

ANN in forecasting Missing Rainfall Data

Jagottam Das Agrawal^{1*},

¹Department of Civil Engineering, Dr D Y Patil Institute of Technology, Pimpri, Pune, India.

Abstract. ANN has been used to estimate rainfall data by analysing patterns and mapping the correlation between historical data and weather patterns. ANNs are a type of machine learning algorithm that is modelled on the structure of the human brain, which makes them particularly effective for solving complex problems that involve large amounts of data. Overall, the use of ANNs for estimating rainfall data is a promising area of research that has the potential to provide valuable insights into weather patterns and their impacts on the environment and human society. The potential of ANN in the estimation of missing precipitation values has been investigated in detail. This study proposes to compare the performance of conventional methods as well as different algorithms of the Artificial Neural Network method to predict missing rainfall in the Upper Tapi catchment area in the West region of India. It was found that the ANN method has an edge over the conventional methods and proved to be a better method of finding the missing rainfall data values.

1. Introduction

Rainfall data in the specified area is the most important variable in hydrological modelling in forecasting the runoff. Rainfall data used shall be complete and reliable for accurate forecasting. Normally ground-based rain gauges are used for collection of such data. A large number of rain gauges are required to be deployed in the area of concern for collection of such data. However, such data may contain missing values due to error/ malfunctioning of the rain-gauges or due to some other reason.

The accuracy of rainfall-runoff models depends upon the consistency and reliability of rainfall data which is sometimes missing for a few days or longer periods. These missing data gaps in rainfall can be filled in several ways depending on the type of data, the distance between the measuring stations, etc. In most of the cases where missing rainfall data gap is of short duration, simple methods such as the Arithmetic Average Method, Normal Ratio (NR) Method, Inverse Distance Method, etc. are commonly employed. In the recent past, Artificial Neural Network (ANN) has been successfully utilized for the analysis and prediction of hydrological processes due to its simplicity and enhanced accuracy in prediction.

* Corresponding author: jagottam.agrawal@dypvp.edu.in

There have been several studies that have used ANN to fill in missing rainfall data. Here are a few examples of literature on the topic: Arithmetic mean, NR method and IDW methods were used for estimation of missing data for a short-term period. It was found that none of these methods were able to accurately estimate the rainfall data for short-term duration [1]. Long-term short-term estimation of missing rainfall data was done using various prevalent methods along with deep neural network model. It was concluded that deep neural network model estimated better results as compared to other models [2]. ANN has been found to be better as compared to other methods of prediction for computation of missing rainfall data [3]. It is not advisable to rely on a data obtained by a single technique for fulfilling the missing data requirements [4]. Using the K-nearest neighbour algorithm for estimation of missing rainfall data, they were able to estimate the data with quite good accuracy [5]. Used artificial neural networks to predict the runoff discharge one day ahead using ANN, model can predict both short- and long-term runoff discharges accurately because of using multi-scale time series of rainfall and runoff data as the ANN input layer. They have used ANN successfully to predict tidal data very accurately. The literature shows that ANN has been successfully used to fill in missing rainfall data in various hydrological datasets [6]. ANN models have been found to accurately predict missing rainfall data, and in some cases, they have been found to outperform other traditional interpolation techniques. However, the performance of the ANN model can vary depending on the complexity of the dataset and the specific modelling approach used. In this study, ANN has been used to predict missing rainfall data for two rain gauge stations in a River catchment, a part of the Upper Tapi catchment using three different ANN algorithms. The results obtained using ANN are then compared with the conventional methods.

2. Study area and data used

The Purna river is a major left-bank tributary of the Tapi river Fig.1. There were 22 rain-gauge stations recording the rain fall data. It was assumed that rain gauge stations at Kholapur and Nandura are not working, accordingly they were selected for the reconstruction of missing rainfall data. Rain fall data of surrounding stations was used for the prediction purposes. These stations were also selected so that the predicted data can be utilized to improve the accuracy in runoff predictions at Burhanpur and Yerli weir. These data can be used to predict the inflow at Hatnur reservoir in the downstream.

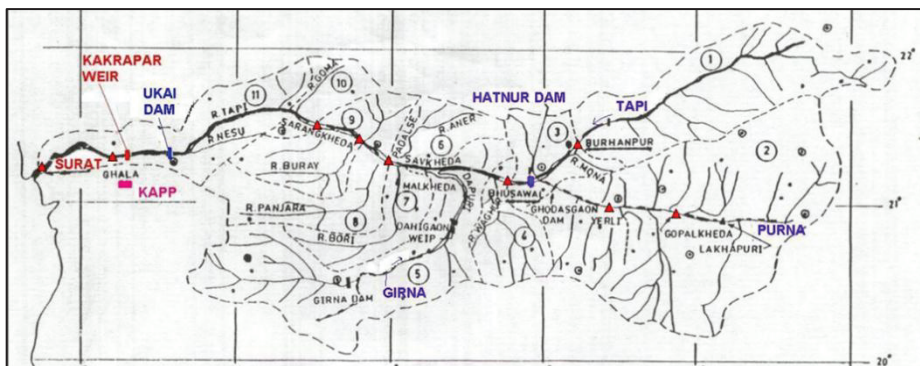


Fig. 1. Tapi river basin map.

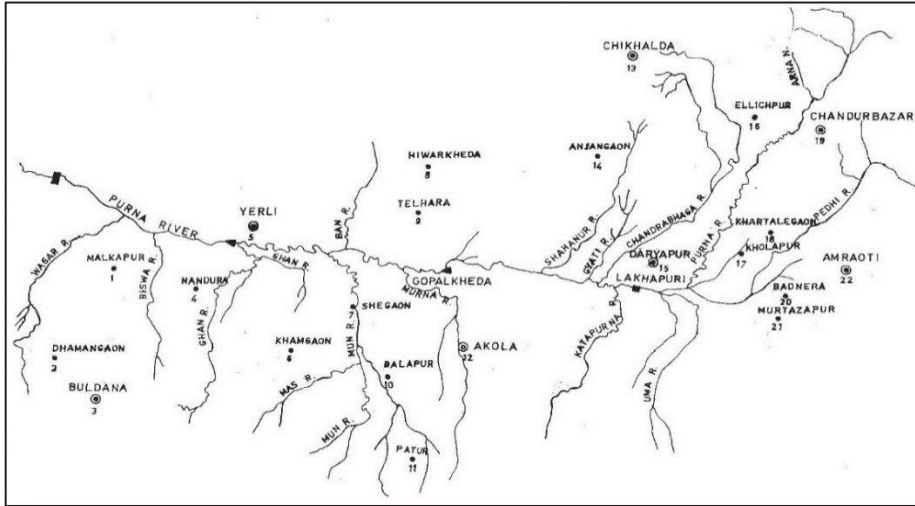


Fig. 2 Rain gauge stations.

2.1 Rainfall data

Rainfall records of the nine stations out of 22 stations were used for this study. Rainfall data during the monsoon period of June to September were only available. As stated earlier, in order to check the accuracy of the prediction using ANN, it was assumed that the rainfall data at these two rain gauge stations are missing. The missing rain gauge data were predicted using conventional techniques as well through ANN. Normal annual rainfall at these 9 stations is as follows:

Table 1. Normal annual rainfall.

Gauging station	Rainfall mm	Gauging station	Rainfall mm	Gauging station	Rainfall mm
Kholapur	601	Nandura	618	Murtazpur	789
Daryapur	502	Malkapur	685	Yerli	683
Amroati	758	Shagaon	653	Khagaon	619

The above data was used for the studies. These stations were divided into two groups of 4 and 5 stations. Kholapur being at a central location (Fig 2) among the four stations, it was assumed that the values of rainfall data at Kholapur are missing. Observed values of daily rainfall, for seven years, for these stations were used for the study. The total available data (854 days) of four stations were divided into a training set and a testing set consisting of (80%), and (20%) respectively. Similarly, Nandura being at a central location (Fig 2) among the five stations, it was assumed that the values of rainfall data at Nandura is missing. Observed values of daily rainfall for nine years for these stations were used for the study. The total available data (1098 days) of five stations were divided into a training set (80%) and a testing set (20%). As stated earlier, the training of rainfall data at Kholapur and Nandura was done, till the desired value of error was achieved by changing the number of neurons in the hidden layer and the number of epochs (iterations). This lower error value was decided by making a sensitivity analysis of trained weights and biases for testing data and comparing

the output of testing with that of observed values, till a satisfactory performance was achieved. These weights and biases were retained for testing.

3. Methodology

3.1 Rainfall prediction

Three conventional methods i.e., Normal Ratio method, Arithmetic Average method, and the Inverse Distance method, along with one data-driven model with three different algorithms have been used for the prediction of missing rainfall data, and the results from these models were compared. The annual precipitation at the adjacent stations near Nandura (surrounding stations are Malkapur, Shagaon, Khamgaon, Yerli) and Kholapur (surrounding stations are Daryapur, Amroati, Murtaazpur) shown in Fig. 2 were estimated by an arithmetic average of the rainfalls at the adjacent gauges using the equation.

$$PA = \frac{1}{n} (P_1 + P_2 + P_3 + P_4 + \dots + P_n) \quad (1)$$

Using Normal ratio method, the observed precipitation at each surrounding station was weighted by the ratio of the normal annual precipitation at station 'A' for which data was missing to the normal annual precipitation at that station. The estimated value at station 'A' was taken as the sum of the weighted values.

$$R_A = \frac{\sum_{i=1}^n \frac{NR_A R_i}{NR_i}}{n} \quad (2)$$

Where R_A is the estimated rainfall at station A, R_i is the rainfall at surrounding stations, NR_A is the normal monthly or seasonal rainfall at station A, NR_i is the normal monthly or seasonal rainfall at station I, n is the number of surrounding stations whose data are used for estimation. In the inverse distance method, the rainfall at station A was estimated as a weighted average of observed rainfall at the neighbouring stations. The weights are taken equal to the reciprocal of the distance or some power of distance of the estimator stations from the estimated stations. The relation used was as follows:

$$R_A = \frac{\sum_{i=1}^N \frac{R_i}{D_i^2}}{\sum_{i=1}^N \frac{1}{D_i^2}} \quad (3)$$

Where R_A is the estimated rainfall at station 'A', R_i is the rainfall at surrounding stations, D_i is the distance of estimator station from the estimated station.

3.2 Artificial neural network (ANN)

The ANN model is trained using a three-layered, feed-forward backpropagation algorithm that adjusts the weights of the connections between the neurons to minimize the error between the predicted and actual values Fig 3. The training process continues until the error between the predicted and actual values reaches a minimum. This algorithm adjusts the interconnection weights during training, based on the generalized delta rule proposed by [7].

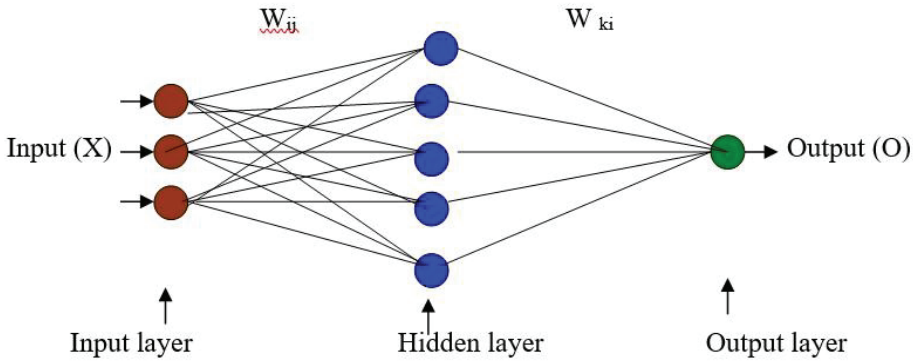


Fig 3. Structure of ANN Model for missing data analysis

3.3 ANN model setup

The ANN model used has been designed considering the number of neurons and number of layers in the model, activation function to be used and optimization algorithm to be used during the training process. ANN model was trained using training data sets. The ANN model learns to predict the missing rainfall data by adjusting the weights of the connections between neurons based on the error between the predicted and actual values. ANN model has been validated using the 20% data sets kept aside for this purpose. The performance of the model has been evaluated using mean squared error, root mean squared error, and correlation coefficient. Model has been tested using the testing dataset to evaluate its performance in predicting the missing rainfall data. The testing dataset is a new dataset that the model has not gone through before. The final step is to evaluate the results of the ANN model and compare them with other interpolation techniques.

The equation for the model developed is as follows.

$$p_x(t) = f [p_1(t), p_2(t), p_3(t), p_4(t)] \tag{4}$$

Where: p_x : Rainfall at station x(Missing), p_1, p_2, p_3, p_4 :Rainfall at four neighboring stations. ANN in this study was trained and simulated using the software MATLAB. All the data used in the study have been normalized between 0 and1 based on log-sigmoidal function. The normalization has been done using the following equation.

$$\bar{X} = \frac{(X-X_{min})}{X_{max}-X_{min}} \tag{5}$$

Where \bar{X} is the normalized value of the input, X_{min} and X_{max} are respectively, the minimum and maximum of the actual values, in all observations and X is the original data set.

3.4 ANN Algorithms used

Three training Algorithms namely Levenberg-Marquardt (LM), Conjugate Gradient Fletcher-Reeves (CGFR), and Broyden-Fletcher-Goldfarb-Shanno (BFGS) were used. Levenberg-Marquardt (LM) provides a powerful optimization technique for training ANNs, enabling them to learn and adapt to complex patterns in the data more efficiently. Backpropagation technique enable the complex patterns and make accurate predictions. It

enables ANNs to learn from labeled training data by iteratively adjusting the network's weights and biases to minimize the error between the predicted output and the desired output. The weight updates are repeated iteratively until a stopping criterion is met. This criterion could be a maximum number of iterations, convergence of the error, or reaching a desired level of performance. The basic gradient descent algorithm adjusts the weights in the direction in which the performance function is decreasing most rapidly, which doesn't lead to faster convergence. In the conjugate gradient Fletcher-Reeves (CGFR) algorithms, a search is performed along conjugate directions, which produces generally faster convergence than steepest descent directions. Conjugate Gradient backpropagation with Fletcher-Reeves is the ratio of the norm squared of the current gradient to the norm squared of the previous gradient. The conjugate gradient algorithms are usually much faster than other algorithms but the result depends on the problem. Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is an optimization algorithm used to solve unconstrained nonlinear optimization problems. It is an iterative method that aims to find an objective function by iteratively updating an estimate of the optimal solution. Once the convergence criterion is satisfied, terminate the algorithm and return the estimated optimal solution.

4. Results

4.1 Performance evaluation

Graphical, numerical performance indicators, have been used for comparison of results. The graphical performance indicators used in this paper are as follows:

- A linear scale plot of the simulated and observed data series
- A scatter plot of the simulated versus observed data series for the validation period.

The root mean square error (R_{mse}), Nash-Sutcliffe Efficiency (NSE), also known as Nash-Sutcliffe model efficiency coefficient or simply 'E' and coefficient of correlation 'R' are the numerical performance estimators. They are defined as follows:

$$R_{mse} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{y}_i)^2}{N}} \quad (6)$$

$$E = 1 - \left(\frac{\sum (x_i - y_i)^2}{\sum (x_i - \bar{x})^2} \right) \quad (7)$$

$$(R) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (8)$$

Where:

- x_i is the observed value at time i
- y_i is the modeled/predicted value at time i
- \bar{x} is the mean of the observed values
- \bar{y} is the mean of the modelled values

All the results were evaluated based on the above criteria. RMSE provides a measure of how well a model's predictions align with the actual observed values. A lower RMSE indicates that the model's predictions are closer to the actual values, suggesting higher accuracy. 'E' value of 1 indicates a perfect match between the predicted and observed values, and values close to 1 indicate good model performance. A negative value of 'E' indicates that the model

predictions are worse than simply using the mean of the observed values. The coefficient of correlation 'R' is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is commonly used to analyse the association between variables and assess the degree of their linear dependence. The correlation coefficient can range from -1 to +1. 'R' = -1 indicates a perfect negative linear relationship, 'R' = 0 indicates no linear relationship (random pattern), 'R' = +1 indicates a perfect positive linear relationship.

4.2 Discussion of results

Three training algorithms namely Levenberg-Marquardt (LM), Conjugate Gradient Fletcher-Reeves (CGFR), and Broyden-Fletcher-Goldfarb-Shanno (BFGS) were applied to these data through MATLAB. The network architecture and the performance goal for each algorithm for Kholapur is presented in Tables 2. It is seen that Kholapur normal annual rainfall is 10% above than the normal annual rainfall of the surrounding stations. This indicates the applicability of the normal ratio method. This can also be verified from the results from Table 3, wherein the normal ratio method has comparatively given better results in all the test parameters than the arithmetic average method.

Table 2. Result for ANN models at Kholapur

Algorithm	Network Architecture	Goal	Epochs	Correlation Coefficient
CGFR	3:10:1	0.009	232	0.933
BFGS	3:8:1	0.009	415	0.918
LM	3:8:1	0.004	255	0.965

Table 3. Comparison of results at Kholapur

Method	Nash-Sutcliffe Efficiency	R _{mse}	Correlation Coefficient
ANN LM Algorithm	0.9241	0.8368	0.965
Arithmetic Average	0.6769	5.4597	0.925
Normal Ratio Method	0.8185	4.0918	0.929
Inverse Distance Method	0.5948	6.1140	0.913

However, ANN was the best among the four methods, and the performance of the inverse distance method was slightly lower among all. Rain gauge station Nandura's normal annual rainfall at is within 10% of the normal annual rainfall of the surrounding stations. This proves the applicability of the arithmetic average method. This can also be verified from the results from tables 4, wherein the arithmetic average method has comparatively given better results in all the test parameters than the normal ratio method. However, ANN was best among the four methods used and the performance of the inverse distance method was in between the ANN and other methods Table 5. The trained network was tested for the remaining 20% of data, yielded satisfactory results for missing rainfall data at the same time step. For ANN model at Kholapur prediction accuracy was very good as the neighboring stations were equally distributed and truly represented the rainfall pattern of the catchment. However, at Nandura the predicted values were in good agreement with the observed values.

Table 4. Comparison of results at Nandura

Method	Nash-Sutcliff Efficiency	R _{mse}	Correlation Coefficient
ANN LM Algorithm	0.800	5.756	0.900
Arithmetic Average	0.632	7.407	0.802
Normal Ratio Method	0.613	7.590	0.797
Inverse Distance Method	0.697	6.713	0.847

Table 5. Result for ANN models at Nandura

Algorithm	Network Architecture	Goal	Epochs	Correlation Coefficient
CGFR	4:11:1	0.02	8	0.699
BFGS	4:16:1	0.01	351	0.797
LM	4:14:1	0.006	212	0.900

Table 6. Comparison of results of various methods at Kholapur for 10-days average data

Method	Nash-Sutcliff Efficiency	R _{mse}	Correlation Coefficient
ANN LM Algorithm	0.896	14.011	0.960
Arithmetic Average	0.830	17.894	0.955
Normal Ratio Method	0.916	12.560	0.960
Inverse Distance Method	0.789	19.947	0.942

Table 7. Comparison of results of various methods at Kholapur for monthly average data

Method	Nash-Sutcliff Efficiency	R _{mse}	Correlation Coefficient
ANN LM Algorithm	0.963	17.871	0.987
Arithmetic Average	0.706	50.436	0.935
Normal Ratio Method	0.868	34.325	0.939
Inverse Distance Method	0.626	56.866	0.924

This is evident from the scatter plot and the accompanying high values of correlation coefficient and rainfall plots showing observed and predicted values. It was possible to compare the lows and peaks in the time history plots. ANN model was also run by taking 10 days' average value and monthly average rainfalls with respect to known values at neighboring stations. This approach yielded improved results as compared to daily rainfall data, the reason could be that the error was averaged due to large data sets. The scatter plot at Kholapur for 10 daily rainfall data and comparison of rainfall data between estimated and actual is given in Fig. 4 and 5 respectively. Performance indicators for 10 days average and monthly average values of rainfall data are shown in above in Tables 6 & 7. Fig 6 shows the rainfall data results obtained by conventional methods and ANN. It can be seen that ANN has edge over other methods.

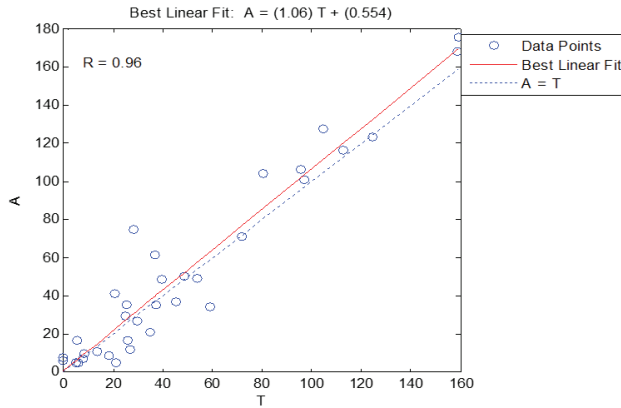


Fig 4. Scatter plot at Kholapur for 10 daily rainfall data

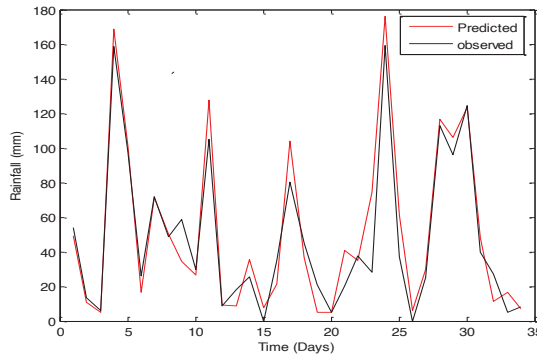


Fig 5. Comparison of rainfall data between estimated and actual at Kholapur.

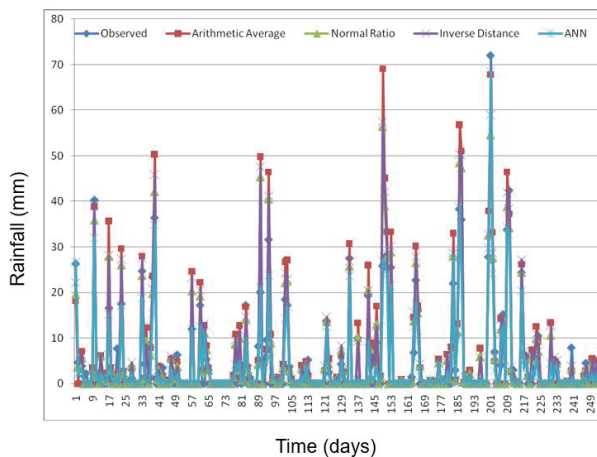


Fig 6. Plot of rainfall data results by conventional methods and ANN at Kholapur.

5. Conclusion

In conclusion, the setup of an ANN model for filling in missing rainfall data involves preparing the dataset, designing the model, training & validating the model, testing the

model, and evaluating the results. The accuracy of the ANN model depends on the complexity of the dataset and the specific modeling approach used. Rainfall data from nine stations were used to predict the missing rainfall data at two stations. The trained network when tested for unseen inputs gave satisfactory results. The results of LM algorithm were found to be the best, among the three ANN algorithms. The accuracy of the results was also verified from the low values of error measures. Hence, it can be concluded that ANN is a better alternative method for missing rainfall estimation.

References

1. A A G Nadilah and H. Hannani. *Comparison of methods to estimate missing rainfall data for shortterm period at UMB Gambang*, in 4th National Conference on Wind & Earthquake Engineering, Earth and Environmental Science **682**(2021).
2. Salim Djerboual, *Missing precipitation data estimating using long term memory deep neural network*. Journal of Ecological Engineering, **23(5)** pp 216-225, (2022).
3. Nur Afiah Ahmad Norazizi & Sayang Mohd. *Comparison of artificial neural network (ANN) and other computation methods in estimating missing rainfall data at Kuantan station*. Deni International Conference of soft computing in data science pp 298-306, Sept (2019).
4. Muluke L E, *Techniques of filling missing values of daily and monthly rainfall data-A review*, Journal of Environmental and Earth Science Vol **3**, Edition 1, (2020).
5. Vasker Sharma, Kezang Yuden. *Imputing missing data in hydrology using machine learning models*, International Journal of Engineering & Technology, Vol **10**, Jan (2021).
6. Vahid Nourani, Mehdi Komasi & Akira Mano, *A Multivariate ANN-Wavelet Approach for Rainfall–Runoff Modeling*. Springer J. Water Resources Management, Volume **23**, pages 2877–2894, (2009).
7. Rumelhart, D. E., Hinton. G. E., and Williams, R. J., *Learning internal representation by error propagation* Parallel Distributed Processing, MIT Press, Cambridge,(1986).