

A Flood Prediction System Developed Using Various Machine Learning Algorithms

Kruti Kunverji
Information Technology
K.J. Somaiya Institute of Engineering and
Information Technology, Sion, Mumbai
University of Mumbai, India
kruti.kunverji@somaiya.edu

Krupa Shah
Information Technology
K.J. Somaiya Institute of Engineering and
Information Technology, Sion, Mumbai
University of Mumbai, India
krupa04@somaiya.edu

Prof. Nasim Banu Shah
Information Technology
K.J. Somaiya Institute of Engineering and
Information Technology, Sion, Mumbai
University of Mumbai, India
nshah@somaiya.edu

Abstract Floods have become the most well-known and lethal cataclysmic events of this century. Absence of a successful flood forecasting framework has brought about grave loss of human existence and infrastructure. This has reiterated on the importance of having in place a flood prediction system. This paper looks at developing the most effective flood determining model. AI calculations and a hearty, productive and precise flood expectation framework will give all the fundamental aid and assistance needed to the residents and government. Hence, the Decision Tree Model is being built. This model actualizes various calculations on datasets with a scope of accuracy. The model utilizes an AI calculation which predicts floods, sending alerts to the local and government authorities using an Android Application. The three Machine Learning Algorithms used for comparison are Decision Tree, Random Forest and Gradient Boost. This model focuses on improving the rate of prediction by dealing with more intricate information and a high-level algorithm.

Keywords Flood forecasting · Machine learning · Gradient boost · Decision tree · Random Forest · Android

1 Introduction

A flood happens when water submerges land that is normally dry, which can happen in an enormous number of ways. Brisk liquefying of ice, outlandish rainfall or a burst dam, can overwhelm a river, spreading over the contiguous land. Ocean front flooding happens when a colossal storm or tsunami makes the ocean flood inland. Floods are considered as the most common natural disaster on Earth, second only to the forest fires.

Environmental change is growing the risk of floods around the globe, particularly in waterfront and low-lying zones, because of its job in extraordinary climate situations and rising sea-level. The increment in temperatures that goes with an unnatural weather change can add to tropical storms that move even more gradually and drop more rainfall, channeling dampness into climatic streams. Thus, it is truly imperative to create frameworks that can foresee floods before the rainfall even hits the land. The recurrence and intensity of floods have been expanding since last century and it is currently perhaps the deadliest catastrophic events on earth. Because of global warming, the frequency at which the floods are expanding is expected to increment exponentially. Thus, in this new time of innovation and technology, it tends that flood prediction systems have been developed. In any case, the capability and accuracy of these models rely upon the information and algorithm utilized.

According to the Organization for Economic Cooperation and Development, floods cause damages of more than \$40 billion worldwide every year. Most nations actually do not have successful flood cautioning frameworks. According to the Central Water Commission, 20% of flood fatalities occur in India. Bihar is the most noticeably awful influenced state, with practically 73% of its complete surface territory getting overwhelmed every year. The cost of damage to infrastructure, crops, and public utilities all over India was reported to be as much as 3% of India's gross domestic product in 2018.

Despite the fact that there are different ways that can be undertaken to forestall floods, quite possibly the best and simplest early warning system is using AI algorithms [1] for the forecast of floods because of substantial rains and flooding

of various water bodies. With the approach of sensor innovation, different attributes have been recorded to anticipate floods. A wide scope of datasets is now available that can be utilized to create expectation frameworks. Machine Learning would guarantee vigorous, proficient, and precise predictions.

2 Literature Survey

Yovan Felix and T. Sasipraba [2] have developed a flood warning system for which data has been collected from remote sensing satellites and ground application. The parameters considered are the amount of rainfall and water-level of nearby water bodies. Gradient Boost Algorithm [3] is used to obtain a non-linear relationship between the total sum of rainfall and runoff, thus reducing the mean-squared error. For the datasets which were not a part of the training dataset, the Decision Tree Algorithm is used to predict floods on datasets which were not used in the training dataset. The architecture is divided into three layers which are interconnected. The three layers are the Physical Layer, Network Layer and Application Layer. Physical Layer collects the data from sensors and the Application Layer is the user interface which is the Mobile/Desktop Application. The historical dataset was collected from the Meteorological Department and Flood Forecasting Commission. At an interval of an hour, the real-time data is collected and sent to the Network Layer using the GSM Module and stored on the Cloud Server to be passed to the Machine Learning Model.

Ding et al. [4] forecasted floods in the region of the Lech river basin in Europe. Using a Spatio-Temporal Attention Long Short-Term Memory (STA-LSTM) Model, floods were forecasted. This is a data-driven model which sets up a connection between verifiable hydrological features and the runoff. LSTM shapes the fundamental piece of the neural network framework and is the adjustment of the Recurrent Neural Organization (RNN). Using the attention model takes out the issue of the impact of the equivalent hydrological highlights on different floods. It likewise analyzes the precision of the STA-LSTM with the Support Vector Machine (SVM), Fully Connected Network (FCN) and the original LSTM [5]. In the STA-LSTM Model, the skewness calculated for the training dataset is 0.58 and for the test dataset is -0.0651. At the point when the correlation at T+3 was done, the FCN model fared better compared to the SVM, LSTM and STA-LSTM network model. Yet, at T+6 and T+9, the STA-LSTM Model performs comparatively the best and FCN performs out the least.

Using the Sparse Bayes Model, Yirui Wu, Yukai Ding and Jun Feng [6] carried out flood prediction experiments in Changhua River. SMOTE [7] algorithm eliminates the issue of lopsided sample distribution; a Sparse Bayesian model is trained using AdaBoost Methodology which improves the model's performance in over-fitting. Using a group of Sparse Bayesian

models accomplishes a high accuracy when compared with a single Sparse Bayesian Model. It was seen that the model performs better compared to the single model. It was likewise presumed that the testing limit plays a significant part in deciding the performance of the model. The dataset for "Sparse Bayesian Flood Forecasting Model based on SMOTEBoost" is the yearly summer flood information of Changhua river basin from 1998 to 2010. The real-time data is recorded each hour. The data credits consist of the Changhua stream and rainfall, and the rainfall of the stations in the upper ranges of Changhua.

An urban flood estimating and checking platform created as a part of a UK Newton Fund project in Malaysia by Karyotis et al. [8] utilizes a hybrid Deep Learning (DL) and Fuzzy Logic (FL) based algorithm. This model uses low-cost sensors to gather real-time data. DL [9] utilizes Artificial Neural Organizations (ANN) for both supervised and unsupervised training, giving a dependable arrangement in time forecasting issues. FL, which depends on the idea of fuzzy sets, can deal with "fractional truth". The most ideal size of the data window is 200 data points for the Deep Learning Model. It was additionally seen that if the rain intensity is high, storm span is high and soil absorption is exceptionally low at that point likelihood of flood is high.

Dola et. al. [10] have developed a system to predict the occurrence of floods by using Machine Learning models. The data of rainfall from previously available data is used to predict the rainfall for the next month. Forecasting can be done for both short-term and long-term rainfall. Data has been gathered from the Indian Meteorological Department. Two distinctive datasets that comprise normal rainfall data from 1951-2000 for each month and district; the subsequent information is of 1901-2015, which comprises average rainfall data for each state. This Low Cost IoT based Flood Monitoring System employs IoT to figure out the time it would take for the flood to reach land. Severity of the rainfall is estimated using ML algorithms. The algorithms applied for the same are Linear Regression, Support Vector Machine and ANN. The various IoT devices used are rain-drop sensors, water-float sensors and IoT Gecko. As and when the water level rises, a buzzer beeps and an alert is sent of an approaching flood in the area. For the linear regression model, the dataset of the last three months is taken to predict the rainfall for the next month. It is the same for SVM. For ANN, CNN 1-D [11] strategy is applied. The mean absolute error for the linear regression algorithm is 40.2467874. The SVM model gave a mean absolute error of 90.606787. A mean absolute error of 21.8097545 was obtained for ANN.

3 Existing System

The Flood Detection and Warning System (FLoWS) [12] is the progression taken by the Malaysian government to help prevent the genuine damage caused to houses, streets, organizations, public offices and individuals by the annual

floods. It helps in monitoring and managing this critical circumstance by giving crucial data like flood conditions, plan and preparation, etc. to the general population and the local authorities at the affected territory. The system is able to measure the water level and alarm people in general and the local authorities by sending a warning through SMS and MMS in regards to the flood conditions. The system additionally empowers general society and the local authorities to see the live graph data of the water level using an Android application. It uses an ultrasonic sensor which quantifies the water level. The Raspberry Pi 3 goes about as a server to process and store every output from the microcontroller and use this data to trigger the Raspberry Pi camera to capture the image of the flood situation. The microcontroller gathers the data including water distance, temperature, humidity, and flood level. Then, the GSM SIM 900/900A is liable for sending the data from the microcontroller to the server utilizing AT command. The GSM is likewise responsible for sending warning messages, flood levels that have been measured from the sensor and image to the targeted mobile phone.

Global Flood Monitoring System (GFMS) is a computer tool which can be utilized for mapping flood conditions around the world. Created by Robert Adler and Huan Wu of the University of Maryland, it is utilized by zooming into an area of interest on the system's global interactive guide to see whether the water is at flood stage, subsiding, or rising. It can also be utilized to figure out whether there is a rain event

upstream, regardless of whether the rain is finished, and how the water is moving downstream. GFMS works 24x7, in any event, when there is cloud cover or other impedance. It depends on precipitation data from NASA's Earth observing satellites. Precipitation data from GFMS is joined with a land surface model that fuses vegetation cover, soil type, and terrain to decide how much water is absorbing and what amount is feeding the streamflow. users can see statistics for rainfall, streamflow, water profundity, and flooding every 3 hours. Users can likewise zoom in further to see inundation maps as fine as 1 km goal.

4 Dataset

This Decision Tree model helps predict floods in the districts of Bihar and Orissa in India. The districts of Bihar taken into account are Patna, Sheohar, Darbhanga, Kishanganj, East Champaran, West Champaran, Gopalganj, Sitamarhi, Muzzafarpur and Saran. Similarly, the districts of Orissa are Bhadrak, Cuttack, Jagatsinghpur, Kalahandi, Kendrapara, Baleshwar, Koraput, Puri, Jajapur and Sambalpur. The data collected is from the year 1992-2002. The data points collected during this period are 2640. The dataset was collected from a verified website called the India Water Portal [13]. The dataset had to be downloaded in a district-wise manner in a CSV format. The dataset collected has been gathered on a monthly basis. All the districts then had to be compiled together.

	A	B	C	D	E	F	G	H	I	J
1	LOCATION	YEAR	MONTH	MIN_TEMP	MAX_TEMP	RAINFALL	CLOUD_COVER	WET_DAY_FREQ	DIURNAL_TEMP	FLOOD OCCURRENCE
2	BIHAR-PATNA	1992	JANUARY	8.067	22.456	20.727	22.672	1.5696	14.389	NO
3	BIHAR-PATNA	1992	FEBRUARY	9.776	24.752	5.904	22.846	1	14.953	NO
4	BIHAR-PATNA	1992	MARCH	17.734	34.124	1.126	29.132	0.9778	16.364	NO
5	BIHAR-PATNA	1992	APRIL	22.842	38.889	4.345	25.191	1	16.021	NO
6	BIHAR-PATNA	1992	MAY	24.716	39.157	8.737	34.006	1.5548	14.394	NO
7	BIHAR-PATNA	1992	JUNE	27.115	37.999	33.17	58.664	3.6736	10.871	NO
8	BIHAR-PATNA	1992	JULY	26.981	34.403	191.148	73.11	11.17	7.423	NO
9	BIHAR-PATNA	1992	AUGUST	25.733	32.456	154.266	68.575	9.8425	6.721	NO
10	BIHAR-PATNA	1992	SEPTEMBER	25.44	33.177	103.491	57.506	7.2156	7.725	NO
11	BIHAR-PATNA	1992	OCTOBER	21.875	32.476	66.13	34.832	3.4364	10.598	NO
12	BIHAR-PATNA	1992	NOVEMBER	15.592	29.46	0.875	20.209	0.5023	13.845	NO
13	BIHAR-PATNA	1992	DECEMBER	10.132	24.75	0	22.926	0	14.62	NO
14	BIHAR-PATNA	1993	JANUARY	10.31	24.721	10.156	22.672	1.0301	14.389	NO
15	BIHAR-PATNA	1993	FEBRUARY	13.85	28.803	0.2	22.602	0.2	14.953	NO
16	BIHAR-PATNA	1993	MARCH	16.233	32.62	10.422	29.132	1.2841	16.364	NO
17	BIHAR-PATNA	1993	APRIL	21.817	37.861	4.268	25.191	1.002	16.021	NO
18	BIHAR-PATNA	1993	MAY	26.417	40.812	24.864	34.144	2.3305	14.394	NO
19	BIHAR-PATNA	1993	JUNE	28.39	39.262	37.736	58.715	4.0813	10.871	NO

Fig 1: Snapshot of Dataset

5 Scope of the System

Only a single flood has the capacity to cause a huge destruction which has again emphasized on the importance of having a flood detection system which is easy to operate and gives faster and accurate predictions. This system is being developed solely for the purpose of detection of floods. On the basis of the real-time data, the occurrence of a flood will be forecasted surrounding the flood prone areas. This system is being developed only for the country of India, particularly the districts of Bihar and Orissa which experience heavy rainfall and flooding almost every year. After detecting the occurrence of the flood, it will send a notification to the local people and the meteorological department. Using this, the government can undertake rescue and relocation operations faster.

6 Methodology

India has seasonal rainfall. It receives the heaviest rainfall during the period from June to September. At other times the rainfall that occurs is not enough to cause deluges. Hence, the attributes are considered to vary. The attributes that have been considered to be used in the model are the month, rainfall, cloud cover, wet day frequency, diurnal temperature. Figure 2 shows the structure of how the Flood Forecasting system is expected to work.

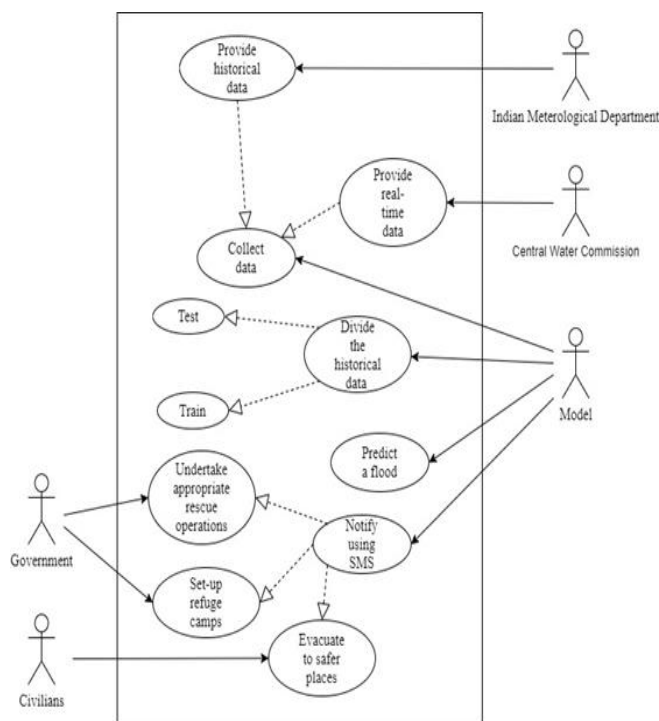


Fig. 2: Use-case diagram of System

Historical dataset was first collected from various sources and compiled. The dataset was then split into training and test data. It was divided in a 75:25 ratio, wherein training data had 75%

of the original data. There were three algorithms that were used to train the model. These were Decision Tree, Gradient Boost and Random Forest. This was done to figure out the algorithm that works best with the dataset. After figuring this out, alerts can be generated. Using an Android Application, the government officials can input the weather conditions and the occurrence of flood will be predicted. The government can then issue an alert. The civilians will receive a notification if an alert regarding an incoming flood has been issued. The civilians present at the scene can also issue alerts. The application also provides the citizens with the emergency steps that need to be taken when a flood occurs. The emergency contacts for the police, ambulance, fire department and the disaster management department have also been provided.

A decision tree is a decision support tool instrument that utilizes a tree-like model and their potential results, including decision event results, resource expenses, and utility. It is one approach to show an algorithm that just contains contingent control statements. The nodes in the graph address an event or decision and the edges of the chart address the choice guidelines or conditions. Random Forest is a supervised learning algorithm which is utilized for both classification as well as regression. Yet, nonetheless, it is mainly used for classification problems. The Random Forest algorithm makes decision trees on data samples and afterward gets the prediction from every one of them. It lastly chooses the best solution by method of voting. Gradient boosting re-defines boosting as a mathematical optimization issue where the objective is to minimize the loss function of the model by adding weak learners utilizing gradient descent. Gradient descent is a first-order iterative optimization algorithm for tracking down a local minimum of a differentiable function.

7 Results

Accuracy is a measurement unit used to evaluate the Machine Learning Algorithm. It indicates the percentage. The Decision Tree Algorithm gave an accuracy of 94.4%. The Gradient Boost Algorithm gave an accuracy of 87.9% whereas the Random Forest Algorithm gave an accuracy of 92.4%. Hence, the Decision Tree Algorithm was chosen for the model.

The scatter plot is a kind of mathematical graph in which two particular variables are mapped along the x-axis and y-axis. The resulting pattern reveals the correlation present between the two variables. For this particular model, the two variables chosen are Temperature and Rainfall. Temperature has been plotted along the y-axis and Rainfall has been plotted along the x-axis. The prediction of flood is heavily dependent on these two variables. Temperature has been calculated as an average of the minimum and maximum temperature mentioned in the dataset. The red points inside the red region indicate that the prediction of flood occurrence of those points at that particular temperature and rainfall has been done correctly by the model. The green points inside the red region are the points which

have been predicted wrong. The model predicted 'NO' for these green points inside the red region when the model should have predicted 'YES'. The green points inside the green region have been predicted correctly for the flood occurrence and the red points inside the green region have been predicted incorrectly.

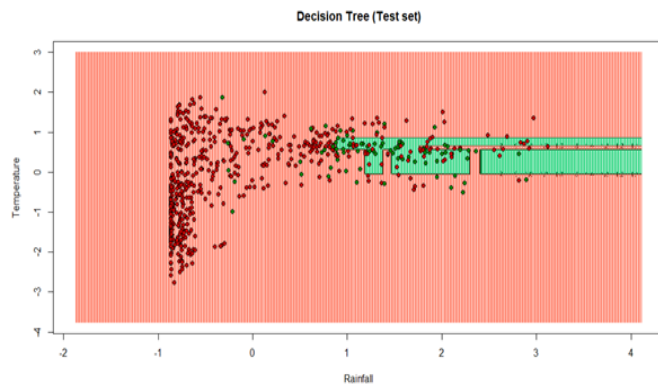


Fig. 3: Result of Decision Tree Algorithm

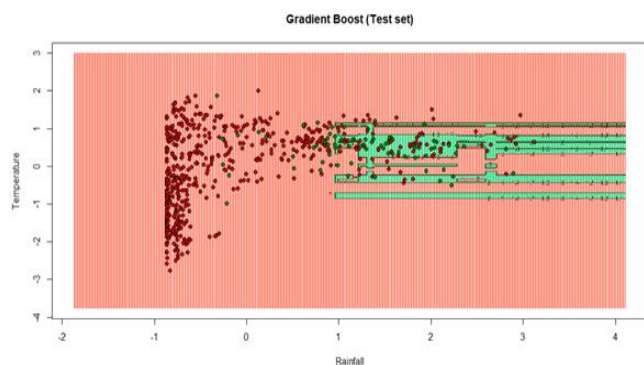


Fig. 4: Result of Gradient Boost Algorithm

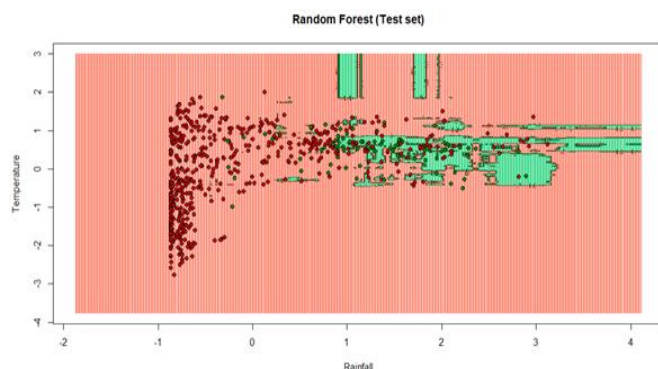


Fig. 5: Result of Random Forest Algorithm

The Decision Tree Algorithm has the highest accuracy. Figure 3 shows the scatter plot of the Decision Tree Algorithm depicting the accuracy. Figure 4 depicts the scatter plot of Gradient Boost Algorithm. The Confusion Matrix of the

Gradient Boost Algorithm showed 525 True Positives (TP), 49 False Negatives (FN), 36 False Positives (FP) and 50 True Negatives (TN). Figure 5 depicts the scatter plot of the Random Forest Algorithm. The Confusion Matrix of the Random Forest Algorithm showed 553 True Positives (TP), 17 False Negatives (FN), 33 False Positives (FP) and 57 True Negatives (TN).

The other existing model which uses Linear Regression, SVM and ANN gave a mean absolute error of 40.24, 90.61 and 21.81 respectively for the three algorithms. The Decision Tree Model explained in this paper gave a mean absolute error of 0.05606. Mean Absolute Error is another way of calculating accuracy for a Machine Learning Model. The lower the value of mean absolute error, the better the model's performance. Hence, the Decision Tree Model fared better when compared to the other proposed system.

The Android Application helps notify citizens of an imminent danger. It also helps the government predict floods and initiate rescue and relocation operations. The various activities included in the Android Application are Menu Activity, Flood Emergency Steps Activity, Issue Common Alerts, Issue Government Activity, Issue Alerts Menu Activity and Issue Government Alert Activity.

8 Conclusion and Future Scope

In the Decision Tree Machine Learning Algorithm, the parameters collected using an architectural set-up allows seamless integration of data. This data is then fed onto a Machine Learning model which is then able to predict the chances of flood. The proposed framework performs analysis with a high and satisfactory fault-tolerant accuracy.

The system has also been built according to the conditions prevalent in a country like India. The system sends out warnings and alerts of an incoming flood to the citizens and helps save the lives of civilians and if possible, the infrastructure. The system also helps the government save money in rescue operations and helps them start the relocation operations before the flood hits the town.

In the future, a collaboration between the forecast of rainfall and flood can be achieved. Using satellite imaging, the civilians can also be informed of safe places that they can relocate to and guide them towards the rehabilitation camps set up by the government.

Acknowledgement

The authors would like to thank everyone who has always been there to help and mentor us throughout this journey. They would like to express special thanks to their Project Guide Prof. Nasim Banu Shah for her kind cooperation, guidance and support in the planning and development of the project. She

has always motivated them to give their best and has helped them in meeting the deadlines and requirements well on time.

References

- [1] Mosavi, A., Ozturk, P., Chau, K.: Flood Prediction Using Machine Learning Models: Literature Review. *Water*, **10**, pp. 1-41 (2018)
- [2] Felix, A., Sasipraba, T.: Flood Detection Using Gradient Boost Machine Learning Approach. In: International Conference on Computational Intelligence and Knowledge Economy, pp. 779-783 (2019)
- [3] Ying, B., Sayed, A.: Diffusion gradient boosting for networked learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2512-2516 (2017)
- [4] Ding, Y., Zhu, Y., Wu, Y., Feng, J., Cheng, Z.: Spatio-Temporal Attention LSTM Model for Flood Forecasting. International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 458-465 (2019)
- [5] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation*, **9**(8), pp. 1-32 (1997)
- [6] Wu, Y., Ding, Y., Feng, J.: Sparse Bayesian Flood Forecasting Model Based on SMOTEBoost. In: International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 279-284 (2019)
- [7] Chawla, N., Bowyer, K., Hall L., Kegelmeyer, W.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, pp. 321-357 (2002)
- [8] Karyotis, C., Maniak, T., Doctor, F., Iqbal, R., Palade, V., Tang, R.: Deep Learning for Flood Forecasting and Monitoring in Urban Environments. In: 18th IEEE International Conference on Machine Learning and Applications, pp. 1392-1397 (2019)
- [9] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature*, **521**(7553), pp. (2015)
- [10] Rani, D., Dr. Jayalakshmi G N, Baligar, V.: Low Cost IoT based Flood Monitoring System Using Machine Learning and Neural Networks. In: Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications, pp. 261-267 (2020)
- [11] Albawi, S., Mohammed, T., Al-Zawi, S.: Understanding of a convolutional neural network. In: International Conference on Engineering and Technology, pp. 1-6 (2017)
- [12] Marzukhi, S., Sidik, M., Nasir, H., Zainol, Z., Ismail, M.: Flood Detection and Warning System (FLoWS). In: 12th International Conference on Ubiquitous Information Management and Communication, pp. 1-4 (2018)
- [13] India Water Portal: Data Finder. (2005) <https://www.indiawaterportal.org/datafinder>
- [14] Schwenk, H., Bengio, Y.: Adaboosting neural networks: Application to on-line character recognition. In: 7th International Conference, pp. 967-972 (1997)
- [15] Puttinaovarat, S., Horkaew, P.: Flood Forecasting System Based on Integrated Big and Crowdsourced Data by Using Machine Learning Techniques. *IEEE Access*, **8**, 5885-5905 (2020)
- [16] Ranit, A., Durge, P.: Flood Forecasting by Using Machine Learning. In: Proceedings of the Fourth International Conference on Communication and Electronics Systems, pp. 166-169 (2019)
- [17] Moniz, N., Ribeiro, R., Cerqueira, V., Chawla, N.: Smoteboost for regression: Improving the prediction of extreme values. In: IEEE 5th International Conference on Data Science and Advanced Analytics, pp. (2018)
- [18] Abdullahi, S., Habaebi, M., Malik, N.: Flood Disaster Warning System on the go. In: 7th International Conference on Computer and Communication Engineering, pp. 258-263 (2018)
- [19] Liu, F., Xu, F., Yang, S.: A Flood Forecasting Model based on Deep Learning Algorithm via Integrating Stacked Autoencoders with BP Neural Network. In: IEEE Third International Conference on Multimedia Big Data, pp. 58-61 (2017)
- [20] Noymanee, J., Nikitin, N., Kalyuzhnaya, A.: Urban Pluvial Flood Forecasting using Open Data with Machine Learning Techniques in Pattani Basin. In: 6th International Young Scientists Conference in HPC and Simulation, pp. 288-297 (2017)
- [21] Menon, K., Kala, L.: Video surveillance system for real-time flood detection and mobile app for flood alert. In: International Conference on Computing Methodologies and Communication, pp. 515-519 (2017)
- [22] Han, S., Coulibaly, P.: Bayesian flood forecasting methods: A review. *Journal of Hydrology*, **551**, 340-351 (2017)
- [23] Segretier, W., Collard, M., Clergue, M.: Evolutionary predictive modelling for flash floods. In: IEEE Congress on Evolutionary Computation, pp. 844-851 (2013)