New Machine Learning Ensemble for Flood Susceptibility Estimation

Romulus Costache¹,² · Alireza Arabameri³ · Iulia Costache⁴ · Anca Crâciun² · Binh Thai Pham⁵

Abstract

Floods are among the most severe natural hazard phenomena that affect people around the world. Due to this fact, the identification of zones highly susceptible to floods became a very important activity in the researcher’s work. In this context, the present research work aimed to propose the following 3 novel ensembles to estimate the flood susceptibility in Putna river basin from Romania: UltraBoost-Weights of Evidence (U-WOE), Stochastic Gradient Descending-Weights of Evidence (SGD-WOE) and Cost Sensitive Forest-Weights of Evidence (CSForest-WOE). In this regard, a sample of 132 flood locations and 14 flood predictors was used as input datasets in the 3 aforementioned models. The modeling procedure performed through a ten-fold cross-validation method revealed that the SGD-WOE ensemble model achieved the highest performance in terms of ROC Curve-AUC (0.953) and also in terms of Accuracy (0.94). The slope and distance from river flood predictors achieved the highest importance in terms of flood susceptibility genesis, while the aspect, TPI, hydrological soil groups, and plan curvature have the lowest influence in terms of flood occurrence. The area with high and very high susceptibility represents between 21% and 24% of the Putna river basin from Romania.

Keywords Flood susceptibility · Romania · Machine learning · Bivariate statistics · Putna river basin

1 Introduction

As part of the hydrological cycle, flooding has always been a natural part of the system, however, over the past decade, a gradual increase has been observed in the frequency and magnitude of flooding (Vojtek and Vojteková 2019; Zhang et al., 2022). In Romania, this has been a challenging task because many countries have found it difficult to develop effective flood prevention measures. As climate and land-use changes continue, flooding is expected to increase. This can largely be attributed to an increase in the occurrence of flood events (Pistrika et al. 2014; Afriyanie et al. 2020). As a result of these anthropogenic
impacts, basins in which there has been substantial anthropogenic activity often experience increased runoff resulting in rapid erosion or increased evaporation processes (Chen et al. 2009; Chakrabortty et al. 2021). The mapping of flooding susceptibility is essential element of early warning systems intended to prevent and mitigate future flood risks since they identify the most prone zones based on the physical factors determining how much flooding will occur (Vojtek and Vojteková 2019; Janizadeh et al. 2021). Therefore, vulnerability assessment may also incorporate the notion of susceptibility. There is a series of conditions that have been recognized as necessary to construct flood susceptibility mapping techniques as they reflect the physical characteristics of the area under investigation (Arnell and Gosling 2016). These physical characteristics refer to several geographical components like lithology, land use, slope angle, river density, and other morphometric variables and hydrological soil groups. In selecting the appropriate conditioning factors for an assessment of flood susceptibility, one has to take into consideration the spatial scale of that analysis. For example, if the study zone is large (for example, the nation’s territory), then using fewer variables would be reasonable. This is because it is more challenging to achieve the same data (thought at the same resolution or scale) for the whole area (Dodangeh et al. 2020). Some studies contend that fewer factors are possibly more likely to reduce the possibility of your receiving some overrated factors if there is a limited number of them. Local-scale (e.g., catchment) studies may integrate a variety of data and factors that are location-specific, allowing for an accurate description of flood predispositions because they utilize local data and variables.

Geographical information systems (GIS) have emerged as powerful tools to facilitate the synthesis of multiple input data sources (Zhao et al. 2020; Wang et al. 2021; Zhou et al. 2021a, 2021b), variables (Gao et al. 2021; Liu et al. 2021a, 2021b; Quan et al., 2022), and ensuing logical and mathematical relationships to generate flood susceptibility maps, as flooding has both spatial and temporal aspects. There have been several methods developed over the years which have been used to identify and assess flood-prone areas in different geographic areas of the world. It should be noted the very strong development after 2010 of machine learning techniques (Chao et al. 2021; Zhang et al. 2020; Zhang et al. 2021; Yin et al. 2022a; Zhan et al. 2022a, 2022b). Thus, they began to be used on a large scale in almost all scientific fields (Jafari-Asl et al. 2021; Rad et al. 2022). One of the factors that determined the accelerated increase in the degree of usage of these modern techniques is represented by the development of software and the implementation of machine learning algorithms with the help of programming languages (Mehta and Devarakonda 2018). There were many advances in machine learning techniques used for mapping flood susceptibility in recent years, and several have revealed powerful features and demonstrated favorable results (Zhang et al. 2019; Liu et al. 2020; Liu et al. 2022; Li et al. 2022). These techniques include decision trees (Chen et al. 2020), artificial neural networks (Hong et al. 2018), random forest (Lee et al. 2017), Naive Bayes (Costache et al. 2019), logistic regression or support vector machine (Tehrany et al. 2015; Allahbakhshian-Farsani et al. 2020). Several studies have recently attempted to build flood susceptibility models by using hybrid strategies to determine their strengths and weaknesses. For example, machine learning was used to combine bootstrapping and subsampling methods (Dodangeh et al. 2020). Moreover, there were attempts to estimate the flood susceptibility using some advanced hybrid models that were generated by the combination of Dagging and Random Subspace algorithms (Tian et al. 2021), on the one hand, with the random forest, support vector machine
and artificial neural network, on the other hand (Islam et al. 2021). Other authors compute the flood susceptibility using the functional trees hybridized with bagging and rotational forest (Arabameri et al. 2020). A common practice in the assessment of flood-prone areas is to combine the bivariate statistics methods, like frequency ratio or weights of evidence, with models from the machine learning category like classification and regression tree, deep learning neural network, random forest, or support vector machine (Costache and Bui 2019; Bui et al. 2020). Hybrid machine learning algorithms provide excellent results for modeling natural hazards and many natural hazard models have been applied to hybrid machine learning algorithms for implementation and creation. While there was an agreement among the scientists that some tools would be useful to evaluate various hazards more effectively (Chen et al. 2021), there was no universal agreement on which tools were best suited to stimulate natural hazards such as floods. Flood susceptibility and other ways of stimulating natural hazards are among the research topics being discussed by researchers.

Therefore, in the present research work, the main scope of the developed workflow was to estimate the flood susceptibility across the Putna river basin from Romania, with the help of novel ensemble algorithms which were generated by the combination of Weight of Evidence bivariate statistics and the next three machine learning models: UltraBoost, Stochastic Gradient Descending and Cost Sensitive Forest.

The accuracy of the results was demonstrated with the Receiver Operating Characteristic (ROC) Curve method and by the computation of several statistical metrics. It should be also mentioned that the following sections are structured as follows: Study area; Data; Methods; Results; Discussions; Conclusions and future directions.

2 Study Area

The Putna river basin in Romania is the subject of this case study (Fig. 1). There are 2509 km² in this catchment, and its boundaries have the following coordinates: 26° 22’ 13.08” and 27° 29’ 57.84” E longitude, and 46° 2’ 44.16” and 45° 31’ 23.16” N latitude (Costache and Bui, 2019).

It is verified by government data in Romania that the worst flooding event which can be attributed to the Putna river catchment took place in 2005, and this catastrophe caused the death of a total of 15 people (Romanescu and Nistor 2011). A total of approximately €150 million was also generated by flooding that year as a result of the floods (Tîrnovan et al. 2014). Supplementary aspects of the Study area are provided in Sect. 1 (S1) of Online Supplementary Material.

3 Data

3.1 Flood Inventory

As part of this study, a map of flood risk was created using information collected from the General Inspection for Emergency Situations of Romania. The study zone was subsequently found to have 132 flood events, as a result of this investigation procedure. It was determined that an additional sample with 132 non-flood locations should be created since flood suscep-
tibility mapping is a process related to the binary classification topic. Furthermore, to ensure that we have the opportunity to test the model’s performance and to verify the accuracy of the prediction results, the flood and non-flood database samples were split into the training data set of 70% and the validation data set of 30%. Supplementary aspects of flood inventory are included in the Sect. 2 (S2) of the Online Supplementary File.

### 3.2 Flood Predictors

The flood predictors we chose to calculate the flood susceptibility for the entire region under study were comprised of 14 flood prediction models after careful consultation of the literature and taking into account the specific characteristics of the Putna River Basin. In order to create the Digital Elevation Model (DEM) across the area on which the study is focused, it was crucial to use the Shuttle Radar Topographic Mission at a resolution of 30 m. As a result, the DEM was used to derive the following nine morphometrical flood predictors, namely: slope angle (Fig. 2a), topographic position index (TPI) (Fig. 3a), elevation (Fig. 2h), convergence index (Fig. 2c), plan curvature (Fig. 3b), stream power index (SPI) (Fig. 3c), profile curvature (Fig. 2b), topographic wetness index (TWI) (Fig. 2g) and aspect (Fig. 3e). The rest of 5 conditioning factors (land use (Fig. 2d), lithological groups (Fig. 2e), distance from river (Fig. 3f), hydrological soil groups (Fig. 2f) and rainfall (Fig. 3d)) were derived using vectorial database such as: Corine Land Cover, 2018 used for land use; Geological Map of Romania 1:20000 used for lithology; National River Cadastre for Romania that was used as database for distance to river factor; Soil Map of Romania, 1:200000 used to extract hydrological soil groups; mean annual sum of precipitation for period 1980–2015.

![Study area location](image)
which was used to generate the rainfall map for the study area. A short description of the predictors is provided in Sect. 3 (S3) and Table S1 of the Online Supplementary Material.
4 Methods

4.1 ReliefF Method

By using the ReliefF method, we will be able to make a preliminary evaluation of the predictive potential of the variables used as inputs to estimate flood susceptibility (Costache et al. 2021). Furthermore, it is important to emphasize that the ReliefF algorithm is capable of operating with continuous and discrete data. It is considered that the attribute value is assigned to the nearest instance of the same or a different class of the given attribute value (Costache et al. 2021). ReliefF was used in the present study as an analysis tool using Weka 3.9 software. Additional aspects of ReliefF method are included in Sect. 4 (S4) of Online Supplementary Material.
4.2 Weights of Evidence

To provide information as to how the Weights of Evidence method can be implemented in GIS environment, it is necessary to establish the following relationships (Armaş 2012):

\[
W^+ = \ln \left( \frac{N_{pix1}}{N_{pix1} + N_{pix2}} \right) \ln \left( \frac{N_{pix3}}{N_{pix3} + N_{pix4}} \right)
\]

(1)

\[
W^- = \ln \left( \frac{N_{pix2}}{N_{pix1} + N_{pix2}} \right) \ln \left( \frac{N_{pix4}}{N_{pix3} + N_{pix4}} \right)
\]

(2)

where \(N_{pix1}\) is the number of flood pixels within a class; \(N_{pix2}\) is the number of flood pixels outside the specific class; \(N_{pix3}\) is the number of pixels in a class without flood locations; \(N_{pix4}\) is the number of non-flood pixels outside a class; \(W^+\) is equal to the positive weight, while \(W^-\) represents the negative weight.

The final values of WOE coefficients (\(W_f\)) will be calculated using the next relation (Dahal et al. 2008):

\[
W_f = W_{plus} + W_{mintotal} - W_{min}
\]

(3)

where \(W_{plus}\) is the positive weight of a class/category from a factor, \(W_{min}\) is the negative weight of a category belonging to a specific factor, \(W_{min\ total}\) represents the sum of all the negative weights in a multiclass map.

The weights of evidence values were used as input data in the data mining models used in the present research. More aspects of Weights of Evidence can be found in Sect. 5 (S5) of Online Supplementary Material.

4.3 UltraBoost (U)

UltraBoost classifier represents a method used to boost heterogeneous learners. In the present research work, the UltraBoost algorithm is trained so as for each case to be assigned a specific probability of flooding (Moustafa et al. 2020). The attribute class forecast with the help of model is flood=1 for the areas highly exposed to flood and 0 for non-flood — so the output probabilities of flood genesis is ranging between 0 and 1. Additional aspects of UltraBoost model are inserted in Sect. 6 (S6) of Online Supplementary Material.

4.4 Stochastic Gradient Descending (SGD)

The stochastic gradient descent method is a stochastic approximation of the gradient descent method, which is an iteration-based method for generating a derivative function (Hoang 2019). In addition to its high efficiency, and ease with which it can be implemented for datasets with redundant samples, this algorithm is very popular. In the flood susceptibility analyses, SGD is seldom applied, which is something that needs to be investigated more in-depth. The application of SGD in the present research work was possible through Weka 3.9 software. Additional aspects of the Stochastic Gradient Descending (SGD) model are presented in Sect. 7 (S7) of Online Supplementary Material.
4.5 Cost Sensitive Forest (CSForest)

The CSForest can be applied through the mean of the Classification Cost Reduction (CCR) function which replaces the Gain Ratio that was used as a splitting criterion (Siers and Islam 2015). The tree built by CSForest, before pruning, is deeper than that by CSTree. CSForest, by contrast, builds an ensemble of trees and then uses CSVoting for classifying records/modules whereas the CSTree builds a single tree. In the present case study, the CSForest, for flood susceptibility modelling, was run in Weka 3.9 software. Additional aspects of Cost Sensitive Forest (CSForest) algorithm are included in Sect. 8 (S8) of Online Supplementary Material.

4.6 Results Validation

4.6.1 Receiver Operating Characteristic (ROC) Curve

ROC Curve suggests that a model can accurately predict flood susceptibility. Area Under Curve (AUC) measures the value of a curve of ROCs which is helpful for the user in getting the most important information (Xie et al., 2021a). The range of the AUC is between 0 and 1 (Xie et al., 2021b). Using the following mathematical relationship, we can estimate the AUC:

\[
AUC = \frac{(\sum TP + \sum TN)}{(P + N)}
\]

where TP (true positive), TN (true negative) are the sums of flood and non-flood locations correctly classified.

Additional aspects of ROC Curve are included in Sect. 9 (S9) of Online Supplementary Material.

4.6.2 Statistical Metrics

The next 5 statistical indicators were involved also in the results validation procedure: Accuracy, Specificity, Sensitivity, Kappa Index, and Precision. In order to calculate the statistical metrics, we will be using the following equations (Canbek et al. 2017):

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{Specificity} = \frac{TN}{FP + TN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]
where $P$ is the sum of flood locations, $N$ is the sum of non-flood locations, $FP$ (false positive) and $FN$ (false negative) are the numbers of flood and non-flood locations erroneously classified, $k$ is the kappa coefficient, $p_o$ is the observed flood locations, and $p_e$ is the estimated flood susceptibility locations.

A short description of the workflow followed in the present study is represented in Fig. 4.

### 5 Results

#### 5.1 Feature Selection Using ReliefF Method

The application of ReliefF method provided the following outcomes: Slope (0.653), Land use (0.390), Distance from river (0.354), Elevation (0.32), Plan curvature (0.318), Rainfall (0.246), SPI (0.22), HSG (0.207), TWI (0.161), Profile curvature (0.140), Lithology (0.138), TPI (0.048), Convergence Index (0.021) and Aspect (0.01) (Fig. S1). According to the obtained values, all the flood predictors have an influence on the flood triggering process, and therefore, all of them were considered in the workflow analysis.
5.2 Weights of Evidence Coefficients

The WOE value is also a positive value as well as a negative value, and consequently, this procedure was necessary. The WOE of 3.56 assigned to the river land use class was the highest value and was followed by elevation below 200 m (3.1) and the class of distance from river lower than 50 m (2.9) (Table S2). The WOE was used as an input dataset in the three machine learning models.

5.3 Ultraboost-Weights of Evidence (U-WOE)

The U-WOE ensemble model was trained through a ten-fold cross-validation method in Weka software. Following the training process, the Mean Absolute Error (MAE) was equal to 0.187, while the Root Mean Squared Error (RMSE) was 0.1991. The Precision revealed after the training procedure was equal to 0.989. A very important output of the training procedure is represented by the flood predictors importance which have the next values: Slope (0.162), SPI (0.115), Convergence index (0.105), Elevation (0.102), Rainfall (0.091), Distance from rivers (0.09), HSG (0.078), TPI (0.077), Plan curvature (0.072), Aspect (0.049), Lithology (0.027), Profile curvature (0.02), Land use (0.007) and TWI (0.004) (Fig. S2).

The importance values were imported in ArcGIS Map Algebra and the Flood Susceptibility Index (FSI) was calculated by the multiplication with WOE coefficients. The results of $FSI_{U-WOE}$ were reclassified into 5 classes through the help of Natural Breaks method. The first class, that indicates the very low flood susceptibility values, appears on around 21% of the study area (Fig. 5a). Further, another 35.38% of the research area is covered by low $FSI_{U-WOE}$ values, while medium flood susceptibility has a total percentage of 19.41%. The high and very high flood susceptibility zones account for approximately 16% of the Putna river basin.

5.4 Stochastic Gradient Descending – Weights of Evidence (SGD-WOE)

The SGD-WOE model was also run in Weka software using the ten-fold cross-validation procedure. The next parameters were established in order to train the SGD-WOE ensemble: batch Size – 100; epochs – 100; epsilon – 0.01; lambda – 1.04E-4; learning rate – 0.01; seed – 1. The training procedure revealed the achievement of a Mean Absolute Error (MAE) of 0.131, while the Root Mean Squared Error (RMSE) was 0.145. The model Precision reached 0.991. The final step in the training of SGD-WOE model was the determination of the next values of flood predictors importance: Slope (0.305), Distance from rivers (0.127), TWI (0.094), SPI (0.087), Convergence index (0.079), Lithology (0.078), Elevation (0.074), Profile curvature (0.061), Land use (0.034), Rainfall (0.032), Aspect (0.012), TPI (0.007), HSG (0.006), Plan curvature (0.002).

These values were used in the GIS environment through Map Algebra in order to estimate the flood susceptibility index. The very low flood susceptibility is presented on 41.62% of the Putna river catchment (Fig. 5b). The low flood susceptibility is spread on 33.04% of the study zone, while the medium values of flood exposure have only 4.28% of the Putna river basin. A total of 21% of the study area is characterized by high and very high flood susceptibility values.
5.5 Cost Sensitive Forest – Weights of Evidence (CSForest-WOE)

In order to train the CSForest-WOE model, a ten-fold cross-validation algorithm was applied. The structure that provides the best performance for CSForest was formed by 5 trees (Fig. S3). The value of Mean Absolute Error (MAE) was 0.095, while the Root Mean Squared Error (RMSE) was equal to 0.127. The very high performance of the CSForest-WOE algorithm was revealed by the Precision of 0.993.

Further, the next hierarchy was revealed among the flood predictors: Slope (0.158), Distance from river (0.119), Elevation (0.108), SPI (0.094), Rainfall (0.086), Convergence index (0.082), Land use (0.068), Lithology (0.062), HSG (0.05), TWI (0.049), TPI (0.042), Plan curvature (0.04), Aspect (0.026) and Profile curvature (0.016).

The flood susceptibility values were split into 5 classes through the mean of Natural Breaks method. The very low values of FSI\textsubscript{CSForest-WOE} occupy around 26.51% of the Putna river basin, while the low flood susceptibility characterizes 34.31% of the research zone (Fig. 5c). Medium flood susceptibility has a total percentage of 14.93%, while the high class of FSI\textsubscript{CSForest-WOE} covers 7.67% of the area on which the present study is focused. The very high class reaches a total percentage of 16.57% of the study zone.

5.6 Results validation

5.6.1 ROC Curve

The Success Rate associated with flood susceptibility maps shows that FSI\textsubscript{SGD-WOE} has the highest performance with an AUC of 0.953, followed by FSI\textsubscript{U-WOE} (AUC=0.951) and FSI\textsubscript{CSForest} (AUC=0.95) (Fig. 6a). The results related to the Prediction Rate revealed that the best flood susceptibility maps were associated to the FSI\textsubscript{SGD-WOE} and FSI\textsubscript{U-WOE} both of them having an AUC of 0.947. The FSI\textsubscript{CSForest-WOE} has the lowest performance being characterized by an AUC of 0.945 (Fig. 6b).

5.6.2 Statistical Metrics

In terms of the training sample, the highest accuracy was achieved by SGD-WOE (0.94), followed by U-WOE (0.935) and CSForest-WOE (0.929). Regarding the results for k-index, it should be noted that the best value was obtained by SGD-WOE (0.88), followed by U-WOE (0.87) and CSForest-WOE (0.859). In regards to the validating sample, it should be seen that the maximum accuracy of 0.938 was attributed to SGD-WOE model, followed by U-WOE (0.925) and CSForest (0.913). The k-index revealed the same hierarchy with SGD-WOE in the first place with 0.875, followed by U-WOE (0.85) and CSForest-WOE (0.825) (Table 1).

6 Discussions

There is no doubt that atmospheric emissions of greenhouse gases have been increasing in the past few decades, and these emissions have become increasingly pronounced because of the occurrence of weather-related and climate-related events such as flooding, droughts,
Fig. 5 Flood Susceptibility Index (a. U-WOE; b. SGD-WOE; c. CSForest-WOE)
Fig. 6 ROC Curves (a. Success Rate; b. Prediction Rate)
storm surges, and sea-level rises (Dimeyeva et al. 2015; Yin et al. 2022b). Natural risk evaluation research works also include the flood susceptibility models, which are among the most challenging. Flood susceptibility prediction accuracy is strongly influenced by the effects of particular conditions relevant to the study area(s) and by the method used in making the prediction (Islam et al. 2021). As in many other similar studies, in this paper, the authors wanted to estimate the susceptibility to floods by comparative use of 3 machine learning models. These were also combined with a bivariate statistical method, the latter allowing uniform input data to be obtained for all flood predictors. The ensemble approach used in flood susceptibility mapping between researchers has been quite useful in some cases. An ensemble model was applied for bivariate and multivariate statistical analysis by Bui et al. (2019), and they found the ensemble model to be better than the individual models. The 3 machine learning models are represented by UltraBoost, Stochastic Gradient Descending and Cost Sensitive Forest, while the bivariate statistics method was represented by Weights of Evidence. The performance of the 3 models of flood susceptibility was performed using the ROC curves, the Area Under the Curve (AUC) and several statistical metrics (sensitivity, specificity, accuracy and k-index) results. All the results showed that the models were able to perform reasonably well in assessing flood-prone areas. All 3 models had associated AUC values greater than 0.945, which indicates a very high performance and a high accuracy of flood susceptibility maps. These performances far exceeded the performance obtained by Nachappa et al. (2020), in which ensemble methods between machine learning and bivariate statistics were also used and in which all AUC values were below 0.9. The aforementioned study focused on the Salzburg province from Austria which has a multitude of common climatic and relief characteristics as in the high and middle altitude regions of the present study area. Another study whose central topic was the determination of flood susceptibility using machine learning combined with bivariate statistics was conducted by Liuzzo et al. (2019) in Devon County, in the United Kingdom. And in this case, the AUC values obtained by the authors of the previously mentioned study were generally more than 0.91, being, therefore, lower than those obtained in the present study. This demonstrates once again that the use of the 3 algorithms in the present study may be more efficient than other machine learning models used in the literature. Another example of a study on flood susceptibility calculated by state-of-the-art algorithms was conducted by Wang et al. (2020).

The authors of the study used a combination of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) to estimate flood exposure in Shangyou County from China. In this case, the highest AUC value was 0.937 and was associated with the 2D-CNN algorithm. Even in this situation where very powerful computational algorithms were used

| Table 1 Statistical metrics for the applied flood susceptibility models |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Metrics        | CSForest-WOE    | SGD-WOE         | U-WOE           | CSForest-WOE    | SGD-WOE         | U-WOE           |
| TP             | 85              | 87              | 88              | 36              | 37              | 38              |
| TN             | 86              | 86              | 84              | 37              | 38              | 36              |
| FP             | 7               | 5               | 4               | 4               | 3               | 2               |
| FN             | 6               | 6               | 8               | 3               | 2               | 4               |
| Sensitivity (%)| 0.934           | 0.935           | 0.917           | 0.923           | 0.949           | 0.905           |
| Specificity (%)| 0.925           | 0.945           | 0.955           | 0.902           | 0.927           | 0.947           |
| Accuracy (%)   | 0.929           | 0.940           | 0.935           | 0.913           | 0.938           | 0.925           |
| k-index        | 0.859           | 0.880           | 0.870           | 0.825           | 0.875           | 0.850           |

© Springer
in estimating flood susceptibility, the performance of the models applied in the present study was not exceeded.

The findings of the present study are generally expected to have a significant impact on refining flood risk management plans and activities as well as the principles in the region for making certain that additional damages caused by flooding are avoided.

7 Conclusions and Future Directions

The central topic of this scientific article is the estimation of flood susceptibility through 3 advanced machine learning models in the hydrographic area of the Putna River Basin in Romania. The combination of the bivariate statistical method represented by Weights of Evidence with the following 3 machine learning models allowed the generation of ensemble models that provided very good results: UltraBoost, Stochastic Gradient Descendent and Cost Sensitive Forest. The application of ROC Curve and Area Under Curve showed that the best performance was obtained by SGD-WOE ensemble whose AUC value was 0.953. Also, statistical metrics showed that the same model obtained the highest Accuracy, equal to 0.94. According to the flood susceptibility maps generated by the three ensemble models, the surfaces with a high and very high degree of flood exposure are found on surfaces that fall between 21.04% (FSI\textsubscript{SGD-WOE}) and 24.24% (FSI\textsubscript{CSForest-WOE}). These areas with high and very high exposure to floods are spread mainly in the plain area of the study area but also along the main valleys in this region.

The main novelty of the present study is the application for the first time in the literature of the three ensemble models in order to estimate the susceptibility to floods. Their very high performance is a strong reason for these three ensemble models to be taken over in other research papers and applied to other study areas. At the same time, the authors of the study consider, as future research directions, the integration of the methodology for estimating the susceptibility to floods within a wider methodology for estimating the risk of these phenomena. Thus, in future works, the socio-economic objectives of the studied area will also be taken into account. It is also considered to carry out detailed case studies on floodplains through hydraulic modeling and the use of digital land models with high spatial resolution.

Supplementary Information  The online version contains supplementary material available at https://doi.org/10.1007/s11269-022-03276-0.

Author Contributions  Conceptualization, R.C., I.C., A.C. and B.T.P.; methodology, R.C., A.A. and I.C.; software, R.C., I.C., and B.T.P.; validation, R.C., I.C., A.C. and B.T.P.; formal analysis, R.C., A.A., A.C. and I.C.; investigation, R.C., A.A. and I.C.; resources, R.C., I.C. and B.T.P.; data curation, R.C., I.C. and B.T.P.; writing—original draft preparation, R.C., I.C. and B.T.P.; writing—review and editing, R.C., A.A. and I.C.; visualization, R.C. and B.T.P.; supervision, R.C., A.A. and I.C.; project administration, R.C.; funding acquisition, R.C. and B.T.P. All authors have read and agreed to the published version of the manuscript.

Funding  This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS – UEFISCDI, project number PN-III-P1-1.1-PD-2019-0424-P, within PNCDI III.

Availability of Data and Materials  The data used in this research are available by the corresponding author upon reasonable request.
Declarations

Ethics Approval The authors confirm that this article is original research and has not been published or presented previously in any journal or conference in any language (in whole or in part).

Consent to Participate and Consent to Publish Not applicable.

Competing Interests The authors have no conflict of interest.

References

Afriyanie D, Julian MM, Riqqi A, Akbar R, Suroso DS, Kustiwan I (2020) Re-framing urban green spaces planning for flood protection through socio-ecological resilience in Bandung City. Indonesia Cities 101:102710

Allahbakhshian-Farsani P, Vafakhah M, Khoosavi-Farsani H, Hertig E (2020) Regional flood frequency analysis through some machine learning models in semi-arid regions. Water Resour Manage 34:2887–2909

Arabameri A, Saha S, Chen W, Roy J, Pradhan B, Bui DT (2020) Flash flood susceptibility modelling using functional tree and hybrid ensemble techniques. Journal of Hydrology 587:125007 https://doi.org/10.1016/j.jhydrol.2020.125007

Armaş I (2012) Weights of evidence method for landslide susceptibility mapping. Prahova Subcarpathians Romania Natural Hazards 60:937–950

Arnell NW, Gosling SN (2016) The impacts of climate change on river flood risk at the global scale. Clim Change 134:387–401

Bui DT, Tsangaratos P, Ngo P-TT, Pham TD, Pham BT (2019) Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. Sci Total Environ 668:1038–1054

Bui Q-T, Nguyen Q-H, Nguyen XL, Pham VD, Nguyen HD, Pham V-M (2020) Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. J Hydrol 581:124379. https://doi.org/10.1016/j.jhydrol.2019.124379

Canbek G, Sagiroglu S, Temizel TT, Baykal N, Science, Engineering (2017) (UBMK), IEEE, pp. 821–826

Chakrabortty R, Pal SC, Janizadeh S, Santosh M, Roy P, Chowdhuri I, Saha A (2021) Impact of climate change on future flood susceptibility: an evaluation based on deep learning algorithms and GCM model. Water Resour Manage 35:4251–4274

Chao L, Zhang K, Wang J, Feng J, Zhang M (2021). A comprehensive evaluation of five evapotranspiration datasets based on ground and satellite observations: Implications for improvement of evapotranspiration retrieval algorithm. Remote Sens 13(12): 2414.

Chen Y, Xu Y, Yin Y (2009) Impacts of land use change scenarios on storm-runoff generation in Xitiaoxi basin, China. Quatern Int 208:121–128

Chen W, Li Y, Xue W, Shahabi H, Li S, Hong H, Wang X, Bian H, Zhang S, Pradhan B (2020) Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Sci Total Environ 701:134979

Chen X, Quan Q, Zhang K, Wei J (2021) Spatiotemporal characteristics and attribution of dry/wet conditions in the Weihe River Basin within a typical monsoon transition zone of East Asia over the recent 547 years. Environ Model Softw 143:105116.

Costache R, Bui DT (2019) Spatial prediction of flood potential using new ensembles of bivariate statistics and artificial intelligence: A case study at the Putna river catchment of Romania. Sci Total Environ 691:1098–1118

Costache R, Hong H, Wang Y (2019) Identification of torrential valleys using GIS and a novel hybrid integration of artificial intelligence, machine learning and bivariate statistics. CATENA 183:104179

Costache R, Pham QB, Arabameri A, Diaconu DC, Costache I, Crăciun A, Ciobotaru N, Pandey M, Arora A, Ali SA (2021) Flash-flood propagation susceptibility estimation using weights of evidence and their novel ensembles with multicriteria decision making and machine learning. Geocarto International 1–33

Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Masuda T, Nishino K (2008) GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. Environ Geol 54:311–324
New Machine Learning Ensemble for Flood Susceptibility Estimation

Dimeyeva L, Sitpayeva G, Sultanova B, Ussen K, Islamgulova A (2015) High-altitude flora and vegetation of Kazakhstan and climate change impacts. Climate Change Impacts on High-Altitude Ecosystems. Springer, pp 1–48

Dodangeh E, Choubin B, Egidir AN, Nabipour N, Panahi M, Shamshirband S, Mosavi A (2020) Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. Sci Total Environ 705:135983

Gao C, Hao M, Chen J, Gu C (2021) Simulation and design of joint distribution of rainfall and tide level in Wuchengxiyu Region, China. Urban clim. 40:101005.

Hoang N-D (2019) Automatic detection of asphalt pavement raveling using image texture based feature extraction and stochastic gradient descent logistic regression. Autom Constr 105:102843

Hong H, Panahi M, Shirzadi A, Ma T, Liu J, Zhu A-X, Chen W, Kougias I, Kazakos N (2018) Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. Sci Total Environ 621:1124–1141

Islam ARMT, Talukdar S, Mahato S, Kundu S, Eibek KU, Pham QB, Kuriqi A, Linh NTT (2021) Flood susceptibility modelling using advanced ensemble machine learning models. Geosci Front 12:101075

Jafari-Asl J, Seghier MEAB, Ohadi S, van Gelder P (2021) Efficient method using Whale Optimization Algorithm for flood susceptibility modeling. Appl Soft Comput 101:107036

Janizadeh S, Vafakhah M, Kapelan Z, Dinan NM (2021) Novel Bayesian additive regression tree methodology for flood susceptibility modeling. Water Resour Manage 35:4621–4646

Lee S, Kim J-C, Jung H-S, Lee MI, Lee S (2017) Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk 8, 1185–1203

Li X, Li X, Wang Y, Hu Y, Zhou C, Zhang H (2022) Numerical Investigation on Stratum and Surface Deformation in Underground Phosphorite Mining Under Different Mining Methods. Front Earth Sci 10:831856. https://doi.org/10.3389/feart.2022.831856

Liu Y, Zhang K, Li Z, Liu Z, Wang J, Huang P (2020) A hybrid runoff generation modelling framework based on spatial combination of three runoff generation schemes for semi-humid and semi-arid watersheds. J Hydrol 590:125440.

Liu S, Liu Y, Wang C, Dang X (2022) The Distribution Characteristics and Human Health Risks of High-Fluoride Groundwater in Coastal Plain: A Case Study in Southern Laizhou Bay, China. Front Environ Sci 10:901637. https://doi.org/10.3389/fenvs.2022.901637.

Liuzzo L, Sammartano V, Freni G (2019) Comparison between different distributed methods for flood susceptibility mapping. Water Resour Manage 33:3155–3173

Li Y, Zhang Z, Liu X, Wang L, Xia X (2021a) Efficient image segmentation based on deep learning for mineral image classification. Adv Powder Technol 32(10): 3885-3903.

Li Y, Zhang Z, Liu X, Wang L, Xia X (2021b) Ore image classification based on small deep learning model: Evaluation and optimization of model depth, model structure and data size. Miner Eng 172:107020.

Mehta N, Devarakonda MV (2018) Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey? J Allergy Clin Immunol 141:2019–2021

Moustafa AF, Cary TW, Sultan LR, Schultz SM, Conant EF, Venkatesh SS, Sehgal CM (2020) Color doppler ultrasound improves machine diagnosis of breast cancer. Diagnostics 10:631

Nachappa TG, Piralilou ST, Gholamnia K, Ghorbanzadeh O, Rahmati O, Blaschke T (2020) Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using Dempster Shafer Theory. Journal of Hydrology125275

Pistrika A, Tsakiris G, Nalbantis I (2014) Flood depth-damage functions for built environment. Environ Processes 1:553–572

Quan Q, Liang W, Yan D, Lei J (2022) Influences of joint action of natural and social factors and hydrological cycle on hydrological cyclic in Inner Mongolia, China. Urban clim. 41:101043.

Rad MJG, Ohadi S, Jafari-Asl J, Vatani A, Ahmadabadi SA, Correia JA (2022) GNDO-SVR: An efficient surrogate modeling approach for reliability-based design optimization of concrete dams. Presented at the Structures, pp 722–733

Romanescu G, Nistor I (2011) The effects of the July 2005 catastrophic inundations in the Siret River’s Lower Watershed, Romania. Nat Hazards 57:345–368

Siers MJ, Islam MZ (2015) Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem. Inform Syst 51:62–71.

Tehrany MS, Pradhan B, Jebur MN (2015) Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stoch Env Res Risk Assess 29:1149–1165

Tian J, Liu Y, Zheng W, Yin L (2021) Smog prediction based on the deep belief-BP neural network model (DBN-BP). Urban Clim. 41:101078

Timovzan A, Cociog GM, Romanescu G, Obreja F (2014) Predicting the potential index of major floods production in the Suha river basin (Suha Bucovineana). Water resources and wetlands
Vojtek M, Vojteková J (2019) Flood susceptibility mapping on a national scale in Slovakia using the analytical hierarchy process. Water 11:364

Wang Y, Fang Z, Hong H, Peng L (2020) Flood susceptibility mapping using convolutional neural network frameworks. J Hydrol 582:124482

Wang S, Zhang K, Chao L, Li D, Tian X, Bao H, Chen G, Xia Y. (2021) Exploring the utility of radar and satellite-sensed precipitation and their dynamic bias correction for integrated prediction of flood and landslide hazards. J Hydrol 603: 126964.

Yin L, Wang L, Keim BD, Konsoer K, Zheng W (2022a) Wavelet Analysis of Dam Injection and Discharge in Three Gorges Dam and Reservoir with Precipitation and River Discharge. Water 14(4):567.

Zhan C, Dai Z, Soltanian MR, Zhang X (2022a) Stage-Wise Stochastic Deep Learning Inversion Framework for Subsurface Sedimentary Structure Identification. Geophys Res Lett 49(1):e2021GL095823. https://doi.org/10.1029/2021GL095823

Yin L, Wang L, Huang W, Tian J, Liu S, Yang B, Zheng W (2022b) Haze Grading Using the Convolutional Neural Networks. Atmosphere 13(4):522.

Xie W, Li X, Jian W, Yang Y, Liu H, Robledo LF, Nie W (2021a) A Novel Hybrid Method for Landslide Susceptibility Mapping-Based GeoDetector and Machine Learning Cluster: A Case of Xiaojin County, China. ISPRS Int J Geo-Inf 10(2):93.

Xie W, Nie W, Saffari P, Robledo LF, Descote P, Jian W (2021b) Landslide hazard assessment based on Bayesian optimization--support vector machine in Nanping City, China. Nat Hazards 109(1):931-948.

Zhan C, Dai Z, Samper J, Yin S, Ershadnia R, Zhang X, Wang Y, Yang Z, Luan X, Soltanian MR (2022b) An integrated inversion framework for heterogeneous aquifer structure identification with single-sample generative adversarial network. J Hydrol 610:127844.

Zhang K, Wang S, Bao H, Zhao X (2019) Characteristics and influencing factors of rainfall-induced landslide and debris flow hazards in Shaanxi Province, China. Nat Hazards Earth Syst Sci 19(1):93-105.

Zhang Z, Luo C, Zhao Z (2020) Application of probabilistic method in maximum tsunami height prediction considering stochastic seabed topography. Nat Hazards 104:2511-2530.

Zhang X, Ma F, Yin S, Wallace CD, Soltanian MR, Dai Z, Ritzi RW, Ma Z, Zhan C, Lü X (2021) Application of upsampling methods for fluid flow and mass transport in multi-scale heterogeneous media: A critical review. Appl Energy, 303:117603.

Zhang K, Shalehy MH, Ezaz GT, Chakraborty A, Mohib KM, Liu L (2022). An integrated flood risk assessment approach based on coupled hydrological-hydraulic modeling and bottom-up hazard vulnerability analysis. Environ Model Softw 148:105279.

Zhao T, Shi J, Lv L, Xu H, Chen D, Cui Q, Jackson TJ, Yan G, Jia L, Chen L, Zhao K, Zheng X, Zhao L, Zheng C, Ji D, Xiong C, Wang T, Li R, Pan J, Wen J, Yu C, Zheng Y, Jiang L, Chai L, Hu L, Yao P, Ma J, Lv H, Wu J, Zhao W, Yang N, Guo P, Li Y, Hu L, Geng D, Zhang Z (2020). Soil moisture experiment in the Luan River supporting new satellite mission opportunities. Remote Sens Environ 240:111680.

Zhou G, Li C, Zhang D, Liu D, Zhou X, Zhan J (2021a) Overview of Underwater Transmission Characteristics of Oceanic LiDAR. IEEE J Sel Top Appl Earth Obs Remote Sens 14:8144-8159.

Zhou G, Long S, Xu J, Zhou X, Song B, Deng R, Wang C (2021b) Comparison Analysis of Five Waveform Decomposition Algorithms for the Airborne LiDAR Echo Signal. IEEE J Sel Top Appl Earth Obs Remote Sens 14:7869-7880.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Romulus Costache1,2 · Alireza Arabameri3 · Iulia Costache4 · Anca Crăciun2 · Binh Thai Pham5

Springer
New Machine Learning Ensemble for Flood Susceptibility Estimation

Iulia Costache
iulia.elena.costache@gmail.com

Anca Crăciun
anca.craciun@ddni.ro

Binh Thai Pham
binhpt@utt.edu.vn

1 Department of Civil Engineering, Transilvania University of Brasov, 5, Turnului Str, 500152 Brasov, Romania
2 Danube Delta National Institute for Research and Development, 165 Babadag Street, 820112 Tulcea, Romania
3 Department of Geomorphology, Tarbiat Modares University, 36581-17994 Tehran, Iran
4 Faculty of Geography, University of Bucharest, 010041 Bucharest, Romania
5 University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, 100000 Ha Noi, Vietnam