Flood susceptibility computation using state-of-the-art machine learning and optimization algorithms

Romulus-Dumitru Costache (✉ romuluscostache2000@yahoo.com)

Universitatea Transilvania din Brasov <https://orcid.org/0000-0002-6876-8572>

Alireza Arabameri

Tarbiat Modares University

Abu Reza Md. Towfiqul Islam

Begum Rokeya University

S.I. Abba

King Fahd University of Petroleum & Minerals

Manish Pandey

Chandigarh University

R.S. Ajin

Government of Kerala

Binh Thai Pham

University of Transport Technology

Research Article

Keywords: flood susceptibility, machine learning, optimization, Romania, bivariate statistics

Posted Date: March 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1405369/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The present study aims to estimate the flood susceptibility degree over the Prahova River basin located in the central-southern part of Romania. To obtain the proposed outcomes, the next 10 flood predictors were used as independent variable in the machine learning models: slope angle, convergence index, distance from river, elevation, plan curvature, hydrological soil group, lithology, topographic wetness index, rainfall and land use. The factors along with 158 flood locations, that represent the dependent variables, were involved in the training procedure of the following four ensemble models: Deep Learning Neural Network –Statistical Index (DLNN-SI), Particle Swarm Optimization-Deep Learning Neural Network–Statistical Index (PSO-DLNN-SI), Support Vector Machine–Statistical Index (SVM-SI) and Particle Swarm Optimization-Support Vector Machine–Statistical Index (PSO-SVM-SI). Through the Statistical Index method, the coefficients of each flood predictor class/category were calculated. The best performance was achieved by PSO-DLNN-SI model for which an AUC-ROC Curve of 0.952 was calculated. It is worth to note the application of PSO algorithm manage to increase the models performance. Also, it is important to note that around 25% of the study area has a high and very high exposure to flood phenomena.

1. Introduction

On a global scale, floods constitute the major natural hazards causing the greatest incidence of damaged property as measured by the number of material damages caused by this phenomenon (Jacinto et al., 2015). The hazards associated with floods therefore pose a great challenge for each and every one of the authorities concerned with the flood risk management activity (Luino et al., 2018). As a result of the climatic changes that have taken place over the past few decades, it has become evident that floods have become more frequent and severe, but it has also contributed to a marked increase in the amount of material damage and human losses (Peng-cheng et al., 2016). This is one of the reasons why the number of scientific publications that address the topic of flood risk management has increased steadily in the past few years. Romania ranks second behind France in terms of the economic damage caused by floods at an EU level. Romania must, like all countries in Europe, ensure that its flood risk management program is in compliance with the Directive 2007/60/EC. This document suggests that, in order to manage flood risks in communities, a great deal of attention should be paid to identifying areas prone to flooding (Afriyane et al., 2020). Furthermore, the evaluation of flood vulnerability is likewise one of the most critical non-structural measures implemented in order to reduce material losses and lives that result from natural disasters such as flooding (Arora et al., 2021). It is possible to quickly determine whether a given area is vulnerable to flooding by using the latest Geographic Information System technologies. It is clear that the efficiency of any GIS process will be determined, in a considerable measure, by the accuracy of its input data, as well as by the combination of GIS models with statistical and machine learning algorithms. A growing number of studies that have used GIS techniques in combination with some models specific to machine learning and bivariate statistics to identify and map flood-exposed zones, have been published in recent years (Ahmadlou et al., 2021). There are three bivariate statistical models most commonly used to develop models for evaluating flood susceptibility: Weights of Evidence,

Statistical Index and Frequency Ratio (Sahana and Patel, 2019). There is a fundamental limitation with the bivariate statistical models because they consider only the spatial relation between the flood locations and the conditioning factors, but they do not consider the causal relationship between the predictors governing the flow of the floods (Costache, 2019a). Among the machine learning methods used to detect those surfaces showing flood prone characteristics, the most commonly used ones are: decision trees algorithms, adaptive neuro-fuzzy inference systems, support vector machines and artificial neural networks (Alizadeh et al., 2018; Khosravi et al., 2018; Xiong et al., 2019). One of the key advantages characteristics for machine learning models resides in the high automation that they provide, as well, as the ease with which they can identify patterns and trends into a data sample. Additionally, machine learning models can handle multiple diversity, as well as multi-dimensional, data sets (Elmahdy et al., 2020). Even though the machine learning models constitute the most advanced aspects of the data processing field, they also have certain disadvantages that make them less attractive. In the literature, there have been a great deal of discussions regarding two of the most significant drawbacks of this kind of technique: the large amount of data required and the high error susceptibility (Albon, 2018; Ao et al., 2019; Baghban et al., 2018).

In most cases, the authors of studies regarding assessment of flood potential for a particular region consider a multitude of geographic factors that influence the nature and severity of this natural disaster. Usually, the authors take into account the following flood predictors: rainfall, distance from rivers, soils, slope angle, land use, plan curvature, topographic wetness index and lithology (Chowdhuri et al., 2020). It is possible to compute the flood susceptibility by overlaying in a simple manner the flood in GIS software and analyzing their contributions. Advance models that allow the weighting of these factors are also able to determine flood susceptibility (Bui et al., 2020). As a general rule, flood conditioning factors are weighted according to their spatial relationships with historical flood occurrences (Azareh et al., 2019). As such, the locations of recent flood events are considered when giving local weights to these factors (Dano et al., 2019).

Taking into account these details, the main purpose of the present study is to identify the flood-prone areas in Prahova river's hilly and mountain basins, by using four models. Thus, the Deep Learning Neural Network (DLNN), Support Vector Machine (SVM) and also these 2 optimized models with Particle Swarm Optimization (PSO) were used to estimate the flood susceptibility.

As a part of the training of the models mentioned above, a number of 10 flood predictors, as well as 158 flood historical events, will be taken into account. One of the huge advantages of this approach is the possibility of evaluating and validating, quantitatively, the precision of the algorithms involved in the methodological workflow and also to assess the accuracy of the flood susceptibility maps. In the present study, the accuracy of the results has been tested by computing several statistical parameters and by plotting the ROC curve and calculating the area under the curve.

2. Study Area

It was chosen to carry out the study on the upper and middle basin of the Prahova river in Romania (Fig. 1). It is estimated that the area of the study zone is roughly 2600 km² and its altitude ranges from 128 to 2505 meters. During the period of the year, the distribution of rainfall on the surface of the river catchments has a great influence on the hydrological regime of the rivers. During the summer season, there is a variable amount of rainfall in the area, which frequently causes flash floods and flooding. Approximately 12.43° is the mean value of slope angle, one of the most powerful factors that influence flooding processes.

It is also important to note that the areas covered by soil groups C and D, which are also present in the mountainous and hilly regions of the Prahova river basin, favor the occurrence of floods as well. Using hydrometric information from the research zone, Costache (2019b) has synthesized a review of the recent flood events.

3. Data

3.1. Flood inventory

An inventory of flood-prone locations affected in the past by flood phenomena was performed to get a clearer understanding of how geographic factors influence the flooding process. This information was taken from the archives of the General Inspectorate for the Situations of Emergencies of Romania. In the study area, there were 158 flood events recorded as flooding events. Also, 158 non-flooding locations were chosen at random to grow the objectivity of this research work. These locations are either located in interfluvial zones or within very steep slopes where flooding is not likely to occur. Taking into account the literature (Tehrany et al., 2017), all the dataset was divided in a random manner into training (70%) and validating points (30%) (Fig. 1). Finally, the training sample contains 110 flood locations and 110 non-flood locations, while the validating sample contains 48 flood locations and 48 non-flood locations.

3.2. Flood predictors

As further step in this attempt to obtain an accurate spatial estimation of flood susceptibility values, the choice of the geographical factors that best explain the flooding process is another important factor (Dahri and Abida, 2017). As part of this analysis, 10 flood predictors, which are described in the next rows, were considered.

According to past research papers (Bui et al., 2019), slope angles obtained with a Digital Elevation Model are one of the most influential factors that control the amount of surface runoff in a given area. The majority of flood events occur on slopes less than three degrees in the study zone. Around forty percent of the research zone have slopes ranging in intensity from seven to fifteen degrees (Fig. 2a).

A number of researchers have documented the effects of elevation on flooding processes (Kanani-Sadat et al., 2019). Lower zones are likely to accumulate water because of their lower elevations. A total of eight classes of elevation were established within the study area (Fig. 2d). The class from 200 m to 400 m

occupies the highest percentage (25%) of the study area, while around 70% of flood locations are found at altitudes from 200 m to 600 m.

The distances from rivers were calculated using the Euclidean Distance algorithm. It has been determined that the distance from rivers factor can be classified into 8 intervals, based on the literature (Chapi et al., 2017) (Fig. 2c). It is interesting to note that more than 40% of the flooding events took place in the first 50 meters from the river network.

Corine Land Cover (2018) European database was used to extract the land use factor. The area under study was categorized into 10 land use categories (Fig. 2e). Approximately 50% of the area consists of forests, while 50% of the flood pixels belong to the built-up areas.

The hydrological soil types play a major role in determining the water infiltration process. There is a presence of all four soil groups across the study area (Fig. 3d). There were more floods in hydrological group C (68%) than any other group.

In the same way that soil groups affect flooding, lithology controls the infiltration of water. There were identified 12 lithological categories within the study zone (Fig. 2f) based on the Geological Map of Romania, scale 1:200,000. Approximately 80% of all flood pixels tend to be located within the clays and sandstones areas.

There is a direct relationship between rainfall and flood genesis. The IDW method was employed to interpolate data from 21 hydrometric stations and nine meteorological stations (1980–2017) for the purpose of analyzing the spatial distribution of the rainfall (Fig. 3a). A range of 531 mm/year to 1250 mm/year was split into seven rainfall classes.

TWI values are calculated by dividing a specific basin area by its slope. Five intervals were defined for the research perimeter (Fig. 3b). With a range from 8.5 to 12, the middle class has the highest percentage in terms of spatial extension (30%) and the greatest proportion of flood pixels (33%).

Flood susceptibility computation workflow can also take into account the convergence index, another morphometric variable. Its values had been classified according to the literature into five different groups (Fig. 2b). It is remarkable that the flood events are all occurring within surfaces that have negative convergence index values.

There are two types of areas on the plan curvature: convergent and divergent. Taking into account all the plan curvature values, 3 classes were identified as follows: 11.8–0, 0.1–0.5, 0.6–8.5 (Fig. 3c). Middle class cover the highest weight in the study area and it accounts for the greatest percentage of the total number of flood incidents (75%).

4. Methods

4.1. Statistical index (SI)

Statistical Index (SI) is a highly used bivariate approach for analyzing the correlation between elements in the environment. It has been used in studies concerned with the detection of zones exposed to different types of natural hazards (Bui et al., 2011). The calculation of SI coefficients can be performed using the next equation (Regmi et al., 2014):

$$SI_{ij} = \ln \left(\frac{f_{ij}}{f} \right) = \ln \left(\frac{\frac{Npix(S_i)}{Npix(N_i)}}{\frac{\sum Npix(S_i)}{\sum Npix(N_i)}} \right)$$

1

where: SI_{ij} – is the coefficient provided for a specific category/class i of parameter j ; f_{ij} – represents the density of flood pixels in the the class i of parameter j , f – represents the density of flood pixels in the entire study area, $Npix(S_i)$ represents the number of flood pixels in class i , while $Npix(N_i)$ represents the total number of pixels in the same parameter class.

4.2. Deep Learning Neural Network (DLNN)

Deep learning is a rapidly growing area of research in machine learning based on artificial neural networks, and is able to detect and classify patterns of data using multidimensional representations (Ngo et al., 2021). In this context, 'deep' refers to the transformation of one level of data representations into a higher level of representation (Ortega Adarme et al., 2020). In a data transformation, the maximum number of layers (depth of network) can be extracted in order to give the best hierarchical representation of the data. It should also be noted that in addition to the many hidden layers, two neurons in the output layer will represent flood and non-flood points, which will capture the effects of flooding. As the assessment of flood susceptibility is really a binary classification process, flooding locations would be assigned the value of 1, while non-flooding locations would be assigned zero (Costache et al., 2021). The sigmoid function $E(Y = i/x)$ is implied in the training process of DLNN. An input neuron associated with classification i will be able to provide a rough approximation of the sigmoid function with the information that resides on it. Moreover, we will be using a SoftMax function whose form is shown below (Costache et al., 2021):

$$softmax(a_i) = \frac{\exp(a_i)}{\sum_k \exp(a_i)}$$

2

where a_i represent the parameter associated to the *softmax* function.

A deep neural network is expressed with many hidden layers using an activation function in the next relation (Costache et al., 2021):

For $h = 1, \dots, H$ (hidden layers),

$$a^{(h)}(x) = b^{(h)} + W^{(h)}p^{(h-1)}(x)$$

3

$$p^{(h)}(x) = k(a^{(h)}(x))$$

4

where, k is a parameter which characterizes the activation function.

4.3. Support Vector Machine

SVMs (Support Vector Machines) are advanced models based on statistics-based learning theory, which has found applications for both classification and regression problems that are associated with large empirical data sets (Gao et al., 2017). Basically, the aim of the algorithm is to determine a hyper-plane that divides the training data in an optimal manner (Kavzoglu et al., 2014). The support vectors, which represent the closest data elements to the hyperplane, are considered to be the most essential elements in training data (Pham et al., 2016). It is thus possible to use the optimized hyperplane to classify the new data once the optimum surface of the hyperplane has been obtained. An equation that describes a linear hyperplane that should be used to classify new data can be written as follows (Tehrany et al., 2019):

$$f(x) = \text{sign} \left(\sum_{i=1}^n y_i \alpha_i x_i + b \right)$$

5

where α_i are Lagrange multipliers, x is the independent variable, y is the dependent variable, while b represents the hyper plane offset determined from the origin. It should be understood that there are many practical situations when a non-linear separable feature hyperplane would be useful, and as a consequence the original input data might have to be transferred into a high-dimension feature space. As a result, whenever a nonlinear kernel function is necessary, the Eq. (5) is rewritten using the form shown here:

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

6

where α_i and α_i^* are Lagrange multipliers ($\alpha_i \geq 0$, $\alpha_i^* \leq C$) and $K(x_i, x_j)$ is the kernel function.

In Fig. 4a is represented the linear hyper-plane, while in Fig. 4b is represented the non-linear hyperplane.

In the present research the Radial Basis Function (RBF) was particularly used in order to apply the SVM models. Also, the determination of parameters C and γ is crucial for the precision of SVM models. The

cross validation is the most common method used to estimate the C and γ optimally values.

4.4. Particle Swarm Optimization

Particle Swarm Optimization (PSO) algorithm was developed as a way to simulate a complex adaptive system (CAS), which represents an evolutionary computation technique (Li et al., 2019). Based on the regularity of activity of birds, the algorithm was initially constructed based on the swarm intelligence principle, and then a simplified version was built based on the swarm learning principle (Jain et al., 2018). Each optimization problem in PSO is solved by one of the particles, which is called a "bird" in the search space (Sachdeva et al., 2017). In a PSO, random particles are used to initialize the system and it is used to search for the ideal solution by iterative evolution (Javidi and Mansoury, 2017). The particles are able to update themselves automatically during each iteration by tracking the extreme values of velocity and position (Bui et al., 2017). In order to express the dynamics of the above-mentioned particles mathematically, let us look at the following formula (Jain et al., 2018):

$$\begin{cases} V_i^{n+1} = t \bullet V_i^n + c_1 \bullet r_1 \bullet (p_i^n - x_i^n) + c_2 \bullet r_2 \bullet (p_g^n - x_i^n) \\ x_i^{n+1} = x_i^n + V_i^n \end{cases}$$

7

where V_i^{n+1} is the velocity of particle i at the n iteration, x_i^{n+1} is the position of particle i at n iteration, $t \bullet V_i^n$ is the weight of inertia, c_1 is the factor of personal learning, c_2 is the factor of social learning, r_1 and r_2 are 2 random values within the range $[0,1]$, p_i^n is the best position of particle i at iteration n , p_g^n is the best position of the swarm at iteration i .

The optimal position of the population in PSO algorithm is reached according the scheme presented in Fig. 5.

The PSO will be applied involved in the training process of Support Vector Machine in order to optimize the C and γ parameter, a procedure intended to increase the precision of SVM model. Also, in terms of DLNN model the PSO algorithm will help in order to determine optimal number of hidden layers and hidden neurons which are associated with the lowest error. The Mealpy (Xie et al., 2022) Python library was used in order to apply the SVM, DLNN and also the PSO for both algorithms.

4.5. Results validation

4.5.1. ROC Curve

For studies that seek to investigate flooding susceptibility it is ever more common to use the receiver operating characteristic (ROC) curve as a means of validating the findings (Bradley, 1997). ROC graphs are designed to represent relationships between sensitivity on the X axis and 1-specificity on the Y axis. Using this model, we show that a model is capable or not to predict the flood exposed areas with considerable accuracy, based on the results of our research. As the most significant statistic indicator in

the field of statistics, AUC represents the area under the ROC curve that indicates the critical information regarding the performance of the model. The Area Under ROC Curve can be calculated using the following relation (Costache, 2019a):

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

8

where P is equal to the flood points number, N is equal to the non-flood points number, TP (true positive), TN (true negative) the points belonging to floods and non-floods that are the correctly classified.

4.5.2. Statistical metrics

According to this study, statistical indices are considered significant if they can provide empirical evidence for a spatial correlation between the location of observed floods and non-floods and the areas that are estimated to be highly susceptible to flooding (Costache and Bui, 2019). The next formulas are used to compute the statistical metrics represented by Specificity, Kappa Index, Sensitivity and Accuracy:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

9

$$\text{Specificity} = \frac{TN}{FP + TN}$$

10

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

11

$$k = \frac{p_o - p_e}{1 - p_e}$$

12

FP (false positive) – flood pixels that are classified in an erroneously manner, FN (false negative) – non-flood pixels that are classified in an erroneously manner, k – kappa coefficient, p_o – observed flood pixels, and p_e – locations that are susceptible to flood.

The six main steps applied for this research are represented in Fig. 6

5. Results

5.1. Multicollinearity assessment and feature selection

In terms of multicollinearity assessment, the Tolerance and Variance Inflation (VIF) indices were used. The lowest value of Tolerance was of 0.301 and was assigned to slope predictor, while the highest one was equal to 0.878 and was attributed to Plan curvature. Instead, the VIF maximum value was assigned to slope (3.035), meanwhile the minimum VIF belongs to Plan curvature (1.14). According to the results of modelling, there is no serious multicollinearity detected among the flood predictors.

Table 1
Multicollinearity assessment

Flood predictors	Collinearity Statistics	
	Tolerance	VIF
Elevation	0.503	1.987
Convergence index	0.756	1.323
Land use	0.353	2.832
Distance from river	0.600	1.666
Lithology	0.402	2.488
Plan curvature	0.878	1.140
Slope	0.301	3.035
Hydrological Soil Group	0.703	1.423
Topographic Wetness Index	0.682	1.465
Rainfall	0.571	1.753

The next values were achieved using the Gain Ratio (GR) method for feature selection: Slope (0.966), Lithology (0.402), Distance from river (0.379), Land use (0.361), Topographic Wetness Index (0.215), Plan curvature (0.206), Elevation (0.197), Hydrological Soil Group (0.135), Convergence index (0.12), Rainfall (0.083). Since all the GR values are higher than 0, all the predictors will be taken into account for the susceptibility modelling.

5.2. Results of Statistical Index (SI)

It should be remarked that a number of 19 flood predictors classes or categories do not contain any flood locations. Since the lowest SI coefficient of a class/category that contains at least 1 flood events was -2.71, it was took the decision that, for those classes/categories which not contain nor flood events, to assign, in a conventional way, the value of -3 (Table 1). The maximum SI value was 0.69, and it was

achieved by a number of 8 flood predictor classes/categories. Further the computed SI coefficients were used as input data in the machine learning models.

Table 2
Statistical Index values for each class/category of flood conditioning factors

<i>Factor</i>	<i>Class</i>	<i>Class pixels (%)</i>	<i>SI</i>
Lithology	1	27.28	-1.35
	2	3.08	-1.02
	3	0.65	-3
	4	1.43	-3
	5	0.10	-3
	6	0.05	-3
	7	0.89	-0.41
	8	28.62	-2.23
	9	14.30	-0.45
	10	23.45	0.55
	11	0.07	-3
	12	0.08	-3
Slope	< 3°	11.46	0.69
	3–7°	13.32	0.69
	7–15°	41.72	-3
	15–25°	26.42	-3
	> 25°	7.09	-3
Plan curvature	-11.8–0	32.68	-0.14
	0.1–0.5	56.65	0.16
	0.6–8.5	10.66	-3
HGS	A	1.52	0.69
	B	44.41	-0.95
	C	37.82	0.18
	D	16.25	-0.20
TWI	-9.7–4.5	21.30	-0.20
	4.6–8.4	24.98	-0.29
	8.5–12	29.70	-0.09

<i>Factor</i>	<i>Class</i>	<i>Class pixels (%)</i>	<i>SI</i>
	12.1–15	21.09	0.19
	15.1–25	2.93	0.69
Land use	Built-up areas	8.61	0.66
	Agriculture zone	11.55	0.31
	Vineyards	0.95	-3
	Fruit trees	9.17	0.47
	Pastures	6.11	-0.81
	Forest	52.23	0.00
	Natural grassland	6.66	-0.23
	Moors and heatland	0.81	-3
	Woodland-shrub	3.13	-3
	Water bodies	0.78	0.69
CI	-96 - -3	21.72	0.67
	-3 - -2	6.34	0.41
	-2 - -1	8.51	-1.52
	-1–0	10.62	-3
	0–95	52.81	-3
Elevation (m)	128–200	2.12	0.69
	200–400	23.38	0.29
	400–600	20.96	0.10
	600–800	10.42	-0.75
	800–1000	11.96	-3
	1000–1200	12.72	-3
	1200–1400	10.33	-3
	> 1400	8.12	-3
Distance from river (m)	0–50	4.58	0.69
	50–100	4.24	0.69
	100–150	4.20	0.61

<i>Factor</i>	<i>Class</i>	<i>Class pixels (%)</i>	<i>SI</i>
	150–200	3.10	0.60
	200–400	14.49	-0.95
	400–700	19.64	-1.33
	700–1000	16.53	-1.71
	> 1000	33.21	-2.71
Rainfall (mm/year)	531–600	9.94	0.38
	600–700	19.82	0.27
	700–800	16.85	0.04
	800–900	7.34	0.18
	900–1000	5.46	-0.10
	1000–1100	14.61	-0.88
	1100–1250	25.99	-0.23

5.3. Computation of FFSI maps

5.3.1. DLNN-SI and PSO-DLNN-SI

The models accuracy and the degree in which the target data managed the output, following the training procedure of both DLNN-SI and PSO-DLNN-SI hybrid models, are highlighted in Fig. 8. It should be remarked that the RMSE value in the case of DLLN-SI for training sample (0.223) is higher than RMSE value in the case of PSO-DLNN-SI also for training sample (0.137). Further, it should be noted that in terms of testing sample the RMSE of DLNN-SI (0.239) is also higher than the value of the same parameter in the case of PSO-DLNN-SI (0.2127). Therefore, it can be state that the optimization of DLNN-SI with the help of PSO algorithm managed to reduce the error of the model.

Also, it should be noted that the DLNN-SI model achieved the lowest error using an architecture containing 3 hidden layers and 100 hidden neurons in each hidden layer (Fig. 9a), while PSO-DLNN-SI achieved the lowest accuracy with an optimal architecture with 3 hidden layers and 83 hidden neurons in each hidden layer (Fig. 9b).

The training process allow the user to determine the importance of flood predictors. Thus, in the case of DLNN-SI ensemble the highest importance was assigned to Slope predictor (0.315), followed by Lithology (0.131), Distance from river (0.124), Land use (0.118), Topographic Wetness Index (0.070), Plan curvature (0.067), Elevation (0.064), Hydrological Soil Group (0.044), Convergence index (0.039) and Rainfall (0.027). The PSO-DLNN-SI ensemble revealed also that Slope predictor achieved the highest importance (0.269) followed by Distance from river (0.129), Land use (0.123), Topographic Wetness Index (0.113),

Lithology (0.110), Convergence index (0.082), Hydrological Soil Group (0.072), Elevation (0.067), Rainfall (0.028) and Plan curvature (0.006) (Fig. 10).

Using the importance of flood predictors, the Flood Susceptibility Index (FSI) was computed for each of the 4 ensemble models. Thus, the FSI DLNN-SI has values between -0.9 and 0.59 which were classified into 5 intervals with the help of Natural Breaks classification method. The very low class of values are spread on 24.7% of the middle and upper Prahova River basin, while the low susceptibility covers 27.36% of the study zone (Fig. 11a). The surfaces with moderate flood susceptibility can be found on around 23.39% of the territory, while the high and very high flood susceptibility manage to account approximately 24.55% of the study area. In terms of FSI PSO-DLNN-SI its values range from -0.97 to 0.62 (Fig. 11b) and were also split into 5 classes with the same Natural Breaks method. The very low susceptibility has a total surface equal to 21.45% of the study zone, while the low susceptibility is present on 30.12% of the middle and upper zone of Prahova River basin. Moderate flood susceptibility has a percentage of 24.67%, while, together, high and very high flood susceptibility span on 23.76% of the entire study zone.

5.3.2. SVM-SI and PSO-SVM-SI

The matching between targets and outputs for both training and testing samples, after the application of SVM-SI and PSO-SVM-SI hybrid models, is shown in Fig. 12. Moreover, there is represented also the RMSE that is lower in the case of optimized PSO-SVM-SI hybrid model (0.147) comparing to SVM-SI (0.252) for training sample. Also, the testing sample revealed the same situation in which the RMSE for PSO-SVM-SI (0.205) is lower than the RMSE associated to SVM-SI (0.244). Again, the optimization procedure managed to decrease the model's error.

Like in the case of DLNN-based ensembles, also for SVM-based ensemble models were derived the importance for each flood predictor. Thus, in terms of SVM-SI model, the highest importance was obtained by Slope (0.287), followed by Lithology (0.120), Distance from river (0.113), Land use (0.107), Elevation (0.088), Hydrological Soil Group (0.070), Convergence index (0.065), Topographic Wetness Index (0.064), Plan curvature (0.061) and Rainfall (0.025). For the PSO-SVM-SI model, the following hierarchy was revealed: Slope (0.283), Land use (0.151), Elevation (0.103), Hydrological Soil Group (0.1), Lithology (0.099), Distance from river (0.091), Convergence index (0.064), Topographic Wetness Index (0.044), Plan curvature (0.039) and Rainfall (0.027).

The calculated importance of each flood predictor was used to estimate the flood susceptibility across the study area. Thus, for FSI SVM-SI the values are situated between -0.92 and 0.59 and, like in the previous 2 cases, were grouped into 5 classes using the Natural Breaks method (Fig. 11c). The first class, which is assigned to very low values, has a total percentage of 23.54% of the study zone, while the low susceptibility is found on 28.2% of the entire middle and upper zone of Prahova River basin. The moderate values quantify approximately 23.39% of the study area, while the high and very high flood susceptibility is spread on 24.87% of the research zone. The FSI PSO-SVM-SI values range from -0.83 to 0.59 . According to the Natural Breaks classification method, the very low class values covers around

24.04% of the research perimeter, while the low flood susceptibility accounts for 28.49% of the same surface. Around 22.1% of the research area is attributed to moderate flood susceptibility, while the high and very high values appear on a percentage of 25.37%.

5.4. Results validation

5.4.1. ROC Curve

The ROC Curve represents the first method used to validate the flood susceptibility results. Thus, the Success Rate built with training data sample, shows that the best model was PSO-DLNN-SI having an AUC of 0.952. In terms of AUC, this model is followed by PSO-SVM-SI (AUC = 0.944), DLNN-SI (AUC = 0.93) and SVM-SI (AUC = 0.928) (Fig. 13a). The Prediction Rate, constructed using the validation sample, highlights also that the best model was PSO-SVM-SI (AUC = 0.926), followed by PSO-SVM-SI (AUC = 0.921), DLNN-SI (AUC = 0.904) and SVM-SI (AUC = 0.858).

5.4.2. Statistical metrics

According to the Table 3, the statistical metrics calculated for each flood susceptibility model show very good performances of this algorithms. Thus, using the training sample the accuracy values is characterized by the following hierarchy: PSO-DLNN-SI (0.959), PSO-SVM-SI (0.955), DLNN-SI (0.931) and SVM-SI (0.919). The same sample highlights the next ranking of K-index values: PSO-DLNN-SI (0.919), PSO-SVM-SI (0.911), DLNN-SI (0.861) and SVM-SI (0.837). In terms of validating sample, the most accurate model was also PSO-DLNN-SI (0.954), followed by PSO-SVM-SI (0.935), DLNN-SI (0.907) and SVM-SI (0.889). The K-index, with the validating sample, has the next values: PSO-DLNN-SI (0.907), followed by PSO-SVM-SI (0.870), DLNN-SI (0.833) and SVM-SI (0.778).

Table 3
Flood potential maps accuracy assessment using statistical metrics

Metrics	Training sample				Validating sample			
	DLNN-SI	PSO-DLNN-SI	PSO-SVM-SI	SVM-SI	DLNN-SI	PSO-DLNN-SI	PSO-SVM-SI	SVM-SI
TP	114	119	117	114	49	51	51	48
TN	115	117	118	112	50	52	50	48
FP	9	4	6	9	5	3	3	6
FN	8	6	5	11	4	2	4	6
Sensitivity	0.934	0.952	0.959	0.912	0.925	0.962	0.927	0.889
Specificity	0.927	0.967	0.952	0.926	0.909	0.945	0.943	0.889
Accuracy	0.931	0.959	0.955	0.919	0.917	0.954	0.935	0.889
K-index	0.861	0.919	0.911	0.837	0.833	0.907	0.870	0.778

6. Discussions

It will never be possible to prevent all natural disasters, including complex natural hazards like floods (Bui et al., 2019). For this reason it is exceptionally important so that flood prediction methods and flood mitigation tactics are improved in order to minimize the risk of loss of lives and the socio-economic effects of floods; this will enable us to minimize the long-term impacts of flooding (Dodangeh et al., 2020). An important part of flood modeling and risk assessment is the process of mapping flood susceptibility in for a specific region. Through flood susceptibility maps it can be identified where floods are increasingly likely to occur within a given area, and also can be determined the best way to prevent flood damages using non-structural and structural measures that can enhance the flood resilience (Hong et al., 2018). The presence research work is an attempt to apply and compare the results of four machine learning state-of-art models (DLNN-SI, PSO-DLNN-SI, SVM-SI and PSO-SVM-SI) in order to estimate the flood susceptibility degree in the middle and upper part of Prahova River basin from Romania.

It should be noted that frequently, flood susceptibility mapping focuses on identifying the most vulnerable areas by considering a variety of factors that affect flooding (Kanani-Sadat et al., 2019). The severity of floods is mainly determined by several geographical variable like hydrological parameters, geological features, meteorological conditions, morphological and topographical characteristics of a specific study region (Chowdhuri et al., 2020). Additionally, only specific flood predictors have an impact on flood susceptibility models; and thus, the choice of the right factors can affect the flood risk estimation. In the current study, a practical method of selecting features (Gain Ratio) for flood susceptibility estimation was was applied and allowed the selection of an adequate set of input flood predictors through which the flood susceptibility models were trained. According to the results of the Gain

Ratio method, slope angle has the highest predictive ability in terms of flooding phenomena, while hydrological soil group, convergence index and rainfall were the least important predictors. These results are in agreement with many studies that were carried out before (Costache et al., 2020; Dano et al., 2019; Khosravi et al., 2019). It is true that rainfall is the primary cause of flooding, but the effect of rainfall on flooding is not always linear (Khosravi et al., 2019). In terms of elevation, there is a general agreement that the higher altitudes are less prone to flooding than the lower ones (Sahana and Patel, 2019).

The construction of ROC curves as well as the computation of a several statistical metrics facilitated the validation of the model's performance. According to the results of the validation phase, both ensembles DLNN-SI and SVM-SI were significantly improved by using Particle Swarm Optimization. Overall, the PSO-DLNN-SI outperformed the other models by achieving an AUC-ROC Curve of 0.952 comparing with PSO-SVM-SI whose AUC was 0.944. It should be noted that there is almost a general agreement among the results of the four models according to which a quarter of study area have a high and very high flood susceptibility. Also, it is evident that these high and very high potential affected zones are located in the extremely southern part of the study zone, but also along the main rivers and depressions of the study area.

Researchers or maybe the persons in charge with flood hazard management may find the outputs of this research study very useful in selecting the most appropriate method for the simulation of flood susceptibility. This type of studies has also a main limitation which refers to the impossibility to calculate the flow velocity or depth.

7. Conclusions

In the present research paper, a bivariate statistical method (SI), two machine learning models (DLNN and SVM) and the Particle Swarm Optimization method were combined in order to create four ensemble models able to estimate the flood susceptibility in Prahova River basin from Romania. The four models were trained using a number of 10 geographical predictors along with 158 flood locations identified based on governmental sources. A first selection of predictors was carried out with the help of Gain Ratio method, this algorithm showing that all 10 predictors should be considered for the analysis. Also, through VIF and Tolerance indices was determined that among the flood predictors there is no serious multicollinearity. The use of ROC Curve and statistical metrics highlights that all the four models (DLNN-SI, PSO-DLNN-SI, SVM-SI, PSO-SVM-SI) achieved very good performances confirmed by AUC ROC Curve higher than 0.85. These very good results recommend the use of the four models also in other study areas that faces the danger of flood phenomena.

Declarations

Author Contribution R. Costache: Conceptualization, Methodology, Software, Validation, Formal analysis, Funding acquisition, Investigation, Writing—Original Draft. A. Arabameri: Methodology, Software, Validation, Formal analysis. A. R. M. T. Islam: Validation, Formal analysis, S. I. Abba: Investigation,

Writing—Original Draft. M. Pandey: Methodology, Software, Validation. R. S. Ajin: Investigation, Writing—Original Draft. B. T. Pham: Methodology, Software, Validation, Formal analysis.

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.

Data Availability

All data can be provided, at request, by the corresponding author.

Consent to Publish

All the authors have approved the submission and consented for publication.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

1. Afriyane, D., Julian, M.M., Riqqi, A., Akbar, R., Suroso, D.S., Kustiwan, I., 2020. Re-framing urban green spaces planning for flood protection through socio-ecological resilience in Bandung City, Indonesia. *Cities* 101, 102710.
2. Ahmadlou, M., Al-Fugara, A., Al-Shabeeb, A.R., Arora, A., Al-Adamat, R., Pham, Q.B., Al-Ansari, N., Linh, N.T.T., Sajedi, H., 2021. Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks. *Journal of Flood Risk Management* 14, e12683.
3. Albon, C., 2018. *Machine learning with python cookbook: Practical solutions from preprocessing to deep learning*. O'Reilly Media, Inc.
4. Alizadeh, M., Ngah, I., Hashim, M., Pradhan, B., Pour, A., 2018. A hybrid analytic network process and artificial neural network (ANP-ANN) model for urban earthquake vulnerability assessment. *Remote Sensing* 10, 975.
5. Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z., 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering* 174, 776–789.
6. Arora, A., Arabameri, A., Pandey, M., Siddiqui, M.A., Shukla, U., Bui, D.T., Mishra, V.N., Bhardwaj, A., 2021. Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga Plain, India. *Science of the Total Environment* 750, 141565.

7. Azareh, A., Rafiei Sardooi, E., Choubin, B., Barkhori, S., Shahdadi, A., Adamowski, J., Shamshirband, S., 2019. Incorporating multi-criteria decision-making and fuzzy-value functions for flood susceptibility assessment. *Geocarto International* 1–21.
8. Baghban, A., Bahadori, M., Lemraski, A.S., Bahadori, A., 2018. Prediction of solubility of ammonia in liquid electrolytes using least square support vector machines. *Ain Shams Engineering Journal* 9, 1303–1312.
9. Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 1145–1159.
10. Bui, D.T., Bui, Q.-T., Nguyen, Q.-P., Pradhan, B., Nampak, H., Trinh, P.T., 2017. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agricultural and forest meteorology* 233, 32–44.
11. Bui, D.T., Lofman, O., Revhaug, I., Dick, O., 2011. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural hazards* 59, 1413.
12. Bui, D.T., Ngo, P.-T.T., Pham, T.D., Jaafari, A., Minh, N.Q., Hoa, P.V., Samui, P., 2019. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping. *Catena* 179, 184–196.
13. Bui, Q.-T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., Pham, V.-M., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. *Journal of Hydrology* 581, 124379.
<https://doi.org/10.1016/j.jhydrol.2019.124379>
14. Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T., Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. *Environmental modelling & software* 95, 229–245.
15. Chowdhuri, I., Pal, S.C., Chakraborty, R., 2020. Flood susceptibility mapping by ensemble evidential belief function and binomial logistic regression model on river basin of eastern India. *Advances in Space Research* 65, 1466–1489.
16. Costache, R., 2019a. Flood Susceptibility Assessment by Using Bivariate Statistics and Machine Learning Models-A Useful Tool for Flood Risk Management. *Water Resources Management* 33, 3239–3256.
17. Costache, R., 2019b. Flash-Flood Potential assessment in the upper and middle sector of Prahova river catchment (Romania). A comparative approach between four hybrid models. *Science of The Total Environment* 659, 1115–1134.
18. Costache, R., Arabameri, A., Moayedi, H., Pham, Q.B., Santosh, M., Nguyen, H., Pandey, M., Pham, B.T., 2021. Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree. *Geocarto International* 1–28.
19. Costache, R., Bui, D.T., 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. *Science of*

The Total Environment 691, 1098–1118.

20. Costache, R., Țîncu, R., Elkhachy, I., Pham, Q.B., Popa, M.C., Diaconu, D.C., Avand, M., Costache, I., Arabameri, A., Bui, D.T., 2020. New neural fuzzy-based machine learning ensemble for enhancing the prediction accuracy of flood susceptibility mapping. *Hydrological Sciences Journal* 65, 2816–2837.
21. Dahri, N., Abida, H., 2017. Monte Carlo simulation-aided analytical hierarchy process (AHP) for flood susceptibility mapping in Gabes Basin (southeastern Tunisia). *Environmental Earth Sciences* 76, 302.
22. Dano, U.L., Balogun, A.-L., Matori, A.-N., Wan Yusouf, K., Abubakar, I.R., Said Mohamed, M.A., Aina, Y.A., Pradhan, B., 2019. Flood susceptibility mapping using GIS-based analytic network process: A case study of Perlis, Malaysia. *Water* 11, 615.
23. Dodangeh, E., Choubin, B., Eigdir, A.N., Nabipour, N., Panahi, M., Shamshirband, S., Mosavi, A., 2020. Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. *Science of the Total Environment* 705, 135983.
24. Elmahdy, S., Ali, T., Mohamed, M., 2020. Flash flood susceptibility modeling and magnitude index using machine learning and geohydrological models: A modified hybrid approach. *Remote Sensing* 12, 2695.
25. Gao, L., Ye, M., Lu, X., Huang, D., 2017. Hybrid method based on information gain and support vector machine for gene selection in cancer classification. *Genomics, proteomics & bioinformatics* 15, 389–395.
26. Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.-X., Chen, W., 2018. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. *Science of the total environment* 625, 575–588.
27. Jacinto, R., Grosso, N., Reis, E., Dias, L., Santos, F., Garrett, P., 2015. Continental Portuguese Territory Flood Susceptibility Index: contribution to a vulnerability index. *Natural Hazards and Earth System Sciences* 15, 1907–1919.
28. Jain, I., Jain, V.K., Jain, R., 2018. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing* 62, 203–215.
29. Javidi, M.M., Mansoury, S., 2017. Diagnosis of the disease using an ant colony gene selection method based on information gain ratio using fuzzy rough sets. *Journal of Particle Science & Technology* 3, 175–186.
30. Kanani-Sadat, Y., Arabsheibani, R., Karimipour, F., Nasser, M., 2019. A new approach to flood susceptibility assessment in data-scarce and ungauged regions based on GIS-based hybrid multi criteria decision-making method. *Journal of Hydrology* 572, 17–31.
31. Kavzoglu, T., Sahin, E.K., Colkesen, I., 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* 11, 425–439.
32. Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Bui, D.T., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment* 627, 744–755.

33. Khosravi, K., Shahabi, H., Pham, B.T., Adamowski, J., Shirzadi, A., Pradhan, B., Dou, J., Ly, H.-B., Gróf, G., Ho, H.L., 2019. A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods. *Journal of Hydrology* 573, 311–323.
34. Li, D., Huang, F., Yan, L., Cao, Z., Chen, J., Ye, Z., 2019. Landslide susceptibility prediction using particle-swarm-optimized multilayer perceptron: Comparisons with multilayer-perceptron-only, bp neural network, and information value models. *Applied Sciences* 9, 3664.
35. Luino, F., Turconi, L., Paliaga, G., Faccini, F., Marincioni, F., 2018. Torrential floods in the upper Soana Valley (NW Italian Alps): Geomorphological processes and risk-reduction strategies. *International journal of disaster risk reduction* 27, 343–354.
36. Ngo, P.T.T., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., Lee, S., 2021. Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geoscience Frontiers* 12, 505–519.
37. Ortega Adarme, M., Queiroz Feitosa, R., Nigri Happ, P., Aparecido De Almeida, C., Rodrigues Gomes, A., 2020. Evaluation of Deep Learning Techniques for Deforestation Detection in the Brazilian Amazon and Cerrado Biomes From Remote Sensing Imagery. *Remote Sensing* 12, 910.
38. Peng-cheng, Q., Min, L., Lan, L., 2016. Application of effective precipitation index in rainstorm flood disaster monitoring and assessment. *Chinese Journal of Agrometeorology* 37, 84.
39. Pham, B.T., Bui, D.T., Dholakia, M.B., Prakash, I., Pham, H.V., 2016. A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area. *Geotechnical and Geological Engineering* 34, 1807–1824.
40. Regmi, A.D., Devkota, K.C., Yoshida, K., Pradhan, B., Pourghasemi, H.R., Kumamoto, T., Akgun, A., 2014. Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. *Arabian Journal of Geosciences* 7, 725–742.
41. Sachdeva, S., Bhatia, T., Verma, A., 2017. Flood susceptibility mapping using GIS-based support vector machine and particle swarm optimization: A case study in Uttarakhand (India). Presented at the 2017 8th International conference on computing, communication and networking technologies (ICCCNT), IEEE, pp. 1–7.
42. Sahana, M., Patel, P.P., 2019. A comparison of frequency ratio and fuzzy logic models for flood susceptibility assessment of the lower Kosi River Basin in India. *Environmental Earth Sciences* 78, 1–27.
43. Tehrany, M.S., Kumar, L., Shabani, F., 2019. A novel GIS-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine: Brisbane, Australia. *PeerJ* 7, e7653.
44. Tehrany, M.S., Shabani, F., Javier, D.N., Kumar, L., 2017. Soil erosion susceptibility mapping for current and 2100 climate conditions using evidential belief function and frequency ratio. *Geomatics, Natural Hazards and Risk* 8, 1695–1714.

45. Xie, C., Nguyen, H., Choi, Y., Armaghani, D.J., 2022. Optimized functional linked neural network for predicting diaphragm wall deflection induced by braced excavations in clays. *Geoscience Frontiers* 13, 101313.
46. Xiong, J., Li, J., Cheng, W., Wang, N., Guo, L., 2019. A GIS-based support vector machine model for flash flood vulnerability assessment and mapping in China. *ISPRS International Journal of Geo-Information* 8, 297.

Figures

Figure 1

Study area location

Figure 2

Flood predictors (a. Slope angle; b. Convergence index; c. Distance from river; d. Elevation; e. Land use; f. Lithology)

Figure 3

Flood predictors (a. Rainfall; b. TWI; c. Plan curvature; d. Hydrological Soil Group)

Figure 4

Optimally hyper-plane (a. linearly separable; b. non-linearly separable)

Figure 5

The PSO procedure used to achieve the optimal position of population

Figure 6

Flowchart of the methodological steps in the present research

Figure 7

Gain Ratio coefficients

Figure 8

Training performance of DLLN-SI model (a. target and output ANFIS value of train data samples; b. target and output ANFIS value of testing data samples; c. MSE and RMSE value of train data samples; d. frequency errors of train data samples; e. MSE and RMSE value of testing data samples; f. frequency errors of testing data samples) and PSO-DLNN-SI model (g. target and output ANFIS value of train data samples; h. target and output ANFIS value of testing data samples; i. MSE and RMSE value of train data samples j. frequency errors of train data samples; k. MSE and RMSE value of testing data samples; l. frequency errors of testing data samples)

Figure 9

DLNN architecture (a. DLNN-SI; b. PSO-DLNN-SI)

Figure 10

Flood predictors importance in flood susceptibility models

Figure 11

Flood Susceptibility Index (a. FSI DLNN-SI; b. FSI PSO-DLNN-SI; c. FSI SVM-SI; FSI PSO-DLNN-SI)

Figure 12

Training performance of SVM-SI model (a. target and output ANFIS value of train data samples; b. target and output ANFIS value of testing data samples; c. MSE and RMSE value of train data samples; d. frequency errors of train data samples; e. MSE and RMSE value of testing data samples; f. frequency errors of testing data samples) and PSO-SVM-SI model (g. target and output ANFIS value of train data samples; h. target and output ANFIS value of testing data samples; i. MSE and RMSE value of train data samples j. frequency errors of train data samples; k. MSE and RMSE value of testing data samples; l. frequency errors of testing data samples)

Figure 13

ROC Curve (a. Success Rate; b. Prediction Rate)