Application of Data Mining and Machine Learning for Weather Forecasting: A Comprehensive Study

Nasimul Hasan

Nayan Chandra Nath

Department of Computer Science and Engineering
International Islamic University Chittagong

This dissertation is submitted for the degree of
Bachelor of Computer Science and Engineering

June 2016

Dedication

To our parents

Declaration

We declare and guarantee that all data, knowledge and information in this document has been obtained, processed and presented in accordance with academic rules and ethical conduct. No portion of the work contain in this thesis has been submitted in support of any other degree or qualification of this or any other university or institute.

Nasimul Hasan
Nayan Chandra Nath
June 2016

Acknowledgements

We would like to express the deepest appreciation to our supervisor, Lecturer Risul Islam Rasel, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research and an excitement in regard to teaching. As our supervisor, Mr. Risul worked closely with us during the proposal writing and during the period of our dissertation. Without his guidance and persistent help, this dissertation would not have been possible. We would also like to thank Mr. Tanveer Ahsan, Associate professor, Department of Computer Science and Engineering for imparting his pearls of knowledge during the course of Research Methodology and his support and guidance in the way of this long process.

Nasimul Hasan, Nayan Chandra Nath

I wish to express my unqualified thanks to my friend, Md. Taufeeq Uddin. I could never have accomplished this dissertation without his continuous support and inspiration. I also wish to thank my parents for doing their best to understand a son who had to be confined to his study for such a long time and who raised me and taught me to study hard and to give priority in my life to the quest for knowledge.

Nasimul Hasan

I wish to thank my parents for supporting the best way all the time. They support me all the time for continuing my study & everything. They raised me, taught me, and cherished me.

Nayan Chandra Nath

Abstract

In this study, we focus on forecasting weather using Data Mining and Machine learning techniques. Weather forecasting for an area where the weather and climate changes spontaneously; is a challenging task and especially weather is a chaotic system. A strong forecasting system can play a vital role in different sectors like making business and agricultural decisions. Since many other things such as drought, tourism, transportation, construction etc. also sometimes depends on whether; a good prediction can help to take decisions regarding these sectors. Weather is basically a non-linear system because of various components of climate such as humidity, wind speed, sea level, the density of air etc. and they have a great impact on the weather. This paper exhibits the performance of different data mining techniques and proposes a robust weather prediction technique in view of the recent weather data of Bangladesh utilizing Support Vector Regression (SVR), a relapse methodology of Support Vector Machine (SVM). The collected raw data wasn't prepared for using as the input of algorithm, thus, it had been pre-processed manually to suit into the algorithm, then fed to the algorithm. The evaluation results of the study conducted on the data show that the projected technique performs higher than the conventional frameworks in term of accuracy and process running time. The proposed approach yielded the utmost prediction of 99.83% for Rainfall and 99.28% for Temperature prediction.

Table of contents

List of figures

List of tables

Chapter 1

INTRODUCTION

1.1 Background

The climate is the condition of the environment, to the extent that it is hot or cool, wet or dry, quiet or stormy, clear or shady. Most climate marvels happen in the troposphere, just beneath the stratosphere. Weather prediction is one of the most challenging tasks to accomplish as many natural and man-made components are involved in weather change such as change of seasons and greenhouse effect. The main challenge is to predict the weather with the most accuracy. Weather prediction plays a significant role in many components in decision making related to many fields as agriculture, energy management and human and animal health. Climate determining includes anticipating how the present circumstance with the air will change in which display atmosphere conditions are taken by ground recognition, for example, from boats, plane, Radiosondes, Doppler radar, and satellites. The gathered information is then sent to meteorological centers in which the data are assembled, examined, and made into a combination of frameworks, maps, and diagrams. Calculations trade countless onto the surface and upper air maps, and draw the lines on the maps with help from meteorologists. Calculations draw the maps and also foresee how the maps will take a gander eventually later on. The determination of atmosphere condition utilizing calculations is plot as numerical or computational weather prediction [1]. Generally the climate and atmosphere expectation issues have been seen as various disciplines. Numerical Weather Prediction (NWP) is urgently subject to characterizing an exact starting state and running at the most astounding conceivable resolutions [2], while atmosphere prediction has tried to incorporate the full multifaceted nature of the Earth framework keeping in mind the end goal to precisely catch long time-scale varieties and inputs deciding the present atmosphere and potential atmosphere change. The idea of a unified or seamless structure for climate and atmosphere expectation has pulled in a lot of consideration in the most recent couple of years [3] [4][5][6][7][8][9]

The field of Data Mining and Machine learning has progressed rapidly over the last few decades. Predictive analysis has gone to a very new level with the use of machine learning techniques. Weather data used in this study data are dependent on their nature and thus, their estimation is not effectively made with numerous quantitative methodologies. However, they can be portrayed, estimate and arranged quantitatively by utilizing probability theory. The goal of this paper is to find the pattern of weather of Chittagong, Bangladesh and predict the weather. Moreover to find the perfect combination of data to predict the weather. We aim to tackle these challenges via a representation that jointly predicts rainfall and temperature across space and time. The study combines a bottom-up predictor for each individual variable with a Support Vector Regression and an Artificial Neural Network model to determine an effective and efficient model that models the joint statistical relationships. The two Machine learning algorithms are well known and strong algorithms. So, we conducted a comparative study between the algorithms. Another key component in the framework is a data-driven kernel, based on a similarity function that is learned automatically from the data. We propose an ANN based model that is highly efficient for temperature prediction and a SVM/ANN based model for forecasting rainfall.

The main contributions of this work can be summarized as follows:

1. We present an efficient model with discriminative and generative components form spatiotemporal inferences about the weather.

2. We design and implement a data-driven kernel function that shapes predictions in accordance with physical laws.

3. We present two different model to forecast rainfall and temperature

4. We compared two commonly used methods to find the best to build a better model

5. We evaluate the methods with a set of experiments that highlight the performance and value of the methodology.

6. We determined the correlation between the features and presented their importance respectively.

In this chapter, we discussed our motivation, objective, scope of the study and utilizations of the study. Chapter 2 includes the Literature Review, where we discussed previous works regarding the problem we worked on and some basic study of our two algorithms.

1.2 Motivation

The weather has a great significance in our day to day life. Lots of prediction model has been developed so far for weather forecasting. Weather forecasting has a great impact on life especially in the countries which depend mostly on agriculture and weather change is a very common and continuous event. Several numerical methods had been used so far for predicting weather previously. But, because of phenomena within the world climate, like the greenhouse effect, conventional methods may come deficient due to lack of adaptation. We wanted to study the trend of weather in Bangladesh. The main motive was to find out the pattern of weather and to develop a feasible prediction method. Another motivation behind this work is to know which data combination is best to predict the weather (Rainfall and Temperature).

A flawless weather forecasting can help us to take choice what to develop and what to not. Since numerous different things, for example, drought, tourism, transportation, development and so forth moreover some of the time relies on upon weather; a great forecast can take choices with respect to these divisions. From the antiquated time, people are attempting to discover the example of climate and foresee climate for their prosperity. From the earliest starting point of science and innovation, climate forecast is an exceptionally fascinating part of the study. Forecasting weather is an intense task because of the entanglement of the material science and distinctive variable which cause precipitation [10].

1.3 Objective

The climate of Chittagong is described by tropical storm atmosphere. The dry and cool season is from November to March; the pre-storm season is from April to May which is exceptionally hot. The sunny and the rainstorm season is from June to October, which is warm, overcast and wet.

1. To propose a weather time series prediction model using Support Vector Machine Regression (SVR) and Artificial Neural Network.

2. To evaluate the model's prediction results with real time dataset.

1.4 Scope of the Study

1. The models used here can be used to study weather of different places.

2. The created model can be utilized for rainfall and temperature analysis in weather forecasting.

3. Proposed model can be connected to monthly and daily prediction in weather forecasting.

4. For this study, Data is gathered from the Bangladesh Meteorological Department (BMD) such that research result can be estimate and evaluated.7 years (2008-2014) raw datasets are gathered and processed then the data isolated into two section, training dataset(2008-2013) and testing dataset(2014).

5. Further research can be done based on this study to improve the weather prediction accuracy.

1.5 Utilization of the Study

The proposed model will be valuable for weather prediction regarding rainfall and temperature time series data. It's like

1. User will be capable to predict the rainfall and temperature data using this model.

2. The proposed model will individuals to farmer settle on choices when they will plan for farming in different season and what weather will be.

3. Weather forecasting department can utilize this model so as to deliver graphical representation of the rainfall and temperature data also make decision what weather will be in next day or next month.

Chapter 2

LITERATURE REVIEW

2.1 Weather Prediction using Machine Learning

With the rapid and impressive improvement of machine learning; its use on a various problem has increased in a significant way. A huge amount of research on this topic has been designed depending on ML techniques. Accurate precipitation forecasts can reduce forcing uncertainty in hydrological (e.g. rainfall-runoff) models and can greatly improve the quality of streamflow forecasts. However, NWP precipitation forecasts are inherently uncertain and subject to three types of error [11] Most of the work in weather forecasting to date depend on the utilization of generative methodologies, where the climate frameworks are recreated through numerical techniques. [12][13][14] or rely on time-series analysis like ARIMA models and simple classifiers based on Artificial Neural Networks [14][15][16][17][18] Regardless of the accomplishment of machine learning in an assortment of undertakings, applications to the issue of climate determining have been constrained. Special cases incorporate the utilization of Bayesian Networks for precipitation prediction [19] and temporal demonstrating by means of Restricted Boltzmann Machines (RBM) [20][21]. A different string of exploration has additionally centered around effective representation of social spatiotemporal data in Random Forests for expectation of serious surface-level climate forms, V. M. Krasnopolsky and M. S. Fox-Rabinovitz introduced a new practical implementation of NN where they formulated e a new paradigm in environmental numerical modeling. They proposed a new kind of a complex hybrid environmental numerical model, an HGCM—based on a synergetic combination of deterministic modeling and the machine learning techniques within such a model [14] Andrei Bautu and Elena Bautu [22] introduced Genetic Algorithm based Meteorological Data analysis technic with a very low error rate. Gwo-Fong Lin and Lu-Hsien Chen [23] developed an NN with two hidden layers to forecast typhoon. Their model is capable of forecasting rainfall when a typhoon is nearby. Soo-Yeon Ji, Sharad Sharma, Byunggu Yu,

and Dong Hyun Jeong [24] proposed a rule based hourly rainfall prediction system. They used CART and C4.5 to generate rules. Their model can predict average accuracies of the chance of rain using CART and C4.5 are 99.2% and 99.3%, respectively.

Artificial Neural Network (ANN) is a commonly used technique to produce a good prediction [25] [26][27][23]. Frequent study shows that ANN can work way better than different regression techniques, MA, and EMA. ANN frequently displays conflicting and unpredictable execution on boisterous data [28]. Support Vector Machine (SVM) is a supervised classification and regression algorithm. An SVM model represents the samples as point spaces. It is mainly based of decision plane concept. It separates objects with different classes by a visible gap as large as possible between the classes. SVM can perform both linear and nonlinear classification with good efficiency. Kesheng Lu and Lingzhi Wang [29] showed in their research that Support Vector Machine can perform an efficient rainfall prediction with a low error rate. Their model can produce almost 99% accurate prediction. A. Mellit, A. Massi Pavan & M. Benghanem [30] developed a SVM model which can produce up to 99% accurate prediction for different models. At the point when utilizing SVM, the fundamental issue is defied: how to pick the appropriate kernel and how to set the best kernel function. The best possible parameters setting can enhance the SVM relapse exactness. Diverse kernel capacity and distinctive parameter settings can bring about huge contrasts in execution. However, there are no investigative strategies, on the other hand, solid heuristics that can direct the client in selecting a fitting part capacity and great parameter values [29]. Perez, and Richard [31] combined three independent validations of global horizontal irradiance (GHI) multi-day forecast models that were conducted in the US, Canada, and Europe. Hall and Tony [9] proposed A neural network model using input from the ETA model and upper air soundings for the probability of precipitation (PoP) and quantitative precipitation forecast (QPF) for the Dallas-Fort Worth, Texas area. Their model forecasts with over 70% of the PoP forecasts being less than 5% or greater than 95%. The forecasts of less than 5% PoP were always associated with no rain and the forecasts of greater than 95% PoP were always associated with rain.

2.2 Time Series Analysis

A time series is a succession of perceptions which are requested in time (or space). In the event that perceptions are made on some marvel all through time, it is most sensible to show the information in the request in which they emerged, especially since progressive data will likely be reliant. Time series are best represented in a scatter plot. The series value

X is plotted on the vertical axis and time t on the horizontal axis. Time is known as the independent variable. There are two kinds of time series data:

1. Continuous, where we have an observation at every instant of time
2. Discrete, where we have an observation at (usually regularly) spaced intervals.

There are two fundamental objectives of time series analysis.

- Recognizing the nature of the phenomenon represented by the sequence of observations, and
- Forecasting

Both of these objectives require that the example of observed time series data is distinguished and pretty much formally depicted. Once the example is set up, we can interpret and integrate it with other data. Regardless of the the profundity of our comprehension and the legitimacy of our understanding (hypothesis) of the marvel, we can extrapolate the recognized example to foresee future occasions.

2.3 SVR

Support Vector Regression (SVR) is a regression technique developed by Vapnik [32], it uses almost the same principles as SVM's do. The principle method of SVM is to map the training data from the input space into a higher dimensional feature space through a function ϕ and draw a hyper plane with maximum margin in the feature space. Given a training set of data $x_i \subset R_n, i = 1,....,l$, where l corresponds to the size of the training data and $y_i = +-1$ class labels, SVM will find a hyper plane direction ω and an offset scalar b such that $f(x) = \omega * \phi(x) + b \geq 0$ for positive examples and $f(x) = \omega * \phi(x) + b \leq 0$ for negative examples. Subsequently, in spite of the fact that we can't locate a linear function in the information space to choose what sort the given information is, we can locate an ideal hyper plane that can precisely separate between the two sorts of information [33]. If we have a training data $\{(x_1, y_1), ..., (x_i, y_i)\}$ where all $x_i \subset R^n$ represents the input space and has a related target value $y_i \subset R^n$ for $i = 1,....,l$ where l represents the size of training data [34][35] ε -insensitive loss function is a good choice where the output is real number and there is numerous possibilities. It reduces $\|\omega\|^2$ to minimize complexity where ξ_i and ξ_i^* are slack variables $i = 1,....,n$ to calculate the difference of training data outside sensitive zone [36] [37]

Minimize:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.1}$$

Subject to:

$$\begin{cases} y_i - f(x_{i,\omega}) \leq \varepsilon + \xi_i^* \\ f(x_{i,\omega}) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 1, ..., n \end{cases} \tag{2.2}$$

Kernel Function:

$$K(X_i, X_j) = \begin{cases} (X_i * X_j) & Linear \\ (\gamma X_i * X_j + C)^d & Polynomial \\ exp\left(-\gamma \left\|X_i - X_j\right\|^2\right) & RBF \\ tanh(\gamma X_i * X_j) + C & Sigmoid \\ exp(-g(x_i - y_i)) & Anova \end{cases} \tag{2.3}$$

Here, $K(X_i, X_j) = \phi(X_i, X_j)$, and γ is the adjustable parameter.

2.4 ANN

Neural Network has its starting points in endeavors to discover numerical representations of data processing in biological systems [38]. Without a doubt, it has been utilized extensively to cover an extensive variety of various models, a lot of them have been the subject of misrepresented cases with respect to their biological credibility. From the viewpoint of applications of pattern recognition, however, biological authenticity would force totally superfluous limitations.

2.4.1 Feed Forward Neural Network

The linear models for regression is based on linear combinations of fixed nonlinear basis functions $\phi_j(x)$ and the form generated from this is:

$$y(x, w) = f\left(\sum_{j=1}^{M} \omega_j \phi_j(x)\right) \tag{2.4}$$

Here, $f(.)$ is the activation function for regression problem.
If the input variables are $x_1,...,x_D$, the M linear combination of the input is:

$$a_j = \sum_{i=1}^{D} \omega_{ji}^{(1)} x_i + \omega_{j0}^{(1)} \tag{2.5}$$

Here $j = 1,...,M$, and the superscript (1) signifies that the respective parameters are from the first layer of the network. $\omega_{ji}^{(1)}$ are the weights and $\omega_{j0}^{(1)}$ are the biases of the network. The quantity a_j are know as activation. Then they get transformed and give:

$$z_j = h(a_j) \tag{2.6}$$

This quantities are the output of basis function and called hidden units. When we linearly combine them for output unit activations, it takes the form:

$$a_k = \sum_{j=1}^{M} \omega_{kj}^{(2)} z_j + \omega_{k0}^{(2)} \tag{2.7}$$

Here k = 1,...,K, and K is the total number of outputs and the superscript (2) signifies that the respective parameters are from the first layer of the network and $\omega_{k0}^{(2)}$ are biases. At this time the output unit activations are transformed using an appropriate activation function to give a set of network outputs y_k. In regression problem, the activation function is the identity, so that $y_k = a_k$. The combination of different steps gives the whole network function for sigmoidal output unit activation functions,

$$yk = (x,w) = \sigma \left(\sum_{j=1}^{M} \omega_{kj}^{(2)} h \left(\sum_{i=1}^{D} \omega_{ji}^{(1)} x_i + \omega_{j0}^{(1)} \right) + \omega_{k0}^{(2)} \right) \tag{2.8}$$

Chapter 3

METHODOLOGY

3.1 Windowing

Windowing operator is an exclusive operator which can perform better for time series prediction. It converts series sample data into single valued data. The series data must be given as Example Set. The parameter "series representation" defines how the series data is represented by the Example Set. It Convert the last row of a window within the time series into a label or target variable [39]. Fed the cross sectional values as inputs to the machine learning technique such as liner regression, Neural Network, Support vector machine and so on. Figure 3.1 shows the mechanism of windowing operator [40].

3.2 Sliding Window Validation

This is a exceptional validation chain which can only be utilized for series predictions where the time points are encoded as examples. It utilizes a certain window of examples for training and uses another window (after horizon examples, i.e. time points) for testing. The window is moved across the example set and all performance measurements are averaged afterwards. The parameter "cumulative_training" indicates if all former examples should be used for training (instead of only the current window). This validation operator provides several values which can be logged by method of a ProcessLogOperator. All performance estimation operators of RapidMiner provide access to the average values calculated during the estimation. Since the operator cannot guarantee the names of the delivered criteria, the ProcessLog operator can access the values via the generic value names:

- performance: the value for the main criterion calculated by this validation operator

- performance1: the value of the first criterion of the
- performance vector calculated
- performance2: the value of the second criterion of the performance vector calculated
- performance3: the value of the third criterion of the performance vector calculated for the main criterion, also the variance and the standard deviation can be accessed where applicable.

3.3 Proposed Support Vector Regression Model

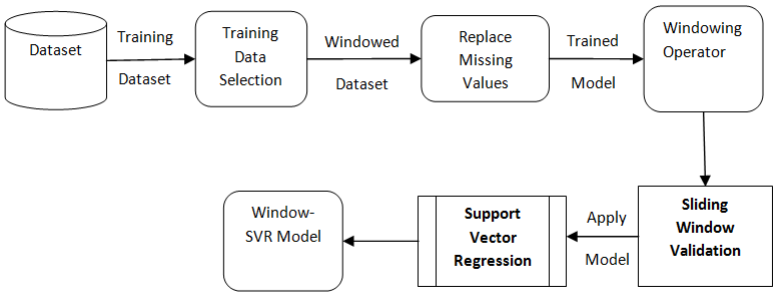


Fig. 3.1 Workflow of Win-SVR Model

Training Phase:

- Step 1: Perused training dataset from local repository and select ID, Level in specific attribute.
- Step 2: Use replace missing value operator for missing value handling with maximum.
- Step 3: Apply window operator which changes a given example set containing series data into another example set containing single esteemed cases. This stride will change over the last row of a window inside the time series into a name or target variable. Last variable is dealt with as level.
- Step 4: Accomplish a sliding window validation which is used for time series prediction with training, window width, training window step size in different horizon then fed to the SVR algorithm from delivered windowed example set.

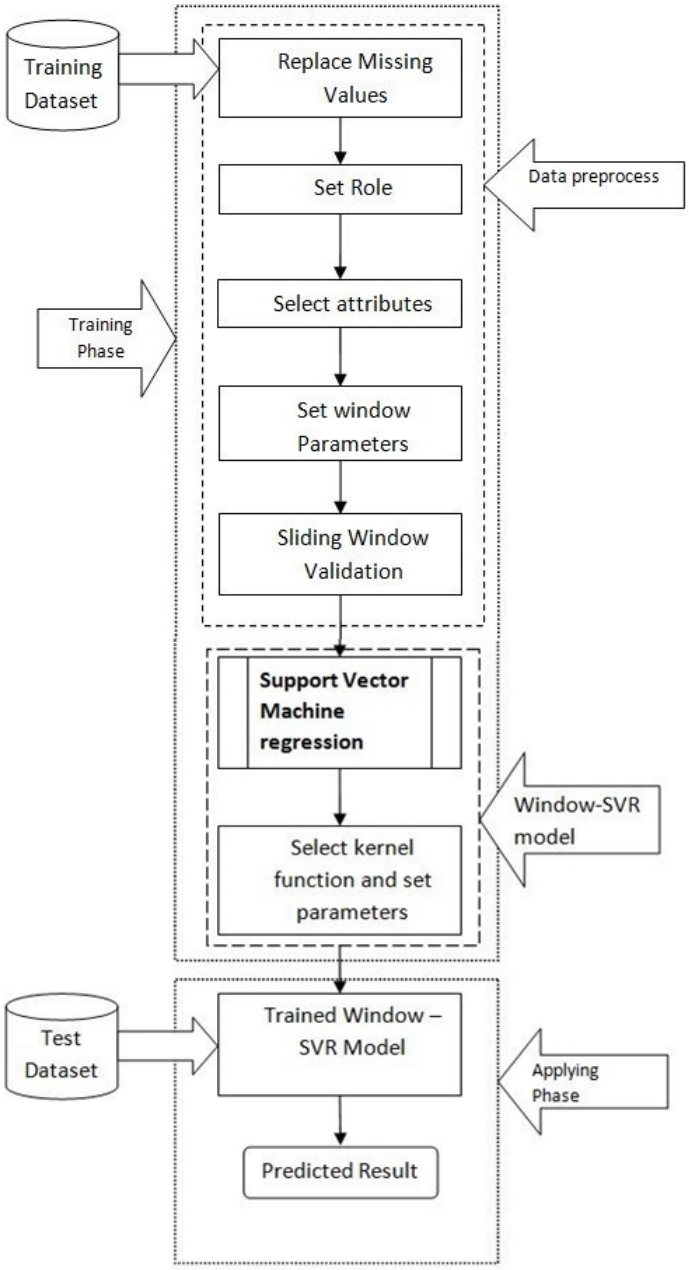


Fig. 3.2 Flowchart of Win-SVR Model

- Step 5: Select SVR special parameters kernel type such as dot, redial, polynomial, neural, anova, epachnenikov, gaussian_combination, miltiquadric with kernel gamma, kernel degree, C, L(+,-),Epsilon(+,-) etc.
- Step 6: Run the model and observe performance with accuracy.
- Step 7: If the performance accuracy is good then save the results otherwise change the validation process again and then fit to the algorithm with SVR parameters in different values.
- Step 8: Exit from the training stage, compare with test dataset for best results then apply prepared model to the testing dataset.

  Applying Phase:

- Step 1: Perused testing dataset from local repository.
- Step 2: Apply the training model dataset results into the test dataset for forecasting performance with RMSE (Root Mean Square Error).
- Step 3: Built the predicted model results and forecast performance.

3.4 Proposed ANN Model

3.4.1 Leaky ReLUs

Leaky Rectified Linear activation is first introduced in acoustic mode [41]. Neural networks commonly utilize a sigmoidal nonlinearity function. Recently, however, there is increasing evidence that other types of nonlinearites can improve the performance of DNNs. The mathematical function is:

$$y_i = \begin{cases} x_i & if\ x_i \geq 0 \\ \frac{x_i}{a_i} & if\ x_i < 0 \end{cases} \tag{3.1}$$

where a_i is a fixed parameter in range $(1, +\infty)$. In original paper, the authors suggest to set ai to a large number like 100.[41]
There are some advantages of using ReLu in ANN: One major benefit is the reduced likelihood of the gradient to vanish. This arises when $a > 0$. In this regime the gradient has a constant value. In contrast, the gradient of sigmoids becomes increasingly small as the absolute value of x increases. The constant gradient of ReLUs results in faster learning.

The other benefit of ReLUs is sparsity. Sparsity arises when $a \leq 0$. The more such units that exist in a layer the more sparse the resulting representation. Sigmoids on the other hand are always likely to generate some non-zero value resulting in dense representations. Sparse representations seem to be more beneficial than dense representations.

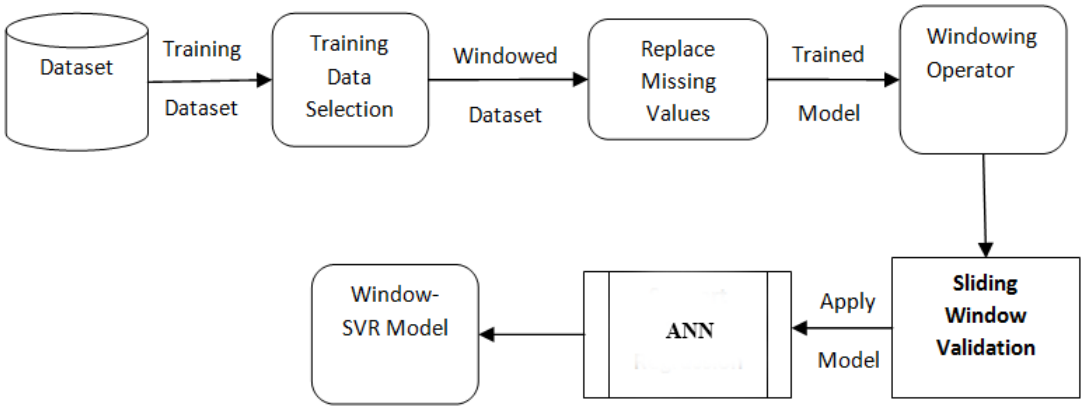


Fig. 3.3 Workflow of Win-ANN Model

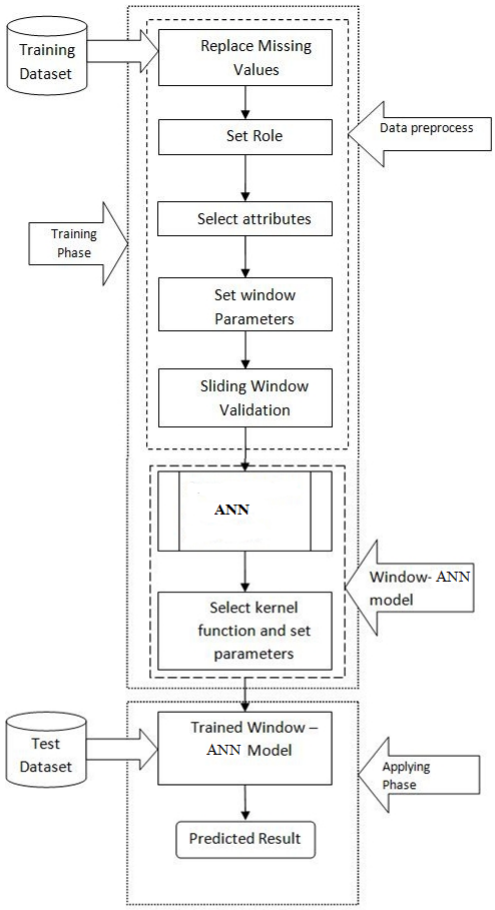


Fig. 3.4 Flowchart of Win-ANN Model

3.5 RMSE

RMSE or RMSD is a commonly used measure of the contrasts between qualities anticipated by a model or an estimator and the qualities really observed. The RMSE shows the example standard deviation of the contrasts between anticipated values and observed values. These individual contrasts are called residuals when the counts are performed over the information test that was utilized for estimation, and are gotten expectation mistakes when processed out-of-sample. The RMSE serves to total the extents of the errors in forecasts for different times into a solitary measure of prescient force. RMSE is a decent measure of precision, however just to analyze estimating blunders of various models for a specific variable and not between variables, as it is scale-dependent [42][43]

$$RMSE = \sqrt{\frac{\sum_n^{i=1} (yt - \hat{y}t^2)}{n}} \tag{3.2}$$

Here, yt is the original value of a point for a given time period t, n is the total number of fitted points, and $\hat{y}t$ is the fitted forecast value for the time period t.

3.6 MAE

The MAE is a typical measure of figure mistake in time series data where the expressions "Mean Absolute Deviation" is once in a while utilized as a part of perplexity with the more standard meaning of mean absolute deviation.

$$MAE = \frac{SAE}{N} = \frac{\sum_n^{i=1} |x_i - \hat{x}_i|}{N} \tag{3.3}$$

Here, x_i is the actual observations time series, $\hat{x}_i$ is the estimated or forecasted time series, SAE is the sum of the absolute errors (or deviations), N is the number of non-missing data points.

Chapter 4

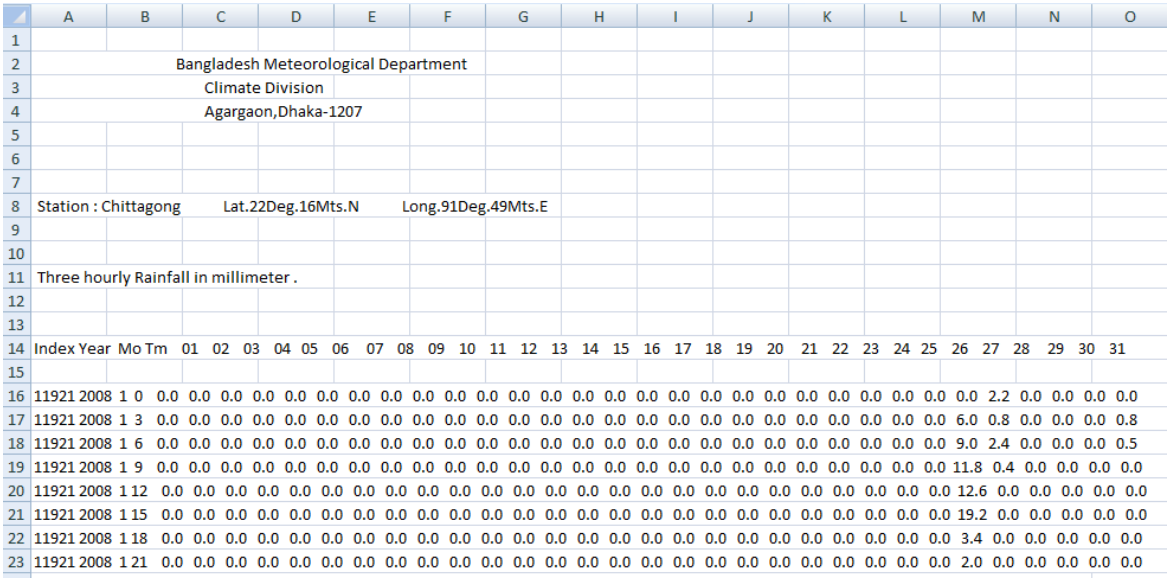
Experiment Design

4.1 Research Data

6 years Meteorological data (2008-2014) of Chittagong, Bangladesh from the Meteorological Department, Bangladesh were collected to perform experiment and result evaluation of this study. The experiment data consists of total 2558 instances. The whole experiment is divided into two major parts. One is for rainfall prediction with only rainfall data and the other is for to predict rainfall and temperature with the combined data of rainfall and temperature. The reason behind splitting the study into two part is to find the best combination of data by which weather can be well predicted. As we have time series data including rainfall and temperature data; we decided to find which data can predict rainfall and temperature, separately only rainfall and temperature data or the combination of both type of data. Six attributes, Date, total, avg, max, min, MA included both the separated (Rainfall and Temperature) data and then split for training and testing. The 'Date' attribute was selected as id for all the models and another attribute 'total' for total rainfall prediction was chosen as label. Figure 4.1 shows the actual rainfall (2008-2014). 80% of the data were considered as training data and the rest 20% as testing data.

4.1.1 Data Preprocessing

The data collected from Meteorological Department of Bangladesh was not ready to use instantly, because it was not arranged to fit in a Machine Learning algorithm. The dataset was firstly arranged in row and columns in spreadsheet, then the unnecessary data were removed. As the dataset was not prepared for Machine Learning; the whole pre-processing was done manually. Figure 4.1 and Figure 4.2 shows some of the raw data collected from the Meteorological Department of Bangladesh. Table 4.1 Displays The Rainfall And



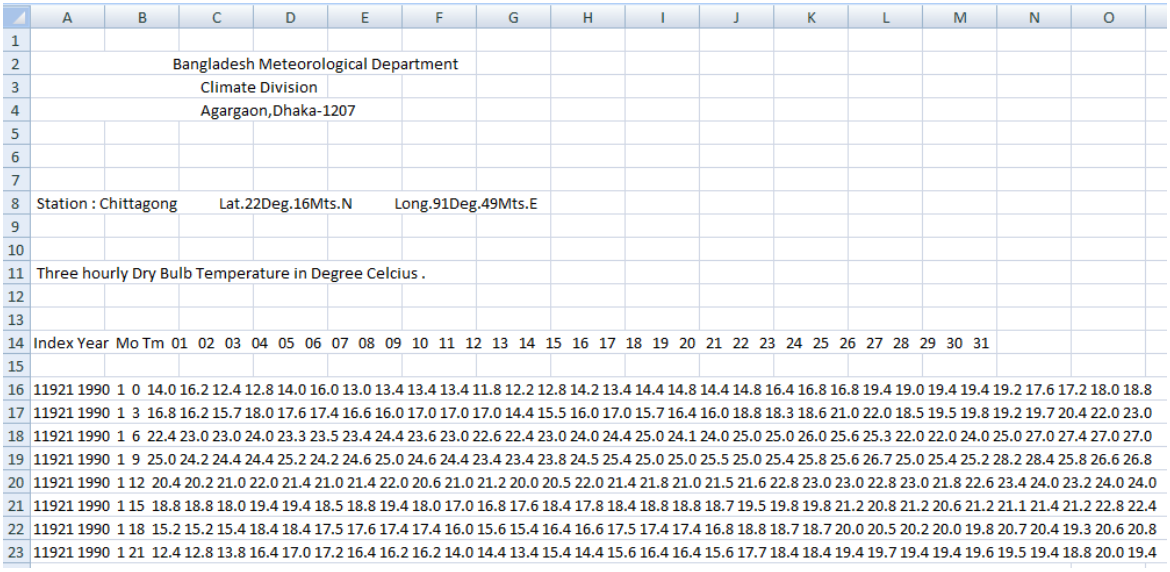
Bangladesh Meteorological Department
Climate Division
Agargaon,Dhaka-1207

Station : Chittagong Lat.22Deg.16Mts.N Long.91Deg.49Mts.E

Three hourly Rainfall in millimeter .

Index	Year	Mo	Tm	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
11921	2008	1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	0.0	0.0	0.0	0.0
11921	2008	1	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.8	0.0	0.0	0.0	0.8
11921	2008	1	6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	2.4	0.0	0.0	0.0	0.5
11921	2008	1	9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.8	0.4	0.0	0.0	0.0	0.0
11921	2008	1	12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.6	0.0	0.0	0.0	0.0	0.0
11921	2008	1	15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.2	0.0	0.0	0.0	0.0	0.0
11921	2008	1	18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.4	0.0	0.0	0.0	0.0	0.0
11921	2008	1	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0

Fig. 4.1 Rainfall Raw Data From Meteorological Department, Bangladesh



Bangladesh Meteorological Department
Climate Division
Agargaon,Dhaka-1207

Station : Chittagong Lat.22Deg.16Mts.N Long.91Deg.49Mts.E

Three hourly Dry Bulb Temperature in Degree Celcius .

Index	Year	Mo	Tm	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
11921	1990	1	0	14.0	16.2	12.4	12.8	14.0	16.0	13.0	13.4	13.4	13.4	11.8	12.2	12.8	14.2	13.4	14.4	14.8	14.4	14.8	16.4	16.8	16.8	19.4	19.0	19.4	19.4	19.2	17.6	17.2	18.0	18.8
11921	1990	1	3	16.8	16.2	15.7	18.0	17.6	17.4	16.6	16.0	17.0	17.0	17.0	14.4	15.5	16.0	17.0	15.7	16.4	16.0	18.8	18.3	18.6	21.0	22.0	18.5	19.5	19.8	19.2	19.7	20.4	22.0	23.0
11921	1990	1	6	22.4	23.0	23.0	24.0	23.3	23.5	23.4	24.4	23.6	23.0	22.6	22.4	23.0	24.0	24.4	25.0	24.1	24.0	25.0	25.0	26.0	25.6	25.3	22.0	22.0	24.0	25.0	27.0	27.4	27.0	27.0
11921	1990	1	9	25.0	24.2	24.4	24.4	25.2	24.2	24.6	25.0	24.6	24.4	23.4	23.4	23.8	24.5	25.4	25.0	25.0	25.5	25.0	25.4	25.8	25.6	26.7	25.0	25.4	25.2	28.2	28.4	25.8	26.6	26.8
11921	1990	1	12	20.4	20.2	21.0	22.0	21.4	21.0	21.4	22.0	20.6	21.0	21.2	20.0	20.5	22.0	21.4	21.8	21.0	21.5	21.6	22.8	23.0	23.0	22.8	23.0	21.8	22.6	23.4	24.0	23.2	24.0	24.0
11921	1990	1	15	18.8	18.8	18.0	19.4	19.4	18.5	18.8	19.4	18.0	17.0	16.8	17.6	18.4	17.8	18.4	18.8	18.8	18.7	19.5	19.8	19.8	21.2	20.8	21.2	20.6	21.2	21.1	21.4	21.2	22.8	22.4
11921	1990	1	18	15.2	15.2	15.4	18.4	18.4	17.5	17.6	17.4	17.4	16.0	15.6	15.4	16.4	16.6	17.5	17.4	17.4	16.8	18.8	18.7	18.7	20.0	20.5	20.2	20.0	19.8	20.7	20.4	19.3	20.6	20.8
11921	1990	1	21	12.4	12.8	13.8	16.4	17.0	17.2	16.4	16.2	16.2	14.0	14.4	13.4	15.4	14.4	15.6	16.4	16.4	15.6	17.7	18.4	18.4	19.4	19.7	19.4	19.4	19.6	19.5	19.4	18.8	20.0	19.4

Fig. 4.2 Temperature Raw Data From Meteorological Department, Bangladesh

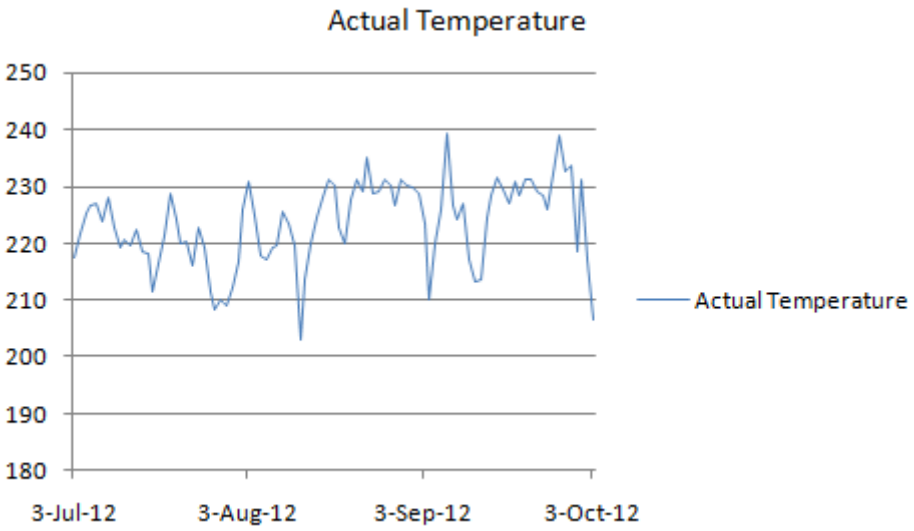


Fig. 4.3 Actual Temperature, 2008-2014

Temperature Attributes After Data Pre-Processing. Max_R, Total_R, Avg_R, Maxma_R, Totalma_R, Avgma_R Are Directly Used In Experiment. Figure 4.3 shows the actual data used in this study.

Moving Average:

A Moving Average (rolling average or running average) is an estimation to analyze data points by making a progression of midpoints of distinctive subsets of the full data set.

$$MA = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{4.1}$$

Here x_1, x_2, x_n are the average values of the monthly rainfall and n is the total number of observation. Moving average of 7 days was considered for all attributes taken as labels in this study.

Table 4.1 Pre-processed Data

Rainfall Processed Data							
Date	Max_R	Min_R	Total_R	Avg_R	MaxMA_R	TotalMA_R	AvgMA_R
1-Jul-12	0	0	0	0	38	94	12
2-Jul-12	0	0	0	0	30	72	9
3-Jul-12	26	0	37.2	4.65	28	67	8
4-Jul-12	0.1	0	0.1	0.0125	4	6	1
5-Jul-12	2	0	2.5	0.3125	4	6	1
6-Jul-12	0	0	0	0	4	6	1
7-Jul-12	1	0	1.6	0.2	4	6	1
8-Jul-12	4	0	6.7	0.8375	4	6	1
Temperature Processed Data							
Date	Max_R	Min_R	Total_R	Avg_R	MaxMA_R	TotalMA_R	AvgMA_R
1-Jul-12	32	26.7	233.2	29.15	29	217	27
2-Jul-12	30	27	226.9	28.3625	30	221	28
3-Jul-12	29.8	25.4	217.4	27.175	30	223	28
4-Jul-12	28.8	25.8	221.7	27.7125	31	226	28
5-Jul-12	30	26.6	225	28.125	31	227	28
6-Jul-12	30.8	26.6	226.6	28.325	30	227	28
7-Jul-12	30.8	26	227	28.375	30	226	28
8-Jul-12	31.2	26	223.6	27.95	30	225	28

Table 4.1 Shows the pre-processed data used in this study. For the 'Rainfall Processed Data' from the second to the last column represents the value of Maximum Rainfall, Minimum Rainfall, Total Rainfall, Average Rainfall, Moving Average for Maximum Rainfall, Moving Average for Total Rainfall, and Moving Average for Average Rainfall respectively.

4.2 SVR Analysis

Table 4.2 Support Vector Kernel Analysis

Model	Horizon	Sliding Window Validation		Kernel Type	Degree	C	G	ε	ε+	ε-
		Window Size	Step Size							
Total Rainfall Prediction	1 day	5	1	Anova	1	100	1	1	1	1
	7 day	5	1		1	120	1	1	1	1
	10 day	5	1		1	200	1	1	1	1
Total Temperature Prediction	1 day	5	1		1	100	1	1	1	1
	7 day	5	1		1	150	1	1	1	1
	10 day	5	1		1	300	3	1	1	1

Table 4.2 Displays The Window Size, Step Size, Kernel Type, Kernel Degree, C, Kernel Gamma, Epsilon Used In Support Vector Kernel Analysis For Total Rainfall And Temperature Combine Dataset.

Table 4.3 SVR Kernel Analysis (Single Dataset)

Model	Horizon	Sliding Window Validation		Kernel Type	Degree	C	G	ε	ε+	ε-
		Window Size	Step Size							
Total Rainfall Prediction	1 day	5	1	Anova	1	300	1	1	1	1
	7 day	5	1		1	400	1	1	1	1
	10 day	5	1		1	700	1	1	1	1
Total Temperature Prediction	1 day	5	1		1	300	1	1	1	1
	7 day	5	1		1	500	1	1	1	1
	10 day	5	1		1	400	1	1	1	1

Table 4.3 Displays The Window Size, Step Size, Kernel Type, Kernel Degree, C, Kernel Gamma, Epsilon In Support Vector Kernel Analysis For Total Rainfall And Temperature Single Dataset.

Table 4.4 Sliding Window Validation For SVR

Model	Horizon	Training Window Width	Training Window Step	Test Window Width	Cumulative Training
Total Rainfall	1	5	1	5	No
	7	5	1	5	No
	10	5	1	5	No
Total Temperature	1	5	1	2	No
	7	5	1	5	No
	10	5	1	5	No

Table 4.4 Displays The Training Window Width, Testing Window Width, Training Window Step Size In Sliding Window Validation For Total Train And Test Dataset

Table 4.5 SVR Model For 1 Day Forecasting

SV 2040				
Bias(b) 80.756				
Weight(w)				
w[Max_R-4]	w[Max_R-3]	w[Max_R-2]	w[Max_R-1]	w[Max_R-0]
12439.72	11324.576	18340.882	18462.71	14945.589
w[Avg_R-4]	w[Avg_R-3]	w[Avg_R-2]	w[Avg_R-1]	w[Avg_R-0]
9529.243	9255.4	14570.532	16793.736	13556.434
w[maxMA-4]	w[maxMA-3]	w[maxMA-2]	w[maxMA-1]	w[maxMA-0]
21105.09	18777.284	17703.346	15852.759	15360.3
w[totalMA_R-4]	w[totalMA_R-3]	w[totalMA_R-2]	w[totalMA_R-1]	w[totalMA_R-0]
19052.581	16318.756	15276.371	13317.361	12661.916
w[avgMA_R-4]	w[avgMA_R-3]	w[avgMA_R-2]	w[avgMA_R-1]	w[avgMA_R-0]
22975.886	21291.032	19628.356	19368.837	19238.515
w[MAX_T-4]	w[MAX_T-3]	w[MAX_T-2]	w[MAX_T-1]	w[MAX_T-0]
24882.456	23734.844	25185.788	24599.531	19732.255
w[TOTAL_T-4]	w[TOTAL_T-3]	w[TOTAL_T-2]	w[TOTAL_T-1]	w[TOTAL_T-0]
31469.161	31324.143	30272.234	30246.077	28700.487
w[AVG_T-4]	w[AVG_T-3]	w[AVG_T-2]	w[AVG_T-1]	w[AVG_T-0]
31469.161	31324.143	30272.234	30246.077	28700.487
w[maxMA_T-4]	w[maxMA_T-3]	w[maxMA_T-2]	w[maxMA_T-1]	w[maxMA_T-0]
21227.876	21726.301	22661.747	23799.884	24066.091
w[totalMA_T-4]	w[totalMA_T-3]	w[totalMA_T-2]	w[totalMA_T-1]	w[totalMA_T-0]
28316.325	28722.436	29397.953	29892.629	30229.62
w[avgMA_T-4]	w[avgMA_T-3]	w[avgMA_T-2]	w[avgMA_T-1]	w[avgMA_T-0]
27734.255	28227.199	29394.881	29884.372	29903.639

Table 4.5 Displays The Support Vector, Bias, Weight In Support Vector Model From Support Vector Machine For One Day Forecasting

4.3 ANN Analysis

Table 4.6 Neural Net Parameter Analysis For Combined Dataset

Model	Horizon	Training Cycles	Learning Rate	M	Hidden Layer	Shuffle	Normalize	Error Epsilon
Total Rainfall	1 day	120	0.3	0.2	2	Yes	Yes	1.00E-05
	7 day	120	0.3	0.2	3	Yes	Yes	1.00E-05
	10 day	110	0.3	0.2	1	Yes	Yes	1.00E-05
Total Temp.	1 day	120	0.3	0.2	2	Yes	Yes	1.00E-05
	7 day	100	0.3	0.2	3	Yes	Yes	1.00E-05
	10 day	110	0.3	0.2	2	Yes	Yes	1.00E-05

Table 4.6 Displays The Training Cycles, Learning Rate, Momentum, Hidden Layer, Shuffle, Normalize, Error Epsilon In Neural Net Parameter Analysis For Total Rainfall And Temperature Combine Dataset

Table 4.7 Neural Net Parameter Analysis For Single Dataset

Model	Horizon	Training Cycles	Learning Rate	M	Hidden Layer	Shuffle	Normalize	Error Epsilon
Total Rainfall	1 day	120	0.3	0.2	3	Yes	Yes	1.00E-05
	7 day	120	0.3	0.2	2	Yes	Yes	1.00E-05
	10 day	90	0.3	0.2	3	Yes	Yes	1.00E-05
Total Temp.	1 day	120	0.3	0.2	2	Yes	Yes	1.00E-05
	7 day	110	0.3	0.2	1	Yes	Yes	1.00E-05
	10 day	120	0.3	0.2	1	Yes	Yes	1.00E-05

Table 4.7 Displays The Training Cycles, Learning Rate, Momentum, Hidden Layer, Shuffle, Normalize, Error Epsilon In Neural Net Parameter Analysis For Total Rainfall And Temperature Single Dataset

Table 4.8 Sliding Window Validation For ANN

Model	Horizon	Training Window Width	Training Window Step	Test Window Width	Cumulative Training
Total Rainfall	1	2	1	2	No
	7	2	1	2	No
	10	2	1	2	No
Total Temperature	1	2	1	2	No
	7	2	1	2	No
	10	2	1	2	No

Table 4.8 Displays The Sliding Window Validation With Training Window Width, Testing Window Width, Training Window Step Size In Total Training And Test Dataset

Table 4.9 ANN Model For 1 Day Forecasting

Threshold -1.001							
Bias(b) with Sigmoid				Regression(Linear) in Output			
Node1	Node2	Node3	Node4	Node1	Node2	Node3	Node4
-0.201	-0.188	-0.266	-0.193	0.063	0.006	0.082	-0.138
Node5	Node6	Node7	Node8	Node5	Node6	Node7	Node8
-0.161	-0.142	-0.154	-0.535	-0.185	-0.061	-0.1	0.461
Node9	Node10	Node11	Node12	Node9	Node10	Node11	Node12
-0.182	-0.177	-0.221	-0.143	-0.328	-0.138	0.545	-0.187
Node13	Node14	Node15	Node16	Node13	Node14	Node15	Node16
-0.231	-0.14	-0.209	-0.143	0.046	-0.297	0.095	-0.034
Node17	Node18	Node19	Node20	Node17	Node18	Node19	Node20
-0.189	-0.164	-0.152	-0.187	-0.093	-0.055	-0.028	0.052
Node21	Node22	Node23	Node24	Node21	Node22	Node23	Node24
-0.156	-0.196	-0.368	-0.148	0.152	0.339	0.658	-0.364
Node25	Node26	Node27	Node28	Node25	Node26	Node27	Node28
-0.188	-0.172	-0.14	0.051	-0.03	0.116	0.205	-1.058
Node29				Node29			
-0.395				0.769			

Table 4.9 Displays The Bias With Sigmoid, Regression (Linear) In Output In Neural Net Model For One Day Forecasting

Chapter 5

Result

5.1 Discussion

It is clear from the study, that SVR can produce better prediction result for rainfall prediction with both single and combined dataset. SVR produce the best result for 1 day ahead model with only 0.95 MAE for the single dataset and for combined dataset the best result is to 7 days ahead model , with 0.17 MAE. For Temperature prediction from both single and combined dataset, ANN showed better performance than SVR. ANN produce the best prediction for 5 days ahead model with the single dataset with only 0.72 MAE and for combined dataset the best result is also for 5 days ahead model. NN produce a result with only 1.72 MAE.
Table 5.1 to Table 5.9 displays different results from the study. Table 5.1 displays the Average Merit and Average Rank for two different (Rainfall and Temperature) label which were generated by I_{GAIN} method.

Table 5.1 Average Merit And Rank

Features	Total Rainfall		Total Temperature	
	Average Merit	Average Rank	Average Merit	Average Rank
Avg	3.575 ± 0.007	1 ± 0	2.626 ± 0.033	6 ± 0
Max	2.967 ± 0.015	2 ± 0	2.361 ± 0.035	8.4 ± 0.49
Total Temp	2.624 ± 0.006	3 ± 0	2.626 ± 0.033	5 ± 0
Avg Temp	2.624 ± 0.006	4 ± 0	9.388 ± 0.006	1 ± 0
Total MA	1.589 ± 0.022	5 ± 0	2.583 ± 0.014	7 ± 0
Max Temp	1.509 ± 0.008	6.4 ± 0.49	4.766 ± 0.007	3 ± 0
Min Temp	1.5 ± 0.011	6.6 ± 0.49	5 ± 0.001	2 ± 0
Total MA Temp	1.413 ± 0.006	8 ± 0	4.732 ± 0.007	4 ± 0
Max MA	1.369 ± 0.018	9 ± 0	2.239 ± 0.013	10 ± 0
Avg MA	0.754 ± 0.015	10 ± 0	1.087 ± 0.011	12 ± 0
Avg MA Temp	0.621 ± 0.009	11 ± 0	2.343 ± 0.01	8.6 ± 0.49
Max MA Temp	0.574 ± 0.006	12 ± 0	2.066 ± 0.008	11 ± 0
Min	0.111 ± 0.014	13 ± 0	0.09 ± 0.004	13 ± 0

Table 5.2 Rainfall and Temperature Prediction Result Using SVM

Total Rainfall & Temperature Test Data (Horizon 1)							
Date	Rainfall Data			Date	Temperature Data		
	Actual	Predicted	RMSE		Actual	Predicted	RMSE
29-Aug-13	24.2	23.22	0.98	26-Sep-13	225.5	226.7	1.2
30-Aug-13	0.6	8.17	7.57	27-Sep-13	221.8	224.23	2.43
31-Aug-13	1.8	2.55	0.75	28-Sep-13	219.4	219.94	0.54
1-Sep-13	0	2.16	2.16	29-Sep-13	222.7	224.27	1.57
2-Sep-13	5.2	7.97	2.77	30-Sep-13	219	219.16	**0.16**
3-Sep-13	3.6	3.92	**0.32**	1-Oct-13	223	219.41	3.59
4-Sep-13	29.2	28.07	1.13	2-Oct-13	223.9	225.35	1.45
5-Sep-13	8.8	10	1.2	3-Oct-13	223	223.87	0.87
6-Sep-13	0	3.39	3.39	4-Oct-13	202.6	200.18	2.42
7-Sep-13	7.2	5.05	2.15	5-Oct-13	205.2	206.63	1.43

Table 5.2 Displays The Actual Data, Predicted Data, Rmse For Total Rainfall And Temperature Experiment Dataset In Horizon 1 Using Support Vector Machine.

Table 5.3 Rainfall and Temperature Prediction Result Using SVM

Total Rainfall & Temperature Test Data (Horizon 7)							
Date	Rainfall Data			Date	Temperature Data		
	Actual	Predicted	RMSE		Actual	Predicted	RMSE
23-Sep-13	1.5	5.39	3.89	15-Nov-13	205.9	207.6	**1.7**
24-Sep-13	0	5.02	5.02	16-Nov-13	200.9	205.68	4.78
25-Sep-13	3.2	2.11	1.09	17-Nov-13	200.9	203.64	2.74
26-Sep-13	0	2.97	2.97	18-Nov-13	202.4	198.12	4.28
27-Sep-13	4.2	5	**0.8**	19-Nov-13	192.6	195.45	2.85
28-Sep-13	0	6.72	6.72	20-Nov-13	179.5	189.82	10.32
29-Sep-13	11	2.77	8.23	21-Nov-13	177.5	188.31	10.81
30-Sep-13	11.2	0.16	11.04	22-Nov-13	195.3	202.18	6.88
1-Oct-13	0	1.41	1.41	23-Nov-13	187	196.25	9.25
2-Oct-13	5.5	1.04	4.46	24-Nov-13	173.1	181.26	8.16

Table 5.3 Displays The Actual Data, Predicted Data, Rmse For Total Rainfall And Temperature Experiment Dataset In Horizon 7 Using Support Vector Machine

Table 5.4 Rainfall and Temperature Prediction Result Using SVM

Total Rainfall & Temperature Test Data (Horizon 10)							
Date	Rainfall Data			Date	Temperature Data		
	Actual	Predicted	RMSE		Actual	Predicted	RMSE
23-Jul-14	25.4	2.02	23.38	29-Oct-14	205.1	206.55	1.45
24-Jul-14	3.6	10.53	6.93	30-Oct-14	210.6	211.5	**0.9**
25-Jul-14	16.4	11.49	4.91	31-Oct-14	214	212.66	1.34
26-Jul-14	1.4	7.67	6.27	1-Nov-14	214	222.93	8.93
27-Jul-14	7.6	16.6	9	2-Nov-14	217.5	225.04	7.54
28-Jul-14	0	11.12	11.12	3-Nov-14	217.6	219.73	2.13
29-Jul-14	0	5.38	5.38	4-Nov-14	220	226.36	6.36
30-Jul-14	0	6.34	6.34	5-Nov-14	226.8	223.84	2.96
23-Jul-14	0	0.52	**0.52**	6-Nov-14	227.9	222.34	5.56

Table 5.4 Displays The Actual Data, Predicted Data, Rmse For Total Rainfall And Temperature Experiment Dataset In Horizon 10 Using Support Vector Machine

Table 5.5 Rainfall and Temperature Prediction Result Using ANN

Total Rainfall & Temperature Test Data (Horizon 1)							
Date	Rainfall Data			Date	Temperature Data		
	Actual	Predicted	RMSE		Actual	Predicted	RMSE
9-Sep-13	13.4	10.89	10.89	11-Aug-13	221.7	221.71	**0.01**
10-Sep-13	0	0.97	0.97	12-Aug-13	228.2	225.63	2.57
11-Sep-13	16.6	10.53	10.53	13-Aug-13	214.9	217.95	3.05
12-Sep-13	0	0.24	**0.24**	14-Aug-13	228.6	225.5	3.1
13-Sep-13	14.8	7.11	7.11	15-Aug-13	222.8	221.48	1.32
14-Sep-13	0	6.2	6.2	16-Aug-13	221.5	222.35	0.85
15-Sep-13	28	16.61	16.61	17-Aug-13	204.4	208.55	4.15
16-Sep-13	0	0.47	0.47	18-Aug-13	209.9	212.24	2.34
17-Sep-13	19.4	11.46	11.46	19-Aug-13	210.7	211.88	1.18
18-Sep-13	0.6	5.95	5.95	20-Aug-13	206.3	208.55	2.25

Table 5.5 Displays The Actual Data, Predicted Data, RMSE For Total Rainfall And Temperature Experiment Dataset In Horizon 1 Using Neural Net

Table 5.6 Rainfall and Temperature Prediction Result Using ANN

Total Rainfall & Temperature Test Data (Horizon 7)							
Date	Rainfall Data			Date	Temperature Data		
	Actual	Predicted	RMSE		Actual	Predicted	RMSE
24-Jul-14	3.6	4.76	1.16	4-Aug-14	215.1	220.71	5.61
25-Jul-14	16.4	5.41	10.99	5-Aug-14	218.8	221.97	3.17
26-Jul-14	1.4	5.79	4.39	6-Aug-14	222.7	223.92	1.22
27-Jul-14	7.6	6.41	1.19	7-Aug-14	225.5	224.96	0.54
28-Jul-14	0	6.29	6.29	8-Aug-14	229.7	225.25	4.45
29-Jul-14	0	5.94	5.94	9-Aug-14	228.9	225.79	3.11
30-Jul-14	0	5.49	5.49	10-Aug-14	226	225.81	0.19
31-Jul-14	0	4.7	4.7	11-Aug-14	228	225.98	2.02
1-Aug-14	8.2	3.61	4.59	12-Aug-14	225.7	225.38	**0.32**
2-Aug-14	1.4	2.49	**1.09**	13-Aug-14	221.8	224.66	2.86

Table 5.6 Displays The Actual Data, Predicted Data, RMSE For Total Rainfall And Temperature Experiment Dataset In Horizon 7 Using Neural Net

Table 5.7 Rainfall and Temperature Prediction Result Using ANN

Total Rainfall & Temperature Test Data (Horizon 10)							
Date	Rainfall Data			Date	Temperature Data		
	Actual	Predicted	RMSE		Actual	Predicted	RMSE
5-Sep-13	8.8	6.16	2.64	11-Aug-13	221.7	222.64	0.94
6-Sep-13	0	5.61	5.61	12-Aug-13	228.2	223.31	4.89
7-Sep-13	7.2	7.09	**0.11**	13-Aug-13	214.9	219.87	4.97
8-Sep-13	2.4	5.58	3.18	14-Aug-13	228.6	222.46	6.14
9-Sep-13	13.4	4.29	9.11	15-Aug-13	222.8	221.26	1.54
10-Sep-13	0	4.83	4.83	16-Aug-13	221.5	222.03	**0.53**
11-Sep-13	16.6	4.04	12.56	17-Aug-13	204.4	214.6	10.2
12-Sep-13	0	5.75	5.75	18-Aug-13	209.9	216.88	6.98
13-Sep-13	14.8	5.49	9.31	19-Aug-13	210.7	216.53	5.83
14-Sep-13	0	8.57	8.57	20-Aug-13	206.3	213.04	6.74

Table 5.7 Displays The Actual Data, Predicted Data, RMSE For Total Rainfall And Temperature Experiment Dataset In Horizon 10 Using Neural Net

Table 5.8 Total Rainfall Prediction For Single And Combined Dataset

Model	Horizon	Total Rainfall Prediction Aug'2013 to Dec'2014 (Rainfall Data)		Total Rainfall Prediction Aug'2013 to Dec'2014 (Rainfall & Temperature Data)	
		RMSE	MAE	RMSE	MAE
SVR	1	20.33	**0.95**	19.88	1.93
	7	27.68	1.71	27.6	**0.17**
	10	28.57	2.25	30.96	4.51
Neural Net	1	21.41	3.54	18.43	2.42
	7	31.97	10.87	27.53	11.33
	10	27.34	1.02	25.84	3.31

Table 5.8 Displays The Total Rainfall Prediction (August 2013 To December 2014) Results RMSE, MAE In SVR And NN For Single And Combine Dataset

Table 5.9 Total Temperature Prediction For Single And Combined Dataset

Model	Horizon	Total Temperature Prediction Aug'2013 to Dec'2014 (Temperature Data)		Total Temperature Prediction Aug'2013 to Dec'2014 (Rainfall & TemperatureData)	
		RMSE	MAE	RMSE	MAE
SVR	1	4.27	5.3	5.03	6.25
	7	9.82	4.18	11.7	5.71
	10	9.98	2.56	12.32	6.34
Neural Net	1	3.31	2.29	4.14	4.86
	7	7.89	**0.72**	8.4	**1.72**
	10	7.96	5.46	8.03	1.98

This Table Displays The Total Temperature Prediction (August 2013 To December 2014) Results RMSE, MAE In SVR And NN For Single And Combined Dataset

5.2 Graphical Representation

Figure 5.1 shows the correlation between the features that were selected to conduct this study. The scale on the right of the figure denotes the strength of relation. Here, the value of relation *1* means there is a strong relation between the features and -1 means the relation between the features are very weak.
Figure 5.3, 5.4, 5.5, 5.6, 5.7, and 5.8 shows the Actual value Vs. Predicted value of Rainfall and Temperature for monthly average prediction

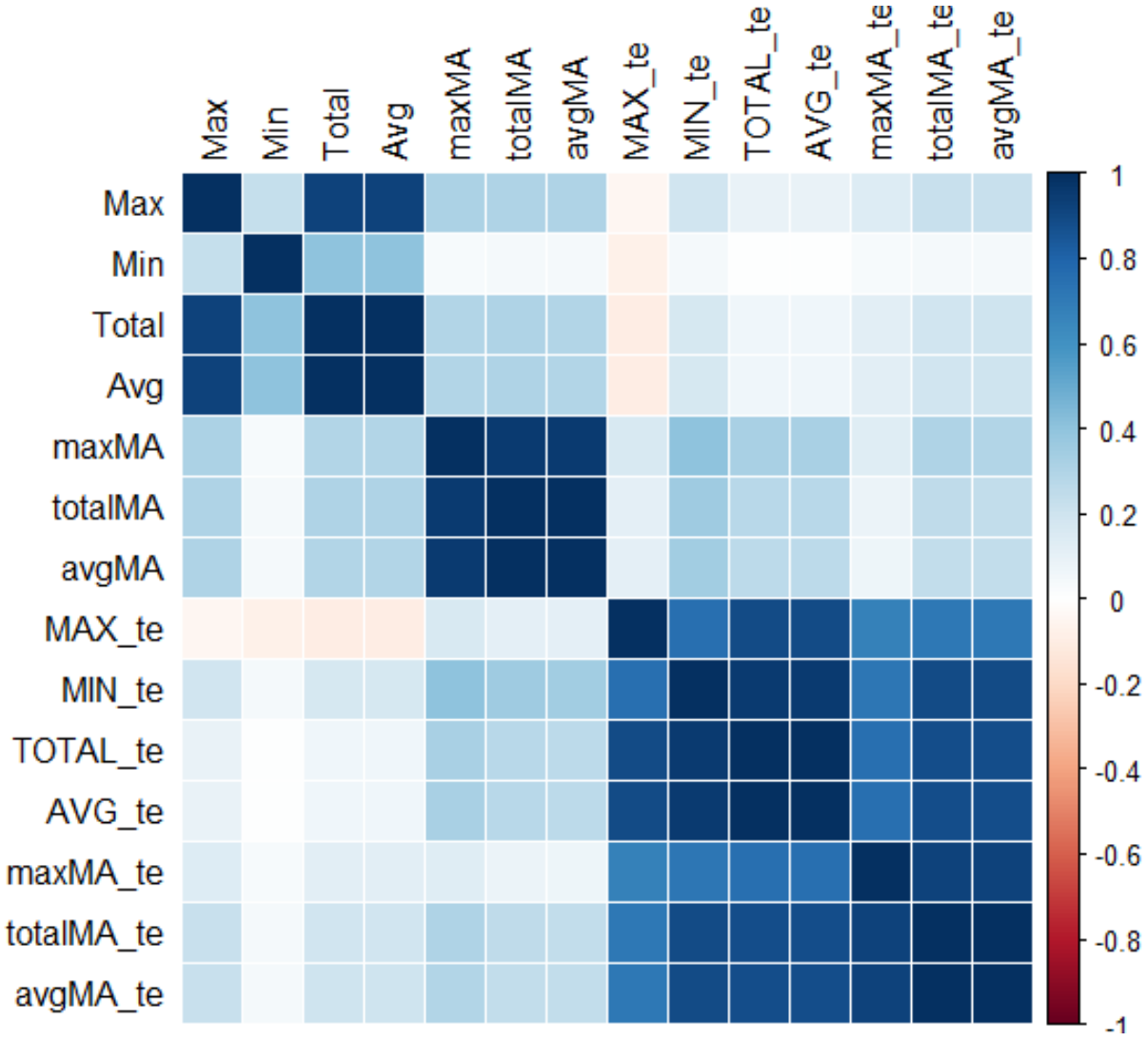


Fig. 5.1 Correlation Between The Attributes

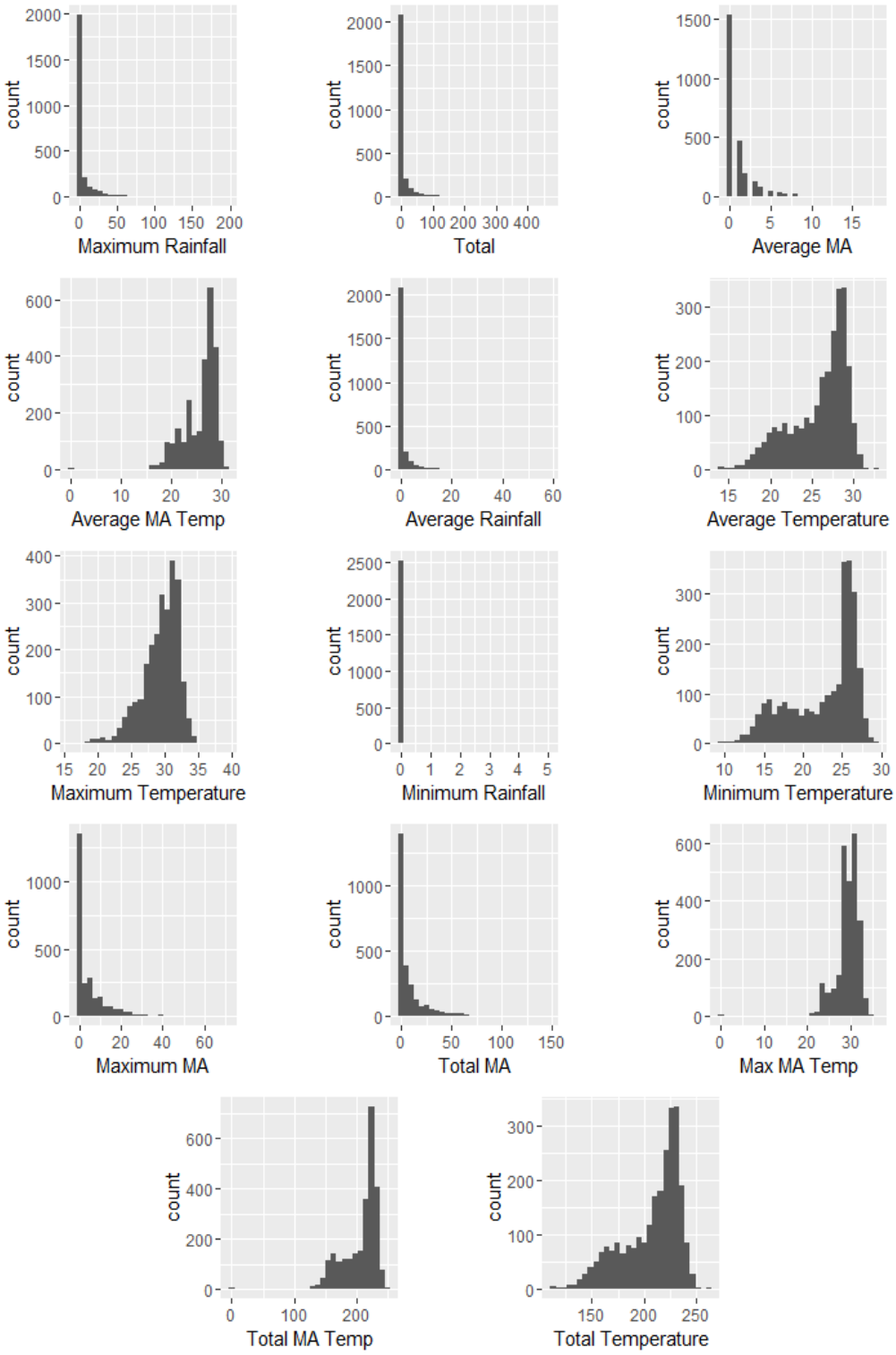


Fig. 5.2 Histograms of the investigated Rainfall and Temperature features on the Meteorological dataset

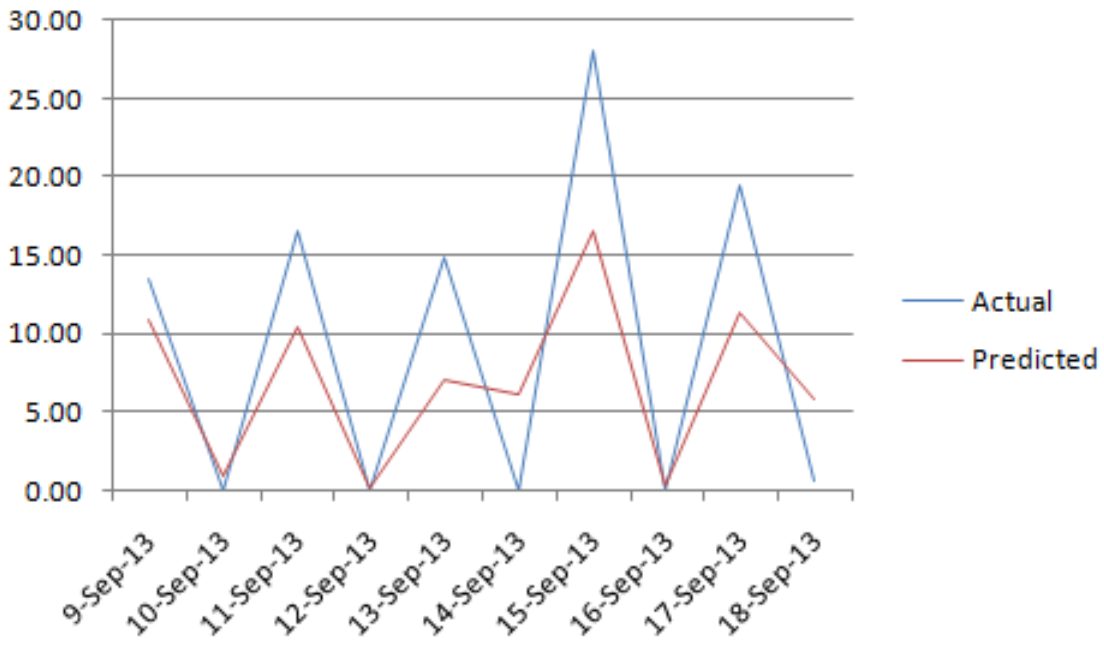


Fig. 5.3 Total Rainfall Prediction Using Combined Dataset In Neural Net, Horizon 1.

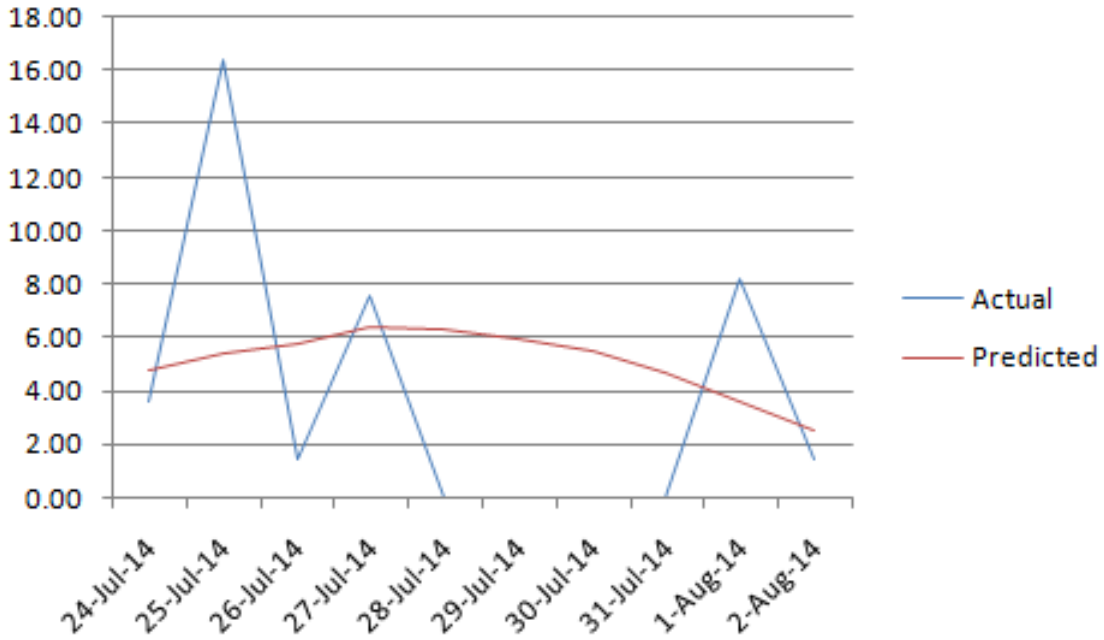


Fig. 5.4 Total Rainfall Prediction Using Combined Dataset In Neural Net, Horizon 7.

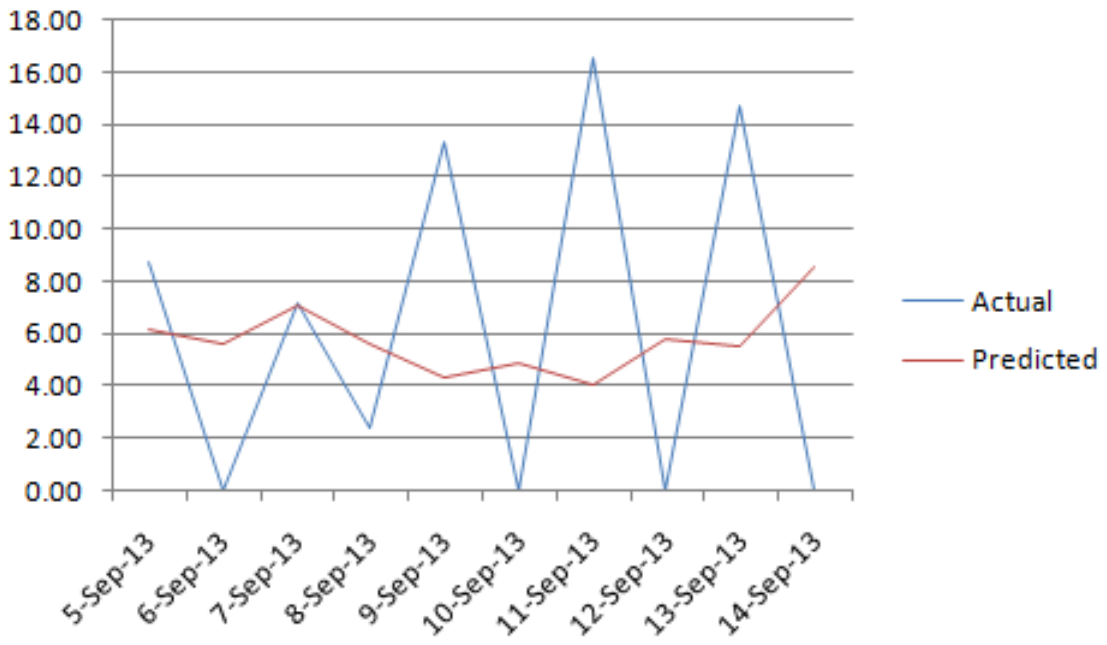


Fig. 5.5 Total Rainfall Prediction Using Combined Dataset In Neural Net, Horizon 10.

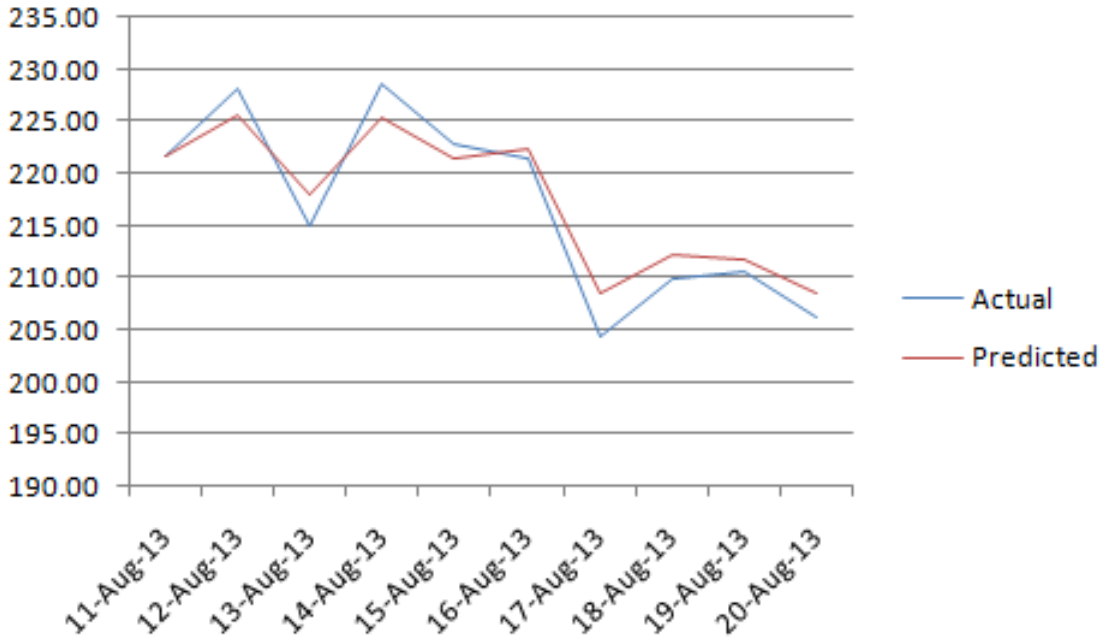


Fig. 5.6 Total Temperature Prediction Using Combined Dataset In Neural Net, Horizon 1.

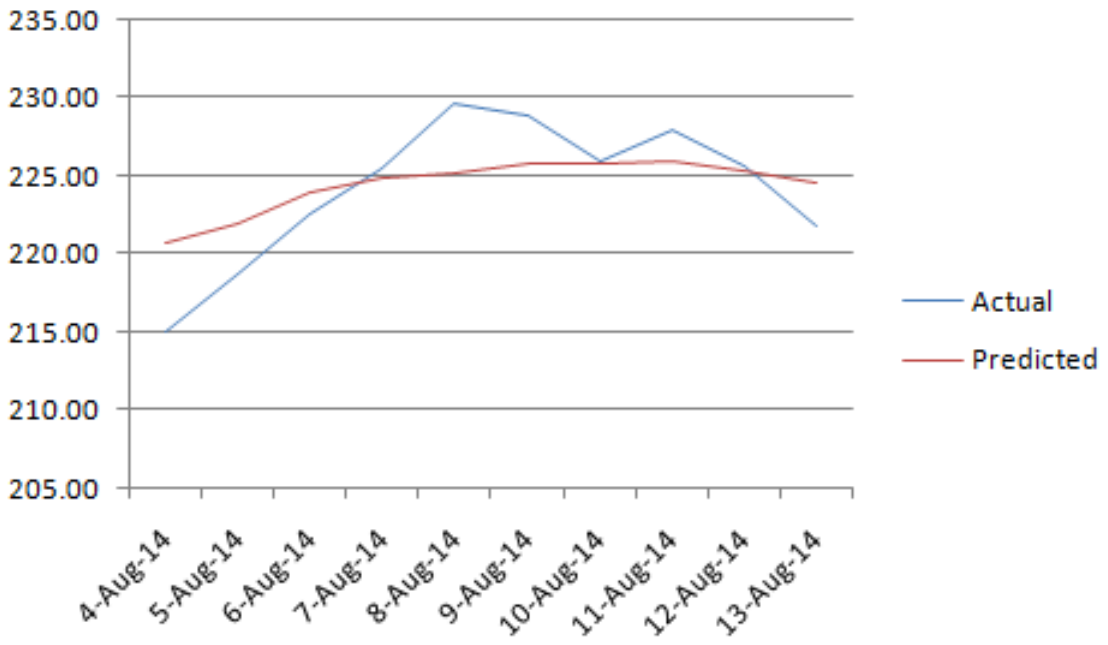


Fig. 5.7 Total Temperature Prediction Using Combined Dataset In Neural Net, Horizon 7.

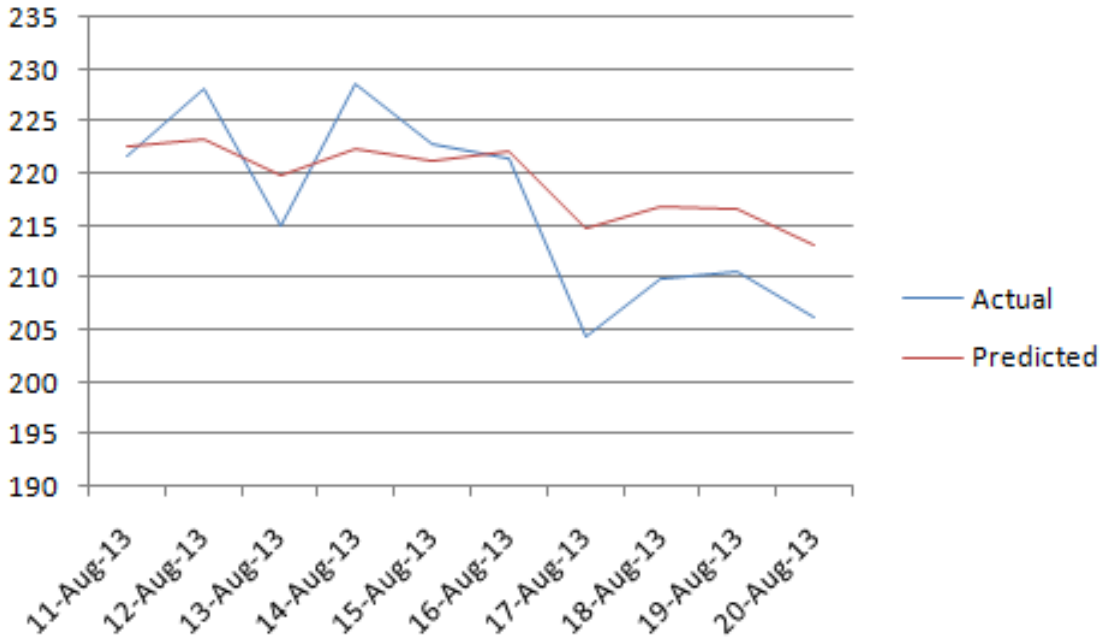


Fig. 5.8 Total Temperature Prediction Using Combined Dataset In Neural Net, Horizon 10.

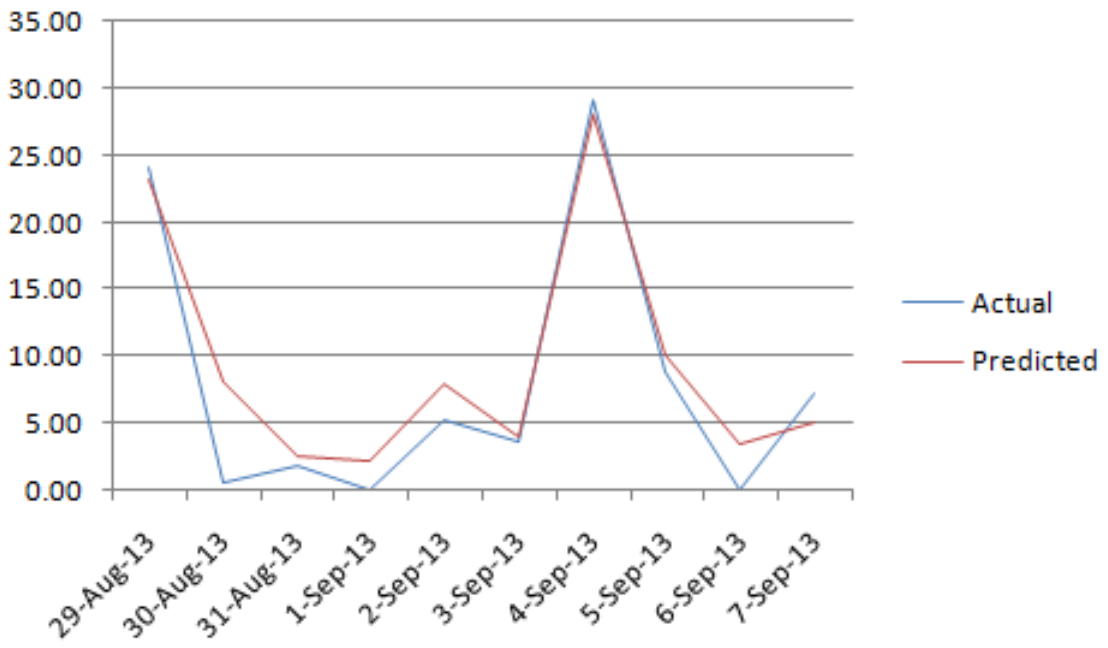


Fig. 5.9 Total Rainfall Prediction Using Combined Dataset In Support Vector Regression, Horizon 1.

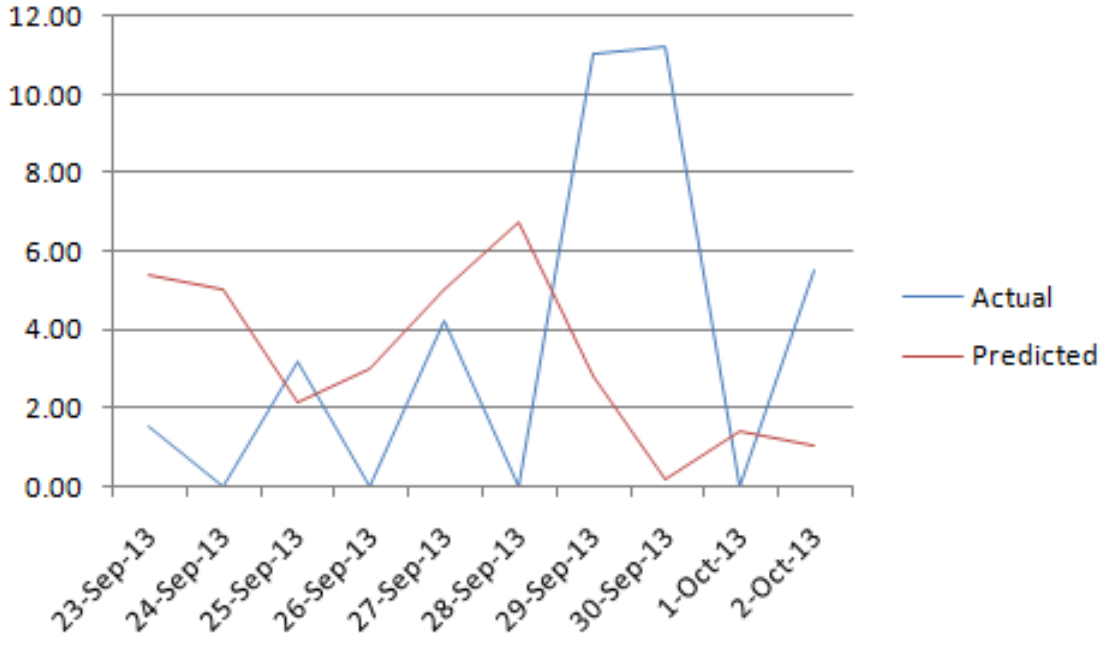


Fig. 5.10 Total Rainfall Prediction Using Combined Dataset In Support Vector Regression, Horizon 7.

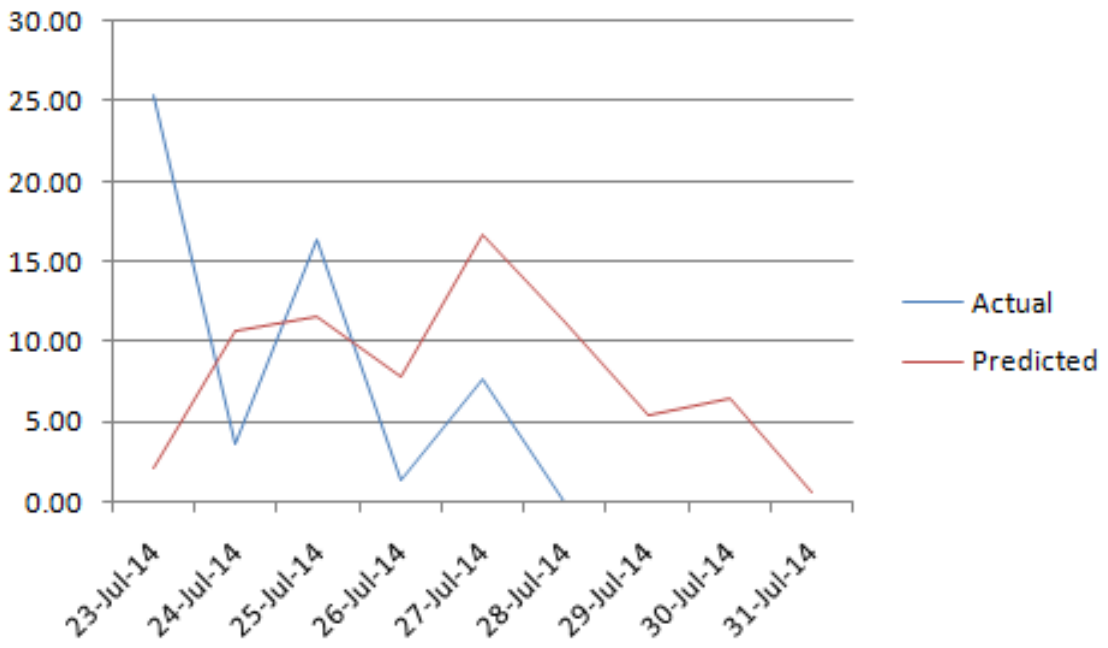


Fig. 5.11 Total Rainfall Prediction Using Combined Dataset In Support Vector Regression, Horizon 10.

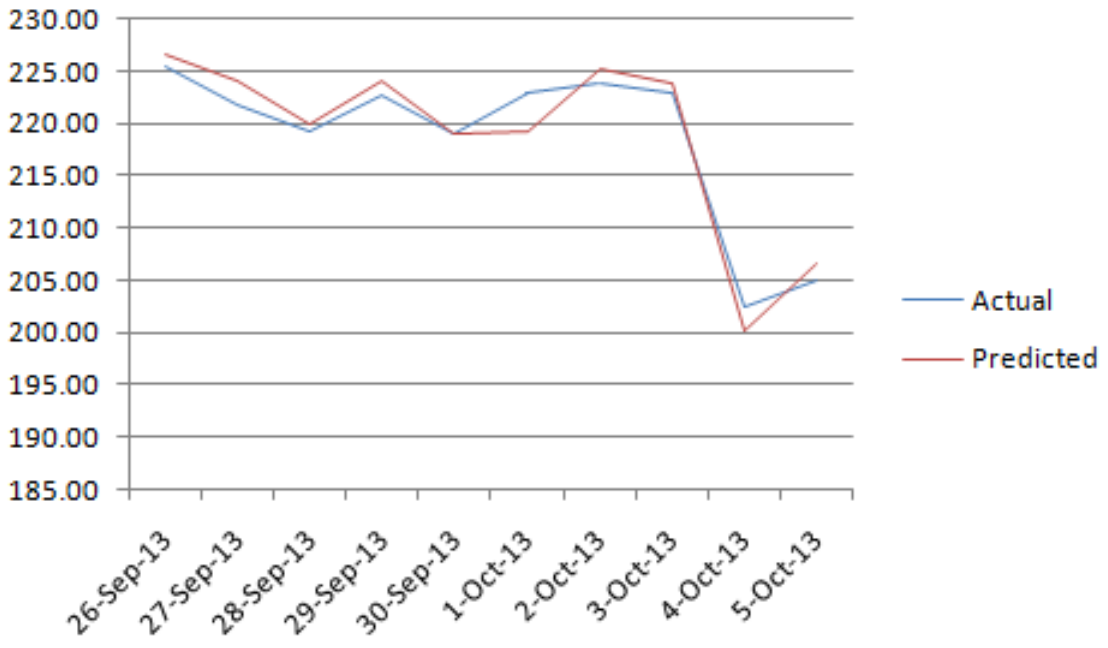


Fig. 5.12 Total Temperature Prediction Using Combined Dataset In Support Vector Regression, Horizon 1.

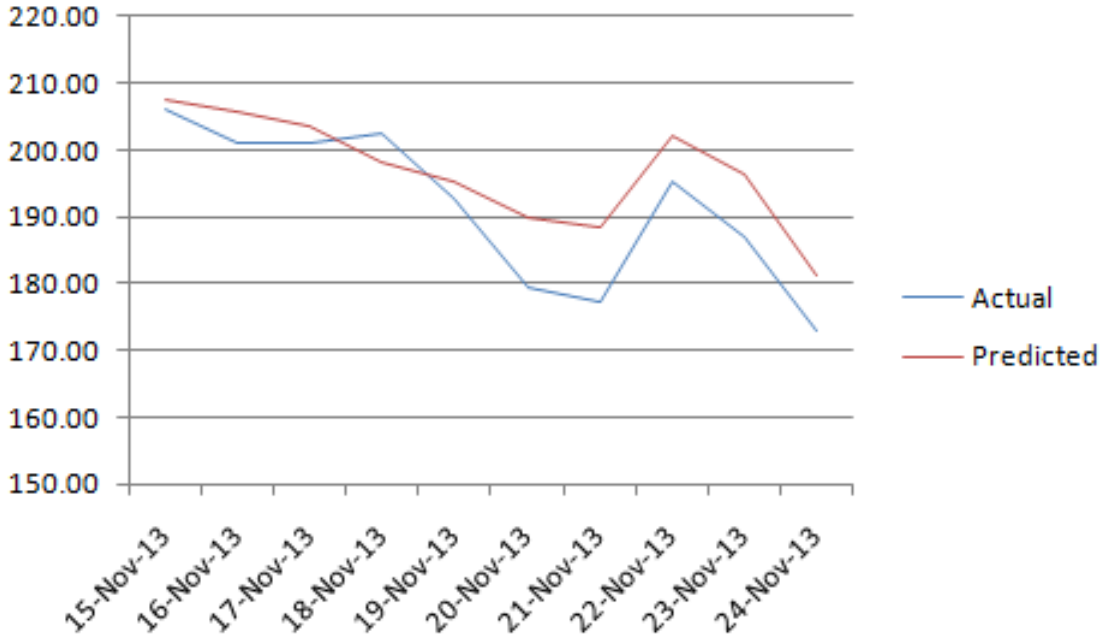


Fig. 5.13 Total Temperature Prediction Using Combined Dataset In Support Vector Regression, Horizon 7.

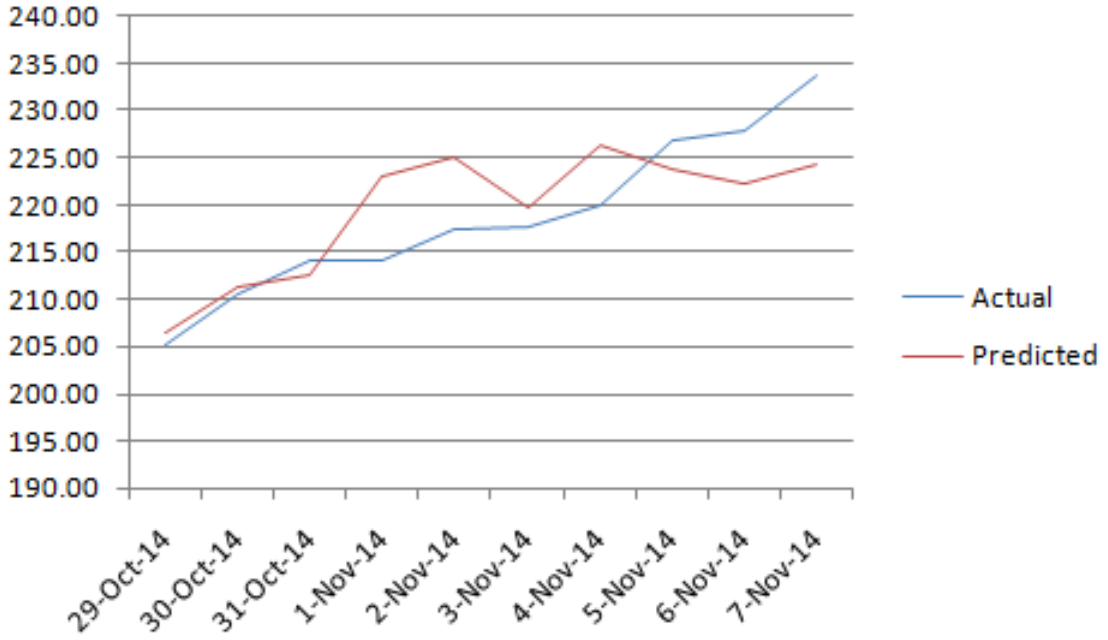


Fig. 5.14 Total Temperature Prediction Using Combined Dataset In Support Vector Regression, Horizon 10.

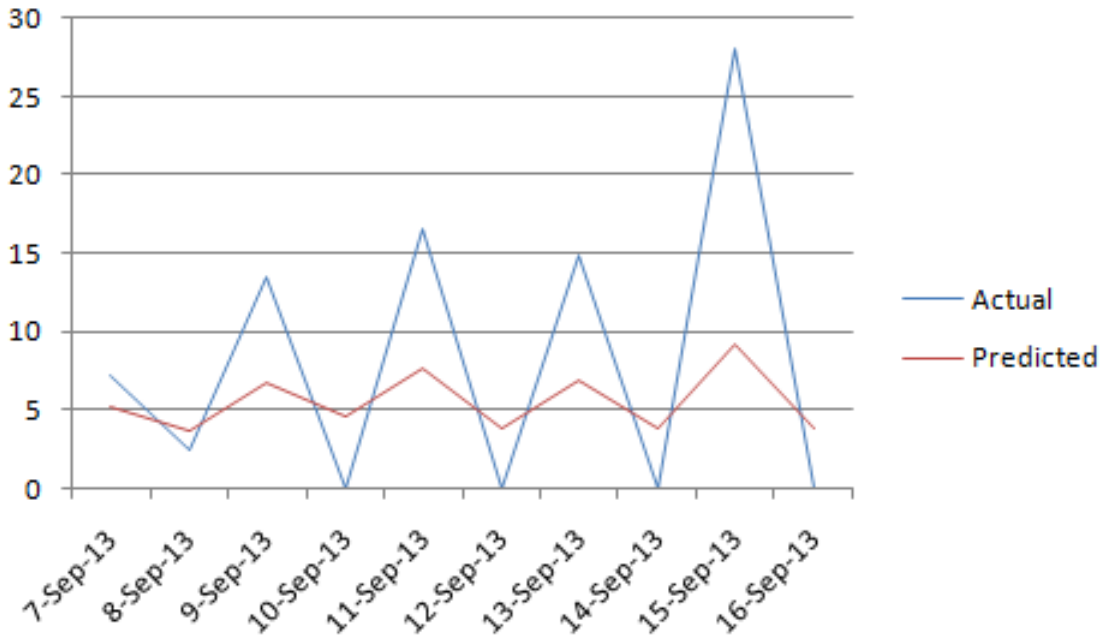


Fig. 5.15 Total Rainfall Prediction Using Single Dataset In Neural Net, Horizon 1.

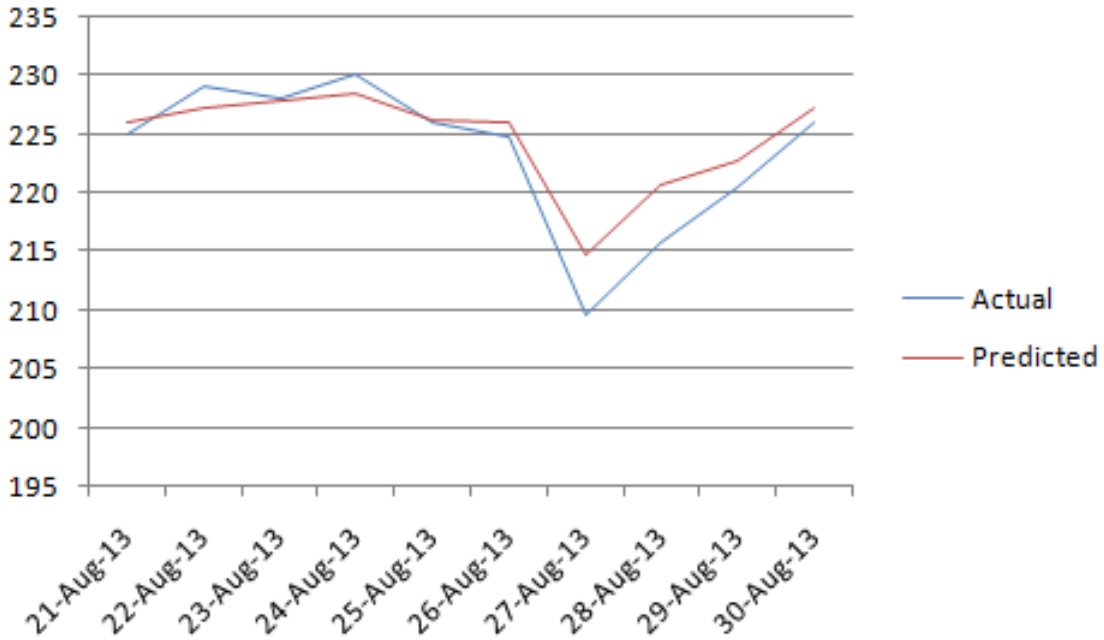


Fig. 5.16 Total Temperature Prediction Using Single Dataset In Neural Net, Horizon 1.

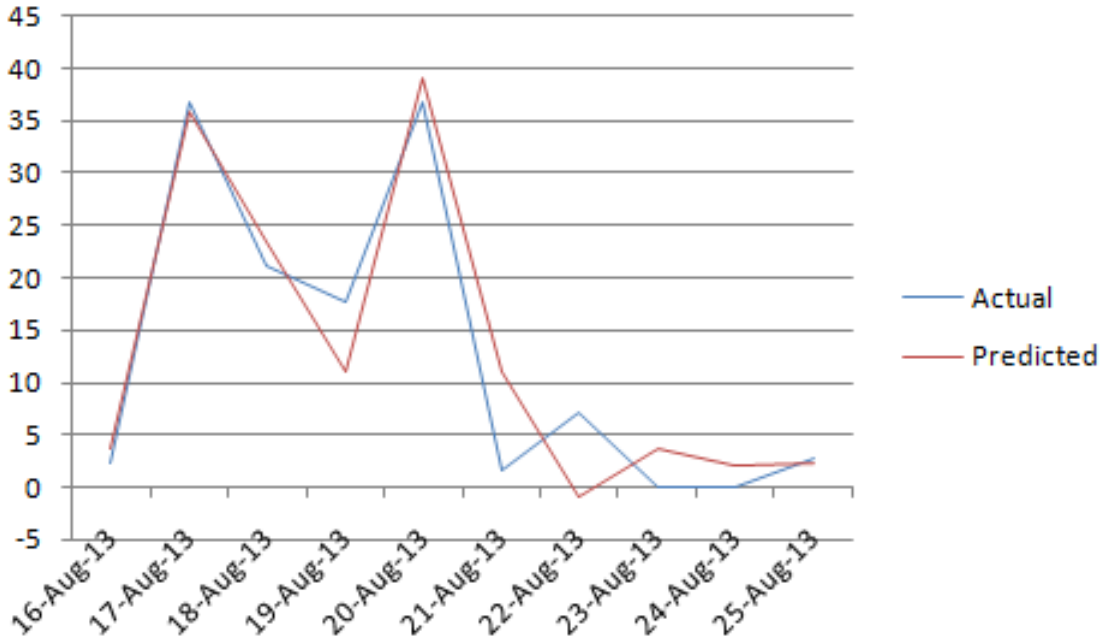


Fig. 5.17 Total Rainfall Prediction Using Single Dataset In Support Vector Regression, Horizon 1.

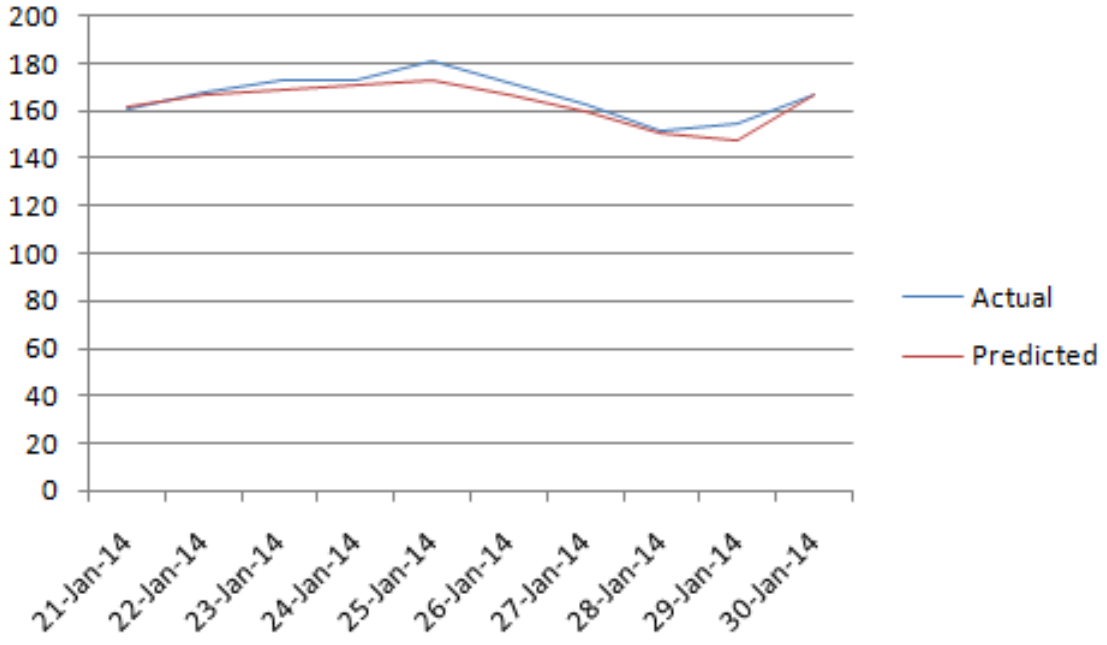


Fig. 5.18 Total Temperature Prediction Using Single Dataset In Support Vector Regression, Horizon 1.

Chapter 6

Conclusion

6.1 Summary

The purpose of the study is to observe performance of different Machine Learning and Data Mining techniques to forecast weather and to propose a weather forecasting model to forecast weather with high accuracy. Two different novel Data Mining techniques, Support Vector Regression- a regression method using Support Vector Machine and conventional Artificial Neural Network were used to conduct the study. At first a feature selection model; I_{GAIN} was used to determine the necessary features. The data was fed to the algorithms in order to train and test the model.

As one of our objectives was to determine the combination of feature to predict the weather, it is visible from the results that SVR can perform better for Rainfall prediction for both single and combined dataset and ANN performs better for Temperature prediction for both single and combined dataset.

6.2 Limitation And Future Work

- In this study we have just used data from a single station of a country. We did not compare our techniques with other dataset. Only six years data was considered to build the models. For ANN models we used maximum 3 hidden layer networks. Only Anova type kernel is used in this study.
- We will use different dataset from different areas of the world and much more data to conduct our future work. We will use different settings for hidden layers in Neural networks and other different kernel types for Support Vector Regression technique.

References

[1] F. Olaiya and A. B. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 1, p. 51, 2012.

[2] W. H. Klein and H. R. Glahn, "Forecasting local weather by means of model output statistics," *Bulletin of the American Meteorological Society*, vol. 55, no. 10, pp. 1217–1227, 1974.

[3] A. Selakov, D. Cvijetinović, L. Milović, S. Mellon, and D. Bekut, "Hybrid pso–svm method for short-term load forecasting during periods with significant temperature variations in city of burbank," *Applied Soft Computing*, vol. 16, pp. 80–88, 2014.

[4] J. Hurrell, G. A. Meehl, D. Bader, T. L. Delworth, B. Kirtman, and B. Wielicki, "A unified modeling approach to climate system prediction," *Bulletin of the American Meteorological Society*, vol. 90, no. 12, p. 1819, 2009.

[5] G. Brunet, M. Shapiro, B. Hoskins, M. Moncrieff, R. Dole, G. N. Kiladis, B. Kirtman, A. Lorenc, B. Mills, R. Morss *et al.*, "Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction," *Bulletin of the American Meteorological Society*, vol. 91, no. 10, p. 1397, 2010.

[6] M. Shapiro, J. Shukla, G. Brunet, C. Nobre, M. Béland, R. Dole, K. Trenberth, R. Anthes, G. Asrar, L. Barrie *et al.*, "An earth-system prediction initiative for the twenty-first century," *Bulletin of the American Meteorological Society*, vol. 91, no. 10, p. 1377, 2010.

[7] C. Nobre, G. P. Brasseur, M. A. Shapiro, M. Lahsen, G. Brunet, A. J. Busalacchi, K. Hibbard, S. Seitzinger, K. Noone, and J. P. Ometto, "Addressing the complexity of the earth system," *Bulletin of the American Meteorological Society*, vol. 91, no. 10, p. 1389, 2010.

[8] W. Hazeleger, C. Severijns, T. Semmler, S. Stefanescu, S. Yang, X. Wang, K. Wyser, E. Dutra, J. M. Baldasano, R. Bintanja *et al.*, "Ec-earth: a seamless earth-system prediction approach in action," *Bulletin of the American Meteorological Society*, vol. 91, no. 10, pp. 1357–1363, 2010.

[9] C. Senior, A. Arribas, A. Brown, M. Cullen, T. Johns, G. Martin, S. Milton, S. Webster, and K. Williams, "Synergies between numerical weather prediction and general circulation climate models," *The development of atmospheric general circulation models: complexity, synthesis, and computation. Cambridge University Press, Cambridge*, 2010.

[10] L. Xiong and K. M. O'Connor, "An empirical method to improve the prediction limits of the glue methodology in rainfall–runoff modeling," *Journal of Hydrology*, vol. 349, no. 1, pp. 115–124, 2008.

[11] F. Habets, P. LeMoigne, and J. Noilhan, "On the utility of operational precipitation forecasts to served as input for streamflow forecasting," *Journal of Hydrology*, vol. 293, no. 1, pp. 270–288, 2004.

[12] L. F. Richardson, *Weather prediction by numerical process*. Cambridge University Press, 2007.

[13] G. Marchuk, "Numerical methods of weather forecasting," DTIC Document, Tech. Rep., 1970.

[14] V. M. Krasnopolsky and M. S. Fox-Rabinovitz, "Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction," *Neural Networks*, vol. 19, no. 2, pp. 122–134, 2006.

[15] R. J. Kuligowski and A. P. Barros, "Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks," *Weather and forecasting*, vol. 13, no. 4, pp. 1194–1204, 1998.

[16] L. Chen and X. Lai, "Comparison between arima and ann models used in short-term wind speed forecasting," in *Power and Energy Engineering Conference (APPEEC), 2011 Asia-Pacific*. IEEE, 2011, pp. 1–4.

[17] I. Horenko, R. Klein, S. Dolaptchiev, and C. Schütte, "Automated generation of reduced stochastic weather models i: simultaneous dimension and model reduction for time series analysis," *Multiscale Modeling & Simulation*, vol. 6, no. 4, pp. 1125–1145, 2008.

[18] C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet, "Numerical weather prediction (nwp) and hybrid arma/ann model to predict global radiation," *Energy*, vol. 39, no. 1, pp. 341–355, 2012.

[19] A. S. Cofıno, R. Cano, C. Sordo, and J. M. Gutierrez, "Bayesian networks for probabilistic weather prediction," in *15th Eureopean Conference on Artificial Intelligence (ECAI)*. Citeseer, 2002.

[20] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted boltzmann machine," in *Advances in Neural Information Processing Systems*, 2009, pp. 1601–1608.

[21] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee, "Structured recurrent temporal restricted boltzmann machines," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1647–1655.

[22] A. Bautu and E. Bautu, "Meteorological data analysis and prediction by means of genetic programming," in *Proceedings of the 5th Workshop on Mathematical Modeling of Environmental and Life Sciences Problems Constanta, Romania*. Citeseer, 2006, pp. 35–42.

[23] G.-F. Lin and L.-H. Chen, "Application of an artificial neural network to typhoon rainfall forecasting," *Hydrological Processes*, vol. 19, no. 9, pp. 1825–1837, 2005.

[24] S.-Y. Ji, S. Sharma, B. Yu, and D. H. Jeong, "Designing a rule-based hourly rainfall prediction model," in *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*. IEEE, 2012, pp. 303–308.

[25] T. Hall, H. E. Brooks, and C. A. Doswell III, "Precipitation forecasting using a neural network," *Weather and forecasting*, vol. 14, no. 3, pp. 338–345, 1999.

[26] J. Wu, L. Huang, and X. Pan, "A novel bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting," in *Computational Science and Optimization (CSO), 2010 Third International Joint Conference on*, vol. 2. IEEE, 2010, pp. 466–470.

[27] J. Wu and E. Chen, "A novel nonparametric regression ensemble for rainfall forecasting using particle swarm optimization technique coupled with artificial neural network," in *Advances in Neural Networks–ISNN 2009*. Springer, 2009, pp. 49–58.

[28] W.-C. Hong, "Rainfall forecasting by technological machine learning models," *Applied Mathematics and Computation*, vol. 200, no. 1, pp. 41–57, 2008.

[29] K. Lu and L. Wang, "A novel nonlinear combination model based on support vector machine for rainfall prediction," in *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*. IEEE, 2011, pp. 1343–1346.

[30] A. Mellit, A. M. Pavan, and M. Benghanem, "Least squares support vector machine for short-term prediction of meteorological time series," *Theoretical and applied climatology*, vol. 111, no. 1-2, pp. 297–307, 2013.

[31] R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. Van Knowe, K. Hemker, D. Heinemann, J. Remund, S. C. Müller, W. Traunmüller *et al.*, "Comparison of numerical weather prediction solar irradiance forecasts in the us, canada and europe," *Solar Energy*, vol. 94, pp. 305–326, 2013.

[32] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines, advances in neural information processing systems 9," 1997.

[33] C.-H. Wu, J.-M. Ho, and D.-T. Lee, "Travel-time prediction with support vector regression," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 5, no. 4, pp. 276–281, 2004.

[34] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Artificial Neural Networks—ICANN'97*. Springer, 1997, pp. 999–1004.

[35] K. Muller, A. Smola, G. Ratch, B. Scholkopf, J. Kohlmorgen, and V. Vapnik, "Using support vector support machines for time series prediction," *Image Processing Services Research Lab, AT&T Labs*, 2000.

[36] L. K. Lai and J. N. Liu, "Stock forecasting using support vector machine," in *2010 International Conference on Machine Learning and Cybernetics*, 2010.

[37] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[38] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of mathematical biology*, vol. 52, no. 1-2, pp. 99–115, 1990.

[39] R. I. Rasel, N. Sultana, and P. Meesad, "An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators," *International Journal of Computational Intelligence Studies*, vol. 4, no. 2, pp. 134–150, 2015.

[40] http://cdn2.hubspot.net/hub/64283/file-15469871-png/images/time-series-forecasting-using-windowing-in-rapidminer-resized-600.png, [Online; accessed 31-May-2016].

[41] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, 2013, p. 1.

[42] T. Hastie, R. Tibshirani, and J. Friedman, *Unsupervised learning*. Springer, 2009.

[43] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

Appendix A

LIST OF ABBRIVIATION

SVR - Support Vector Regression
SV - Support Vector
ANN - Artificial Neural Network
MA - Moving Average
NN - Neural Network
M - Momentum
G - Kernel Gamma
SVM - Support vector Machine
RMSE - Root Mean Square Error
RMSD - Root Mean Square Deviation
MAE - Mean Absolute Error
Max - Maximum
Min - Minimum
Avg - Average
Temp - Temperature
DNN - Deep Neural Network