin E. P., *Prediction of soil composition from CPT data using general regression neural network* in Journal of Computing in Civil Engineering, 20(4), 2006, pp. 281-289

Robertson, P. K., *Soil classification using the cone penetration test* in Canadian Geotechnical Journal, 27(1), 1990, pp. 151-158

SEC4TD - Securing tailings dam infrastructure with an innovative monitoring system – project website sec4td.fbk.eu

Stefanek, P, Engels, J, Wrzosek, K, Sobiesak, P & Zalewski, M 2017, *Surface tailings disposal at the Żelazny Most TSF, today and into the future*, in A Wu & R Jewell (eds), Paste 2017: Proceedings of the 20th International Seminar on Paste and Thickened Tailings, University of Science and Technology Beijing, Beijing, pp. 213-225

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 15(1), 1929-1958.