Development of advanced Artificial Intelligence models for daily rainfall prediction

A Project Report Submitted

for the Course

MA498 Project I

by

Himadri Boro

(Roll No. 200123024)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

November 2023

CERTIFICATE

This is to certify that the work contained in this project report entitled **"Development of advanced Artificial Intelligence models for daily rainfall prediction"** submitted by **Himadri Boro (Roll No.: 200123024)** to the Department of Mathematics, Indian Institute of Technology Guwahati towards the partial requirement of Bachelor of Technology in Mathematics and Computing has been carried out by him/her under my supervision.

It is also certified that this report is a survey work based on the references in the bibliography.

Turnitin Similarity: 40%

Guwahati - 781 039	(Dr./Prof. Bhupen Deka)
November 2016	Project Supervisor

ABSTRACT

The main objective of this study was to develop and test two Artificial Intelligence (AI) models, especially artificial neural networks (ANN) and Support vector machines (SVM) for prediction of daily rainfall. in order to achieve this goal, we collected various weather variables, including maximum temperature, minimum temperature, precipitation, wind speed, relative humidity, and solar radiation which formed the input parameters for the models. Daily rainfall was used as an output parameter in our study. We did the test Modeling using quality assessment parameters, such as correlation coefficient (R) and mean square error(MSE).To improve rainfall accuracy Forecasting, this paper presents an artificial neural network (ANN) with the Design of Backpropagation Neural Network (BPNN) and SVM. The results showed that While the ANN model provided reasonable forecasts of daily rainfall, the SVM method proved to be the most effective method for rainfall forecasting.

Contents

Chapter 1

Introduction

Rainfall plays an important role in various industries such as agriculture, hydroelectric production, and water resource management. It is especially important for Agriculture in developing countries which depends heavily on rainfall. Rainfall forecasting is important, especially in the context of climate change, which changes rainfall patterns, causing extreme events like floods and droughts. Therefore, a quick and accurate rainfall forecasting method is needed To maintain water resources management.

Timely and accurate rainfall forecasts help in planning agricultural activities, even during exceptional rainfall. Accurate predictions are also important To minimize the damage caused by natural disasters like droughts, floods, and landslides. Therefore, Artificial Neural Network (ANN) models and Support Vector Machines(SVM) will be used in this paper to forecast rainfall, which exhibits a nonlinear data structure. AI rainfall forecasting models have been prepared in the methodology. To determine if they are efficient With different data sets, we use sensitivity analysis. If a model is not strong enough, it can make mistakes. Sensitivity analysis also tells us which data are important and What doesn't. Here, we will use the Datasets from Kamrup district in Assam For our research. We will prepare the training dataset to improve the performance of ANN and SVM models.

1.1 Dataset

In this study, rainfall data were collected from a location at Latitude 26.1445° and Longitude 91.7362° in the Kamrup district, Assam. Weather data were obtained from the Power Data Access Viewer at https://power.larc.nasa.gov/data-access-viewer/. In modeling, a total of six parameters, namely maximum temperature, minimum temperature, wind speed, relative humidity, precipitation, and sky surface shortwave downward irradiance, were used as input variables. Daily rainfall data were used as an output variable for generating training and testing datasets. In total, 7761 data samples were collected during the period from January 01, 2001, to March 31, 2021.

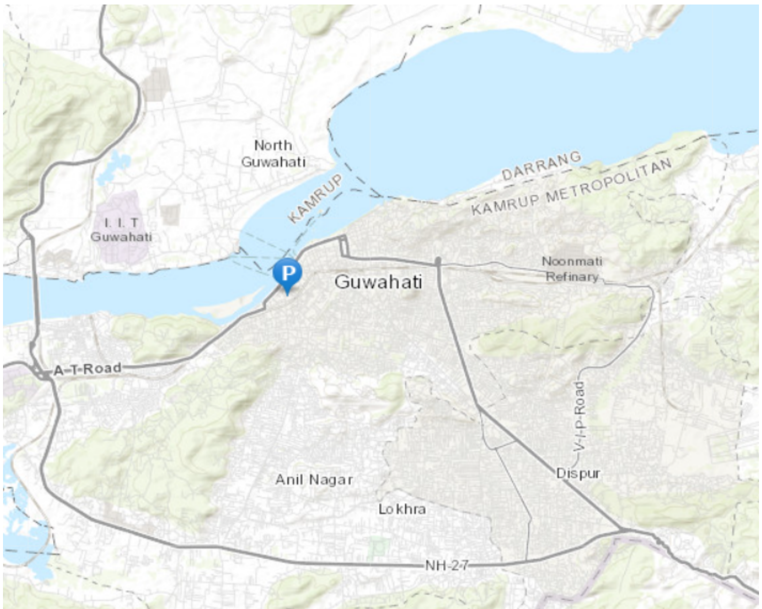
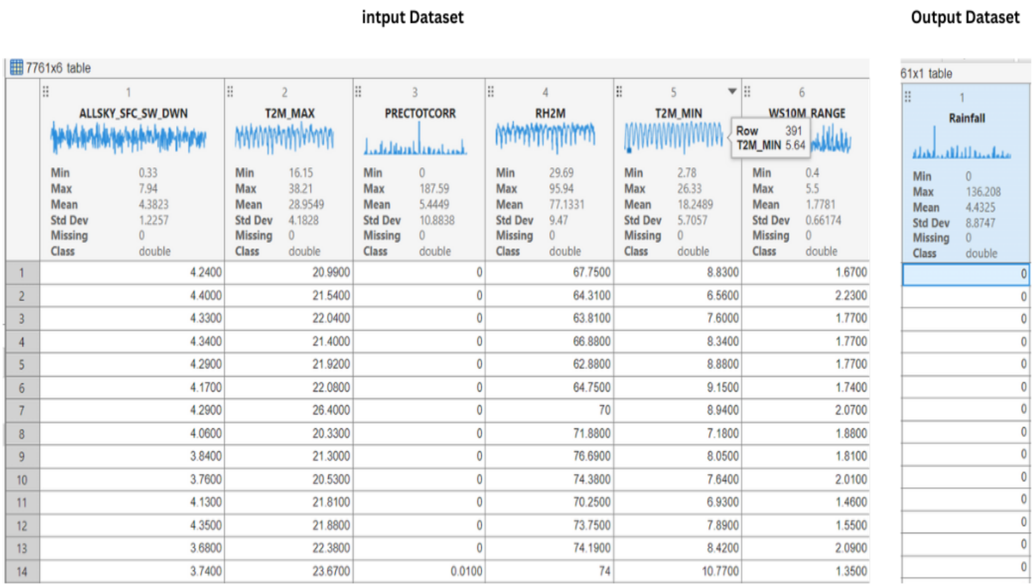


Fig. 1. Geographical Location of Kamrup, Assam



intput Dataset

7761x6 table

Row 391 T2M_MIN 5.64

	1 ALLSKY_SFC_SW_DWN	2 T2M_MAX	3 PRECTOTCORR	4 RH2M	5 T2M_MIN	6 WS10M_RANGE
Min	0.33	16.15	0	29.69	2.78	0.4
Max	7.94	38.21	187.59	95.94	26.33	5.5
Mean	4.3823	28.9549	5.4449	77.1331	18.2489	1.7781
Std Dev	1.2257	4.1828	10.8838	9.47	5.7057	0.66174
Missing	0	0	0	0	0	0
Class	double	double	double	double	double	double
1	4.2400	20.9900	0	67.7500	8.8300	1.6700
2	4.4000	21.5400	0	64.3100	6.5600	2.2300
3	4.3300	22.0400	0	63.8100	7.6000	1.7700
4	4.3400	21.4000	0	66.8800	8.3400	1.7700
5	4.2900	21.9200	0	62.8800	8.8800	1.7700
6	4.1700	22.0800	0	64.7500	9.1500	1.7400
7	4.2900	26.4000	0	70	8.9400	2.0700
8	4.0600	20.3300	0	71.8800	7.1800	1.8800
9	3.8400	21.3000	0	76.6900	8.0500	1.8100
10	3.7600	20.5300	0	74.3800	7.6400	2.0100
11	4.1300	21.8100	0	70.2500	6.9300	1.4600
12	4.3500	21.8800	0	73.7500	7.8900	1.5500
13	3.6800	22.3800	0	74.1900	8.4200	2.0900
14	3.7400	23.6700	0.0100	74	10.7700	1.3500

Output Dataset

61x1 table

	1 Rainfall
Min	0
Max	136.208
Mean	4.4325
Std Dev	8.8747
Missing	0
Class	double
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0

Fig. 2. Some Initial Observation from Datasets

1.2 Materials and Methods

Methods Used

Artificial Neural Network

Artificial Neural Network (ANN) is like a computer system inspired by our own brains. Just as our brains use nerve cells to process information, ANNs use units called neurons to do the same thing quickly. There are different ways to train ANNs, such as Perceptron, Backpropagation, Self-Organizing Map (SOM), and Delta.

Artificial neural networks have been comprised of two types:

1.Single-layered neural networks.

2.Multi-layered neural networks.

In this research, we're using the Backpropagation Neural Network (BPNN) algorithm to predict rainfall. We're looking at past data patterns to make better predictions about rain, even when the data is not in a straight line. Our goal is to be as accurate as possible and minimize errors in our predictions. In the process of training a neural network, the initial step involves providing the network with a dataset comprising both input and output values. The network's weights are set to random numbers within the range of -1 to +1. Subsequently, the input data is fed into the neural network, and the weights are adjusted based on the comparison between the produced output and the expected result. This iterative training process continues until the neural network consistently generates outputs that closely resemble the desired outcomes. Once a satisfactory level of accuracy is achieved, the weight adjustment ceases. The effectiveness of the trained weights is then validated by assessing their performance on another dataset. If the network's outputs are accurate, the training is considered successful; otherwise, the process is repeated until the neural network demonstrates reliable predictive capabilities across various datasets.

Backpropagation NN consists of 3 layers:

1.Input layer

2.Hidden layer

3.Output layer

In backpropagation, Hidden layered neural networks are calculated by the sigmoid function. It has one neuron in the output layer. All these neurons are interrelated with each layer. In Backpropagation NN input signals are captured by input neurons and output signals are captured by output neurons.

So, it is called acyclic nature. When a neural network is constructed then the weightage of the input has to be adjusted according to the outputs. Among them, it is an effective approach that is driving from the input layer to the output layer.

Sigmoid function: This function is used to bring linearity in the provided data. The three types of learning—supervised, semi-supervised, and unsupervised—fall within the broader category of machine learning, a subfield of artificial intelligence.

1.Supervised learning

2.Semi-supervised learning

3.Unsupervised learning

Supervised learning: It includes a method of ensuring that the network has the expected output by mathematically categorizing the network's production or by understanding the expected output with the inputs. Input and output information is specified to the NN to help in future data processing of the dataset.

Semi-supervised learning: semi-supervised learning makes use of unlabeled data for training. It has having lower percentage of labeled data. It has had a higher percentage of unlabeled data. The network is trained with unlabeled data to define the limits.

Unsupervised learning: In unsupervised learning, the NN should decide by avoiding external help. The data is grouped by the NN by using unsupervised learning.

The Backpropagation Neural Network (BPNN) is a supervised learning method. BPNN adjusts weights as information goes from the output to the input layer. It's similar to how humans learn through practice and examples. BPNN has three parts in each layer: input, hidden, and output layers.

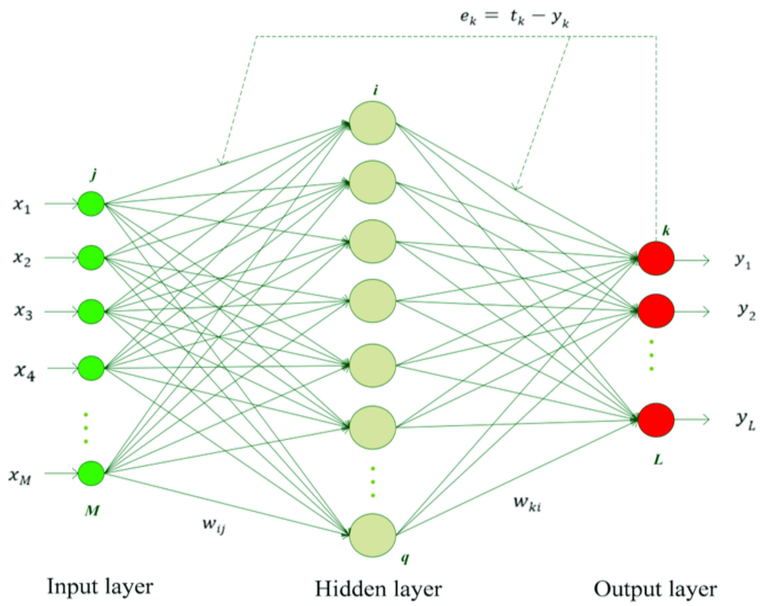


Fig. 3. A typical structure of BPNN architecture

Support Vector machines

SVM is a supervised learning algorithm that analyzes data for classification purposes and regression, designed with statistical learning theory. The SVM algorithm attempts to find a particular line(hyperplane) in N-dimensional space. It's used for sorting things into groups and making predictions. This line does a great job of dividing the data points into two different groups and trying to keep those groups as far apart as possible. Some of the data points are super important to help determine exactly where this line should go. We call them support vectors. The SVM uses a "kernel" to help it do its job. The main objective of the SVM is to ensure that this separation distance is in the right place, and it does this by trying to maximize the distance between the line and the data points.

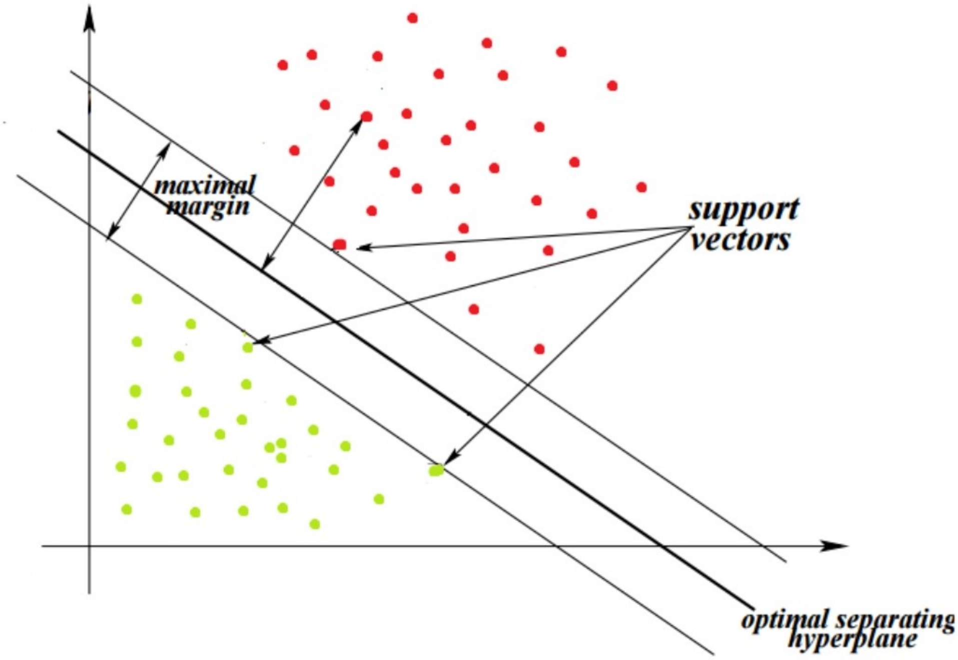


Fig. 4. Illustration of SVMs algorithm

Chapter 2

Rainfall Modeling

2.1 Methodology

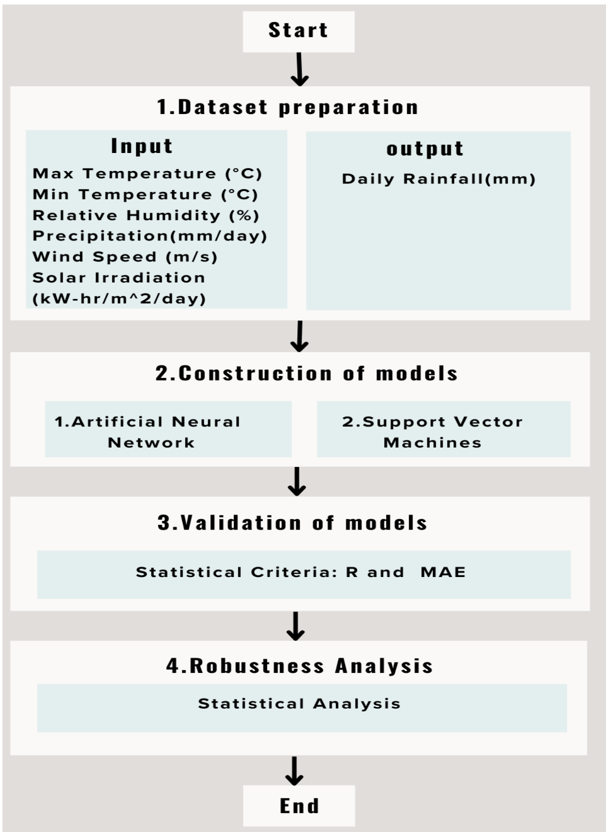


Fig. 5. Methodology Flow chart

In this paper, The Rainfall dataset is distributed in three different kinds: Training set, Validation set, and Testing set. The rainfall has been predicted using deep learning techniques. Two deep learning techniques that were used are Multilayer Perceptron and Auto-Encoders. Auto-encoders are responsible for time series forecasting by performing the feature extraction and the MLP is used in prediction and classification tasks. Firstly the auto-encoders extract all the non-linear features from the data. This auto-encoder consists of three layers:

1.Input Layer

2.Hidden Layer

3.Output Layer

The auto-encoder extracts all the non-linear features and then send them to MLP, this helps in making a prediction. The performance of the methodology is also evaluated by using R-value and MSE(Mean Square Error).

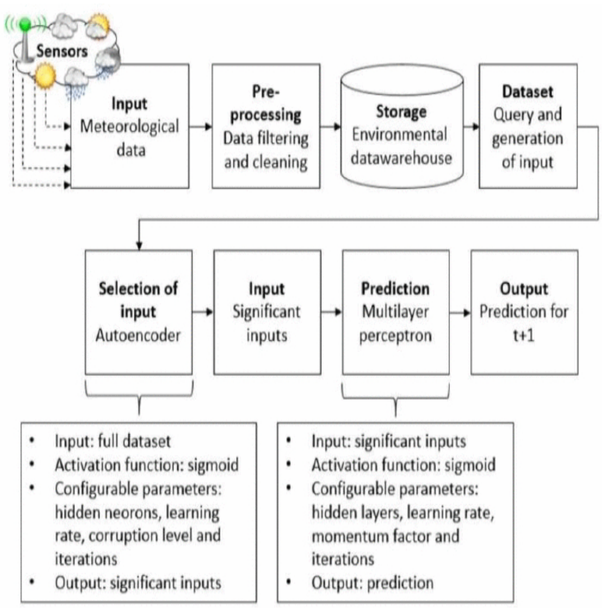


Fig. 6. Flow of architecture

To construct a support vector machine (SVM) for rainfall forecasting Different kernel functions were used: linear, radial basis function (RBF), sigmoid, and polynomial. The aim is to compare the performance of these kernels to determine which is more efficient for a given task. this The method allows a detailed analysis of the intrinsic SVM model Mathematical transformations, providing insight into forces and weaknesses of each kernel.

2.2 Result and Analysis

In Multilayer perceptron input is taken from mult i-levels and predicts the future data from the past data. In this, the data is separated into training and testing data. the number of iterations is being calculated and checked for the best Epoch value.

Model Summary

Train a neural network to map predictors to continuous responses.

Data

Predictors: input - [7761x6 double]

Responses: output - [7761x1 double]

input: double array of 7761 observations with 6 features.

output: double array of 7761 observations with 1 features.

Algorithm

Data division: Random

Training algorithm: Levenberg-Marquardt

Performance: Mean squared error

Training Results

Training start time: 09-Nov-2023 23:03:50

Layer size: 15

	Observations	MSE	R
Training	5433	47.1197	0.6231
Validation	1164	52.2079	0.6018
Test	1164	56.6873	0.5833

Fig. 7. Flow of architecture

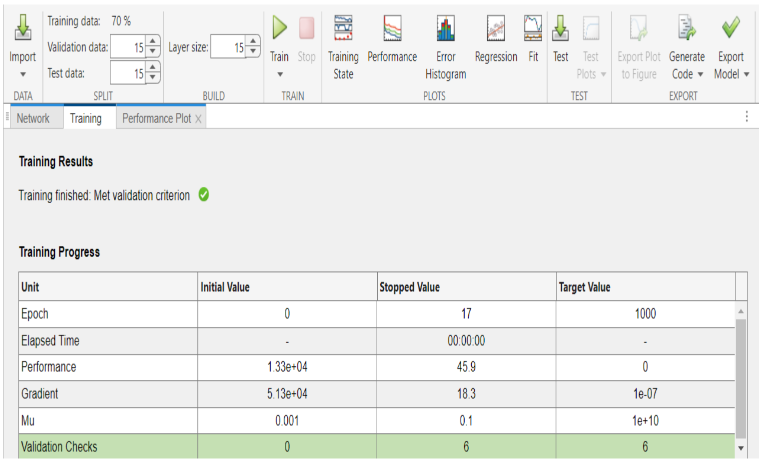


Fig. 8. Neural network training

The best Epoch value results at the 17th iteration as seen in Fig(8), got the best performance at 20. When the training and testing are performed there is a slight variation. It indicates Mean square error is less but when performed validation got the best performance at Epoch 11 as shown in Fig(10).

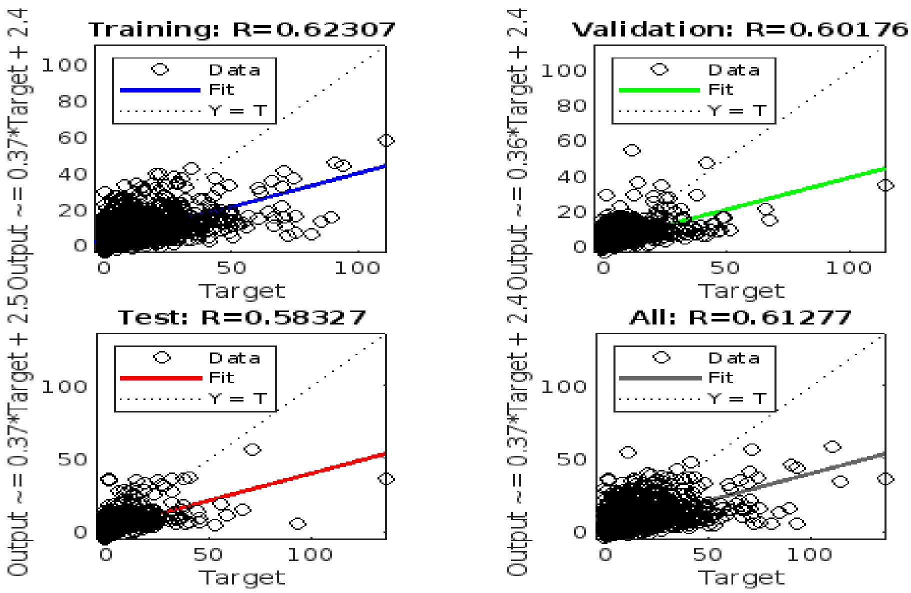


Fig. 9. Finding the best epoch value

Now the training testing and validation set results are combined to get the best result at the targeted value as shown in Fig(5)

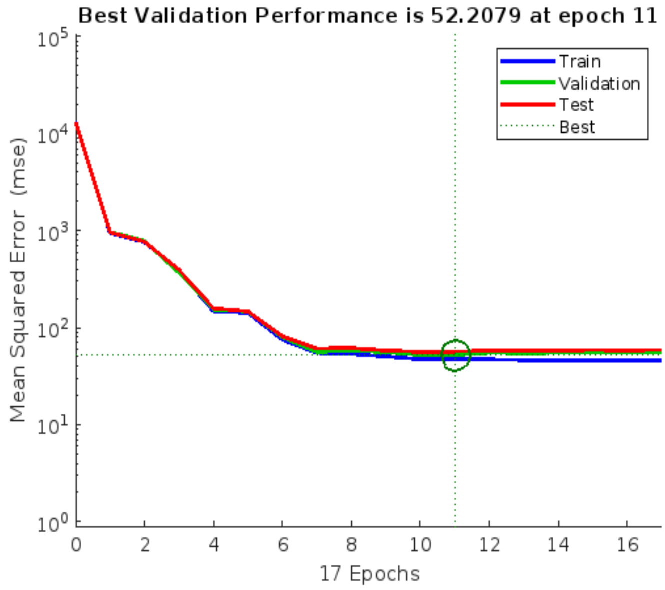


Fig. 10. Targeted vs prediction

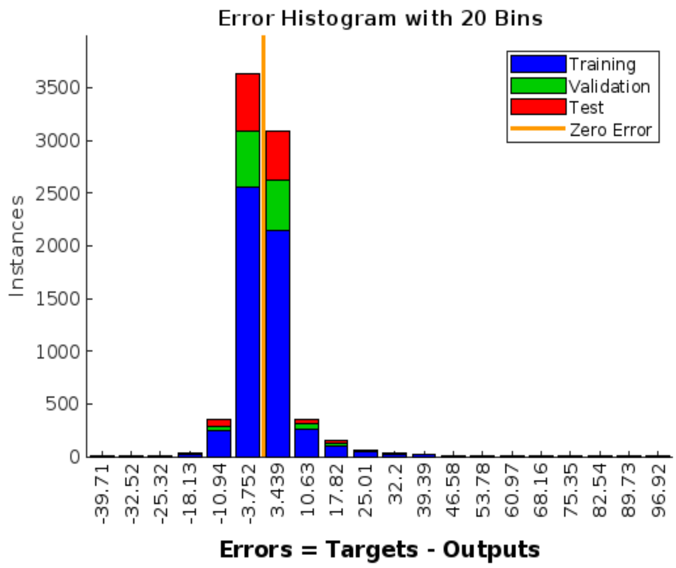


Fig. 11. Error

In our Support Vector Machine (SVM) classification task, we classified our dataset, which includes 75% of the data in the training set and the remaining 25% in the testing set. We tested four types of kernel functions: linear, radial basis function (RBF), sigmoid, and polynomial. After careful analysis, the linear kernel stood out as the most efficient one for our data set. In particular, setting the hyperparameter C to 0.001 resulted in optimal performance. The choice of C in SVM is important because it deals with the trade-off between achieving a smooth decision boundary and Equitable distribution of training points. This was carefully done with the SVM model. A linear kernel and C= 0.001 showed high predictive power, indicating the importance of choosing an appropriate kernel implementation and fine-tuning hyperparameters for optimal model performance.

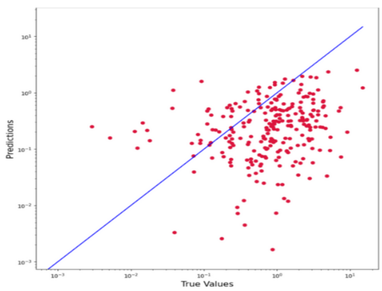


Fig: 12. Plot using Linear Kernel

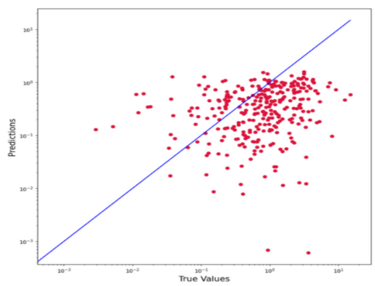


Fig: 13. Plot using Radial basis function Kernel

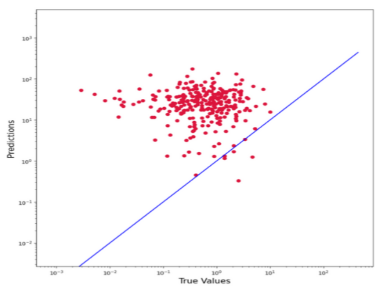


Fig: 14. Plot using Sigmoid Kernel

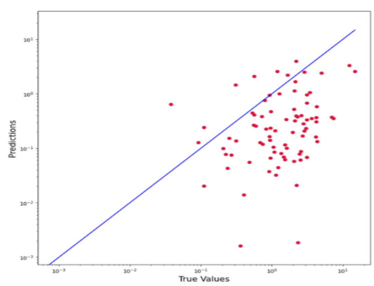


Fig: 15. Plot using Polynomial Kernel

Kernal	R Value	MSE
Linear	0.5509	0.8793
Rbf	0.5249	0.8683
Sigmoid	-0.2488	6624.1514
Polynomial	0.5165	0.9806

Fig: 16 . Comparision between different Kernels of SVM

Validation of the models: The constructed models were validated using both training and testing datasets. R and MSE were used to validate and compare the predictive capability of the models.

Chapter 3

Conclusion

In the present study, two advanced artificial intelligence (AI) models, ANN and SVM, were used to predict daily rainfall in the Kamrup district of Assam. The model used maximum temperature, minimum temperature, wind speed, precipitation, relative humidity, and solar radiation as inputs and daily rainfall as output. Validation of the models was done using R and MSE. The performance of this model in forecasting daily rainfall is shown in the figure above (i.e. R(SVM) = 0.5509, MSE(SVM) = 0.8793, R(ANN) = 0.61277, MSE(ANN) = 52.2079).In comparison, the performance of SVM for daily rainfall is better because this method yielded slightly lower values of R than those obtained from ANN but with a much lower value of MSE. AI-based analysis will help to make daily rainfall forecasts quick and more accurate. Rainfall forecasting is of great importance in the context of water availability and management, human livelihoods, and climatic conditions. Issues of incorrect or incomplete measurements may be encountered because the rainfall measurement is affected by spatial and local change and variability characteristics. This paper presented a review of various approaches to rainfall forecasting and identified various methods that can be used to predict rainfall.

Chapter 4

Future Work

The future work of the project would be focusing on the Spatial Interpolation Methods for Estimating Rainfall Distribution in Kamrup, Assam. Due to climate change, rainfall patterns are changing worldwide and therefore the study of hydrological processes has become important in the management of water resources. Rainfall distribution plays an important role in understanding hydrological processes. In environmental watershed planning, many geographic information system (GIS) models require rainfall as an input, either discrete or continuous formate. More often spatially explicit rainfall data are required by the hydrologic models. However, most meteorological data (especially precipitation) are collected using rain gauges, thus reflecting biological data. Spatially explicit data are often obtained by interpolation methods. The representation and accuracy of rainfall data in the digital world are shaped by the spatial distribution of weather stations and interpolation methods that may or may not reflect reality. This proposal, therefore, compares and evaluates the performances of well-established interpolation techniques that can be used to estimate daily rainfall in Assam.

Bibliography

[1] Emilcy Hern´andez1, Victor Sanchez-Anguix2, Vicente Julian3, Javier Palanca3, and N´estor Duque4. Rainfall prediction: A deep learning approach. 2016.

[2] Dr. Rajeev Kumar and Mohammed Hatim Ferhan Siddiqui. Interpreting the nature of rainfall with ai and big data models. 2020.

[3] Mislana, Haviluddinb, Sigit Hardwinartoc, Sumaryonod, and Marlon Aipassa. Rainfall monthly prediction based on artificial neural network: A case study in tenggarong station, east kalimantan - indonesia. 237:142–151, 2015.

[4] N.A.Charaniya and S.V.Dudul. Design of neural network models for daily rainfall prediction. 61, 2013.

[5] Binh Thai Pham, Lu Minh Leb, Tien-Thinh Lec, Kien-Trinh Thi Buid, Vuong Minh Leb, Hai-Bang Lya, and Indra Prakash. Development of advanced artificial intelligence models for daily rainfall prediction. 237, 2020.