Journal of
Engineering and
Applied Sciences

SINCE 2000
MEDWELL PUBLICATIONS

Classification of Remote Sensor Data for Flood Disaster Forecasting using Data Mining Hybrid Techniques: A Proposed Model

Hasmeda Erna Che Hamid, Nurjannatul Jannah Aqilah Md Saad, Noor Afiza Mat Razali, Muslihah Wook, Mohammad Adib Khairuddin and Mohd Nazri Ismail
National Defence University of Malaysia, Kem Perdana Sungai Besi, 57000 Kuala Lumpur, Malaysia

Key words: Dimensionality reduction, remote sensor data, classification, government, SVM

Corresponding Author:
Hasmeda Erna Che Hamid
National Defence University of Malaysia, Kem Perdana Sungai Besi, 57000 Kuala Lumpur, Malaysia

Page No.: 2542-2548
Volume: 15, Issue 12, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences
Copy Right: Medwell Publications

Abstract: Based on the National Security Council (NSC) Directive No. 20 that concern in coordinating responsible agencies and committee, the Malaysian government has established a disaster management coordination and preparedness agency. Among the natural disasters that occurred in Malaysia, floods are the most destructive. Thus, research to develop the flood forecasting model tailored to Malaysia requirements is crucially needed. Nowadays, neural network, SVM and decision tree have been used extensively as the data mining models. Support Vector Machine (SVM) is greatly popular, robust and efficient in flood modeling and prediction. SVM has been also extended as a regression tool, known as Support Vector Regression (SVR). However, the increasing volume and varying format of collecting data from remote sensing presents challenges on the efficiency of data classification for forecasting. Data that are obtained are high dimensional in nature and dimensionality reduction needs to be improved by reducing random variable in classification techniques. This research aims to propose flood disaster forecasting using data mining classification techniques by reducing random variable for efficient result in flood forecasting. This research will investigates/identify the data mining technique in disaster that being research by the researchers and proposed a conceptual model to analyze flood data. SVR will be employed to select nearby sensors and develops a linear model for a target sensor. Neural network will be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. The proposed model will be tested using data from Malaysian disaster management agencies. The result of this study shall create a new model that is expected to improve flood disaster forecasting and contribute to enhancement of early warning system and decision making during a disaster in Malaysia.

INTRODUCTION

Floods are among the most frequent natural disasters in worldwide. This flood has great impact on placement, agriculture, livelihoods and so on. Floods in 2014 have a major impact on the economy and the general public in several states, especially in Kelantan, Malaysia. It affects the long-term, severe and unpredictable and cause great loss of life and severe damage. Floods in Malaysia have been classified in two categories by the Malaysian Drainage and Irrigation Department which are flash flood and monsoon floods[1].

According to the Sendai framework program for disaster risk reduction adopted by the Third United Nation World Conference on Disaster Risk Reduction (WSCRR) in 2015, it is necessary to improve the monitoring system and better understand various types of threats as well as enhance the forecasting and early warning systems. This should increase the awareness of responding agencies and the general population[2]. The flood disaster has brought a lot of impact to the locals and its surroundings. Statistic shows that the flood events that occurred in 2005-2015 have cause over 700 thousand people had lost their lives and approximately, 23 million had been made homeless people[3].

In order to build resilience and reduce losses and damages, Sendai framework prioritize its actions in the following four areas: understanding disaster risk; strengthening disaster risk governance to manage disaster risk; investing in disaster risk reduction for resilience; enhancing disaster preparedness for effective response and to "Build Back Better" in recovery, rehabilitation and reconstruction[3]. The availability of access to early warning systems and disaster risk information and assessment to people should substantially increase by 2030. For this, governments should take into account the needs of different categories of users and data dissemination in order to enhance disaster preparedness for effective response. In addition, space and in situ information including Geographic Information Systems (GIS) are needed to be fully utilized in order to enhance disaster analysis tools and to support real-time access services of reliable disaster data.

Data analytics for disaster management and response requires a large variety of heterogeneous datasets that are related each other and show different aspects of the changes caused by a disaster. Integration of that datasets are needed and many types of sensors outputting different format of data ranging from time series data to semi structured data and textual data. Sensors in IoT applications sense the complicated environment and generate an enormous data that must be filtered and cleaned so that it can be interpreted and user will be provided with insights of the data collected in form of patterns. As for now, an applications to manage countless sensor information in Internet of Things (IoT) massively increasing due to a wide assortment of sensor gadgets on detecting layer. However, the number of sensors and the choice of topology determine the accurate prediction. It results into an improved effectiveness, exactness and better economic outcomes.

Problem statement: Flood has been a common occurrence in Malaysia and many had suffered loss financially. Without proper flood disaster management, the impacts can be severe ranging from damaging properties to endangering lives. In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated and hence redundant. This is where dimensionality reduction algorithms come into play which later can help in data compression and hence reduced storage space. It also can reduce computation time and remove redundant features, if any during forecasting process. Forecasting the flood is essential to help preparing for the best and the worst of the disaster. Therefore, a more efficient flood management system is proposed to better manage flood disasters. By leveraging the technology of data mining, real-time updates and notifications for flood events can be delivered directly to the community.

LITERATURE REVIEW

Flood disaster: Based on the National Security Council (NSC) Directive No. 2.0 that concern in coordinating responsible agencies and committee, the Malaysian government has established a disaster management coordination and preparedness agency. The agency task is to coordinate preparedness of the nation to face disaster. According to Department of Irrigation and Drainage (DID), average 143 number of floods occurred in Malaysia every year and 90% of it is a flash flood. Figure 1 shows the number of flood events occurring in Malaysia from 2001-2015.

In Malaysia, there were 76 disaster consists of various type has been recorded in the period of 1965 to 2016 and more than half are flood disaster[4]. Timely decision-making to direct and coordinate the activities of other people is important to achieve disaster management goals[5]. However, decision making in government usually takes much longer and is conducted through consultation and mutual consent of a large number of diverse personnel[6]. This may be due to standard operating procedure, top management discussion and so forth.

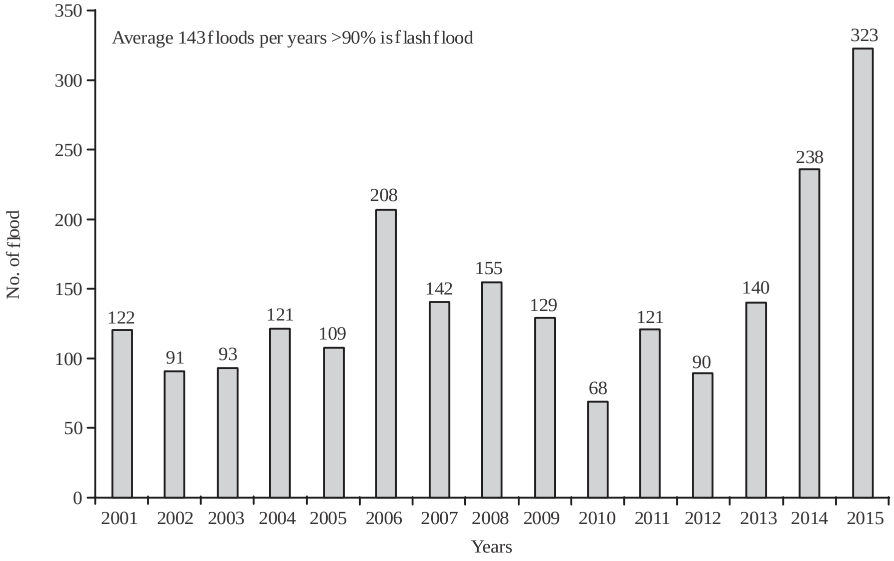

Fig. 1: Flood records between 2001-2015 in Malaysia[1]

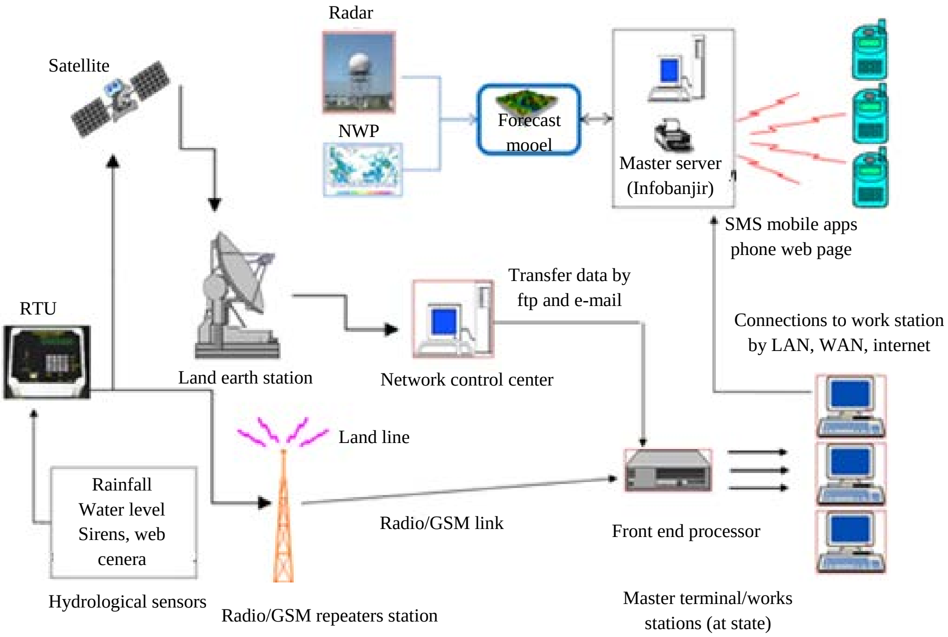

Fig. 2: System architecture for data collection and dissemination in Malaysia[1]

Figure 2 shows the system architecture for data collection and dissemination in Malaysia developed by Malaysian Centre for Remote Sensing (MACRES). MACRES is responsible to disseminate information obtained from remote sensing or other related technologies via. the early warning detection and monitoring.

Remote sensor data: A large number of network sensors are embedded into various devices and machines in the real world[7]. Sensors are used in different fields can collect a variety of data such as environmental data, geographic data, astronomical data, logistics data, etc. The vast variety of data sources present in times of a disaster creates a need for integration and aggregation of data and to make effective visualizations for it[8-10].

Sensor network has been able to be applied to save livelihood among flood affected area in recent times. Ancone, etc., accumulate IoT-based flood monitoring research in terms of efficiency, scalability and reliability[11]. An integrated weather and flood detection and notification systems are also proposed where audible alarms, Short Message Service (SMS)-based notification, web portal-based visualization and status of the flood situation is facilitated[6].

A Netduino Plus 2-based water level mentoring system is recently designed to measure the water level in a river, pond, lake and lagoon[12]. The system was developed using water level sensor to estimate the depth of water level by using sensor as an important tool which the information about the water level is sent to a local machine via. the local WiFi. Information received on local machine can be obtained by any smartphone and other digital devices.

In Malaysia environment, application developed by MACRES using sensor technology, GPS, GIS, remote sensing to obtain real time data show that the latest technology has been adapted in disaster management. However, the technology and resources was not fully utilized between organizations due to lack of tools and data management[7].

CLASSIFICATION TECHNIQUE FOR FLOOD FORECASTING

It is very important for the organizations involved in disaster management to obtain real-time disaster data that has been processed as quickly as possible in order to respond and coordinate efficiently. Big data tools and techniques can assist disaster management officials to optimize decision-making procedures[13]. Even after the disaster, the organizations should make future plans to reduce the impact of disasters and forecast the ongoing disaster events using existing historical data and information. But effective planning and management hugely depends on the quality as well as quantity of the data available[14]. Efficiently managed data will not only empower decision-makers to make accurate assessment during the disaster but also help to take suitable actions for effective disaster response and recovery as well as disaster forecasts.

Prediction is concerned with estimating the outcomes for unseen data while forecasting is a sub-discipline of prediction in which are making predictions about the future on the basis of time-series data. Weather forecasting predicts the weather in the future using temporal information. One of the challenging of forecasting is finding the number of previous events that should be considered when making predictions about the future. This also depends on whether making about the immediate or the distance future. So, for a forecasting model with exogenous inputs need to model two things; model the exogenous, non-temporal features; model the historical, temporal data. To obtain accurate forecasts, these models have to be combined judiciously such that the estimates of the temporal model are adjusted by the estimates from the feature model.

Sit and Demir[15] proposed a flood prediction benchmarking data for future applications in machine learning and scalable approaches to forecast river stage for individual survey points on rivers. Gated Recurrent Unit (GRU) networks are utilized in this study. The purpose of this study is to predict the water level for certain points in the river, the river network structure and the connection should be understood by the framework. Data used in this study consists of gage height from United States Geological Survey (USGS) and UFC sensors, NOAAs Stage IV radar rainfall and metadata and sensor information from Iowa Flood Center (IFC). The results show that an artificial neural network for the state of Iowa anticipates the stage height very close to the actual height measurements[15].

Ensemble modeling is a process of combining the predictions of single models into an integrated model to increase prediction accuracy. Choubin *et al*.[10] propose a framework for assessing flood vulnerability where only models with <80% accuracy are allowed to be used in the ensemble model. The results from this study show that the MDA Model has the highest predicted accuracy (89%), followed by the SVM (88%) and CART (0.83%) models. Sensitivity analysis shows that the percentage of slopes, drainage density and distance from the river is the most important factor in mapping the vulnerability of the flood[10]. The ensemble modeling approach often results in better classification than individual models.

Yang *et al*.[16] has proposed a time-series forecasting model based on the imputation of missing values and variable selection. This model used five imputation methods like median of nearby points, series mean, mean of nearby points, linear and regression imputation to compares the findings to estimate the missing values by using the delete strategy. Then, it lists the importance of the atmospheric variables once identifying the key variable through factor analysis. The key variable will influence the daily water levels and sequentially removes the unimportant variables. This model uses the Random Forest machine learning method to build forecasting model.

Linear Discriminant Analysis (LDA) has been used by Ponnanna *et al*.[17] in producing a model to reduce the dimensionality of the dataset independent variables. The independent variables in dataset include parameters such as temperature, humidity, atmospheric pressure, current value from the solar panel which is set to three different levels, precipitation levels, water level, the dependent variable which is possibility of a flood is divided into 3 states, namely no flood, light flooding, heavy flooding which are assigned certain values. Different model

Table 1: Comparison of a data mining techniques for weather prediction[18]

Application	Techniques	Variables	Accuracy (%)	Advantages	Disadvantages
Weather prediction for ship navigation	Decision tree	Climate, humidity, stormy, temperature	-	Verifiable performance	Do not handle continuous range data directly
Weather prediction	Decision tree	Pressure, clouds, quantity, humidity, precipitation, temperature	83	Good prediction accuracy	Data transformation is required. Extra computation required
Hourly rain-fall prediction	Decision tree	Temperature, wind direction, speed, gust, humidity, pressure	99, 93	High prediction accuracy	Small data is left for prediction
Daily rainfall prediction in river basin	Decision tree, clustering	Temperature, pressure, wind, rainfall	-	Grouping of multisite rainfall data in clusters	Small data is left for prediction. No verification is done
Weather prediction and climate change studies	Decision tree, ANN	Temperature, rainfall, evaporation, wind speed	82	Best network is selected for prediction	Accuracy varies highly with size of training data-set
Weather prediction general	ANN	Temperature, humidity, wind speed	-	Combining both gives better prediction accuracy	Attribute normalization is required
Climate prediction in Sri Lanka	ANN	Temperature, humidity, precipitation, wind speed	-	Beneficial for dynamic data	Need to integrate feature selection techniques
Meteorological data analysis	Clustering	Temperature, humidity, rain, wind speed	-	Good prediction accuracy	Dynamic data mining required
Rainfall prediction	Regression	Min and max temperature, wind direction, humidity, rainfall	63	Acceptable accuracy	Attribute elimination required for better accuracy
Short term rainfall prediction	Regression	Min and max temperature, wind direction, humidity, rainfall	52	Can work even with small dataset	Instead of accurate, an approximated value is retrieved
Drought prediction	Regression	Rainfall, sea level, humidity, temperature	-	Correlation and statistical analysis is also applied	Verification is not done

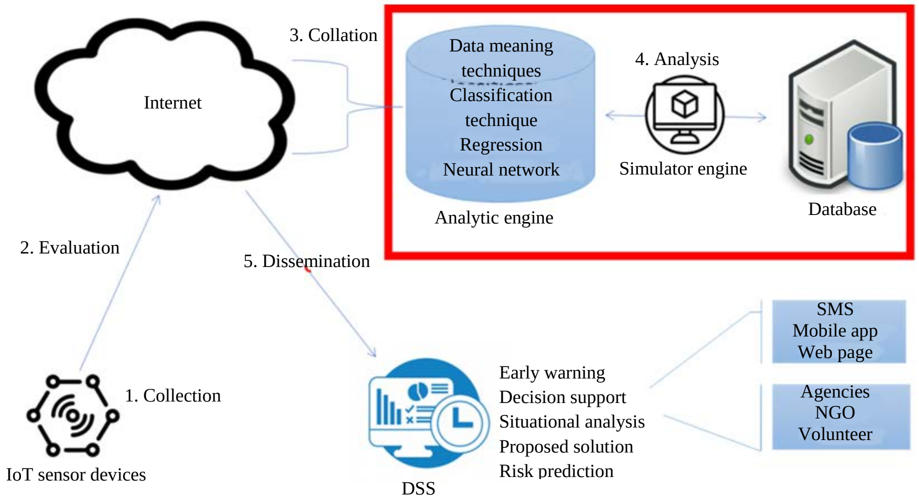

Fig. 3: Conceptual model for forecasting the floods occurring in Malaysia

classifiers are modeled to the same dataset and the best model is chosen. The constant real-time stream is taken and the new prediction values are calculated constantly based on the optimum model (Table 1).

A PROPOSED MODEL

This study proposed a conceptual model for forecasting the floods occurring in Malaysia using a hybrid classification data mining techniques. The model proposed for this study is shown in Fig. 3. The aims of this study to propose flood disaster forecasting using data mining classification techniques by reducing random variable for efficient result in flood forecasting. The most effective approach to reduce irrelevant variables is dimensionality reduction. Dimensionality reduction have two major techniques; feature selection is a process that chooses an optimal subset of features according to an

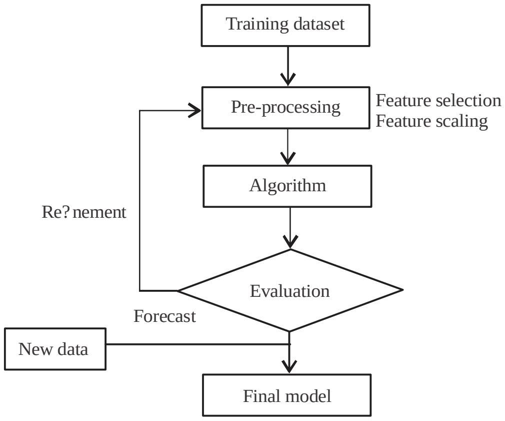

Fig. 4: Process flow for proposed conceptual model

objective function; feature extraction refers to the mapping of the original high-dimensional data into a lower-dimensional space. This study proposes to use feature selection techniques as described. Support Vector Regression (SVR) and neural network will be used based on previous studies and a few variables have been identified in the process of forecasting. Among the variables used in previous studies are temperature, rainfall, water level, pressure, wind speed, direction, humidity and sea level and most of the studies require more than four variables to get accurate and effective forecast. Therefore, this proposed model will use a water level, rainfall and wind speed as our features to ensure that the forecasts are almost accurate with existing models.

Throughout the process of data analysis using the identified variables, we will go through the process as in Fig. 4 to get the best result in flood forecasting.

CONCLUSION

This study emphasizes on reducing the variables while maintaining the accuracy and efficiency of data processing coming from sensors, flood disaster forecasting could be improve and contribute to enhancement of early warning system and decision making during a disaster in Malaysia. The accuracy and efficiency will be compared to existing model according to the variables used.

RECOMMENDATIONS

The next step, we will conduct an analysis of the data obtained using the proposed techniques according to the suggested variables based on the stated process flow. The data that have been analyzed will be validated by the suitable organization and at the same time they will validate the proposed conceptual model through interview sessions.

ACKNOWLEDGEMENT

This research is supported by Ministry of Higher Education, Malaysia under RACER FRGS Grant and Universiti Pertahanan Nasional Malaysia under Short Term grant.

REFERENCES

01. Mohd Anip, M.H. and S. Osman, 2016. Flash flood forecasting and warning in Malaysia. Proceedings of the 1st Steering Committee Meeting on Southeastern Asia-Oceania Region Flash Flood Guidance, July 10-12, 2016, Jakarta, Indonesia, pp: 1-14.
02. Bugaets, A., B. Gartsman, A. Gelfan, Y. Motovilov, O. Sokolov, L. Gonchukov and A. Kalugin *et al*., 2017. The integrated system of hydrological forecasting in the Ussuri River basin based on the ECOMAG model. Geosci., Vol. 8, 10.3390/geosciences8010005
03. Nations, U., 2015. Sendai framework for disaster risk reduction 2015-2030. United Nations Office for Disaster Risk Reduction (UNDRR), Geneva, Switzerland. https://www.unisdr.org/we/coordinate/sendai-framework
04. Amin, M.Z.M., 2016. Applying Big Data Analytics (BDA) to diagnose hydro-meteorological related risk due to climate change. GeoSmart Asia, Adelaide, Australia. https://geosmartasia.org/presentation/applying-big-data-analytics-BDA-to-diagnose-hydro-meteorological-related-risk-due-to-climate-change.pdf
05. Othman, S.H. and G. Beydoun, 2013. Model-driven disaster management. Inf. Manage., 50: 218-228.
06. Kim, G.H., S. Trimi and J.H. Chung, 2014. Big-data applications in the government sector. Commun. ACM., 57: 78-85.
07. Baharin, S.S.K., A.S. Shibghatullah and Z. Othman, 2009. Disaster management in Malaysia: An application framework of integrated routing application for emergency response management system. Proceedings of the International Conference on Soft Computing and Pattern Recognition SOCPAR'09, December 4-7, 2009, IEEE, Ayer Keroh, Malaysia, ISBN:978-1-4244-5330-6, pp: 716-719.
08. Chen, F., P. Deng, J. Wan, D. Zhang and A.V. Vasilakos *et al*., 2015. Data mining for the internet of things: Literature review and challenges. Intl. J. Distrib. Sens. Networks, 2015: 1-14.
09. Choi, S. and B. Bae, 2015. The Real-Time Monitoring System of Social Big Data for Disaster Management. In: Computer Science and its Applications: Lecture Notes in Electrical Engineering, Park, J., I. Stojmenovic, H. Jeong and G. Yi (Eds.). Springer, Berlin, Heidelberg, ISBN:978-3-662-45401-5, pp: 809-815.

10. Choubin, B., E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini and A. Mosavi, 2019. An Ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees and support vector machines. Sci. Total Environ., 651: 2087-2096.
11. Janis, I.L. and L. Mann, 1977. Emergency decision making: A theoretical analysis of responses to disaster warnings. J. Hum. Stress, 3: 35-48.
12. Kumar, A.V., B. Girish and K.R. Rajesh, 2015. Integrated weather & flood alerting system. Intl. Adv. Res. J. Sci. Eng. Technol., 2: 21-24.
13. James, L., 2018. Cracking the data science interview: The 10 Neural Network Architectures Machine Learning Researchers Need To Learn. Medium Website, USA. https://medium.com/cracking-the-data-science-interview/a-gentle-introduction-to-neural-networks-for-machine-learning-d5f3f8987786
14. Anonymous, 2019. Prediction vs forecasting. Data Science Blog, Egypt. https://www.datascienceblog.net/post/machine-learning/forecasting_vs_prediction/
15. Sit, M. and I. Demir, 2019. Decentralized flood forecasting using deep neural networks. Machine Learning, Vol. 1, 10.31223/osf.io/e9xqr
16. Yang, J.H., C.H. Cheng and C.P. Chan, 2017. A time-series water level forecasting model based on imputation and variable selection method. Comput. Intell. Neurosci., 2017: 1-11.
17. Ponnanna, P.B.L., R. Bhakthavathsalam and K. Vi-shruth, 2017. Urban flood forecast using machine learning on real time sensor data. Trans. Mach. Learn. Artif. Intell., 5: 69-75.
18. Chauhan, D. and J. Thakur, 2014. Data mining techniques for weather prediction: A review. Intl. J. Recent Innovation Trends Comput. Commun., 2: 2184-2189.