ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse581022021.pdf
https://doi.org/10.30534/ijatcse/2021/581022021

Analysis of Model Performance and Implementation of an Optimized Flood Prediction Model Using Data Mining Techniques

Jittima Silprachawong[1*], Pornthip Wimonsong[2], Kannika Kaewchuea[3]
Faculty of Science and Technology, SuratthaniRajabhat University, SuratThani, Thailand
[1*]Corresponding author,jittima.sil@sru.ac.th
[2]pornthip.wim@sru.ac.th
[3]kannika.kae@sru.ac.th

ABSTRACT

Flooding is a major problem globally, and especially in SuratThani province, Thailand. Along the lower Tapeeriver in SuratThani, the population density is high. Implementing an early warning system can benefit people living along the banks here. In this study, our aim was to build a flood prediction model using artificial neural network (ANN), which would utilize water and stream levels along the lower Tapeeriver to predict floods. This model was used to predict flood using a dataset of rainfall and stream levels measured at local stations. The developed flood prediction model consisted of 4 input variables, namely, the rainfall amounts and stream levels at stations located in the PhraSeang district (X.37A), the Khian Sa district (X.217), and in the Phunphin district (X.5C). Model performance was evaluated using input data spanning a period of eight years (2011–2018). The model performance was compared with support vector machine (SVM), and ANN had better accuracy. The results showed an accuracy of 97.91% for the ANN model; however, for SVM it was 97.54%. Furthermore, the recall (42.78%) and f-measure (52.24%) were better for our model, however, the precision was lower. Therefore, the designed flood prediction model can estimate the likelihood of floods around the lower Tapee river region.

Key words: data mining techniques, flood prediction model, artificial neural network, support vector machine, model performance.

1. INTRODUCTION

Natural disasters are adverse events caused by natural hazards, such as volcanoes, earthquakes, tsunamis, storms, floods etc. Since, they are a global concern, we need to develop instruments and technologies to predict these events, such that we can safeguard people, animals, environment, and infrastructure during natural calamities. In particular, for disaster management, data mining techniques are used to build prediction models.

The incidences of natural disasters in Thailand, such as droughts and floods, are increasing. Floods are among the most destructive water-related natural disasters that cause property damage and loss of human lives [1]. Flooding occurs due to heavy rainfall in regions where water absorption is low, and river channels fail to control overflow. The more rapidly the rainwater reaches a river channel, the more likely it is to cause flooding [2]. In particular, flooding during the northeast and southwest monsoon seasons, generally results in widespread and significant damage in different regions in Thailand, which has also affected the overall economy of the country.

During the 2011 flood crisis, severe flooding occurred in several provinces, including SuratThani, which is in the upper central southern region. The basin area of the lower Tapeeriver is 3,180.787 km^2,and it flows through Wiang Sa, Ban Na San, Khian Sa, Phunphin, and Mueang districts. The floods destroyed farmlands and several houses. To aid disaster management, this study applied data mining techniques to build a flood prediction model.

ANNs and SVMs are among the most popular machine learning methods having applications in flood prediction [3]. These methods provide fast processing, and can be used for classification and grouping, forecasting, and developing relationship models. In this study, we aimed to determine which model could perform well in predicting floods. For evaluation, the four determinants of model performance used were accuracy, precision, recall, and f-measure. Herein, we determined which type of model yields realistic predictions.

2. LITERATURE REVIEW

Flood prediction models are classified into three categories, namely, physical or deterministic, conceptual or lumped, and empirical, metric, or black box models [3]. The advantages and disadvantages of each model are different. Physical and conceptual models have high accuracy because they mainly utilize physical data. However, the cost of data collection is quite high. Such models require the availability of tools, equipment, regular maintenance, and special personnel during fieldwork. Moreover, a considerable amount of time is required to update and process data in the field. The black box models are advantageous due to their faster processing times. The model database can be updated immediately and easily, however, their accuracy is not as high as the physical and conceptual models [4]. ANNs are classified as black-box

models. ANNs have structures and operating principles similar to those of a human brain. Using up-to-date datasets within a short time, ANN can learn and remember various forms and patterns [5]. Thus, it is an important modeling approach. However, support vector machines (SVMs) are one of the most popular approaches for flood prediction in addition to ANNs, and they use supervised learning models with associated learning algorithms. SVMs use the same concept of the well-known ability in any multivariate function with the desired accuracy level as the ANN [6].

There are numerous studies in which ANNs have been implemented to predict floods. For instance, a model developed to predict river floods, determined river water levels from rainfall[7]. Although, a number of factors can alter water levels, only two (one for rainfall and another for water level) were considered. A multi-layer perceptron network, using an ANN's feed-forward and back-propagation algorithms was used. Further, the authors analyzed the model-data fit and performed simulations to predict water levels. This model successfully predicted flood water levels 24 h ahead of time [7]. In addition, artificial neural network-based model was also investigated for the prediction of maximum water levels during a flash flood event [8]. Backpropagation based on an ANN can also be used to develop an appropriate forecasting model.The rainfall-runoff amounts from the gauge station in a municipal area (Chaiyaphum province), were utilized in the modeling procedure [9]. Records (1,824) of rainfall-runoff during 2007–2012 had been used to select stations, variables, and time lags. This predictive model, which used 15 input variables, had a mean absolute error of 1.008. Further, using ANN Sulafa et al. (2014) simulated flows at certain locations in the river reach, based on the flows at upstream locations. They utilized readings from stations along the Blue Nile, White Nile, Main Nile, and Atbara river between 1965 and 2003 to predict the likelihood of flooding at Dongola station (Sudan). This analysis also suggested that ANNs can provide a reliable estimate of flood hazards along the Nile river [10]. Additionally, factors like temperature and rainfall were used to develop an ANN model with a deep feed-forward neural network to predict floods in Nigeria [1]. The network had three hidden layers between the input and output layers. The two input parameters were temperature and rainfall, and the predicted standard precipitation index was the output. During this study, network training was performed using a back-propagation algorithmand two-thirds (67%) of the dataset, while the rest was used to test and validate the network Further, they implemented Adam's algorithm as an optimizer, and a loss function for categorical cross entropy. This model provides the predicted standard precipitation index as an output. However, the average accuracy of the model was 76%, which was sufficient for generating predictions. Moreover, introducing an extended Kalman filter at the output of the back-propagation neural network can show significant improvement in the prediction and tracking performance of models. A case study, which modeled flood water levels by collecting data at the upstream and downstream stations of a river in Sungai BatuPahat, Johor, Malaysia presented this approach [11].

In addition to ANNs, SVMs that are supervised learning models with associated learning algorithms, are also popular approaches in flood prediction. SVM uses the concept of a decision plane to define decision boundaries. A decision plane separates sets of objects that have different class memberships [6]. In this respect, a model using data mining techniques like SVM, naïve Bayes, k-nearest neighbor, decision tree, and multilayer perceptron, could predict rainfall in Lahore city [12]. The dataset consisting of several atmospheric attributes had been obtained from a weather forecasting website. Here, the authors utilized preprocessing techniques like cleaning and normalization to improve the prediction efficacy. Further, model performance was analyzed using precision, recall, and f-measure with different ratios of training and test data. In case of an ANN-based model, the climate change data is used to estimate the rainfall for the future century. The results showed that the training stage and more than 10 instances of high floods are forecasted for the future using climate change inputs [13]. In order to improve the efficiency, neural network algorithm was reported with the high efficiency methods for predicting the potential of floods after integrating utilization of GIS spatial analysis [14].

3. RESEARCH METHODOLOGY

Initially, to design a flood prediction model, we compared the performance of an ANN approach based on the NeuralNet algorithm with a SVM model. A data science platform, RapidMiner, which allows one to select an optimum model for predictive modeling, was used for evaluation. RapidMiner also provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. Secondly, we predicted flooding using the selected flood prediction model. Rainfall and stream level data collected from stations on the banks of lower Tapeeriver (SuratThani province) were used in our analyses.

3.1 Building the flood prediction model

Six major steps of the data mining process, which include business and data understanding, data preparation, model creation, model performance evaluation, and deployment, were used to build the flood prediction model (Figure 1).

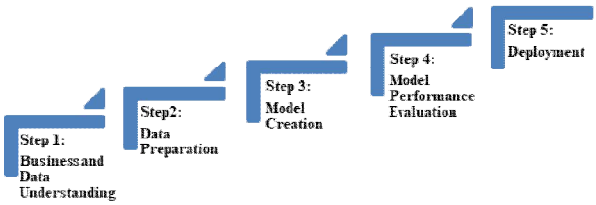

Figure 1: Data Mining Process

Step 1: Business and data Understanding

To facilitate business and data understanding, we focused on the flood prediction model for predicting floods in the area along the Lower Tapee River in SuratThani Province, which has been subjected to a number of major floods, as depicted in Figure 2. In the present study, we considered input that

spanned a period of eight years (2011–2018). Based on our prediction model, a warning can be issued to people living in this area and with the help of preemptive measures, extensive damage during floods can be avoided. For this purpose, we considered water quantities and levels based on rainfall and stream levels, as well as floods, in the validation dataset.

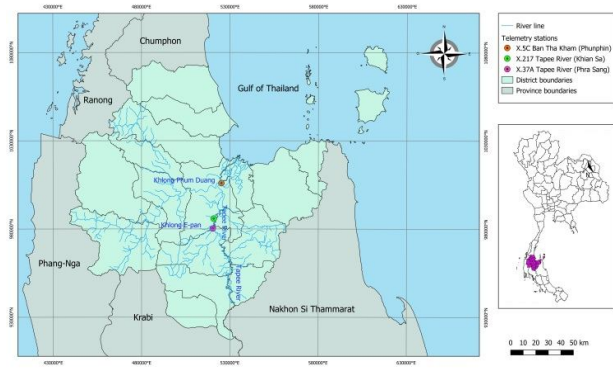

Figure 2: Location of the lower Tapeeriver.

Step 2: Data preparation

Hydrological data were obtained from the Southern Region Irrigation Hydrology Center (SIHC), Southern Irrigation Department (Phatthalung province). Daily water levels were additionally obtained from Phrasaeng station (X.37A), KhianSa station (X.217), and Phunphin station (X.5C), and rainfall data was obtained from the meteorological department in SuratThani (Figure 3). From the collected dataset, daily water levels and rainfall data over the period (2011–2018) were used to predict the amount of water. The collected data was organized into a single dataset using Microsoft excel, during the data integration process. This resulted in an input data, which had four attributes, namely, rainfall amounts and daily water levels from three locations including the X.37A, X.217, and X.5C locations (see Figure 3). The single output data provides the flooding attribute.The missing values are replaced with the average of the attributes, during the data cleaning process. Then, the data file was transformed into the csv format. Subsequently, RapidMiner was used for modeling.

Step 3: Model creation

Using RapidMiner, an ANN and a SVM model were generated (Figure 4). A data file with 2,922 records was imported into the software tool for data preparation. A cross-validation technique was used to separate the data into training and test groups. This was performed using the K-fold method, where the data file is randomly divided into k groups, or folds, of approximately equal size. The first fold was used for testing, and the remaining folds were trained on k-1 folds. To reduce bias, we selected 10 folds to build the model. The cross-validation operator connected the data file to the results of each model. The neural network and SVM algorithms were used for predictive modeling. In order to obtain an optimized model, patterns were executed and run with the training and test groups.

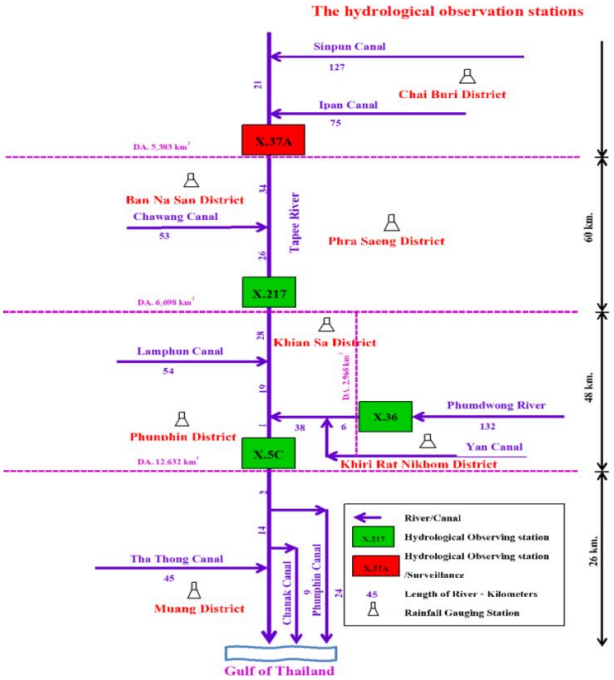

Figure 3: Stream level stations in the study area along the lower Tapeeriver.

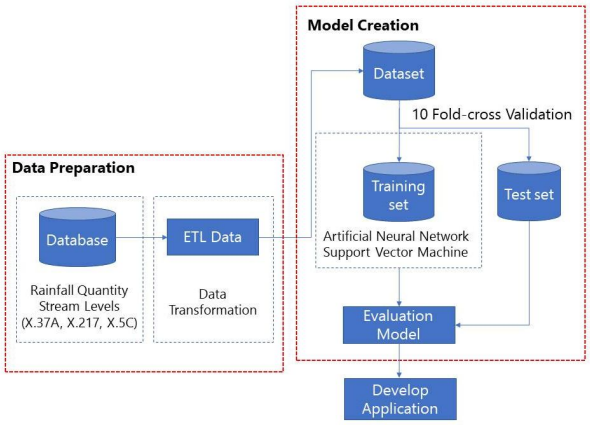

Figure 4: Schematic ofthe modeling procedure.

Step 4: Model performance evaluation

During analysis and comparison, performance of the models was evaluated using four parameters, namely, accuracy, precision, recall, and f-measure [15], [16]. All parameters were derived from the confusion matrix, as shown in Figure 5.

Predicted Values	Actual Values: Class = Yes	Actual Values: Class = No
Class = Yes	True Positive (TP)	False Positive (FP)
Class = No	False Negative (FN)	True Negative (TN)

Figure 5: The confusion matrix is represented here.

Accuracy is a measurement of the integrity of a particular model by considering every class.

$$Accuracy = TP + TN/TP + TN + FP + FN, \quad (1)$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Precision of a particular model is determined by,

$$Precision = TP/(TP + FP). \quad (2)$$

Further, recall was calculated as,

$$Recall = TP/(TP + FN). \quad (3)$$

The F-measure is a measurement of the combination of the precision and recall values for a particular model by considering each class:

$$F\ \ measure = \frac{Precision*Recall*2}{(Precision+Recall)} (4)$$

In addition, the ANN approach for the flood prediction model was derived from equations (5) and (6) as follows [17]:

$$f(x) = \sum_{i=1}^{n} x_i w_i + b, (5)$$

where x, w, n, and b denote each node of the input layer, the weight of each node, the number of input layers, bias or threshold, respectively. Next, a sigmoid function operates on the weighted sum of each hidden node to determine the output values from the ANN. The values were restricted between 0 and 1. For all values of the sigmoid function x ≥ 0.5, 1 is returned; otherwise, the value 0 is returned. The sigmoid function used is as follows [16]:

$$y(x) = \frac{1}{1+e^{-x}}, (6)$$

where x is the weighted sum of each node.

Step 5: Model deployment

Based on the accuracy, precision, recall, and f-measure values obtained for the models, we selected an optimized approach for data mining. Subsequently, the optimized approach was used to build the flood prediction model.

3.2 Possible Application of the Flood Prediction Model

The application for flood prediction was developed using the python programming language (PPL). We used python for machine learning, since it is simple, user-friendly provides efficient data mining tools, and can support modules scikit-learn to build models. Therefore, the flood prediction model was implemented using the processes from the previous analyses. The system runs with a window allowing the entry of the input data (rainfall quantity, stream levels) for each station (X.37A, X.217, and X.5C). The output value obtained indicates whether flooding will occur.

4. RESULTS

The amount of rainfall and stream levels at X.37A, X.217 and X.5C were used in the present study. Results for both the developed flood prediction model and application software used in this study are presented. ANN and SVM approaches were compared for their model performances. Then, the best model was used for developing flood prediction model and for utilizing the application software.

4.1Flood prediction model performance

We observed that the accuracy, recall, and the f-measure for the ANN model was better than SVM. However, the precision of the SVM model was better than ANN (Table 1).

Table 1: Model performance evaluation.

Model	Accuracy	Precision	Recall	F-Measure
ANN	97.91%	73.20%	42.78%	52.24%
SVM	97.54%	87.50%	17.22%	28.00%

Therefore, we developed the flood prediction model using the neural network approach of ANN, which had three layers, namely, the input, hidden, and output layers [16] (Figure 6). Each node was connected to each node in the next layer, and each connection was assigned a particular weight. The weight specifies the impact of a given node on a node in the next layer. When all node values from the input layer are multiplied by their weights and combined, a value for a hidden node is obtained.

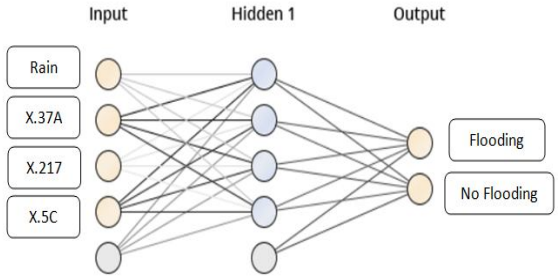

Figure 6: Schematic representation of the flood prediction model

The weights for each hidden node are listed in Table 2. Further, we can represent equation f(x) as follows:

Node 1 = Rain*-1.120)+(X.37A*-4.004)+(X.217*0.525) +(X.5C*-3.958)-1.742

Node 2 = (Rain*-1.115)+(X.37A*-3.706)+(X.217*0.565) +(X.5C*-3.671)-1.715

Node 3 = (Rain*-1.041)+(X.37A*-3.355)+(X.217*0.515) +(X.5C*-3.316)-1.606

Node 4 = (Rain*-1.120)+(X.37A*-3.973)+(X.217*0.665) +(X.5C*-3.980)-1.760

Table 2: Weights for each hidden node.

Input Node	Node 1	Node 2	Node 3	Node 4
Rain	-1.120	-1.115	-1.041	-1.120
X.37A	-4.004	-3.706	-3.355	-3.973
X.217	0.525	0.565	0.515	0.665
X.5C	-3.958	-3.671	-3.316	-3.980
Bias	-1.742	-1.715	-1.606	-1.760

Further, the weights of the output nodes are multiplied by the results of the sigmoid function for each hidden node. The weighted sums of the output nodes are calculated and with the help of the sigmoid function, we obtained the output values for flood prediction (Table 3).

Table 3: Weights for the output values.

Hidden Node	Output 1 "No"	Output 2 "Yes"
Node 1	3.107	-3.040
Node 2	2.831	-2.857
Node 3	2.519	-2.538
Node 4	3.087	-3.113
Threshold	-3.065	3.067

Output 1 (No) = (Node 1*3.107)+(Node 2*2.831)+(Node 3*2.519)+(Node 4*3.087)-3.065

Output 2 (Yes) = (Node 1*-3.040) + (Node 2*-2.857)+(Node 3*-2.538)+(Node 4*-3.113)+3.067

Therefore, the flood prediction model provided the output 1 (NO) and output 2 (yes) values, which were then utilized in the sigmoid function as follows:

$$y(x) = \frac{1}{1+e^{-(Output1\ or\ Output2)}}. \qquad (7)$$

When one of these values has y(x)=1, this is the prediction value. This suggests that if output 1 (No) is 1, it predicts "No Flooding", whereas if output 2 (Yes) is 1, it predicts "Flooding."

4.2Flood Prediction Model Application

Using the ANN approach and PPL, an application for utilizing the flood prediction model was developed. The system processes the input data, such as the rainfall amounts and stream levels for X.37/a, X.217, and X.5C, as depicted in Figure 7. Subsequently, the system calculates the result when the input values are submitted by clicking the button, "Please click on this button."

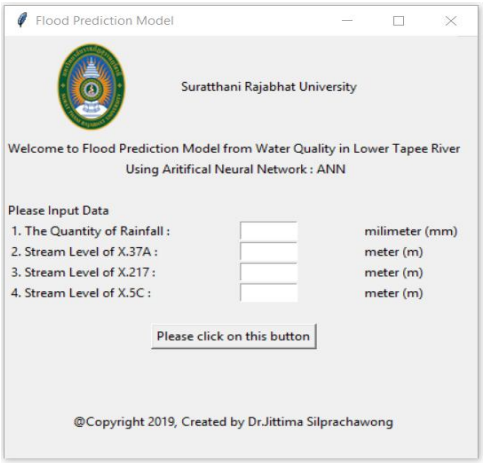

Figure 7: Flood prediction model for the input of data

5. DISCUSSION AND CONCLUSION

A model to predict floods based on ANN, and an application to utilize it was developed. Since, model evaluation suggested that ANN performs better than SVM we utilized the ANN approach. The similarity of this study was revealed in the prediction of the flood water level from rainfall and current river water level, which resulted in better model performance when compared with the others [4], [9]. Previous studies have shown that it is possible to predict floods using ANN models based on rainfall and temperature [1], [2]. Earlier, the developed ANN model predicted rainfall four months in advance [18], [19]. However, our flood prediction application can provide a warning 24 h in advance. The application was developed such that one can obtain accurate predictions in real life situations. In future, we would like to utilize additional parameters like geographic information system (GIS) data, satellite images, weather data etc., to build more accurate flood prediction models.

ACKNOWLEDGEMENT

The Research and Development Institute, SuratThaniRajabhat University supported this study. We would like to give special thanks to SIHC, Southern Irrigation Department, Phatthalung province, and the Meteorological Department, SuratThani province for the provided data. We would also like to thank the Science and Technology faculty for contributing to and facilitating this study. We would like to acknowledge Dr. AongartAun-a-nan and Mr. AsokSrisawat for their suggestions and assistance.

REFERENCES

1. EsiefarienrheBukohwo Michael andOfikwuEne Patience, **Flood prediction in Nigeria using artificial neural network**, American Journal of Engineering Research, e-ISSN: 2320-0847, Vol. 7(9), pp. 15-21, 2018.
2. Amir Mosavi, Pinar Ozturk, and Chau Kwok-wing,**Flood prediction using machine learning models: literature review**, Water, Vol. 10, pp. 1536, 2018.
3. Gayathri, K. Devi, B. P. Ganasri, and G. S. Dwarakish, **A review on hydrological models**, Aquatic Procedia, 4pp. 1001-1007, 2015.
4. ThaveesakVangpaisalandJakkaritThreenat,**Factors affecting the accuracy of water level forecasting at M.7 gauge station using artificial neural network model**, TU Science Journal of Science, 6(1), pp. 50-60, 2013.
5. ChaipimonplinTawee, See, M. Linda, andKneale, E.Pauline,**Using radar data to extend the lead time of neural network forecasting on the river Ping**, Disaster Advances, Vol. 3(3), pp. 35-43, 2010.
6. Vojislav Kecman, **Support Vector Machines: Theory and Applications,** Springer,pp. 1-47, 2005.
7. Abhijit PaulandProdipto Das,**Flood prediction model using artificial neural network**, International Journal of Computer Applications Technology and Research, Vol. 3(7), pp. 473-478, 2014.
8. Simon Berkhahn, Lothar Fuchs and InsaNeuweiler, **An ensemble neural network model for real-time prediction of urban floods**. Journal of Hydrology, Vol. 575, pp. 743-754, 2019.
9. MuninWanatada and PunneeSittidech, **Runoff forecasting using back-propagation neural network technique: case study of municipally of chaiyaphum**, The 9th national conference on computing and information technology (In Thai), Thailand, NCCIT3, pp.179-184, 2013.
10. Sulafa Hag Elsafi,**Artificial neural networks (ANNs) for flood forecasting at dongola station in the river Nile, Sudan**, Alexandria Engineering Journal, Vol. 53, pp. 655-662, 2014.

11. Rahmiaulia Adnan, FazlinaAhmatRuslan, A.M. Samad, and ZainazlanMdZain,**Flood water level modelling and prediction using artificial neural network: case study of Sungai BatuPahat in Johor**, IEEE Control and System Graduate Colloquium (ICSGRC), pp. 22-25, 2012.
12. ShabibAftab, Munir Ahmad, NoureenHameed, Muhammad Salman Bashir, Iftikhar Ali and Zahid Nawaz,**Rainfall prediction in Lahore city using data mining techniques**, International Journal of Advanced Computer Science and Applications, Vol. 9(4), pp. 254-260, 2018.
13. B. G. Rajeev Gandhi, Dilip Kumar and Hira Lal Yadav,**An Artificial Neural Network Model for Estimating the Flood in Tehri Region of Uttarakhand Using Rainfall Data,** Soft Computing: Theories and Applications, pp. 467-477, 2020.
14. Mohammad Hossein Jahangir, SeyedehMahsaMousaviReineh and MahnazAbolghasemi,**Spatial predication of flood zonation mapping in Kan River Basin, Iran, using artificial neural network algorithm,** Weather and Climate Extremes, Vol. 25, pp. 100215, 2019.
15. Jiawei Han, MichelineKamber, and Jian Pie, **Data mining: concepts and techniques (The Morgan Kaufmann Series in Data Management Systems)**, 3rd editionWaltham: MA, 2012.
16. Witten, Ian H., Frank, Eibe, Hall, Mark A. and Pal, Christopher J.,**Data mining: practical machine learning tools and techniques (Morgan Kaufmann Series in Data Management Systems)**, published by Morgan Kaufmann publishers, 4th Edition, Cambridge: MA, 2017.
17. **The mathematics of neural networks**, retrieved June 19, 2016, available at https://medium.com/coinmonks/the-mathematics-of-neural-network-60a112dd3e05.
18. Jeongwoo Lee, Chul-Gyum Kim, JeongEun Lee, Nam Won Kim, and Hyeonjun Kim, **Application of artificial neural networks to rainfall forecasting in the Geum river basin, Korea,**Water, Vol. 10, pp. 1448, 2018; doi:10.3390/w10101448.
19. RakeshTanty, and Tanweer S. Deshmukh, **Application of artificial neural network in hydrology – a review**, International Journal of Engineering research & Technology, Vol. 4(6), 2015.