2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)

Early Flood Risk Assessment using Machine Learning: A Comparative study of SVM, Q-SVM, K-NN and LDA

Talha Ahmed Khan
British Malaysian Institute, Universiti Kuala Lumpur
Usman Institute of Technology
Karachi, Pakistan
Talha.khan@s.unikl.edu.my

Zeeshan Shahid
Institute of Business Management (IoBM), Pakistan
Karachi, Pakistan

Muhammad Alam
Institute of Business Management (IoBM), Pakistan

M.M Su'ud
Malaysian France Institute(MFI),
Universiti Kuala Lumpur,
Kuala Lumpur, Malaysia

Kushsairy Kadir
British Malaysian Institute, Universiti Kuala Lumpur
Kuala Lumpur, Malaysia

Abstract— **Abundant floods and cyclones are the major cause of large emergency and acute ruin of properties in various countries. Usually floods are acknowledged as one of the most crucial problem in Malaysia, Indonesia, Bangladesh and France etc. Diverse techniques were carried out for a robust prediction system to investigate the flash floods. A dynamic system for the identification of run offs involves the computation of water peak, rainfall velocity, Global Positioning System-Precipitable Water Vapor (GPS PWV), wind speed, orientation, complex levels of river, land humidity, oceanic basement pressure and flash flood color with authentic cognizance algorithms. Accurate and precise forecasting of floods is very complex as it depends on many factors like precipitation, cloud to ground flashes, geo-magnetic field, color of water, wind velocity, wind direction, temperature and others. In this research paper classification approaches like Linear Support vector machine, Quadratic Support vector machine, K-nearest neighbor and Linear discriminant analysis have been implemented to classify the true positive event of flash floods accurately and precisely. Comparative analysis has been performed between these three algorithms to determine the highest accuracy algorithm. Parametric comparison and results of training and testing proved that Support Vector Machine (SVM) performed very well.**

Keywords— Flash flood forecasting, false alarm rate, natural disaster, Support Vector Machine, K-NN, QSVM

I. INTRODUCTION

Flash flood can be regarded as a leading instinctive hazard all over the world that profoundly affect living things safety, worthwhile accouterments, infrastructure, longhorn and farming production. Techniques can be categorized into two types of engineering (building of waiter bank and non-engineering (early identification investigation of flash floods using AI). These types were further advanced and developed by using instrumentation and AI based algorithms [1]. Automatic Target Recognition Algorithms produced more false alarms as they depend on the climatic conditions where they are installed. False alarms are produced due to the harsh environment and complex anisotropy of the sea-floor Williams et. al. [2-4]. Dual Tree Complex Wavelet Transform (DTCWT) was used to measure the textural characteristics like areas of the sea having same design and patterns Clutter characteristics were investigated by Markov Random Field (MRF). ATR was compared with the sea-floor filter. A lesser effect was observed in the results proving that this approach can be used to minimize the false alarm in the ATR. An adequate system requires early warning of runoffs and exit routes guidance [2]. To design a disaster management model is quite complex as sewages are the combination of various irregular objects. [3-5]. Many sensors like acoustic sensors, seismic sensors, passive infrared sensors etc.) have been deployed for sensing the flash flood and seismic activity. UGS are cost effective solution. Commonly Unattended Ground sensors have large fake alerts ratio because of inadequate algorithms. [6-7]. Smaller batteries backup is also a crucial problem for sensed data transmission [8]. Parameters for the uttermost identification of flash floods includes computation of the rainfall intensity, water level, lake levels, wind velocity,

978-1-7281-4956-1/19/$31.00 ©2019 IEEE

wind pressure, temperature, color of the water, distance, blockage, wave current pattern, wind direction, GPS-Precipitable Water Vapor, moisture level, Flickers from cloud to ground surface and bathymetry assessment, hydrological and climatic differences have been taken as the criterion to evaluate the deadly floods [9]. Fuzzy logic, Multi-layer perceptron(MLP), back propagation neural network, Support vector machine (SVM), extended Kalman filtering (EKF), Adaptive neural fuzzy interference system (ANFIS), Neural network autoregressive model with exogenous input (NNARX) and Particle Swarm Optimization (PSO) based structure have been developed to forecast the floods. WARN receives input from pressure spectator (lower surface of ocean) and computes to disclose the signs of shockwaves from earth [10-12]. To minimize the false alarms in forecasting floods, Multi-layer perceptron(MLP) based neural network was configured [13]. Arithmetic method was enforced and three hours' data were processed. The second procedure was applied for the calculation of rainfall [14-15]. Ultrasonic sensors with camera have been used to measure the false alarm rate in forecasting [16]. A fresh novel research was found in which geomagnetic field changes were measured using tesla meters during the extreme flash floods [17-18]. Results showed that during the flood event magnetic field that is usually propagated from the earth center rapidly reduced [19]. Scaled conjugate gradient worked better in the forecasting of floods [20].

II. PROBLEM STATEMENT

Flash floods are very abrupt and causes infrastructural and human loss. Accurate early forecasting of flash floods with less false alarms has been a crucial topic for researchers since decade. Many diversified approaches were studied but they were not up to the mark due to the inadequate algorithms or missed data. Accuracy and reliability of sensors deteriorated due to the poor sensitivity in harsh environment. In this research paper the data has been collected from Pakistan Meteorological Department and Machine Learning Classifiers have been implemented to diminish the false alarm rate. Forecasting flood at a particular location depend essentially only on measurements that could be measured on location. Estimation of the sky, wind and temperature conditions as well as the previous data of the regional climate history may vary according to the change of venue.

III. METHODOLOGY

A. Basic Flow Diagram

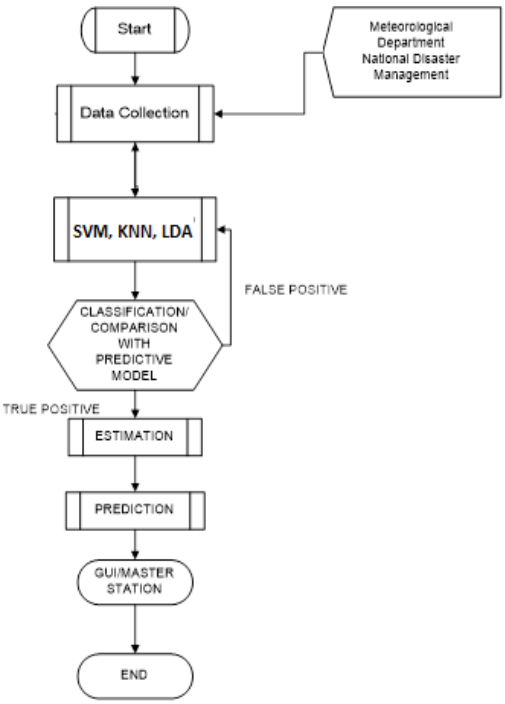

Fig. 1. Basic Flow Diagram

Initially data was collected from the Pakistan Meteorological Department. The data was in raw form it was filtered initially and then machine learning classifiers were applied.

B. Risk Levels

Risk levels were decided for the forecasting of floods. Low level was defined as class 0 flood. Class 1 has some moderate risk for the flood event and class 3 is the High alert for the flood risk.

CLASS 0 = Low

CLASS 1 = Moderate

CLASS 2 = High

C. Data Collection

Table 1: PMD Dataset for 2016-2017

Precipitation Past 24 hrs.(mm)	Temperature Max °C	Temperature Min °C	Humidity 1200 UTC (%)	Wind speed 1200 UTC (knots)	Cloudiness 1200 UTC (okta)	Wind direction 1200 UTC (degree)	Average Temperature °C	Class
0	29.4	16.5	29	6	0	225	23	1
0	31.5	14	50	6	5	225	22.8	1
0	29.8	14	50	6	0	225	21.9	1
0	30.5	14.5	63	8	5	225	22.5	1
0	28.7	15.3	37	4	0	180	22	1
0	29.5	16	58	6	5	225	22.8	1
0	27.5	16.6	52	6	4	225	22.1	1
0	28	14.5	55	10	3	225	21.3	1
0	28.2	13.5	55	10	2	225	20.9	1
0	28.5	14	56	8	3	225	21.3	1
0	28	15.5	62	8	4	225	21.8	1
0	28	17	19	4	0	315	22.5	1
0	28.5	15	19	8	0	45	21.8	1
0	26.5	14	32	4	0	225	20.3	1
0	27	14	24	0	0	0	20.5	1
0	27.5	13.5	47	12	1	225	20.5	1
0	27	14.5	46	6	4	225	20.8	1
0	28	15	45	4	2	180	21.5	1
3.1	28	14	39	12	2	225	21	1
0	27.3	10.5	32	8	0	225	18.9	0
0	27.5	11.5	45	10	0	225	19.5	0
0	27	13	54	10	4	225	20	1
0	27.5	13.5	39	6	0	135	20.5	1
0	26	13.5	47	8	0	225	19.8	0
0	26.5	13	57	6	3	225	19.8	0
0	25.5	13.5	59	14	4	225	19.5	0
0	27.3	17	49	6	6	135	22.2	1
0	24	16.5	62	8	3	225	20.3	1
0	27.5	16.5	56	16	4	225	22	1
0	37	25	46	10	0	248	31	2
0	37.5	25	61	12	4	225	31.3	2
0	35	26	59	16	6	225	30.5	2

The collected dataset that has been used for research paper comprised of eight attributes and the time duration of the dataset is for previous two years 2016-2017. Each of the attribute of dataset contributed to the training of model and helped in training of algorithm as they are logically and mathematically co-related with each other. Usually climate has tremendous repetition, so the main task of the model is to find models that often appear in the data set and learn about changes that will allow you to predict these common models. The forecast requires a large amount of data and the number of data is directly proportional to the efficiency and reliability of system. It is clear that data is never normally present, but always includes irregularities and layoffs. Therefore, we must eliminate these irregularities and layoffs to prepare the data for processing. Data was received in a clear and standardized way.

D. Support Vector Machine (SVM)

Support Vector Machine (SVM) can be acknowledged as supervised machine learning algorithm. Usually SVMs are applied to solve the classification and regressions problems. SVMs classifies the data by creating Hyperplanes which segregates the data. Identification of Hyperplane is also mandatory in order to classify the data accurately and precisely. In easy words Hyperplane can be considered as the decision surface for the support vectors. Support vectors are the nearest points to the hyperplane which creates classification boundary. Input for the SVMs can be taken as x_1, x_2, x_3…..x_n and output would be processed as Y_n according to the value of weights W_i. In our case study three hyperplanes were created

HP1, H2 and H3 are the planes:

HP1: $w \bullet x_i + b = 0$ (1)

HP2: $w \bullet x_i + b = 1$ (2)

HP3: $w \bullet x_i + b = 2$ (3)

Plane H0 is median in between, where $w \bullet xi + b = 0$

$w^T x + b \geq 0$ for di = 0 (4)

$w^T x + b \geq 1$ for di = 1 (5)

$w^T x + b \geq 2$ for di = 2 (6)

For the maximization of the margin, ||w|| will be minimized. Having the case that there will be no data values between HP1 and HP2

Non-Linear SVMs also used to separate the classes linearly by using the quadratic equation.

$(y-a)(y-b) = y^2 - (m+n)y + mn$ (7)

Optimization issue of the weight values can be resolved by using the following equations for SVMs:

For the maximization;

$$\frac{1}{\|w\|} \quad (8)$$

Min. $| w^T x + b | = 0$ for n = 1,2,3……n

For the minimization;

$$\frac{1}{2} W^T.W \quad (9)$$

$y_n = | w^T x + b | = 0$ for n = 1,2,3……n

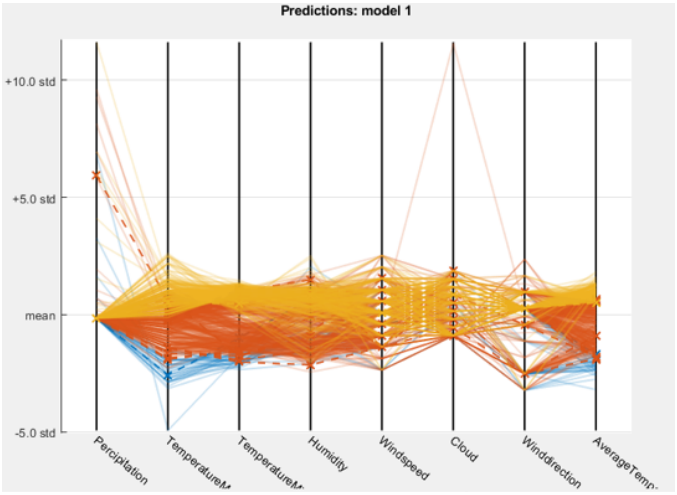

Fig. 2. SVM Predictions Model

Fig. 2 shows the prediction model for Support vector machine (SVM). Data values have been graphically illustrated in this figure.

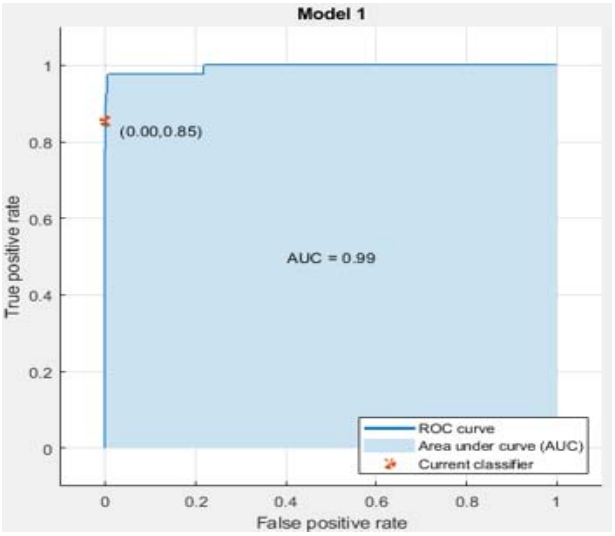

Fig. 3. SVM ROC

ROC is the graphical evaluation and visualization of the multi-class classifier performance. The higher the area under curve(AUC) the higher the classification of classes. Receiver operating characteristics (ROC) curve has been sketched with True positive rate at y-axis with False positive rate(FPR) at x-axis. Area under the curve (AUC) achieved 0.99 value which is very high. Generally, it is said that the more the AUC is closer to one the better performance of classification would be achieved therefore AUC can be acknowledged as yardstick to gauge the multi-class classification.

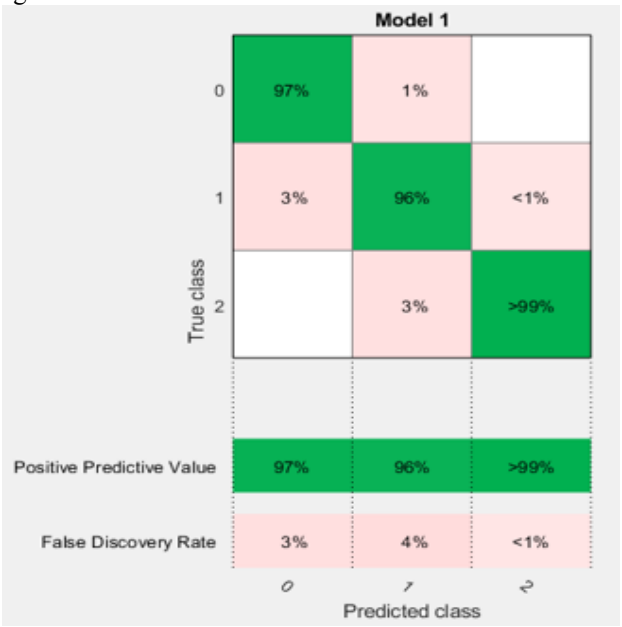

Fig. 4. Linear SVM Confusion Matrix

Fig. 4 presents the confusion matrix for the linear support vector machine (SVM). True positive, True negative, False positive and false negative can easily be estimated by the confusion matrix. Performance measurement of any classifier depends upon the AUC, precision, f-measure, recall and accuracy parameters which can be calculated by the True positive, True negative, False positive and false negative. In this figure confusion matrix shows that class 0 has achieved 97% accuracy in classifying true discovery and false positive rate was around 3%. Class 1 has achieved 96% accuracy of classifying true discovery rate and false positive rate was 45. Moreover, greater than 99% was achieved in class 2 while false discovery rate was found to be less than 1%.

Recall is the measurement of corrected classified values out of the all positive classes. The higher the recall the better the performance. Recall can be measured by using eq. (10).

$$Recall = \frac{Tp}{TP + FN} \tag{10}$$

F-measure can be acknowledged as the comparative analysis or to know the comparison between recall and precision.

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{11}$$

The estimation of actual positive out of all positive can be classes can be found be precision.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

Table 2: SVM Parameters

SVM Parameters	Results
Accuracy	97.4%
Prediction Speed	5300 obs/sec
Training time	7.5003 seconds
Precision	0.97
Recall	0.98
F-measure	0.98

Fig. 5. Quadratic SVM Confusion Matrix

Fig. 5. Shows the confusion matrix of quadratic SVM. In quadratic SVM ‖w‖ has to be minimized. The following quadratic function was applied:

$$\min f(n) = \frac{1}{2}\|w\|^2 \quad (13)$$

$$g(n)=yi*(w.n_i)–b = 0 \quad (14)$$

$$g(n)=yi*(w.n_i)–b = 0 \quad (15)$$

$$g(n)=yi*(w.n_i)–b = 0 \quad (16)$$

Quadratic SVM was applied to the data set and produced 92% maximum accuracy in classifying positive values in class 1 and 8% false error rate was achieved in classifying positive values. Class 2 achieved 96% efficiency in predicting true positive values and 3% false classification was found. In class 2 positive predictive rate was around 99% and less than 1% false discovery rate was calculated.

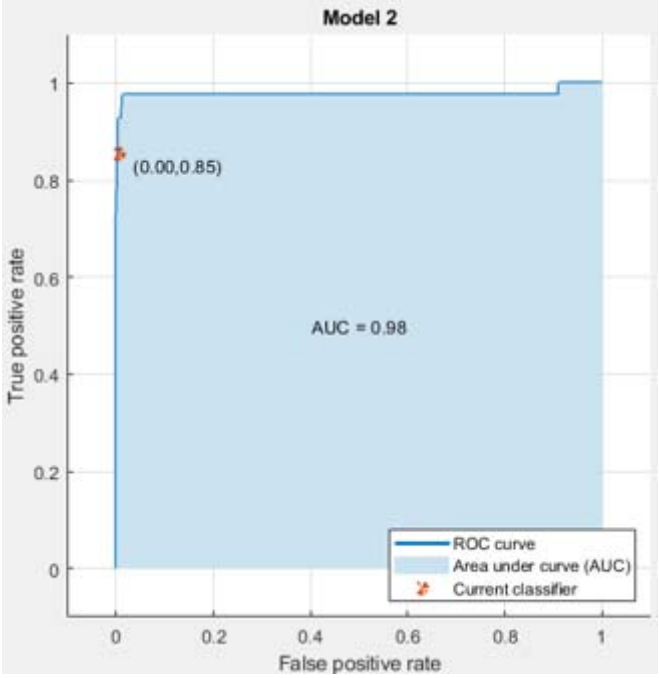

Fig. 6. Quadratic SVM ROC

Generally, ROC is the graphical illustration and assessment of the multi-class classifier performance. The maximum the area under curve(AUC) the maximum the classification of classes would be achieved. Receiver operating characteristics (ROC) curve has been plotted with True positive rate(TPR) at y-axis against False positive rate(FPR) at x-axis. AUC achieved 0.98 slightly less than the linear SVM.

Table 3: Q-SVM Parameters

Quadratic-SVM Parameters	Results
Accuracy	96.9%
Prediction Speed	24000 obs/sec
Training time	0.76163 seconds
Precision	0.95
Recall	0.96
F-measure	0.96

E. K-Nearest Neighbour (KNN)

"K" can be used as a controlling variable parameter for the K- nearest neighbor classifier. Selection of suitable "K" parameter value is mandatory like in proper learning rate is required in the cost function. Usually Euclidean distance is calculated to find out the closest distance with the value of the K. For the classes 0,1 and 2 three values of K would be calculated. The nearest value of the Euclidean distance with the value of "k" would be selected.

$$d = \sqrt{(x1 - xA1)^2 + (x2 - xA2)^2} \quad (17)$$

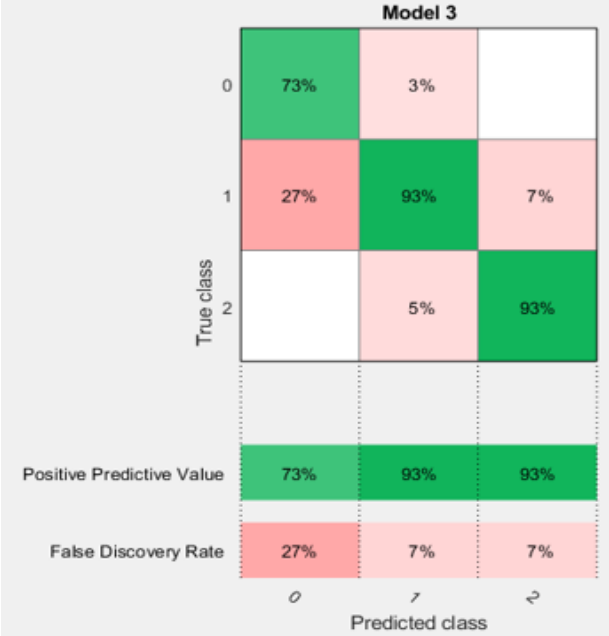

Fig.7. KNN Confusion Matrix

KNN confusion matrix has been produced to evaluate the KNN algorithm classification performance. It has been observed through the confusion matrix that 73% classification rate was found to be positive while 27% false discovery rate was determined for class 0. Class 1 achieved 93% positive true rate and 7% false discovery rate. In class 2 same results were achieved for the classification like in class 2.

Table 3: k-NN Parameters

K-NN Parameters	Results
Accuracy	91.7%
Prediction Speed	13000 obs/sec
Training time	1.3133 seconds
Precision	0.91
Recall	0.89
F-measure	0.89

F. Linear Discriminant Analysis

Linear discriminant analysis is usually preferred when there are more than two classes for the classification. LDA works in a new dimension by using increased distance between means of classes and minimum scattered variations known as co-variance by using the following equation:

$$max \frac{(\mu 1 - \mu 2)^2}{\alpha 1^2 + \alpha 2^2} \quad (18)$$

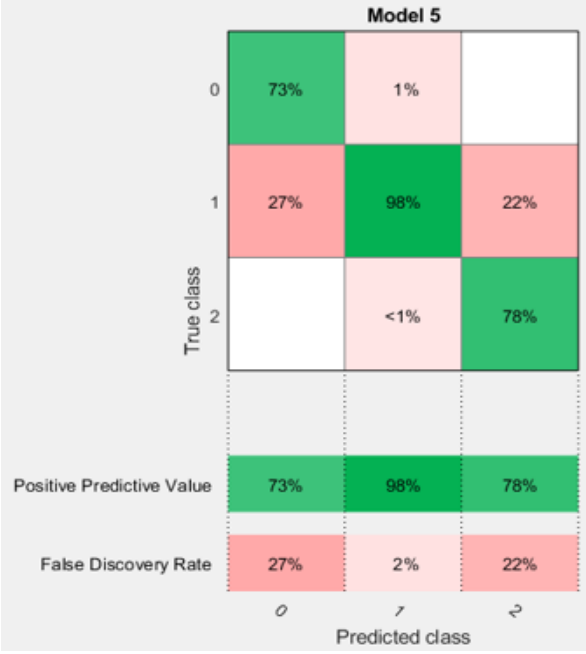

Fig.8. LDA Confusion Matrix

73% positive discovery rate was determined in classifying class 0 using Linear discriminant analysis (LDA) while 27% of false discovery rate was found. In class 1, 98% true positive was classified and 2% of them were acknowledged as false discovery rate. In class 2, 78% of true positive rate efficiency was determined and rest of the 22% was found to be false discovery rate.

IV. TESTING AND VALIDATION

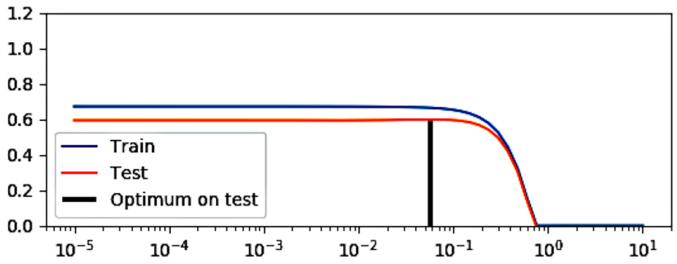

Fig.9. Training and Testing Graph

Actually the real performance of proposed algorithm for classification or regression depends upon the training and testing assessment. Sometimes proposed algorithms achieve best training score but in testing they failed as the new data is processed during testing. Here in our case Training and testing curves proved that Support vector machine (SVM) performed very well compared to the other existing classification approaches with less false alarm rate. Training and cross validation scores can be easily observed here. Cross validation curves are used to evaluate the proposed classification or regression model.

V. RESULTS AND DISCUSSION

Parameters	Linear SVM	Q-SVM	K-NN	LDA
Accuracy	97.4%	96.9%	91.7%	87.1%
Prediction Speed	5300 obs/sec	24000 obs/sec	13000 obs/sec	17000 obs/sec
Training time (sec)	7.5003 seconds	0.76163 seconds	1.3133 seconds	1.4741
Precision	0.97	0.95	0.91	0.87
Recall	0.98	0.96	0.89	0.81
F-measure	0.98	0.96	0.89	0.83

The learning process involves training with known data and then tests to make sure the training process works as planned. Results and verifications are defined in a confusion matrix that demonstrates the correct classification and misclassification of data. It also produces the accuracy of the model performance test.

VI. CONCLUSION

Flood risk prediction system have been developed using machine learning that eliminates the extra efforts involved in manual estimation and provides the desired precision for decision-making. The system has access to flood information and the weather information, therefore it calculates the event before time that can help concerned authorities plan. It has been concluded that Support Vector Machine is the best choice for numerical prediction if accuracy and precision is the highest priority.

ACKNOWLEDGMENT

Authors are extremely thankful to the Pakistan Meteorological Department (PMD) for providing the reliable data set.

REFERENCES

[1] S. Zhang, L. Lu, J. Yu and H. Zhou, "Short-term water level prediction using different artificial intelligent models," *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, Tianjin, 2016, pp. 1-6.doi: 10.1109/Agro-Geoinformatics.2016.7577678

[2] K. W. Chau, "Particle swarm optimization training algorithm for ANNs in stage prediction of ShingMun River, "*J. Hydrol.*, 2006.

[3] P.C. Nayak, B. Venkatesh, B. Krishna, and S. K. Jain, "Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach," *J. Hydrol.* ,2013.

[4] T. A. Khan, M. Alam, Z. Shahid and M. M. Suud, "Prior investigation for flash floods and hurricanes, concise capsulization of hydrological technologies and instrumentation: A survey," *2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS)*, Bangkok, 2017, pp. 1-6.

[5] L. Basin, S. France, A. Johannet, V. Taver, D. Bertin, and a Non, "Comparison Between Inverse Modelling and Data Assimilation to Estimate Rainfall from Runoff using the Mlutilayer Perceptron, "pp.1-8, 2015.

[6] G. Goodman, "Detection and Classification for unattended ground sensors.," in proceedings of information Decision and Control 99, pp. 419-424, 1999.

[7] S. Sarkar, T. Damarla and A. Ray, "Real-time activity recognition from seismic signature via multi-scale symbolic time series analysis (MSTSA)," *2015 American Control Conference (ACC)*, Chicago, IL, 2015, pp. 5818-5823.

[8] Dan Li, K. D. Wong, Yu Hen Hu and A. M. Sayeed, "Detection, classification, and tracking of targets," in *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 17-29, Mar 2002. doi: 10.1109/79.985674

[9] Mohamed Abdelkader, Mohammad Shaqura, Christian G. Claudel and Wail Gueaieb. "A UAV based system for real time flash flood monitoring in desert environments using Langrangian microsensors", *International Conference on Unmanned aircraft systems (ICUAS)*, May 28-31, 2013, Grand Hyatt Atlanta, Atlanta, GA

[10]] B. Pirenne, A. Rosenberger, E. Guillemot and R. Jenkyns, "The web-enabled awareness research network (WARN) project early earthquake and tsunami warning at Ocean Networks Canada," *2014 Oceans - St. John's*, St. John's, NL, 2014, pp. 1-7. doi: 10.1109/OCEANS.2014.7003176

[11] C. R. Barnes, M. M. R. Best, B. D. Bornhold, S. K. Juniper, B. Pirenne and P. Phibbs, "The NEPTUNE Project - a cabled ocean observatory in the NE Pacific: Overview, challenges and scientific objectives for the installation and operation of Stage I in Canadian waters," *2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies*, Tokyo, 2007, pp. 308-313.

[12] T. Ahmed Khan, K. Kadir, M. Alam, Z. Shahid, and M. Mazliham, "Identification of Flash floods using Soil Flux and CO2: An implementation of Neural Network with Less False Alarm Rate", *ijie*, vol. 10, no. 7, Nov. 2018.

[13] T. A. Khan, M. Alam, K. Kadir, Z. Shahid and S. M Mazliham, "A Novel Approach for the Investigation of Flash Floods using Soil Flux and CO2: An Implementation of MLP with Less False Alarm Rate," *2018 2nd International Conference on Smart Sensors and Application (ICSSA)*, Kuching, 2018, pp. 130-134.

[14] S. Meckelnburg, A. Jurczyk, J. Szturc, K. Osrodka, "Quantitative Precipitation Forecasts (QPF) Based on Radar Data for Hydrological Models", in cost action, 2002.

[15] ELIZABETH E. EBERT, * LAURENCE J. WILSON,1 BARBARA G. BROWN, # PERTTI NURMI," Verification of Nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project", Weather and Forecasting, 2004.

[16] T. Khan *et al.*, "Foreign objects debris (FOD) identification: A cost effective investigation of FOD with less false alarm rate," *2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, Putrajaya, 2017, pp. 1-4.

[17] B. Philips and V. Chandrasekar, "The Dallas Fort Worth urban remote sensing network," *2012 IEEE International Geoscience and Remote Sensing Symposium*, Munich, 2012, pp. 6911-6913. doi: 10.1109/IGARSS.2012.6352574

[18] Chandrasekar, V., and Coauthors, 2010: The CASA IPA test-bed after 5 years' operation: Accomplishments, breakthroughs, challenges and lessons learned. Sixth European Conf. on Radar Meteorology, Sibiu, Romania, September 6 -10,

[19] T. Khan, K. Kadir, M. Alcm, Z. Fchiihid and M. S. Mazliham, "Geomagnetic field measurement at earth surface: Flash flood forecasting using tesla meter," *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*, Kuala Lumpur, 2017, pp. 1-4.

[20] T. Khan, M. Alam, and M. Mazliham, "Artificial Intelligence Based Multi-modal Sensing for Flash Flood Investigation", *jictra*, pp. 40-47, Jun. 2018.

[21] M. Yeary, B. L. Cheong, J. M. Kurdzo, T. y. Yu and R. Palmer, "A brief overview of weather radar technologies and instrumentation," in *IEEE Instrumentation & Measurement Magazine*, vol. 17, no. 5, pp. 10-15, Oct. 2014.

[22] IEEE Standard Radar Definitions," in *IEEE Std 686-2008 (Revision of IEEE Std 686-1997)*, vol., no., pp.c1-41, May 21 2008 doi: 10.1109/IEEESTD.2008.4530766