Research Article

Firas Mohammed Aswad, Ali Noori Kareem, Ahmed Mahmood Khudhur, Bashar Ahmed Khalaf, and Salama A. Mostafa*

Tree-based machine learning algorithms in the Internet of Things environment for multivariate flood status prediction

https://doi.org/10.1515/jisys-2021-0179
received August 09, 2021; accepted October 06, 2021

Abstract: Floods are one of the most common natural disasters in the world that affect all aspects of life, including human beings, agriculture, industry, and education. Research for developing models of flood predictions has been ongoing for the past few years. These models are proposed and built-in proportion for risk reduction, policy proposition, loss of human lives, and property damages associated with floods. However, flood status prediction is a complex process and demands extensive analyses on the factors leading to the occurrence of flooding. Consequently, this research proposes an Internet of Things-based flood status prediction (IoT-FSP) model that is used to facilitate the prediction of the rivers flood situation. The IoT-FSP model applies the Internet of Things architecture to facilitate the flood data acquisition process and three machine learning (ML) algorithms, which are Decision Tree (DT), Decision Jungle, and Random Forest, for the flood prediction process. The IoT-FSP model is implemented in MATLAB and Simulink as development platforms. The results show that the IoT-FSP model successfully performs the data acquisition and prediction tasks and achieves an average accuracy of 85.72% for the three-fold cross-validation results. The research finding shows that the DT scores the highest accuracy of 93.22%, precision of 92.85, and recall of 92.81 among the three ML algorithms. The ability of the ML algorithm to handle multivariate outputs of 13 different flood textual statuses provides the means of manifesting explainable artificial intelligence and enables the IoT-FSP model to act as an early warning and flood monitoring system.

Keywords: flood prediction, Internet of Things, multivariate classification, machine learning, explainable artificial intelligence

1 Introduction

Natural disasters have caused a lot of damages to mankind, causing huge material and moral losses that affected the lives of 200 million people and affected the economy, with a loss of about $95 billion. Its impact also included other life aspects, including transportation, electricity, water, and ecosystems [1]. The main

* **Corresponding author: Salama A. Mostafa,** Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat 86400, Johor, Malaysia, e-mail: salama@uthm.edu.my
Firas Mohammed Aswad: Computer Department, College of Basic Education, University of Diyala, 32001, Diyala, Iraq, e-mail: drfirasaswad@gmail.com
Ali Noori Kareem: Computer Engineering Department, Bilad Alrafidain University College, 32001, Diyala, Iraq, e-mail: Dr.alinoori@bauc14.edu.iq
Ahmed Mahmood Khudhur: Computer Engineering Department, Bilad Alrafidain University College, 32001, Diyala, Iraq, e-mail: Dr.ahmedm@bauc14.edu.iq
Bashar Ahmed Khalaf: Department of Medical Instruments Engineering Techniques, Bilad Alrafidain University College, 32001, Diyala, Iraq, e-mail: bashar@baus14.edu.iq

factors that cause flood are intense or extreme rainfall events [2], hurricanes, and sewage blockages. In addition, other factors related to humans, such as land-use changes, urbanization, and mineral resource exploitation [3]. The risk of natural disasters increases, especially with the rapid growth in urban areas, where there is an increase in the density of human structures, which causes a lack of efficient water resources management [4], sanitation networks, and poor management of solid waste. This may result in health problems, floods, and landslides. According to ref. [5], the percentage of human losses in the Asian continent as a result of natural disasters is about 90%, which is often caused by floods. Due to these issues, floods and mitigating the damage they cause are essential and important strategies to consider [2,6]. One of the main solutions in managing flood disasters and mitigating their future severity is the identification of flood and torrential risk areas using effective and highly accurate methods [7]. Hydrological models are used to determine areas at risk of flooding; hence, forecasting the severity of the floods and assessing the anthropogenic mitigation measures will be required in the future [8].

Studies on rainfall and floods help establish the correct procedures in the natural disaster alert and response systems and improve preparedness to face these situations [4]. Due to the number of complex variables related to floods, it is important to apply the Internet of Things (IoT) technology along with the data mining techniques to assess flood sensitivity accurately [9]. The IoT technology provides an advanced data acquisition infrastructure that satisfies the needs of distributed and dynamic environments of the flooding areas. Data mining and machine learning (ML) techniques are effective tools to investigate and develop various models for floods prediction. Data mining techniques can be used to illustrate the mechanism between specific events and related variables [10]. Several research and studies have been performed in the field of floods, specifically the creation of models that predict floods occurring in different regions of the world [11,12]. In general, supervised ML or statistical data mining methods have recently been used for flood predictions such as Logistic Regression (LR) [13], Artificial Neural Networks (ANNs) [14,15], Naïve Bayes (NB) [16], Random Forest (RF) [17], Support Vector Machine (SVM) [18], Decision Trees (DT) [2], and Decision Jungle (DJ). Among these methods, DT is a good and effective method for mapping susceptibility to floods and has proven effective with the high predictive performance of floods [19]. SVM is another effective tool for a range of hydrological modeling applications across many continents [20]. However, in explainable artificial intelligence, there is a need to propose a solution model that can produce results that are understood by humans. This issue can be resolved by enabling the supervised ML algorithm to predict a bigger set of class labels that have contextual form.

This article has presented the various methods, techniques, and models that are used to achieve this work. It provides an extensive overview of the various machine learning and data mining approaches in the flood predictions field. Subsequently, this article contributes to the following:

1. An Internet of Things-based flood status prediction (IoT-FSP) model that facilitates the prediction of the river's flood situation based on three ML algorithms of DT, DJ, and RF.
2. A complete design of flood alert system that has the feature of explainable artificial intelligence.
3. An evaluation of three different ML algorithms of DT, DJ, and RF in terms of accuracy, precision, and recall of flood prediction.

The structure of the article is arranged in six sections as follows. Section 2 reviews all works related to flood data acquisition and predictions. Section 3 presents the research framework, methods used to perform the data mining task, along the dataset and the evaluation metrics. Section 4 illustrates the main components of the IoT-FSP model. Section 5 presents the results and discusses the outcomes, and Section 6 states the concluding remarks and future work.

2 Related works

The related works are divided into two sections based on the research scope. The first section focuses on flood data acquisition using IoT technology. The second section focuses on flood status prediction using ML. The

review aims to identify the state-of-the-art models and methods that are used to solve the problems of multivariate classification for pre-disaster flood early warning, as presented in the last section of this review.

2.1 Flood data acquisition

The IoT technology has been recently implemented in a wide range of projects as a smart solution for data collection and processing, especially in dynamic and complex systems. Ghapar et al. [9] propose the use of IoT architecture for flood data management. They suggest the usage of different sensors for collecting flood-related data, such as sensors for measuring hydrological, geological, and meteorological data. The project confirms the ability of the IoT architecture to facilitate data collection, transmission, and management. However, this architecture remains conceptual and is never implemented.

Noymanee et al. [21] propose a conceptual framework based on an IoT platform for flood early-warning systems in an urban environment. A context-aware module supports the framework to characterize the flood conditions. The IoT architecture for the context-aware system consists of five layers. These layers are application, storage management, processing or reasoning, raw data retrieval, and sensors. They intend to make the framework to predict the flood situations from observations without using an explicit prediction method. However, this framework also remains conceptual and is never implemented. Similarly, Chen and Chen [22] propose the use of a context-oriented IoT platform for data acquisition and integration. The integration process entails converting the raw data to a semantic context for easy storage, understanding, and sharing.

Balakrishna et al. [23] propose a sensor data acquisition and analysis framework based on an IoT platform. The framework is implemented in a traffic monitoring system by using the ThingSpeak IoT Cloud platform that provides data analysis and visualization services. A median filter is used to improve the context of the data. The data processing is performed by the Gaussian mixture model and context construction method. The test results show that the framework achieves an accuracy of 84.56% on average for traffic condition description.

Fang et al. [24] propose an IoT-based integrated information system for snowmelt flood early warning. The IoT architecture is used for data acquisition, sharing, and management of multi-source information. The integrated information system has been developed as a web application. The test case study is the Quergou River Basin, which is located in Hutubi County, Xinjiang, China. The resulted warning recommendations are compared between the actual and presented water levels.

2.2 Flood status prediction

Various research articles investigate the use of ML algorithms in the prediction of river flooding situations, and some of them are presented in this section. Chen et al. [2] use three techniques for spatial forecasting of floods in the Quannan region of China. They conduct an evaluation and comparison of these techniques, the Naive Bayes Tree (NBTree), the Alternative Decision Tree (ADTree), and RF methods, to determine their ability to predict floods. Their approach is based on producing a flood inventory with 363 flood sites and dividing them into training and verification datasets with a 70/30 random selection. Thirteen factors are used to create the spatial flood database to explain and understand the floods. Their results show that RF is an effective and reliable model for assessing flood vulnerability.

Hong et al. [6] examine four DT-based ML models, namely Logistic Model Trees (LMT), Reduced Error Pruning Trees (REPT), NBT, and Alternating Decision Trees (ADT) for flash flood susceptibility mapping in Iran. They construct a spatial database with 201 present and past flood locations and 11 flood-influencing factors. The capability of these models for flood predictions is evaluated and compared using statistical evaluation measures, the receiver operating characteristic curve, and Freidman and Wilcoxon signed-rank

tests. The results show that the ADT model has the highest prediction capability for flash flood susceptibility assessment, followed by the NBT, the LMT, and the REPT, respectively. These strategies have proven to be effective in the rapid determination of flood-prone areas.

Hong et al. [6] introduce a new approach to building a flood susceptibility map in China by applying fuzzy proof weight (fuzzy-WofE) and data-mining methods. The thing that distinguishes the proposed approach is the use of fuzzy-WofE, which creates a preliminary flood sensitivity map and determines the variables associated with floods. LR, RF, and SVMs are implemented, taking into account the 11 flood-related variables. They evaluate the efficiency of their approach using the area under the curve (AUC). Their results show that the fuzzy WofE-SVM model produces the highest predictive performance (AUC value, 0.9865), which also appears to yield statistically significant differences from the other predictive models.

Tehrany et al. [19] examine and validate the hypothesis that the accuracy of the final susceptibility mapping result improves by adding more conditioning variables to the dataset used in river flood modeling. In addition, this research assesses the effect of individual conditioning influences on flood susceptibility mapping and their significance in the construction of accurate mapping of possible flood regions. They use DT and SVM to test spatial correlations between flood conditioning factors and rate their degree of importance for flood-prone mapping. They assess the accuracy for two ML approaches, SVM and DT, using the AUC method. The results show that SVM and DT provide the highest predictive accuracy levels of 85.52 and 88.47%, respectively, using DS1 (LiDAR dataset). Finally, it is concluded that the use of additional variables in the simulation does not necessitate the achievement of higher accuracy. Suliman et al. [20] review the related works of flood forecasting. The review focuses on the two most popular ML algorithms, which are ANN and SVM.

Choubin et al. [25] use two new algorithms, namely Multivariate Discriminant Analysis (MDA) and Classification and Regression Trees (CART) combined with the SVM algorithm in flood susceptibility analysis. They use these models with a flood inventory map and many factors of flood conditioning to develop a flood susceptibility map. A new framework for flood susceptibility assessment is proposed to ensure a more accurate ensemble model where only those models with an accuracy of 80% are permissible for use in ensemble modeling. The results show that the MDA model produces the highest predictive accuracy (89%), followed by the SVM (88%) and CART (83%) models. The ensemble modeling approach indicates that areas with a high population density are more vulnerable to floods, and therefore these areas should be given priority to flood prevention and treatment.

Liu et al. [26] combine Stacked Autoencoders (SAE) and Backpropagation Neural Networks (BPNN) to implement a new deep learning approach for floods prediction. To further develop the ability to model nonlinearity, their architecture performs two processes. First, K-means clustering is applied for data classification into various categories. Then, they represent their related data categories by using multiple SAE-BP modules. The results of the comparison between their approach and other approaches, which are SVM, BPNN, Radial-based Functions, and Extreme Learning Machine, show that their approach performs much better.

Widiasari et al. [27] provide a definition of the main model of the ANN that is useful in time series forecasting and a basic procedure for the practical implementation of the ANN in this form of mission. The model analyzed is the Multilayer Perceptron (MLP). To assess the degree of precision of the flood prediction, Mean Absolute Percentage Error (MAPE) is used in which the system predicts the lower the MAPE value, the more accurate the results. By using this, MLP achieves a MAPE value of 3.64%, which means that the error caused by the built-in device is 3.64% compared to the actual value used for testing. Also, MLP has a greater effect on the expected water level than multiple linear regression.

Widiasari et al. [28] use the Long Short-Term Memory (LSTM) algorithm, which is common and powerful at managing long-run periods of temporal dependencies for complex time-series data like precipitation and water elevation degree that can be sensed by using sensors. The model produces a MAPE value of 3.6% through the LSTM algorithm, which indicates that the error that is produced to predict the water level within the downstream river is 3.6% compared with the real water elevation value. LSTM produces more correct predictions of water elevation level inside the downstream if compared to MLR models that produce a MAPE prediction value of 10.55%.

From the previous contributions, we observe that ML algorithms are mainly used to predict floods. In Chen et al. [2], the RF model achieves the highest accuracy of 91.5% in predicting the flood status of five classes. The work of Khosravi et al. [5] achieves the highest accuracy of 94.3% in predicting the flood status of four classes through the ADT model. The work of Hong et al. [6] achieves the highest accuracy of 92.2% in predicting the flood status of five classes through the Fuzzy WofE-SVM model. However, none of these models introduces a complete architectural design of the flood alert system. They only focus on evaluating the ML ability to predict flood status from a maximum of five classes and neglect flood data acquisition.

3 Materials and methods

3.1 Anglian river basin district (RBD) dataset

The dataset that is obtained from a specific source, such as the Internet, might be incomplete and inconsistent. Hence, selecting appropriate data is an important research issue [29,30]. The Anglian RBD includes 27,900 km^2 in which a total of 7.1 million people settled in this area. It includes the cities of Northampton, Lincoln, Chelmsford, and Milton Keynes as shown in Figure 1. The RBD dataset that is used in the research for flood prediction is an open dataset taken from the environmental agency [31]. The dataset includes a collection of datasets related to flood measurement, classification, environmental effect, and protection.

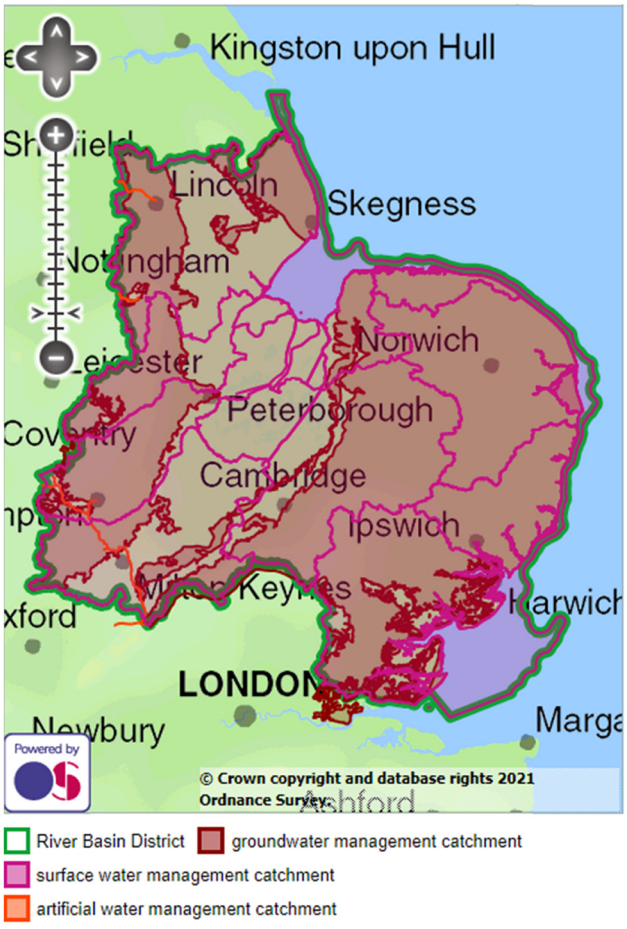

Figure 1: The RBD map [31].

The last update of the RBD dataset is on September 17, 2020. This dataset has 19 attributes and 149,676 instances, and the multivariate target classes of the dataset are 13 river states, as displayed in Table 1.

Table 1: Multivariate class labels of the RBD dataset

Class	Target river status
1	Active
2	Bad
3	Does not require assessment
4	Does not support good
5	Fail
6	Good
7	High
8	Moderate
9	Moderate or less
10	No trend
11	Poor
12	Supports good
13	Upward trend

3.2 ML algorithms

Data mining is a process of classification, audit, and semi-automatic analysis of very large amounts of data to obtain useful information and explore patterns and links [32]. Data mining is used in several different fields as a method in predicting data appropriately [33]. This section provides a summary of some data mining algorithms, which are DT, RF, and DJ algorithms.

DT is a structure of a branching tree that is used to determine the course of work or show the possibilities of a solution. Each internal node represents a test on an attribute, and each branch represents a potential DT [34]. Usually, an entropy function is deployed to control the DT splitting the data. The entropy affects the boundaries of a solution that the DT draws. The entropy formula is given in (1).

$$H(s) = -\text{sum}(\text{each } k \text{ in } K\ p(k)\star\ \log(p(k))). \tag{1}$$

The DT provides a transparent tree-like structure with effective, easy rules that are easily interpreted and understood [35]. DT is used in flood forecasting of high-water levels and water flow [36]. Compared with other methods of classification, DT can be built quickly [34]. It has also been widely used for both continuous and discrete datasets. Variables screening and features selection are good enough in DT [37]. Regarding its performance, nonlinearity does not impact any of the DT parameters [38]. Hence, it is good to solve problems with multiple alternatives or situations that are related to risks and uncertainties. The basic rule in building a DT is to find the best question for each branch of the tree so that these questions divide the data into two sections. The first section applies to the question, and the second section does not apply so that through a series of questions, the DT is built with its chain of branches. Although the DT is used for exploration and data preparation for statistical operations, it is also used more often to predict the values of other cases not found in the training group [39].

RF is considered as an ensemble learning algorithm. It is utilized mainly for classification and regression tasks [40]. The RF architecture is represented as a multitude of DTs in which the output of the RF is the class that has an average or a majority selection among the DTs [41]. The prediction of a new case x after the training phase by using the average function is represented in (2).

$$y = \frac{1}{N}\sum_{i=1}^{N} x_i(\grave{x}). \tag{2}$$

For many classification problems, RF usually outperforms the DT for linear and nonlinear problems with an efficient mapping of input into forecasted spaces [42]. The design of the RF allows it to overcome the main problem of the DT, which is its tendency of training overfitting [43]. However, the RF performance is highlighted and affected by the characteristics of the data [44]. RF is currently considered as a popular classifier and has been integrated into many applications due to its stable performance in a wide range of classification problems [45].

DJ algorithm is an ensemble learning approach for classification. The algorithm works by utilizing and building different selection trees and then voting on the maximum famous output magnificence [46]. The trees that have high prediction have a greater weight in the final decision of the ensemble. A large number of applications have been developed using decision forests and trees in data science, although these approaches have certain drawbacks, such as given a large amount of data, the number of nodes in DTs may grow exponentially in size [47]. The DJ differs from DTs by the directed acyclic graph (DAG) group that permits multiple paths from a root to the leaves, whereas traditional DTs allow only one path per node [48]. The DJ method has a strong contrast between the two algorithms as it combines two modern nodes, thereby improving the function and structure of the DAG.

3.3 Evaluation metrics

To compare the performance of each of the classifications, preprocessing is carried out based on all the values of the taken 19 attributes in which no feature selection has been performed. A comparative study of DT, DJ, and RF classification results is performed with a three-fold validation method for training and testing [49–51].

- Accuracy: It is the most commonly used metric to judge a model and is not a clear indicator of performance. The worse happens when classes are imbalanced.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}). \tag{3}$$

- Precision or positive predictive value: It is the ratio of correctly classified attack flows (TP) in front of all the classified flows (TP + FP).

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}). \tag{4}$$

- Recall or sensitivity: It is the ratio of correctly classified attack flows (TP), in front of all generated flows (TP + FN).

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}), \tag{5}$$

where TP, TN, FP, and FN have their usual meaning.

4 IoT-FSP model

The IoT-FSP model is proposed to provide suitable flood data acquisition and prediction mechanisms. The IoT-FSP model employs an IoT architecture to manage the data flow and processing from the source to the destination. It uses three tree-based ML algorithms of multivariate classification for analyzing the extracted data and performing flood status prediction. The basic IoT architecture consists of three core layers: application layer, network layer, and data acquisition or sensor layer. This architecture can play an important role in facilitating flood early detection; hence, it is selected to advance this research. Figure 2 shows a general view of the data acquisition of the IoT environment.

From the top to bottom of this architecture, a set of sensors and devices are used to collect data related to the RBD, operational and catchment management, waterbody, water level, and water quality assessment items. These data are transmitted to the network layer using the client–server local area network. Because

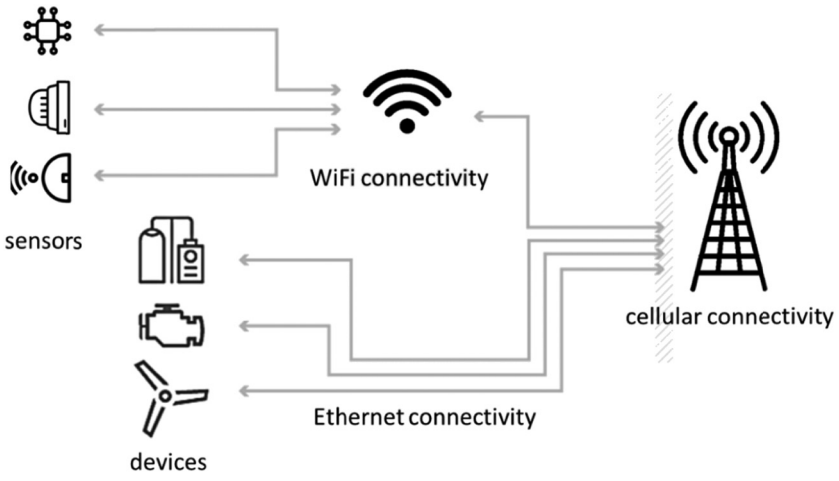

Figure 2: The data acquisition of the IoT environment.

this is an outdoor environment, cellular connectivity is applied with Wi-Fi and Ethernet as subnetworks. They are known as a gateway or edge computing that connects the sensors and devices with the cloud computing system, as shown in Figure 2. Subsequently, the network layer reads the required data and sends these data through the Wi-Fi and cellular connectors that allow access to the cloud services. Hence, the main task of the network layer connects the IoT acquisition layer with the application platform. The application platform engages with the collected data from the previous layers. In the real world, data are generally incomplete as they lack certain behaviors or trends, or contain only aggregated data without meaning, and they are likely to contain many errors. Data preprocessing is a proven method for solving such problems to obtain consistent, meaningful data that are understandably arranged. The application layer starts with the data preparation that includes three sub-phases: data selection, data preprocessing, and data partitioning. In this phase, the RBD dataset is retrieved from the database and get selected to be processed. Data transformation is the process of converting data from one format or structure into another format or structure. It is performed after the data selection process to ensure that the chosen data are complete and verified. Data partition forms as part of preprocessing. It helps to split the data using three-fold cross-validation so that the resulting mining process is more effective and the patterns found are robust. Figure 3 shows the IoT-FSP model based on the IoT architecture.

The data mining paradigm applies ML technology aims to extract knowledge from huge amount of data as it works to convert the incomprehensible raw data into data that people can read and understand. Finally, the proposed ML algorithms, which are DT, DJ, and RF, are applied to the datasets using MATLAB ML Toolbox. It is a platform that facilitates the work with ML algorithms. It has a complete reference and module to help experiments and scoring workflow. ML algorithms require parameter setting phase, and the parameters are set based on the need of each algorithm to provide the best performance. The fused results from the ML identify the river flood initial status (flood or no flood). Then, it determines the target 13 river final status. When the river status implies flood, then the model triggers an alarm and provides notification through the user interface regarding the flood status conditions.

5 Results and discussion

The implementation of the IoT-FSP model considers MATLAB and Simulink as development platforms in which MATLAB is used for the soft computing process and Simulink is used to stimulate the online data transmission process through the cloud computing services. The implemented ML algorithms to predict the flood status are DT, DJ, and RF algorithms. The algorithms are tested using the three-fold cross-validation

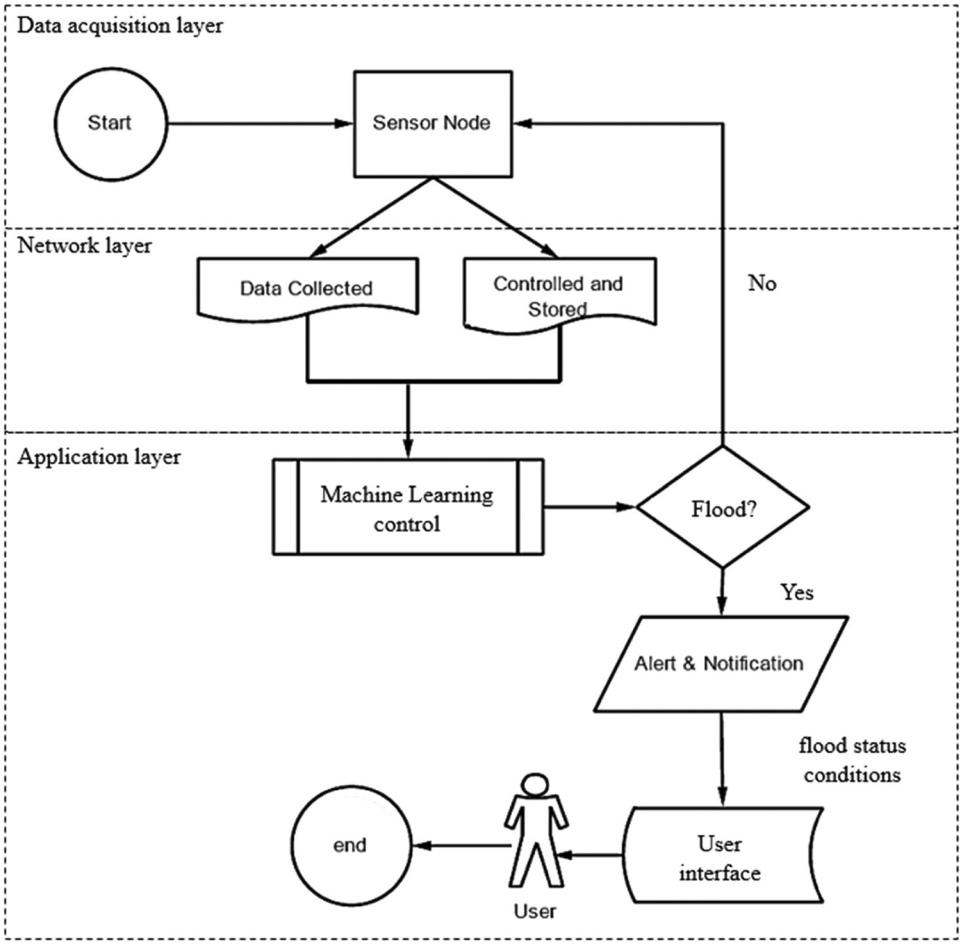

Figure 3: The IoT-FSP model based on the IoT architecture.

approach to check the ability of the algorithms in predicting the 13 flood status. Every algorithm has its separate process at the data simulation phase to perform data splitting, data preprocessing, and training. In the test phase, the ML algorithms of the IoT-FSP model are trained after the training parameters have been set correctly. The test set is used to calculate and evaluate the performance of every algorithm's prediction. Subsequently, verified results on the prediction of tested data are demonstrated by those three algorithms.

Finally, the results and analysis phase consist of testing and evaluating the performance of the model of the selected algorithms. During the evaluation, the results from the testing phase are evaluated by comparing the results of the algorithms to see which algorithm produces lower error rates and higher accuracy. This phase then shows which is the most efficient algorithm to get the best flood status results of accuracy, precision, and recall. Table 2 shows the results of the three tested algorithms of the IoT-FSP model in terms of accuracy, precision, and recall.

The results in Table 2 shows that the IoT-FSP model successfully performs the data acquisition and prediction tasks. The three algorithms achieve an average accuracy of 85.72% for the three-fold cross-validation results. The statistical analysis of the accuracy results shows that there is a small variance between the three-fold cross-validation results of 0.004409 for the DT, 0.0334622 for the DJ, and 0.0018556 for the RF. The highest accuracy result is achieved by the DT, which is 93.22%, followed by the RF that achieves an accuracy of 86.57% and the DJ achieves the lowest accuracy of 77.38%. Moreover, the DT has the lowest time complexity followed by the RF, and the DJ has the highest time complexity among the three [52]. Figure 4 shows the three-fold test results of the three ML algorithms.

Table 2: The results of the IoT-FSP model

Algorithm	Data split	Accuracy		Precision		Recall	
		Overall	Avg.	Overall	Avg.	Overall	Avg.
DT	40:60	0.9325	0.9473	0.9245	0.9260	0.9261	0.9126
	50:50	0.9276	0.9432	0.9317	0.9207	0.9286	0.9041
	60:40	0.9364	0.9446	0.9295	0.9454	0.9297	0.9211
DJ	40:60	0.7424	0.8501	0.7376	0.7223	0.7374	0.6532
	50:50	0.7700	0.8970	0.7659	0.7546	0.7733	0.7303
	60:40	0.8090	0.8656	0.8079	0.8043	0.7960	0.7915
RF	40:60	0.8675	0.9299	0.8686	0.8524	0.8630	0.7856
	50:50	0.8638	0.9364	0.8615	0.8549	0.8674	0.8191
	60:40	0.8659	0.9306	0.8577	0.8503	0.8693	0.8209

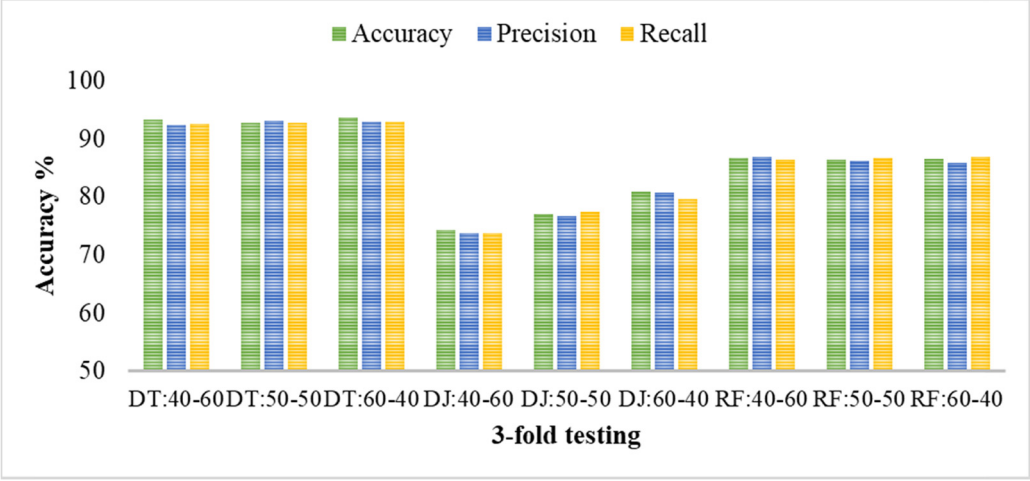

Figure 4: The visualization of the IoT-FSP model results.

Table 3: Comparison with the related work

Ref.	Method/Model	Description
[2]	ML	For forecasting of floods in the Quannan region of China. Multivariate classification of five classes. RF model achieves the highest accuracy of 91.5%
[5]	ML	For flash flood susceptibility mapping in Iran. Multivariate classification of four classes. ADT model achieves the highest accuracy of 94.3%
[6]	ML, fuzzy logic	For flood susceptibility map in China. Multivariate classification of five classes. Fuzzy WofE-SVM model achieves the highest accuracy of 92.2%
[9]	IoT architecture	For flood data management. It is never implemented
[19]	ML	For assessing the effect of individual conditioning influences on flood susceptibility mapping. Multivariate classification of three classes. LR model achieves the highest accuracy of 90.6%
[21]	IoT, context-aware	For flood early-warning system, only conceptual. It is never implemented
[22]	IoT, context-oriented	For wireless sensor networks to be used in ambient systems. It is never implemented
[23]	IoT, Gaussian mixture	For traffic flow analysis, implemented ThingSpeak IoT Cloud platform. It is never implemented
[24]	IoT	For snowmelt flood early warning. It is only used for data acquisition
IoT-FSP	ML, IoT	For forecasting of floods of RBD dataset. Multivariate classification of 13 classes. DT model achieves the highest accuracy of 93.22%

The literature review presents several research examples that are conducted to handle flood disasters, including flood data acquisition, data visualization, flood prediction, flood detection, early warning, and flood monitoring. The scope of this article covers the pre-disaster phase for early warning comprising flood data acquisition using IoT technology and multivariate classification for flood status prediction using ML. The related works in flood research are summarized in comparison to our work in Table 3.

Based on Table 3, the IoT-FSP model implementation-wise outperforms its alternative models, predicting complex multivariate of 13 classes and achieving higher prediction accuracy. Moreover, the model utilizes both IoT and ML technologies in its design to form a flood alert system that has the feature of explainable artificial intelligence. The limitation of this model is that it needs to be tested in a real-world multivariate flood status prediction environment.

6 Conclusion

Flooding is a dangerous event whose risks should be controlled properly. It is mainly attributed to heavy rain, melting snow, or events emerging frequently as a consequence of climate change. The impacts of the floods are not limited to individuals only but extend and encompass whole societies, affecting many aspects, most notably the economic, environmental, and social aspects. However, the negative aspects of a flood differ in its intensity and impact depending on several natural and contingency factors. Therefore, there are many special measures to mitigate its impact, including early flood detection, warning, and monitoring systems. These systems should be implemented in the wider management of floodplains. Flood prediction or forecasting is made to ensure that the risk of flooding is reduced. The variables and methodology used in this study apply to flood susceptibility mapping for the different study areas. Subsequently, this article proposes an IoT-FSP model that facilitates the prediction of the river's flood situation.

The IoT-FSP model encompasses an IoT architecture for data collection and three different ML algorithms of DT, DJ, and RF for flood prediction. The performance of each ML algorithm is assessed and compared in terms of accuracy, precision, and recall. The experimental RBD dataset is used in this project to evaluate the performance of the IoT-FSP model and determine the best ML algorithm for flood prediction. The results of the three-fold cross-validation method show that the DT achieves the highest accuracy score of 93.22%, precision score of 92.85, and recall score of 92.81 among the three ML algorithms. The DT algorithm is found to be better than RF and DJ and is more reliable in predicting the flood status. The main contribution of this article is proposing a model that is able to handle multivariate flood statuses, providing the means of manifesting explainable artificial intelligence and enabling it to act as an early warning and flood monitoring system.

The future work will closely study the IoT aspects of the project, including managing data of heterogeneous sensors, energy consumptions, and efficiency of performance. Additionally, integrating contextual description modules in the IoT-FSP model to improve the explainable artificial intelligence ability of the model in this domain shall be implemented to be represented as a report form.

Acknowledgments: The authors would like to thank the Center of Intelligent and Autonomous Systems (CIAS), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this work. Also, the authors would like to thank the College of Basic Education, University of Diyala for supporting this work.

Funding information: Communication of this research is made possible through monetary assistance by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

Conflict of interest: The authors have no conflicts of interest to declare. The authors certify that the submission is an original work and is not under review at any other publication. All authors have seen and agree with the contents of the manuscript, and there is no financial interest to report.

Data availability statement: The used dataset of this research is available online and has a proper citation within the article's contents.

References

[1] Chapi K, Singh VP, Shirzadi A, Shahabi H, Bui DT, Pham BT, et al. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ Model Softw. 2017;95:229–45.

[2] Chen W, Li Y, Xue W, Shahabi H, Li S, Hong H. Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Sci Total Env. 2020;701:134979.

[3] Guo C, Qin Y, Ma D, Xia Y, Chen Y, Si Q, et al. Ionic composition, geological signature and environmental impacts of coalbed methane produced water in China. Energy Sources A Recov Utilization Environ Eff. 2021;43(10):1259–73.

[4] Bangera CS, Kotian PS, Dias C, Divya T, Aithal G. Flood and heat wave prediction using weighted moving average, anomaly detection and K-nearest neighbours for the city of Mangalore. In 2018 IEEE distributed computing, VLSI, electrical circuits and robotics (DISCOVER). IEEE; 2018. p. 93–7. doi: 10.1109/discover.2018.8674113.

[5] Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, et al. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci Total Environ. 2018;627:744–55.

[6] Hong H, Tsangaratos P, Ilia I, Liu J, Zhu A-X, Chen W. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. Sci Total Environ. 2018;625:575–88.

[7] Shaaban NN, Hassan N, Mustapha A, Mostafa SA. Comparative performance of supervised learning algorithms for flood prediction in Kemaman, Terengganu. J Computer Sci. 2021;17(5):451–8.

[8] Lundin LC, Bergstrom S, Eriksson E, Seibert J. Hydrological models and modeling. 2015. Retrieved from: [online] http://www.balticuniv.uu.se/index.php/component/docman/doc_download/270-the-waterscape-11hydrological-models-and-modelling [27, June, 2015].

[9] Ghapar AA, Yussof S, Bakar AA. Internet of Things (IoT) architecture for flood data management. Int J Future Gener Commun Netw. 2018;11(1):55–62.

[10] Lee S, Lee MJ, Jung HS. Data mining approaches for landslide susceptibility mapping in Umyeonsan, Seoul, South Korea. Appl Sci. 2017;7:683.

[11] Hong H, Panahi M, Shirzadi A, Ma T, Liu J, Zhu AX, et al. Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. Sci Total Environ. 2018;621:1124–41. doi: 10.1016/j.scitotenv.2017.10.114.

[12] Khosravi K, Nohani E, Maroufinia E, Pourghasemi HR. A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decisionmaking technique. Nat Hazards. 2016;83(2):947–87.

[13] Tien Bui D, Khosravi K, Shahabi H, Daggupati P, Adamowski JF, Melesse AM, et al. Flood spatial modeling in northern Iran using remote sensing and GIS: a comparison between evidential belief functions and its ensemble with a multivariate logistic regression model. Remote Sens. 2019;11(13):1589.

[14] Zhao G, Pang B, Xu Z, Yue J, Tu T. Mapping flood susceptibility in mountainous areas on a national scale in China. Sci Total Environ. 2018;615:1133–42.

[15] Sulaiman J, Wahab SH. Heavy rainfall forecasting model using artificial neural network for flood prone area. Lecture Notes Electr Eng. 2017;68–76. doi: 10.1007/978-981-10-6451-7_9.

[16] Tang X, Li J, Liu M, Liu W, Hong H. Flood susceptibility assessment based on a novel random Naïve Bayes method: a comparison between different factor discretization methods. Catena. 2020;190:104536. doi: 10.1016/j.catena.2020.104536.

[17] Muñoz P, Orellana-Alvear J, Willems P, Célleri R. Flash-flood forecasting in an Andean mountain catchment – development of a step-wise methodology based on the random forest algorithm. Water. 2018;10(11):1519. doi: 10.3390/w10111519.

[18] Shi Y, Taalab K, Cheng T. Flood prediction using support vector machines (SVM). In Proceedings of the 24th GIS research UK (GISRUK) conference. London, UK: GIS Research UK (GISRUK); 2016.

[19] Tehrany MS, Pradhan B, Jebur MN. Spatial prediction of flood susceptible areas using rule-based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol. 2013;504:69–79.

[20] Suliman A, Nazri N, Othman M, Abdul M, Ku-Mahamud KR. Artificial neural network and support vector machine in flood forecasting: a review. In Proceedings of the 4th international conference on computing and informatics, ICOCI; 2013. p. 28–30.

[21] Noymanee J, San-Um W, Theeramunkong T. A conceptual framework for the design of an urban flood early-warning system using a context-awareness approach in internet-of-things platform. In information science and applications (ICISA) 2016. Singapore: Springer; 2016. p. 1295–305.

[22] Chen YS, Chen YR. Context-oriented data acquisition and integration platform for internet of things. In 2012 conference on technologies and applications of artificial intelligence. IEEE; 2012, November. p. 103–8.
[23] Balakrishna S, Thirumaran M, Solanki VK. A framework for IoT sensor data acquisition and analysis. EAI Endorsed Trans Internet Things. 2018;4(16):1–13.
[24] Fang S, Xu L, Zhu Y, Liu Y, Liu Z, Pei H, et al. An integrated information system for snowmelt flood early-warning based on internet of things. Inf Syst Front. 2015;17(2):321–35.
[25] Choubin B, Moradi E, Golshan M, Adamowski J, Sajedi-Hosseini F, Mosavi A. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci Total Environ. 2019;651:2087–96.
[26] Liu F, Xu F, Yang S. A flood forecasting model based on deep learning algorithm via integrating stacked autoencoders with BP neural network. In 2017 IEEE third international conference on multimedia big data (BigMM). IEEE; 2017. p. 58–61. doi: 10.1109/bigmm.2017.29.
[27] Widiasari IR, Nugroho LE, Widyawan. Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network-based hydrology time series data mining. In 2017 International conference on innovative and creative information technology (ICITech). IEEE; 2017. p. 1–5 doi: 10.1109/innocit.2017.8319150.
[28] Widiasari IR, Nugoho LE, Widyawan, Efendi R. Context-based hydrology time series data for a flood prediction model using LSTM. In 2018 5th International conference on information technology, computer, and electrical engineering (ICITACEE). Semarang, Indonesia: IEEE; 2018. p. 385–90. doi: 10.1109/icitacee.2018.8576900.
[29] Mostafa SA, Gunasekaran SS, Mustapha A, Mohammed MA, Abduallah WM. Modelling an adjustable autonomous multi-agent internet of things system for elderly smart home. In International conference on applied human factors and ergonomics. Cham: Springer; 2019, July. p. 301–11.
[30] Dali AD, Omar NA, Mustapha A. Data mining approach to herbs classification. Indonesian J Electr Eng Computer Sci. 2018;12:570–6. doi: 10.11591/ijeecs.v12.i2.pp570-576.
[31] Environment Agency-Catchment Data Explorer, updated: October 16 2019. environment.data.gov.uk, Retrieved from https://environment.data.gov.uk/catchment-planning/RiverBasinDistrict/5.
[32] Banu GR. A role of decision tree classification data mining technique in diagnosing thyroid disease. Int J Computer Sci Eng. 2016;4(11):111–5.
[33] Nafi SNMM, Mustapha A, Mostafa SA, Khaleefah SH, Razali MN. Experimenting two machine learning methods in classifying river water quality. In Applied Computing to Support Industry: Innovation and Technology. Cham, Ramadi: Springer; 2020.
[34] Sharma Himani, Kumar Sunil. A survey on decision tree algorithms of classification in data mining. Int J Sci Res (IJSR). 2016;5(4):2094–7.
[35] Tien Bui D, Ho T-C, Pradhan B, Pham B-T, Nhu V-H, Revhaug I. GIS-based modeling of rainfall-induced landslides using data mining based functional trees classifier with AdaBoost, bagging, and Multi Boost ensemble frameworks. Env Earth Sci. 2016;75(14):1–22.
[36] Han D, Cluckie I, Karbassioun D, Lawry J, Krauskopf B. River flow modelling using fuzzy decision trees. Water Resour Manag. 2002;16:431–45.
[37] Lior R. Data mining with decision trees: theory and applications. Hackensack, New Jersey, United States: World Scientific; 2014 September 3.
[38] Patel Harsh, Prajapati Purvi. Study and analysis of decision tree based classification algorithms. Int J Computer Sci Eng. 2018;6:74–8. doi: 10.26438/ijcse/v6i10.7478.
[39] Bramer Max. Principles of data mining. London: Springer; 2007. doi: 10.1007/978-1-84628-766-4.
[40] Maseer ZK, Yusof R, Bahaman N, Mostafa SA, Foozy CFM. Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset. IEEE Access. 2021;9:22351–70.
[41] Azizan AH, Mostafa SA, Mustapha A, Foozy CFM, Abd Wahab MH, Mohammed MA, et al. A machine learning approach for improving the performance of network intrusion detection systems. Ann Emerg Technol Comput (AETiC). 2021;5(5):201–8.
[42] Elhoseny M, Mohammed MA, Mostafa SA, Abdulkareem KH, Maashi MS, Garcia-Zapirain B, et al. A new multi-agent feature wrapper machine learning approach for heart disease diagnosis. Comput Mater Contin. 2021;67:51–71.
[43] Mostafa SA, Mustapha A, Khaleefah SH, Ahmad MS, Mohammed MA. Evaluating the performance of three classification methods in diagnosis of Parkinson's disease. In: Ghazali R, Deris M, Nawi N, Abawajy J, (eds.). SCDM 2018. AISC. Vol. 700, Cham: Springer; 2018. p. 43–52.
[44] Mostafa SA, Mustapha A, Mohammed MA, Ahmad MS, Mahmoud MA. A fuzzy logic control in adjustable autonomy of a multi-agent system for an automated elderly movement monitoring application. Int J Med Inf. 2018;112:173–84.
[45] Zulhilmi A, Mostafa SA, Khalaf BA, Mustapha A, Tenah SS. A comparison of three machine learning algorithms in the classification of network intrusion. In International conference on advances in cyber security. Singapore: Springer; 2020, December. p. 313–24.
[46] Shotton J, Sharp T, Kohli P, Nowozin S, Winn J, Criminisi A. Decision jungles: compact and rich models for classification. In Advances in neural information processing systems. Lake Tahoe Nevada: Curran Associates Inc; 2013. p. 234–42.
[47] Mosavi A, Ozturk P, Chau K. Flood prediction using machine learning models: literature review. Water. 2018;10(11):1536. doi: 10.3390/w10111536.

[48] Alam TM, Khan MMA, Iqbal MA, Wahab A, Mushtaq M. Cervical cancer prediction through different screening methods using data mining. Int J Adv Comp Sci Appl(ijacsa). 2019;10(2):1–9.
[49] Fadel AH, Hameed RS, Hasoon JN, Mostafa SA, Khalaf BA. A light-weight ESalsa20 Ciphering based on 1D logistic and chebyshev chaotic maps. Solid State Technol. 2020;63(1):704–17.
[50] Khalaf BA, Mostafa SA, Mustapha A, Mohammed MA, Mahmoud MA, Al-Rimy BAS, et al. An adaptive protection of flooding attacks model for complex network environments. Sec Commun Netw. 2021;2021:1–17.
[51] Babatunde OS, Ahmad AR, Mostafa SA. A smart network intrusion detection system based on network data analyzer and support vector machine. Int J Emerg Trends Eng Res. 2020;8(1):213–20.
[52] Juman ZAMS, Hoque MA. An efficient heuristic to obtain a better initial feasible solution to the transportation problem. Appl Soft Comput. 2015;34:813–26.