Article

Not peer-reviewed version

A Novel Approach for Hydropower Real-Time Critical Data Detection

Maria Viorela Muntean * and Daniela Onita

Posted Date: 7 May 2026

doi: 10.20944/preprints202605.0347.v1

Keywords: classification; forecasting; partitioning clustering; critical data; real-time data

 doi:10.20944/preprints202605.0347.v1

Article

A Novel Approach for Hydropower Real-Time Critical Data Detection

Maria Viorela Muntean * and Daniela Oniță

Informatics and Engineering Department, 1 Decembrie 1918 University of Alba Iulia, Romania

* Correspondence: mmuntean@uab.ro

Highlights

- Real-time monitoring of hydropower critical data requires high-performing AI tools
- Existing models are built on already collected data
- The new proposed approach consists of building models on the forecasted data
- Achieving comparable or higher classification accuracy than other existing models.

Abstract

Real-time system monitoring without human intervention is an important issue nowadays. The challenge is to find the learning model that best suits each system. In hydropower systems, critical situations occur when the water reaches the spill level or the minimum exploitation level. The actual learning models use past data to detect such instances. Our approach is to build models on future data, which is more appropriate for learning from real data. Given that the current forecasting methods are well developed and have proven their performance (the RBF Regressor achieved an RMSE of 0.291 in the current work), we propose forecasting data stored within the next month and using it to build clustering and classification models. The results show that our proposed approach achieves higher classification accuracy (99.51%) and higher or comparable Precision, Recall, F-Measure, and MCC than those of other models trained on similar datasets.

Keywords: classification; forecasting; partitioning clustering; critical data; real-time data

1. Introduction

In hydropower developments, continuous monitoring and forecasting of reservoir quotas are among the most important tasks performed by both the staff of the hydropower dispatch and the operating personnel of hydroelectric power plants and reservoir dams. The monitoring of the level of the quota has main implications in the functional ensemble of a hydropower development and not only, namely: electricity production, calculation of mechanical power transformed into electrical power, establishing the hydropower reserve, management of critical situations (reaching the spill level or reaching the minimum exploitation level), scheduling of electricity production, monitoring the correlation between the elevation and the flow of infiltration from the dam body, scheduling of certain maintenance actions, efficient management of flood periods, preventing emergencies such as flooding. Monitoring the quota is essential for establishing the water reserve, and reservoir elevations are important parameters for the Romanian Waters National Administration's activities.

Recent studies aim to enhance energy consumption forecasting [1], [2], [3], [4], hydropower unit vibration forecasting [5], [6], optimization of hydropower reservoir operations [7], [8], [9], [10] or the fusion of preprocessing methods and deep learning to predict deformation in a damaged arch dam [11]. Regarding abnormalities, these are identified by comparing the joint cumulative distribution function value against a predefined alarm limit [12].

Anomalies are discovered by combining k-means partitioning clustering with quadratic discriminant analysis, yielding a 0.23 silhouette score [13], or by applying HDBSCAN density-based

 doi:10.20944/preprints202605.0347.v1

clustering, yielding a 0.4935 score [14]. Clustering methods do not predict human-defined labels. They learn from past data and are used to determine which cluster center is closest to the new instances.

Forecasting hydropower data is performed using Cascaded Adaptive Network-Based Fuzzy methods that reach a root mean square error of 1.01, and with the Gated Recurrent Unit algorithm with a root mean square error of 6.5 [15]. The reported results remain quite high, and forecasting is useful for observing future trends in data, but it cannot be used to label newly collected data. For the PC-MTS-GCN prediction model, the authors report a determination coefficient (R^2) of 0.98 and a Kline Gupta Efficiency Coefficient (KGE) of 0.96 for water level [16].

Other studies report coupling large language models, specialized intelligent agents, and traditional automation technologies to enhance the automation of hydropower operations [17] and a framework that integrates structural fatigue and flood modeling to identify probabilistic risk data [18]. Hydropower stations are monitored using Artificial Intelligence and the Internet of Things/Cloud-native technology [19], [20], [21], [22] or a time-series Transformer method that integrates vibration, electrical, and hydraulic monitoring data [23].

The main objective of the proposed intelligent system is to provide the most precise assistance to the operating personnel and hydropower dispatchers. The proposed procedure consists of applying several machine-learning steps to the collected data. First, we applied various forecasting methods to accurately predict the evolution of Oașa Lake in Romania. Second, a clustering procedure was applied to obtain the cluster centroids corresponding to the three water-level types: normal water level, the minimum technical level of exploitation, and the spillway level. Next, the labelled data were classified. We have trained the classifiers to identify the most suitable one, which can accurately detect critical data in real time. Finally, the system was able to make decisions and send notifications to operating personnel and the hydropower dispatcher regarding critical water levels.

2. Methodology

The proposed approach consists of an architecture that enables the construction of classification models from forecasted data. Given that existing forecasting methods are highly performant, the predicted data is closer to the data that will be collected and classified in the short term. In this way, the classification model proved to be more performant than models built on past data.

The proposed system is composed of a set of iterative steps, as follows (Figure 1): collecting data from a SCADA system, forecasting the collected data to find the next instances of the system, clustering the previously forecasted data to identify the critical data, building a classifier that can predict such data, and activating a decision/notification system when critical data occur. Once the system is built and trained, new data from the SCADA system can be used to monitor the hydropower system in real time.

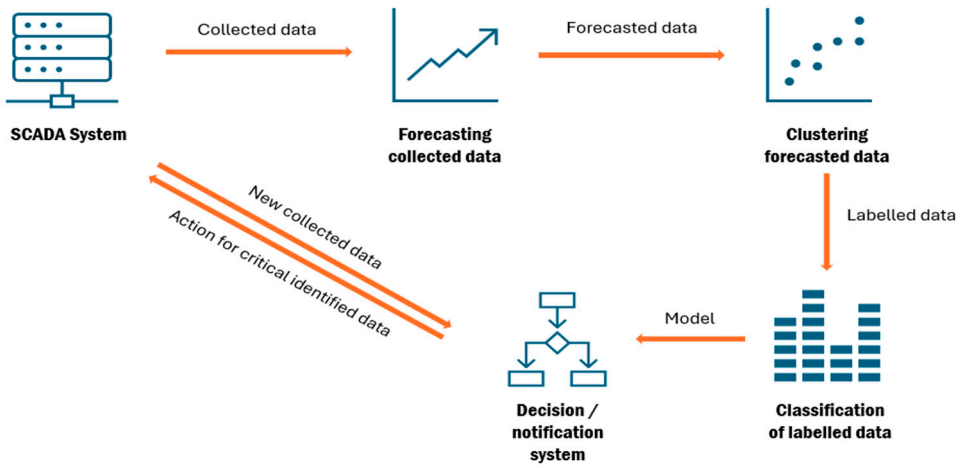

Figure 1. System's architecture.

 doi:10.20944/preprints202605.0347.v1

In particular, the proposed process includes (Figure 2) forecasting the entire dataset to analyze forecasting methods on a large amount of data stored over a longer period, then splitting the dataset into a training set (80% of instances) and a testing set (20% of instances). The proposed process then proceeds with two parallel operations: first, the system uses the 80% training instances to predict the remaining 20% instances with the most suitable classifier discovered in the previous step, and then applies partitioning clustering methods and evaluates the discovered clusters; on the other hand, the system clusters the testing set (20% of instances) and evaluates the discovered clusters.

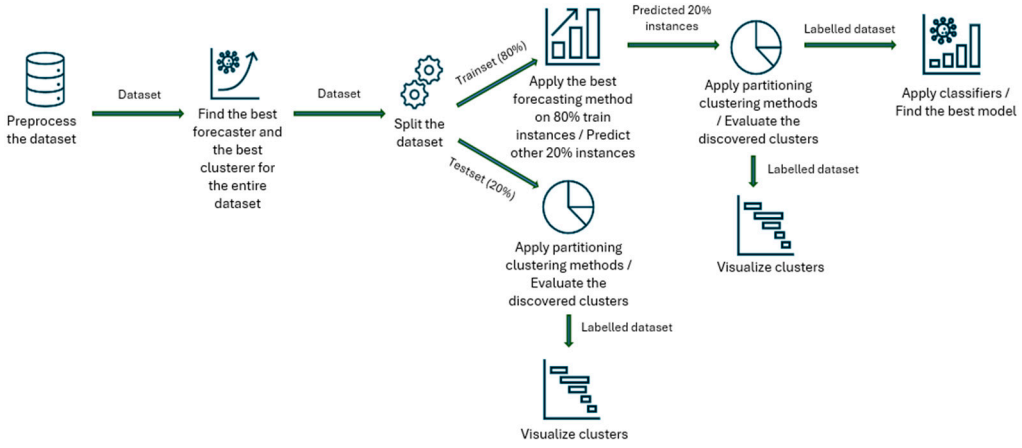

Figure 2. The workflow of the process.

The results obtained from the two parallel cluster evaluations are compared, and the proposed approach (the first proposed clustering procedure) is validated. At the end of the learning process, the best classifier is discovered using the dataset clustered with the proposed procedure.

Models were trained and tested on 4122 instances using a workstation with an Intel i7-1165G7 CPU and 8 GB of RAM.

3. Experimental Results

3.1. Hydropower System Description

The proposed system aims to monitor and predict reservoir level and infiltration within the Oaşa Dam body in Romania. The purpose is to assist hydropower dispatchers and operating personnel in monitoring and predicting the Oaşa Lake elevation and infiltration within the dam body, especially for the most efficient management of critical situations (reaching the discharge level, reaching the minimum technical level of exploitation, and reaching the critical levels of infiltration within the dam body).

The complex arrangement of the Sebeș River includes four hydropower plants arranged in a cascade, a pumping station, and two micro-hydropower plants. The first and largest of the lakes that form the waterfall, located along the course of the Sebeș River, is the Oașa Lake. It concentrates the waters of the Sebeș River and other local watercourses through secondary catchments. At the normal retention level (1255 m³/m), Oașa Lake has a volume of 136 million cubic meters and an area of 454 hectares. The dam, which is 91 meters high, is made of rock. The Gîlceag underground power plant, commissioned in 1980, harnesses the energy potential stored in Lake Oașa. The plant, equipped with two vertical hydrogenators driven by turbines, has an installed capacity of 150 MW. In 2003, the Gîlceag Pumping Station was put into operation, which, with the help of two 10 MW pumps, transfers the water accumulated in Lake Cugir, through the penstock of the Gîlceag hydropower plant, to Lake Oașa, to supplement the energy produced in the Sebeș Hydropower Development [24], [25].

In hydropower developments, the SCADA system is essential for monitoring and controlling dam and hydropower plant operations. At a reservoir dam, SCADA collects data from a diverse

 doi:10.20944/preprints202605.0347.v1

network of sensors that measure parameters such as water level, flow, pressure, and the dam's structural condition. Terminal stations (RTUs) at the measurement points convert the analog or digital signals from the sensors into data transmitted over the hydropower development's internal network. This data reaches the central SCADA servers, where it is processed and stored. Within the internal network, the SCADA server uses specialized software to analyze and display information in real time via HMI interfaces, providing operators with a clear picture of the dam and hydraulic system status [26]. The system also allows automatic or manual control of equipment (e.g., turbines, exhaust valves), ensuring optimal operation and maximum safety.

3.2. Dataset Description

The main data collected by the SCADA system for the reservoir and the Oașa dam include: recording date, lake level (water level in the lake), temperature, rainfall amount, snow layer thickness, total influent flow, turbine flow, power group1, function group1, power group2, function group2, pump1 running time, pump1 flow, pump2 running time, pump2 flow, and infiltration flows from the body of the Oașa dam (DM22, downstream spillway 1).

We simulated data for the period from January 1, 2025, to June 21, 2025, and collected 4122 instances. The data was preprocessed in the Attribute-Relation File Format (.arff) for use with machine learning algorithms. The attribute types and a sample of the preprocessed data are shown in Figure 3.

```
@relation hydropower_data
@attribute rec_date date "yyyy-MM-dd HH:mm:ss"
@attribute 'water_level (mdM)' numeric
@attribute 'temperature (C)' numeric
@attribute 'rainfall (mm/h)' numeric
@attribute 'snow_layer (cm)' numeric
@attribute 'total_influent_flow (mc/s)' numeric
@attribute 'turbine_flow (mc/s)' numeric
@attribute 'power_group1 (MW)' numeric
@attribute 'function_group1 (min)' numeric
@attribute 'power_group2 (MW)' numeric
@attribute 'function_grup2 (min)' numeric
@attribute 'pump1_running_time (min)' numeric
@attribute 'flow_pump1 (mc/s)' numeric
@attribute 'pump2_running_time (min)' numeric
@attribute 'flow_pump2 (mc/s)' numeric
@attribute 'DM22 (mc/s)' numeric
@attribute 'dev_aval1 (cm)' numeric
@data
"2025-01-01 00:00:00",1248.01,-5,0,50,3.52,0,0,0,0,0,60,4.5,0,0,-0.000005,13.34
"2025-01-01 01:00:00",1248,-5,0,50,3.52,17.32,0,0,65,60,60,4.5,60,4.5,-0.000006,13.33
"2025-01-01 02:00:00",1248.01,-4,0,52,3.52,0,0,0,0,0,0,0,60,4.5,-0.000005,13.34
"2025-01-01 03:00:00",1248.02,-3,0,53,3.52,0,0,0,0,0,0,60,4.5,0,0,-0.000005,13.34
"2025-01-01 04:00:00",1248,-3,0,53,3.52,18.65,0,0,70,60,0,0,0,0,-0.000006,13.33
"2025-01-01 05:00:00",1248.01,-4,0,54,3.53,0,0,0,0,0,0,0,60,4.5,-0.000005,13.34
"2025-01-01 06:00:00",1248.03,-4,0,55,3.53,0,0,0,0,0,60,4.5,60,4.5,-0.000005,13.34
"2025-01-01 07:00:00",1248.01,-4,0,54,3.53,15.99,60,60,0,0,0,0,0,0,-0.000005,13.34
"2025-01-01 08:00:00",1248.02,-4,0,54,3.53,0,0,0,0,0,0,0,0,0,-0.000005,13.34
"2025-01-01 09:00:00",1248.01,-4,0,53,3.53,10.66,40,60,0,0,0,0,0,0,-0.000005,13.34
"2025-01-01 10:00:00",1247.99,-4,0,54,3.53,17.32,65,60,0,0,0,0,0,0,-0.000006,13.33
```

Figure 3. Sample of preprocessed data.

3.3. Evaluation of the Applied Forecasting Methods

To identify the most accurate forecasting method for predicting the evolution of the Oașa Lake water level, we applied LinearRegression, RBFRegressor, and SMOReg from the Weka Machine Learning Tool [27], [28]. The obtained results are presented in Table 1, Figure 4, Figure 5, and Figure 6. Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are the main benchmarks for evaluating forecasting models. For all metrics, lower values are better; 0 indicates a perfect forecast.

 doi:10.20944/preprints202605.0347.v1

The results indicate that RBFRegressor accurately predicted the data. The MAE for predicting 2000 instances with RBFRegressor (50% more instances than the initial training set) was equal to 0.178, and the MAE for predicting 824 instances (20% more instances) was equal to 0.172. The MSE for the 824 predicted instances was 0.086, indicating that the RBFRegressor is a suitable forecaster for the considered dataset.

Table 1. Forecasting results for predicting 824 instances (20% more instances than the initial dataset).

Forecaster	MAE	MSE	RMSE
LinearRegression	0.338	0.260	0.457
RBFRegressor	0.172	0.086	0.291
SMOReg	0.411	0.297	0.502

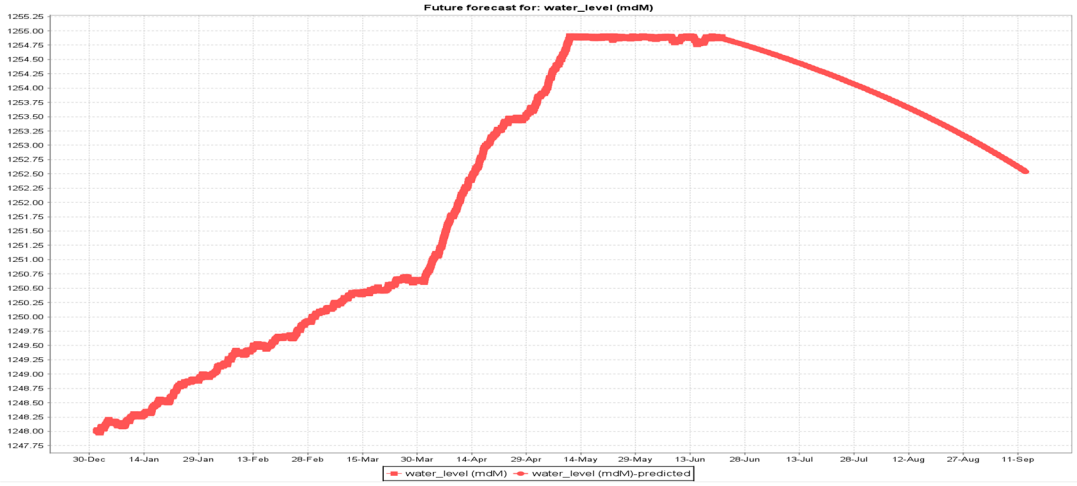

Figure 4. Water-level forecasting using LinearRegression to predict 2000 instances.

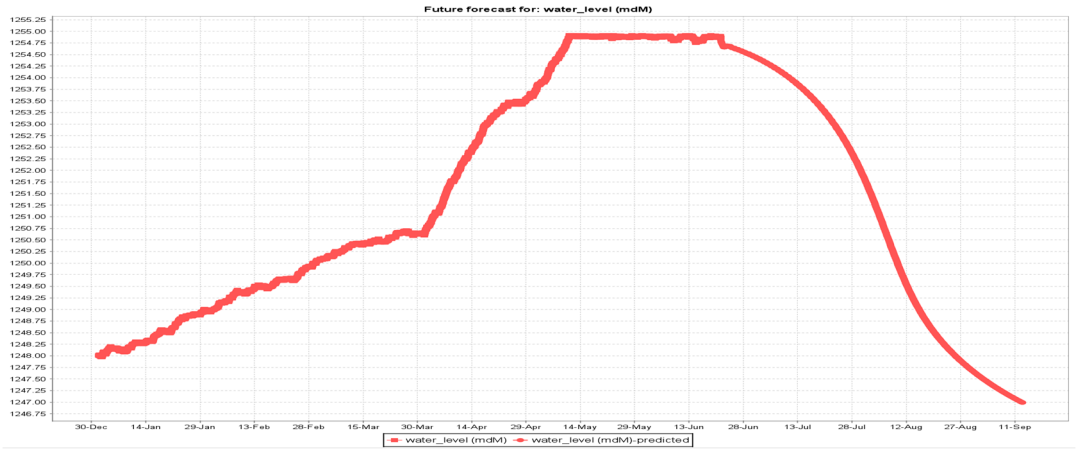

Figure 5. Water-level forecasting using an RBFRegressor to predict 2000 instances.

 doi:10.20944/preprints202605.0347.v1

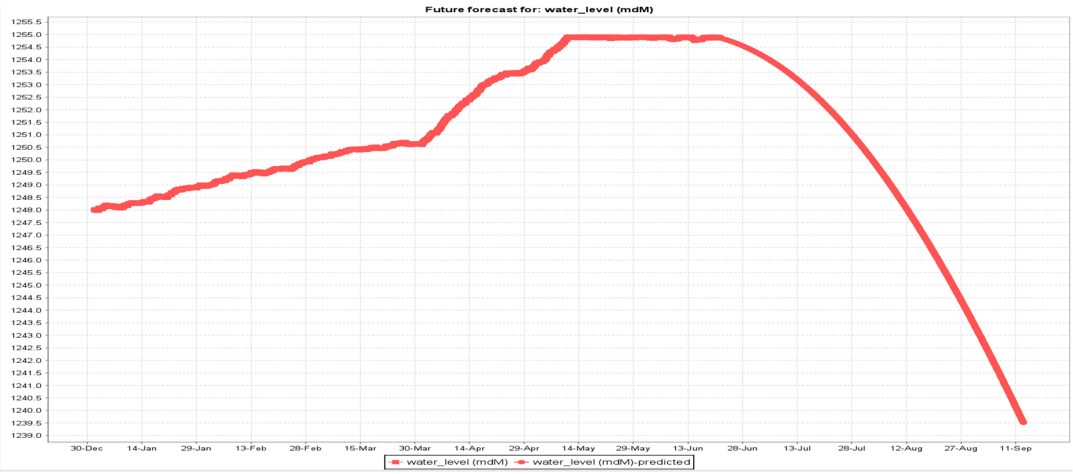

Figure 6. Water-level forecasting using SMOReg to predict 2000 instances.

3.4. Clustering the Proposed Dataset

For the clustering procedure, we used the XMeans partitioning method and the FarthestFirst heuristic method, both from the Weka Machine Learning tool [27], [28]. The cluster centroids returned by the applied methods corresponded to three water-level types: normal water levels, those reaching the minimum technical exploitation level, and those reaching the spillway level. The XMeans model discovered on the entire dataset is described below:

```
Cluster 0

1.742702055052265E12 1251.39 -0.04 0.15 39.24 7.61 8.36 18.25 18.44 13.34 13.22 60.0 4.5

35.12 2.63 3.215679442508741E-5 14.46

Cluster 1

1.7475842978313252E12 1254.30 8.15 0.14 1.45 13.12 14.27 30.12 30.02 23.89 23.71 0.0 0.0

4.24 0.31 6.450923694779208E-5 15.43

Cluster 2

1.7401356936957778E12 1249.74 -4.00 0.16 60.27 4.33 4.04 9.95 10.61 5.26 5.55 0.0 0.0 4.12

0.30 1.376865240023133E-5 13.91
```

In XMeans clustering, Distortion and the BIC (Bayesian Information Criterion) are used together to balance how well clusters fit the data against how many clusters are used. Distortion measures the compactness of the clusters and is defined as the sum of the squared distances from each data point to the cluster centroid to which it is assigned. The ideal value for distortion is lower. BIC is used in XMeans to decide whether to split a cluster into two smaller ones and combines the likelihood (related to distortion) with a penalty for the number of clusters to prevent overfitting.

In Weka's implementation of XMeans, the highest (least negative) BIC value identifies the best clustering structure. For the considered dataset, the Distortion value for the discovered clusters was 5134.19, and the BIC value was -10749.93 (Figure 7), indicating a less accurate model. Also, the graphical representation of the water-level and downstream-spillway clusters shows that they are not well separated, and the model did not distinguish between their instances.

 doi:10.20944/preprints202605.0347.v1


```
Distortion: 5134.193336
BIC-Value : -10749.931908


Time taken to build model (full training data) : 0.1 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       1064 ( 26%)
1       1329 ( 32%)
2       1729 ( 42%)
```

Figure 7. XMeans performance metrics for the entire dataset.

The Farthest-First heuristic method discovered the following cluster centroids:

```
Cluster 0

1.7453268E12 1253.33 -5.0 0.0 0.0 8.9 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 5.4E-5 15.11

Cluster 1

1.7386344E12 1249.15 -5.0 0.0 101.0 3.58 33.25 65.0 60.0 60.0 60.0 60.0 4.5 60.0 4.5 7.0E-

6 13.72

Cluster 2

1.7471052E12 1254.88 17.0 3.0 0.0 15.04 39.6 75.0 60.0 75.0 60.0 60.0 4.5 0.0 0.0 7.1E-5

15.63
```

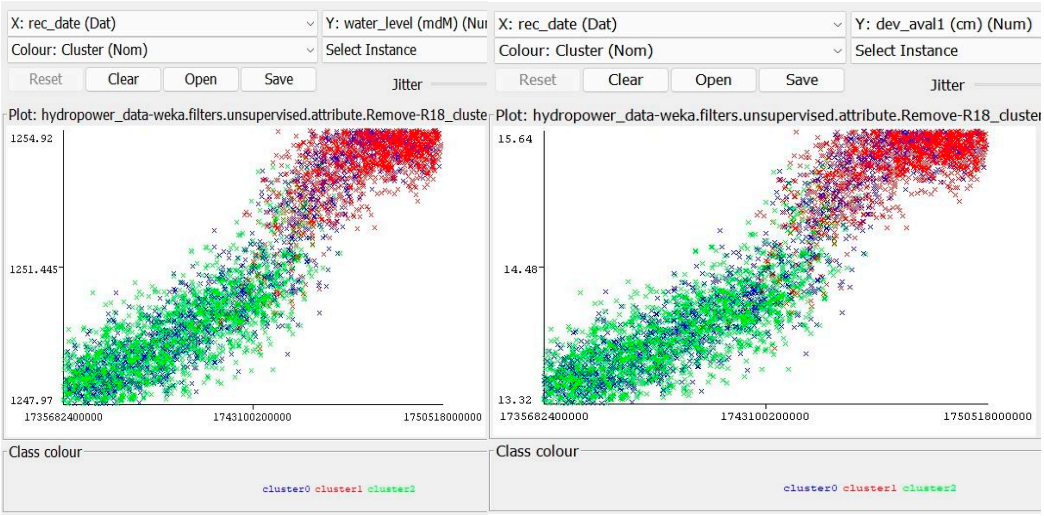

Figure 8. Water level clusters and downstream spillway clusters were discovered on the entire dataset using XMeans.

The cluster distribution (Figure 9) and the graphical representation of clustered instances (Figure 10) show that for the monitored attributes, namely water level and downstream spillway, the clusters are not well separated.

 doi:10.20944/preprints202605.0347.v1


```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       3149 ( 76%)
1        554 ( 13%)
2        419 ( 10%)
```

Figure 9. FarthestFirst performance metrics for the entire dataset.

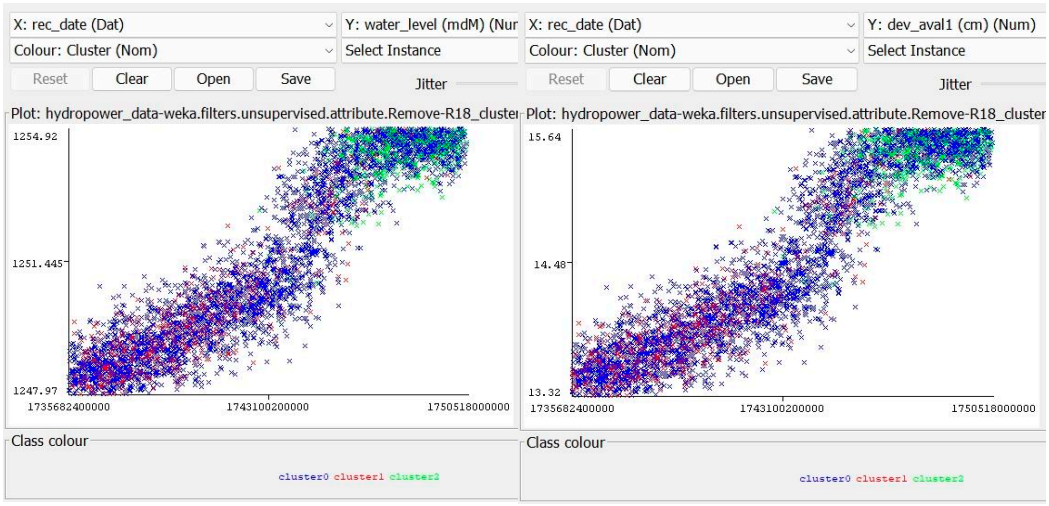

Figure 10. Water level clusters downstream spillway clusters discovered on the entire dataset using FarthestFirst.

3.5. Splitting the Dataset into a Training Set and a Testing Set

Next, we decided to use 80% split percentage method to divide the dataset into a training set (80% of instances, 3298 instances, Figure 11) and a testing set (the remaining 20% of instances, 824 instances, Figure 12), due to the results returned by the RBFRegressor forecasting model. The obtained datasets will be used in the next phases of the proposed process.

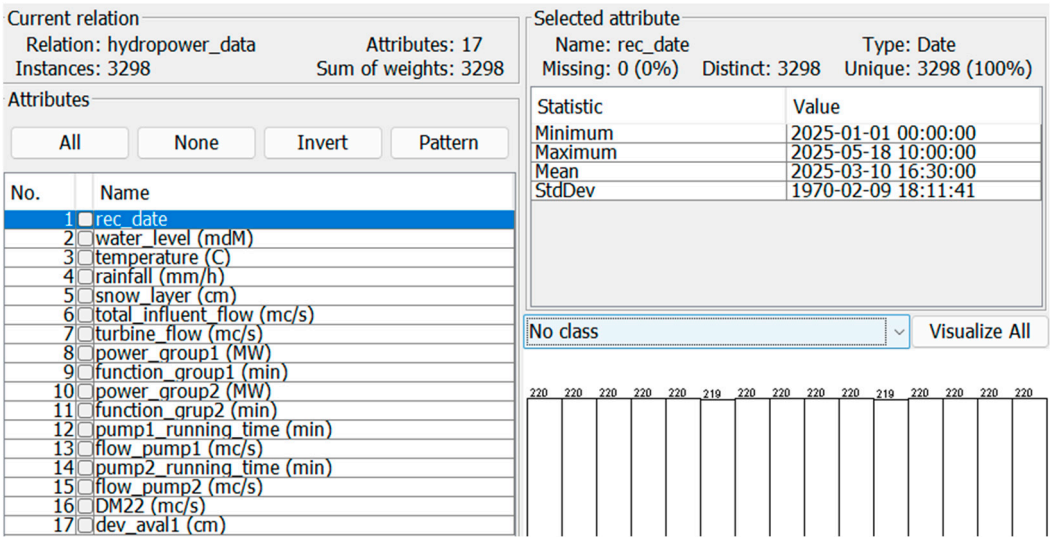

Figure 11. Training set description.

 doi:10.20944/preprints202605.0347.v1

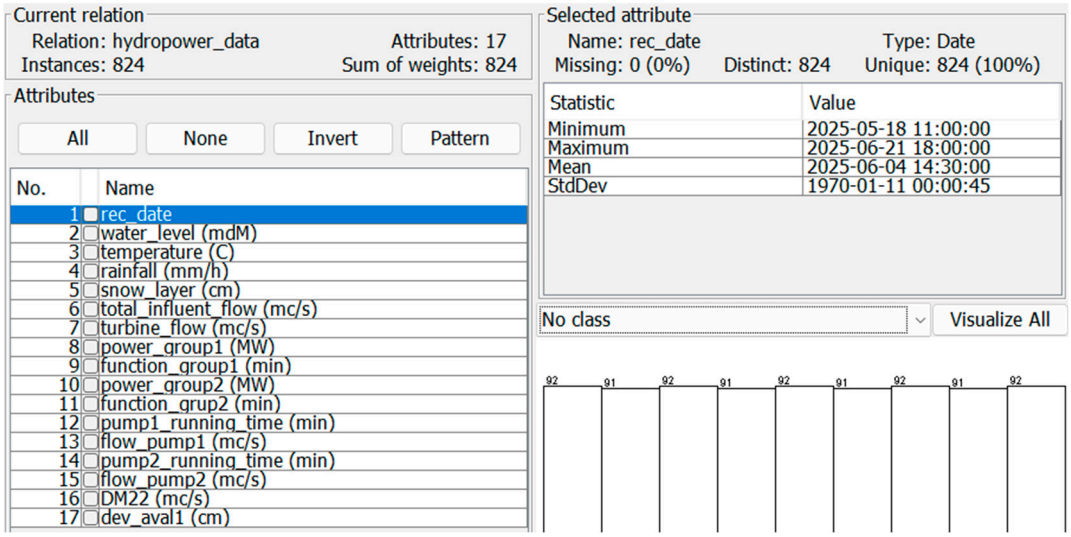

Figure 12. Testing set description.

3.6. Clustering the Actual Testing Set

The cluster centroids returned by the XMeans partitioning method at this stage, corresponding to the three analyzed situations, are described below:

```
Cluster 0

1.7491001098265896E12 1254.88 11.27 0.14 0.0 14.35 4.24 0.0 0.0 16.09 17.16 6.93 0.523

5.026 0.37 7.087861271676348E-5 15.62

Cluster 1

1.7491582229508196E12 1254.86 10.83 0.14 0.0 14.32 15.10 57.21 60.0 0.0 0.0 6.55 0.49 3.93

0.29 7.081420765027323E-5 15.62

Cluster 2

1.7488866630508474E12 1254.87 10.10 0.15 0.0 14.46 30.53 57.10 53.08 58.55 54.71 22.37 1.67

18.71 1.40 7.084745762711896E-5 15.623
```

For the considered testing set, the Distortion value was 957.33, and the BIC value was -2339.63 (Figure 13).

```
Distortion: 957.336353
BIC-Value : -2339.635784


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       357 ( 43%)
1       195 ( 24%)
2       272 ( 33%)
```

Figure 13. XMeans results for clustering the actual testing set.

As a remark, the discovered clusters were not well separated for either the water level or the downstream spillway attributes (Figure 14), given the temporal variation in both attributes.

 doi:10.20944/preprints202605.0347.v1

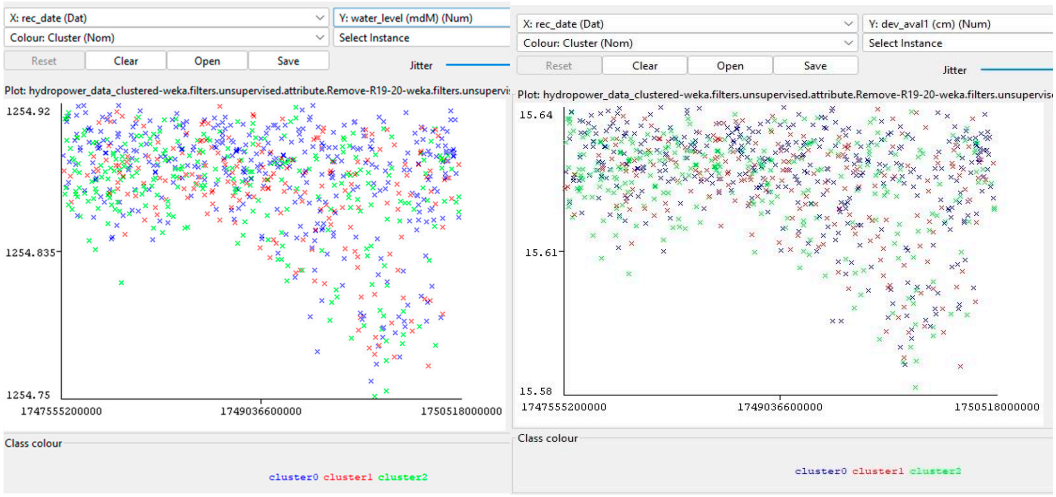

Figure 14. Water level clusters downstream spillway clusters discovered on the actual testing set.

3.7. Applying the Best Forecasting Method to the Train Instances. Predict the Test Instances

RBFRegressor was used on the training set, and 824 instances were forecast. The actual and predicted water level values were evaluated using graphical representations (Figure 15), and the predicted values for all attributes (Figure 16) were sent to the clustering component.

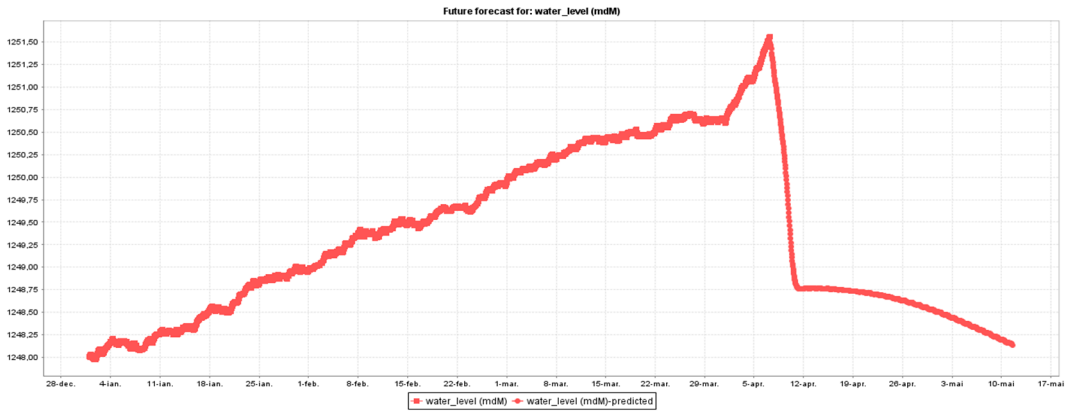

Figure 15. Actual (train set) and predicted values for water level.

```
@relation hydropower_forecasted_data
@attribute rec_date date 'yyyy-MM-dd HH:mm:ss'
@attribute 'water_level (mdM)' numeric
@attribute 'temperature (C)' numeric
@attribute 'rainfall (mm/h)' numeric
@attribute 'snow_layer (cm)' numeric
@attribute 'total_influent_flow (mc/s)' numeric
@attribute 'turbine_flow (mc/s)' numeric
@attribute 'power_group1 (MW)' numeric
@attribute 'function_group1 (min)' numeric
@attribute 'power_group2 (MW)' numeric
@attribute 'function_grup2 (min)' numeric
@attribute 'pump1_running_time (min)' numeric
@attribute 'flow_pump1 (mc/s)' numeric
@attribute 'pump2_running_time (min)' numeric
@attribute 'flow_pump2 (mc/s)' numeric
@attribute 'DM22 (mc/s)' numeric
@attribute 'dev_aval1 (cm)' numeric
@data
'2025-05-18 11:00:00',1254.7943,7.0207,-0.0949,-0.1696,15.0178,19.1311,39.5383,41.7518,32.5951,26.4691,2.1096,0.6092,-0.9028,-0.0677,0.0001,15.6067
'2025-05-18 12:00:00',1254.7827,7.2756,0.5599,-0.4537,15.0594,13.7884,14.3483,19.0261,34.5429,31.8501,2.2629,0.5122,-1.1428,-0.0857,0.0001,15.6038
'2025-05-18 13:00:00',1254.7729,7.5061,0.2093,-0.6279,15.0786,15.5266,20.2241,10.3748,35.0216,31.5851,4.2549,0.6252,0.4692,0.0352,0.0001,15.6013
'2025-05-18 14:00:00',1254.7536,7.3238,-0.1601,-0.0857,15.0362,19.5818,33.8082,26.84,36.6126,30.7463,2.5699,0.7903,0.3385,0.0254,0.0001,15.5961
'2025-05-18 15:00:00',1254.7345,7.3815,0.2413,-0.1661,15.0407,20.0807,33.2,35.4982,37.683,37.3283,-3.6674,0.4789,-2.2007,-0.1651,0.0001,15.5914
'2025-05-18 16:00:00',1254.7283,7.7578,0.2402,-0.3119,15.0173,12.1565,18.1783,20.8553,21.2889,28.9631,8.0451,1.2481,8.2545,0.6191,0.0001,15.59
'2025-05-18 17:00:00',1254.6962,7.441,0.7297,-0.1969,15.1078,16.5002,26.8905,21.6333,28.0282,29.0169,1.5712,0.5147,2.0594,0.1545,0.0001,15.5827
'2025-05-18 18:00:00',1254.6773,7.6457,0.4483,-0.154,15.2569,8.0074,19.5255,23.6128,0.4258,13.1682,-10.3868,-0.2775,3.1072,0.233,0.0001,15.5803
'2025-05-18 19:00:00',1254.6703,7.9033,0.5773,0.0524,15.3059,10.6442,7.8853,9.5731,20.776,27.9234,-1.7742,0.4433,9.905,0.7429,0.0001,15.5809
'2025-05-18 20:00:00',1254.6594,7.8253,0.7869,-0.1178,15.5089,10.9481,23.3291,25.6061,6.2545,19.5498,3.7593,1.1093,11.7108,0.8783,0.0001,15.5808
'2025-05-18 21:00:00',1254.6479,7.6052,1.1342,-0.3138,15.5976,7.4639,4.3524,-1.1098,11.0802,25.4893,5.4753,0.9696,11.0165,0.8262,0.0001,15.5767
'2025-05-18 22:00:00',1254.6374,7.9303,1.1779,-0.351,15.5335,6.9794,12.6003,10.8793,1.119,17.5002,20.4121,1.7617,29.7194,2.229,0.0001,15.5739
'2025-05-18 23:00:00',1254.6153,7.6499,1.061,-0.2523,15.657,6.7647,-3.6037,-4.8597,12.8731,28.8265,34.3908,2.9628,35.304,2.6478,0.0001,15.5701
```

Figure 16. Sample of forecasted data.

 doi:10.20944/preprints202605.0347.v1

3.8. Clustering the Forecasted Data

At this stage, we applied the XMeans clustering method to the dataset provided in the previous step. The discovered model consisted of the following cluster centroids:

```
Cluster 0

1.7500069484536082E12 1254.80 2.90 1.24 0.82 18.35 6.92 -1.27 0.96 11.55 37.00 11.06 1.73

5.90 0.44 9.999999999999968E-5 15.58

Cluster 1

1.7494624472727273E12 1254.69 2.42 1.28 1.75 19.05 5.02 -4.83 -1.19 8.41 33.37 13.29 1.75

6.40 0.48 9.999999999999965E-5 15.55

Cluster 2

1.748348956097561E12 1254.53 2.26 1.16 3.16 18.95 3.32 -6.05 -1.59 5.07 26.40 14.89 1.71

7.75 0.58 1.0000000000000036E-4 15.51
```

The Distortion (equal to 309.96) and BIC-Value (equal to 6014.03) indicate that the obtained model is better than the model obtained on the actual data (Figure 17), and the identified clusters are well separated for both water-level and downstream-spillway attributes (Figure 18).

```
Distortion: 309.965059
BIC-Value : 6014.037975


Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      194 ( 24%)
1      273 ( 33%)
2      357 ( 43%)
```

Figure 17. XMeans results for clustering the forecasted dataset.

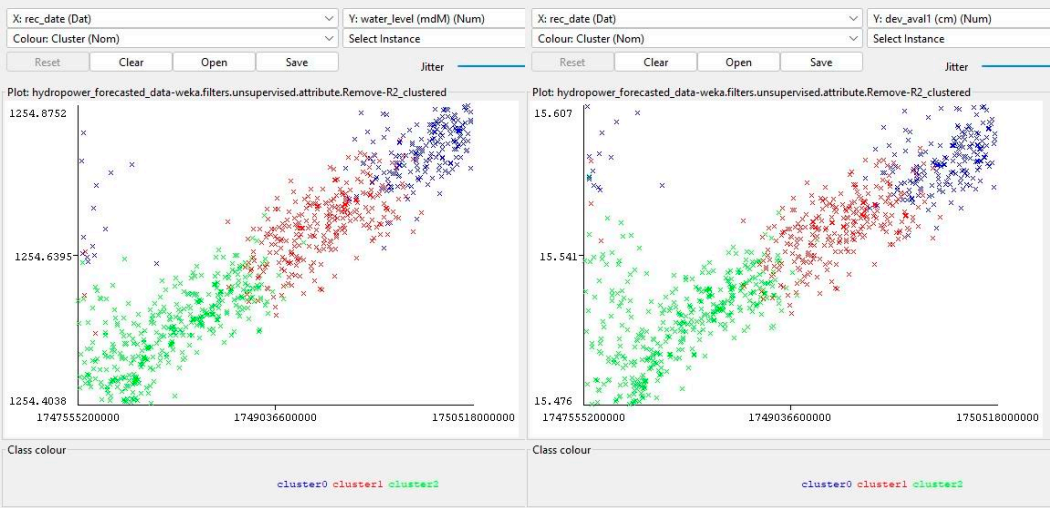

Figure 18. Water level clusters and downstream spillway clusters discovered on predicted data.

Our proposed approach proved more accurate at discovering clusters for the considered dataset, and the labelled data from this step were used to identify the most suitable classifier for accurately detecting critical data in real time.

 doi:10.20944/preprints202605.0347.v1

3.9. Classification of the Labelled Forecasted Data

The next step was to identify the most suitable classifier for the data cluster under consideration. The classification model was evaluated on a dataset comprising 824 instances. The evaluation results for each classifier are presented in Table 2 and Figure 19.

Table 2. Evaluation results of the classification models.

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistics	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
IB5	98.67%	1.34%	0.9794	0.0201	0.0894	4.66%	19.2541
PART	98.91%	1.09%	0.983	0.0084	0.0833	1.96%	17.93%
JRip	98.67%	1.34%	0.9794	0.0103	0.0926	2.39%	19.93%
J48	98.79%	1.21%	0.9813	0.0091	0.0896	2.10%	19.29%
RandomForest	99.15%	0.85%	0.9869	0.0081	0.0599	1.88%	12.90%
MLP	98.54%	1.46%	0.9775	0.0099	0.0786	2.29%	16.92%
SimpleLogistic	98.67%	1.34%	0.9794	0.0156	0.0882	3.62%	19.00%
CostSensitive1	99.27%	0.73%	0.9887	0.0087	0.0591	2.03%	12.72%
CostSensitive2	99.51%	0.49%	0.9925	0.0088	0.06	2.04%	12.92%

Figure 19 shows the performance of each classifier measured by Precision, Recall, F-Measure, and MCC. CostSensitive2 and SimpleLogistic achieved the highest overall predictive quality, while simpler models like IB5 and PART performed moderately.

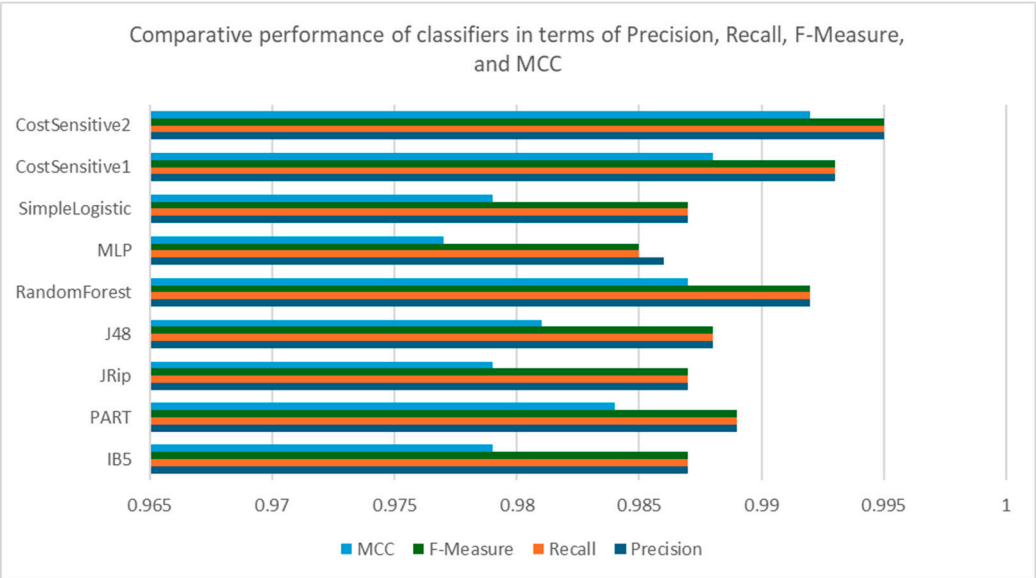

Figure 19. Comparative performance of classifiers in terms of Precision, Recall, F-Measure, and MCC.

The scatter plot from Figure 20 illustrates the trade-off between Precision and Recall. Classifiers closer to the top-right corner exhibit balanced high Precision and Recall.

To assess the classifiers' ability to distinguish between classes, we selected the ROC Area and PRC Area for analysis. Higher values indicate better performance. RandomForest, CostSensitive1, and CostSensitive2 achieved the highest ROC and PRC Areas.

 doi:10.20944/preprints202605.0347.v1

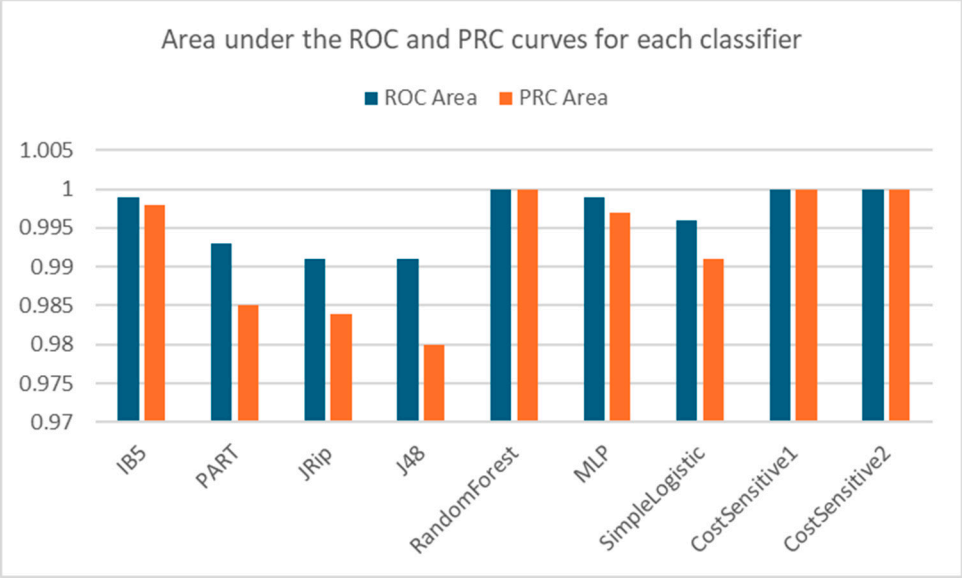

Figure 20. Area under the ROC and PRC curves for each classifier.

Figure 21 shows the computational cost of each classifier. Some models, such as MLP, SimpleLogistic, and RandomForest, take longer to build than rule-based classifiers.

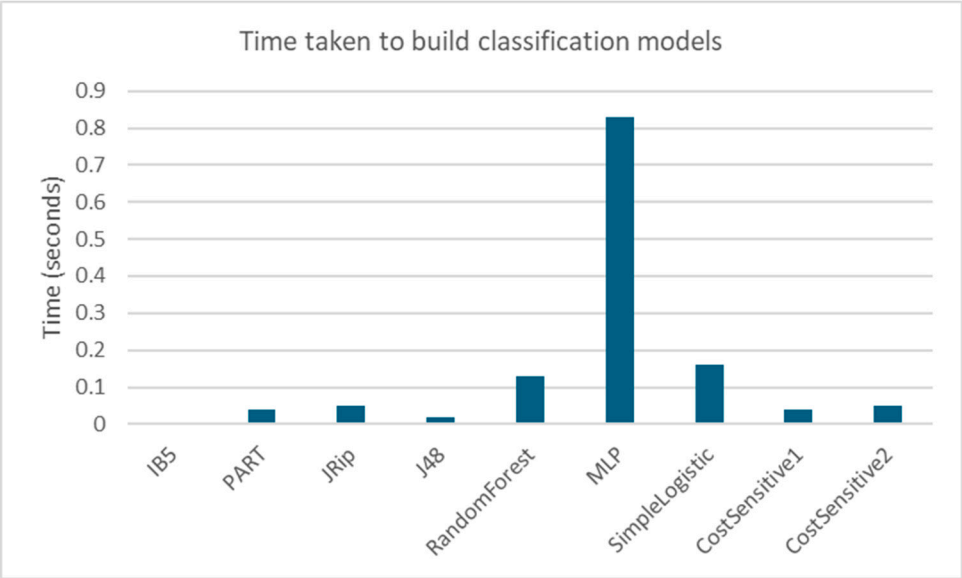

Figure 21. Time taken to build each classifier (seconds).

The results using the CostSensitive2 classifier (Figure 22) indicate high overall predictive performance, with 99.51% of instances correctly classified and 4 (0.48%) incorrectly classified. The Kappa statistic, which measures agreement between predicted and actual classifications beyond chance, was 0.9925. Error metrics, including the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE), were computed to assess the magnitude of prediction errors.

 doi:10.20944/preprints202605.0347.v1


```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Cost Matrix
 0    1    1
 1    0    2.2
 2    3    0


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         820               99.5146 %
Incorrectly Classified Instances         4                0.4854 %
Kappa statistic                          0.9925
Mean absolute error                      0.0088
Root mean squared error                  0.06
Relative absolute error                  2.0421 %
Root relative squared error             12.917  %
Total Number of Instances              824

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.000    1.000      0.995    0.997      0.997    1.000     1.000     cluster0
                 0.993    0.002    0.996      0.993    0.994      0.992    1.000     1.000     cluster1
                 0.997    0.006    0.992      0.997    0.994      0.990    1.000     1.000     cluster2
Weighted Avg.    0.995    0.003    0.995      0.995    0.995      0.992    1.000     1.000

=== Confusion Matrix ===

   a   b   c   <-- classified as
 193   0   1 |   a = cluster0
   0 271   2 |   b = cluster1
   0   1 356 |   c = cluster2
```

Figure 22. Classification results using the Cost-Sensitive classifier and the best-discovered Cost Matrix.

These metrics collectively demonstrate that the classifiers achieved highly reliable predictions with minimal deviation from the true class labels.

4. Conclusions

In this paper, we propose a novel approach for detecting critical hydropower data in real time. Instead of learning from past data, our approach builds a model from forecasted future data. The experimental results show that our proposed approach achieves higher classification accuracy, meaning 99.51% and higher or comparable Precision, Recall, F-Measure, and MCC than those of other models trained on similar datasets. Real-time water level monitoring is performed using classification techniques and reaches high accuracy levels in other works as well: in [29] the authors applied the k-nearest neighbour algorithm and found models that detect data with 90% accuracy, in [30] a neural network was trained, leading to 98.6% accuracy, in [31] deep learning models were applied (YOLOv5 and CNNs) leading to a mean average precision of 96.8%, a quantum-inspired neural network algorithm is proposed in [32] with an accuracy rate of 93%, a recall rate of 89%, and an F1 value of 91%, and in [33] Long Short-Term Memory models are proposed and the authors report an accuracy rate as high as 98.50%, and the F1 score is 96.14%. The Hydrostatic-Seasonal-Time model and transfer learning detected anomalous sensor data with a false detection rate below 0.11% and an average F1-Score of 97.7% [34]. An F-score of 99.73% is reported in [35], where the authors propose an optimized auto-learning and class-detection network model.

As further work, we propose building a multi-agent system to automate the proposed tasks. The multi-agent system will include a monitoring agent that communicates with the other agents, makes decisions, and sends notifications to the end user.

Author Contributions: Conceptualization, M.V.M.; methodology, M.V.M.; software, M.V.M.; validation, M.V.M. and D.O.; formal analysis, M.V.M. and D.O.; investigation, M.V.M. and D.O.; resources, M.V.M. and

 doi:10.20944/preprints202605.0347.v1

D.O.; data curation, M.V.M. and D.O.; writing—original draft preparation, M.V.M. and D.O.; writing—review and editing, M.V.M. and D.O.; visualization, M.V.M. and D.O.; supervision M.V.M.; project administration, M.V.M.; funding acquisition, M.V.M. and D.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received financial support from the funds for scientific research of 1 Decembrie 1918 University of Alba Iulia, Romania.

Data Availability Statement: Data is available upon request.

Acknowledgments: We acknowledge Nicolae Roman (roman.nicolae.pabd25@uab.ro) for providing the simulated dataset for the current research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu, D.-M., Hu, X.-X., Wang, W.-C., Chau, K.-W., Zang, H.-F., & Wang, J. (2024). A new hybrid model for monthly runoff prediction using ELMAN neural network based on decomposition-integration structure with local error correction method. *Expert Systems with Applications*, 238(Part A), Article 121719. https://doi.org/10.1016/j.eswa.2023.121719
2. Park, J. S., Park, J. H., Choi, J. H., & Kwon, H. Y. (2026). Self-supervised inference of missing energy sources for enhancing energy consumption forecasting. *Applied Soft Computing*, 192, Article 114732. https://doi.org/10.1016/j.asoc.2026.114732
3. Ullah, B.; Kamran, M.; Rui, Y. Predictive Modeling of Short-Term Rockburst for the Stability of Subsurface Structures Using Machine Learning Approaches: t-SNE, K-Means Clustering and XGBoost. *Mathematics* 2022, 10, 449. https://doi.org/10.3390/math10030449
4. Lu, Z., Tao, R., Xiao, R., & Li, P. (2024). Forecasting the hydropower unit vibration based on adaptive variational mode decomposition and neural network. *Applied Soft Computing*, 150, 111040. https://doi.org/10.1016/j.asoc.2023.111040
5. Castro-Freibott, R.; García-Sánchez, Á.; Espiga-Fernández, F.; González-Santander de la Cruz, G. Deep Reinforcement Learning for Intraday Multireservoir Hydropower Management. *Mathematics* 2025, 13, 151. https://doi.org/10.3390/math13010151
6. Wu, X.; Guo, J.; Shan, Y.; Jin, G. GeoFormer: Geography-Aware Adaptive Transformer with Multi-Scale Temporal Fusion for Global Reservoir Water Level Forecasting. *Mathematics* 2026, 14, 676. https://doi.org/10.3390/math14040676
7. Chong, K. L., Lai, S. H., Ahmed, A. N., Jaafar, W. Z. W., & El-Shafie, A. (2021). Optimization of hydropower reservoir operation based on hedging policy using Jaya algorithm. *Applied Soft Computing*, 106, Article 107325. https://doi.org/10.1016/j.asoc.2021.107325
8. Li, S., Ahmadianfar, I., Farooque, A. A., & Yaseen, Z. M. (2025). An efficient mathematical-based optimization method to optimize multi-hydropower operating rules. *Applied Soft Computing*, *185*(Part A), Article 113846. https://doi.org/10.1016/j.asoc.2025.113846
9. Zhou, C.; Shen, Z.; Xu, L.; Sun, Y.; Zhang, W.; Zhang, H.; Peng, J. Global Sensitivity Analysis Method for Embankment Dam Slope Stability Considering Seepage–Stress Coupling under Changing Reservoir Water Levels. *Mathematics* 2023, 11, 2836. https://doi.org/10.3390/math11132836
10. Pan, H.; Yang, J.; Yu, Y.; Zheng, Y.; Zheng, X.; Hang, C. Intelligent Low-Consumption Optimization Strategies: Economic Operation of Hydropower Stations Based on Improved LSTM and Random Forest Machine Learning Algorithm. *Mathematics* 2024, 12, 1292. https://doi.org/10.3390/math12091292
11. Xu, B., Zhang, H., Su, H., & Chen, Z. (2026). Fusion of feature selection, feature extraction, and deep learning for displacement prediction of arch dams with cracks. *Applied Soft Computing*, 190, 114566. https://doi.org/10.1016/j.asoc.2026.114566
12. Lu, Y., & Wu, Z. (2024). A vine-copulas based multi-sensor fusion structural damage monitoring method and its application in dam engineering. *Applied Soft Computing*, 167(Part B), 112356. https://doi.org/10.1016/j.asoc.2024.112356

 doi:10.20944/preprints202605.0347.v1

13. Yi, T., Zhao, X., Shi, Y., Jing, X., Lei, W., Guo, J., Meng, Y., & Zhang, Z. (2025). Anomaly detection method for hydropower units based on KSQDC-ADEAD under complex operating conditions. *Sensors*, 25(13), 4093. https://doi.org/10.3390/s25134093
14. Zhang, F., Guo, J., Yuan, F., Qiu, Y., Wang, P., Cheng, F., & Gu, Y. (2024). Enhancement methods of hydropower unit monitoring data quality based on the hierarchical density-based spatial clustering of applications with a noise–Wasserstein slim generative adversarial imputation network with a gradient penalty. *Sensors*, 24(1), 118. https://doi.org/10.3390/s24010118
15. Rathnayake, N., Rathnayake, U., Dang, T. L., & Hoshino, Y. (2022). A cascaded adaptive network-based fuzzy inference system for hydropower forecasting. Sensors, 22(8), 2905. https://doi.org/10.3390/s22082905
16. Kou, Y., & Lin, Z. (2025). Research on the joint prediction model of lake water level and water resource usage trend based on deep learning. *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence*, 1329–1336. https://dl.acm.org/doi/10.1145/3773365.3773575
17. Li, M., Zhang, B., Hu, L., Ning, Z., Chen, L., & Zhang, M. (2025). Application of artificial intelligence in hydropower operation automation. *2025 25th Conference of the Electric Power Supply Industry (CEPSI)*, 899–904. https://doi.org/10.1109/CEPSI66359.2025.11403280
18. Wu, Z., Hu, K., Dong, Y., Guo, P., & Zhang, F. (2025). Multi-dimensional risk prevention and emergency management for hydropower and pumped storage projects: A framework and review. *2025 IEEE 9th Conference on Energy Internet and Energy System Integration (EI2)*, 1944-1950. https://doi.org/10.1109/EI268505.2025.11425262
19. Chen, Q., Ran, F., Wen, Z., & Wang, H. (2025). Intelligent inspection method of hydropower station based on artificial intelligence and internet of things. *2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, 573–576. https://doi.org/10.1109/AIoTC66747.2025.11198799
20. Wang, Y., Cheng, W., & Li, S. (2025). Design of water supply plant parameter monitoring system based on Internet of Things. *Proceedings of the 9th International Conference on Electronic Information Technology and Computer Engineering*, 74–79. https://doi.org/10.1145/3766671.3766685
21. Chen, J., Li, A., Wang, K., Yan, N., Liu, Q., & Ling, S. (2025). Design and implementation of hydropower and new energy big data platform based on cloud-native technology. *Proceedings of the 2025 11th Annual International Conference on Network and Information Systems for Computers*, 154–162. https://doi.org/10.1145/3776942.3776995
22. Yue, Y., Xia, S., & Sun, Q. (2026). Inversion analysis of thermodynamic parameters of concrete dam based on intelligent algorithm. *Proceedings of the 2025 International Conference on Artificial Intelligence and Sustainable Development*, 405–410. https://doi.org/10.1145/3786484.3786546
23. Ouyang, B., Shi, R., Zhang, F., & Cai, Q. (2025). Research on intelligent fault diagnosis method of multi-source monitoring data of hydropower units based on time series transformer. *2025 5th International Conference on New Energy and Power Engineering (ICNEPE)*, 835–840. https://doi.org/10.1109/ICNEPE67923.2025.11384173
24. Video Lake Oașa, the “pearl” on the Transalpina. Sadoveanu's corner of heaven, hidden in the heart of the Șureanu Mountains. https://cluj-napoca.xyz/video-lake-oasa-the-pearl-on-the-transalpina-sadoveanus-corner-of-heaven-hidden-in-the-heart-of-the-sureanu-mountains/
25. Hidroelectrica. https://www.hidroelectrica.ro/article/42
26. Grigoras, G., Gârbea, R., & Neagu, B.-C. (2024). Toward smart SCADA systems in the hydropower plants through integrating data mining-based knowledge discovery modules. *Applied Sciences*, 14(18), 8228. https://doi.org/10.3390/app14188228
27. Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA workbench (4th ed.). Online appendix for Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.
28. University of Waikato. (n.d.). Weka 3: Data mining software in Java. https://ml.cms.waikato.ac.nz/weka/
29. Yang, R., Tian, Z., Hu, D., Jiang, X., & Dong, M. (2024, August 4–7). The development of information management system for live detection data in hydro power station. In Proceedings of the 2024 IEEE 14th International Conference on the Properties and Applications of Dielectric Materials (ICPADM) (pp. 33–36). IEEE. https://doi.org/10.1109/ICPADM61663.2024.10750721

 doi:10.20944/preprints202605.0347.v1

30. Guo, S., Huang, J., Yan, Y., Zhang, P., Wang, B., Shen, H., & Yuan, Z. (2025). Secure indoor water level monitoring with temporal super-resolution and enhanced Yolov5. Sensors, 25(9), 2835. https://doi.org/10.3390/s25092835
31. Pawar, M., Ranjan, N., Patil, V., Naik, S., Shetty, R., & Bhoyar, P. (2025). Real-time detection and extraction of invasive aquatic weeds using YOLO and autonomous robotics: A sustainable solution. *2025 IEEE 6th India Council International Subsections Conference (INDISCON)*, 1–6. https://doi.org/10.1109/INDISCON66021.2025.11251619
32. Li, R., & Zheng, D. (2025). Research on data mining and analysis of hydropower station monitoring based on intelligent system. *2025 4th International Conference on Energy and Electrical Power Systems (ICEEPS)*, 65–70. https://doi.org/10.1109/ICEEPS66790.2025.11239750
33. Jian, X., & Chen, J. (2025). Real-time monitoring and early warning system for hydropower safety based on the Internet of things. *Intelligent Decision Technologies*, 19(4), 2295–2309. https://doi.org/10.1177/18724981251332530
34. Tian, J., Chen, J., Wang, J., Huang, H., & Li, Y. (2026). Real-time classification model for anomalous sensor data in dam safety monitoring based on convolutional neural networks and transfer learning. *Expert Systems with Applications*, 296(Part D), Article 129135. https://doi.org/10.1016/j.eswa.2025.129135
35. Lee, J., Kim, K., & Sohn, H. (2023). The unknown abnormal condition monitoring method for pumped-storage hydroelectricity. Sensors, 23(14), 6336. https://doi.org/10.3390/s23146336