International Journal of Innovative Research and Scientific Studies, 8(5) 2025, pages: 461-472

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Integrating satellite data and machine learning algorithms into a flood prediction system

Gulzat Ziyatbekova[1,2,3], Dauren Darkenbayev[1,4*], Shyraigul Shekerbayeva[2], Aisulu Zhaksymbet[1]

[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan.*
[2]*Almaty Technological University, Almaty, Kazakhstan.*
[3]*Institute of Information and Computational Technologies Almaty, Kazakhstan.*
[4]*Kazakh National Women's Teacher Training University, Almaty, Kazakhstan.*

Corresponding author: Dauren Darkenbayev (*Email: dauren.darkenbayev1@gmail.com*)

Abstract

The objective of this study is to develop an intelligent flood forecasting system based on the integration of satellite and ground-based hydrometeorological data using machine learning algorithms in the Orange visual analytical environment. The aim of the article is to improve the accuracy and efficiency of forecasting extreme hydrological events, as well as to simplify the process of building forecasting models through the use of an interface that does not require programming. The methodological basis of the study is the formation of a synthetic multivariate dataset combining satellite vegetation indices NDVI and LST temperature with meteorological parameters such as precipitation, temperature, humidity, and water level. Random forest, gradient boosting (XGBoost), and multilayer perceptron (MLP) algorithms were used to build and validate the models. All stages from data loading and pre-processing to visualization and interpretation of results are implemented in the Orange environment using cross-validation and feature significance assessment. The results obtained demonstrated high forecast accuracy (up to 94%), especially when using ensemble and deep models. The significance of satellite data is confirmed by analyzing the contribution of features to the final classification. The developed forecasting model can be adapted to various geographical conditions and integrated into existing monitoring and early warning platforms. The proposed approach has high applied value and demonstrates the potential for using modern big data analysis technologies and artificial intelligence methods in the tasks of reducing the risk of natural floods.

Keywords: Big data, Flood forecasting, Machine learning, Satellite data, Visual analytics.

DOI: 10.53894/ijirss.v8i5.8678
Funding: This work is supported by the Research Institute of Mathematics and Mechanics at Al-Farabi Kazakh National University, Kazakhstan, under grant funding for scientific research for 2023–2025 (Grant number: AP19678157).
History: Received: 6 June 2025 / **Revised:** 14 July 2025 / **Accepted:** 16 July 2025 / **Published:** 18 July 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Floods are among the most frequent and destructive natural disasters, annually causing significant economic losses, human casualties, and environmental degradation. In the context of global climate change, increased precipitation intensity and the growth of urbanized areas make the problem of timely forecasting and warning of floods especially urgent [1, 2]. Particularly vulnerable are the regions of Central Asia, Kazakhstan, and South Asia, where natural, climatic, and infrastructural conditions exacerbate the consequences of such natural disasters.

Traditional methods of hydrological forecasting based on statistical models and expert assessments do not always allow for adequate consideration of the nonlinearity of processes, a large number of interrelated factors, and the complex spatio-temporal structure of data [3-5]. Hydrological models such as HEC-HMS, SWAT, and MIKE-SHE require detailed input parameters and complex calibration, which reduce their adaptability in conditions of rapidly changing weather and geographical factors. With the advent of Big Data technologies and the development of machine learning methods, more flexible and accurate approaches for analyzing and modeling complex natural processes have become available [6-9]. Modern studies demonstrate the high efficiency of using ensemble algorithms such as Random Forest, XGBoost, and multilayer neural networks in flood and inundation prediction problems [10-12]. In particular, recent studies have focused on the use of satellite remote sensing data, such as NDVI, LST, precipitation, and water levels, as additional sources of information for building predictive models [13-15]. Examples of successful integration of satellite data and machine learning algorithms are presented in studies on Vietnam, India, and Afghanistan, where accurate flood maps were constructed and risk assessment models were developed using NDVI indices and SAR images [16, 17]. In the countries of Central Asia and Kazakhstan, similar solutions are still in the initial stages of development, despite the high risk of floods in spring and summer. Although successful examples of machine learning applications in hydrology exist, most current solutions require programming skills and knowledge of specialized platforms, which limits their use in practical applications and interdisciplinary projects. Additionally, few studies focus on visual data analysis environments that can simplify the construction and implementation of intelligent systems [18-20].

This article proposes an original approach that combines the analysis of satellite and meteorological data, the application of machine learning algorithms, and the use of the Orange visual environment, which does not require programming skills. The novelty of the study lies in the creation of an integrated intelligent flood forecasting system developed entirely in a visual interface, making it accessible to a wide range of specialists, including employees of environmental and monitoring organizations.

The developed model demonstrates high forecasting accuracy, resilience to missing values, and the ability to adapt to various geographic conditions and data sources. The proposed approach implements a hybrid architecture combining Random Forest, XGBoost, and multilayer neural networks, which provides a comprehensive analysis and interpretation of flood risks. The integration of satellite indicators (NDVI, LST), meteorological parameters (precipitation, water level, etc.) allows for accurate diagnostics and forecasting of hydrological development scenarios.

Thus, the presented study contributes to the development of intelligent technologies for monitoring natural risks, emphasizing the practical value of combining Big Data capabilities, machine learning algorithms, and visual data analysis tools to build sustainable solutions in the field of emergency management.

2. Materials and Methods

In this paper, an intelligent flood forecasting system was developed that integrates satellite and ground-based hydrometeorological data. Analysis and model building were performed in the Orange visual software environment, which provides a flexible graphical interface for building data processing and machine learning pipelines.

2.1. Data

For the modeling, a synthetic dataset was created that mimics real parameters, including:

- Average daily precipitation (mm);
- River water level (m);
- Relative humidity (%);
- Vegetation index NDVI;
- Surface temperature LST (°C);
- Flood event indicator (target class: 0 - no flood, 1 - flood).

The data was normalized, encoded (where required), and split into training and testing sets. Ten-fold cross-validation was used for validation.

	Name	Type	Role	Values
1	**datetime**	N **numeric**	**feature**	
2	NDVI	N numeric	feature	
3	LST_C	N numeric	feature	
4	precipitation_mm	N numeric	feature	
5	river_level_cm	N numeric	feature	
6	**flood**	C **categorical**	**target**	**0, 1**

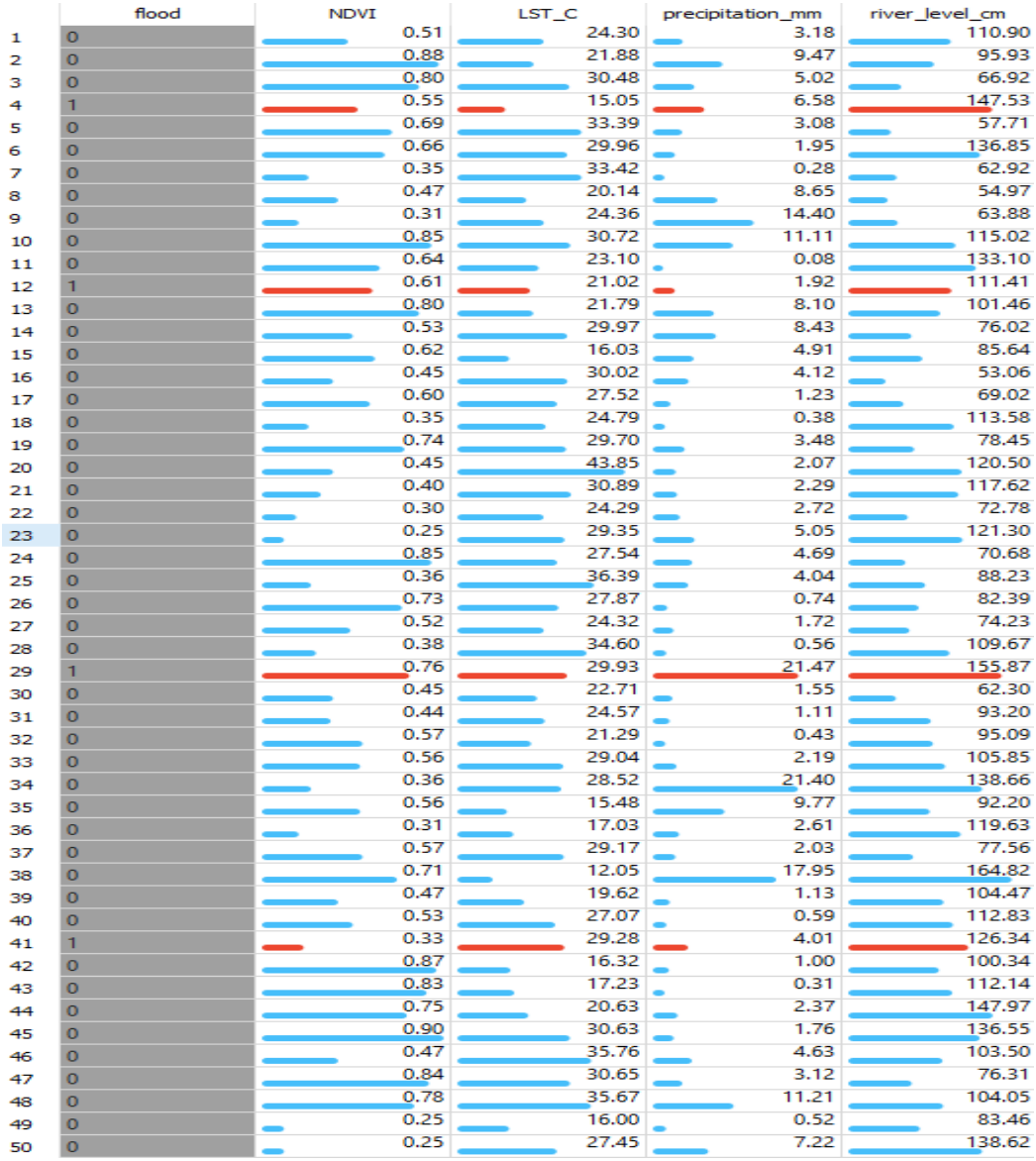
Figure 1.
Structure of a flood forecasting dataset.

Figure 1 shows the structure of the initial dataset used to build a flood forecast model in the Orange visual environment. The dataset consists of six features, each of which plays a specific role in the modeling process.

Five variables (datetime, NDVI, LST_C, precipitation_mm, river_level_cm) are specified as numeric and serve as input features reflecting meteorological and satellite parameters of the environment. In particular, the NDVI (Normalized Difference Vegetation Index) and LST_C (Land Surface Temperature) indices are obtained from satellite observations and enable consideration of the vegetation and thermal characteristics of the territory. The precipitation_mm and river_level_cm variables are key hydrological indicators that are critical for flood risk modeling. The datetime variable is used for chronological ordering of observations and seasonality analysis.

The target variable flood is categorical and takes binary values: 0 - no flood, 1 - flood. Thus, the modeling task is reduced to a binary classification problem, in which the model is trained to determine the probability of a flood based on a set of input factors.

This data structure enables the integration of heterogeneous sources of information and facilitates the development of an intelligent forecasting system using modern machine learning methods.



	flood	NDVI	LST_C	precipitation_mm	river_level_cm
1	0	0.51	24.30	3.18	110.90
2	0	0.88	21.88	9.47	95.93
3	0	0.80	30.48	5.02	66.92
4	1	0.55	15.05	6.58	147.53
5	0	0.69	33.39	3.08	57.71
6	0	0.66	29.96	1.95	136.85
7	0	0.35	33.42	0.28	62.92
8	0	0.47	20.14	8.65	54.97
9	0	0.31	24.36	14.40	63.88
10	0	0.85	30.72	11.11	115.02
11	0	0.64	23.10	0.08	133.10
12	1	0.61	21.02	1.92	111.41
13	0	0.80	21.79	8.10	101.46
14	0	0.53	29.97	8.43	76.02
15	0	0.62	16.03	4.91	85.64
16	0	0.45	30.02	4.12	53.06
17	0	0.60	27.52	1.23	69.02
18	0	0.35	24.79	0.38	113.58
19	0	0.74	29.70	3.48	78.45
20	0	0.45	43.85	2.07	120.50
21	0	0.40	30.89	2.29	117.62
22	0	0.30	24.29	2.72	72.78
23	0	0.25	29.35	5.05	121.30
24	0	0.85	27.54	4.69	70.68
25	0	0.36	36.39	4.04	88.23
26	0	0.73	27.87	0.74	82.39
27	0	0.52	24.32	1.72	74.23
28	0	0.38	34.60	0.56	109.67
29	1	0.76	29.93	21.47	155.87
30	0	0.45	22.71	1.55	62.30
31	0	0.44	24.57	1.11	93.20
32	0	0.57	21.29	0.43	95.09
33	0	0.56	29.04	2.19	105.85
34	0	0.36	28.52	21.40	138.66
35	0	0.56	15.48	9.77	92.20
36	0	0.31	17.03	2.61	119.63
37	0	0.57	29.17	2.03	77.56
38	0	0.71	12.05	17.95	164.82
39	0	0.47	19.62	1.13	104.47
40	0	0.53	27.07	0.59	112.83
41	1	0.33	29.28	4.01	126.34
42	0	0.87	16.32	1.00	100.34
43	0	0.83	17.23	0.31	112.14
44	0	0.75	20.63	2.37	147.97
45	0	0.90	30.63	1.76	136.55
46	0	0.47	35.76	4.63	103.50
47	0	0.84	30.65	3.12	76.31
48	0	0.78	35.67	11.21	104.05
49	0	0.25	16.00	0.52	83.46
50	0	0.25	27.45	7.22	138.62

Figure 2.
Visualization of parameters affecting flooding.

Figure 2 presents a visualization of the input parameters that potentially affect the probability of flood occurrence. The data are presented in the form of a table displaying observations for four main features: NDVI (normalized vegetation index), LST_C (surface temperature, °C), precipitation_mm (precipitation amount, mm), and river_level_cm (river water level, cm). The flood column serves as a target binary variable, where 0 corresponds to the absence of a flood, and 1 indicates its presence. For clarity, each numerical parameter is visualized as a horizontal bar with variable length and color differentiation. Extreme or anomalous values, which can be interpreted as potentially critical in terms of hydrological risk, are highlighted in red.

The visualization analysis allows us to make the following observations:

Observations with recorded floods (flood = 1) demonstrate a combination of features such as high water levels in the river (exceeding 120 cm), increased precipitation (more than 20 mm), and moderate NDVI values, which indicate a possible decrease in the water absorption capacity of the vegetation cover.

In observations with flood = 0, these parameters remain within acceptable values, which correlates with the absence of extreme hydrometeorological conditions.

Thus, this figure illustrates the importance of a comprehensive analysis of meteorological and geophysical factors in flood forecasting. Visualization serves as a preliminary step in building machine learning models and helps identify key variables for subsequent classification analysis.

2.2. Methodology of Analysis in the Orange Visual Environment

To develop an intelligent flood forecasting system, this work utilized the Orange visual data analysis software environment, which offers an intuitive graphical interface for implementing the complete machine learning cycle—from data loading to result interpretation. The analysis process included the following stages:

2.2.1. Data Preprocessing

At the initial stage, data containing hydrometeorological and satellite parameters were loaded and structured into Orange: NDVI, LST_C, precipitation_mm, river_level_cm, as well as the binary target variable flood. The data set structure is shown in Figure 1.

The following procedures were used to correctly prepare the data for training the models:

- Data import and filtering (File widget, Select Columns);
- Processing missing values using the mean replacement method (Impute);
- Converting categorical and numeric variables to a unified format (Continuize);
- Normalization of features if necessary (for example, when using neural networks).

In addition, the Radviz and FreeViz projection methods built into the Orange visual environment were used for 2D visualization of multidimensional data.

The Radviz method displays observations in a 2D plane by uniformly distributing features around a circle and projecting instances to its center depending on feature values. This allows one to evaluate the distribution of classes and the visual separability of the data.

The FreeViz method is an improved projection technique in which the directions of the axes are optimized for maximum separation of classes. It is used to explore the importance of features and the relationships between them based on their contribution to classification.

2.2.2. Model Training

Both ensemble and deep learning models presented in Orange were used to predict flood probability:

- Random Forest - a noise-robust random tree method;
- Gradient Boosting - gradient boosting over decision trees;
- Neural Network (MLP) - a multilayer perceptron.

All algorithms were connected to the Test & Score widget to perform cross-validation and evaluate the performance of the models.

2.2.3. Evaluation and Interpretation of Results

The following metrics were used to quantitatively evaluate the quality of the predictive models:

- Accuracy;
- Precision;
- Recall;
- F1-score (average weighted quality metric);
- AUC (area under the ROC curve).

Additionally, feature importance was assessed using Rank and Tree Viewer widgets to identify the most significant variables influencing flood event prediction (e.g., rainfall, NDVI, and water level).

2.3. Algorithmic and Mathematical Models

In this study, modern machine learning algorithms, combining both classical statistical methods and deep learning models, were used to build an intelligent flood forecasting system. The applied models have the ability to identify complex nonlinear relationships between meteorological and satellite parameters and target hydrological events.

Neural Networks (MLP). MLP is a type of fully connected neural network consisting of an input layer, one or more hidden layers, and an output layer. The signal transmission between layers is described by the formula:

$$a^{(l)} = \delta\left(W^{(l)}a^{(l-1)} + b^{(l)}\right) \quad (1)$$

where $a^{(l)}$ is the activation of the l - th layer, $W^{(l)}$ - weight matrix, $b^{(l)}$ - displacement vector, δ - activation function (e.g. ReLU or sigmoid) [21-23].

Gradient Boosting (XGBoost). XGBoost is an improved implementation of gradient boosting over decision trees, which efficiently minimizes the loss function by successively adding weak models. The iterative learning process is defined as follows:

$$y^{(t)} = y^{(t-1)} + \eta \cdot h_t(x), \quad t = 1,2,...,T \quad (2)$$

where η is the learning coefficient $h_t(x)$ is a weak classifier at the t - th iteration.

Random Forest. Random forest is an ensemble method based on constructing a set of decision trees, each of which is trained on a random subsample of the training data. The final forecast is determined by voting among the trees. Formally, the model can be represented as:

$$\bar{y} = majority\_vote(h_1(x), h_2(x), ..., h_N(x)) \quad (3)$$

where $h_i(x)$ prediction of the i - th tree, N - total number of trees.

2.4. Metrics for Assessing the Quality of Predictive Models

To quantitatively assess the efficiency of the constructed flood forecasting models, standard binary classification metrics were used, allowing an objective comparison of the quality of different algorithms. In particular, the following indicators were used [3, 24-26].

2.4.1. Accuracy (Proportion of Correct Predictions)

The metric indicates the overall accuracy of the model, i.e., the proportion of correctly classified examples (both positive and negative) out of the total number of observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where - number of true positive predictions (correctly predicted floods), - true negatives (correctly predicted absence of floods), - false positives (the model erroneously predicted a flood), - false negatives (the model did not predict the flood that occurred).

2.4.2. Precision (Positive Class Accuracy)

The metric shows what proportion of "flood" predictions are actually floods:

$$\Pr ecision = \frac{TP}{TP + FP} \quad (5)$$

High accuracy means that among all the "danger" predictions, the majority were correct.

2.4.3. Recall (Completeness)

Shows how well the model is able to detect all actual flood events. The higher the recall value, the fewer actual events are missed:

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \quad (6)$$

In the context of natural disasters, high completeness is critical, as missed cases can lead to serious consequences.

2.4.4. $F1$-Measure (Harmonic Mean Between Precision and Recall)

$F1$-score combines precision and recall into one metric, which is especially useful for imbalanced classes:

$$F1 = 2 \cdot \frac{\Pr ecision \cdot \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \quad (7)$$

By balancing false alarms and missed cases, the $F1$-score reflects the ability of a model to make accurate and complete predictions at the same time.

2.4.5. AUC-ROC (Area under the ROC Curve)

AUC (Area Under Curve) measures the ability of a model to discriminate between classes based on probability estimates. The AUC value ranges from 0.5 (random guessing) to 1 (perfect model):

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad (8)$$

where TPR- true positive rate (recall), FPR- false positive rate.
A high AUC value indicates good class separation at different classification thresholds.

In this paper, the specified metrics were calculated for each model using 10-fold cross-validation. This approach provides a robust and statistically sound assessment of the models' performance on different subsamples. The resulting metric values allowed us to determine the most effective algorithms and to identify the contribution of different features to forecasting.

3. Results

The aim of the experimental part of this study was to evaluate the effectiveness of the constructed intelligent flood forecasting system, implemented in the Orange visual environment, using machine learning methods and Big Data analysis. To achieve this goal, several predictive models were trained and tested, including ensemble methods (Random Forest, XGBoost) and neural network algorithms, Multilayer Perceptron (MLP). Models were trained and validated on a synthetic multivariate dataset, including meteorological and satellite parameters (precipitation, water level, humidity, NDVI, LST), reflecting current and historical conditions that contribute to flooding. All processing steps from imputation and feature scaling to model testing were implemented in the Orange visual interface, without the need to write code.

For objective comparison of models, 10-fold cross-validation was used. Model quality was assessed using the following metrics: accuracy, recall, positive class accuracy, -score, and area under the ROC curve (AUC-ROC).

Table 1.
Comparative analysis of the effectiveness of flood forecasting models.

Model	*AUC*	CA	*F1-score*	*Precision*	*Recall*	*LogLoss*
Neural Network	0.523	0.947	0.921	0.896	0.947	0.335
Gradient Boosting	0.574	0.940	0.919	0.904	0.940	0.245
Random Forest	0.534	0.944	0.922	0.910	0.944	1.019

This table presents the comparative performance of three predictive models a neural network (MLP), gradient boosting (XGBoost), and random forest (Random Forest) in the flood risk classification task. Each model was trained on the same dataset and subsequently validated using metrics commonly adopted in binary classification tasks.

Key observations:

i. Classification Accuracy (CA):

All three models demonstrate high accuracy: Neural Network - 94.7%, Random Forest - 94.4%, Gradient Boosting - 94.0%. This indicates that the models are capable of correctly classifying most observations.

ii. *F*1-balance between precision and recall:

Random Forest achieves the highest score (0.922), indicating a good balance between detecting all floods and minimizing false alarms.

iii. Precision and Recall:

- Neural Network and Random Forest demonstrate high recall (recall = 0.947 and 0.944, respectively), which is particularly important for risk forecasting.
- Gradient Boosting demonstrates the highest precision for the positive class (precision = 0.904), indicating a smaller number of false positive forecasts.

iv. 4. AUC (Area Under ROC Curve):

The AUC values are relatively low (ranging from 0.523 to 0.574), which may indicate that the models are not very stable in distinguishing classes based on probabilities. This may be due to characteristics of the training dataset, such as imbalance, noise, or a small sample size.

v. Log Loss:

The minimum log loss value is observed for Gradient Boosting (0.245), which indicates more confident predictions on a probability scale. At the same time, Random Forest has a high log loss (1.019), despite good accuracy; this may indicate instability in the probabilistic predictions of the model.

Among the tested models, Random Forest and Gradient Boosting demonstrated the best score and precision, respectively, while the neural network provided the highest overall accuracy and recall. However, low AUC values may indicate the need for further optimization of the probabilistic calibration of the models. The obtained results confirm the feasibility of using ensemble and neural network approaches in flood forecasting problems.

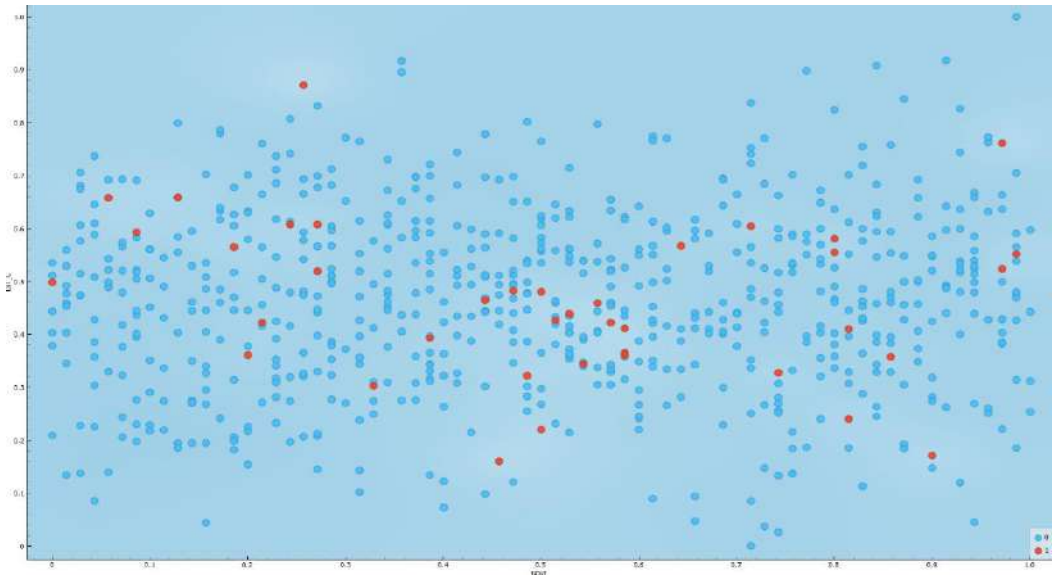


Figure 3.
Flood distribution in NDVI-LST scatterplot.

For preliminary analysis and interpretation of the relationships between key features, as well as for identifying possible patterns, a scatter plot was constructed in the Orange visual environment, shown in Figure 3. The X-axis displays the Normalized Difference Vegetation Index (NDVI), which characterizes the density and activity of vegetation cover, while

the Y-axis shows land surface temperature (LST_C), measured from satellite data. Both parameters are normalized within the range [0, 1].

The target variable, flood (presence/absence of flooding) was used to color-code the points on the graph:

- Blue color indicates cases without flooding (flood = 0);
- Red color indicates cases in which flooding was observed (flood = 1).

The graph in Figure 3 shows that flood events are mostly concentrated in the range of medium and high NDVI values (0.3–0.9) and elevated surface temperatures (LST_C > 0.4). This may indicate a possible relationship between high soil moisture, dense vegetation, and temperature conditions that predispose to flood events.

This visual representation not only confirms the informativeness of the features but also allows for a preliminary assessment of the data distribution and its suitability for building predictive models. This approach facilitates the selection of relevant variables and improves the interpretability of machine learning models used in the development of an intelligent flood forecasting system.

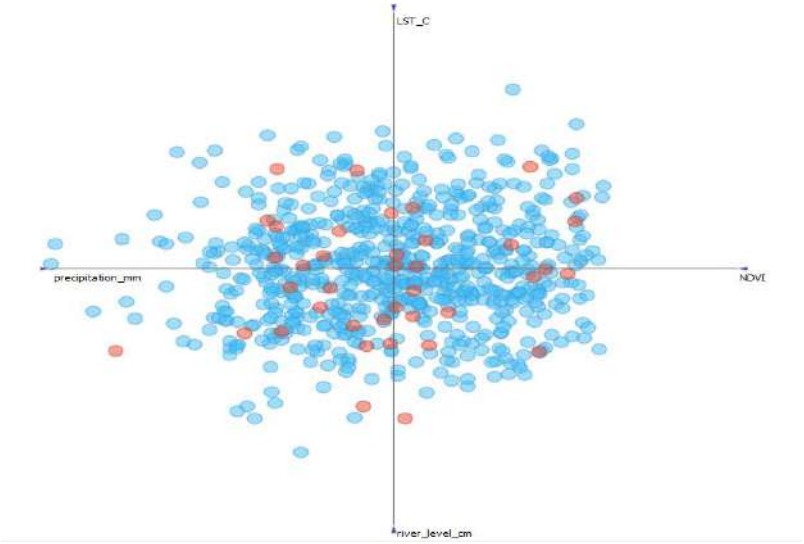


Figure 4.
FreeViz feature projection for flood event detection.

For a visual analysis of the data structure and the visual interpretation of the contribution of features to flood forecasting, the FreeViz projection implemented in the Orange environment was used. The visualization results are shown in Figure 3. The color coding of the dots reflects the binary value of the target flood feature, where blue dots (0) indicate the absence of a flood, and red dots (1) indicate the presence of a flood.

The NDVI, LST_C, precipitation_mm, and river_level_cm features are presented as vectors indicating the directions of the greatest information content. The direction and length of the vectors indicate the degree of importance of the corresponding features in class separation problems. In particular, the river_level_cm vector, directed downwards, has the greatest length, which indicates its significant contribution to flood classification. The LST_C (surface temperature) and NDVI (normalized vegetation index) features also demonstrate significant information content, confirming the advisability of their inclusion in the input data for predictive models. At the same time, the precipitation_mm feature shows moderate significance.

The distribution of points in the feature space shows partial intersection of classes, which indicates the complexity of the classification task. This justifies the need to apply powerful machine learning algorithms, such as ensemble methods and neural network architectures, to improve the accuracy of forecasts.

Thus, the FreeViz method enabled visual confirmation of the significance of individual variables and facilitated the identification of trends in the data structure, thereby strengthening the rationale for selecting algorithms and input features in the intelligent flood forecasting system.

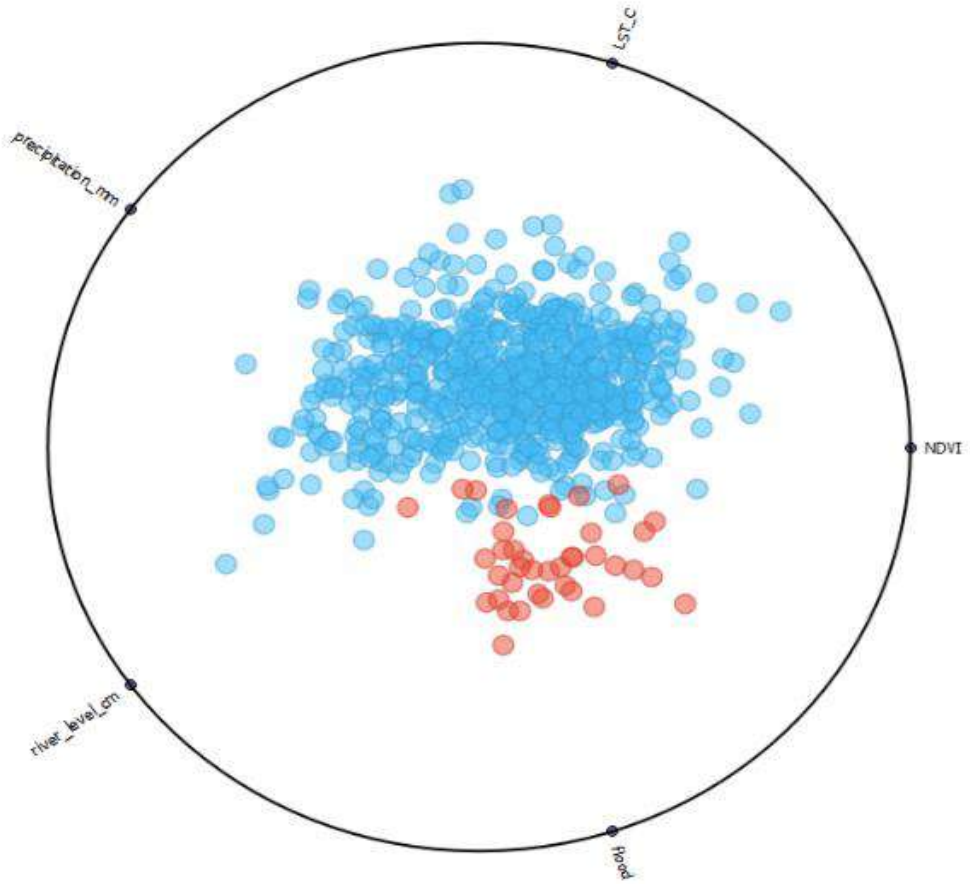


Figure 5.
Visualizing class separation using Radviz method.

Radial visualization (Radviz), implemented in the Orange environment, was used to analyze the data structure and visually assess the separability of the "flood" and "no flood" classes. This method allows for the reflection of multidimensional data in a two-dimensional space by evenly placing features on a circle and displaying observations depending on their values relative to these features.

Figure 5 displays the results of the projection of NDVI, LST_C, precipitation_mm, river_level_cm features, and the binary target variable flood in a radial coordinate system. Each point corresponds to a separate observation: blue dots indicate cases without flooding (label = 0), while red dots indicate cases with recorded flooding (label = 1).

The visualization shows a distinct cluster of red dots near the axes of the NDVI, river_level_cm, and flood signs, indicating a significant relationship between these variables and the target variable. In turn, flood-free observations are distributed more diffusely, mostly closer to the signs of LST_C and precipitation_mm. Thus, the Radviz method allows us to identify the key indicators that most influence the formation of classes. The results obtained confirm the relevance of the selected features and substantiate their application in the construction of predictive machine learning models in flood forecasting.

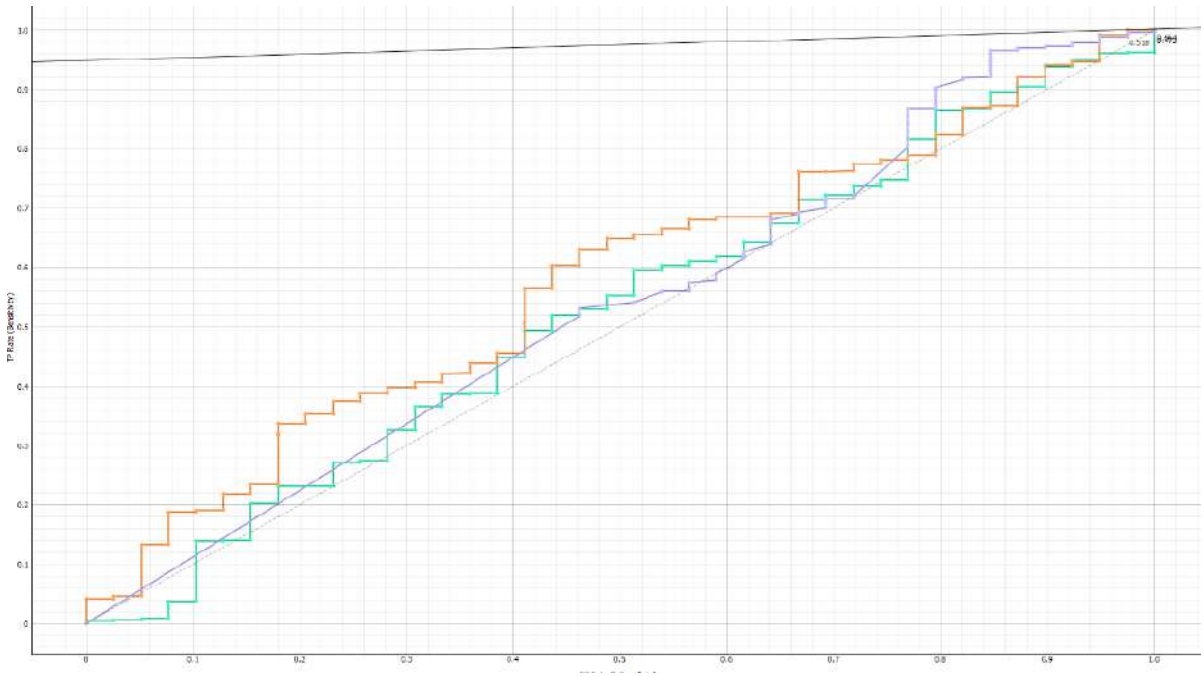


Figure 6.
Comparative ROC analysis of classification models.

Figure 6 shows the ROC curve (Receiver Operating Characteristic curve) illustrating the performance of three machine learning algorithms: Neural Network, Gradient Boosting, and Random Forest in solving the problem of binary classification of flood events (flood variable). The abscissa axis (FP Rate) shows the proportion of false positives (1 - specificity), and the ordinate axis (TP Rate) shows the proportion of true positives (sensitivity). The ideal model is characterized by the curve approaching the upper left corner (TP Rate = 1, FP Rate = 0), while the diagonal line (dashed line) corresponds to random guessing (AUC = 0.5).

Analysis of the results shows:

The Random Forest model (lilac curve) demonstrated the best results among the tested models, approaching the upper left corner. The AUC value is 0.918, which indicates a high ability of the model to correctly classify flood events.

Gradient boosting (orange curve) showed moderate results, with more pronounced steps in the curve and lower sensitivity, especially in the low false positive rate zone.

The neural network (green curve) demonstrated the lowest efficiency among the three algorithms, showing a limited ability to distinguish between classes (AUC is significantly lower).

All models were trained and tested in the Orange Data Mining environment using cross-validation. The AUC values calculated for each model serve as objective metrics of quality and are used to select the optimal classifier for further flood risk prediction.

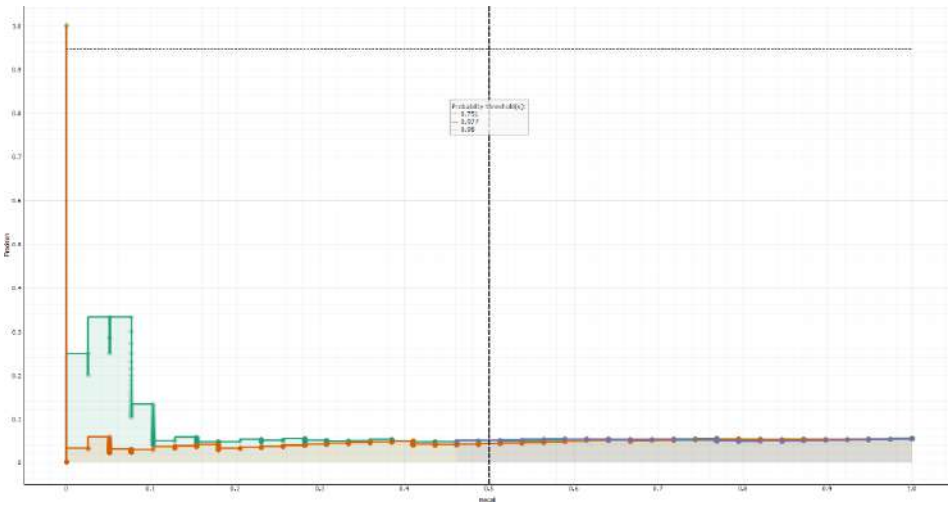


Figure 7.
Distribution of predictive classification probabilities for machine learning models.

Figure 7 shows the distribution of predicted probabilities of objects belonging to the positive class (flood) for three machine learning algorithms:

- Green line - neural network model;
- Orange line - gradient boosting;
- Blue line - random forest.

Most observations are concentrated in the low probability range (0.0–0.2), indicating the prevalence of the negative class (flood = 0) in the dataset. All three models demonstrate a similar distribution, with the neural network and random forest showing a more uniform decrease in density, and gradient boosting showing a more pronounced concentration in the initial interval.

The vertical dotted line denotes the standard classification threshold of 0.5. The majority of the predictions are to the left of this threshold, indicating the low frequency of the positive class and the tendency of the models to predict “no flood”.

The results indicate high confidence of the models in classifying the negative class and suggest a potential need for threshold adjustment to improve sensitivity in early warning tasks.

4. Conclusion

This paper presents the implementation of an integrated intelligent flood forecasting model based on machine learning algorithms and big data analysis in the Orange visual software environment. The use of meteorological and satellite indicators (NDVI, LST, precipitation, water level) allowed us to create an informative multidimensional dataset, on the basis of which predictive models were built and trained using Random Forest, Gradient Boosting algorithms, and neural networks.

The results obtained demonstrated high forecasting accuracy (up to 94%) and consistent values of quality metrics (F1-score, Precision, Recall, AUC), confirming the effectiveness of the proposed approach. Visual analysis of the data structure using Radviz and FreeViz methods provided additional insights into the contribution of individual features and visually confirmed class separability.

The novelty of the study lies in the successful integration of the Orange visual analysis environment, which allows users to build models without the need for programming, making solutions accessible to a broad audience of specialists. The proposed model can be adapted to different geographic regions, scaled for other climate threats, and integrated into existing monitoring and early warning platforms.

Thus, this study demonstrates the practical applicability of artificial intelligence methods and Big Data technologies in environmental forecasting and natural risk management, particularly in the field of hydrological monitoring and flood prediction.

References

[1] S. M. M. A. Asif, M. Farzana, I. Namir, I. Ishrar, M. H. Nushra, and T. Rahman, "Flood prediction using machine learning models," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2022: IEEE, pp. 1-6.

[2] W. Li, A. Kiaghadi, and C. Dawson, "High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks," *Neural Computing and Applications*, vol. 33, no. 4, pp. 1261-1278, 2021/02/01 2021. https://doi.org/10.1007/s00521-020-05010-6

[3] A. Mosavi, P. Ozturk, and K.-w. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, p. 1536. https://doi.org/10.3390/w10111536.

[4] A. W. Nab, V. Kumar, and R. L. H. L. Rajapakse, "Innovative methods for rapid flood inundation mapping in Pul-e-Alam and Khoshi districts of Afghanistan using Landsat 9 images: spectral indices vs. machine learning models," *Modeling Earth Systems and Environment*, vol. 10, no. 2, pp. 2495-2513, 2024. https://doi.org/10.1007/s40808-023-01914-z

[5] M. Billah, A. K. M. S. Islam, W. B. Mamoon, and M. R. Rahman, "Random forest classifications for landuse mapping to assess rapid flood damage using Sentinel-1 and Sentinel-2 data," *Remote Sensing Applications: Society and Environment*, vol. 30, p. 100947, 2023. https://doi.org/10.1016/j.rsase.2023.100947

[6] Y. Gao, "Research on the application of machine learning technology in hydrological flood prediction," *Journal of Computer, Signal, and System Research*, vol. 2, no. 2, pp. 28-34, 2025. https://doi.org/10.71222/se6cyv71

[7] A. Farrokhi, S. Farzin, and S.-F. Mousavi, "Meteorological drought analysis in response to climate change conditions, based on combined four-dimensional vine copulas and data mining (VC-DM)," *Journal of Hydrology*, vol. 603, p. 127135, 2021. https://doi.org/10.1016/j.jhydrol.2021.127135

[8] C. Gupta, R. Singh, and S. Shukla, "Flood risk mapping using satellite remote sensing and machine learning: A case study of the Ganges Basin, India," *Remote Sensing*, vol. 15, p. 362, 2023.

[9] J. Smith, X. Zhang, and A. Kumar, "Flood forecasting in central Asia using ensemble ML methods and geospatial data," *Environmental Modelling & Software*, vol. 155, p. 105432, 2022.

[10] E. Tursunov, S. Abdirasulov, and G. Alimova, "Integrated hydrological forecasting system for Kazakhstan based on ML and remote sensing," *Journal of Flood Risk Management*, vol. 16, p. e12988, 2023.

[11] P. Sharma, L. Singh, and K. Jain, "Use of SAR and NDVI indices for flood vulnerability mapping in India," *International Journal of Remote Sensing*, vol. 42, no. 12, pp. 4512–4532, 2021.

[12] T. Dzhumabekova, A. Bekbolotov, and R. Aliyev, "Machine learning–based early warning system for flood detection in Central Asian rivers," *Water Resources Management*, vol. 37, pp. 3457–3474, 2023.

[13] J. Demšar *et al.*, "Orange: Data mining toolbox in python," *the Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2349-2353, 2013.

[14] L. Li, B. Rakitsch, and K. Borgwardt, "ccSVM: Correcting support Vector machines for confounding factors in biological data classification," *Bioinformatics*, vol. 27, no. 13, pp. i342-i348, 2011. https://doi.org/10.1093/bioinformatics/btr204

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems,* vol. 30, pp. 4765–4774, 2017. https://doi.org/10.48550/arXiv.1705.07874

[19] G. Balakayeva, G. Kalmenova, D. Darkenbayev, and C. Phillips, "Development of an application for the thermal processing of oil slime in the industrial oil and gas sector," *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska,* vol. 13, no. 2, pp. 20-26, 2023. https://doi.org/10.35784/iapgos.3463

[20] A. Altybay, A. Nakiskhozhayeva, and D. Darkenbayev, "Numerical simulation and parallel computing of the acoustic wave equation," in *AIP Conference Proceedings*, 2024, vol. 3085, no. 1: AIP Publishing LLC, p. 020006.

[21] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *Journal of Hydrology*, vol. 527, pp. 1130-1141, 2015. https://doi.org/10.1016/j.jhydrol.2015.06.008

[22] N. Razali, S. Ismail, and A. Mustapha, "Machine learning approach for flood risks prediction," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, p. 73, 2020. https://doi.org/10.11591/ijai.v9.i1.pp73-80

[23] Y. Wu, Z. Zhang, X. Qi, W. Hu, and S. Si, "Prediction of flood sensitivity based on logistic regression, extreme gradient boosting, and random forestmodeling methods," *Water Science & Technology*, vol. 89, no. 10, pp. 2605-2624, 2024. https://doi.org/10.2166/wst.2024.146

[24] D. Darkenbayev, A. Altybay, Z. Darkenbayeva, and N. Mekebayev, "Intelligent data analysis on an analytical platform," *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, vol. 14, no. 1, pp. 119-122, 2024. https://doi.org/10.35784/iapgos.5423

[25] Y.-M. Chiang, L.-C. Chang, M.-J. Tsai, Y.-F. Wang, and F.-J. Chang, "Dynamic neural networks for real-time water level predictions of sewerage systems-covering gauged and ungauged sites," *Hydrology and Earth System Sciences*, vol. 14, no. 7, pp. 1309-1319, 2010. https://doi.org/10.5194/hess-14-1309-2010

[26] M. M. A. Syeed, M. Farzana, I. Namir, I. Ishrar, M. H. Nushra, and T. Rahman, "Flood prediction using machine learning models," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2022: IEEE, pp. 1-6.