1. Krzysztof KRÓL[1], 2. Tomasz RYMARCZYK[1,2], 3. Michał GOŁĄBEK[1], 4. Dariusz WÓJCIK[1], 5. Konrad NIDERLA[1,26], 6. Edward KOZŁOWSKI[3]

Research and Development Center, Netrix S.A (1), University of Economics and Innovation, Lublin (2), Lublin University of Technology (3)
ORCID. 1. 0000-0002-0114-2794, 2. 0000-0002-3524-9151, 3. 0000-0002-3524-9151, 4. 0000-0002-4200-3432, 5. 0000-0003-1280-0622, 6. 0000-0002-7147-4903

doi:10.15199/48.2022.02.44

Random forest method to identify seepage in flood embankments

***Abstract**. The paper presents research on the effectiveness of testing infiltration in flood embankments using electrical impedance tomography. The usefulness of the algorithm was verified and also the best results were checked. In order to test the reconstructive algorithms obtained during the research, images were generated based on simulation measurements. For this purpose, a special model of the embankment was built. In order to obtain feedback on the degree of infiltration in the flood embankment, prediction by means of the Random Forest method was used.*

***Streszczenie.** W artykule przedstawiono badania nad efektywnością badania infiltracji w wałach przeciwpowodziowych za pomocą elektrycznej tomografii impedancyjnej. Zweryfikowano przydatność algorytmu, a także sprawdzono najlepsze wyniki. W celu przetestowania uzyskanych w trakcie badań algorytmów rekonstrukcyjnych wygenerowano obrazy na podstawie pomiarów symulacyjnych. W tym celu zbudowano specjalny model wału przeciwpowodziowego. W celu uzyskania informacji zwrotnej o stopniu przesiąkania w wale przeciwpowodziowym zastosowano predykcję za pomocą metody Random Forest (**Metoda drzew losowych do identyfikacji przesiąkania w wałach przeciwpowodziowych**).*

Keywords: Random forest, tomography, seepage in flood embankments
Słowa kluczowe: Las losowy, tomografia, przesiąkanie wałów przeciwpowodziowych

Introduction

Tomographic imaging of objects provides a unique opportunity to discover the complexity of the structure without damaging the object. In many situations, there is a growing need for information about what is happening inside various objects. Therefore, observations should be carried out in a non-invasive way using tomographic apparatus. Conventional measuring instruments may not be suitable for such processes.

Investigations to identify the problem of seepage or moisture inside flood embankments cannot be carried out by ordinary destructive methods. Non-destructive methods should be used for this purpose. One of such methods is the impedance tomography method, which works very well in this type of research. What is used here is the reconstruction of what is happening inside the examined area. It can be said that the task is to define a model in which the conductivity must be calculated from the x-signals from the individual electrodes. Finite elements are used for the calculations in this work [1-7]. Many different methods are used to solve optimisation problems [8-18]. The random forest method corresponding to the finite elements in the analysed field of view can be used for the reconstruction [19-20].

Classification tree

In tomography, the task consists of reconstruction of the viewing area. In other words, the task is to define a model, where based on signals $x \epsilon R$ from electrodes, we should estimate a conductivity for the finite elements. The viewing area is created as a set of finite elements. Thus the reconstruction model can be defined as a set of classification trees corresponded to finite elements in the analysed viewing area. Let $D=\{(x_{(i)},y_i): x_{(i)} \epsilon R^m, y_i \epsilon \{0,1\}, 1 \le i \le n\}$ be the learning set for established finite element. For i-th observation, $1 \le i \le n$ the vector $x_{(i)} \epsilon R^m$ denotes the realisations of independent (input) variables, in tomography usually there are the signals obtained from the electrodes, whereas $y_i \epsilon \{0,1\}$ denotes the belonging to appropriate class to the conductivity. The tree-based method consists of partition (splitting) of the feature space into a set of separable regions and fitting constant values to appropriate regions. This method is straightforward and powerful. Below we consider a classification problem for response variable Y. Entire space of features Rm we split into $S_1, S_2, \ldots, S_k$ regions, where $S_i \cap S_j = 0$ for $1 \le i \ne j \le k$ and $S_1 \cup S_2 \cup \ldots \cup S_k = R^m$. Based on input vector $x \epsilon R$ we predict the output variable Y as follows

(1) $$f(x) = \sum_{j=1}^{k} c_j I_j(x)$$,where

(2) $$I_j(x) = \begin{cases} 1, \; for \quad x \in S_j \\ 0, \; for \quad x \notin S_j \end{cases}$$

and constant $c_i \epsilon R$ for $1 \le j \le k$ denotes the mean value of response variable when the features belong to S_j region, i.e

(3) $$c_j = \frac{1}{N_j} \sum_{x_{(i)} \in S_j} y_i$$

N_j is a length of subsequence for which the features belong to S_j region. From (1) we see, that the main task of building the classification tree is splitting the entire space of features into separated regions [23-27].

Random Forest

The decision trees have a high variance. One of the possible techniques to improve the predictions obtained from decision trees is bagging [28]. The main idea is to create an ensemble of decision trees based on several bootstrapped training sets (such a set of trees can be compared as a Concilium of independent experts). These training sets are chosen randomly with replacement from the data set and are used to train the decision trees. The variance is reduced by aggregating a set of predictions obtained from an ensemble (set) of trees. For classification trees, we take a majority vote (a class most often indicated) from the obtained predictions by each tree. The random forest is an extension of the bagging method [29], where for each tree, the training set contains a subset of predictors (features) that are randomly chosen from the complete set of predictors. Thus, we make an ensemble of random trees. A collection of random trees is called a random forest. This technique avoids the problem of selecting the dominant predictor in the split of space for each tree. As a result, the predictions obtained from trees with randomly selected

features are less correlated. It causes the prediction obtained from random forest (as average of the predictions obtained from a set of regression trees or majority vote from a set of classification trees) to be less variable and more reliable. The random forest usually is applied in the case where the set of predictors is large.

Receiver Operating Characteristic (ROC)

Below we analyse the model of the embankment, where the viewing area consists of kϵN finite elements. For each finite element, we determine the classification tree (1). Based on signal xϵR^m obtained from electrodes, we estimate $P_j(Y=1|X)$ probability that the j-finite element belong to the seepage area, 1≤j≤k. The reconstruction of viewing area is defined as a sequence $\{\hat{z}_j\}_{1\le j\le k}$ where

$$\hat{z}_j = \begin{cases} seepage & P_j(Y=1\mid X) > l \\ no\_seepage & P_j(Y=0\mid X) \le l \end{cases} \tag{4}$$

for the level lϵ(0,1).

The basic terminology and coefficients describe the quality of recognition of the viewing area. We assume: the finite element does not belong to the seepage area, and we accept as a negative case (N) (elements belong to ground), while the finite elements are included in the seepage area, we accept as a positive case (P). To estimate the quality of reconstruction, we determine the values of the confusion matrix as follows: TP (True Positive) means the number of finite elements which properly belong to the seepage area, TN (True Negative) - the number of finite elements which correctly recognised as belonging to the background, FP (False Positive) - the number of finite elements belonging to the background and have been recognised as belonging to seepage area (false alarm), FN (False Negative) - the number of finite elements belonging to seepage area and have been recognised as background. The confusion matrix is presented as follows

Table 1. The confusion matrix

	Positive	Negative
Positive Prediction	TP	FP
Negative Prediction	FN	TN

To describe the quality of seepage detection of the viewing area based on classification tree (1) we calculate the classical characteristics (ratios) [21-22]. Accuracy represents the part of the visual area that the model correctly recognised. On the other hand, it is only one possible measure that presents the correctness of recognition. In EIT, the possibilities of seepage finding in the viewing area should also be described during reconstruction. To determine the ability, we determine the Receiver Operating Characteristic (ROC curve) curve [21-22]. This curve shows the relationship between sensitivity and specificity during the reconstruction. The diagonal in the ROC curve describes a strategy based on guessing seepages into reconstruction. If the ROC is above the diagonal, the recognition technology is much better than guess. The area under the ROC curve in the literature is called AUC (Area under ROC curve) and is a measure of predictivity (predictability). In the EIT, the possibilities of seepage finding should also be described during recognition of the embankment state. To determine the ability of the classifier based on the application of the classification tree [21-22], we determine the Receiver Operating Characteristic (ROC curve) curve. This curve shows the relationship between sensitivity and specificity during the reconstruction. The diagonal of the square on the ROC curve describes a strategy based on guessing the seepages during the recognition. If the ROC is above the diagonal, it means that the recognition technique is much better than guess.

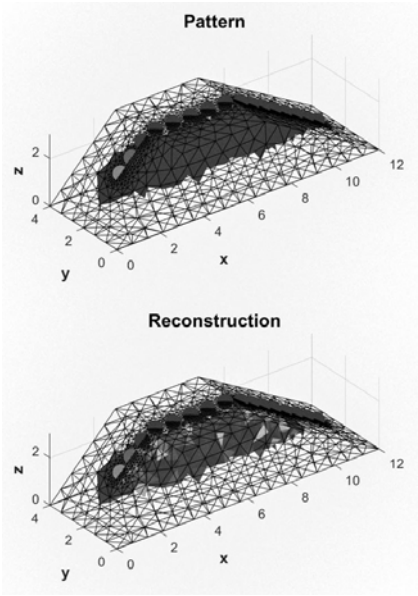

Fig.1. Example 1

Table 2. Confusion Matrix for example 1

	Positive	Negative
Positive Prediction	2909	7
Negative Prediction	20	6917

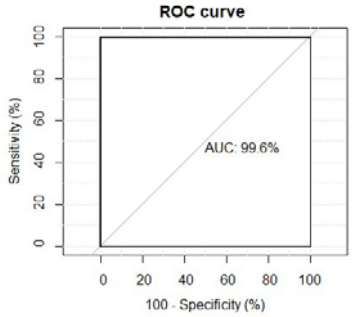

Fig.2. Classifier evaluation for example 1

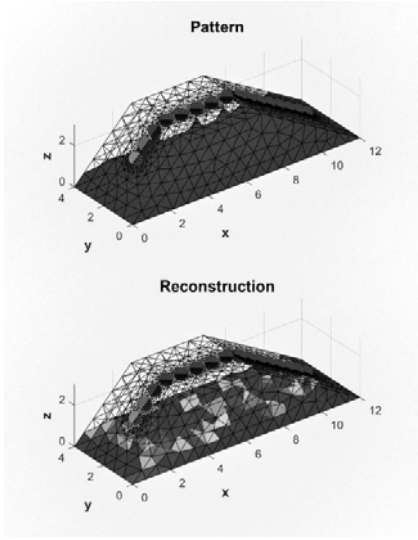

Fig.3. Example 2

Results

The model of embankment contains 16 electrodes located on the outside. This model includes 9853 finite

elements. After obtaining the signal from electrodes (signal contains 96 measurements) the main aim consists in recognition of the seepage of flood embankment. A random forest of only 20 trees was constructed for each element. The results of the reconstruction and classifier evaluation are shown in Figures 1-9.

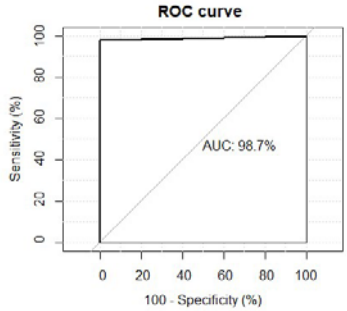

Fig.4. Classifier evaluation for example 2

Table 3. Confusion Matrix for example 2

	Positive	Negative
Positive Prediction	1150	25
Negative Prediction	27	8651

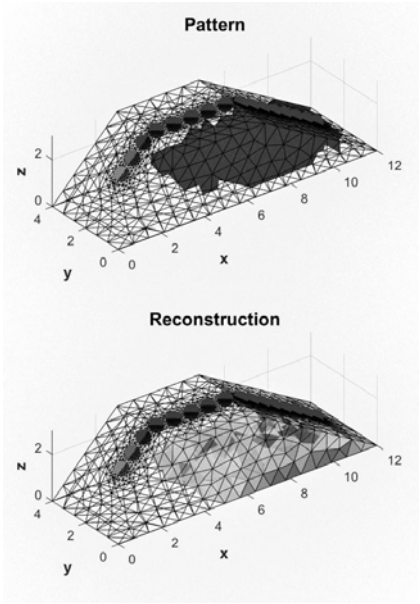

Fig.5. Example 3

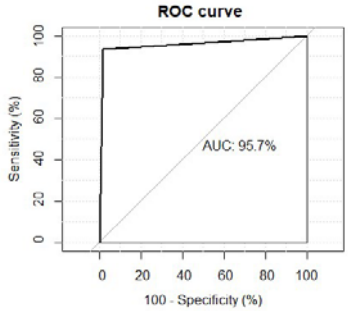

Fig.6. Classifier evaluation for example 3

Table 4. Confusion Matrix for example 3

	Positive	Negative
Positive Prediction	915	167
Negative Prediction	65	8706

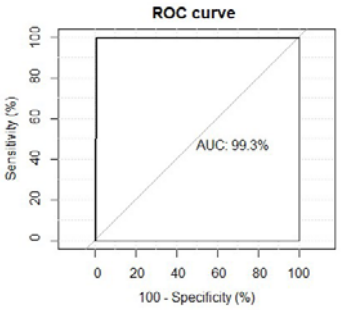

Fig.7. Classifier evaluation for example 4

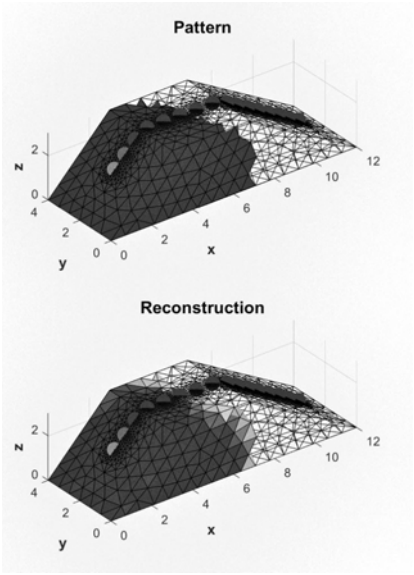

Fig.8. Example 4

Table 5. Confusion Matrix for example 4

	Positive	Negative
Positive Prediction	4474	47
Negative Prediction	20	5312

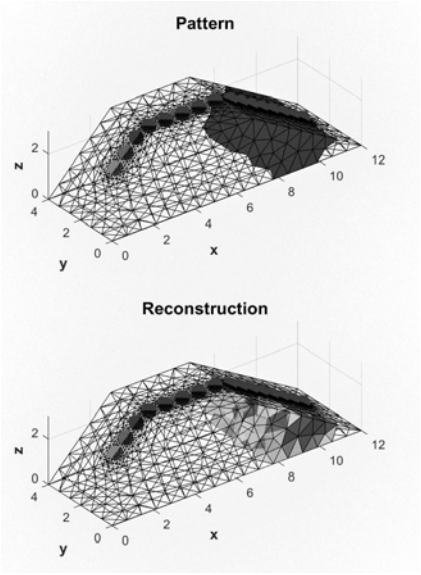

Fig.9. Example 5

Table 6. Confusion Matrix for example 5

	Positive	Negative
Positive Prediction	1876	122
Negative Prediction	66	7789

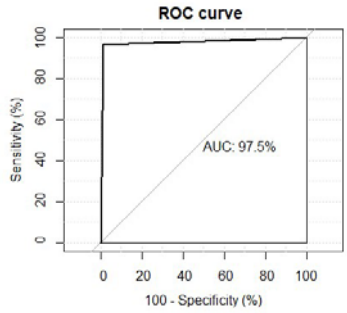

Fig.10. Classifier evaluation for example 5

Table 7. Evaluation metrices for examples

	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5
Accuracy	0.9973	0.9947	0.9765	0.9932	0.9809
Sensitivity	0.9932	0.9771	0.9337	0.9955	0.9660
Specificity	0.9990	0.9971	0.9812	0.9912	0.9846

Conclusion

The paper presents a method of leakage control in flood embankments. The Random Forest forecast was used to provide feedback on the extent of dike infiltration. The research on the effectiveness of the application of machine learning methods in electrical impedance tomography is presented. Verifying the suitability of the algorithm as well as checking the best results. Due to the high non-linearity of transformations accompanying the solution of the inverse problem, we believe that there is a high probability of success in the application of this type of algorithm.

Authors: *Tomasz Rymarczyk, D.Sc Ph.D. Eng., University of Economics and Innovation, Projektowa 4, Lublin,. E-mail: tomasz@rymarczyk.com; Krzysztof Król, Research & Development Centre Netrix S.A., Email: krzysztof.krol@netrix.com.pl; Michał Gołąbek, Research &Development Centre Netrix S.A., E-mail:michal.golabek@netrix.com.pl; Dariusz Wójcik, Research & Development Centre Netrix S.A., E-mail: dariusz.wojcik@netrix.com.pl; Konrad Niderla, University of Economics and Innovation, Projektowa 4, Lublin, Poland E-mail konrad.niderla@netrix.com.pl; Edward Kozłowski Ph.D.Eng.,Lublin University of Technology, Nadbystrzycka 38, Lublin, E-mail: e.kozlowski@pollub.pl;*

REFERENCES

[1] Rymarczyk T., Kłosowski G., Hoła A., Sikora J., Wołowiec T., Tchórzewski P., Skowron S., Comparison of Machine Learning Methods in Electrical Tomography for Detecting Moisture in Building Walls, Energies, 14(10), 2777; 2021.
[2] Rymarczyk T., Kozłowski E., Kłosowski G., Electrical impedance tomography in 3D flood embankments testing – elastic net approach, Transactions of the Institute of Measurement and Control, 42(4), 680-690, 2020.
[3] Rymarczyk T., Nita P., Vejar A., Woś M., Stefaniak B., Adamkiewicz P.: Wearable mobile measuring device based on electrical tomography, Przegląd Elektrotechniczny, 95(4), 211-214, 2019.
[4] Rymarczyk T., Kłosowski G., Tchórzewski P., Cieplak T., Kozłowski E.: Area monitoring using the ERT method with multisensor electrodes, Przegląd Elektrotechniczny, 95(1), 153-156, 2019.
[5] Koulountzios P., Rymarczyk T., Soleimani M., A quantitative ultrasonic travel-time tomography system for investigation of liquid compounds elaborations in industrial processes, Sensors, 19(23), 5117, 2019.
[6] Kłosowski G., Rymarczyk T., Kania K., Świć A., Cieplak T., Maintenance of industrial reactors based on deep learning driven ultrasound tomography, Eksploatacja i Niezawodnosc – Maintenance and Reliability; 22(1), 138–147, 2020.
[7] Kłosowski G., Rymarczyk T., Cieplak T., Niderla K., Skowron Ł., Quality Assessment of the Neural Algorithms on the Example of EIT-UST Hybrid Tomography, Sensors, 20(11), 3324, 2020.
[8] Miłak, M., Leszczyńska, A., Grudzień, K., Romanowski, A., & Sankowski, D. (2019). Slug flow velocity estimation during pneumatic conveying of bulk solid materials based on image processing techniques. Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska, 9(1), 11-14.
[9] Kłosowski G., Rymarczyk T., Wójcik D., Skowron S., Adamkiewicz P., The Use of Time-Frequency Moments as Inputs of LSTM Network for ECG Signal Classification, Electronics, 9(9), 1452, 2020.
[10] Kryszyn, J., Wanta, D., Smolik, W. T. (2019). Evaluation of the electrical capacitance tomography system for measurement using 3d sensor. Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska, 9(4), 52-59.
[11] Korzeniewska E., Krawczyk A., Stando J., Torsion field - an example of pseudo-scientific concept in physics, Przeglad Elektrotechniczny, Volume97, Issue1, Page196-199, DOI10.15199/48.2021.01.41, 2021.
[12] Korzeniewska, E; Szczesny, A; Lipinski, P; Drozdz, T; Kielbasa, P; Miernik, A, Prototype of a Textronic Sensor Created with a Physical Vacuum Deposition Process for Staphylococcus aureus Detection, SENSORS Volume: 21 Issue: 1 Article Number: 183, 2021.
[13] Wajman, R; Banasiak, R; Babout, L, On the Use of a Rotatable ECT Sensor to Investigate Dense Phase Flow: A Feasibility Study, SENSORS Volume: 20 Issue: 17 Article Number: 4854, 2020.
[14] Banasiak, R.; Wajman, R.; Jaworski, T.; Fiderek, P.; Fidos, H.; Nowakowski, J.; Sankowski, D. Study on two-phase flow regime visualization and identification using 3D electrical capacitance tomography and fuzzy-logic classification. Int. J. Multiph. Flow 2014, 58, 1–14.
[15] Jan Dusek, Jan Mikulka, Measurement-Based Domain Parameter Optimization in Electrical Impedance Tomography Imaging, Sensors 2021, 21(7), 2507
[16] Daniewski K., Kosicka E., Mazurkiewicz D., Analysis of the correctness of determination of the effectiveness of maintenance service actions. Management and Production Engineering Review 9 (2018); No. 2, 20-25.
[17] Romanowski, A. Contextual Processing of Electrical Capacitance Tomography Measurement Data for Temporal Modeling of Pneumatic Conveying Process. In Proceedings of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Poznan, Poland, 9–12 September 2018; 283–286.
[18] Chen, B.; Abascal, J.; Soleimani, M. Extended Joint Sparsity Reconstruction for Spatial and Temporal ERT Imaging. Sensors 2018, 18, 4014.
[19] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, Springer-Verlag New York Inc., 2009.
[20] James G., Witten D., Hastie T., Tibshirani R., An introduction to statistical learning, Springer-Verlag GmbH, 2013.
[21] Fawcett T., An introduction to ROC analysis, Pattern Recognition Letters. 27 (2006) 861–874.
[22] Hand D.J., Till ., R.J., A simple generalisation of the area under the ROC curve for multiple class classification problems, Machine Learning. 45 (2001) 171–186.
[23] Friedman J., Hastie T., Tibshirani R., Regularisation paths for generalised linear models via coordinate descent, Journal of Statistical Software. 33 (2010) 1.
[24] Hosmer Jr D.W., Lemeshow S., Sturdivant R.X., Applied logistic regression, John Wiley & Sons, 2013.
[25] Breiman L., Friedman J., Stone C.J., Olshen R.A., Classification and regression trees, CRC press, 1984.
[26] Kozłowski E., Mazurkiewicz D., Sęp J., Żabiński T., The use of principal component analysis and logistic regression for cutter state identification, in: Innovations in Industrial Engineering, Springer International Publishing, 2021: pp. 396–405.
[27] Antosz K., Mazurkiewicz D., Kozłowski E., Sęp J., Żabiński T., Machining process time series data analysis with a decision support tool, in: Lecture Notes in Mechanical Engineering, Springer International Publishing, 2021: pp. 14–27.
[28] Breiman L., Bagging predictors, Machine Learning. 24 (1996) 123–140
[29] Breiman L., Random forests, Machine Learning. 45 (2001) 5–32.