Rainfall Prediction using Data Mining Techniques: A Systematic Literature Review

Shabib Aftab, Munir Ahmad, Noureen Hameed, Muhammad Salman Bashir, Iftikhar Ali, Zahid Nawaz
Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

***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. Data mining techniques can effectively predict the rainfall by extracting the hidden patterns among available features of past weather data. This research contributes by providing a critical analysis and review of latest data mining techniques, used for rainfall prediction. Published papers from year 2013 to 2017 from renowned online search libraries are considered for this research. This review will serve the researchers to analyze the latest work on rainfall prediction with the focus on data mining techniques and also will provide a baseline for future directions and comparisons.**

Keywords*—Rainfall prediction; data mining techniques; SLR; systematic literature review

I. INTRODUCTION

Analysis of time series data is one of the important aspects of modern research in the domain of knowledge discovery [28]. Time series data is collected over a specific period of time such as hourly, daily, weekly, monthly, quarterly or yearly [23], [40]. Data mining techniques can use this data to predict upcoming situations in various domains such as climate change, education, and finance etc. These techniques can be used to extract hidden knowledge from time series data for future use [23], [27], [29], [40]. Weather forecasting is very benificial but challenging task [26]. Weather data consists of various atmospheric features such as wind speed, humidity, pressure and temperature etc. Data mining techniques have the capacity to extract the hidden patterns among available features of past weather data and then these techniques can predict future weather conditions by using extracted patterns [40]. Rainfall is a complex atmospheric process, which depends upon many weather related features. Accurate and timely rainfall prediction can be helpful in many ways such as planning the water resources management, issuance of early flood warnings, managing the flight operations and limiting the transport & construction activities [24], [25]. Accurate rainfall prediction is more complex today due to climate variations. Reseacrhers consistently have been working to predict rainfall with maximum accuracy by optimizing and integrating data mining techniques [41]. Data mining algorithms are classified as supervised and un-supervised. Supervised methods get trained first with pre-classified data (training data) and then classify the input data (test data) [7], [38], [39]. Un-supervised methods on the other hand do not require any training, instead of pre-classified data these techniques use algorithms to extract hidden structure form un-labled data. It has been observed from latest research that for high accuracy, researchers prefer the integrated techniques for the rainfall prediction. To reflect the latest research, this study provides a systematic literature review by focusing on latest papers, which are published in last five years (2013-2017). Three renowned online search libriries are selected for literarure extraction: Elsvier, IEEE and Springer. Initally 4844 papers are extracted and then through a systemaic research process 8 most relevent research articles are selected for critical review.

Further organization of this paper is as follows. Section II elaborates the related work. Section III presents the research protocol, which is followed in this research. Section IV presents the review of shortlisted articles. Section V discusses the review findings. Section VI finally concludes this study.

II. RELATED WORK

Reseachers have been working to improve the accuracy of rainfall prediction by optimizing and integrating data minig techniques. Some of the selected studies are discussed in this section. In [1], author 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 forecastingfor 1 hour to 6 hoursand; (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. Researchers in [2] 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], author 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. Researchers in [4] 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. Researchers in [5] 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. Researchers in [6] 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. Researchers in [8] 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 [9], 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. Researchers in [10] 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 [11], researchers 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 [12], researchers 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.

III. Research Protocol

High quality SLR is one which attains its objective by providing the compact information of required research topic for a particular time span. A detailed research methodology with step by step guidance is needed to conduct an effective SLR. In this research a systematic research process is formulated by following the guidelines extracted from [13]-[18]. Usually SLR consists of three basic steps: plan review, conduct review and document review moreover further nested steps can be included from modern and state of the art research papers for an effective presentation. For this study, a step by step systematic review process is extracted from the latest review articles of software engineering domains [19]-[22]. The systematic review process of this research consists of the following steps: A) Identification of research questions, B) Keywords selection for query string, C) Selection of search space, D) Outlining the selection criteria, E) Literature extraction, F) Quality assessment, G) Literature Analysis and H) Results and Discussion (Fig. 1).

A. Identification of Research Questions

Research objectives are identified and presented in the form of research questions. The ultimate purpose of SLR is to find the answers of those questions via critical review. Flowing are the research questions identified for this research.

RQ1: Which data mining techniques are used / proposed for rainfall prediction?

RQ2: How the performance of prediction techniques is evaluated?

RQ3: Which type of data is used for prediction?

RQ4: For which location the rainfall prediction is performed?

RQ5: Which factors affect the prediction results?

RQ6: Which are the latest research trends in the domain of rainfall prediction?

Fig. 1. SLR process.

B. Keywords Selection for Query String

Second step is to formulate the query string and for that purpose, keywords are extracted first from the research questions and then arranged in a particular sequence to form a query. Following keywords are extracted for query:

Improved, Customized, Integrated, Data Mining, Techniques, Methods, Algorithms, Rainfall, Prediction, Forecasting, Estimation, Performance, Evaluation, Assessment.

The finalized query string is given below:

("Performance" AND ("Evaluation" OR "Assessment") AND / OR ("Improved" OR "Customized" OR "Integrated") AND ("Data Mining") AND ("Techniques" OR "Methods" OR "Algorithms") AND "Rainfall" AND ("Prediction" OR "Forecasting" OR "Estimation")).

C. Selection of Search Space

This step deals with the selection of libraries from where the related literature will be extracted through query string. Three well known and widely used online libraries are selected to extract the literature: IEEE, Elsevier and Springer. All three libraries have different options to search the relevant material, so few adjustments were made in query strings to extract the appropriate and most relevant literature. The Query was searched multiple times with various combinations of key-words. Results of search queries are available in Table I.

TABLE I. SEARCH SPACE AND QUERY RESULTS

Sr. #	Digital Library	Date Searched	Total Results
1	Elsevier	2018-24-02	1819
2	IEEE	2018-24-02	1119
3	Springer	2018-24-02	1906

D. Outlining the Selection Criteria

This step aims to outline the selection boundary so that most relevant research papers can be selected. This activity consists of two steps, IC (inclusion criteria) and EC (exclusion criteria).

1) Inclusion Criteria

Below are the rules of Inclusion criteria.

IC1: Papers which are published from 2013 till 2017.

IC2: Papers which are available in journals, conferences, proceedings of conferences or workshops.

IC3: Papers which have predicted the rainfall using data mining techniques.

IC4: Papers which have performed comparison of data mining techniques on rainfall prediction.

IC5: Papers which have presented improved or customized data mining techniques to predict rainfall.

IC6: Papers which have integrated data mining technique with any other technique.

2) Exclusion Criteria (EC)

Below are the rules of exclusion criteria.

EC1: Papers which are not in English.

EC2: Papers published before 2013 or after 2017.

EC3: Papers which did not perform rainfall prediction.

EC4: Paper which did not use any data mining technique in proposed model/method?

EC5: Paper which did not use any weather data for prediction.

EC6: Papers which did not evaluate the performance of used/proposed technique.

E. Literature Extraction

The purpose of selection criteria is to extract the most relevant literature for the review. After applying IC and EC, 18 articles were shortlisted. Complete process of literature extraction is given in Fig. 2.

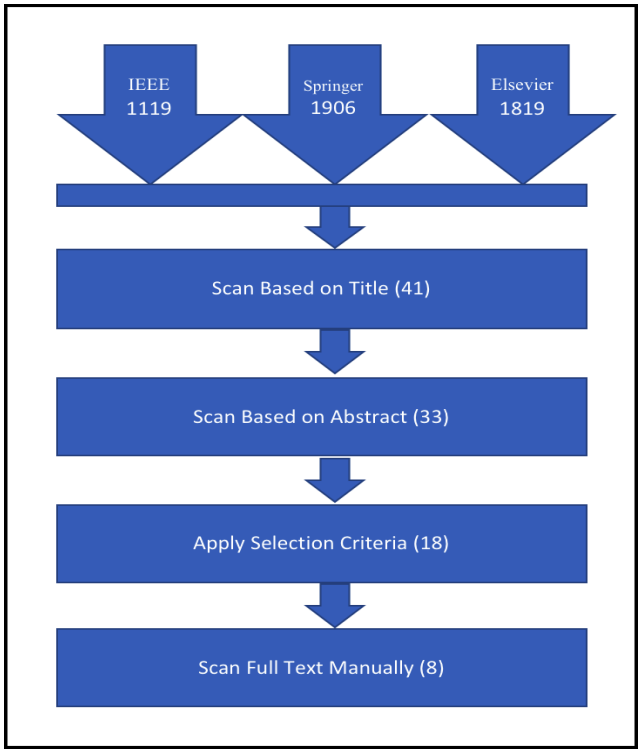

Fig. 2. Search process.

F. Quality Assessment

To meet the research objectives, it was make sure to follow the quality parameters throughout the systematic research process. To ensure the quality of results, following measures were taken.

- Authentic and renowned online libraries were selected to extract research articles.
- Latest research papers were selected to reflect latest research.
- The process of selection was un-biased.
- Complete steps of Systematic Research Process were followed in the true sense.

IV. Literature Analysis

Full text of 18 selected articles were analyzed and then 8 most relevant research papers are shortlisted for critical review as shown in Table II. The Review of shortlisted articles is given below.

TABLE II. Most Relevent Research Literature

Sr. #	Digital Library	Selected Research Literature	No. of Researches
1	Elsevier	[30]-[33]	4
2	IEEE	[34]	1
3	Springer	[35]-[37]	3

A. Indian Summer Monsoon Rainfall (ISMR) Forecasting using Time Series Data: A Fuzzy-Entropy-Neuro based Expert System

In [30], authors presented a model to forecast Indian Summer Monsoon Rainfall on the basis of monthly and seasonal timescales. To forecast, time series dataset was used, spanning from 1871 till 2014. The dataset was classified in two parts (1) 1871-1960 used as training data, and (2) 1961-2014 used as test data. Statistical analysis reported the dynamic nature of rainfall in monsoon, which could not be predicted effectively with mathematical and statistical models. So, the authors in this research recommended to use three techniques for this type of prediction: Fuzzy Set, Entropy and Artificial Neural Network. By using these three techniques, a forecasting model is developed to deal with the dynamic nature of the ISMR. In proposed model, fuzzy set theory is used to handle uncertainties which are inherited in dataset. The entropy computational concept was modified in this model and used to provide the input as a degree of membership in the entropy function. That entropy function was referred as Fuzzy Information-Gain (FIG). Then, each fuzzified rule was defuzzified using the ANN. The value of FIG of each fuzzy- set was then used as input into ANN. The proposed model was named as “Fuzzy-Entropy-Neuro Based Expert System for ISMR Forecasting” because it is the integration of fuzzy set, entropy and ANN. To evaluate the performance of proposed model following accuracy measures were used: Standard Deviations (SDs), Correlation Coefficient (CC), Root Mean Square Error (RMSE) and Performance Parameter (PP). According to results the proposed model is effective and efficient in comparison with other existing models.

B. An Extensive Evaluation of Seven Machine Learning Methods for Rainfall Prediction in Weather Derivatives

The researchers in [31] compared the predictive performance of latest and state of the art method named “Markov chain extended with rainfall prediction” with the other widely used machine learning techniques: Support Vector Regression, Genetic Programming, M5 Rules, M5 Model trees, Radial Basis Neural Networks, and k-Nearest Neighbours. Daily rainfall datasets were collected from 42 cities of two continents, with very diverse climatic features. 20 cities were selected from around the Europe and 22 from around the USA. There were two reasons of choosing two continents for data extraction, first is to perform the experiment on different climates having diverse weather and second was the geographical locations as the selected cities were very far apart from each other. The ultimate goal was to not bias the experiment to particular climate type or for particular geographic location. According to results the accumulating rainfall amounts can bring good results as compared to prediction using daily rainy data. While using the accumulated data, Support Vector Regression, Radial Basis Functions, and Genetic Programming overall performed well however Radial Basis Functions performed better then modern technique of “Markov chain”. For all selected datasets, each technique used the same parameters so it was not guaranteed that the best possible set of parameters was used for all the techniques. During the experiment, the researchers have noted

a relationship between predictive accuracy and climatic attributes such as: volatile nature of rainfall, amount of maximum rainfall and the interquartile range of rainfall. Moreover no significant difference was noted in algorithms' prediction error among the cities of both the continents (USA and Europe). Issue regarding the discontinuity in rainfall data was solved with the help of accumulated rainfall amounts.

C. A Hybrid Model for Statistical Downscaling of Daily Rainfall

Authors in [32] proposed a hybrid technique to downscale the daily rainfall by integrating two methods: 1) Random Forest, and 2) Support Vector Machine. RF was selected due to its robustness in classification and it was used to predict whether it will be rain or not whereas SVM were selected due to its feature to fit in non-linear data and it was used to predict the amount of rainfall, if it will occur. The proposed model was evaluated by downscaling daily rainfall at three stations, Dungun, Besut, and Kemaman on the east cost of peninsular, Malaysia. Daily rainfall time series data from 1961 till 2000 was collected from Department of Irrigation and Drainage Malaysia. Total of 26 climatic features were collected from National Centre for Environmental Prediction re analysis dataset, which were used as predictors for downscaling the models. To assess homogeneity in rainfall time series, various quality control activities were performed. Histograms for the dataset were created to reflect the problems moreover Student's t test was also used to identify any variance in the means between two segments of dataset which finally found homogeneous at all three locations. According to results the hybrid technique is capable to downscale the rainfall with Nash-Sutcliff efficiency within range of 0.90-0.93, which is much higher than RF and SVM models.

D. Prediction of Monthly Rainfall in Victoria, Australia: Clusterwise Linear Regression Approach

In [33], researchers presented a technique named Clusterwise Linear Regression for monthly rainfall prediction in Victoria, Australia. The proposed CLR is an integrated method of clustering and regression techniques. CLR incrementally extracted the subsets from dataset and then those subsets could be easily estimated with linear function one by one. Dataset which was used for prediction obtained from eight different weather stations for the period of 1889 - 2014 and consisted of five meteorological variables. The selected weather stations included three from east region, two from central region and three from the west region of Victoria. The ultimate goal for the selection of geographical apart stations was to evaluate the performance of proposed model on multiple locations having different atmospheres. The metrological variables which were used as predictors included Vapor Pressure, Solar Radiation. Evaporation, Minimum Temperature, and Maximum Temperature. This proposed technique was compared with following: SVM Reg, ANNs, CLR with CR-EM, and MLR. The model was developed first for each weather station with each technique using training data and then evaluated with test data. To analyze the performance of proposed technique, observed and predicted rainfall measures were compared and four accuracy parameters were used for evaluation: Mean Absolute Scaled Error, Mean Absolute Error, Root Mean Squared Error, and coefficient of efficiency. According to results, the proposed technique outperformed other prediction methods in most of the locations.

E. Prediction and Anomaly Detection of Rainfall using Evolving Neural Network to Support Planting Calender in Soreang (Bandung)

Authors in [34] proposed Evolving Neural Network for the prediction and anomaly detection of rainfall to Support Planting Calendar in Soreang. Dataset was obtained from Department of Agriculture and Department of Water Resources spanning from 1999-2013. The proposed ENN used Artificial Neural Networks and Genetic Algorithm to identify the best weights and biases. The proposed framework consisted of various steps starting from the obtaining of raw data which then gone through the pre-processing phase which consisted of following steps: Integration, Transformation, Reduction and Cleaning of data. Dataset was divided in three scenarios: scenario 1 as dry season from April to September, scenario 2 as wet season from October to March and scenario 3 as the complete data from January to December. Each scenario was further sub divided for training and test data as 9, 12, 14 years for training data and 6, 3, 1 years for testing data, respectively. Learning process of proposed framework used integrated technique and then the result was used for rainfall prediction and anomaly detection followed by the final result which was the predicted starting time for planting. The starting week of January, April and October was selected as beginning time for planting activity in year 2014. According to results, by using all data from 1999-2013 shown the accuracy of 84.6%, for dry season the reported accuracy was 66.02% and for wet season the accuracy was 79.7%.

F. Rainfall Prediction: A Deep Learning Approach

In [35], authors presented a Deep Learning based architecture to predict the daily accumulated rainfall for next day. Proposed architecture consists of two techniques: Auto encoder Network and the Multilayer Perceptron Network. Auto encoder is an unsupervised network which performed the feature selection activity and the Multilayer Perceptron Network was assigned the classification and prediction tasks. Dataset for prediction was obtained from Instituto de Estudios Ambientales (IDEA) of Universad Nacional de Colombia which is located in Manizales Colombia. Dataset spanned from 2002 to 2013 and consisted of 47 weather attributes. IDEA extracted the data from a meteorological station located in the central area of same city and stored in an environmental DWH. As ETL steps were performed on data so pre processing was not needed. Obtained 2952 data samples were classified into subsets for the purpose of training, validation and testing, with 70%, 15% and 15%, respectively. Normalization process was then performed to keep the values of data in to the range of 0 to 1 for better working. Results of the experiment were compared with other methods such as: naive approach which predicts the accumulated rainfall of $t - 1$ for t, MLP with optimized parameters for training & validation set and with some other published techniques. Performance was evaluated in terms of measurement errors: Mean Square Error and Root Mean Square Error.

G. A novel approach for Optimizing Climate Features and Network Parameters in Rainfall Forecasting

Authors in [36] presented a Genetic Algorithm-based approach to identify the best combination of input features and Neural Network parameters to achieve most accurate result. Dataset for prediction spanning of 107 years, from 1908 to 2015, taken from Innisfail, Queensland, Australia and consisted of various weather attributes including rainfall values, mean maximum temperature, mean minimum temperature, and Southern Oscillation Index etc. Data went through a preprocessing stage where couple of tasks was performed. In preprocessing, missing values were replaced with the mean of that attribute and when not applicable the value of that record was taken from closely available weather station. Genetic algorithm usually picks the best chromosome from last iteration but in proposed approach it is customized to select the best chromosome in each of the iteration. The best network which was saved in current iteration was compared to the other generated networks in each coming iteration. The proposed model reflected the highest scores, when compared to climatology and alternative selection methods. Selection of Climatic attributes and network parameters by using proposed hybrid genetic algorithm reflected better performance with 141.67 mm RMSE for a location with 3553.0 mm annual average rainfall whereas climatology, climate input parameters selection-based genetic algorithm, and climate features selection-based genetic algorithm showed 200.32, 171.34, and 178.22 mm consecutively.

H. Early Prediction of Extreme Rainfall Events: A Deep Learning Approach

Authors in [37] presented a framework for the prediction of extreme rainfall by using past climatic features. The proposed model consisted of following phases: Feature Learning, Feature Compression, and the classification process. Stacked Auto-encoder was used for the compression of feature-set. Support Vector Machines and Neural Network were used for classification. Parameters of selected classifier were tuned for the best performance and the issue of biased dataset was dealt effectively by Cost-Sensitive SVM. Presented technique showed the ability to predict extreme rainfall before 6 to 48 hours from occurrence; however some false positives were also reported. The proposed technique also reduced the false alarms which were raised due to the rainfall in surroundings. This method had the capability to generate warnings for rain in surroundings as well. Dataset for rainfall prediction was collected from National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR), for the following months: June, July, August and September. The obtained dataset spanned from 1969 to 2008 for Mumbai, and from 1980 to 2000 for Kolkata. Rainfall data was also obtained for the same period from India Meteorological Department. Weather variables for prediction were taken for entire Indian sub-continent region which was divided in to 255 grids. Total of 21 variables were obtained for each grid; 4725 for entire region (255*21) in case of daily data which could further increased in case of 24 h and 48 h data. The results of experiment were compared with other methods from literature and found the proposed one much better.

V. Results and Discussions

Eight research papers are finally shortlisted by applying the literature extraction criteria, explained in Section III. Below are the answers of Research Questions which are extracted during in-depth analysis and review of shortlisted papers.

RQ1: Which data mining techniques are used/proposed for rainfall prediction?

Authors in all selected papers presented customized/integrated/modified mining techniques for effective rainfall prediction. In each research, multiple climatic attributes/variables from past weather data were used as predictors for the purpose of prediction/forecasting. The ultimate purpose of each research was to increase the accuracy of rainfall prediction. Detail review of selected papers is available in previous section.

RQ2: How the performance of prediction techniques is evaluated?

The selected papers [30]–[37] have compared the proposed technique/model with one or more published techniques. The performance was evaluated by comparing the predicted results with the observed (actual) measures. Information retrieval metrics and statistical techniques were used for performance analysis of proposed techniques in comparison with other methods from the literature.

RQ3: Which type of data is used for prediction?

Each of the selected paper used past weather data for rainfall prediction and for the training purpose of used supervised data mining techniques. Un-supervised data mining techniques were also used in combination of supervised techniques. Various climatic attributes were used as predictors including rainfall polarity, rainfall measure, minimum temperature, maximum temperature, wind speed, and humidity etc. According to researchers, using more features is not the guarantee for more accuracy in prediction instead irrelevant attributes could affect the performance. So the combination of relevant attributes is needed for accurate rainfall prediction moreover these combinations varies upon case to case.

RQ4: For which location the rainfall prediction is performed?

According to shortlisted articles, rainfall was predicted in locations situated in India, Australia, Columbia, Indonesia, Malaysia. USA and Europe.

RQ5: Which factors affect the prediction results?

After the critical review of shortlisted papers, it has been observed that following factors could affect the rainfall prediction results: Past weather data: which is selected for training the mining algorithm, climatic attributes: which are used as predictors, location: for which the rainfall prediction has to be performed, surrounding environment, pre-processing techniques and most importantly the used model/technique/method.

RQ6: Which are the latest research trends in the domain of rainfall prediction?

The ultimate goal of all the shortlisted articles was to improve the prediction accuracy, for this purpose some researchers have explored the correlation among weather features and prediction accuracy and tried to find the best combinations of those features to tune the performance. Few researchers on the other hand worked to train the mining technique well to achieve the high accuracy in prediction. Few have compared the modern techniques with the conventional ones. However most of the researchers presented/used integrated techniques and focused on using the combination of two or more techniques for prediction and claimed that this could bring more accurate results. Each research has provided the justification for the presented/proposed/used technique by means of performance evaluation through quality metrics.

A. Limitations of Research:

This research has the following limitations.

1) The literature was extracted with a rigorous and thorough research process which indicates the quality and completeness of this study however some important relevant work might have been missed.

2) Most of the integrated and modified techniques were evaluated by authors, so the real results may not be as accurate as explained. This may affect the analysis and results of this study.

VI. Conclusion and Future Work

Rainfall prediction is a beneficial but challenging task. Data mining techniques have the ability to predict the rainfall by extracting and using the hidden knowledge from past weather data. In the last decade, many researchers have worked to increase the accuracy of rainfall prediction by optimizing and integrating data mining techniques. Various models and techniques are available today for effective rainfall prediction but still there was a lack of a compact literature review and systematic mapping study which could reflect the current problems, proposed solutions and the latest trends in this domain. This research provided a comprehensive systematic mapping as well as the critical review of latest research from 2013 till 2017 in the area of rainfall prediction by focusing on data mining techniques. In this research a list of significant research questions was identified and then a systematic research process was followed to extract and shortlist the most relevant research articles from renowned digital search libraries. Answers of the identified questions were explored by critically reviewing the shortlisted articles. The research focus on the domain of rainfall prediction has been increasing since last decade and so are the problem areas. So it was concluded that enhancements, optimizations and integrations of data mining methods are vital to explore and solve these problems.

References

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

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

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

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

[5] N. Tyagi and A. Kumar, "Comparative analysis of backpropagation and RBF neural network on monthly rainfall prediction," Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 1, 2017

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

[7] M. Ahmad, S. Aftab, and I. Ali, "Sentiment Analysis of Tweets using SVM," Int. J. Comput. Appl., vol. 177, no. 5, pp. 25-29, 2017

[8] C. S. Thirumalai, “Heuristic Prediction of Rainfall Using Machine Learning Techniques,” no. May, 2017.

[9] 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.

[10] 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.

[11] 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.

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

[13] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, “Lessons from applying the systematic literature review process within the software engineering domain,” J. Syst. Softw., vol. 80, no. 4, pp. 571–583, 2007.

[14] B. a. Kitchenham et al., “Preliminary guidelines for empirical research in software engineering,” IEEE Trans. Softw. Eng., vol. 28, no. 8, pp. 721–734, 2002.

[15] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3," Engineering, vol. 45, no. 4ve, p. 1051, 2007.

[16] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, “Systematic Mapping Studies in Software Engineering,” 12th Int. Conf. Eval. Assess. Softw. Eng., pp. 1–10, 2008.

[17] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering – A systematic literature review,” Inf. Softw. Technol., vol. 51, pp. 7–15, 2008.

[18] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, “Sentiment Analysis using SVM : A Systematic Literature Review,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 182–188, 2018.

[19] F. Selleri Silva et al., “Using CMMI together with agile software development: A systematic review,” Inf. Softw. Technol., vol. 58, pp. 20–43, 2015.

[20] F. Anwer and S. Aftab, “Latest Customizations of XP : A Systematic Literature Review,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 12, pp. 26–37, 2017.

[21] S. Ashraf and S. Aftab, “Scrum with the Spices of Agile Family : A Systematic Mapping,” I.J. Mod. Educ. Comput. Sci., vol. 9, no. 11, pp. 58–72, 2017.

[22] S. Ashraf and S. Aftab, “Latest Transformations in Scrum: A State of the Art Review,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 7, pp. 12–22, 2017.

[23] 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,” J. ICT Res. Appl., vol. 11, no. 2, p. 168, 2017.

[24] 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.

[25] 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.

[26] 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.

[27] D. Gupta and U. Ghose, “A Comparative Study of Classification Algorithms for Forecasting Rainfall,” pp. 0–5, 2015.

[28] 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.

[29] M. Ahmad, S. Aftab, and S. S. Muhammad, “Machine Learning Techniques for Sentiment Analysis: A Review,” Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, pp. 27-32, 2017.

[30] P. Singh, “Indian summer monsoon rainfall (ISMR) forecasting using time series data: A fuzzy-entropy-neuro based expert system,” Geosci. Front., vol. 2002, 2017.

[31] 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.

[32] S. H. Pour, S. Shahid, and E. S. Chung, “A Hybrid Model for Statistical Downscaling of Daily Rainfall,” Procedia Eng., vol. 154, pp. 1424–1430, 2016.

[33] 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.

[34] Gunawansyah, T. H. Liong, and Adiwijaya, “Prediction and anomaly detection of rainfall using evolving neural network to support planting calender in soreang (Bandung),” 2017 5th Int. Conf. Inf. Commun. Technol. ICoIC7 2017, vol. 0, no. c, 2017.

[35] F. Martínez-Álvarez, A. Troncoso, H. Quintián, and E. Corchado, “Rainfall Prediction: A Deep Learning Approach,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9648, pp. 151–162, 2016.

[36] A. Haidar and B. Verma, “A novel approach for optimizing climate features and network parameters in rainfall forecasting,” Soft Comput., 2017.

[37] S. G. B, S. Sarkar, P. Mitra, and S. Ghosh, “Early Prediction of Extreme Rainfall Events: A Deep Learning Approach,” vol. 9728, pp. 154–167, 2016.

[38] 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.

[39] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, "Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 4, 2017.

[40] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, "Rainfall Prediction in Lahore City using Data Mining Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, pp. 254-260, 2018

[41] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, "SVM Optimization for Sentiment Analysis," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, 2018.