Enhanced Rainfall Predictions using Stacking Technique

[1]Pamulapati Sri Madhu Chowdary, [2]Anbarasi M

[1]Student, [2]Professor,
[1]Computer Science and Engineering,
[1]Vellore Institute Of Technology, Vellore, India.

Abstract : Rainfall is the most crucial process of nature. All the living beings rely on water and rainfall is a process that is responsible for the continual process of the water cycle. Many human activities like agriculture are dependent on rainfall, especially in a country like India. Thus, it is essential and necessary to predict the rainfall patterns to estimate the flooding and drowning events. Application of data mining algorithms is the best way to forecast rainfall. This paper outlines the use of stacking technique to predict rainfall. In stacking technique, there are two layers - the first layer is considered as the meta layer and the second layer is considered as the base layer. Classification techniques such as Bayesian, decision tree, support vector machine, and K nearest neighbor are used in the base layer and in the meta layer, dip neural networks provide greater accuracy for ensemble techniques.

Index Terms **- Stacking; Data Mining; Ensemble; Rainfall; Deep Neural Network.**

I. INTRODUCTION

Data mining is a process of mining data from a database with specific techniques and process. It consists of identifying the pattern in the given dataset (often referred as training sample), or analyzing a set of already classified objects, whose results can be used to predict the outcomes of other data with missing attributes (often referred as test sample). It aims at the accurate analysis of data and generation of precise results. This analysis and prediction of data in data mining can be done using multiple methods such as classification, clustering, and regression to mention a few. Algorithms like stacking can make rainfall prediction reliable with the techniques such as artificial neural networks, fuzzy logic, decision tree, deep learning as they provide a methodology to classify and predict data better than the traditional statistical techniques.

Stacking is a classification method in which the output of one level classifier is used as input for the next level. Stacking also uses meta-learners to combine predictions of base learners, where the base layer (level-0 model) is the input for the meta layer (level-1 model). Stacking technique has been chosen as the preferred technique for the study after evaluating the data set with different measures of the model. This method provides a better ability to relate the attributes within the dataset.

Knowing the rainfall patterns helps in drought and flood management; however, considering suitable attributes for the prediction is not an easy task. The intensity and frequency estimation from the data set is very useful for an agriculture-dependent country like India.

In this paper, values for atmospheric variables like pressure, temperature and humidity are considered and the dataset looks at these values for a period of one year. This study uses the base layer as deep neural network. The study uses decision tree, CART, Support Vector network (SVM), K-nearest neighbor (KNN) and Bayesian algorithm; the second model utilizes deep learning / deep neural networks.

II. LITERATURE SURVEY

(S. Kannan, et.al. 2011) states that K-means clustering technique is used to describe the variation in the patterns of rainfall in a region and how the clusters can be used to predict the possibility of rainfall according to the atmospheric conditions. Besides, supervised classification and regression tree techniques are also used on the data to get the rainfall states. After the application of the above techniques, various clusters were obtained and a relationship was established between them. (B. Kavitha Rani et.al 2016) explains about uses of both statistical methods and data mining techniques to predict rainfall from the database. It involves data-driven models like a multi-layer neural network and time series modelling involving crucial temporal dimensions. In addition to this, parameters like root mean square, the coefficient of correlation, absolute mean error are also obtained from the data. These parameters are used to compare the results obtained from the above data mining techniques. As best practice, almost a third of the data set was used for testing purposes. The paper concludes that prediction of rainfall using back propagation neural network is entirely accurate while taking additional inputs like sea surface temperature mainly around southern parts of India.

In this paper (Valmik B. Nikam et al. 2013), the authors explain about proposed rainfall prediction model that was made to undergo data pre-processing to extract seven attributes relevant to rainfall prediction from a total of given 36 attributes. The values for the variables were a result of high-performance computing models. This paper implemented naïve bayes classification for data demanding model which offers the best accurateness, using various properties for computation. (Jyothis Joseph et al. 2013) explained about the two main types of data mining approaches – Empirical and dynamic. This paper outlines the implementation of a combination of various classification and clustering techniques to predict rainfall occurrence. This further applies neural network and neural network Bayesian regularization on the implementation. (R.Senthil Kumar et al. 2016) indicates that rainfall prediction model implementation must be estimated statistically. In this, authors mainly focused on K nearest neighbour, naive bayes, neural network and did a comparative study between them. The paper also included the recent algorithms like SLIQ, ANFIS, ARIMA. It also analyzed which method gave the prediction a better accuracy. It also explains how the accuracy of results changes based on the size of the dataset; although the volume of the dataset increases the accuracy initially, it decreases after a certain limit. In Khaled (K. Hammouda et al. 2013) explains about various clustering techniques that can be used. This includes mountain clustering, fuzzy C clustering, and subtractive clustering. In the method of mountain clustering, the mass function is designed at each place in the data set and the situation with more mass is chosen as the center of the original cluster. Similarly, the center of other groups is identified after destroying the effect of its previous clusters. This iteration continues until the required clusters

formed. In this method of reductive clustering, the places of each data point are used to evaluate mass function which reduces calculation time. (M. Kannan et al. 2010) explains that rainwater is more significant for producing food, drinking water and all activities in the landscape. Due to heavy rain or dry period in critical stages crop growth may reduce crop yield. Rainfall prediction is more significant in agricultural countries like India where the monsoon rainfall is also highly complex. Using this prediction method, we can forecast rainfall for our states.

(Shoba G et al. 2015) explained there are many algorithms for rainfall prediction. Here various algorithms were compared to find the suitable candidate for rainfall prediction. These algorithms can result in deciding or predicting a model based on a given data set. (M.S.Chaudhari et al. 2015)expresses on the accurate and exact estimated rainfall prediction and states that all techniques in data mining are not ideal to predict rainfall. Data mining is a much-needed opportunity to deliver information for stakeholders and decision makers. Rainfall estimation and prediction varies based on the different techniques. (Dhawal et al. 2016) explained the most used algorithm for estimation is deterioration analysis. They conclude that the method which gives more significant result is artificial neural networks amongst various other methods that have been used in many previous prediction studies. (Emilcy Juliana Hernandez et al. 2016) explains about deep learning for the estimation of rainfall even used in cases where prediction is possible for the next day. It also includes aspects about autoencoder for dipping and this does not require a relationship among the attributes. (Zeyi Chao et al. 2018) explains how to apply MEMS sensors in real-time rainfall prediction. It used algorithm called as Seasonal trend decomposition using Loess (STL)and also compared with support vector machine (SVM), random forest (RF). (Shabib Aftab et al. 2018) explains about the challenging tasks in rainfall prediction. Exact rainfall estimation can be useful to yield actual methods in expansion observing: carrying actions, cultivated duties and even in aircraft navigations. Algorithms can successfully estimate the rainfall by unseen mining patterns between existing structures from earlier data and the recent data set. This will provide a rainfall estimation and also a reference point for future guidelines and comparison.

(Yu Xiang et al. 2018) explains that several estimating algorithms are planned in the current years. The information holding long and short deviation private innovative rainfall analysis is traveled using ensemble empirical mode decomposition according to the inquiry on three rainfall data together by forecasting places situated in Andhra Pradesh, Tamilnadu and Kerala. By analyzing estimation & exactness along with time & efficiency, the model created information mined using the decomposition method. This method assumes several directed ways for few workings of input data, which uses SVM for short-time jobs of estimation, where NN is used for a longtime working estimation. Researchers show better operations compared to old techniques which offer the latest methods in the rainfall analysis area.

(Changhwan kim et al. 2018) explains about the parity over several rainfall situations that were valued through a rough explanation created on a kinematic idea. The overflow of hydrographs was pretended using a kinematic wave overflow method. Rough and added outcomes are used to explore the relationship between altitudinal–progressive rainfall features and rainfall altitudinal determination. Assumed outcomes were, it is to establish that (i) essential rainfall evaluation is majorly to predict the accuracy. (ii) The rainfall time besides rainfall altitudinal flow is a central influence for defining an essential rainfall altitudinal firmness for exact overflow estimation. (iii) Rainfall altitudinal spreading period which is previously ignored should be measured when taking the additional measurement as the most factor. (Zhifeng et al. 2018) explain load estimation severe power system arrangement and effective decision making. This estimating method depends on deep learning, this solution will establish the following (1) deep learning depending on the methodology that results in precise calculations taking the current consumption into consideration for forest and ascent boosting methods (2) Possibility mass estimation that will enable estimating high value to determine duration.

III. METHODOLOGY

This paper uses a stacking technique for rainfall estimation, where the meta layer is used for input and base layer is used for output. In meta layer, inputs are given as decision tree, support vector machine (SVM), K nearest neighbours (KNN), naive Bayesian classification and in base layer output is provided as meta-learning. Through the stacking ensemble approach, two strategies are used for combining the models. The other plans in ensemble methods are bagging and boosting. Bagging has n models and uses similar methods. In a strange case, every method is estimated. Boosting when compared to bagging, is a method that is used to make changes. Occurrences that are regularly misidentified are permitted to get more analysis worked out. There are n classifications, where between themselves, a discrete weightage for their exactness is applied and the preference is given to the highest weightage. Bagging is best when compared to boosting. The reason for this is because boosting agonizes which is not relevant when relevance provides a better data set. This is because a trained dataset is best in this scenario. There are two ways of using joining techniques - Voting or Stacking. The difference among stacking and voting is: voting does not take place at the meta layer, as the final classification is absolute by usual votes by the base layer classification whereas in stacking meta layer take place.

Stacking has combined method for different techniques produced by changed methods-D1, D2,..Dn on a single data. The first layer in base layer methods M1, M2...Mn is produced. The second layer in meta layer methods is combined by base layer classification. Meta layer dispersed in the mining field is the frequently used ensemble techniques. Among them, the approach to combine methods is referred as the so-called meta-decision trees, where it compacts joining one level of classification as decision trees. In this there are various approaches for combining classifications like voting, bagging and boosting. Stacking is an equivalent joining classification where all methods are completed in equivalence and it will increase revenue at meta layer. In ensemble of homogenous or heterogeneous classification, this has proved to produce better operation. It offers those are dependent on the particular presentation of top joining different classification it makes best compared to homogenous classification. The better classification between base layers classifications are at certain, respected data providing with each other classification has been unnoticed. Classification ensembles are identified as joining or teams, base layer classification shows which are joined at another mode is stacking. The advantage of stacking is an alternative of picking a definite common at various types. Stacking trusts those for taking given outcomes data as inputs into the latest place. Stacking has common guests at the latest place. The voting approach has its restrictions because of capability in detention direct affiliation.

Stacking, as ensemble classification, is initially qualified by bootstrapping sample data set constructing layer 0 classifiers output as base layer classifier and that taken to qualify a meta-classification. Its target is to achieve the next layer to make sure the given dataset exactly has a process. Merging of models usually modelled in layer 0(base layer) is worked at equivalent to join with next layer classification in a so-called meta layer classification shown in figure 1. Here a decision tree is used and so are SVM, KNN and naive Bayesian classification at base layer classifications, various meta layer classifications are tested and practically deep learning is considered for meta layer classification; this provides the best outcome. These techniques have established, most important distress in erecting ensemble in such a way that the exact mining methods as ensembles are selected. Ensemble classification mostly offers various methods of classification at the base layer and meta layer and are dependent on the level of submission. This paper also recommends effective combination of the base and meta layer classifications to estimate rainfall.

A. Base Level

Decision tree uses a tree-like model for decisions. One way to exhibit classification is to use definite regulator reports. The decision tree is commonly used in decision analysis. The decision tree has a source node and child nodes. Uppermost is a tree node whereas the leaf node represents the class. Benefits of the decision tree are that it does not require any field facts. Classification steps in the decision tree are modest and fast. The decision tree is a flowchart-like structure. Decision tree takes the inputs as a set of working out tuples, associated class labels, set of candidate attributes as the selection method. Tree lopping approach is pre-lopping (the tree is lopped by uncurtaining construction earlier), post-loping (it removes sub-tree from the grown tree). Cost complexity measured by a number of leaves in tree, error rate of the tree.

Support vector method is the methods used for classification and regression. SVM model generates non-overlapping partitions. SVM algorithm gives an output with an optimal hyperplane that categorizes new examples. When data is unlabeled, supervised learning is not possible. SVM algorithm has two steps, and they are separable case (infinite boundaries are possible to different data) and Non-separable case (Two classes are not unrelated, but they join with each other). We can use linear SVM for the outcome of the most significant and smallest edge hyperplane which divides the data when there are two hyperplanes. They are biased hyperplane (which does not pass through the origin of the coordinate system) and unbiased hyperplane (which passes through the origin of the coordinate system). The advantage of SVM is that it is disconnected sensations by choice of an actual kernel mode. It generally provides accurate predictions. Hyperplane gives optimal nearest points but not by detached points.

KNN is a method which provisions all the suitable methods and classifies new cases based on similarity measures like euclidean distance, manhattan, and minkowski. The major drawback in this algorithm is calculating distance measures directly from the data set where variables have different criteria. KNN does not have limitation managing method where it will sort records for taken data. Non-parametric does not give a fixed number of parameters in the model. The advantage of KNN is easy to understand, no assumptions in data, it can be applied for both regression and classification, and it works on multi-class problems. Using KNN in R software is more optimism, pessimism, it is used in the past, at present, for future.

Naive Bayesian methods used for text categorization with word frequencies as the feature. Naive bayesian is an exact processor which have more aggressive at that province with fundamental techniques together with SVM. This method is highly scalable, requiring more number of parameters. Naive Bayes models are knowns as independent bayes or bayesian networks methods. It allows class conditional independencies to be definite between the subdivisions of variables. It gives a graphical model of a normal relationship where learning can perform. Naïve used for trained network classification with a set of conditional probability tables.

B. Meta Level

Deep learning:

Among all old generations, deep learning is the first class of algorithm. Deep learning is scalable which gives better performance. It is a supervised algorithm. Deep refers to the number of hidden layers. Deep learning also called as future learning. Deep learning is the duct of all modules that are trained because it has multiple stages in recognizing an object as shown in figure 2. Deep learning is achieving more result in devices like phones, tablets, tv's. Deep learning requires more substantial amounts of label data. Deep learning needs extensive computing power. GPU's use more deep learning where it combines cloud computing. Deep learning uses most neural networks architecture. Deep learning is more inspired by the biological nervous system. Every succeeding level gives the best outcome through the input that is taken. Deep learning can be supervised (classification) or unsupervised (pattern analysis). Deep learning architecture is created by layer by layer. The algorithms in supervised and unsupervised are 1) Autoencoders, denoising 2) stacked deionizing auto-encoders 3) Restricted Boltzmann machines 4) Deep belief networks. In deep learning, the layers between any two layers are called as hidden layers. Each hidden layer is composed with neurons in neural network. Network consumes large amount of input data; it operates multiple layers increasing complex features of data in each layer.

IV. FLOW CHART

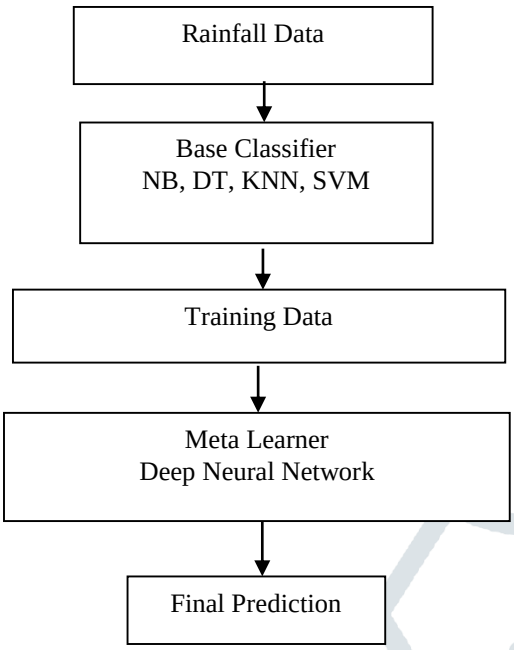

Figure 1: Proposed model for stacking with deep neural network

V. ARCHITECTURE OF DEEP LEARNING

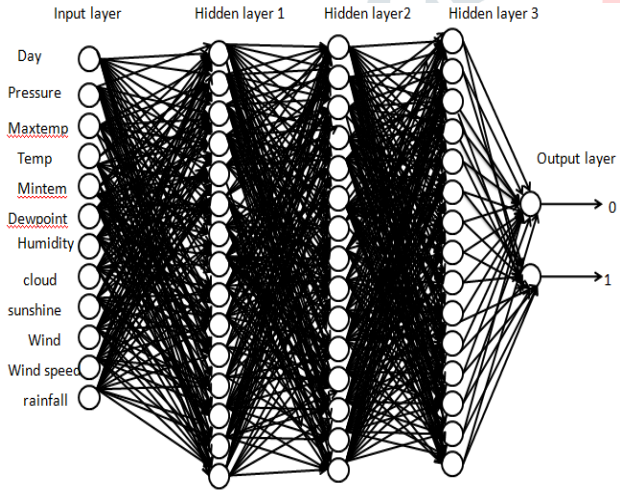

Figure 2: Deep Neural Network

VI. Results And Discussions

TABLE 1: RESULTS

Stacking	Classification Techniques	Accuracy	Classification error	F measure	Sensitivity	Specificity
Single Classifer	NB	78.18%	21.82%	64.71%	62.86%	85.33%
	DT	79.09%	20.91%	58.18%	45.71%	94.67%
	KNN	70.00%	30.00%	52.17%	51.43%	78.67%
	SVM	76.36%	23.64%	63%	62.86%	82.67%
Base Classifier	Deep NN	80.00%	20.00%	58%	42.86%	97.33%

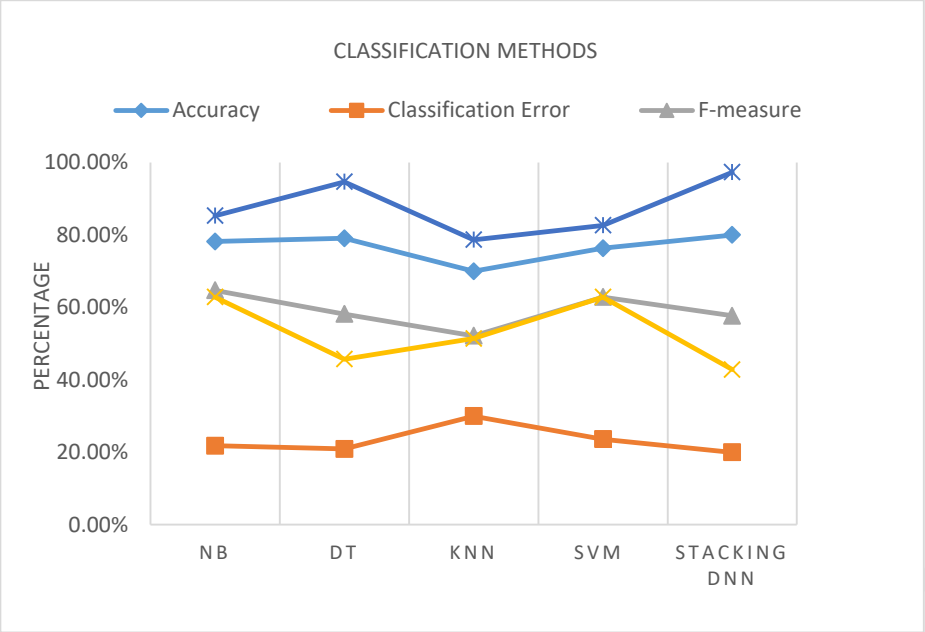

FIGURE 3: Results for Single classifier and stacking with deep Neural network

Discussion:

Analysis of rainfall prediction using stacking method in the base layer and meta layer considered with various classification techniques results shown in figure 3 and table 1 such that for naïve Bayesian classification accuracy is 78.18%, classification error is 21.82%, F-measure is 64.71%, sensitivity is 62.86%, specificity is 85.33%. Decision tree accuracy is 79.09%,classification error is 20.91%,f-measure is 58.18%,sensitivity is 45.71%,specificity is 94.67%.k nearest neighbour accuracy is 70.00%,classification error is 30.00%,f-measure is 52.17%,sensitivity is 51.43%,specificity is 78.67%.support vector machine accuracy is 76.36%,classification error is 23.64%,f-measure is 63%,sensitivity is 62.86%,specificity is 82.67%.in base layer, for deep neural network the accuracy is 80.00%,classification error is 20.00%,f-measure is 58%,sensitivity is 58.00%,specificity is 97.33%.in base layer specificity is more. The study found that the meta layer decision tree has more accuracy, f-measure, knn has more classification error. Naive Bayesian, support vector machine has more sensitivity and decision tree has more specificity.

VII. Conclusion:

In this paper, we predicted the best way to forecast rainfall. Here we used the stacking technique to predict rainfall. In Stacking, there are two kinds of approaches for it. In Stacking, there are two layers, first level considered meta layer and second level base layer we have used four classifiers in meta layer, they are naive bayesian, decision tree, support vector machine, and K nearest neighbour clustering and in base layer we have used deep neural network based on stacking in ensemble technique to improve the accuracy. The effects of rainfall prediction on overflow and rainfall prompted erosion were studied by using the stacking method. We found accuracy, F-measure, sensitivity, specificity all are better in stacking technique for rainfall prediction.

VIII. Acknowledgment

I would like to thank our professor Anbarasi M who guided me through the project with her insights. We would also like to thank our institution for providing us with the equipment and opportunity to conduct this project.

IX. REFERENCES

[1]. Changhwan Kim,Dae-Hong Kim, “Effect of rainfall spatial distribution and duration on minimum spatial resolution of rainfall data for accurate surface runoff prediction”, Journal of Hydro-environment Research, Volume 20, Pages 1-8, 2018.

[2]. Chaudhari M. S., N. K. Choudhari, “Study of Various Rainfall Estimation and Prediction Techniques Using Data Mining”, American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-6, Issue-7, pp-137-139,2017.

[3]. Dhawal Hirani, Nitin Mishra, “A Survey on Rainfall Prediction Techniques”, International Journal of Computer Application (2250-1797) Volume 6. No.2, 2016.

[4]. Hernandez, Emilcy & Sanchez-Anguix, Víctor & Julián, Vicente & Palanca, J & Duque, Néstor, “Rainfall Prediction: A Deep Learning Approach”, 151-162. 10.1007/978-3-319-32034-2_13, 2016.

[5]. Joseph, Jyothis & T K, Ratheesh. (2013). Rainfall Prediction using Data Mining Techniques. International Journal of Computer Applications, 83, pp. 11-15, 10.5120/14467-2750, 2013.

[6]. Kannan S., Subimal Ghosh, "Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output, Stochastic Environmental Research and Risk Assessment", May 2011, Volume 25, Issue 4, pp 457-474.

[7]. Khaled Hammouda, Fakhreddine Karray, "A Comparative Study of Data Clustering Techniques" SYDE 625: Tools of Intelligent Systems Design. Course Project, 2010.

[8]. Kannan, M & Prabhakaran, S & Ramachandran, P, "Rainfall Forecasting Using Data Mining Technique", International Journal of Engineering and Technology, 2, pp. 397-401,2010.

[9]. Rani B, Kavitha & Govardhan Dr. (2013). Rainfall Prediction Using Data Mining Techniques - A Survey. Computer Science & Information Technology. 3. 23-30. 10.5121/csit.2013.3903

[10]. Sendhil Kumar, K.S. and Jaisankar, N., (2017), Towards data centre resource Scheduling via hybrid cuckoo search algorithm in multi-cloud environment, International journal of Intelligent Enterprise, Inderscience publishers, 4(1), pp.21-35.

[11]. Shoba G, Dr. Shobha G(2014), Rainfall Prediction using Data Mining Techniques A Survey International Journal of Engineering and Computer Science, ISSN:2319-7242, 3(5), pp. 6206-6211.

[12]. Valmik B. Nikam, B.B.Meshram (2013), Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach, published in "CIMSIM '13 Proceedings of the 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation", pp 132-136.

[13]. Xiang, Yu & Gou, Ling & He, Lihua & Xia, Shoulu & Wang, Wenyong (2018), A SVR-ANN combined model based on ensemble EMD for rainfall prediction, Applied Soft Computing. 73. 10.1016/j.asoc.2018.09.018.

[14]. Zeyi Chao,Fangling Pu,Yuke Yin, Bin Han, and Xiaoling Chen.(2018), Research on Real-Time Local Rainfall Prediction Based on MEMS Sensors", Journal of Sensors, Article ID 6184713.

[15]. Zhifeng Guo,Kalie Zhou,Xialong Zhang,Shanlin Yang (2018), A deep learning model for short-term power load and probability density forecasting, Energy, 160, 1186-1200.