RAINFALL PREDICTION USING SUPERVISED MACHINE LEARNING AND RNN TECHNIQUES: A CASE STUDY ON COBAR & COFFS HARBOUR, AUSTRALIA

Dr. VALARMATHI J SENIOR PROFESSOR,
School of Electronics Engineering,
VIT, INDIA.

DWARSALA SUJITH KUMAR REDDY- 18BEC0972, SENSE,
VIT, INDIA.

ARKAJYOTI SAHA- 18BEC0938,
SENSE, VIT, INDIA.

TIRTHAJIT DAS- 18BEC0919,
SENSE, VIT, INDIA.

ABANTEE GANGOPADHYAY-
19BEC0363, SENSE, VIT, INDIA.

Abstract— **The Rainfall Prediction is also a major prediction technology of this era, from which we predict the rainfall over a region or any particular area for any given time. It is essential to precisely decide the rainfall for better use of water resources, crop productiveness and making plans for water structures. It additionally enables human beings to take preventive measures. Since many human beings are stricken by it, high level of accuracy is extremely essential. As many know there are two types of prediction, which are short term and long-term rainfall prediction. from through research, short term predictions have been more accurate. Rainfall is also dependent on many other factors and predict them is also equally important. The dynamic nature of ecosystem and implemented arithmetic strategies fail to offer practical accuracy for those parameters. For the prediction of rainfall, we've used Supervised Machine Learning and Recurrent Neural network. Our intention for this undertaking is to create an exceedingly correct prediction version and offer a comparative look among the various machine learning and deep learning techniques. Here we are using SVM, Random Forest, Logistic Regression and RNN for the Rainfall Prediction.**

Index Terms-- Support Vector Machine, Random Forest, Logistic Regression, Supervised Learning, Recurrent Neural Networks and Accuracy.

I. INTRODUCTION

In rainfall prediction we have many different methods and most of them used deep learning techniques. Here we are doing research and making a comparative analysis of various methods used for predicting rainfall and checking their accuracy to avoid some difficulties that are faced by deep learning techniques. We are making use of Support Vector Machine (SVM) alongside Random Forest and Logistic regression as they are powerful and efficient methods to use for classification and regression related problems and also, we are using Recurrent Neural Networks to compare all the four models. Doing proper research and implementation, we can figure out when to make use of which methodology. It will also help others understand the logic and merit of the work-done and also it will help higher studies to develop in this particular study to further benefit the field.

Our research aims to evaluate the precision and accuracy of SVM, Random Forest, Logistic Regression and RNN in the field of rainfall prediction. We are using Receiver Operating Curve and Area Under the Curve for evaluating the performance of each method, we primarily did for individual methods and did a comparative one for Random Forest, Logistic Regression and

RNN. To see how ROC and AUC curves differ from one another. Pre-processing for all the 4 models will remain same to better validate the performance of each in Rainfall prediction for the given dataset.

II. LITERATURE REVIEW

Based on the collection of large number of data present like: humidity, wind-direction, temperature, and wind speed, air pressure, etc., today's weather forecast determines future weather condition. The future is unpredictable, even weather forecasts due to the disorder of atmospheric process future weather forecast predictions make some errors that might be risky. There are mainly three types of weather forecast short-term weather forecast that ranges from 2 to 3 days, then medium-term weather forecast ranges 4 to 9 days, and then long-term weather forecast ranges more than 10 to 15 days. Similarly based on coverage of area the weather forecast can be divided into the large-scale forecast that refers to the forecast of a continent or country, medium-scale forecast that refers to the forecast of a province such as a region and states, and small-scale forecast refers to forecasts of the county, city, areas, etc. Our study is carried out based on a short-medium-scale city forecast.

Studies on hybrid techniques using data-driven models, like Genetic Programming (GP), Adaptive Neuro-fuzzy Inference Systems (ANFISs), Artificial Neural Networks (ANNs), integrated with different optimization methods like Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Differential Evolution (DE) algorithm) are done by Researchers. Over the past decades' research studies on them have been published with positive results for solving problems on hydrology and water resources such as rainfall, evaporation, groundwater, river stage, and sediments.

III. DIFFERENT ALGORITHMS

Improved Quantum Genetic Algorithm: -

This process is based on the combination product of quantum computation and genetic algorithms and is based on the concept of quantum superposition states and quantum bits. Its encoding is very complex and can organize, adapt and learn by itself. It simply uses fitness to get the required solutions thus causing genes of individuals whose fitness can spread rapidly among the population causing loss of diversity at a very early stage.

Support Vector Machine: -

Support vector machine (SVM) is a purely classification method in which each data items are plotted as a point in a Nth dimension space where N being the number of features in the dataset so that the datapoints are put in their correct categories and this boundary is called the hyperplane. This process is used in the recognition of handwritten digits or letters also it can be used in facial recognition.

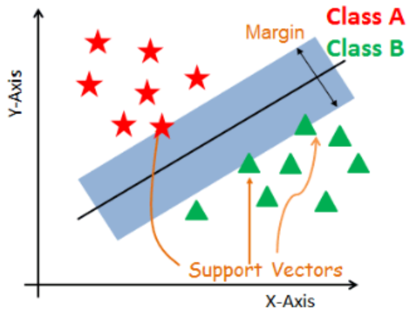


Figure 1

Random Forest: -

This process can also be defined as a group of decision trees that makes up a forest. It combines a number of classifiers to give solutions for problems that are complicated. This process can handle large datasets even with datasets with high dimensionality. It also improves the accuracy of the model and prevents issues related to overfitting. Random forest is more suitable in handling regression tasks than classification task.

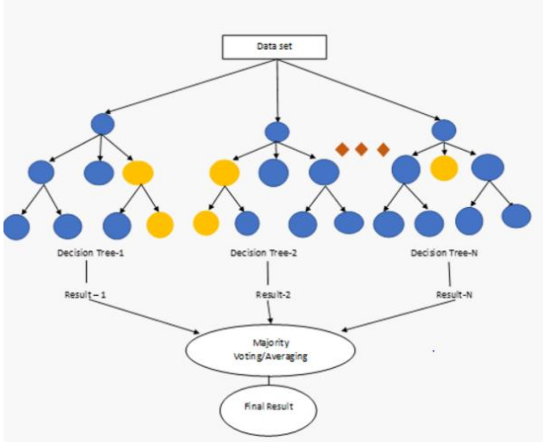


Figure 2

Logistic Regression: -

This process is used in estimating discrete values on the basis of a given set of independent variables by predicting the probability of occurrence of a particular event by fitting data to a logit function. It is a predictive analysis and is used to describe the data to explain relationships between a variable which is binary dependent and one or more independent variables which are either nominal, ordinal or interval.

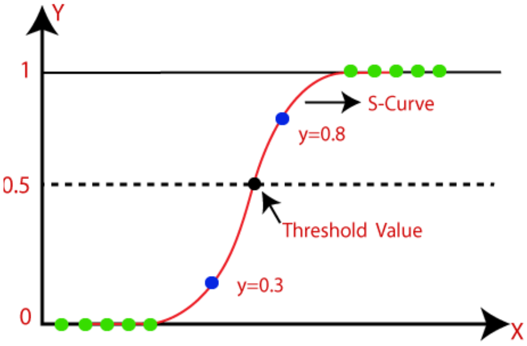


Figure 3

Recurrent Neural Network: -

A recurrent neural network or RNN is a sort of an artificial neural network that uses sequential or statistic information or data. This type of deep learning algorithms is normally used for common problems, resembling language translations, speech recognition, and image recognition. Here, we are using many to one, type of RNN as we have multiple inputs to get the particular output.

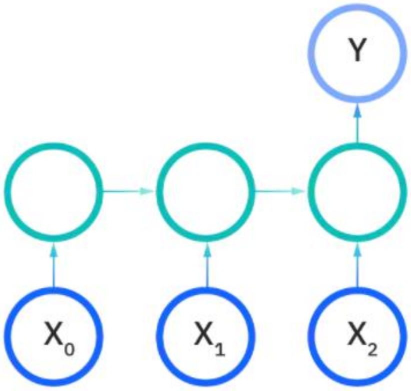


Figure 4

IV. INFORMATION AND HYPOTHESIS TO BE ADDRESSED

All the three algorithms used here are supervised machine learning algorithms mostly used in classification and regression-based problems. SVM which generates hyperplane to classify the given data instances. They are all suited to be trained on small and also average-sized complex datasets and can avoid overfitting, unlike some deep learning algorithms. Logistic regression is easy to implement, interpret, and very efficient to train. Random forest produces easily understandable predictions with high level of accuracy. In this project they will be primarily used for binary classification, into ‘Yes’ and ‘No’ or ‘0’ and ‘1’. The Rainfall Prediction is also a major prediction technology of this era, from which we predict the rainfall over a region or any particular area for any given time.

The various parameters leading to rainfall will be studied by the algorithms and the next day's rainfall occurrence will be predicted. Lastly their accuracy will be tested using relevant evaluation metrics in machine learning.

V. METHODS AND DATASET

1. THE DATASET

This database incorporates about 10 years of each day climate forecasts spanning from 2008 to 2017 from many places all through Australia, along with Albury, Badgerys Creek, Cobar, Sydney, Canberra etc. More than 145,000 sightings have been obtained withinside the workplace at approximately 50 climate stations. Rain Tomorrow is the target/output variable that desires to be predict. It may be either "Yes" or "No". This output will be "Yes" if the rain for that day become 1mm or more, else it shows "No". Dependent variables, namely, Rain Tomorrow relies upon on different factors which includes Temperature, Humidity, Wind Pressure, Direction and wind strength etc. at specific times.

2. THE CLASSIFICATION SYSTEM

- **Parameters**

In this project we have used Classification algorithm of Supervised Learning. The various factors on which rainfall is dependent and can be predicted for the near future have been taken into consideration. Date is the date on which the data were measured. But since Date doesn't affect the future possibility of rainfall it was discarded during Feature Selection process. It was also found to have high Cardinality (Uniqueness) value, which could pose many serious problems like increasing the number of dimensions of data when that feature is encoded. Next is Location, which is the common name of where the weather station is situated. Min-temp is the minimum temperature in degree Celsius and Max-temp is the maximum temperature in degree Celsius. Rainfall is the amount of rainfall recorded on that particular day in mm. Evaporation was measured using The Class A Evaporation Pan which is a standard device for measurement of evaporation. Sunshine is the number of hours of bright sunshine in the day. WindGustDir is the direction of the strongest wind gust in the 24 hours. WindGustSpeed is the speed in km/h of the strongest wind gust in 24 hours. WindDir9am and WindDir3pm correspond to the direction of the wind at 9am and 3pm respectively. WindSpeed9am and WindSpeed3pm correspond to wind speed (km/hr) averaged over 10 minutes prior to 9am and 3pm respectively. Humidity9am and Humidity3pm correspond to humidity percent at 9am and 3pm. Pressure9am and Pressure3pm correspond to the atmospheric pressure reduced to mean sea level. Cloud9am and Cloud3pm are the fraction of sky obscured by clouds. This is measured in "oktas", which is a unit for the number of eights of the sky obscured by clouds. 0 measure indicates completely clear sky whilst 8 indicates that it is completely overcast. Temp9am and Temp3pm are Temperature in degree Celsius(°C). Rain Today is a Boolean data which shows 1 if precipitation (mm) in the 24 hours (till 9am) exceeds 1mm, otherwise it shows 0, and Rain Tomorrow is the target variable, also Boolean.

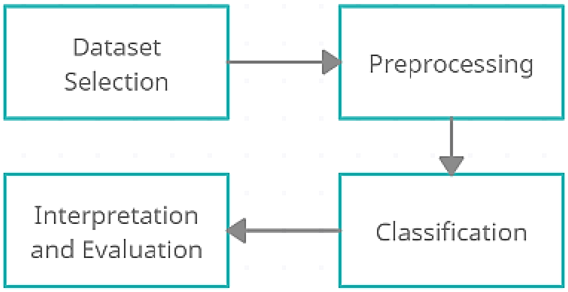


Figure 5

Sunshine is the number of hours of bright sunshine in the day. WindGustDir is the direction of the strongest wind gust in the 24 hours to midnight. WindGustSpeed is the speed (km/h) of the strongest wind gust in the 24 hours to midnight. WindDir9am and WindDir3pm correspond to the direction of the wind at 9am and 3pm respectively. WindSpeed9am and WindSpeed3pm correspond to wind speed (km/hr) averaged over 10 minutes prior to 9am and 3pm respectively. Humidity9am and Humidity3pm correspond to humidity percent at 9am and 3pm. Pressure9am and Pressure3pm are the atmospheric pressure reduced to mean sea level. Cloud9am and Cloud3pm are the fraction of sky obscured by cloud. This is measured in "oktas", which is a unit of eights. It records how many eights of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast. Temp9am and Temp3pm are Temperature in Degree Celsius(°C). Rain Today is a Boolean data which shows 1 if precipitation (mm) in the 24 hours (till 9am) exceeds 1mm, otherwise it shows 0, and Rain Tomorrow is the target variable, also Boolean.

➢ Pre-Processing

After creating a dataset with all the relevant parameters was created and removing outliers, the missing values (NaN) were replaced by the mode value for the particular column using Simple Imputer function. Then the Categorical (non- numeric) data and the dependent column were encoded using label encoder, resulting the “Yes” to be encoded as 1 and “No” to be encoded as 0. This was followed by feature scaling and dividing into training and testing data which marks the end of the pre-processing. After this data was fed into the classifiers and they were trained. Each classifier is then tested with test data to evaluate its performance.

3. PROPOSED MODEL

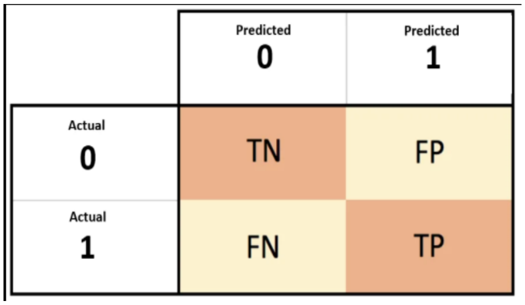
The classifiers used here are SVM, Random Forest, RNN and Logistic Regression. For this particular dataset we found Random Forest to be more accurate than the other three methodologies. We used 100 trees for the final simulation since after trying with 150 or 200 it didn’t improve the accuracy much. In RNN we used 10 epochs after than these is no significant increase in accuracy. Linear kernel was used for SVM since the data here is linearly separable. Linear Kernel is also the fastest among all other Kernels.

4. ASSESSMENT METRICS

Accuracy gives the number of correctly classified data instances over total number of data instances.

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

Confusion matrix is a method to summarize the performance of a classification algorithm. Calculating the confusion matrix can give a better idea of what the classification model is right for and what types of errors it makes. Classification report is a useful performance evaluation metric of classification-based machine learning model. It tells us our model’s recall, precision, F1 score and support. It helps us to analyze the overall performance of our trained model better. For better visualization of the performance, we have also plotted Receiver Operating Curves (ROC) along with Area Under the Curve (AUC) values for every algorithm. ROC is a probability curve and AUC represents the degree or measure of separability, i.e., its capability to separate or distinguish. It gives an estimate of how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting the classes accurately. 0 or "NO" will be classified as 0 and 1 or "YES" will be classified as 1.



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$\text{Recall} \quad \text{Sensitivity} \quad \text{True positive rate (TPR)} \quad \frac{TP}{FN + TP}$$

$$\text{False positive rate (FPR)} \quad \text{False alarm rate} \quad \frac{FP}{TN + FP}$$

An efficient model has AUC closer to 1 which means it has a good measure of separability. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. Every point on the ROC curve corresponds to a confusion matrix and it varies with the threshold value. The dotted diagonal line shows where the True Positive rate is equal to the False Positive Rate. For RNN we also used are model loss and model accuracy.

VI. RESULTS AND ANALYSIS

Here, we have Correlation matrix for the given data set, Histogram, Receiver Operating Curve (ROC) and Area Under the Curve (AUC) for all the 4 models (Random Forest, Logistic Regression, Support Vector Machine and Recurrent Neural Network). Performance of RNN and a combined ROC for Random Forest, Logistic Regression, and Recurrent Neural Network

Co-Relation plot for the Dataset:

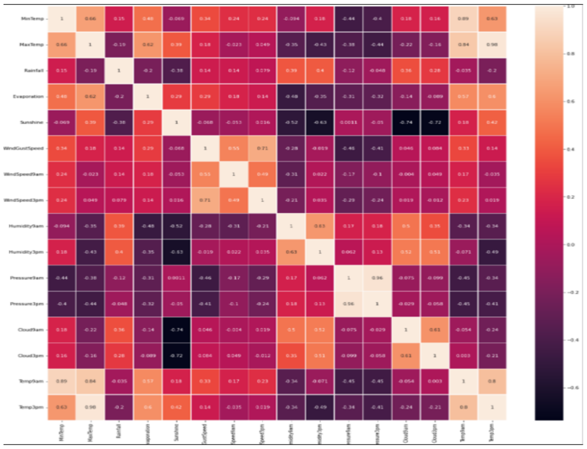


Figure 6

Histogram for the Dataset:

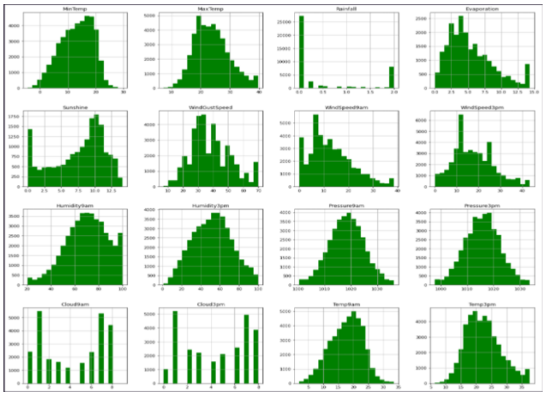


Figure 7

ROC for Random Forest:

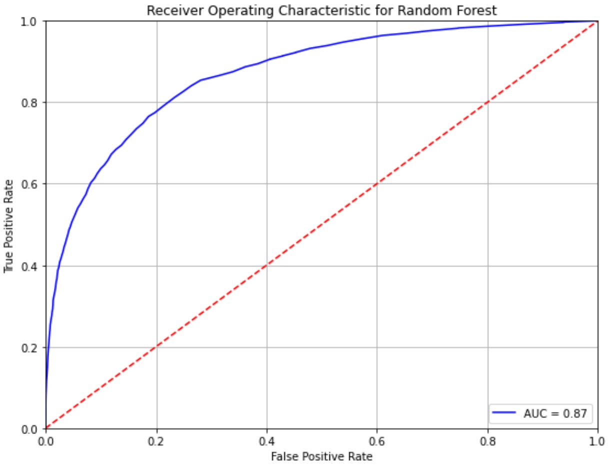


Figure 8

ROC for Logistic Regression:

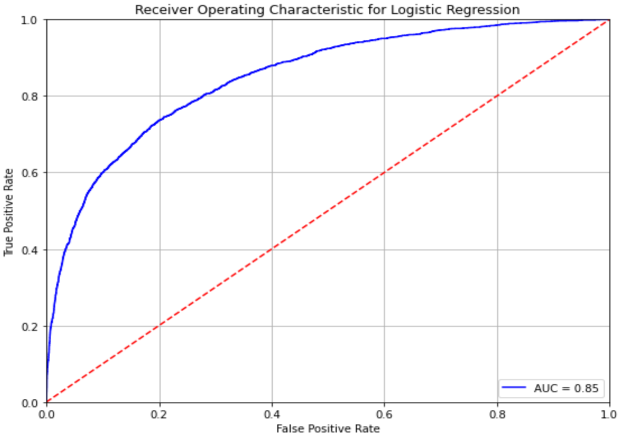


Figure 9

ROC for Support Vector Machine:

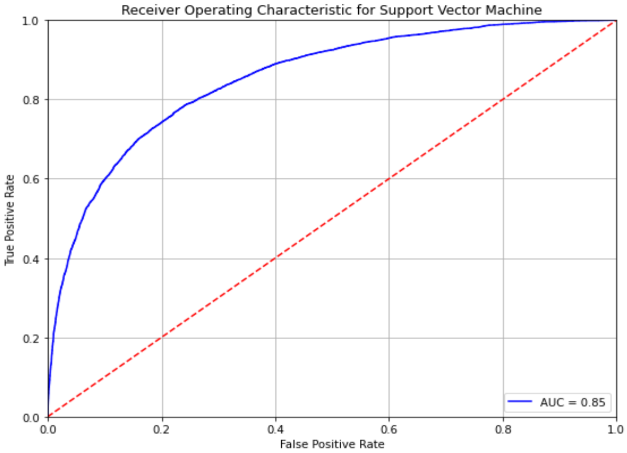


Figure 10

ROC for Recurrent Neural Network:

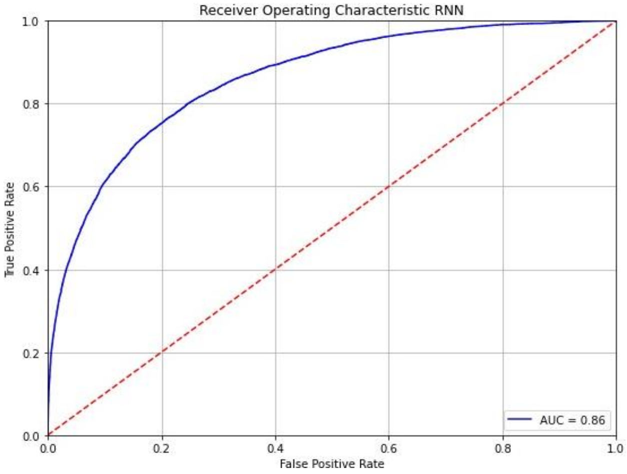


Figure 11

Recurrent Neural Network Model loss:

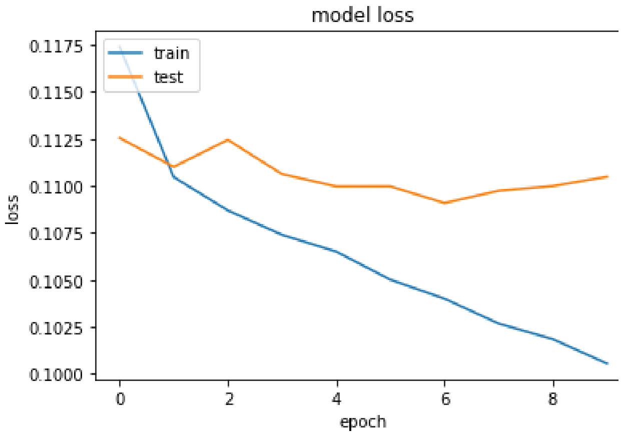


Figure 12

Recurrent Neural Network Model accuracy:

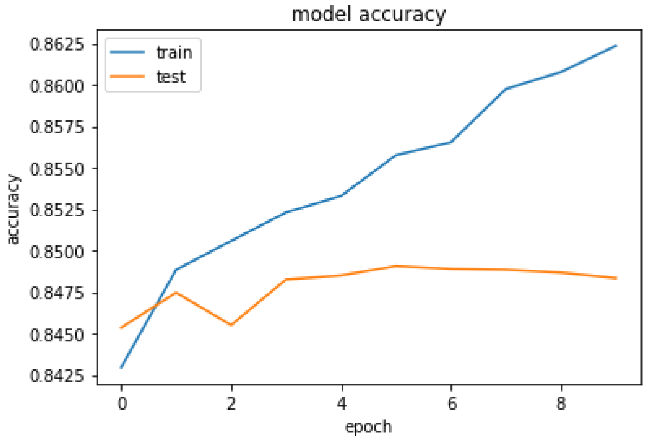


Figure 13

ROC curve comparison:

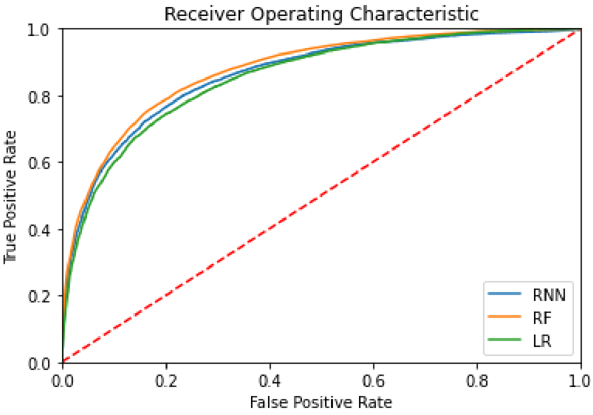


Figure 14

Here, Figure 6 is a correlation matrix [1] through which one can infer the relationship between any two parameters. The lighter colors represent positive correlation and darker shades represent negative correlation. Figure 7 depicts the values (in respective unit) and the corresponding counts of all the parameters in a histogram format [16] We observe a Gaussian distribution for Temperature, Wind, Pressure and Humidity and a Random Distribution for Cloud. Figure 8, 9, 10 and 11 represent the comparative Receiver Operating Curve (ROC) and Area Under the Curve (AUC) values for Random Forest, Logistic Regression, Support Vector Machine and

Recurrent Neural Network. In Figure 12 and 13 we can see how the RNN is performing with each epoch and after 10 we didn't see any noticeable increase in the model so we took up to 10 epochs.

In Figure 14, We can see that the ROC curves most towards the left for Random Forest with respect to others. Moreover, highest AUC value was found to be of Random Forest. Accuracy score was also found to be highest for Random Forest. Hence, we may conclude that Random Forest is the best classifier for this particular model followed by RNN, Logistic regression and SVM.

VII. CONCLUSION

In this paper, we have presented a classification-based rainfall prediction system using four models, three of them being supervised machine learning models and the last one is based on neural networks. Sine this is a large dataset with many parameters, it had to undergo intricate preprocessing, including removal of outliers, null values, imputation and encoding. We studied the importance of every parameter and how they affect the possibility of rainfall the next day. We also studied their interdependence on each other. We have used numerous metrics to evaluate the effectiveness of each model pertaining to the given dataset. All the models were found to be highly useful since all showed accuracy levels of more than 80%. However, after carefully evaluating every metric, Random Forest has been chosen as the most preferred method.

VIII. REFERENCES

[1] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, “An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives,” Expert Syst. Appl., vol. 85, pp. 169– 181, 2017.

[2] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, “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, 2018.

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

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

[5] K. K. Htike and O. O. Khalifa, “Delay Neural Networks,” Int. Conf. Comput. Commun. Eng., no. May, pp. 11–13, 2010.

[6] A. Paniagua-Tineo, S. Salcedo-Sanz, C. Casanova-Mateo, E. G. OrtizGarcía, M. A. Cony, and E. Hernández-Martín, “Prediction of daily maximum temperature using a support vector regression algorithm,” Renew. Energy, vol. 36, no. 11, pp. 3054–3060, 2011.

[7] C. Sivapragasam, S. Liong, and M. Pasha, “Rainfall and runoff forecasting with SSA-SVM approach,” J. Hydroinformatics, no. April 2016, pp. 141–152, 2001.

[8] M.J.C., Hu, Application of ADALINE system to weather forecasting, Technical Report, Stanford Electron, 1964

[9] Kalogirou, S. A., Neocleous, C., Constantinos, C. N., Michaelides, S. C. & Schizas, C. N.,"A time series construction of precipitation records using artificial neural networks. In: Proceedings of EUFIT '97 Conference, 8-11 September, Aachen, Germany. pp 2409-2413 1997.

[10] Lee, S., Cho, S.& Wong, P.M.,"Rainfall prediction using artificial neural network.",J. Geog. Inf. Decision Anal. 2, 233-242 1998.

[11] A rainfall prediction model using artificial neural network. Kumar Abhishek. 27

August 2012

[12] Monthly Rainfall Prediction Using Wavelet Neural Network Analysis. R. Venkata Ramana, B. Krishna, S. R. Kumar & N. G. Pandey. Water Resources Management volume 27, pages3697–3711 (2013)

[13] Development of advanced artificial intelligence models for daily rainfall prediction by binh ThaiPham, June 2020, 104845

[14] A. M. Bagirov, A. Mahmood, and A. Barton, "Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach," Atmos. Res., vol. 188, pp. 20–29, 2017.

[15] S. S. Monira, Z. M. Faisal, and H. Hirose, "Comparison of artificially intelligent methods in short term rainfall forecast," Proc. 2010 13th Int. Conf. Comput. Inf. Technol. ICCIT 2010, no. Iccit, pp. 39–44, 2010.

[16] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar, "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine," IEEE Trans. Knowl. Data Eng., vol. 20, no. 9, pp. 1264–1272, 2008.

[17] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "A Comprehensive Survey of Data Mining Techniques on Time Series Data for Rainfall Prediction," vol. 11, no. 2, pp. 168–184, 2017.

[18] K. W. Chau and C. L. Wu, "A hybrid model coupled with singular spectrum analysis for daily rainfall prediction," J. Hydroinformatics, vol. 12, no. 4, p. 458, 2010.

[19] J. Wu, J. Long, and M. Liu, "Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm," Neurocomputing, vol. 148, pp. 136–142, 2015.

[20] W. C.L. and K.-W. Chau, "Prediction of Rainfall Time Series Using Modular Soft Computing Methods," Eng. Appl. Artif. Intell., vol. 26, no. 852, pp. 1–37, 2012.

[21] D. Gupta and U. Ghose, "A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy," pp. 0–5, 2015.

[22] K. Abhishek, A. Kumar, R. Ranjan, and S. Kumar, "A rainfall prediction model using artificial neural network," 2012 IEEE Control Syst. Grad. Res. Colloq., no. Icsgrc, pp. 82–87, 2012.

[23] M. A. Nayak and S. Ghosh, "Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier," Theor. Appl. Climatol., vol. 114, no. 3–4, pp. 583–603, 2013.

[24] M. Ahmad and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[25] Halide, H. and Ridd P. (2002): "Modeling interannual variation of a local rainfall data using a fuzzy logic technique": Proceedings of International Forum on Climate Prediction, 2002, James Cook Universiy, Australia, pp: 166-170.