Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7641958

FORECASTING OF RAINFALL USING MACHINE LEARNING ALGORITHM

DR. MD. ATHEEQ SULTAN GHORI
ASSOCIATE PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE & ENGG
TELANGANA UNIVERSITY
NIZAMABAD.
atheeqsultan@gmail.com

ABSTRACT

Precipitation expectation is one of the difficult undertakings in weather conditions estimating. Exact and ideal precipitation expectation can be exceptionally useful to go to viable security lengths ahead of time with respect to: progressing development projects, transportation exercises, horticultural errands, flight tasks and flood circumstance, and so on. Storm expectation is obviously vital for India. Two kinds of precipitation forecasts should be possible, They are - Long haul expectations: Foresee precipitation more than few weeks/months ahead of time. - Momentary forecasts: Foresee precipitation a couple of days ahead of time in unambiguous areas. Indian meteorological office gives guaging information expected to the forecast. This framework is intended to deal with long haul forecasts of precipitation. The vitally thought process behind the improvement of this model is to foresee how much precipitation in a specific division or state well ahead of time. We foresee how much precipitation utilizing past information.

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

INTRODUCTION

Exact guaging of precipitation has been one of the main issues in hydrological research on the grounds that early admonitions of extreme weather conditions can assist with forestalling setbacks and harms brought about by catastrophic events, if ideal and precisely anticipated. To build a prescient framework for precise precipitation, estimating is one of the best difficulties to specialists from different fields, for example, climate information mining, ecological AI, functional hydrology, and factual estimating. A typical inquiry in these issues is the manner by which one can dissect the past and utilize future forecast. The boundaries that are expected to foresee precipitation are massively intricate and unobtrusive in any event, for a momentary period. Actual cycles in precipitation are for the most part made out of various sub-processes. An exact demonstrating of precipitation by a solitary worldwide model is now and then unrealistic. To conquer this trouble, the idea of particular demonstrating and consolidating various models has drawn in more consideration as of late in precipitation estimating. In measured models, a few sub-processes are first distinguished, and afterward separate models are laid out for every one of them. Up to this point, different secluded models have been proposed, contingent upon delicate or hard parting of preparing information. Delicate dividing

implies that the dataset can be covered, and the general anticipating yield is the weighted normal of every neighborhood model.

Review of Literature

Researchers have been attempting to work on the precision of precipitation expectation by advancing and coordinating AI procedures. A portion of the chose studies are examined in this segment.

The study by Zhang S., Lu L., et al played out a similar investigation of Help Vector Machine (SVM), Counterfeit Brain Organizations (ANN), and Versatile Neuro Fluffy Induction Framework (ANFIS) on precipitation expectation. The creators have thought about the forecast models in four terms:

- by involving various slacks as displaying inputs;
- by utilizing preparing information of weighty precipitation occasions just;
- execution of anticipating for 1 hour to 6 hours and;
- execution examination in top qualities and all qualities.

As per results ANN performed better when prepared with dataset of weighty precipitation. For 1 to 4 hour ahead estimating, the past 2-hour input information was recommended for every one of the three displaying strategies. ANFIS reflected better capacity in keeping away from data commotion by utilizing various slacks of sources of info. Lastly during top qualities, SVM ended up being more powerful under outrageous tropical storm occasions.

In the study of Zainudin S., Jasim D. S., and Bakar A. A. played out a relative examination of different information digging procedures for precipitation expectation in Malaysia, for example, Arbitrary Backwoods, Backing Vector Machine, Innocent Bayes, Brain Organization, and Choice Tree. For this analysis, dataset was acquired from different weather conditions stations in Selangor, Malaysia. Before grouping process, Pre-handling errands were applied to manage the commotion and missing qualities in dataset. The outcomes showed huge execution of Irregular Woods as it accurately grouped enormous measure of occasions with modest quantity of preparing information.

Nayak D., Mahapatra A., and Mishra P. played out a study on different Brain Organization models which were utilized for precipitation expectation in most recent 25 years. The creators featured that the majority of the specialists obtained critical outcomes in precipitation expectation by utilizing Spread Organization, also the determining strategies which utilized SVM, MLP, BPN, RBFN, and SOM are more reasonable than other measurable and mathematical methods. A few limits have likewise been featured.

According to Rani B.K. and Govardhan A. involved Counterfeit Brain Organization for precipitation expectation in Thailand. They involved Back Spread Brain Organization for expectation which detailed an OK exactness. For future course it was recommended that couple of extra highlights would be remembered for input information for precipitation expectation, for example, Ocean Surface Temperature for the areas around Andhra Pradesh and Southern piece of India.

Tyagi N. and Kumar A. anticipated month to month precipitation by utilizing Back Spread, Spiral Premise Capability and Brain Organization. For expectation, the dataset was gathered from Coonoor area in Nilgiri region (Tamil Nadu). Execution was assessed concerning Mean Square Blunder. As per results higher exactness was accounted for in Outspread Premise Capability Brain Organization with more modest Mean Square Blunder. Besides the analysts additionally involved these strategies for future precipitation expectation.

Solanki N. introduced a Crossover Shrewd Framework by incorporating Counterfeit Brain Organization and Hereditary Calculation. In ANN, MLP functions as the Information Mining motor to perform expectations while the Hereditary Calculation was used for inputs, the association structure between the data sources, the result layers and to make the preparation of Brain Organization more successful.

According to the study by Thirumalai T.S., he examined precipitation pace in earlier years concerning different harvests seasons like rabi, Kharif, zaid and afterward anticipated (precipitation) for future seasons by means of Direct Relapse Technique. For expectation, input dataset was chosen by specific corps times of earlier years.

Mishra N., Soni H.K., Sharma S., and Upadhyay A.Y. one month and multi month anticipating models were created for precipitation expectation by utilizing Fake Brain Organization (ANN). The info dataset was chosen from different stations in North India, crossed on beyond 141 years. Feed Forward Brain Organization utilizing Back Spread and Levenberg-Marquardt preparing capability were utilized in these models. Execution of the two models was assessed by utilizing Relapse Investigation, Mean Square Blunder and Extent of Relative Mistake. The outcomes showed that one month guaging model can foresee the precipitation more precisely than multi month determining model.

Vathsala H. and Koolagudi S. G. introduced a calculation by incorporating Information Mining and Factual Procedures. The proposed procedure anticipated the precipitation in five distinct classifications, for example, Flood, Abundance, Typical, Deficiency and Dry spell. The indicators were chosen with most elevated certainty level, in light of affiliation runs and got from neighborhood and worldwide climate. From nearby climate: wind speed, ocean level tension, most extreme temperature, and least temperature were taken. From worldwide climate: Indian sea dipole conditions and southern wavering were taken.

In [10], R. Venkata Ramana, et al, anticipated the precipitation by utilizing proposed Wavelet Brain Organization Model (WNN), a joining of Wavelet Method and Fake Brain Organization (ANN). To dissect the presentation, month to month precipitation expectation was performed with both the methods (WNN and ANN) by involving dataset of Darjeeling precipitation check station in India. Factual methods were utilized for execution assessment and as indicated by results WNN performed better compared to ANN.

Darji M.P. et al, gave a point by point overview and played out a near investigation of different brain networks on precipitation estimating. As indicated by overview RNN, FFNN, and TDNN are reasonable for precipitation expectation when contrasted with other measurable and mathematical anticipating strategies. In addition TDNN, FFNN and slack FFNN performed well for yearly, month to month and week by week precipitation estimating separately. This

exploration additionally examined the different proportions of exactness utilized by various scientists to assess the ANN's exhibition.

Sharma et al, proposed Bayesian organization model for mean month to month precipitation expectation of 21 stations in Assam, India. This work can be valuable for better administration of water assets. Month to month information of a long time from 1981 to 2000 for every one of the climatic boundaries is utilized for this study which was taken from various sources. Precipitation at a station is taken as a variable for this model and conditions between rainfalls at various station is shown by Bayesian organization. In this work, the creator utilized K2 calculation and contingent likelihood is tracked down utilizing greatest probability approximations. Five unique climatic boundaries viz. Temperature, Overcast cover, Relative dampness, Wind speed and Southern Wavering List (SOI) are utilized. The outcomes uncovered that temperature is seen as generally productive and wind speed least. SOI is likewise tracked down significant in working on the outcomes. Some station got effectiveness above 95% though other station obtained good outcomes.

Akash D Dubey, proposed a precipitation forecast model utilizing fake brain organizations (ANN). In this work the creator has utilized the climate information of Pondicherry, India. Three different preparation calculations viz. feed-forward back proliferation calculation, layer repetitive calculation and feed-forward disseminated time postpone calculation were utilized to make ANN models and saving number of neurons for every one of the models to 20. Of the relative multitude of calculations, the outcomes showed that feed-forward dispersed time postpone calculation has best precision and MSE esteem as low as 0.0083.

Data Collection

For the expectation model, climate information of India, This dataset has normal precipitation from 1951-2000 for each region, for each month. The crude climate information gathered comprises of nine estimated ascribes which are date, temperature (high, low, normal) in ºc , Dew point (high, low, normal) in ºc , Dampness (high, low, normal) in % age, ocean level strain (high, low, normal) in hPa, perceivability (high, low, normal) in Km, wind (high, low, normal) in Km/h, precipitation (high, low, normal) in mm, Occasions (Precipitation snow, rainstorm, mist). For this work out of these 9 highlights we have utilized the Normal temperature, Normal Mugginess, Normal Ocean level tension, Normal breeze and Occasions highlights as displayed in table I. We have disregarded less applicable elements in the dataset for better model calculation and expectation.

Table-1. Weather Data Description

Attribute	Type	Description
Temperature	Numerical	Temp is in ºC
Humidity	Numerical	Humidity in P ercentage
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 primary test in climate expectation is the unfortunate information quality and determination. Therefore, preprocessing of information is painstakingly finished to get precise and right forecast results. In this stage undesirable information or commotion is eliminated from the gathered informational index which is finished by eliminating the undesirable credits and keeping the most significant characteristics that assistance in better forecast. Another significant issue that will be redressed is the missing qualities in the gathered informational index. Missing qualities in the informational collection is filled by utilizing different strategies. The missing qualities for credits in the dataset are supplanted with the modes and means in view of existing information. Adding the missing qualities gives a more complete dataset to the classifiers to be prepared.

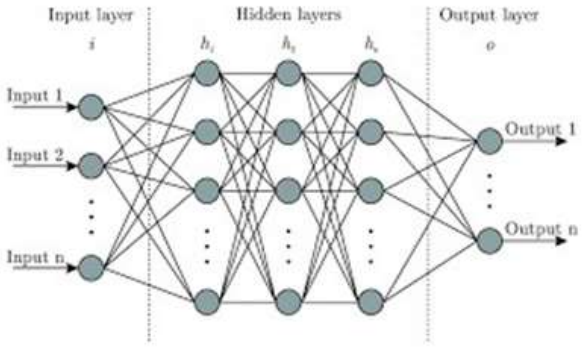
Research Methodology

There are two fundamental kinds of AI draws near; managed learning and unaided learning. Directed learning calculations are utilized for building prescient models. The Characterization calculations ANN, Strategic Relapse, Guileless Bayes, and Random Forest are tentatively executed and thought about against one another.

- **Artificial Neural Networks -**

A brain network is a huge dispersed processor which works in an equal technique and comprises of straightforward handling units which store exact information and have it prepared for dynamic use. The usefulness of brain network is comparable to the human cerebrum, as in they can get data sources, and cycle the data utilizing different processing hubs. An important result is created in light of the ideal application.The fundamental benefit of Brain Organizations is its capacity to show non-linearity presence between the information and result factors.

Fig. 1. Neural Network



Major concerns with neural networks

- **Number of Hidden Layers and Nodes:** The overall Fake Brain Organization portrayed in Fig. 1 comprises of 3 common layers, which are the Information, Stowed away, and Result Layer. Presently, the layer that we are worried about is the Secret Layer. This layer is essentially answerable for the estimations that happen with brain organizations and is additionally the layer where genuine nonlinear planning among information and result happens. Assuming any wrong step were to be taken in this layer, the whole consequence

of the brain organization would be strayed and may devastatingly affect forecast. Because of this, we should consider the number of stowed away layers we that need and the hubs inside them. Such boundaries can support the exactness of the whole organization.

- **Over fitting:** One more serious issue that exists inside the brain network is called overfitting. It obstructs the brain network while attempting to a draw speculation for a particular given input. At the point when there are countless information boundaries that are being taken care of to the brain network during its preparation stage, it can cause a speculation mistake. The model doesn't know which explicit class to group the information in. While, in the event that we don't give it adequate information, a misstatement could happen which prompts more terrible estimate yields.
- **Choice of Enactment Capability:** One of the fundamental worries while utilizing brain networks is choosing a suitable enactment capability. The enactment capability in brain network assumes an imperative part in deciding the way of behaving of the whole brain organization. The fundamental occupation of the enactment capability is to arrange helpful data and eliminate commotion that is viewed as in the ongoing set. The enactment capability is additionally liable for ascertaining the loads of the given contributions at every hub. This permits the capability to decide whether a neuron can be enacted or not. There are numerous actuations works that can be utilized for gauging models like Parallel step, Sigmoid, SoftMax, Relu, and Tanh. Notwithstanding, picking the particular enactment capability is exclusively subject to the issue.

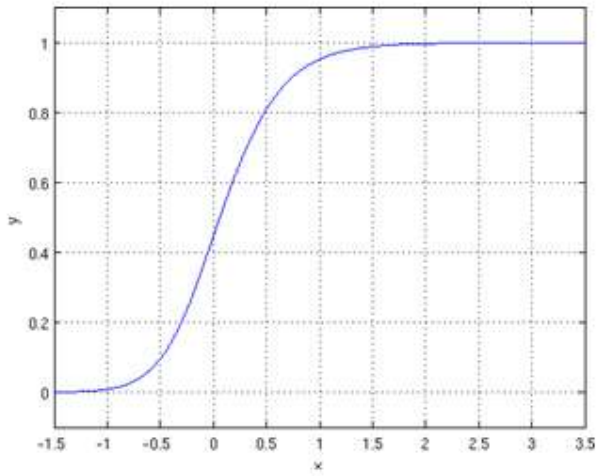
For instance, The Sigmoid capability is more qualified towards double characterizations undertakings while the SoftMax capability is equipped towards multi-groupings assignments. The repeating issue for this multitude of frameworks is the slope float on the brain organization. In some cases, the slopes are too steep in a particular heading and different times it tends to be excessively low or zero. This makes an issue for the ideal determination strategy for the learning boundaries. The slopes of the enactment capability are innately the main pressing concern while utilizing a brain organization. If you somehow happened to pick an unacceptable initiation capability for the brain organization, the last determining model will be very off base and can cause an overwhelming impact on the general expectations. Nonetheless, we can't utilize the Step, and Character procedures as they are known to be steady direct strategies. Since the model would work the other way under back engendering during learning stage, the slope of straight capability stays consistent. This causes the utilization of direct capabilities in back propagation organizations to be wasteful. Slope capabilities are intended for mistake estimation and upgrading last information sources. While moving in reverse inside the back proliferated brain organization, the slope of sigmoid and tanh capabilities gets more modest. This utilizes Sigmoid and TanH as initiation capabilities to be futile as it causes the disappearing angle problem. There would be no extra improvement as the slope model would continue as before. To tackle this issue, Relu actuation capability is utilized for two secret layer and sigmoid capability for the result layer.

- **Logistic Regression -**

Strategic relapse is one of the methods that is utilized to direct a relapse examination on specific information boundaries which are paired. The calculated relapse capability portrayed in following figure is basically centered on the sigmoidal capability, which is at the center of the

whole strategy. It delineates the qualities of some random info information in a S-molded design between the upsides of 0 and 1. It was initially evolved by analysts to concentrate on human populace development inside in a controlled climate. It has been improved from that point forward to suit different spaces and precisely map their boundaries.

Fig. 2 Logistic Function



$y = p + e$

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

Where p =probability of outcome with the range 0 to1

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

- **Naive Bayes -**

The Credulous Bayesian classifier was first portrayed in 1973 and afterward in 1992.Bayesian classifiers are factual classifiers. Gullible Bayes calculation is one of the heartiest AI calculations for precipitation expectation. The Credulous Bayes classifier [16] depends on Bayes rule of contingent likelihood. It investigation each characteristic exclusively and accepts that every one of them are autonomous and significant. Credulous Bayes classifiers have been utilized widely in shortcoming inclination expectation, for instance in [17]. A benefit of the credulous Bayes classifier is that it requires a modest quantity of preparing information to gauge the boundaries essential for characterization.

- **Random Forest-**

Arbitrary Woodland is additionally one more methodology under troupe classifier. Irregular Woods is a classifier in light of choice trees which shows extraordinary execution in PC designing examinations by Guoet al.,. Arbitrary woodland enjoys one significant benefit that it is quick and can deal with huge number of info credits. It incorporates tens or many trees. In the development of choice tree an irregular selection of qualities is involved. The trees are come up with utilizing the accompanying system:

- Each tree's root hub has example bootstrap information which is equivalent to the genuine information. There is an alternate bootstrap test for each tree.
- Using best parted technique subset of factors is haphazardly chosen from input factors.
- Each tree is then developed to the most extreme degree conceivable without pruning.
- When all trees are underlying the backwoods, new occurrences are appended to every one of the trees then; at that point, casting a ballot interaction happens to choose the order with most extreme votes as the new instance(s) forecast.

Experimental Study

Tests are led on climate information of India which is first pre-moved by cleaned. The analyses are directed to analyze different AI calculation for precipitation expectation. In the gathered climate informational collection, Occasion is anticipated variable which tells regardless of whether it will rain on a specific day. The cross approval test is picked for the trials which arbitrarily split the information into preparing and test information. By applying different calculations on the cleaned informational collection models are produced which are otherwise called classifiers. The level of accurately grouped occurrences by the classifier (model) known as order exactness gives us the presentation proportion of the classifier (model).

A. Disarray Lattice Forecast results are normally made sense of utilizing disarray network and related execution measures. Disarray grid is the network representation of result of AI expectation model.

Disarray framework comprises of two lines and two segments that comprise of Genuine Negatives, Genuine Up-sides, Misleading Positive and Bogus Negative.

1) True-Positive (TP), are the quantity of occasions which are really certain and are likewise anticipated positive by the model.

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

2) Genuine Negative (TN), are the quantity of occurrences which are really negative and are additionally anticipated negative by the model.

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

3) Misleading Positive (FP), are the quantity of occurrences which are really negative and are anticipated positive by the model.

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

4) False-Negative (FN), are the quantity of examples which are really sure and are anticipated negative by the model.

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

B. Execution Estimates there are numerous presentation measures for arrangement calculations. In this work we have carried out following execution measures: Exactness, Accuracy, Review, F-measure,

1) Accuracy: Exactness is the level of accurately arranged modules. It is one the most generally utilized grouping execution measurements.

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

2) Precision: This is the quantity of arranged shortcoming inclined modules that really are shortcoming inclined modules.

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

3) Recall: This is the level of shortcoming inclined modules that are accurately grouped.

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

4) F-measure: It is the consonant mean of accuracy and review. F-measure has been broadly utilized in data recovery.

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

Table-2 Performance Measure of Algorithms

Algorithm	Precision	Recall	F- Measure	Accuracy
Naïve Bayes	0.852	0.859	0.853	85.47%
Logistic Regression	0.873	0.881	0.873	87.57%
ANN	0.851	0.853	0.851	85.16%
Random Forest	0.883	0.886	0.882	88.37%

Conclusion

Tests were completed to think about well-known AI calculations for precipitation expectation utilizing different execution estimates over climate information of India. The different estimating credits assume a significant part in giving exact precipitation expectation. It is seen that Irregular Woodland creates best precipitation expectation results with a precision of 87.76% and furthermore shows most noteworthy qualities in Review, and F-Measure when contrasted with other characterization calculations. For this situation, Irregular Woodland approach ends up being a proficient and satisfactory strategy for precipitation expectation. The degree of exactness and expectation exceptionally relies upon the information being utilized as contribution for arrangement and expectation. Each calculation enjoys its benefits and impediments; picking the best algorithm is troublesome. The expectation precision of the model can be expanded by fostering a mixture forecast model where various AI calculations are assembled to work. For our climate dataset, it was closed subsequent to breaking down different

models of regulated discovering that the Irregular Woodland grouping calculation has obvious degree of exactness and acknowledgment.

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 Reliability 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.