

Landslide susceptibility mapping in Denmark – A Machine Learning approach

Angelina AGEENKO, Lars BODUM, Denmark

Key words: landslides, predictive modelling, susceptibility mapping

SUMMARY

The first Danish national landslide mapping and the resultant comprehensive landslide inventory, produced by The Geological Survey of Denmark and Greenland (GEUS), comprises more than 3200 landslides, indicating that landslide hazard might present a more serious problem in Denmark than earlier estimated, requiring methods to map areas at risk.

This study proposes a Machine Learning approach to identify places that might be vulnerable to landslides based on topographic, hydrological, geological, and anthropogenic exogenous variables in a region of interest with a relatively high number of landslide occurrences situated around Vejle Fjord, Denmark, using publicly available data and open-source software. The supervised, tree-based machine learning algorithm Random Forest has been applied for a binary classification of the sample data as landslide presence (centroids from the landslide inventory) and randomly sampled absence points and the classification has been validated through test data unseen by the model.

The results have been presented in the form of a landslide susceptibility map divided into several probability classes. The overall predictive accuracy of 94% indicates that the applied model has prospects to be applied for mapping areas in Denmark that might be prone to landslides. The mapping can be useful for decision-makers and can potentially pave the way to a legislative framework and land management practices for areas vulnerable to landslides and for preventive decisions as well as mitigative measures of the potential risks associated with landslide events.

Landslide susceptibility mapping in Denmark – A Machine Learning approach

Angelina AGEENKO, Lars BODUM, Denmark

1. INTRODUCTION

Professionals like geologists, engineers and others use different definitions for landslides, reflecting the complexity of studying this phenomenon. Even though there are different definitions, basic and common terms associated with the process of landslides exist, regardless of which type of landslide it is. The term landslide generally describes a downslope movement of soil, rock, and other materials, caused by the effects of gravity (Highland and Bobrowsky, 2008).

Many people believe that landslide is a geohazard that takes place in regions with steep slopes. Landslides, however, can be triggered by a wide range of mechanisms and are not solely found in the places with steep terrain (Highland and Bobrowsky, 2008). The stability of the slopes can be influenced by several different natural phenomena including precipitation, melting snow, changes in temperature and various human modifying activities, which can result in landslides (Gariano and Guzzetti, 2016).

Landslides affect natural and built environments, where the latter can suffer economical damage. Residential areas built on or near unstable slopes can be destroyed by landslides and physical infrastructure can suffer damage, affecting a large amount of people. Furthermore, the world population continue to expand, increasing the vulnerability to landslides, as people tend to move to new lands which might have been previously deemed hazardous (Highland and Bobrowsky, 2008). Landslides is one of the most widespread geophysical hazards, leading to substantial economic loss, affecting millions of people, and causing casualties (Wallemacq and House, 2018; Froude and Petley, 2018).

Until recently, there has been little attention paid to the risks connected to landslides in Denmark. Only sparse amount of research has been conducted on the subject until now, where most of the studies were concentrated on field investigations of single local landslides (Svennevig et al., 2020). In 2015 GEUS reported 10 landslides in total to the European landslide databases, being below the rest of Europe (Herrera et al., 2017). With the emerged high quality, nationwide digital elevation models (DEM) in Denmark, knowledge of the terrain and landslides increases. This has made it possible to conduct detailed mapping of landslides in Denmark, that indicates that the extent of landslides has been underrepresented and there are more potentially dangerous landslides in Denmark (Svennevig et al., 2020). The preliminary mapping of landslides in Denmark, seen in Figure 1, conducted by GEUS in 2020 (Svennevig et al., 2020), and which has recently led to the first comprehensive national landslide database

(Lützenburg et al., 2021), has identified more than 3000 distinct landslide cases, indicating that landslides in Denmark are more common than previously realized.

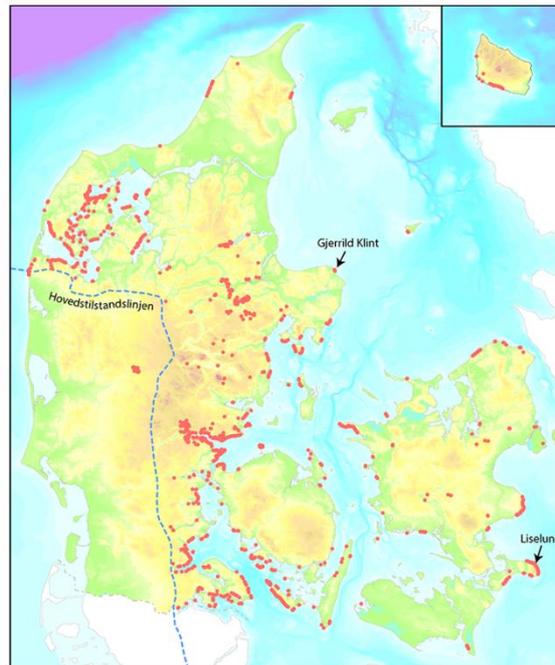


Figure 1. Result of preliminary mapping of landslides in Denmark (Svennevig, 2020)

2. LANDSLIDE SUSCEPTIBILITY MAPPING

It is crucial to translate a natural hazard and to communicate information about it in a way understandable by policy- and decision-makers, planners, landowners, non-technical users and others concerned. This is because the level of concern with a potential hazard among them will be low, if the likelihood of hazardous events is low, the occurrence location is unknown, and the severity is mild. This process of translating and communicating natural hazards normally contains the following elements (Highland and Bobrowsky, 2008):

- Likelihood of the occurrence of a hazardous event that would result in damage or a challenge to existing safety standards.
- Expected location and area effected physically or socioeconomically by the hazardous event.
- Expected severity of physical or socioeconomic effects of the hazardous event.

A tool that can help authorities and decision makers to identify landslide-prone areas and plan for future landslides is susceptibility assessment. Landslide susceptibility is an expression of

the spatial probability of landslide events based on certain geo-environmental conditions (Li et al., 2017). Landslide susceptibility models and derived maps with landslide-prone locations can be regarded as an initial phase and a starting point on the way to a landslide hazard and risk assessment, or it can also be a final product used in land management or environmental impact assessment (Corominas et al., 2014). Even though several different methods for landslide susceptibility assessment and mapping exist, they are based on several common assumptions (Fell et al., 2008; Reichenbach et al., 2018):

- Landslide events leave recognizable traces that can be classified and mapped through field work or remote sensing products.
- Landslide events are controlled by physical laws. Conditions and causative factors can be used to model and predict landslide spatial occurrences.
- The past can explain the future. Areas that have experienced landslides in the past will likely be exposed to landslides in the future unless the source of the landslides is exhausted.
- Future landslide events are more likely to happen in the areas with similar topographical, environmental, geological, and geomorphological conditions as to the areas that have been affected by landslides in the past.
- Spatial landslide occurrence can be derived from heuristic investigations, computed using data, or derived from physical models. Consequently, an area of interest can be divided into susceptibility classes and assigned zones based on the likelihood of landslide occurrence.

Methods for landslide susceptibility assessment are divided into qualitative, quantitative, and semi-quantitative (Shano et al., 2020). The qualitative and semi-quantitative approaches are highly based on expert knowledge and decisions and are generally regarded as being subjective, while the quantitative approaches are based on a mathematical approach i.e., statistics and hereby considered as more objective approaches. The approaches based on expert knowledge and decisions are generally more popular than the mathematical approaches because of their simpler usage and evaluation, while these approaches are highly subjective, and can therefore change, depending on who evaluates landslide susceptibility.

The quantitative approach such as machine learning has gained popularity for landslide susceptibility modelling over bigger areas, which might be explained by deficient and scarce detailed data or its complexity. Landslide susceptibility mapping using machine learning is normally performed under the key assumptions that landslides are likely to occur in similar conditions as in the areas earlier affected by past and present landslide events. The exact relation between landslide and their preconditioning factors is not always known, and they can be difficult to measure over larger areas. Due to this, they are represented by several exogenous variables, functioning as proxies for those factors (Goetz et al., 2015).

3. REGION OF INTEREST

The selected region of interest, seen in Figure 2, is situated in the eastern part of Jutland, Denmark, bounded by Horsens and Kolding from the North and South, and by Vejle and the Kattegat in the West and East, respectively. This region of interest has been selected due to the relatively high number of historical landslide occurrences. According to the national landslide inventory, the area has been historically affected by 189 inland landslides and 264 coastal landslides of various sizes, whose total area is estimated to be 3.33 km².

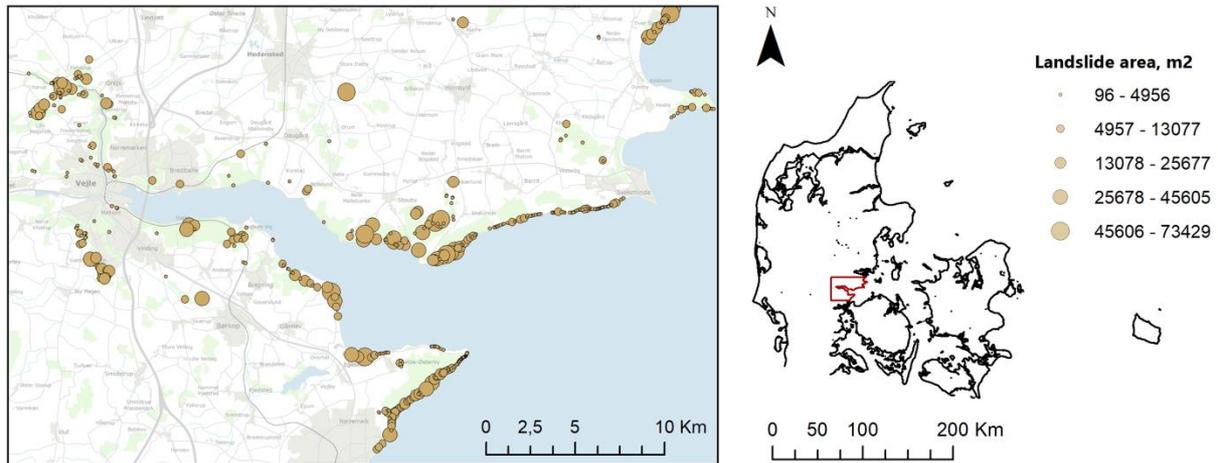


Figure 2. Area of interest with the landslides from the national inventory visualized according to their size from the Danish landslide inventory (Svennevig and Lützenburg, 2021). Base map: The Danish Agency for Data Supply and Efficiency, SDFE.

4. PROPOSED METHOD

4.1 General workflow

Supervised ML classification models usually produce two types of predictions. The first one is in the form of a discrete category, or classes. The second one is a valued prediction in the form of probability, where each class gets a predicted probability value between 0 and 1 and the sum is 1 (Kuhn and Johnson, 2013). In this study, the supervised classification model proposes a relevant approach, since landslide classes are treated as a binary classification problem, where the class is either non- landslide (class 0) or landslide (class 1), and a susceptibility map is produced by generating a probability map. Furthermore, with the landslide data obtained in the project, a supervised binary approach is chosen as this allows for pre-labelling of landslide/non- landslide areas.

The method in this study consists of several stages. First, the definition of the task for machine learning is formulated, followed by data preprocessing of the dependent (landslide presence and absence locations) and independent (predictive) variables and feature selection using Pearson’s correlation coefficient. Then the machine learning model is set up, optimized, and its performance is assessed. As the final step, the resultant landslide susceptibility map is visualized. The workflow is seen in Figure 3.

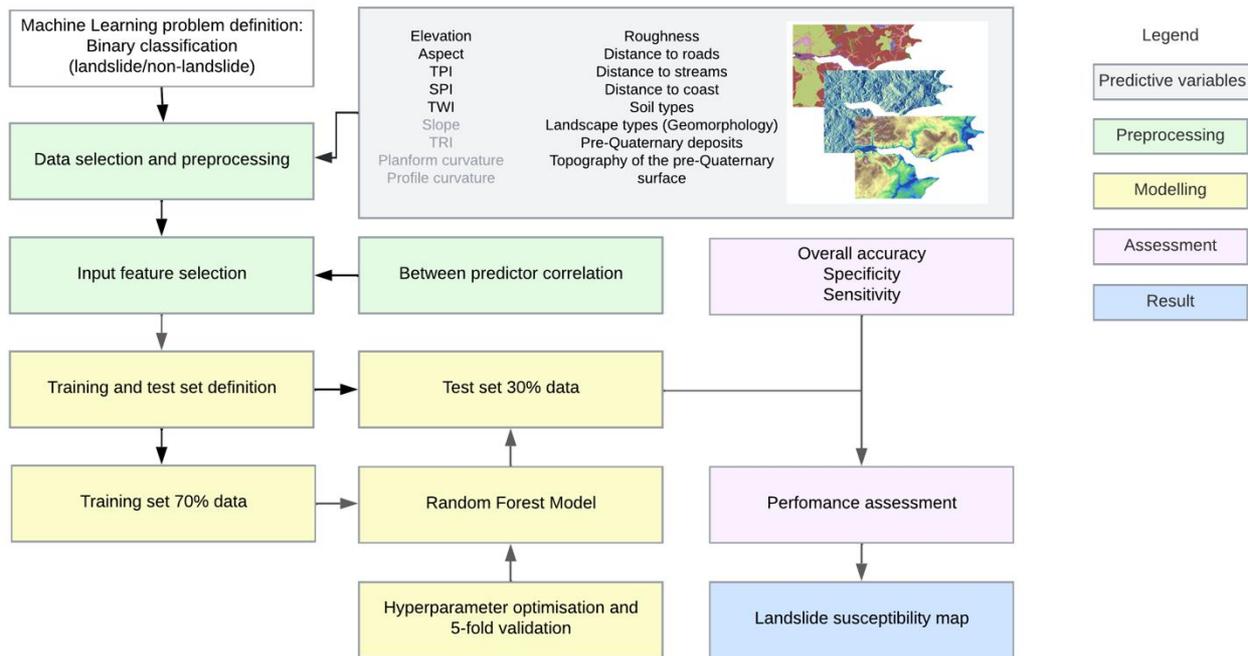


Figure 3. Illustration of the workflow. The light gray variables are not included into the final model.

4.2 Preprocessing

The step of data preparation for this study involves exploration of the data at hand and organizing it into a structured and appropriate form and making it ready for analysis. The available data has been reviewed, and its quality, up-to-datedness, and suitability for the task at hand have been assessed. Detected anomalies and inconsistencies in the data sets such as missing values have been removed. The data integration process consisted of combining data from the different sources and of various types and making it more homogeneous by bringing it to the same projected spatial reference system (ETRS89/UTM zone 32N), same extent of the area of interest, the raster data type, and the same resolution of 10m. For the categorical variables the nearest-neighbour interpolation method is used, while the continuous data is resampled bilinearly.

Two types of variables are essentially needed to produce a landslide susceptibility model and a map. The first type of variable is a target variable representing the landslide occurrences, which are normally obtained through a landslide inventory database. The second type of variable is connected to landslide predisposing factors.

As the target variable, the centroids of landslides from the Danish national landslide inventory are used along with an equal amount of randomly sampled points representing areas that have not been affected by landslides (non-landslide points) to avoid class imbalance and to make the ML algorithm distinguish between these two classes (Ma et al., 2020; Brock et al., 2020).

The predictor variables, that can be proxies for landslide predisposing factors, have been selected based on the availability of data and based on the recommendations in the literature within landslide susceptibility studies (Azarafza et al., 2021; Brock et al., 2020; Saleem et al., 2019; Zhou et al., 2017). Open Danish geodata was used as a source for the predictor variables. Elevation was obtained directly from the Danish DEM, while slope angle, aspect, planform and profile curvature, roughness, TRI (Terrain Ruggedness Index) were computed as DEM derivatives using GDAL terrain processing tools. Aspect was further transformed into sine and cosine components representing terrain's easternness and northerness to avoid the discontinuity of this circular parameter (Brock et al., 2020). TWI (Topographic Wetness Index) was calculated according to the formula (Saleem et al., 2019):

$$TWI = \ln \left(\frac{a}{\tan\beta} \right)$$

SPI (Stream Power Index) is calculated by (Saleem et al., 2019):

$$SPI = a * \tan\beta$$

Where a is the contributing catchment area and β is the slope in radians.

Distance from coast, streams and roads is computed using a proximity tool with the vector files representing the corresponding features obtained from GeoDanmark. Categorical variables, obtained from GEUS, such as soil types, geomorphology, the map of the pre-Quaternary deposits and the pre-Quaternary topography are converted from vector files to raster data type.

Dimensionality reduction is then carried out with the purpose of eliminating redundant and irrelevant data, increasing computational speed and accuracy of the model. Dimensionality reduction in this study is conducted through a general technique - feature selection. Feature selection reduces dimensionality by selecting a subset of the most informative and important variables without any transformation applied to them (Khalid et al., 2014). The pairwise Pearson's correlation between the variables has been calculated and visualized with the help of a correlation matrix plot seen in Figure 4, where the variables the high degree of correlation over the threshold of ± 0.75 (*slope, TRI, plan and profile curvature*) has been eliminated indicating the redundancy of the information (Kuhn and Johnson, 2019).

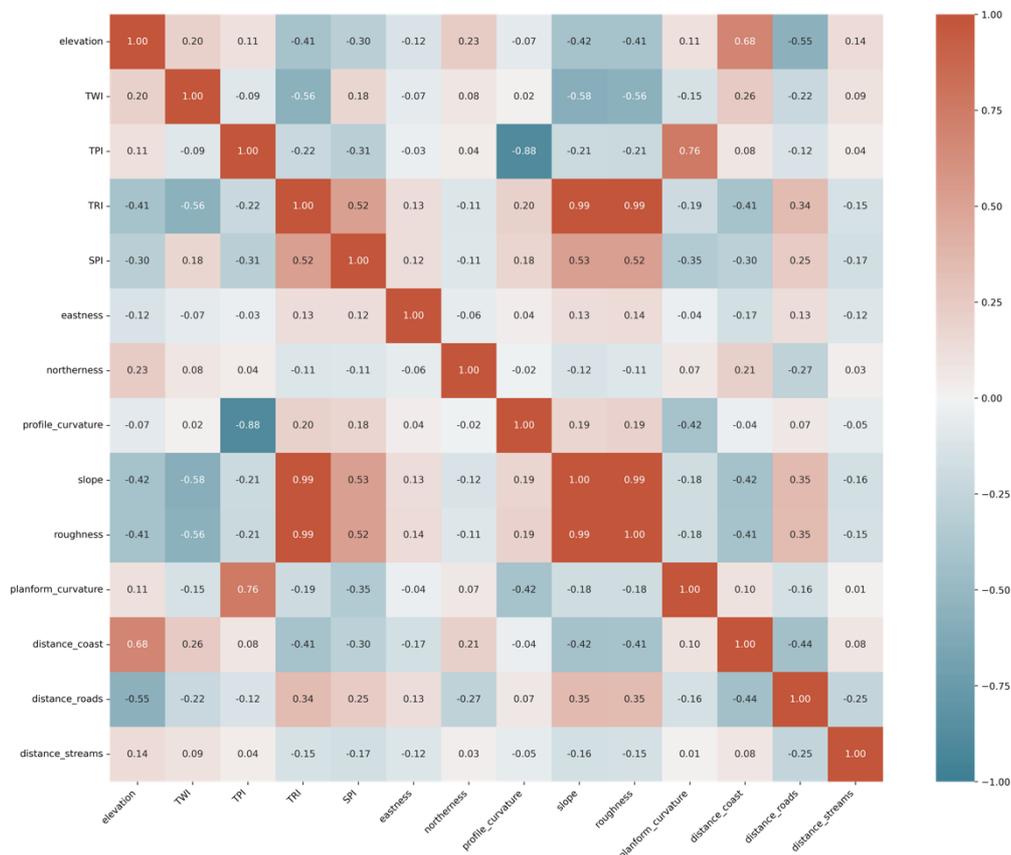


Figure 4. Correlation matrix

Data transformation has been performed to change the data from one representation to another and to give it better interpretability and make them more suitable for machine learning tasks. In this study, numeric to numeric transformation in form of z-score normalization is used, while binarization, or “one-hot” encoding for categorical variables has been applied. The visualization of the final predictive variables selected for modelling is seen in Figure 5.

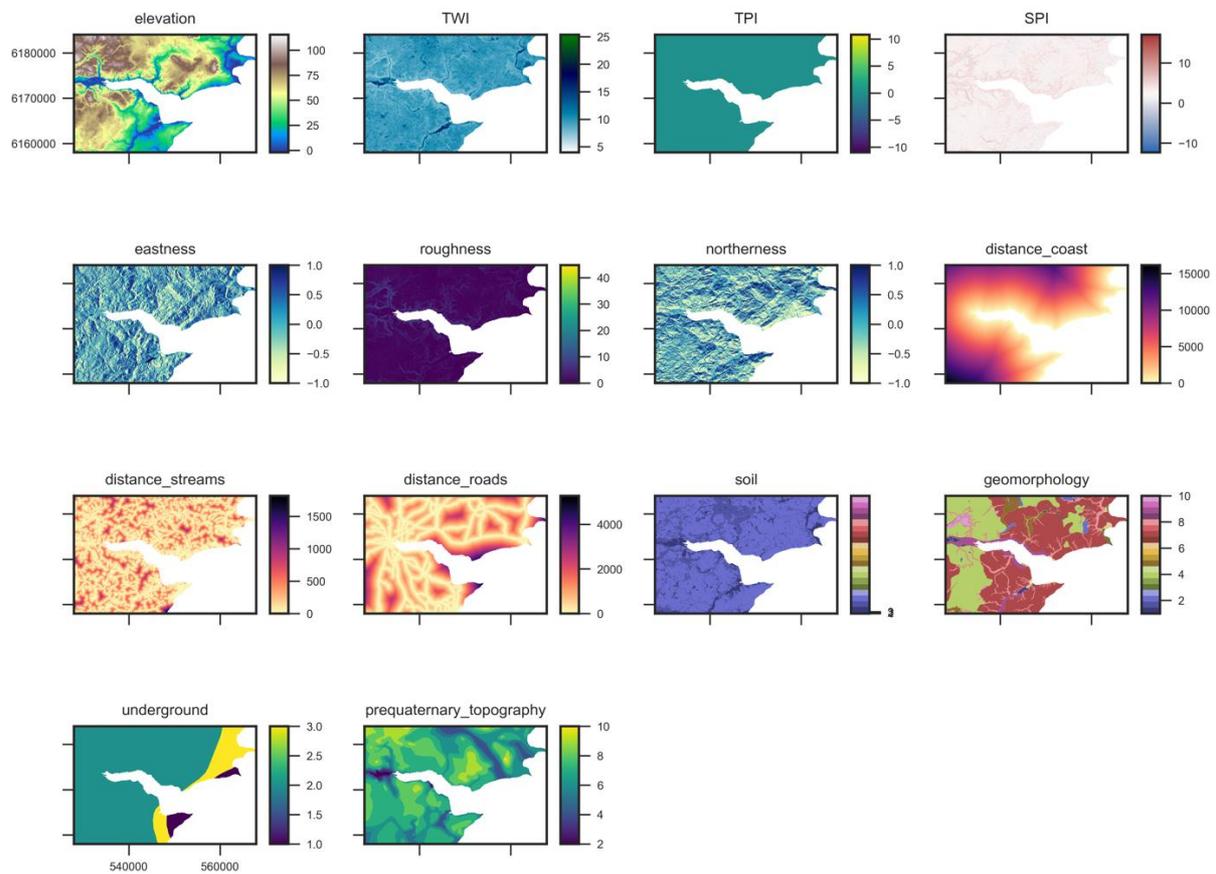


Figure 5. Overview over the selected variables

4.3 Modelling

The preprocessed data is partitioned into training and testing subsets to produce an unbiased validation of the model's predictions. In this project, data set is divided into a training subset 70% of the data to train the model and a testing subset consisting of 30% of the data, seen in Figure 6, to validate the model by making predictions on the unseen data.

Random Forest is a supervised ensemble tree-based ML algorithm, which uses a collection of individual decision trees to enhance their individual performance (Breiman, 2001). The algorithm generates a randomly drawn sample from the original data to train an individual tree model on, where each decision tree randomly selects a sub-sample of the predictors at each node to avoid correlation between the decision trees. The model output is a prediction, which is, in case of a classification problem, a distinct class. Each tree generates a prediction (vote) for a new sample of the observed data and a different set of predictors, and these outcomes then constitute the algorithm's final prediction for the classification task based on the majority vote (Kuhn and Johnson, 2013). The optimal model is created by using hyperparameter optimization through Grid Search (Kuhn and Johnson, 2019) and 5-fold cross-validation technique (Berry et

al., 2020). The number of estimators used to train the classifier yielding to the best prediction results turned out to be 300.

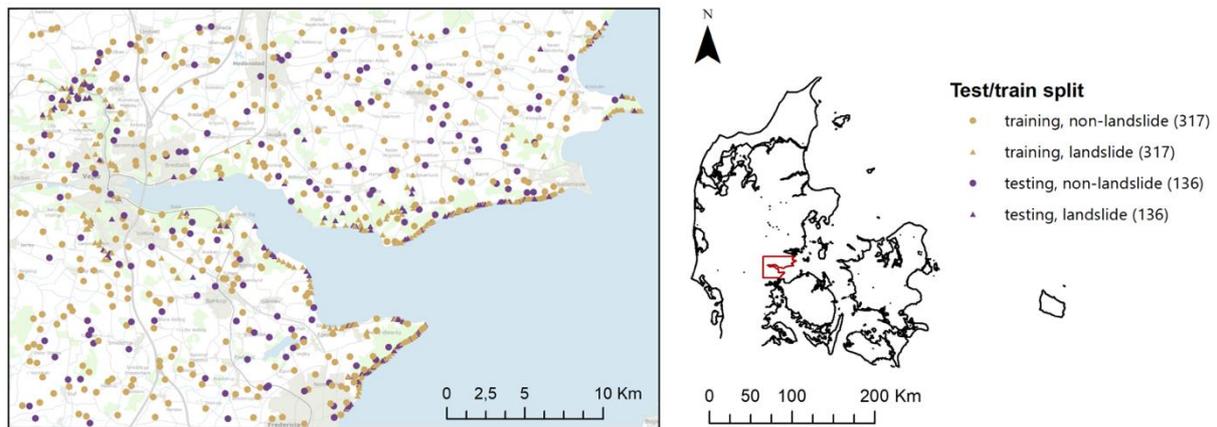


Figure 6. Overview of the data used for training and testing

4.4 Performance assessment

One of the most common methods of conducting accuracy assessments for classification tasks is producing a classification error matrix, also known as confusion matrix. Confusion matrices compare the ground truth with the results from the classification, where the correctly predicted samples are in the matrix diagonal, while errors correspond to non-diagonal elements (Lillesand et al., 2008). In the given case, a false negative implies that the model was not able to recognize an actual landslide, while a false positive indicates that a non-landslide point has been classified as a landslide by the model. The first error is least desirable since some potential landslide susceptible regions might be overlooked and will be mapped as safe. Once the confusion matrix is computed, it is further interpreted to draw several characteristics regarding the performance of the classification such as overall accuracy, sensitivity, and specificity. The produced confusion matrix based on the model predictions on the test dataset is seen in Figure 7.

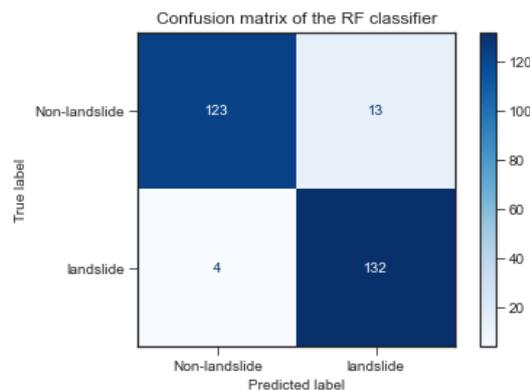


Figure 7. Confusion matrix for the model predictions on the test data.

Overall accuracy shows the percentage of all the data samples that were correctly classified. It is computed by dividing the total number of correctly classified samples, with the total number of reference samples (Lillesand et al., 2008):

$$\text{Overall Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Sample}} = \frac{123 + 132}{123 + 132 + 4 + 13} = 0,94 = 94\%$$

Sensitivity corresponds to the proportion of landslide points which are correctly classified as landslides out of all actual landslide points. Sensitivity is computed using the following formula (Kotu and Deshpande, 2015):

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{132}{132 + 4} = 0,97 = 97\%$$

In comparison to sensitivity, specificity expresses the percentage of non-landslide samples which are correctly classified as non-landslides in relation to all actual landslide absence points (true negatives and false positives), and is computed by the following formula (Kotu and Deshpande, 2015):

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} = \frac{123}{123 + 13} = 0,90 = 90\%$$

5. RESULTS

The model has performed with the overall accuracy of 94% indicating a rather successful classification. The higher number of false positives and lower specificity compared to sensitivity indicate that the model is likely to overpredict landslide occurrences, resulting in a susceptibility map that potentially exaggerates the extent of landslide prone areas.

The trained and validated Random Forest model is used to make predictions for the whole region of interest based on the selected variables. The percentile landslide probability intervals are sorted into 5 following classes <50%, 50-65%, 65-80%, 80-95%, and >95%. The majority filter is used for postprocessing the result for generalization and smoothing to reduce the noise and single cell misclassifications. The susceptibility map, seen in Figure 8, indicates high risk of landslides in certain places along the coast and in the river valleys to the North-West in the region of interest.

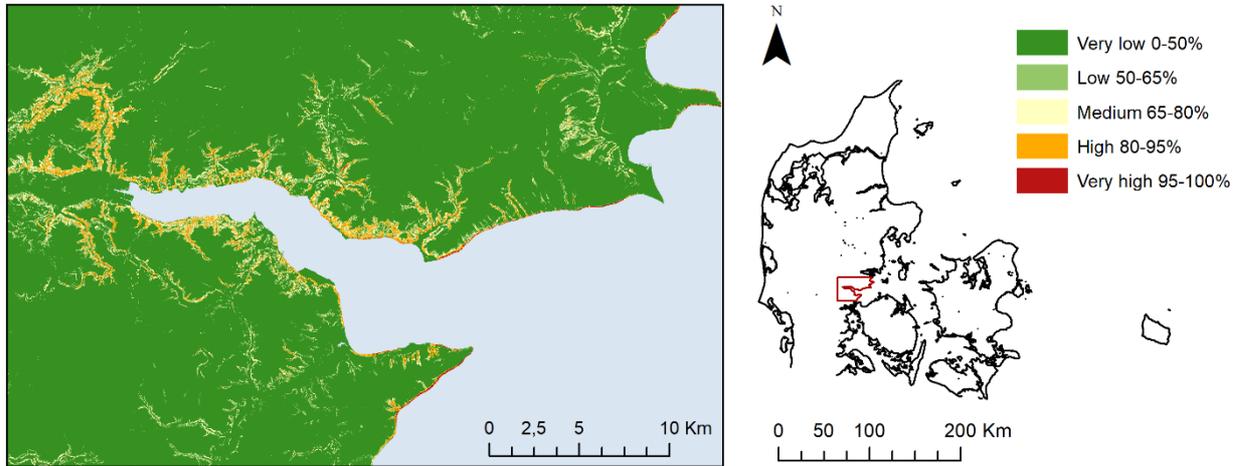


Figure 8. The final landslide susceptibility map

6. CONCLUSION

This study has applied a data-driven approach such as Machine Learning to predict landslide susceptibility in an area in Denmark, based on the main assumptions that the past is the key to the future and that landslides are likely to occur in areas with similar characteristics as in regions that were earlier impacted by landslides. The Random Forest classifier has demonstrated a promising result of overall accuracy at 94% on the test data set, indicating that the method might have the potential for landslide susceptibility mapping in Denmark.

At this moment no planning or legislation for geo-hazard such as landslide exists in Denmark. Planning for landslides can potentially be included into climate adaptation plans and be conducted in a similar manner as planning for inundation. In 2007 the EU made a floods directive, which requires all member states to assess and map flood risks. In Denmark, the Danish coastal authority is responsible for mapping flood risk areas, and the municipalities must create risk management plans for these affected areas. A similar initiative could be implemented for landslides, where the method used in the project proposes an option of creating landslide risk areas. The results of the current mapping should be considered as a decision support and should be interpreted with caution, rather than substituting professional expertise. An expert-based validation would be required to assess the susceptibility maps.

REFERENCES

- Azarafza, M., Azarafza, M., Akgün, H. et al. 2021. Deep learning-based landslide susceptibility mapping. *Sci Rep* 11, 24112 (2021)
- Berry, M. W., Azlinah M. and Yap, B.W. 2020. *Supervised and Unsupervised Learning for Data Science* (Springer, Cham, 2020), 1st edition.
- Breiman, L. 2001. Random Forest. *Machine Learning* 2001, 45, 5–32.
- Brock, J.; Schratz, P.; Petschko, H.; Muenchow, J.; Micu, M.; Brenning, A. 2020. The performance of landslide susceptibility models critically depends on the quality of digital elevation models. *Geomatics, Natural Hazards and Risk* 2020, 11, 1075–1092.
- Corominas, J.; van Westen, C.J.; Frattini, P.; Cascini, L.; Malet, J.P.; Fotopoulou, S.; Catani, F. et al. 2014. Recommendations for the quantitative analysis of landslide risk. *Bulletin of engineering geology and the environment* 2014, 73(2), 209–263.
- Fell, R.; Corominas, J.; Bonnard, C. et al. 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering geology* 2008, 3, 85–98.
- Froude, M. J. and Petley, D. N. 2018. Global fatal landslide occurrence from 2004 to 2016, *Natural Hazards and Earth System Sciences*, 18, 2161–2181.
- Gariano, S. L. and Guzzetti, F. 2016. Landslides in a changing climate. In: *Earth-Science Reviews*, volume 162.
- Goetz, J. N.; Brenning, A.; Petschko, H.; Leopold, P. 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers and Geosciences* 2015, 81, 1–11.
- Herrera, G.; Mateos, R.M.; García-Davalillo, J.C.; Grandjean, G.; Poyiadji, E.; Maftai, R.; Filipciuc, T.C. et al. 2017. Landslide databases in the Geological Surveys of Europe. *Landslides* 2017, 15, 359–379.
- Highland, L. M. and Bobrowsky, P. 2008. *The Landslide Handbook—A Guide to Understanding Landslides*, 1st ed.; U.S. Geological Survey Circular 1325: Colorado, USA.
- Khalid, S.; Khalil, T. and Nasreen, S. 2014. A survey of feature selection and feature extraction techniques in machine learning. In: *2014 Science and Information Conference*, pages 372–378.

Kotu, V. and Deshpande, B. 2015. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner, 1st ed.; Morgan Kaufmann.

Kuhn, M. and Johnson, K. 2013. Applied Predictive Modeling, 1st ed.; Springer: New York, USA.

Kuhn, M. and Johnson, K. 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models, 1st ed.; CRC Press, Taylor and Francis, 2019.

Li, L.; Lan, H. et al. 2017. A modified frequency ratio method for landslide susceptibility assessment. In: Landslides, volume 14, no. 2, pages 727–741.

Lillesand, T.M.; Kiefer, R.W.; Chipman, J.W. 2008. Remote sensing and image interpretation; John Wiley & Sons: Hoboken, US.

Lützenburg, G.; Svennevig, K.; Bjørk, A. A.; Keiding, M.; Kroon, A. 2021. A national landslide inventory of Denmark. Earth System Science Data Discussions 2021, 1–13. Under review

Ma, Z.; Mei, G. and Piccialli, F. 2020. Machine learning for landslides prevention: a survey. In: Neural Computing and Application, (2020).

Reichenbach, P.; Rossia, M.; Malamud, B.D.; Mihirb, M.; Guzzetta, F. 2018. A review of statistically-based landslide susceptibility models. Earth-Science Reviews 2018, 180, 60–91.

Saleem, N.; Huq, M.E.; Twumasi, N.Y.D.; Javed, A.; Sajjad, A. 2019. Parameters Derived from and/or Used with Digital Elevation Models (DEMs) for Landslide Susceptibility Mapping and Landslide Risk Assessment: A Review. ISPRS international journal of geoinformation 2019, 8(12), 545.

Shano, L.; Raghuvanshi, T.K.; Meten, M. 2018. Landslide susceptibility evaluation and hazard zonation techniques – a review. Geoenvironmental Disasters 2018, 7(1), 1–19.

Svennevig, K. 2020. 3000 landskredsområder kortlagt i Danmark. URL <https://www.geus.dk/om-geus/nyheder/nyhedsarkiv/2020/nov/skred/>

Svennevig, K. and Lützenburg, G. 2021, Danish landslide inventory 211104 [dataset], <https://doi.org/10.6084/m9.figshare.16965439.v1>

Svennevig, K. and Keiding, M. 2020. En dansk nomenklatur for landskred. Geologisk Tidsskrift 2020, side 19–30, København.

Svennevig, K, Lützenburg, G.; Keiding, M.K.; Pedersen, S.A.S. 2020. Preliminary landslide mapping in Denmark indicates an underestimated geohazard. GEUS Bulletin, 44.

Wallemacq, P. and House, R. 2018. Economic Losses, Poverty and Disasters 1998-2017. Centre for Research on the Epidemiology of Disasters United Nations Office for Disaster Risk Reduction.

Zhou, C.; Yin, K.; Cao, Y.; Ahmed, B.; Li, Y.; Catani, F.; Pourghasemi, H.R. 2017. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China, Computers and Geosciences (2017), pages 23–37.

BIOGRAPHICAL NOTES

Lars Bodum is a land surveyor, Ph.D., and an Associate Professor of Planning for Urban Sustainability in the Department of Planning at Aalborg University. His academic focus is on the increasing levels of urbanization and digitalization, emphasizing how technologies can create an impact and promote sustainable development of cities. He works with involvement and inclusion in data gathering, modelling, analysis, visualization, and dissemination of knowledge about the place for the benefit of the citizens.

Angelina Ageenko is in her final year of Master's in Surveying, Planning and Land Management (cand.geom.) at Aalborg University. Her interests include geoinformatics, mapping and land management. She focuses on expanding her skills within predictive modelling and on applying them to the projects that have positive effect on the real world and promote the sustainable development goals.

CONTACTS

Lars Bodum
Department of Planning
Rendsburggade 14 9000 Aalborg
Denmark
Phone: +45 9940 8078
E-mail: lbo@plan.aau.dk

Angelina Ageenko
Aalborg
Denmark
Tel. +45 3140 5077
Email: angelinkatula@gmail.com