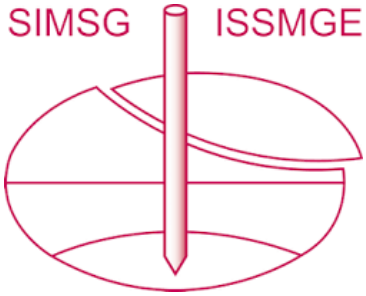
INTERNATIONAL SOCIETY FOR
SOIL MECHANICS AND
GEOTECHNICAL ENGINEERING




*This paper was downloaded from the Online Library of the International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE). The library is available here:*

*https://www.issmge.org/publications/online-library*

*This is an open-access database that archives thousands of papers published under the Auspices of the ISSMGE and maintained by the Innovation and Development Committee of ISSMGE.*

*The paper was published in the proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering and was edited by Mizanur Rahman and Mark Jaksa. The conference was held from May 1st to May 5th 2022 in Sydney, Australia.*

*Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering– Rahman and Jaksa (Eds)*
*© 2022 Australian Geomechanics Society, Sydney, Australia, ISBN 978-0-9946261-4-1*

# When artificial neural networks and genetic algorithms work together on soil embankments stability condition identification

## Quand les réseaux neuronaux artificiels et les algorithmes génétiques travaillent ensemble dans l'identification de l'état de stabilité des remblais

**Joaquim Tinoco** & António Gomes Correia
*University of Minho, ISISE, Department of Civil Engineering, Guimarães, Portugal, jtinoco@civil.uminho.pt*

Paulo Cortez
*University of Minho, ALGORITMI Research Center, School of Engineering, Guimarães, Portugal*

David Toll
*School of Engineering and Computing Science, University of Durham, Durham, UK*

ABSTRACT: The railway networks dimension associated with the budgetary constraints for maintenance purposes are among the main factors that make the management of such transportation infrastructure a big challenging task. Accordingly, aiming to support such task a data-driven model is proposed for stability condition identification of soil embankment slopes. One of the key points of the proposed system is to avoid information from complex monitoring equipment or laboratory expensive testes. Hence, only information taken from routine inspections and complemented with geometric and geologic data are used to feed the models. The learning capabilities of Artificial Neural Networks (ANN) were used to fit a data-driven model for Earthwork Hazard Category identification. For feature selection purposes, the Genetic Algorithms (GA) were applied to select the best set of input variables among the huge number of visual features collected during routine inspections. The achieved modeling results are presented and discussed, detailing GA effectiveness and ANNs performance.

RÉSUMÉ : La dimension des réseaux ferroviaires associée aux contraintes budgétaires pour l'entretien sont parmi les principaux facteurs qui font de la gestion de ces infrastructures de transport un grand défi. En conséquence, afin de répondre à cette demande, un modèle basé sur les données est proposé pour l'identification des conditions de stabilité des talus. L'un des points clés du système proposé est d'éviter les informations provenant d'équipements de surveillance complexes ou de tests coûteux en laboratoire. Ainsi, seules les informations provenant d'inspections de routine et complétées par des données géométriques et géologiques sont utilisées pour alimenter les modèles. Les capacités d'apprentissage des réseaux neuronaux artificiels (ANN) ont été utilisées pour adapter un modèle basé sur des données pour l'identification de la catégorie de risque de terrassement. Pour la sélection des caractéristiques, les algorithmes génétiques (AG) ont été appliqués pour sélectionner le meilleur ensemble de variables d'entrée parmi le grand nombre de données visuelles recueillies lors des inspections de routine. Les résultats de la modélisation sont présentés et discutés, mettant en évidence l'efficacité des GA et la performance des ANN.

KEYWORDS: slopes stability, soil embankments, soft computing, neural networks, genetic algorithms.

## 1 INTRODUCTION

Transportation infrastructures play a key role in our modern society. Indeed, countries often invest in keeping or enhancing their transportation network, aiming at a safer and more functional infrastructure. Particularly for developed countries that already have a very well complete transportation network, the main challenge today is how to keep it operational under all conditions. From the point of view of a transportation network management, a key issue is how to identify the critical parts of the network that require budget allocation for their maintenance or repair. Therefore, and in order to optimize the available budget, it is important to have a set of tools that help decision makers to identify such critical network parts and then make the best decision about how to apply the available budget. In the framework of transportations networks, namely the railways, slopes are perhaps the element for which their failure can have the strongest impact at several levels. As such, it is important to develop new ways to identify potential problems before they result in failures.

Although there are some models and systems to detect slope failures (Pourkhosravani & Kalantari 2011, Suchomel et al. 2010, Sivakumar Babu & Murthy 2005, Pinheiro et al. 2015, Power et al. 2016) most of them were developed for natural slopes, presenting some constraints when applied to engineered (human made) slopes. Moreover, they have limited applicability at network level as most of the existing systems were developed based on specific case studies or using small databases (Cheng & Hoang 2016, Cheng et al. 2012, Kang et al. 2017). Furthermore, some of the existing systems requires data from complex tests or from expensive monitoring systems, which can represent an important limitation for their application.

From a point of view of large-scale network management, these and many other approaches found in the literature are confined to a limited applicability domain and dependent on information that is sometimes difficult and expensive to obtain. Moreover, the prediction of whether a slope will fail or not is often a complex multi-variable modelling problem that is characterized by a high dimensionality. Accordingly, aiming to overcome these types of limitations and help on decision making process concerning to large-scale network management, a first attempt was recently made (Tinoco et al. 2018) by taking advantage of the learning capabilities of flexible Data Mining (DM) algorithms, such as the Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). Although the achieved

results are yet far away from the desired performance, this first iteration allow to measure how DM algorithms perform on slopes stability data and design new ways to improve the achieved results.

One of the observations taken from this first study was that ANNs works better than SVMs in slope stability identification based on visual features. It was also observed that the application of a resampling approach to overcome the problem of imbalance data inherent to slope stability data (since most of the slopes are in a good conditional condition) not always evidenced significant gains, but in some cases can help DM algorithms to learn better the problem. Moreover, it was also observed that the problem of slope stability identification is controlled by core set of features. As discussed in Tinoco et al. (2018), and using recall metric as reference, in all cases a very high performance was achieved for class "A". Also, for class "B" a good performance is observed. However, for class "C" and particularly for class "D" for which the probability of failure is higher, the achieved results are not so good as desired (a detailed analysis is latter presented in section 4 "Main results").

Keeping in mind the same principle that guide the first experiments, i.e., the use of information that can be easily obtained through visual routine inspections, a new attempt was made aiming to improve model's performance in stability condition identification of rock cuttings. Thus, in this new iteration, it was combined the non-linear learning capabilities of ANNs and the optimization power of Genetic Algorithms (GA). It is expected that by using a selected set of features for models training, a better performance will be achieved. According, GA were implemented in order to find the best set of attributes (among more than fifty variables related with data collected during routine inspections) that maximizes the overall performance in slope stability identification.

## 2 DATA CHARACTERIZATION

As previously mentioned, in this work ANNs and GA are combined in the development of new models for stability condition identification, from this point referred to as EHC (Earthwork Hazard Category, (Power et al. 2016), of rock cuttings.

The EHC system comprises 4 classes ("A", "B", "C" and "D") where "A" represents a slope with good stability condition and "D" a slope with bad stability condition. In other words, the expected probability of failure is higher for class "D" and lower for class "A". For model's developments, a database with 25673 records was compiled containing information collected during routine inspections and complemented with geometric, geological and geographic data of each slope. The data were gathered by NetworkRail workers and are concerned with the railway network of the UK. For each slope a class of the EHC system was defined by the NetworkRail Engineers based on their experience, which will be assumed as a proxy for the real stability condition of the slope for year 2015.

Figure 1 depicts the distribution of EHC classes. It shows a high asymmetric distribution (imbalanced data) of the records for each EHC class. For example, more than 63% of the embankments are classified as "A" and only 2.5% belongs to class "D". Although this type of asymmetric distribution, where most of the slopes present a low probability of failure (class "A"), is normal and desirable from the safety point of view and slope network management, it can represent an important challenge for DM models learning. In fact, when approaching imbalanced classification tasks, where there is at least one target class label with a smaller number of training samples when compared with other target class labels, the simple use of a DM training algorithm will often lead to data-driven models with better prediction accuracies for the majority classes and worst classification accuracies for the minority classes. Thus, techniques that adjust the training data in order to balance the output class labels, such as Oversampling and SMOTE, are commonly used with imbalanced datasets. Oversampling is a simple technique that randomly adds samples (with repetition) of the minority classes to the training data, such that the final training set is balanced. SMOTE is a more sophisticated technique that creates "new data" by looking at nearest neighbours to establish a neighbourhood and then sampling from within that neighbourhood. It operates on the assumptions that the original data is similar because of proximity. In the framework of SMOTE approach, and concerning to the nearest neighbour $k$ value, several values were tried and at the end a $k$ = 3 was adopted, which lead to the best overall performance.

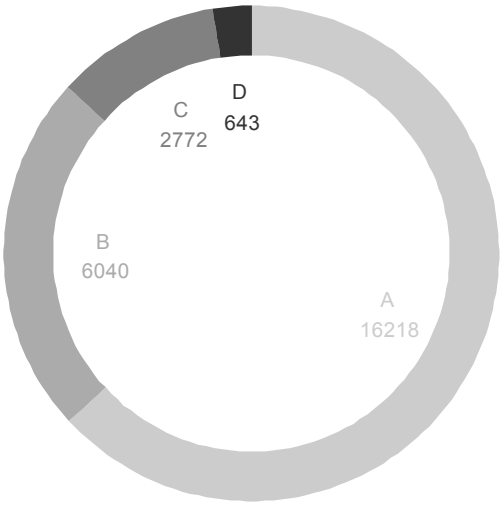



Figure 1. Soil Embankments slopes data distribution by EHC class.

For modelling purposes, 53 variables normally collected during routine inspections as well as geometric, geographic and geological information were considered as models attributes. A complete list of all variables considered in each study can be found on (Tinoco et al. 2018).

## 3 METHODOLOGY

### 3.1 *Artificial neural networks*

One of the main observations taken from the first experiments, which are summarized on Figure 5, was that the best overall performance was achieved by ANN algorithm. Accordingly, in this new attempt only ANNs were trained for EHC prediction.

In fact, ANNs are one of the most successful DM algorithms to solve complex problems in different knowledge domains (Liao et al. 2012, Garb et al. 2014) including in civil engineering (Tinoco et al. 2014a, Tinoco et al. 2014b). Even in slope stability analysis there are also some examples of ANNs applications (Yao et al. 2008, Cheng et al. 2012).

In few words ANNs are learning machines that were initially inspired in functioning of the human brain (Kenig 2011). The information is processed using iteration among several neurons. ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled in order to find patterns. This technique is capable of modelling complex non-linear mappings and is robust in exploration of data with noise. In this study the multilayer perceptron that contains only feedforward connections, with one hidden layer containing H processing units, was adopted.

Because the network's performance is sensitive to H (a trade-off between fitting accuracy and generalisation capability), we adopt a grid search of {0; 2; 4; 6; 8} under an internal (i.e.,

applied over training data) 3-fold cross validation during the learning phase to find the best H value. Such grid search only considered training data, dividing it into fitting (70%) and validation data (30%), where the validation error was used to select the best H. In other words, a different model (under a cross validation approach) was trained for each H value. Then the best model is selected according to the lowest validation error. After selecting the best H value, the ANN is retrained with the whole training data. The neural function of the hidden nodes was set to the popular logistic function $1/(1+e^{-x})$.

### 3.2 *Genetic algorithms*

The use of heuristic search algorithms, such as Genetic Algorithms (GA), for feature selection purposes in a data driven project is a common approach (Chandrashekar & Sahin 2014). In fact, when compared with exhaustive search methods that can become computationally intensive for larger datasets (Chandrashekar & Sahin 2014), iterative algorithms such as sequential search or evolutionary algorithms (e.g., GA) (Golberg 1989) tend to obtain interesting optimization results under a reasonable usage of computational resources.

As above described in section 2, the number of input features available to feed models is significant (more than 50 variables). Since some of these features might be scarcely relevant, feeding the learning models with all these attribute features will increase the model complexity and reduce its predictive accuracy, as shown in the first attempt carried out (Tinoco et al. 2018). However, the search space for all input combinations is too large. Thus, in this work we applied a GA in order to search for the best set of input features that optimize EHC prediction. These modern and flexible optimization methods come forward as some of the most effective optimization tools, on account of their efficiency when dealing with extensive search spaces within reasonable computational effort, as well as their ease of interpretation and implementation. Based on artificial intelligence and natural selection processes, GA start by generating random solutions for a problem, which are iteratively improved, tending towards optimal or near-optimal solutions. Thus, GA can be used to find the subset of features (Sun et al. 2006, Alexandridis et al. 2005) wherein the chromosome bits represent if the feature is included or not. The global maximum for the objective function can be found which gives the best suboptimal subset. Here again the objective function is the predictor performance.

In terms of optimization objectives, GAs are not limited to a single objective optimization. This is an important point, because often there is not a single optimal tradeoff solution, but rather a set of trade-offs with conflicting objectives. In the context of slope stability identification, it is important to achieve a high performance for each one of the four classes of the EHC classification system and not only for a given class. This aspect is of paramount relevance due to the asymmetric distribution of the data, which could lead to an overall high performance just based on a high accuracy for class A (most common class). Thus, a much natural multi-criteria optimization approach is to optimize a Pareto front of solutions, where each solution is called nondominated, or Pareto optimal, if none of the objectives can be improved in value without worsening the other (Bonissone et al. 2009). In the context of slope management, all Pareto optimal solutions can be considered equally and the main choice criteria for selecting one solution over the other can be decided by the decision maker or project manager based on the characteristics of the via (e.g., urban or rural location) where network of slopes is related.

Considering that Pareto front multi-optimization requires the tracking of a population of solutions, populations based meta-heuristics such as evolutionary multiobjective optimization (EMO), have become natural and popular solutions. Evolutionary computation methods, including GA and EMO, work by maintaining a population of individuals (potential solutions), where a chromosome denotes individual data representation of a solution and gene is a value position in such representation. In this study, each chromosome is a sequence of 1 and 0 genes, each indicating if a feature is included (1) or not (0).

This work adopts a popular EMO search engine, the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) algorithm (Chou & Tseng 2011), to tackle the multicriteria optimization problem related to slope stabilization identification. The NSGA-II algorithm was adopted due two main reasons. Firstly, NSGA-II is a popular and standard method for multicriteria evolutionary optimization. Secondly, NSGA-II is easily available for a computational use in the R statistical tool (Team, 2021) via the package mco (Mersmann et al. 2014), which is the same tool adopted for the ANNs implementation.

The default parameterization of NSGA-II method, as implemented in the R tool, was used, namely: population size of 100, stop after 100 generations and crossover probability of 0.7. The mutation probability was set to 0.2. Given that in the mco package of the R tool adopts a real value representation for the NSGA-II method, all genes were first rounded to the nearest integer (1 or 0) as the first step of the fitness function.

### 3.3 *Models evaluation*

For model's evaluation and comparison, three classification metrics were adopted: recall, precision and F1-score. The recall measures the ratio of how many cases of a certain class were properly captured by the model. In other words, the recall of a certain class is given by $TruePositives/(TruePositives + FalseNegatives)$. On the other hand, the precision measures the correctness of the model when it predicts a certain class. More specifically, the precision of a certain class is given by $TruePositives/(TruePositives + FalsePositives)$. The F1-score was also calculated, which represents a trade-off between the recall and precision of a class. The F1-score corresponds to the harmonic mean of precision and recall, according to the following expression $2 \cdot ((precision \cdot recall)/(precision + recall))$. All three metrics can range from 0% to 100%, where the higher the value, the better the predictor.

The generalization capacity of ANNs models was accessed through a 5-fold cross-validation approach under 5 runs (Hastie et al. 2009), excepting during the "Optimization" step where only a 3-fold cross-validation approach under 2 runs was adopted instead. Also, the three prediction metrics are always computed on test unseen data (as provided by k-fold cross-validation procedure).

### 3.4 *Modelling*

In terms of modelling procedure, EHC prediction was approached following a nominal classification strategy, which was those that showed the best overall performance during the first study experiments (Tinoco et al. 2018). Moreover, as above underlined, in this new attempt only ANNs algorithm were trained, which learned best EHC problem. In order to address the problem of imbalance data and considering that during the first study no conclusive observation was taken about which resample approach works better, SMOTE and Oversampling as well as no resampling were tested also in these new experiments (only during modelling step). Thus, the adopted procedure comprises two main steps as following described:

1. Optimization: GA are applied in order to find the best set of variables that minimizes the objective function. Accordingly, GA starts to randomly define an initial population were each individual (sequence of 1 and 0 indicating if a feature is included or not) represents a possible solution (set of variables) to the problem. The

objective function corresponds to the maximization of the recall metric for all four EHC classes (multiobjective problem). The fitness function corresponds to a predictive model based on ANN algorithm. For that, two ANNs were fitted to the database following the methodology described in "Artificial Neural Networks" section (i.e., feedforward network with one hidden layer, logistic function for hidden nodes, grid search of $\{0; 2; 4; 6; 8\}$ for H definition under a 3-fold cross-validation schema as validation procedure). Then, for each EHC class is calculated recall metric (average of both ANNs), which in turn is then used by the GA to optimize the best set of attributes toward EHC identification.

2. ANN training: After the optimization, for each Pareto-optimal solution (set of optimal variables), five ANNs are fitted to the database according to the methodology described in "Artificial neural networks" section. Here, in addition to the normal approach (no resampling), ANNs were also trained using a resampling database according to the SMOTE and Oversampling approaches. This means that for each Pareto-optimal solution, $3 \times 5 \times 5 = 75$ ANNs are trained.

As further analysed and discussed, all these experiments allow us to understand how many variables are in average used for EHC prediction and which of them are more frequently selected by the GA. At the end, considering the overall performance of each model (measured by recall, precision and F1-score for all four EHC classes) as well, a single model was selected and analyzed in detail.

All experiments were conducted using the R statistical environment (Team 2021) and supported through the *rminer* (Cortez 2010) and *mco* (Mersmann et al. 2014) packages, which facilitates the implementation of ANNs and NSGA-II algorithms, as well as different validation approaches such as cross-validation.

## 4 MAIN RESULTS

Figure 2 depicts the trade-off between recall value for each one of the four EHC class for all Pareto-optimal solutions. As expected, the best solution is always a performance commitment between all EHC classes. This means that, for example, a model with a very high performance for class "A" will perform lower on the remains classes, particularly on classes "C" and "D". The same behaviour is also observed for the other possible EHC classes combination, although in some cases, namely between "B" and "C" classes, such commitment is not so evident.

Figure 2 also shows the lower performance of all Pareto-optimal solutions for classes "B", "C" and "D" when compared to class "A".

In addition to the main objective subjacent to the application of GA, that is, feature selection toward an efficient EHC prediction, it was also intended to identify the number of inputs required for an accurate identification of slope stability, as well as which variables should be taken as models attributes.

Accordingly, Figure 3 shows that any Pareto-optimal solution have considered less than 5 features and no more than 37 inputs. In fact, the most frequent number of attributes was 15, which happened on 12 situations. It is also observed that most of the Pareto-optimal solution have considered between 14 and 16 features or between 33 and 35.

From Figure 4, that shows the 25 more frequently used variables by any of the Pareto-optimal solutions, it is observed that none of the 53 features have been considered as model attributes by all Pareto-optimal solutions. "SR" was the most frequently used feature, have been selected in 97 of all Pareto-optimal solutions. It is also observed that only a restrict group of 14 features were selected by 50% of all Pareto-optimal solutions.

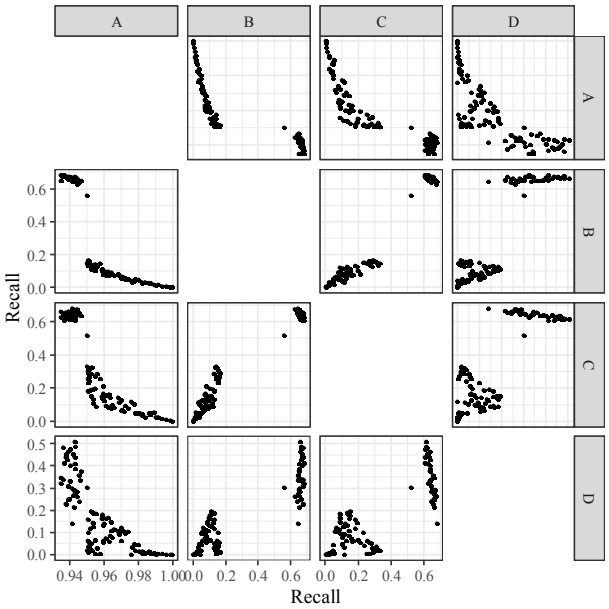



Figure 2. Correlation matrix of recall for all Pareto-optimal solutions.

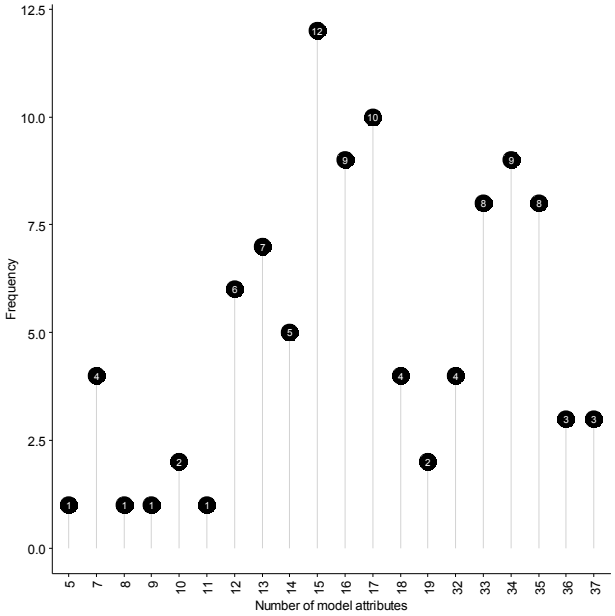



Figure 3. Number of features used as models' attributes over all Pareto-optimal solutions.

Based on recall, precision and F1-score metrics, a single model was selected among all Pareto-optimal solutions according to their performance over all EHC classes. It should be noted that, as described on section "3.4 Modelling", all Pareto-optimal solutions were retrained following a SMOTE and Oversampling approach as well as with no resampling (Normal). Below are presented and discussed the achieved performance of such Pareto-optimal solution.

Figure 5 compares ANN models from the first attempt (from now referred as "Old", Tinoco et al. 2018) and the best ANN model selected among all Pareto-optimal solutions (from now referred as "Updated"), based on recall, precision, and F1-score metrics. From its analysis, it is observed that although in overall the difference between the "Old" and "Updated" models are residual, there are some important advantages of the "Updated" models when compared with the "Old" ones. One on hand, and taken recall metric as reference, the "Updated" models, particularly those following a SMOTE approach, present a

considerable improvement on class "D" prediction. On the other hand, the number of features used by this "Updated" model is considerably low. Only 35 variables were considered instead of the 53 of the "Old" models. These two aspects represent an important achievement from a practical point of view. First, to achieve the same overall performance the required information is lower. Second, a better performance on class "D" prediction is reached with almost no loss of performance for the remaining classes.

Figure 4. Number of times that each feature is selected as model attribute over all Pareto-optimal solutions.

Figure 6 show the relation between observed and predicted EHC classes according to the "Updated" ANN model following a SMOTE approach. From its analysis, it is clear the high performance for classes "A" and "B". However, one can also observe that when a given slope is not correctly predicted, it is classified as belonging to the closest class. For example, when a slope of class "D" is not predicted as "D", in most of the cases is classified as "C".

## 5 FINAL REMARKS

Following a first attempt on stability condition identification of soil embankments through the application of soft computing techniques and considering as models attributes information usually collected during routine inspections (visual features), a new attempt was made by combining Artificial Neural Networks (ANNs) and Genetic Algorithms (GA) for features selection purposes. With this approach, although the overall performance had remained almost the same, the new models present some important advantages. On one hand, they can achieve the same overall performance by considering significantly less input variables. On the other hand, a higher performance is achieved particularly on class "D" prediction, which present a higher probability of failure, without compromise significantly the performance for the remaining classes.

This study underlines once more the high complexity involved on slopes stability condition identification, namely by considering predominantly visual features, which represent a challenging task particularly for developed countries that need to keep operational large-scale transportation network under limited budgets for maintenance and network operation.

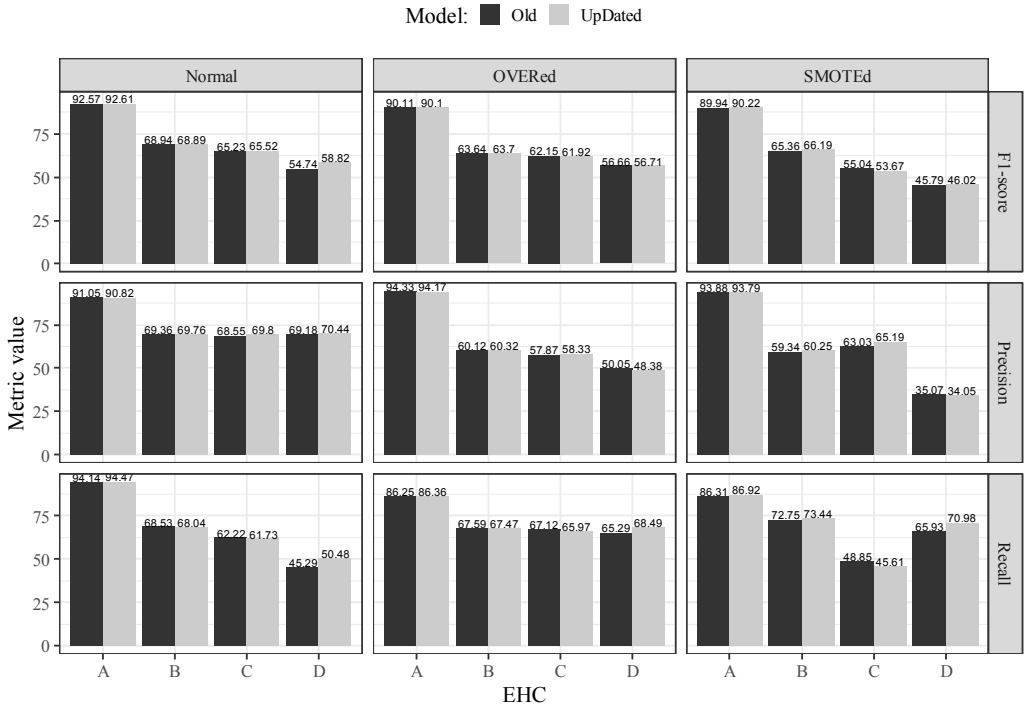



Figure 5. ANN models comparison ("Old" vs "Updated") based on recall, precision and F1-score, according to a nominal classification strategy in slope stability prediction of rock cuttings.

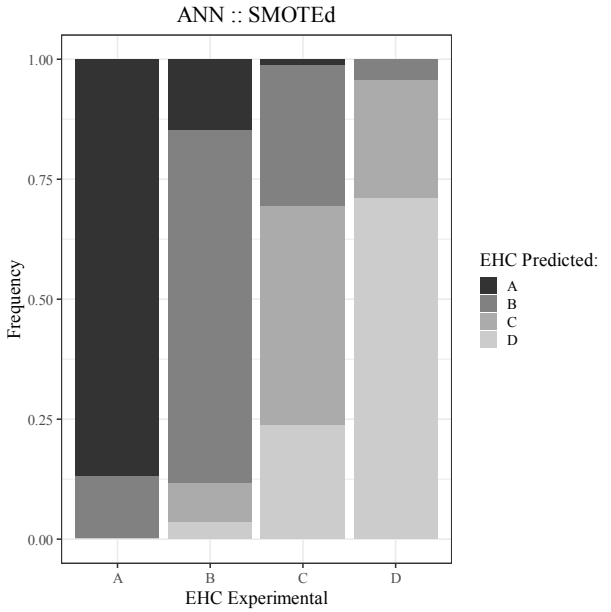



Figure 6. ANN performance ("Updated" model) in EHC prediction of rock cuttings according to a nominal classification strategy and following a SMOTE approach.

## 6 ACKNOWLEDGEMENTS

This work was partly financed by FCT / MCTES through national funds (PIDDAC) under the R&D Unit Institute for Sustainability and Innovation in Structural Engineering (ISISE), under reference UIDB / 04029/2020. This work was also financed by national funds through FCT - Foundation for Science and Technology, under grant agreement SFRH/BPD/94792/2013 attributed to the 1st author. A special thanks goes to NetworkRail that kindly made available the data (basic earthworks examination data and the Earthworks Hazard Condition scores) used in this work.

## 7 REFERENCES

Pourkhosravani A. and Kalantari, B. 2011. A review of current methods for slope stability evaluation. *Electronic Journal of Geotechnical Engineering* 16(Jan), 1245–1254.

Suchomel R. et al. 2010. Comparison of different probabilistic methods for predicting stability of a slope in spatially variable c-φ soil. *Computers and Geotechnics* 37(1), 132–140.

Sivakumar Babu G. and Murthy D. 2005. Reliability analysis of unsaturated soil slopes. *Journal of Geotechnical and Geoenvironmental Engineering* 131(11), 1423–1428.

Pinheiro M. et al. 2015. A new empirical system for rock slope stability analysis in exploitation stage. *International Journal of Rock Mechanics and Mining Sciences* 76(Jun), 182–191.

Power C., Mian J., Spink T., Abbott S. and Edwards M. 2016. Development of an evidencebased geotechnical asset management policy for network rail, great britain. *Procedia Engineering* 143(Sep), 726–733.

Cheng M.-Y. and Hoang N.-D. 2016. Slope collapse prediction using bayesian framework with knearest neighbor density estimation: Case study in taiwan. *Journal of Computing in Civil Engineering* 30(1), 04014116.

Cheng M.-Y., Wu Y.-W. and Chen K.-L. 2012. Risk preference based support vector machine inference model for slope collapse prediction. *Automation in Construction* 22(Mar), 175–181.

Kang F., Li J.-S., Wang Y. and Li J. 2017. Extreme learning machine-based surrogate model for analyzing system reliability of soil slopes. *European Journal of Environmental and Civil Engineering*, 1–22.

Tinoco J., Gomes Correia A., Cortez P. and Toll D. 2018. Data-driven model for stability condition prediction of soil embankments based on visual data features. *Journal of Computing in Civil Engineering* 32(4), 04018027.

Liao S., Chu P. and Hsiao P. 2012. Data mining techniques and applications. A decade review from 2000 to 2011. *Expert Systems with Applications* 39(12), 11303–11311.

Garg A., Garg A., Tai K. and Sreedeep, S. 2014. An integrated srm-multi-gene genetic programming approach for prediction of factor of safety of 3-d soil nailed slopes. *Engineering Applications of Artificial Intelligence* 30, 30–40.

Tinoco J., Gomes Correia A. and Cortez P. 2014a. A novel approach to predicting young's modulus of jet grouting laboratory formulations over time using data mining techniques. *Engineering Geology* 169(Feb), 50–60.

Tinoco J., Gomes Correia A. and Cortez P. 2014b. Support vector machines applied to uniaxial compressive strength prediction of jet grouting columns. *Computers and Geotechnics* 55(Jan), 132–140.

Yao X., Tham L. and Dai F. 2008. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of hong kong, china. *Geomorphology* 101(4), 572–582.

Kenig S., Ben-David A., Omer M. and Sadeh A. 2011. Control of properties in injection molding by neural networks. *Engineering Applications of Artificial Intelligence* 14(6), 819–823.

Chandrashekar, G. and Sahin F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40(1), 16–28.

Golberg D. E. 1989. *Genetic algorithms in search, optimization, and machine learning*. Addion wesley.

Sun Y., Babbs C. and Delp, E. 2006. A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm. *In: 2005 IEEE Engineering in Medicine and Biology, 27th Annual Conference*, IEEE, 6532–6535.

Alexandridis A., Patrinos P., Sarimveis H. and Tsekouras G. 2005. A two-stage evolutionary algorithm for variable selection in the development of rbf neural network models. *Chemometrics and Intelligent Laboratory Systems* 75(2), 149–162.

Bonissone P., Subbu R. and Lizzi J. 2009. Multicriteria decision making (mcdm): a framework for research and applications. *Comp. Int. Mag.* 4(3), 48–61.

Chou J. and Tseng H. 2011. Establishing expert system for prediction based on the project oriented data warehouse. *Expert Syst. Appl.* 38(1), 640–651.

Team, R. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Viena, Austria. http://www.r-project.org/. (Accessed on June 2021).

Mersmann O., Trautmann H., Steuer D., Bischl B. and Deb K. 2014. *Package "mco": Multiple criteria optimization algorithms and related functions*. https://cran.rproject.org/package=mco.

Hastie T., Tibshirani R. and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer-Verlag, New York.

Cortez P. 2010. Data mining with neural networks and support vector machines using the r/rminer tool. *In: Advances in Data Mining: Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, 572–583. Springer, Berlin, Germany.