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.2022.140643
ARCHIVES OF CIVIL ENGINEERING
Vol. LXVIII ISSUE 2 2022

pp. 297 –311

**Research paper**

# The Tailings Storage Facility (TSF) stability monitoring system using advanced big data analytics on the example of the Żelazny Most Facility

**Wioletta Koperska[1], Maria Stachowiak[2], Natalia Duda-Mróz[3], Paweł Stefaniak[4], Bartosz Jachnik[5], Bartłomiej Bursa[6], Paweł Stefanek[7]**

**Abstract:** Approximately 30 million tons of tailings are being stored each year at the KGHMs Zelazny Most Tailings Storage Facility (TSF). Covering an area of almost 1.6 thousand hectares, and being surrounded by dams of a total length of 14 km and height of over 70 m in some areas, makes it the largest reservoir of post-flotation tailings in Europe and the second-largest in the world. With approximately 2900 monitoring instruments and measuring points surrounding the facility, Zelazny Most is a subject of round-the-clock monitoring, which for safety and economic reasons is crucial not only for the immediate surroundings of the facility but for the entire region. The monitoring network can be divided into four main groups: (a) geotechnical, consisting mostly of inclinometers and VW pore pressure transducers, (b) hydrological with piezometers and water level gauges, (c) geodetic survey with laser and GPS measurements, as well as surface and in-depth benchmarks, (d) seismic network, consisting primarily of accelerometer stations. Separately a variety of different chemical analyses are

[1] M.Sc., KGHM Cuprum Research and Development Centre, gen. W. Sikorskiego 2-8, 53-659 Wrocław, Poland, e-mail: wioletta.koperska@kghmcuprum.com, ORCID: 0000-0002-5882-362X

[2] M.Sc., KGHM Cuprum Research and Development Centre, gen. W. Sikorskiego 2-8, 53-659 Wrocław, Poland, e-mail: maria.stachowiak@kghmcuprum.com, ORCID: 0000-0001-7501-3437

[3] M.Sc., KGHM Cuprum Research and Development Centre, gen. W. Sikorskiego 2-8, 53-659 Wrocław, Poland, e-mail: natalia.duda@kghmcuprum.com, ORCID: 0000-0002-5320-5822

[4] D.Sc., KGHM Cuprum Research and Development Centre, gen. W. Sikorskiego 2-8, 53-659 Wrocław, Poland, e-mail: pawel.stefaniak@kghmcuprum.com, ORCID: 0000-0002-1772-5740

[5] M.Sc., KGHM Cuprum Research and Development Centre, gen. W. Sikorskiego 2-8, 53-659 Wrocław, Poland, e-mail: bartosz.jachnik@kghmcuprum.com, ORCID: 0000-0002-7050-4373

[6] M.Sc., GEOTEKO Serwis Ltd., ul. Wałbrzyska 14/16, 02-739 Warszawa, Poland, e-mail: bartlomiej.bursa@geoteko.com.pl, ORCID: 0000-0001-8076-7006

[7] D.Sc., KGHM Polska Miedź S.A., M. Skłodowskiej-Curie 48, 59-301 Lubin, Poland, e-mail: pawel.stefanek@kghm.com, ORCID: 0000-0003-3357-0053

conducted, in parallel with spigotting processes and relief wells monitorin. This leads to a large amount of data that is difficult to analyze with conventional methods. In this article, we discuss a machine learning-driven approach which should improve the quality of the monitoring and maintenance of such facilities. Overview of the main algorithms developed to determine the stability parameters or classification of tailings are presented. The concepts described in this article will be further developed in the IlluMINEation project (H2020).

**Keywords:** hydrotechnics, tailing dam, data mining, risk analysis, strength parameters

# 1. Introduction

Tailings Storage Facility (TSF) is one of the largest known geotechnical facilities composed of earth embankments developed for the storage of non-cost effective, post-flotation ore and water. As a typical example of TSF, we can investigate Zelazny Most in Poland (see Fig. 1a). TSF is a structure served to store the fine residual from mining activities and it is normally surrounded by tailings dams. Tailings are the materials left over after the process of separating the valuable fraction from the non-economic fraction of an ore. Ore is crushed and milled to fine sand in the plant to enable the extraction of precious materials. We can distinguish three main construction methods usually used in tailings dam construction: upstream, centerline, and downstream as depicted in Fig. 1b below.

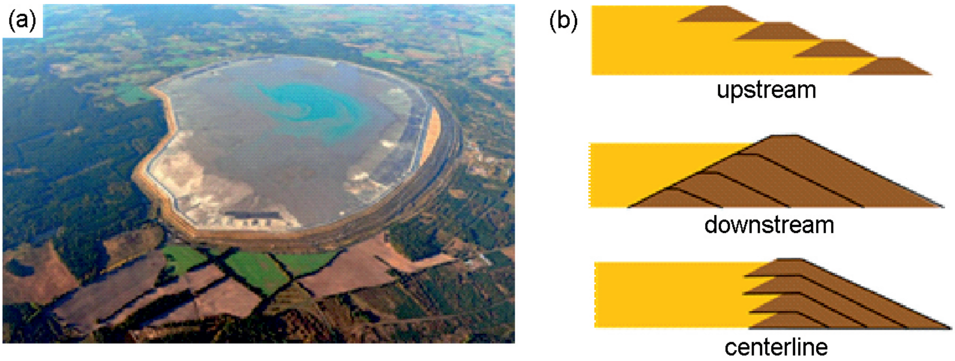

Fig. 1. a) Overview of Zelazny Most TSF. b) TSF construction methods: upstream, downstream, centerline

To design and construct tailings dams, the geotechnical properties need to be known in particular density, grain size distribution, mechanical and hydrogeological properties. The geotechnical tests can be divided into two groups: laboratory and field tests. When it comes to laboratory tests, they are more precise and many geotechnical parameters can be derived from these tests, moreover, they are carried out in well know testing conditions, which simplify the analysis. Unfortunately, they are very expensive. On the other hand, the field tests are very simple to perform. The entire soil profile can be examined in one field test. However, it is not possible to directly estimate the geotechnical parameters based on this type of survey. They need to be correlated with laboratory tests, therefore being able to obtain a good correlation between them will help to examine the large structure like Zelazny Most TSF more thoroughly. In the article, the classification methods will be presented based on the grain size distribution laboratory tests and CPT field tests. The presented algorithms are the main analytical blocks developed in the cyber–physical system as part of the Illumineation project [6]. Additionally, there are over 40,000 measurement points at the ZM, that are a part of four monitoring networks: geotechnical, hydrological, geodetic, and seismic. Ultimately, all data sources will be included in the data fusion process developed as part of Big Data analytics to support TSF's real-time stability assessment and risk prediction.

## 2. Smart stability analysis for TSF based on IOT technologies

Due to the spatial extent of TSF, the amount of recorded data in real–time, and the current analytical challenges of the managers, the natural direction was to develop a robust, multi-level IIoT platform that will incorporate cloud computing and distributed cloud management. The platform will connect using wireless communication with the physical mining world, which will be defined by an extensive, low-cost, all-embracing network of sensors. The use-case concentrates around the automation of the analytical process of the huge amount of data that is collected in the facility, which will be achieved by utilizing the machine-learning techniques to assist engineers in the data analysis and interpretation. The main problems and challenges related to the development of analytics for such huge mining areas are presented in [4]. TSF poses a serious threat to the local environment and society. We know of several structural failures from previous years that led to extensive disasters. Due to the enormous requirements in terms of safety indicators, great emphasis is placed on monitoring the TSF itself as well as its surroundings [7]. Dozens of thousands of parameters are recorded from different acquisition layers and stored in various forms in several places. Currently, there are certain operational limitations in the field of analysis and interpretation of results on an ongoing basis by the human resources. The dynamic development of the Internet of Things technology in terms of the size and performance of sensors, their integration, throughput, and wireless transmission speed have created new opportunities for the development of a cyber–physical system supporting the management of TSF in real–time [8]. The combination with online monitoring and artificial intelligence provides automation of many component analyzes ensuring the detection of anomalies, identification of spatio-temporal patterns, indirect determination of strength parameter, and finally support of the decision-making process [15, 17–19]. Fig. 2 shows the main modules feeding the analytical and decision-making process of the cyber-physical system for TSF.

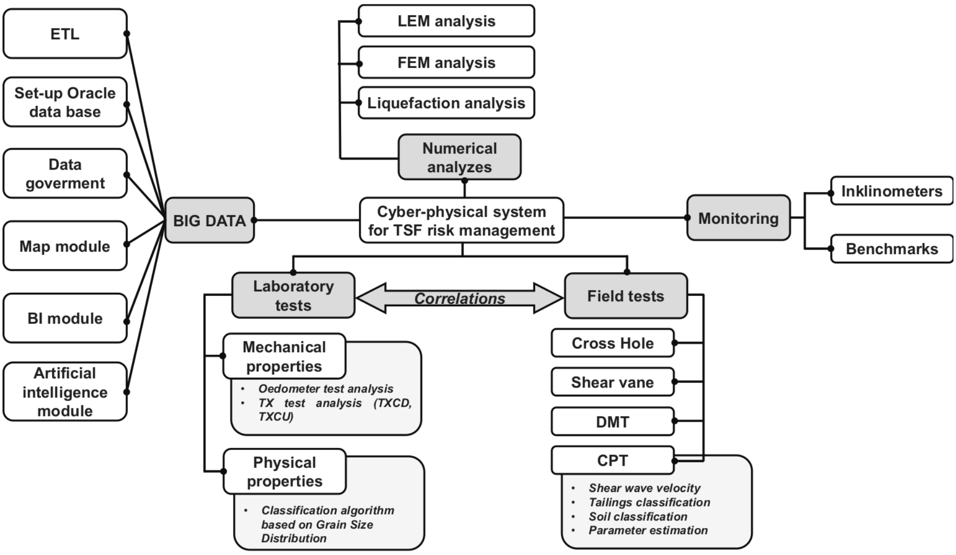

Fig. 2. Main functional modules of a cyber-physical system dedicated to the security of TSF

## 3. Piezocone penetration field test

Cone Penetration Tests (CPT) and Piezocone Penetration Tests (CPTU) containing the measurement of water pressure in the pores have been used in geotechnics for many years [1, 12]. These tests allow the determination of various soil properties, such as soil type, strength, and formability levels. CPTU is based on the introduction of the cone penetrometer into the ground surface at a constant speed – 2 cm/s. The so-called cone penetrometer is a cylindrical probe attached to the drill rod [11]. The probe is pushed from the ground using the hydraulic pushing ring or the conventional drill rig using hydraulics to the static thrust. In the case of this test, a probe with the tip area equal to 15 cm$^2$, an area of the friction sleeve equal to 225 cm$^2$, and a cone tip angle of 60°was used. The diagram of the cone penetrometer is presented in Fig. 3. The test measures the resistance of the

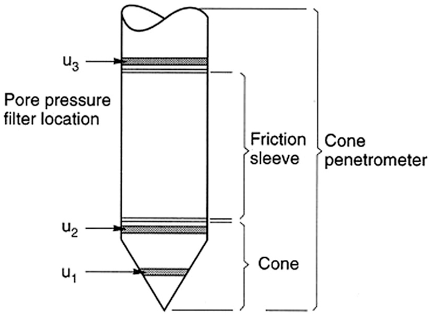

Fig. 3. The schema of cone penetrometer [10]

cone tip and the friction sleeve as well as the water pressure in the pores – in this case, the filter is located behind the cone – $u_2$. The measurement of these parameters is recorded every 2 cm. The registered resistance parameters will correspond to the type of soil and its properties, thanks to which it is possible to create an indirect method for soil classification. In addition, the water pressure in the pores allows taking into account ground moisture, which affects the strength of the soil and the slip of the cone.

## 3.1. Parameters calculation from CPTU test

The CPTU allows measuring three variables dependent on depth. The first of them is unit sleeve friction resistance ($f_s$) that is got by normalizing the measured sleeve force ($F_s$) by the area of sleeve – $A_s$ [11, 13]:

$$f_s = \frac{F_s}{A_s}$$

Similarly, for the cone tip is the next measure – unit cone tip resistance:

$$q_c = \frac{F_T}{A_T}$$

where $F_T$ is the tip force and $A_T$ is tip area. The important parameter that is dependent between the above two values of resistance is so-called the fraction ratio:

$$R_f = \frac{f_s}{q_c} \cdot 100\%$$

Moreover, it is measured the pore water pressure. In this case, the measurement is behind the cone and it is marked as $u_2$. Water allows the cone to glide more easily in the ground, so parameter $q_c$ is dependent on the pore water pressure and this variable allows for proper correction of it. Therefore, the corrected cone resistance is as follows:

$$q_t = q_c + u_2(1 - a)$$

where $a$ is strain index, in this case, it is $a = 0.75$. As it is known in the ground is the stress depends on depth. Assuming that h is the thickness of a given soil layer and $\gamma$ is unit weight depend on the depth, the total overburden stress can be described by:

$$\sigma_{v_0}(k) = \sum_{i=0}^{k} \gamma_i h_i$$

and for normalization by the pore water pressure – the effective overburden stress is as follows:

$$\sigma'_{v_0}(k) = \sum_{i=0}^{k} \gamma_i h_i - u_0$$

The parameter $u_0$ is in situ pore water pressure that can be determined from an additional dissipation test. To avoid its impact stress on parameters it can be used following normalizations:

$$Q_t = \frac{q_t - \sigma_{v0}}{\sigma'_{v0}}$$

$$q_{t_1} = \frac{q_t}{P_a} \cdot \sqrt{\frac{P_a}{\sigma'_{v0}}}$$

$$F_r = \frac{f_s}{q_t - \sigma_{v0}} \cdot 100\%$$

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}}$$

Sequentially $Q_t$ is normalized cone resistance, $q_{t_1}$ is dimensionless normalized cone resistance, $F_r$ is normalized friction ratio, and $B_q$ is pore pressure parameter. In addition, as with any test, measurements may be subject to some error and outliers may appear. Especially when multiple parameter transformations are performed, the impact of outliers may increase. That is why it is so important to analyze outliers, detect and remove them. For this purpose, it will be proposed own technique. The values are considered incorrect if there are long distances to the preceding and the following value. It also was noticed that the variance in the signal is not constant, so the average distance will be changing and the constant threshold to detection cannot be used without proper normalization. In addition, it can be assumed that the values cannot be negative. The method to detect outlier for signal $x$ (parameters $f_s$ or $q_c$) can be described by the following steps:

1. Find negative values $x < 0$ and converting them to zero.
2. Calculate the distances between point $i$ and the next value ($d_1$) and the previous one ($d_2$) as:

$$d_1(i) = |x(i+1) - x(i)|$$

$$d_2(i) = c|x(i) - x(i-1)|$$

3. Calculate the moving average distance in a window of 200 samples (D) and using it to normalize the distances $d_1$, $d_2$;
4. Check if both distances $d_1(i)$, $d_2(i)$ are higher than the set threshold H and change to the average of neighboring points:

$$x(i) = \begin{cases} x(i), & d_1(i) < H, \quad d_2 < H \\ \dfrac{x(i-1) + x(i+1)}{2}, & \text{otherwise} \end{cases}$$

The example of one of the distances with threshold and the result of outliers detection is in Fig. 4.

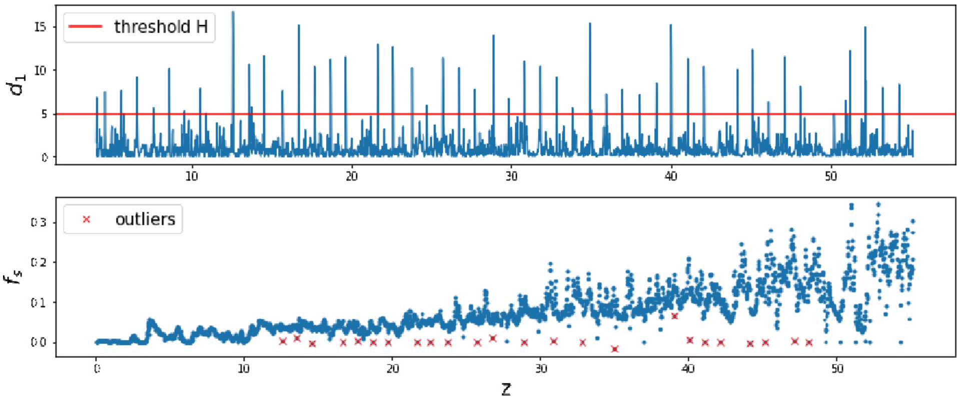

Fig. 4. The example of the outliers detection for unit sleeve friction resistance

## 3.2. Granulation analysis

The popular method to classification various types of ground is granulation analysis. This requires the grain size distribution laboratory test, so it cannot be widely used for soil classification in the field but can be a good starting point for the development of another classification method. Sieves with different hole sizes are used for the test, which allows measuring the percentage of particles of a given size. Three standards allow classification of the ground into five groups (Table 1).

Table 1. The five groups classified by grain size [mm] for three standards

| Standard | Clay | Silt | Sand | Gravel | Cobble |
|---|---|---|---|---|---|
| PN-86 B-02480 | 0–0.002 | 0.002–0.05 | 0.05–2 | 2–40 | > 40 |
| ISO | 0–0.002 | 0.002–0.063 | 0.063–2 | 2–63 | > 63 |
| ASTM | 0–0.002 | 0.005–0.075 | 0.075–4.74 | 4.75–75 | > 74 |

The content of individual soil types can be inferred by constructing the so-called grain size distribution curves. The example with boundaries for the ISO standard is presented in Fig. 5.

In addition, two important parameters can be defined using this method. The first is *SFR* which is the percentage of clay, silt, and sand divided by the percentage of gravel and cobble. The second is the percentage of clay. Mechanical research allowed the identification of six groups of tailings that can be determined using the following limits for the *SFR* parameter (Table 2). On the other hand, the cohesive ground is about $SFR < 0.7$ and non-cohesive for $SFR > 0.7$.

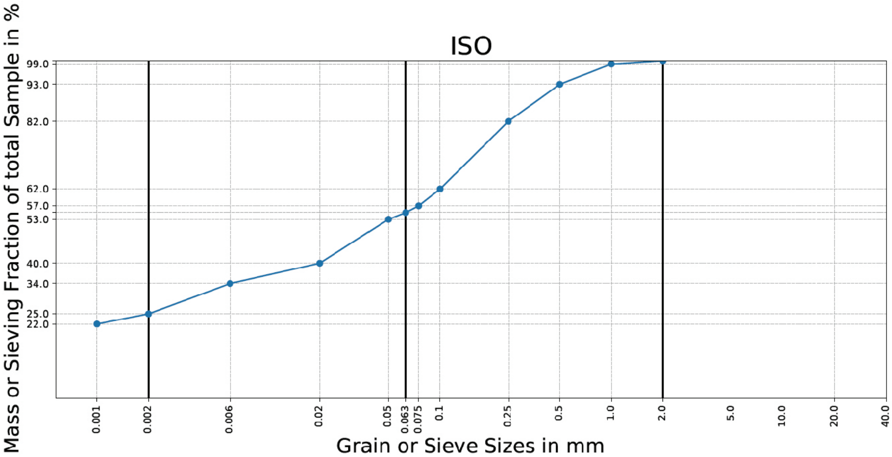

Fig. 5. The example of grain size distribution curve

Table 2. *SFR* values separating tailings groups

| I | II | III | IV | V | VI |
|---|---|---|---|---|---|
| $SFR \leq 0.001$ | $0.001 < SFR \leq 0.6$ | $0.6 < SFR \leq 1.5$ | $1.5 < SFR \leq 2.5$ | $2.5 < SFR \leq 7.4$ | $SFR > 7.4$ |

# 4. Construction of a tailings classifier model

We focus on building a tailings classifier based on CPT data to indirectly estimate strength parameters in a more cost-effective and faster way in comparison to laboratory tests. In the literature, this approach is commonly known generally for natural soils for several decades. The primary methods for classifying natural soils are based on two parameters from the CPT test and established partition limits. For example, methods using partition curves on two-dimensional plots can be found in [3, 12]. Examples of machine learning applications, as decision trees, ANN, and SVM can be found in the article [2] or general regression neural network [9]. However, there is no guide on how to do this for tailings or other anthropological grounds. With the help of the previously described field study, an attempt was made to analyze the collected statistics to build a tailings classifier model.

## 4.1. Granulation analysis

First, the characteristics of the variables were examined: their availability and ranges. After that, we calculated all the indicators needed for the analysis. As a result, we got 282.359 rows with 18 variables. Table 3 shows the availability of the individual variables.

In most cases the variables are full, the exceptions are $R_f$, *SFR* and clay. In $R_f$ case, there are unique situations where the coefficient after calculation gave the infinite value.

Table 3. Availability of data in particular variables

| | $CPT_{\mathrm{ID}}$ | $z$ | $q_c$ | $f_s$ | $u_2$ | $q_t$ | $R_f$ | $\gamma$ | $\gamma_d$ | $\sigma_{v0}$ | $\sigma'_{v0}$ | $u_0$ | $Q_t$ | $F_r$ | $B_q$ | $Q_{t1}$ | $SFR$ | $Clay$ | $q_n$ | $Class$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data absence [%] | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.007 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 96.871 | 96.871 | 0.0 | 0.0 |

The variables *SFR* and clay come from a granulation test. This information contains most of the information about the group to which the sample belongs. However, this test can only be performed rarely, hence the small number of samples. Therefore, this reduces the number of lines to 8.836, because the model can only be trained on lines that contain information from these two variables. Therefore, this reduces the number of lines, because the model can only be trained on lines that contain information from these two variables. Next, let's take a look at the areas (see Table 4). Some variables have very wide ranges of values. High values can be noticed, for example, for the variables $F_r$ and $R_f$ which are also presented in percentage units. Anomalous values correspond to errors that occurred while executing tasks and were filtered before further analysis. Moreover, it is worth adding that the assumed distribution for $u_0$ is greatly simplified. At the moment, it was impossible to consider this problem in more depth. Therefore, in the article, we will consider models based solely on variables unrelated to $u_0$.

Table 4. Variable ranges for selected data

| | $q_n$ [MPa] | $q_c$ [MPa] | $f_s$ [MPa] | $q_t$ [MPa] | $\sigma$ [MPa] | $\sigma_d$ [MPa] |
|---|---|---|---|---|---|---|
| min | –0.31 | 0.00 | 0.00 | –0.01 | 17.84 | 14.69 |
| mean | 6.05 | 6.69 | 0.13 | 6.74 | 19.96 | 16.10 |
| max | 55.92 | 56.48 | 0.51 | 56.50 | 21.58 | 17.18 |
| | $Q_t$ [–] | $B_q$ [–] | $q_{t1}$ [–] | $SFR$ [–] | $\sigma'_{v0}$ [kPa] | $u_0$ [kPa] |
| min | –1.07 | –0.28 | –0.04 | 0.00 | 11.76 | 0.00 |
| mean | 11.56 | 0.04 | 27.85 | 2.10 | 610.23 | 74.07 |
| max | 785.79 | 1.06 | 355.55 | 61.50 | 1078.61 | 161.90 |
| | $F_t$ [%] | $R_f$ [%] | $Clay$ [%] | $z$ [m] | $u_2$ [kPa] | $\sigma_{v0}$ [kPa] |
| min | –731.82 | –10746.27 | 0.00 | 0.66 | –69.47 | 11.76 |
| mean | 3.41 | –0.43 | 20.74 | 35.92 | 187.76 | 684.30 |
| max | 68.70 | 60.35 | 78.00 | 63.00 | 2016.42 | 1240.51 |

## 4.2. Selection of input parameters to the classifier

The first statistic presented is the correlation matrix, that shows the values of the correlation coefficients for the corresponding pairs of variables. It is shown in Figure 6a. As can be seen, many variables are strongly correlated with each other. This is an expected result as their mathematical formulas are often tightly intertwined. Some groups can be distinguished: (1) $z$, $\gamma$, $\gamma_d$, $\sigma_{v_0}$ and (2) $F_r$, $R_f$. This allows to significantly reduce the number of variables. Another method often used to reduce the size of a statistical dataset by discarding recent factors is principal component analysis (PCA). PCA is one of the statistical methods of factor analysis. The goal of PCA is to rotate the coordinate system in such a way as to construct a new observation space in which the most variability is explained by the initial factors. The PCA may be based on either a correlation matrix or a covariance matrix constructed from the input set. When using a covariance matrix, the fields in the input set with the greatest variance have the greatest impact on the result. Here, since the possessed variables differ widely in terms of ranges, standardization is required before calculating the PCs. The PCA reduction is performed to capture 95.0% explained variance. The first four components cover that amount of explained variance (Fig. 6b) and they are: $\gamma_d, F_r, f_s, u_2$. As mentioned before, using the expert consultations groups can be designated by *SFR*. Applying this information to data allows we can designate groups for selected data. The next step looked at the distribution of variables in individual groups (Fig. 6c). As can be noticed in the graphs below, some variables have very similar distributions. These are the highly correlated variables. Unfortunately, none of them separates the groups.

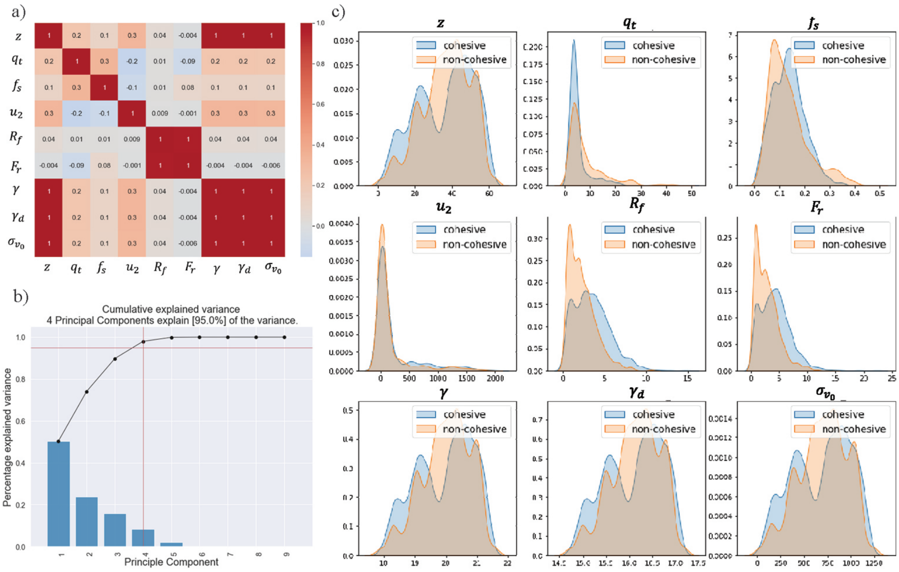

Fig. 6. Main results of exploratory analysis: a) correlation matrix for selected data. b) graph of percentage of explained variance for individual components from PCA. c) the distribution of selected variables divided into groups

## 4.3. Selection of the classifier

To examine a wide range of potential decision boundaries between individual classes, it was decided to test six different classification models: k-nearest neighbors (KNN), quadratic discriminant analysis (QDA) [16], support vector machine (SVM) [14], single classification tree, the random forest model [5] and simple neural networks. K-nearest neighbor classifier assigns the considered observation $X = x$ based on a plurality of classes of K training observations that are closest to the $x$. KNN is a completely non-parametric method, meaning that no assumptions regarding the shape of decision boundaries are being made. Therefore it should outperform many other methods when decision boundaries are highly non-linear. The QDA method is a generalization of the LDA model, which was is on finding linear combinations of features introduced into the model, to distinguish the occurring classes best way possible. But the LDA method assumes that observations from each of the classes have a Gaussian distribution, with identical covariances, which severely limits its use in the considered case. Therefore it was decided to use the QDA, being a generalization of the LDA model, that can be used also when the assumption of equal covariance is not met. Another model considered is the SVM, which enables enlarging the feature space by using different kernels. In this case, two different kernels have been tested – linear and radial (RBF – radial basis function). The possibility to select different kernels and other parameters makes the SVM model perform well in a variety of different settings. Two different tree-based methods also have been tested. The first one is a single decision tree, which tends to perform well in simple problems. Unfortunately, single, deep decision trees very often tend to overfit the training data, especially when lots of input parameters are considered. This means, that despite often having great accuracy on the training dataset, they tend to underperform on the new data. Additionally, single trees are often very non-robust, meaning that relatively small changes in the training sample can greatly impact the final estimated trees. The random forest model counteracts this phenomenon, by utilizing the predictions from many different trees. In the binary classification problem, this means, that each of the trees gives its prediction for the single observation, and then the final decision is made based on the number of predictions assigning an observation to each class (majority vote). Additionally, when building the decision trees, at each split only a random sample of predictors is chosen as split candidates, which contributes to decorrelating the trees and reducing the variance of the model. Finally, a simple neural network classification model has been considered in the article. The network structure consists of one dense layer with rectified linear unit (ReLU) activation function. All of the models have been implemented in Python programming language using the sklearn library.

## 4.4. Application to real data

After selecting the tested classifiers, they are applied to the real data. Finally, the 4948 samples were taken from the “cohesive” class, and 3771 from “non-cohesive”. It should be highlighted, that the groups are fairly evenly divided, so classifiers can be trained and tested right away. Table 5 shows the accuracy scores for selected classifiers. Each of them was tested 10 times for a random training test set and then the average value of the statistic was put into the table. Each of them used the variables $z$, $F_r$, $f_s$ and $u_2$ for the classification task. After a series of trials, these variables were selected as the most informative.

Table 5. Accuracy score for selected classifiers

| Classifier | Nearest Neighbors | Linear SVM | RBF SVM | Decision Tree | Random Forest | Neural Net | QDA |
|---|---|---|---|---|---|---|---|
| Accuracy score | 0.858 | 0.645 | 0.822 | 0.701 | 0.729 | 0.663 | 0.660 |

Summing up, the best results were obtained for the Nearest Neighbors and RBF SVM classifiers, with the first being slightly better. A more in-depth analysis of the Nearest Neighbor classifier is provided in the next table.

Table 6 Summary of the precision, recall, F1 score for each class for chosen classifier (Nearest Neighbors)

Table 6. Accuracy score for selected classifiers

| | Precision | Recall | F1 score | Support |
|---|---|---|---|---|
| Cohesive | 0.84 | 0.86 | 0.85 | 1112 |
| Non-cohesive | 0.86 | 0.84 | 0.85 | 1151 |
| Accuracy | – | – | 0.85 | 2263 |
| Macro avg | 0.85 | 0.85 | 0.85 | 2263 |
| weighted avg | 0.85 | 0.85 | 0.85 | 2263 |

Table 6 shows a summary of the precision, recall, F1 score for each class from the one run. The report includes also the macro mean (unweighted average per label) and weighted average (weighted average supporting per label). The development of the obtained classifier is planned to obtain more satisfying results. One of the considered approaches is to build a classifier based on the statistics from whole sample under the granulation test. To determine the value of *SFR*, under the laboratory tests the sample about one meter long should be taken, what corresponding at least 50 observation of the CPT parameters. This number of samples is sufficient from the statistical point of view to calculate statistics such as mean, median, or IQR. Above mentioned approach allows avoiding using repeating observations of *SFR* corresponding to measurements collected from each 0.02 m of the one–meter sample. It seems to be a promising solution, which can improve the accuracy score and precision of the new classifier.

## 5. Conclusions

The main purpose of TSF is the storage of post–flotation waste. In practice, it is a huge and complex technical object, which in the event of a geotechnical failure constitutes a serious threat to the local environment and society. Due to the serious requirements for maintaining high stability indexes with a large safety margin, a dynamic increase in monitoring and research in this field is observed. Unfortunately, in the case of large TSFs, the amount of recorded data has reached a critical point from the perspective of processing this data by geotechnical experts. The developing trend of IoT technology applications made it possible to develop a cyber–physical system that can analyze this data in real–time, estimate stability parameters, forecast risk, and further support the decision-making process. The article presents the key scope of applications of machine learning algorithms in estimating, among others, physical parameters of soil based on field tests. An example is a ground classification based on CPT surveys commonly known in natural soils. The novelty is to develop such a classifier for anthropogenic soils on the example of TSF. In this regard, a validation procedure has been proposed and thorough correlation analysis has been performed as well to recognize appropriate input vectors. In the next step, a comparative analysis of the various classifiers with their application to real data has been examined. The obtained results were presented and discussed.

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 869379.

## References

[1] I. Bagińska, W. Janecki, M. Sobótka, "On the interpretation of seismic cone penetration test (SCPT) results", *Studia Geotechnica et Mechanica*, 2013, vol. 35, no. 4, pp. 3–11, DOI: 10.2478/sgem-2013-0033.

[2] B. Bhattacharya, D.P. Solomatine, "Machine learning in soil classification", *Neural networks*, 2006, vol. 19, no. 2, pp. 186–195, DOI: 10.1016/j.neunet.2006.01.005.

[3] B.J. Douglas, "Soil classificaion using electric cone penetrometer", in *Symposium on Cone Penetration Testing and Experience*. Geotechnical Engineering Division. ASCE, 1981, pp. 209–227.

[4] H. Dudycz, P. Stefaniak, P. Pyda, "Advanced data analysis in multi-site enterprises. Basic problems and challenges related to the IT infrastructure", in *International Conference on Computational Collective Intelligence*. Cham: Springer, 2019, pp. 383–393.

[5] T.K. Ho, "The random subspace method for constructing decision forests", *IEEE transactions on pattern analysis and machine intelligence*, 1998, vol. 20, no. 8, pp. 832–844, DOI: 10.1109/34.709601.

[6] Illumineation. [Online]. Available: https://www.illumineation-h2020.eu.

[7] M. Jamiolkowski, W.D. Carrier, R.J. Chandler, K. Høeg, W. Swierczynski, W. Wolski, "The geotechnical problems of the second world largest copper tailings pond at Zelazny Most, Poland", in *1st Za Chieh-Moh Distinguished Lecture keynote speech, Proceedings of the 17th SEAGC South East Asian Geotechnical Conference, Taipei, Taiwan*, vol. 2, J.C.C. Li, M.L. Lin, Eds. Taipei, 2010, pp. 12–27.

[8] P. Kruczek, N. Gomolla, J. Hebda-Sobkowicz, A. Michalak, P. Śliwiński, J. Wodecki, R. Zimroz, "Predictive maintenance of mining machines using advanced data analysis system based on the cloud technology", in *Proceedings of the 27th International Symposium on Mine Planning and Equipment Selection-MPES 2018*, Cham: Springer, 2019, pp. 459–470.

[9] P.U. Kurup, E.P. Griffin, "Prediction of soil composition from CPT data using general regression neural network", *Journal of Computing in Civil Engineering*, 2006, vol. 20, no. 4, pp. 281–289, DOI: 10.1061/(ASCE)0887-3801(2006)20:4(281).

[10] T. Lunne, P.K. Robertson, J.J.M. Powell, *Cone penetration testing in geotechnical practice*. CRC Press, 1997.

[11] A.J. Lutenegger, *In Situ Testing Methods in Geotechnical Engineering*. Oxon: CRC Press, 2021, pp. 103-167.

[12] P.K. Robertson, "Soil classification using the cone penetration test", *Canadian Geotechnical Journal*, 1990, vol. 27, no. 1, pp. 151–158, DOI: 10.1139/t90-014.

[13] P.K. Robertson, K.L. Cabal, *Guide to cone penetration testing for geotechnical engineering*. Gregg Drilling & Testing, 2010.

[14] B. Schölkopf, A.J. Smola, R.C. Williamson, P.L. Bartlett, "New support vector algorithms", *Neural computation*, 2000, vol. 12, no. 5, pp. 1207–1245, DOI: 10.1162/089976600300015565.

[15] D.L. Skiles, F.C. Townsend, "Predicting Shallow Foundation Settlement in Sands from DMT", in *Vertical and Horizontal Deformations of Foundations and Embankments. ASCE Geotechnical Special Publications*, no. 40. 1994, vol. 1, pp. 132–142.

[16] S. Srivastava, M.R. Gupta, B.A. Frigyik, "Bayesian quadratic discriminant analysis", *Journal of Machine Learning Research*, 2007, vol. 8, pp. 1277–1305.

[17] P. Stefanek, J. Engels, K. Wrzosek, P. Sobiesak, M. Zalewski, "Surface tailings disposal at the Żelazny Most TSF, today and into the future", in *Proceedings of the 20th International Seminar on Paste and Thickened Tailings*. University of Science and Technology Beijing, 2017, pp. 213–225, DOI: 10.36487/ACG_rep/1752_24_Stejanek.

[18] K. Stefaniak, M. Wróżyńska, "On possibilities of using global monitoring in effective prevention of tailings storage facilities failures", *Environmental Science and Pollution Research*, 2018, vol. 25, no. 6, pp. 5280–5297.

[19] M. Wróżyńska, "Prediction of Postflotation Tailings Behavior in a Large Storage Facility", *Minerals*, 2021, vol. 11, no. 4, p. 362, DOI: 10.3390/min11040362.

# System monitorowania stabilności składowiska odpadów poflotacyjnych z wykorzystaniem zaawansowanej analizy big data na przykładzie obiektu Żelazny Most

**Słowa kluczowe:** hydrotechnika, zbiornik poflotacyjny, eksploracja danych, analiza ryzyka, parametry wytrzymałościowe

## Streszczenie:

W składowisku odpadów poflotacyjnych KGHM Żelazny Most składuje się rocznie około 30 milionów ton odpadów przeróbczych. Zajmujący powierzchnię prawie 1,6 tys. ha i otoczony zaporami o łącznej długości 14 km i wysokości na niektórych obszarach ponad 70 m, czyni go największym zbiornikiem odpadów poflotacyjnych w Europie i drugim co do wielkości na świecie. Z około 2900 urządzeniami monitorującymi i punktami pomiarowymi otaczającymi obiekt, Żelazny Most jest przedmiotem całodobowego monitoringu, co ze względów bezpieczeństwa i ekonomicznych ma kluczowe znaczenie nie tylko dla najbliższego otoczenia obiektu, ale dla całego regionu. Sieć monitoringu można podzielić na cztery główne grupy: (a) geotechniczna, składająca się głównie z inklinometrów i przetworników ciśnienia porowego VW, (b) hydrologiczna z piezometrami i miernikami poziomu wody, (c) geodezyjne z pomiarami laserowymi i GPS oraz jako repery powierzchniowe i gruntowe, (d) sieć sejsmiczna, składająca się głównie ze stacji akcelerometrów. Oddzielnie przeprowadza się szereg różnych analiz chemicznych, równolegle z procesami spigotingu i monitorowaniem studni odciążających. Prowadzi to do dużej ilości danych, które są trudne do analizy konwencjonalnymi metodami. W tym artykule omawiamy podejście oparte na uczeniu maszynowym, które powinno poprawić jakość monitorowania i utrzymania takich obiektów. Przedstawiono przegląd głównych algorytmów opracowanych do wyznaczania parametrów stateczności lub klasyfikacji odpadów. Do analizy i klasyfikacji odpadów wykorzystano pomiary z testów CPTU. Klasyfikacja gruntów naturalnych z wykorzystaniem badań CPT jest powszechnie stosowana, nowością jest zastosowanie podobnej metody do klasyfikacji odpadów na przykładzie zbiornika poflotacyjnego. Analiza eksploracyjna pozwoliła na wskazanie najistotniejszych parametrów dla modelu. Do klasyfikacji wykorzystano wybrane modele uczenia maszynowego: k najbliższych sąsiadów, SVM, RBF SVM, drzewo decyzyjne, las losowy, sieci neuronowe, QDA, które porównano w celu wytypowania najskuteczniejszego. Koncepcje opisane w tym artykule będą dalej rozwijane w projekcie IlluMINEation (H2020).

Received: 23.08.2021, Revised: 31.09.2021