Wireless Networks (2021) 27:5429–5434
https://doi.org/10.1007/s11276-019-02168-3

Study of short term rain forecasting using machine learning based approach

M. S. Balamurugan[1] · R. Manojkumar[1]

Published online: 26 October 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
Weather forecasting has been still dependent on statistical and numerical analysis in most part of the world. Though statistical and numerical analysis provides better results, it highly depends on stable historical relationships with the predict and predicting value of the predict and at a future time. On the other hand, machine learning explores new algorithmic approaches in prediction which is based on data-driven prediction. Climatic changes for a location are dependent on variable factors like temperature, precipitation, atmospheric pressure, humidity, wind speed and combination of other such factors which are variable in nature. Since climatic changes are location-based statistical and numerical approaches result in failure at times and needs an alternate method like machine learning based study of understanding about the weather forecast. In this study it has been observed that percentage in departure of rainfall has been ranging from 46 to 91% for the month of June 2019 as per Indian Meteorological Department (IMD) by using the traditional forecasting methods, but whereas based on the following study implemented using machine learning it has been observed that forecast was able to achieve much better rainfall prediction comparative to statistical methods.

Keywords Internet of things (IoT) · LoRaWAN · Artificial neural network · Logistic Regression · Binary classification

1 Introduction

Weather forecasting is an act of deploying scientific calculations based on the climatic condition by using the latest technologies and supercomputers to predict the outcome. In past millennium also people have attempted to predict weather using hand-based tools like deploying and measuring the changes in barometric pressure, weather and cloud conditions [1]. Nowadays based on the quantitative data available with us, forecasting the weather was practiced using latest tools available. The changes in the atmospheric conditions, is recorded as initial and boundary conditions for the respective mathematical methods and based on history of data weather forecasting is done. At present, weather forecasting relies on computer-generated models and the following factors: (1) Past weather data (2) Present atmospheric conditions (3) Pattern of climatic conditions based on history. But here atmospheric conditions are chaotic, which is non-linear in nature. All the past methods of weather forecasting whether one or the other will mostly rely on fitting the data and predicting an outcome or a range of outcome. Forecasting methods that rely on linear characteristics but whereas their input parameters are non-linear in nature will forecast inaccurate results.

The following are the methods that are widely used for weather forecasting (1) Statistical methods, (2) Numerical weather prediction (NWP) and (3) Ensemble weather forecasting. Statistics and Numerical Methods are the most chosen methods of study that has been used in weather forecasting [2, 3] by the respective agencies in their countries. Statistics rely on carefully chosen specific predictive model for forecasting the respective weather parameter. In summary and short it can be refined as identifying the best fit for the model and the given data. Statistical methods in simple suffers identifying best predictors [4].

✉ M. S. Balamurugan
balamurugan.ms@vit.ac.in

R. Manojkumar
manojkumar.r@vit.ac.in

[1] School of Electronics Engineering, Vellore Institute of Technology, Chennai, India

Statistics draws inference from a sample [5, 6], which translates to statistical methods assumes that future will be a repetition of past weather data. Based on past weather data a study is carried out to find patterns of that may relate to future events. Based on these relationships, and present weather conditions future conditions are arrived at. Statistical Methods are of great importance particularly for long range weather forecasting [7]. Sometimes the NWP aids in statistical analysis [8] or this method supplements NWP. Statistical Methods are widely successful because of its project specific probabilistic model.

Ensemble weather forecasting is the method of forecasting the weather based on different initial conditions the numerical weather forecasting model is run multiple times to attain a most likely outcome for the uncertainties in the weather forecasting [9]. The ensemble methods use more than one model for the different combination of physical parameterizations schemes [10]. This is because of the complexity involved in computations, an error introduced during the imperfect initial conditions and due to the chaotic nature of atmospheric conditions. So, using this method a range of various parameter are studied and observed as result. This method has been successfully deployed in multiple weather prediction facilities right from US, UK, India, Brazil, Australia etc., However here since initial condition of chaotic nature of atmosphere plays a vital role in forecasting models, slight changes in initial condition will have significant impact in the output of the model. That is why in ensemble forecasting a range of possible outcomes is being obtained.

NWP models depend on mathematical modelling of atmospheric and oceanic data sets, and further on the current climatic factors [11, 12]. Climatic conditions are usually non-linear in nature. NWP employs (1) set of equations as defined in science, (2) numerical methods, (3) parameterizations of other physical processes. The set of equations that are represented in NWP from science are Conservation of momentum, Ideal gas law, Conservation of Energy and Mass. Numerical methods representations are nothing but data which are derived from spatial and temporal derivative into numerical methods that can be modelled as scientific equations [13] which can be further processed by computers. The physical process here refers to various factors like atmospheric conditions, radiations from the earth surface and incoming Solar radiations, Topography, Evaporation, Rain, clouds and other such factors. All these factors which cannot be directly equipped are referred to using parametric equations and finally referred in NWP as parameterizations. Also, these weather models consider current atmospheric conditions as initial conditions to apply in numerical models.

In addition to the above mentioned methods there are time series methods which is based on the rule that future is based on the history of trends except for any unforeseen circumstances. The climatology methods observe inference based on averaging the weather statistics observed over multiple years. The analog method which predicts weather based on identifying an occurrence in the past which is like the present conditions and predicting the outcome.

From Fig. 1 it is observed that variation is indeed very high during the January to April and again rises above 50 during November to December. These values illustrate that there needs more improvised study about weather forecasting. Figure 2 shows percentage of deviation in rainfall from 2011 to 2014 for various districts of the state which shows the range of rainfall forecasting to be around − 50–80%. In summary chaotic nature of atmosphere, complexity involved in numerical weather predictions, assumptions of initial conditions which are prone to error, issues involved in parameterizations makes the forecast less accurate. It has been also established by Pulak [14] from IMD that no model except the ANN predicts the long-range forecasting so accurately districts wise. ANN's also work effectively for a non-liner input condition [15, 16].

So, the objective of this work is to analyze the significance of machine learning approaches over other traditional methods in weather forecasting. In this scenario if we can apply machine learning algorithms for forecasting which can be modeled using knowledge of non-linear input characteristics the results can be much accurate than the traditional methods. machine learning models can establish the structural relationships between various parameters in study [17, 18]. In this work a comparison study of Logistic Regression, Decision Tree and Random Forest algorithm are executed to forecast rain in comparison with Statistical approach.

2 Related works

Dating back to 1995 there has been considerable work using Artificial Neural Networks in Weather forecasting especially for Rain Forecasting. In the past there are various machine learning based approaches attempted to build a model for weather forecasting. ARIMA models using shallow models for studying precipitation has been attempted Robert et al. [19] where 6-h precipitation forecasting prediction was validated successfully in comparison with linear regression based approaches. Physical Parameterization of climatology models have been attempted using ANN by Vladimir et al. [20] as well as by Chen et al. [21–23]. These authors studied various models in ANN like ARIMA, Random Forest algorithm. The observation was that for short term forecasting the model performance was better for long term forecast these model performances suffered in terms of variance. It needed

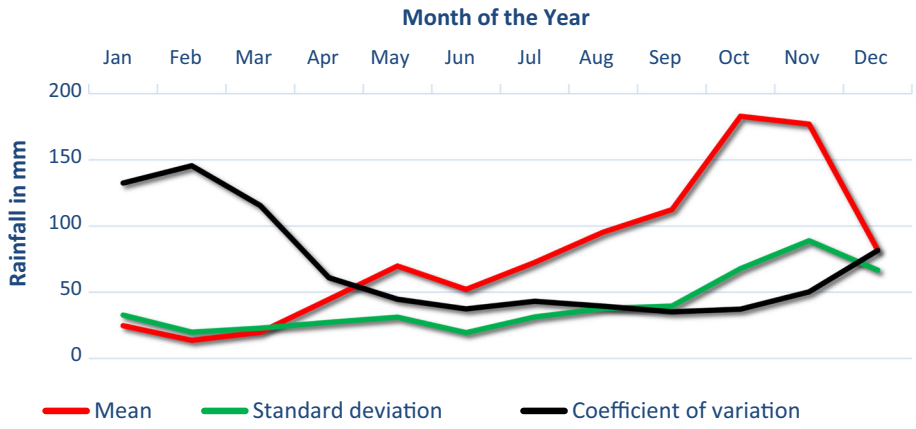

Fig. 1 Rainfall statistics with mean, standard deviation and co-efficient for variation of rainfall for the year 1901–2015

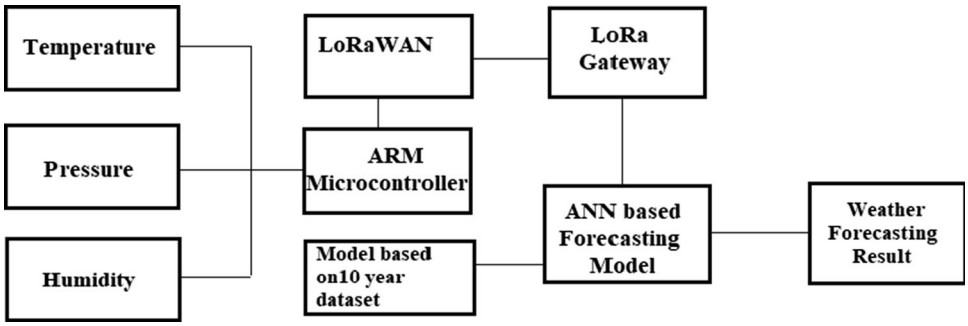

Fig. 2 Experimental setup using Lora WAN to acquire weather data

combination of models with NWP for better forecasting. Prasad et al. [24], used Multi-predictor Logistic Regression model for probabilistic forecasting of weather three regions of India. To study rainfall a convolutional approach using LSTM [25] based methods has been applied by Xingjian [26] This model shows better performance in capturing the extreme rainfall events of India. There has been number of works in predicting various weather parameters like wind, precipitation, temperature, rainfall etc., using various machine learning and ANN based algorithms [27].

3 Experimental setup and data sets

3.1 Experimental setup

A node to monitor temperature, atmospheric pressure and humidity was setup using at Vellore Institute Technology, Chennai campus with Latitude 12.84100876610025 and Longitude 80.1538896560669. The system was setup using a ST Microcontroller processor running on ARM M4 as shown in Fig. 3. The data processed by this system is wirelessly communicated to a remote server using a LoRaWAN, which uses sub 1 GHz spectrum to transfer data wirelessly without the need of backhaul internet. The data is received in the remote server using a nano gateway enabled to receive LoRaWAN packets. LoRa achieves long range communication dues to its implementation of chirp spread spectrum modulation, which is widely used in military based applications [28]. A single gateway using LoRa can cover practically 2 km distance [29], but the coverage can be further increased, but depends on factors like obstruction. LoRaWAN's link budget is practically higher than any communication technology that is standardized.

3.2 Data specification

The data sets are retrieved from open data portal [15]. The objective of the work is to train a system which forecasts various weather parameters. The node measures temperature, pressure and humidity using a DHT11 and BMP 180 sensor interfaced to an ARM based microcontroller. The monitored data is sent through a LoRa based gateway to server. The server is running a model based on the datasets from open portal which has data of 10 years with 1-day interval of data. The datasets contain data of 10 years in the form of Maximum and Minimum Temperature in degree Celsius at 9 AM and 3 PM, Rainfall in mm, Direction of Wind at 3 PM, Wind speed at 9AM at 3 PM, Humidity at 9 AM and 3 PM, Atmospheric Temperature at 9 AM and 3PM, and whether it rained today and tomorrow. The input datasets have 1,23,328 data samples from 2008 to 2017.

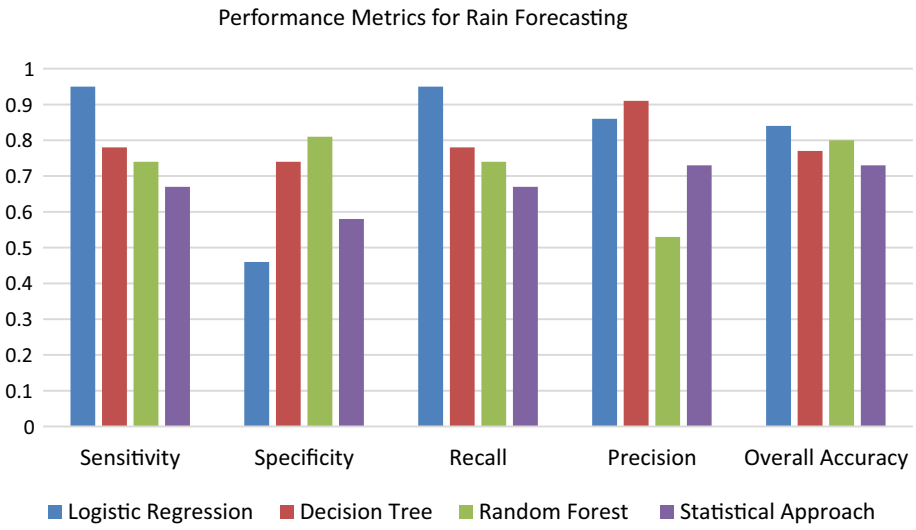

Fig. 3 Comparison of performance metrics of Logistic Regression, Decision Tree, Random Forest and Statistical Approach

3.3 Neural network models under evaluation

3.3.1 Logistic Regression

The logistic function which can also be referred as sigmoid function that takes any real number as an input and maps it to the binary value 0 to 1, using 1/(1 + e^(−value)) [30]. Binary classification gives more qualitative prediction that multiple linear regression. For decision making a simple Binary logistics classification based neural network is applied since the statistically modelled input data belonging to different data sets are being applied. Unlike Linear regression where the output is modelled to numerical value the logistic function described above models the output between 0 to 1(binary values). However Logistic Regression can be represented as

$$y = \frac{e^x}{1 + e^x} \tag{1}$$

where y is the predicted output and any real value can be fitted into this logistic function to be transformed into a binary output. Logistic Regression model as in Eq. 2

$$\pi = \begin{cases} 0, & \beta_0 + \beta_1 X < 0 \\ \beta_0 + \beta_1 X, & 0 \le \beta_0 + \beta_1 X \le 1 \\ 1, & \beta_0 + \beta_1 X > 1 \end{cases} \tag{2}$$

where π denotes the probability of that $P(Y_i = 1|X_i = x)$. But however here due to abrupt changes in slopes, this function is further transformed into $\pi_i = P(\beta_0 + \beta_i x_i)$. The S curve can be modelled for prediction of Y as $P(Y = 1|X = x) = \frac{1}{1+e^{-x}}$. So using this Logistic Regression any real value between $-\infty$ to $+\infty$ can be fit into f(x) having either 0 or 1. Since the intention is to understand whether there will an event of occurrence of rain or not of all the Logistic Regression Binary Logistic Regression is preferred since the output is either yes or no.

3.3.2 Decision Tree

The Decision Tree uses Classification and Regression Tree (CART) and always classifies binary based approach. Here Gini Index is a measure of impurities at the node. If a sample is completely homogenous it is considered less impure. If sample is equally split, then it is considered more impure.

$$i(t) = Gini(t) = \sum_{j=1}^{J} P(j|t) * (1 - P|t))$$

*where P(j|t) is the proportion of category at node t.

The changes in impurities are measured using $[i(t) - P_L * i(t(L)) - P_R * i(t(R))]$, where P_L Proportion of observation in the left branch and P_R is the Proportion of observation in Right Branch.

3.3.3 Random Forest

Random Forest is another form of Classifier which exhibits performance like Decision Tree. The important aspect of Random Forest Algorithm is that it can handle larger attributes. The Trees here are created using the following strategy [19]:

a. Each Tree's root node has a bootstrap data.
b. Using a best split method subset of variables is selected in a random order.
c. Without pruning the tree is grown to its maximum extent possible.
d. Upon all the trees grown using this algorithm new instances are attached to all these trees.

e. A prediction is run using a voting process to select a classification with maximum votes.

4 Results and discussion

Based on the experiment conducted and model arrived at for study of rain forecasting, the models are compared with statistical approach and machine learning Logistic Regression based approach to predict whether it will rain tomorrow based on the monitored parameters. Statistical Approach was carried out using the same Logit function used in machine learning based Logistic Regression. The Maximization of the likelihood function was carried out using the Newton–Raphson Algorithm. In the machine learning model for Logistic Regression the epoch size was set to 100. Prediction results in machine learning are studied using Confusion Matrix with performance measures compared by the parameters like sensitivity, specificity, Recall and Precision and Overall Accuracy. Overall accuracy is the measure of how best the model classifies the true values to be true out of true and false scenarios. The Overall accuracy of Logistic Regression is observed as 84% in comparison to 72.6% of statistical approach through ROC and 77% using Decision Tree. The True Positive rate—Sensitivity is measured at 0.95 using Logistic Regression in comparison to 0.76 using Statistical methods and 0.76 using Decision Tree which indicates that Logistic Regression based approach predicts the outcome Rain at 95% accuracy. However, the performance of Decision Tree was little better in comparison to Logistic Regression and statistical methods in terms of precision. The precision using Decision tree was 0.91, whereas using Logistic Regression was 0.86 and using statistical methods it was 0.72.

The Root Mean Square Error reported by each of the function is shown in Table 1.

5 Conclusion

One of the key objectives of this work is to examine the results of machine learning based models to study the rain forecasting in comparison to existing methods. It is observed through the results that machine learning models fare better in short term rain forecasting of rain as rain and also has minimal error as observed in RMSE calculation. The model was deployed in a real time node set up using a Lora WAN and forecasting was done using Logistic Regression to find the probability of Rain. We also find that during this study we were able to forecast better results using machine learning but still it is observed that forecasting machine learning based approach works good only 2-day rain forecasting beyond which the forecasting throws lot of deviation. The future work will be carried out using Deep Learning approach to study the relation between the parameters hour wise which may more insights. The works of the results has been published in the following github link: https://github.com/balams81/Rain.

Table 1 Summary of root mean square error of machine learning models in comparison

Method	RMSE
Logistic Regression	0.1126
Decision Tree	0.1126
Random Forest	0.1742
Statistics	0.2346

References

1. Iseh, A. J., & Woma, T. Y. (2013). Weather forecasting models, methods and applications. *International Journal of Engineering Research & Technology, 2*(12), 1945–1956. **(ISSN: 2278-0181)**.
2. Ren, Y., Suganthan, P. N., & Srikanth, N. (2015). A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods. *IEEE Transactions on Sustainable Energy, 6*(1), 236–244. https://doi.org/10.1109/TSTE.2014.2365580.
3. Anderson, J., van den Dool, H., Barnston, A., Chen, W., Stern, W., & Ploshay, J. (1999). Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bulletin of the American Meteorological Society, 80*(7), 1349–1362. https://doi.org/10.1175/1520-0477.
4. Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the fourteenth conference on uncertainty in artificial intelligence*, UAI'98, 43–52, ISBN: 1-55860-555-X.
5. Bzdok, D., Altman, N., & Krzywinski, Martin. (2018). Statistics versus machine learning. *Nature Methods, 15*(4), 233–234. https://doi.org/10.1038/nmeth.464210.1038/nmeth.4642.
6. Zwiers, F. W., & Von Storch, H. (2004). On the role of statistics in climate research. *International Journal of Climatology, 24,* 665–680. https://doi.org/10.1002/joc.1027.
7. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *Plos One, 13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889.
8. Lorenc, A. C. (1986). Analysis methods of numerical weather prediction. *Royal Meteorological Society, 112*(474), 1177–1194. https://doi.org/10.1002/qj.49711247414.
9. Kwon, I-H., English, S., Bell, W., Potthast, R., Collard, A., & Ruston, B. (2018). Assessment of progress and status of data assimilation in numerical weather prediction, Bulletin of the American Meteorological Society, Vol 99, no. 5, ES75-ES79, https://doi.org/10.1175/BAMS-D-17-0266.1.
10. Herman, G. R., & Schumacher, R. S. (2018). "Dendrology'' in numerical weather prediction: What random forests and Logistic Regression tell us about forecasting extreme precipitation. *Monthly Weather Review*, *146*, 1785–1812. https://doi.org/10.1175/MWR-D-17-0307.1.

11. https://doi.org/10.1109/iecon.2001.976448.
12. Lorenz, E. N. (1970). Climatic change as a mathematical problem. *Journal of Applied Meteorology, 9*(3), 325–329. https://doi.org/10.1175/1520-0450.
13. Lorenz, E. N., Mitchell, J. M. (1968), Causes of climatic change: A collection of papers derived from the INQUA—NCAR. In *Symposium on causes of climatic change*, https://doi.org/10.1007/978-1-935704-38-6_1.
14. Shrivastava, G., Karmakar, S., Kowar, M., & Guhatakurta, P. (2012). Application of artificial neural networks in weather forecasting: A comprehensive literature review. *International Journal of Computer Applications, 51*(18), 17–29.
15. Somasundar, Ministry of earth science, Data.gov.in, Jan 2017.
16. Crosby, D. S., & Ferraro, R. (1995). Estimating the probability of rain in an SSM/I FOV using logistic regression. *Journal of Applied Meteorology, 34,* 2476–2480.
17. Tetko, I. V., Livingstone, D. J., & Luik, A. I. (1995). Neural network studies comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences, 35,* 826–833. https://doi.org/10.1021/ci00027a006.
18. Lam, H. K., Ling, S. H., Leung, F. H. F., & Tam, P. K. S. (2001). Tuning of the structure and parameters of neural network using an improved genetic algorithm. In *IECON'01. 27th annual conference of the IEEE industrial electronics society (Cat. No.37243)*, Denver, CO, USA (Vol. 1, pp. 25–30).
19. Jiang, Y., Cukic, B., Menzies T., & Bartlow, N. Comparing design and code metrics for software quality prediction. In *Proc. Fourth Int. workshop on predictor models in software engineering, PROMISE'08*, New York, USA, 2008, pp. 11–18.
20. Radhika, Y., & Shashi, M. (2009). Atmospheric temperature prediction using support vector machines. *Int. J. Comput. Theory Eng., 1*(1), 55.
21. Krasnopolsky, V. M., & Fox-Rabinovitz, M. S. (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Network., 19*(2), 122–134.
22. Chen, L., Lai, X. Comparison between arima and ann models used in short-term wind speed forecasting. In Power and energy engineering conference (APPEEC), 2011 Asia-Pacific, IEEE, 2011, pp. 1–4.
23. Kuligowski, R. J., & Barros, A. P. (1998). Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather Forecast, 13*(4), 1194–1204.
24. Prasad, K., Dash, S. K., & Mohanty, U. C. (2010). A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts. *International Journal of Climatology, 30,* 1577–1588. https://doi.org/10.1002/joc.2019
25. Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, 2015 (pp. 802–810).
26. Mcgovern, A., Supinie, T., Gagne, I., Collier, M., Brown, R., Basara, J., Williams, J. Understanding severe weather processes through spatiotemporal relational random forests. In *2010 NASA conference on intelligent data understanding*, 2010.
27. Kumar Abhishek, M. P., Singh, S. G., & Anand, A. (2010). Weather forecasting model using artificial neural network. *Elsevier Procedia Technology, 4,* 311–318.
28. Franzke, C. L., O'Kane, T. J., Berner, J., Williams, P. D., & Lucarini, V. (2015). Stochastic climate theory and modeling. *WIREs Climate Change, 6,* 63–78. https://doi.org/10.1002/wcc.318.
29. Abd Rahman, N. H., Yamada, Y., Husni, M. H., Abdul Aziz N. H. (2018). Feasibility of LoRa implementation for remote weather monitoring system through field measurement and case study analysis. *International Journal of Integrated Engineering*, Vol. 10, No. 7 **(ISSN: 2229-838X)**
30. Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications, 16*(3), 361–368.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

M. S. Balamurugan is working as Assistant Professor in Vellore Institute of Technology, Chennai and has over 10+ years of experience in Industry, Research and Academic. He has published over 10+ papers and has been actively working in the area of Internet of Things.

R. Manojkumar is working as Associate Professor in Vellore Institute of Technology, Chennai and has over 10+ years of experience in Research and Academic. He completed his Ph.D. from TELECOM Sud-Paris, France in 2012 and has been serving in Vellore Institute of Technology, Chennai since then. He has been working as Research Assistant in IIT Kanpur. He has published various research papers in the area of Humar Computer Interaction and Computer Vision.