Methodology

Hari Prasad
College of Engineering Karunagappally
Kollam, Kerala

1. Data Analysis

The Data given had 20,000 samples, each representing the amount of rainfall per day, from 13-07-1966 to 14-04-2021. A simple line plot and histogram plot revealed a trend in the data set. In order to make the data more Gaussian and suitable for regression, Box-Cox Power transform was applied on the data set. The fig.2 shows the plots before and after applying the transformation.

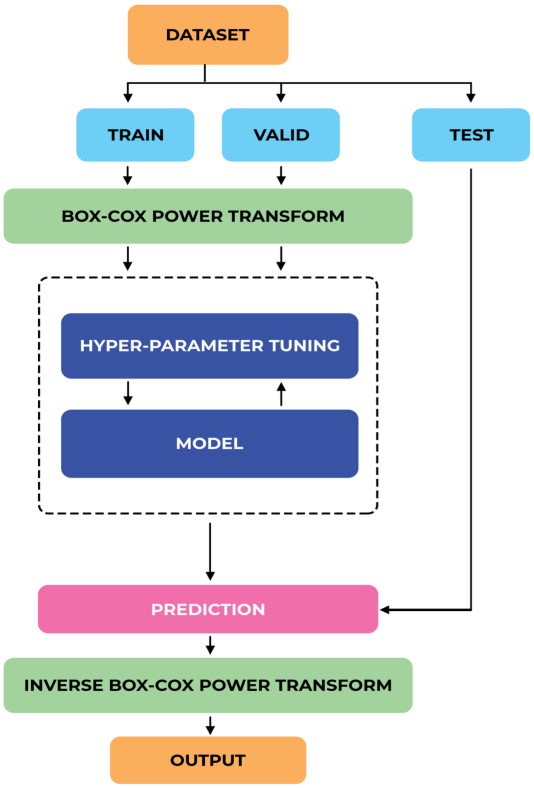

Fig. 1 Methodology

Model	Nodes / Filters	Activation
MLP	140	elu
CNN	126	tanh
LSTM	90	relu
Bi-LSTM	180	elu

Table.1 Hyper parameter tuning results

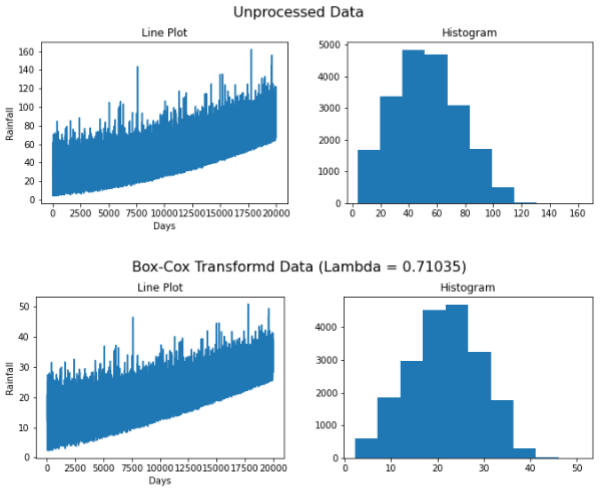

Fig. 2 Processed v/s unprocessed data

2. Dataset preparation

A split of 60% and 40% of 19,940 samples were done for the Training and Validation sets, respectively. Comparisons were made in a step of 30 samples, to reduce the training time. Training and prediction on a step of 365 was done on the best model. The sample split and sizes are given in Table.2.

Instance	Train	Validation	Test
Comparison	11964	7967	60
Final Model	11689	7580	731

Table.2 Sample sizes

3. Model Selection and Hyper Parameter tuning

A comparative study of 3 types of networks were done, including MLP, CNN and 2 variants of LSTM (Bi-directional and Simple). In order to choose the right hyper parameters for each model, tuning was performed with the keras-tuner library. 2 sample spaces were used, one with the number of nodes/filters ranging from 50 - 200. Another parameter was the activation function, including relu, elu, tanh and sigmoid. The results are shown in Table.1.

4. Training and Metrics

The models were trained for 100 epochs each, with a step size of 30. The optimizer used was Adam with a learning rate of 0.001. Since the problem was of regression, Mean squared error (MSE) and R2 Score were considered in order to determine the best models and evaluate the performance. The loss function used was also MSE.

5. Post Processing and Final Model selection

After the training and prediction, the predicted data underwent inverse Normalization and Inverse Box-Cox transform, to get the predicted data in its actual form. Based on the comparison results, the right model was chosen for the yearly prediction, with a step size of 365. A comparison of process data, unprocessed data, different data splits for training, validation and testing were done with various tweaking in the hidden layers, to find the best model. Also with unprocessed data, training showed overfitting.