International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 9, ISSN No. 2455-2143, Pages 365-371
Published Online January 2020 in IJEAST (http://www.ijeast.com)

RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES AND AN ANALYSIS OF THE OUTCOMES OF THESE TECHNIQUES

Mr. Sandeep B
Asst. Prof, Dept. of CSE,
JNNCE, Shimoga, Karnataka, India

Ms. Jahnavi K.S
Dept. of CSE,
JNNCE, Shimoga, Karnataka, India

Abstract—**Rainfall prediction is one of the challenging tasks in weather forecasting. Accurate and timely rainfall prediction can be very helpful to take effective security measures in advance regarding: ongoing construction projects, transportation activities, agricultural tasks, flight operations and flood situation, etc. Monsoon prediction is clearly of great importance for India. Two types of rainfall predictions can be done, They are - Long term predictions: Predict rainfall over few weeks/months in advance. - Short term predictions: Predict rainfall a few days in advance in specific locations. Indian meteorological department provides forecasting data required for the prediction. This system is designed to work on long term predictions of rainfall. The main motive behind the development of this model is to predict the amount of rainfall in a particular division or state well in advance. We predict the amount of rainfall using past data.**

Keywords—***forecasting, security, projects, transportation, prediction, long term, short term, data.***

I. INTRODUCTION

Accurate forecasting of rainfall has been one of the most important issues in hydrological research because early warnings of severe weather can help prevent casualties and damages caused by natural disasters, if timely and accurately forecasted. To construct a predictive system for accurate rainfall, forecasting is one of the greatest challenges to researchers from diverse fields such as weather data mining (Yang et al., 2007), environmental machine learning (Hong, 2008), operational hydrology (Li and Lai, 2004), and statistical forecasting (Pucheta et al., 2009). A common question in these problems is how one can analyse the past and use future prediction. The parameters that are required to predict rainfall are enormously complex and subtle even for a short term period. Physical processes in rainfall are generally composed of a number of sub-processes. A accurate modelling of rainfall by a single global model is sometimes not possible (Solomatine and Ostfeld, 2008). To overcome this difficulty, the concept of modular modelling and combining different models has attracted more attention recently in rainfall forecasting. In modular models, several sub-processes are first identified, and then separate models (also called local or expert models) are established for each of them (Solomatine and Ostfeld, 2008). So far, various modular models have been proposed, depending on soft or hard splitting of training data. Soft splitting means that the dataset can be overlapped, and the overall forecasting output is the weighted average of each local model (Shrestha and Solomatine, 2006; Wu et al., 2008).

II. RELATED WORK

Reseachers have been working to improve the accuracy of rainfall prediction by optimizing and integrating Machine Learning techniques. Some of the selected studies are discussed in this section.

In [1] S. Zhang, L. Lu, J. Yu, and H. Zhou performed a comparative analysis of Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Adaptive Neuro Fuzzy Inference System (ANFIS) on rainfall prediction. The authors have compared the prediction models in four terms: (i) by using different lags as modeling inputs; (ii) by using training data of heavy rainfall events only; (iii) performance of forecasting for 1 hour to 6 hours and; (iv) performance analysis in peak values and all values. According to results ANN performed better when trained with dataset of heavy rainfall. For 1 to 4 hour ahead forecasting, the previous 2-hour input data was suggested for all three modeling techniques

(ANN, SVM and ANFIS). ANFIS reflected better ability in avoiding information noise by using different lags of inputs. And finally during peak values, SVM proved to be more robust under extreme typhoon events.

In [2], S. Zainudin, D. S. Jasim, and A. A. Bakar performed a comparative analysis of various data mining techniques for rainfall prediction in Malaysia such as: Random Forest, Support Vector Machine, Naive Bayes, Neural Network, and Decision Tree. For this experiment, dataset was obtained from various weather stations in Selangor, Malaysia. Before classification process, Pre-processing tasks were applied to deal with the noise and missing values in dataset. The results showed significant performance of Random Forest as it correctly classified large amount of instances with small amount of training data.

In [3], D. Nayak, A. Mahapatra, and P. Mishra performed a survey on various Neural Network architectures which were used for rainfall prediction in last 25 years. The authors highlighted that most of the researchers got significant results in rainfall prediction by using Propagation Network, moreover the forecasting techniques which used SVM, MLP, BPN, RBFN, and SOM are more suitable than other statistical and numerical techniques. Some limitations have also been highlighted.

In [4] B. K. Rani and A. Govardhan used Artificial Neural Network for rainfall prediction in Thailand. They used Back Propagation Neural Network for prediction which reported an acceptable accuracy. For future direction it was suggested that few additional features would be included in input data for rainfall prediction such as Sea Surface Temperature for the areas around Andhra Pradesh and Southern part of India.

In [5], N. Tyagi and A. Kumar predicted monthly rainfall by using Back Propagation, Radial Basis Function and Neural Network. For prediction, the dataset was collected from Coonoor region in Nilgiri district (Tamil Nadu). Performance was evaluated in terms of Mean Square Error. According to results higher accuracy was reported in Radial Basis Function Neural Network with smaller Mean Square Error. Moreover the researchers also used these techniques for future rainfall prediction.

In [6], N. Solanki and G. P. B presented a Hybrid Intelligent System by integrating Artificial Neural Network and Genetic Algorithm. In ANN, MLP works as the Data Mining engine to perform predictions whereas the Genetic Algorithm was utilized for inputs, the connection structure between the inputs, the output layers and to make the training of Neural Network more effective.

In [7], C. S. Thirumalai, discussed rainfall pace in previous years with respect to various crops seasons like rabi, Kharif, zaid and then predicted (rainfall) for future seasons via Linear Regression Method. For prediction, input dataset was selected according to particular corps seasons of previous years.

In [8], N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay one month and two month forecasting models were developed for rainfall prediction by using Artificial Neural Network (ANN). The input dataset was selected from multiple stations in North India, spanned on past 141 years. Feed Forward Neural Network using Back Propagation and Levenberg-Marquardt training function were used in these models. Performance of both models was evaluated by using Regression Analysis, Mean Square Error and Magnitude of Relative Error. The results showed that one month forecasting model can predict the rainfall more accurately than two month forecasting model.

In [9], H. Vathsala and S. G. Koolagudi presented an algorithm by integrating Data Mining and Statistical Techniques. The proposed technique predicted the rainfall in five different categories such as: Flood, Excess, Normal, Deficit and Drought. The predictors were selected with highest confidence level, based on association rules and derived from local and global environment. From local environment: wind speed, sea level pressure, maximum temperature, and minimum temperature were taken. From global environment: Indian ocean dipole conditions and southern oscillation were taken.

In [10], R. Venkata Ramana, et al, predicted the rainfall by using proposed Wavelet Neural Network Model (WNN), an integration of Wavelet Technique and Artificial Neural Network (ANN). To analyze the performance, monthly rainfall prediction was performed with both the techniques (WNN and ANN) by using dataset of Darjeeling rain gauge station in India. Statistical techniques were used for performance evaluation and according to results WNN performed better than ANN.

In [11], M. P. Darji et al, provided a detailed survey and performed a comparative analysis of various neural networks on rainfall forecasting. According to survey RNN, FFNN, and TDNN are suitable for rainfall prediction as compared to other statistical and numerical forecasting methods. Moreover TDNN, FFNN and lag FFNN performed well for yearly, monthly and weekly rainfall forecasting respectively. This research also discussed the various measures of accuracy used by different researchers to evaluate the ANN's performance.

In [12], Sharma *et al*, proposed Bayesian network model for mean monthly rainfall prediction of 21 stations in Assam, India. This work can be useful for better management of water resources. Monthly data of 20 years from 1981 to 2000 for all the atmospheric parameters is used for this study which was taken from different sources. Rainfall at a station is taken as a variable for this model and dependencies between rainfalls at different station is shown by Bayesian network. In this work, the author used K2 algorithm and conditional probability is found using maximum likelihood approximations. Five different atmospheric parameters viz. Temperature, Cloud cover, Relative humidity, Wind speed and Southern Oscillation Index (SOI) are used. The results revealed that temperature is found most efficient and wind speed least. SOI is also found important in improving the results. Some station got efficiency above 95% whereas other station got satisfactory results.

In [13], Akash D Dubey, proposed a rainfall prediction model using artificial neural networks (ANN). In this work the author has used the weather data of Pondicherry, India. Three different training algorithms viz. feed-forward back propagation algorithm, layer recurrent algorithm and feed-forward distributed time delay algorithm were used to create ANN models and keeping number of neurons for all the models to 20. Of all the algorithms, the results showed that feed-forward distributed time delay algorithm has best accuracy and MSE value as low as 0.0083.

III. DATA COLLECTION

For the prediction model, weather data of India, This dataset has average rainfall from 1951-2000 for each district, for every month. The raw weather data collected consists of nine measured attributes which are date, temperature (high, low, average) in ºc , Dew point (high, low, average) in ºc , Humidity (high, low, average) in % age, sea level pressure (high, low, average) in hPa, visibility (high, low, average) in Km, wind (high, low, average) in Km/h, precipitation (high, low, average) in mm, Events (Rainfall snow, thunderstorm, fog). For this work out of these 9 features we have used the Average temperature, Average Humidity, Average sea level pressure, Average wind and Events features as shown in table I. We have ignored less relevant features in the dataset for better model computation and prediction.

Table-1. Weather Data Description

Attribute	Type	Description
Temperature	**Numerical**	**Temp is in ºC**
Humidity	**Numerical**	**Humidity in Percentage**
Sea Level Pressure	**Numerical**	**Sea Level Pressure in hPa**
Windy	**Numerical**	**Wind Speed in km/h**
Events	**Numerical**	**Rainfall in mm**

A. Data Preprocessing and Data Cleaning

The main challenge in weather prediction is the poor data quality and selection. For this reason preprocessing of data is carefully done to obtain accurate and correct prediction results. In this phase unwanted data or noise is removed from the collected data set which is done by removing the unwanted attributes and keeping the most relevant attributes that help in better prediction. Another major issue that is to be rectified is the missing values in the collected data set. Missing values in the data set is filled by using various techniques. The missing values for attributes in the dataset are replaced with the modes and means based on existing data. Adding the missing values provides a more complete dataset for the classifiers to be trained.

IV. RESEARCH METHODOLOGY

There are two main types of Machine Learning approaches; supervised learning and unsupervised learning. Supervised learning algorithms are used for building predictive models. The Classification algorithms ANN, Logistic Regression, Naïve Bayes, and RandomForest are experimentally implemented and compared against each other.

A. Artificial Neural Networks -

A neural network is an immense distributed processor which works in a parallel methods and consists of simple processing units which store empirical knowledge and have it ready for active use. The functionality of neural network is analogous to the human brain, in the sense that they can receive inputs, and process the information using various computing nodes. A relevant output is produced based upon the desired application.

The main advantage of Neural Networks is its ability to display non-linearity existence between the input and output variables.

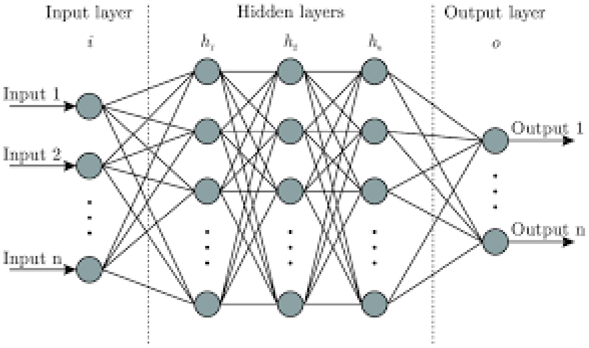

Fig. 1. Neural Network

Major concerns with neural networks

1) **Number of Hidden Layers and Nodes:** The general Artificial Neural Network depicted in Fig. 1 consists of 3 typical layers, which are the Input, Hidden, and Output Layer. Now, the layer that we are concerned with is the Hidden Layer. This layer is mainly responsible for the calculations that occur with neural networks and is also the layer where actual nonlinear mapping between input and output takes place. If any incorrect step were to be taken in this layer, the entire result of the neural network would be deviated and may have devastating effects on prediction. Due to this, we must consider how many hidden layers we need and the nodes within them. Such parameters can boost the accuracy of the entire network.

2) **Overfitting:** Another major problem that exists within the neural network is called overfitting. It hinders the neural network when trying to a draw generalization for a specific given input. When there are a high number of input parameters that are being fed to the neural network during its training phase, it can cause a generalization error. The model doesn't know which specific class to classify the data in. While, if we don't provide it sufficient data, an underestimation could occur which leads to worse approximation outputs.

3) **Selection of Activation Function:** One of the main concerns when using neural networks is selecting an appropriate activation function. The activation function in neural network plays a vital role in determining the behavior of the entire neural network. The main job of the activation function is to classify useful information and remove noise that is found in the current set. The activation function is also responsible for calculating the weights of the given inputs at each node. This allows the function to determine if a neuron can be activated or not. There are many activation functions that can be used for forecasting models such as Binary step, Sigmoid, Softmax, Relu, and Tanh. However, choosing the specific activation function is solely dependent upon the problem. For example, The Sigmoid function is better suited towards binary classifications tasks while the Softmax function is geared towards multi-classifications tasks. The reoccurring issue for all these systems is the gradient drift on the neural network. Sometimes, the gradients are too steep in a specific direction and other times it can be too low or zero. This creates an issue for the optimal selection technique for the learning parameters. The gradients of the activation function are inherently the main issue when using a neural network. If you were to choose an unsuitable activation function for the neural network, the final forecasting model will be extremely inaccurate and can cause a devastating effect on the overall predictions. However, we cannot use the Step, and Identity techniques as they are known to be constant linear techniques. Since the model would work in the opposite direction under back propagation during learning phase, the gradient of linear function remains constant. This causes the use of linear functions in backpropagation networks to be inefficient. Gradient functions are meant for error calculation and optimizing final inputs. When moving backwards within the back propagated neural network, the gradient of sigmoid and tanh functions gets smaller. This makes the use of Sigmoid and TanH as activation functions to be useless as it causes the vanishing gradient problem[12]. There would be no additional improvement as the gradient model would remain the same. To solve this issue, Relu activation function is used for two hidden layer and sigmoid function for the output layer.

B. **Logistic Regression -**

Logistic regression is one of the techniques that is used to conduct a regression analysis on certain input parameters which are binary. The logistic regression function depicted in Fig 2 is mainly focused on the sigmoidal function, which is at the core of the entire method. It maps out the characteristics of any given input data in a S-shaped format between the values of 0 and 1. It was originally developed by statisticians in order to study human population growth within in a controlled environment. It has been enhanced since then to suit other domains and accurately map their parameters.

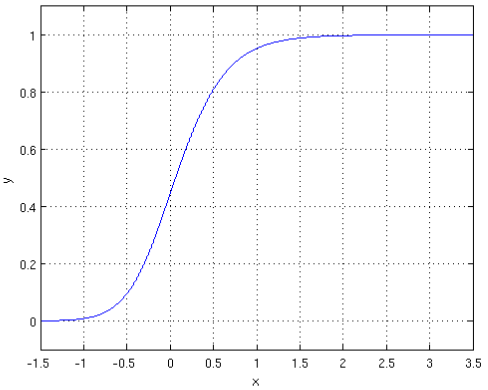

Fig. 2 Logistic Function

$$y = p + e$$

$$\text{Log}[p/(1-p)] = a + b_1x_1 + b_2x_2 + \ldots + e \qquad (1)$$

Where p =probability of outcome with the range 0 to1

Equation (1) is suitable for binary data and predicts a probability.

C. Naive Bayes -

The Naive Bayesian classifier was first described in [14] in 1973 and then in [15] in 1992.Bayesian classifiers are statistical classifiers. Naïve Bayes algorithm is one of the most robust machine learning algorithms for rainfall prediction [11]. The Naïve Bayes classifier [16] is based on Bayes rule of conditional probability. It analysis each attribute individually and assumes that all of them are independent and important. Naive Bayes classifiers have been used extensively in fault-proneness prediction, for example in [17]. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification.

D. Random Forest-

Random Forest [18] is also another approach under ensemble classifier. Random Forest is a classifier based on decision trees which exhibits great performance in computer engineering studies by Guo*et al.*, [19]. Random forest has one important advantage that it is fast and is able to handle large number of input attributes. It includes tens or hundreds of trees. In the construction of decision tree a random choice of attributes is involved. The trees are created using the following strategy [20]:

1) Each tree's root node has a sample bootstrap data which is equal to the actual data. There is a different bootstrap sample for each tree.

2) Using best split method subset of variables is randomly selected from input variables.

3) Each tree is then grown to the maximum extent possible without pruning.

4) When all trees are built in the forest, new instances are attached to all the trees then voting process takes place to select the classification with maximum votes as the new instance(s) prediction.

V. EXPERIMENTAL STUDY

Experiments are conducted on weather data of India which is first pre-possessed and cleaned. The experiments are conducted in order to compare various machine learning algorithm for rainfall prediction. In the collected weather data set, EVENT is predicted variable which tells whether it will rain on a particular day or not. The cross validation test is chosen for the experiments which randomly split the data into training and test data. By applying various algorithms on the cleaned data set models are generated which are also known as classifiers. The percentage of correctly classified instances by the classifier (model) known as classification accuracy gives us the performance measure of the classifier (model).

A. Confusion Matrix Prediction results are usually explained using confusion matrix and related performance measures. Confusion matrix is the matrix visualization of outcome of machine learning prediction model.

Confusion matrix consists of two rows and two columns that consist of True Negatives, True Positives, False Positive and False Negative.

1) True-Positive (TP), are the number of instances which are actually positive and are also predicted positive by the model.

$$\text{True Positive rate/Sensitivity} = \frac{TP}{TP + FN}$$

2) True-Negative (TN), are the number of instances which are actually negative and are also predicted negative by the model.

$$\text{True Negative rate/Specificity} = \frac{TN}{TN + FP}$$

3) False-Positive (FP), are the number of instances which are actually negative and are predicted positive by the model.

$$\text{False Positive rate} = \frac{FP}{FP + TN}$$

4) False-Negative (FN), are the number of instances which are actually positive and are predicted negative by the model.

$$\text{False Negative rate} = \frac{FN}{FN + TP}$$

B. Performance Measures There are many performance measures for classification algorithms. In this work we have implemented following performance measures: Accuracy, Precision, Recall, F-measure,

1) **Accuracy:** Accuracy is the percentage of correctly classified modules. It is one the most widely used classification performance metrics.

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + FN + TN}$$

2) **Precision:** This is the number of classified fault-prone modules that actually are fault-prone modules.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3) **Recall:** This is the percentage of fault-prone modules that are correctly classified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

4) **F-measure:** It is the harmonic mean of precision and recall. F-measure has been widely used in information retrieval.

$$\text{F-measure} = \frac{2 \text{ x Precision } \text{ x Recall}}{\text{Precision } + \text{Recall}}$$

Table-2 Performance Measure of Algoritms

Algorithm	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Naïve Bayes	**0.841**	**0.85**	**0.842**	**85.01%**
Logistic Regression	**0.865**	**0.872**	**0.864**	**87.15%**
ANN	**0.844**	**0.847**	**0.845**	**84.70%**
Random Forest	**0.874**	**0.878**	**0.875**	**87.76%**

VI. CONCLUSION

Experiments were carried out to compare popular machine Learning algorithms for rainfall prediction using various performance measures over weather data of India. The different measuring attributes play a pivotal role in giving precise rainfall prediction. It is observed that Random Forest produces best rainfall prediction results with an accuracy of 87.76% and also exhibits highest values in Recall, and F-Measure as compared to other classification algorithms. In this case, Random Forest approach proves to be an efficient and acceptable method for rainfall prediction. The level of accuracy and prediction highly depends on the data being used as input for classification and prediction. Every algorithm has its advantages and limitations; it is difficult to choose the best algorithm. The prediction accuracy of the model can be increased by developing a hybrid prediction model where multiple machine learning algorithms are put to work together. For our weather dataset, it was concluded after analyzing various models of supervised learning that the Random Forest classification algorithm has appreciable level of accuracy and acceptance.

VII. REFERENCES

[1] Zhang S., Lu L., Yu J., and Zhou H. (2016). “Short-term water level prediction using different artificial intelligent models,” in 5th International Conference on Agro-Geoinformatics, Agro-Geoinformatics.

[2] Zainudin S., Jasim D. S., and Bakar A. A.(2016). “Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, (pp. 1148–1153).

[3] Nayak D., Mahapatra A., and Mishra P.(2013). “A Survey on Rainfall Prediction using Artificial Neural Network,” Int. J. Comput. …, vol. 72, no. 16, (pp. 32–40).

[4] Rani B. K., and Govardhan A.(2013). “RAINFALL PREDICTION USING DATA MINING TECHNIQUES - A SURVEY,” (pp. 23–30).

[5] Tyagi N., and Kumar A.(2017). "Comparative analysis of backpropagation and RBF neural network on monthly rainfall prediction," Proc.

Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 1.

[6] Solanki N. and G. P. B. (2018).“A Novel Machine Learning Based Approach for Rainfall Prediction,” Inf. Commun. Technol. Intell. Syst. (ICTIS 2017) - Vol. 1, vol. 83, no. Ictis 2017.

[7] Thirumalai C. S.(2017). “Heuristic Prediction of Rainfall Using Machine Learning Techniques”.

[8] Mishra N., Soni H. K., Sharma S., and Upadhyay A. K.(2018). “Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data,” Int. J. Intell. Syst. Appl., vol. 10, no. 1,(pp. 16–23).

[9] Vathsala H., and Koolagudi S. G.(2017). “Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches,” Comput. Geosci., vol. 98, (pp. 55–63).

[10] R. VenkataRamana, B. Krishna, S. R. Kumar, and N. G. Pandey .(2013).“Monthly Rainfall Prediction Using Wavelet Neural Network Analysis,” Water Resour. Manag., vol. 27, no. 10,(pp. 3697–3711).

[11] Darji M. P., Dabhi V. K., and Prajapati H. B. (2015).“Rainfall forecasting using neural network: A survey,” 2015 Int. Conf. Adv. Comput. Eng. Appl., no. March,(pp. 706–713).

[12] Sharma, Ashutosh and Manish Kumar Goyal.(2015).“Bayesian network model for monthly rainfall forecast”, Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE International Conference.

[13] Dubey and Akash D.(2015). “Artificial neural network models for rainfall prediction in Pondicherry”, International Journal of Computer Applications, Vol. 120, No. 3.

[14] Duda R. O.,and Hart P. E.(1973). Pattern classification and scene analysis, John Wiley and Sons.

[15] Langley P., Iba W., and Thompson K. (1992). “An analysis of Bayesian Classifiers”, *in Proceedings of the Tenth National Conference on Artificial Intelligence,* San Jose, CA.

[16] Mccallum A., and Nigam K.(1998). “A Comparison of Event Models for Naive Bayes Text Classification”, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)-Workshop on Learning for Text Categorization, (pp. 41-48).

[17] Ibrahim Raaed K., Kadhim Roula A.J. .(2016). “Incorporating SHA-2 256 with OFB to realize a novel encryption”, IEEE paper on image encryption.

[18] T. Menzies, J. Greenwald and A. Frank. (2007). “Data Mining Static Code Attributes to Learn Defect Predictors”, *IEEE Transactions on Software Engineering,* Vol. 33, No. 1, 2-13.

[19] L. Breiman.(2001). “Random forests”, Machine Learning, Vol. 45, No. 1,(pp. 5-32).

[20] Guo L., Ma Y., Cukic B. and Singh H.(2004). Robust prediction of fault-proneness by random forests, *In Proc. of the 15th International Symposium on Software Relaibility Engineering ISSRE’04,*(pp. 417-428).

[21] Jiang Y., Cukic B., Menzies T., and Bartlow N.(2008). “Comparing design and code metrics for software quality prediction”, *Proc. Fourth Int. Workshop on Predictor Models in Software Engineering,* PROMISE’08, New York, USA, (pp. 11-18).

[22] Pradeep Nijalingappa and Sandeep B (2015). “Machine learning approach for the identification of diabetes retinopathy and its stages”. International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), IEEE International Conference.