Temperature forecasting using air pollutants' concentration

Abstract— **Forecasting temperature is an important aspect of everyone's life in today's world. It is a key factor in the climate impact studies on several sectors such as industries, agriculture, etc. Accurate temperature forecasts would help us to protect ourselves from global warming. In this paper, we have used the concentration of air pollutants such as NOx, NO2, CO, SO2, PM2.5, PM10, etc. of the last 50 hours, as an input to our model, further, the next 10 hours' temperature is obtained in the output. This paper aims to study different machine learning techniques used in temperature forecasts. Here, different models such as SRNN, LSTM, GRU are used and further Keras hyper-tuner tool was employed to obtain the optimum number of units and best activation function for each hidden layer in the above models. Using those tuned models, we predicted the temperature for the next 10 hours. Then the paper obtained different error functions - MSE, MAE, and MAPE for training and testing sets for each model. For the SRNN model, testing and training MAPE is 14.35% and 18.41%. Similarly, for the LSTM model, it is 7.64% and 2.40% and for GRU it is 7.08% and 2.89%.**

Keywords—AQI (Air quality index), GRU (Gated recurrent units), LSTM (long short-term memory, Keras-tuner, MAE (mean absolute error), MSE (mean square error), MAPE (mean absolute percentage error), SRNN (simple recurrent neural networks), SVM (support vector machine), neural networks, pandas, sklearn, ReLu (rectified linear unit),

I. INTRODUCTION

As human beings, we must be very careful about the environment and the ecosystem. As we have progressed with information, science, and technology, we have exploited nature for our own benefits and needs. We sometimes exploit nature without even considering the consequences. This results in the emergence of many problems including pollution, climate change, global warming, etc. which in turn affect not only us but every living organism on the earth. For instance, due to changes in climatic conditions, seasonal rainfall is getting affected. Also due to global warming, the temperature of the earth is rising faster than ever at an alarming rate. As a result, the polar region of the ice is melting and hence the increase in ocean level is being reported. Due to the changing sea level, ocean circulation is also getting disturbed which has made the survival of aquatic life in their habitat, difficult. In the last 3 decades, due to rising temperatures an increase in heat waves cases is also been observed.

The objective of this paper is, to predict the temperature of the nearby future using machine learning techniques. Air quality index (AQI) is the measure of pollutant level in the atmosphere, it is particularly very important to measure the AQI, as it tells how the air we breathe, affects our body. It is observed that pollutants concentration such as NOx, NO2, CO, SO2, etc., and temperature have some direct relationship i.e., higher the pollutants concentrations, higher is the AQI, and hence higher the temperature of the polluted region [3]. The major pollutants available in the air are carbon monoxide (CO), nitrous oxide (NO), particulate matter of size 2.5 µm (PM 2.5), particulate matter of size 10 µm (PM 10), (O3) ozone, and (SO2) sulphur dioxide, etc. It is also generally observed that PM10 and PM2.5 content is higher in the air and hence they dictate the overall AQI. In this paper presented, the future temperature is predicted by using the past pollutants concentration and temperature data. In this paper, different algorithms like LSTM, SRNN, and GRU are compared. This analysis was done only for the Indian union territory of Delhi keeping in mind its rising levels of pollutants in recent years.

In section II, the past works are done in this and the allied topics are discussed, further, in section III, information about data sources is mentioned, in section IV, the data pre-processing done is discussed. In further section results and conclusion of this study is mentioned.

II. LITERATURE REVIEW

Prediction of temperature using pollutants concentrations is a challenging task for researchers and many researchers have shown their interest in this field in recent years. There are many algorithms so far that have been used for forecasting of temperatures for e.g. ARIMA (auto regressive integrated moving average models), ANN (artificial neural networks), SRNN (Simple recurrent neural networks), SVM, etc. Y. Radhika et al. [6] have used SVM (Support Vector Machine) for the prediction of the atmospheric temperature at a particular location. But the prediction of temperature by using SVM at a location need proper selection of the parameters such as the compositions of the air, relative humidity, pollutant level, etc.

Nithyashree et al. [2] have used LSTM to predict just the future AQI and has trained the LSTM network on time series data of AQI and individual pollutant concentrations. The model accuracy for the training set was about 0.004 %.

Inyoung Park et al. [5] have used LSTM network to predict the make accurate forecasts. However, the missing values in the data poses many problems in accuracy of the algorithm so they have used the trained model for refinement of data and further they have again trained the algorithm and compare various other model. Finally, it was obtained that RMSE for the proposed model was 0.79 for 24 hours' prediction of temperature.

In the paper presented, first the pre-processing of data was done so that any missing or wrong data can be corrected. Then heat map of correlation matrix was plotted using the seaborn library, so that the correlation between temperatures and pollutants' concentration can be obtained. Further, using the input data of pollutants concentrations and temperature of last 50 hours, and output data of next 10 hours' temperature, best set of hyper-parameters were obtained using Keras Hyper tuning. The tuning was done for each model i.e. GRU, SRNN, LSTM and optimal set of hyper-parameters such as number of hidden layers, learning rate, number of units in each hidden layer, activation functions etc. were obtained.

III. Data Source

The entire air pollutant composition and temperature dataset of this is obtained from the official site of CPCB (Central pollution control board, India). The data is contributed by the following agencies: CPCB, DPCC, HSPCB, UPPCB and SAFAR- India. This dataset contains the hourly composition of the pollutant such as PM2.5, PM10, NO, NO2, NOx, CO, SO2 and temperature. In this paper presented, we have used the dataset of the Dwarka Sector 8, one of the AQI monitoring stations in Delhi, from the 5th May 2020 to 6th June 2021. Its data is maintained by DPCC [7]. The dataset contained 9543 rows and 10 columns where the rows contain hourly data of each pollutant and temperature.

IV. Exploratory Data Analysis

The dataset was extracted from the official website of CPCB [7] in xlsx format and was read using pandas data frame. The dataset contained some missing values which were filled by using Simple Imputer, a tool in sklearn library. It was checked whether the data contained NaN or null values if it was so, then the missing values were replaced by the mean of the respective column. Further, a correlation matrix was obtained and the density plot of temperature is obtained.

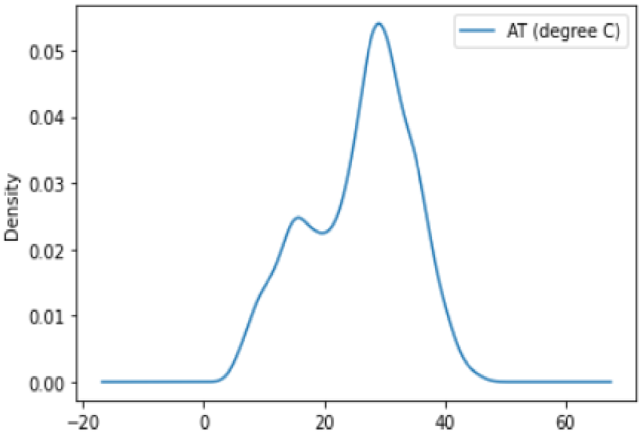

Fig. 1 Density plot of temperatures

From the density plot of temperature shown in figure 1 it can be seen that the density values for temperatures tending to 0 are 0. This information will help in deciding the metrics used for finding the accuracy of the machine learning models.

From the correlation matrix shown in figure 2, it can be seen which parameters or pollutants' concentration has a high relation with the temperature variable. So only PM2.5, PM10, NO2, NO, NOx and CO are observed to have high correlations with AT (Absolute temperature), while SO2's correlation with AT was -0.08, which was very low, so the SO2 concentration data is not included in the analysis. For all the models trained, the data was split into 2 portions i.e., training set, containing 80% of the total data and testing set containing 20% of the total data. It was done using sklearn library.

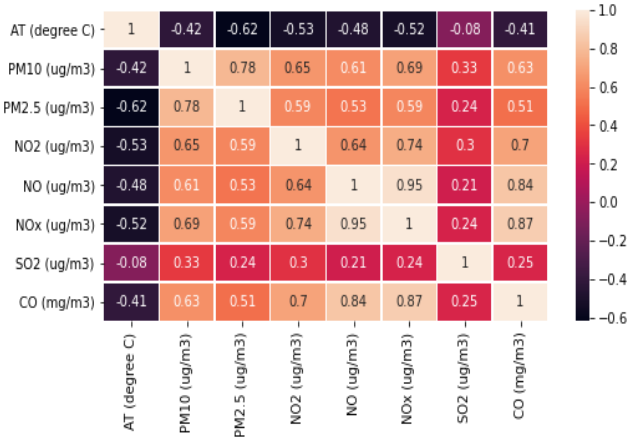

Fig. 2 Correlation matrix

V. Machine Learning Model Architecture

The machine learning models were first trained on time series data of pollutant concentration and temperature. So, the input data consisted of the last 50 hours' concentrations of PM2.5, PM10, NO2, NO, NOx and CO, and temperature using this past 50 hours of data the next 10 hours' data of temperature was predicted. Thus, the shape of the input vector was (None,50,7). Similarly, the shape of the output was obtained to be (None,10). Here, Mean Squared Error (MSE) was taken as the loss function.

$$MSE = \frac{\sum(y_i - \bar{y})^2}{N}$$

To measure the performance of the network MAE, MSE and MAPE were used. It can be seen from figure 1 that for temperatures tending to 0 have density values as 0, indicating that there are no such temperatures in the dataset whose values are tending to 0 ^{o}C, so in that case high values of MAPE will not be observed. So, MAPE can be used here as a metric.

$$MAE = \frac{\sum |y_i - \bar{y}|}{N}$$

$$MAPE(\%) = \frac{100}{N}\frac{\sum |y_i - \bar{y}|}{y_i}$$

Here y_i represents the actual value, $\bar{y}$ represents the predicted value, while N is the total number of fitted points.
In the network architecture, the optimized set of hyperparameters i.e., the activation functions, number of layers, number of units in each layer, learning rate etc. were found out using Keras tuner, a library used for hyper tuning. The MSE was minimized using the Adam optimizer and choice of the best learning rate from 0.01, 0.001, 0.0001 was done using the same tuner. Further, LSTM, SRNN and GRU models were trained using Keras hypermodel builder on the dataset with the best activation functions, the optimum number of units in each hidden layer, the optimum number of hidden layers and best learning rate obtained from the Keras tuner.
Each model will consist an input layer consisting of 7 SRNN/ LSTM/ GRU units depending on the model type. The output layer will consist of 10 dense units.

VI. Results

In Table I, it can be seen that the optimized RNN model for temperature prediction from the concentration dataset consists of 304 RNN units in the first hidden layer and 368 RNN units in the second. It is having a validation MSE of 0.0063 and a validation MAE of 0.0585. Thus the model is of total 4 layers, consisting of 2 hidden layers.
Also, in Table IV, the training and testing MAPE and learning rates for different models are given. The optimum learning rate for the SRNN model was found to be 0.001. The testing MAPE was obtained to be 14.35 % while training MAPE to be 18.41 %. In Table V the training MSE and MAE are given for each model.

Similarly, from Table II, the optimized LSTM model consists of 48 LSTM units in the first hidden layer, 112 LSTM units in the second layer, and 208 units in the last hidden layer with validation MSE of 0.0063. The optimum learning rate was found to be 0.01. The MAPE of 2.40 % was obtained for the training set while for the testing set the MAPE was 7.63 %.
From Table III, the optimized GRU model consists of 16 GRU units in the first layer, 240 units in the second hidden layer, and 432 units in the last hidden layer and it has a validation MSE of 0.0056 with relu activation function in each unit. The optimum learning rate was found to be 0.001. The MAPE for the training set was 2.89 % while for the testing set it was found to be 7.08 %.

Thus for both GRU and LSTM the model is found to have 5 layers consisting of 3 hidden layers.
These optimal set of hyper-parameters were obtained by 11 trials of hyper-parameter tuning.

All the models were trained on 50 epochs. The models were tested by obtaining the plots of temperature vs time, for the input data from 16:00, 5th May 2020 to 17:00, 7th May 2020 and output for the next 10 hours i.e., from 18:00, 7th May to 03:00, 8th May 2020.

TABLE I: SRNN MODEL AFTER 11 TRIALS OF TRAINING IN KERAS HYPER-TUNER

Simple RNN model				
Hidden Layer	No. of Units	Activation Function	Val-Loss -MSE	Val-MAE
1	304	Relu	0.0063	0.0585
2	368	Relu		

TABLE II: LSTM MODEL AFTER 11 TRIALS OF TRAINING IN KERAS HYPER-TUNER

LSTM model				
Hidden Layer	No. of Units	Activation Function	Val-Loss -MSE	Val-MAE
1	48	Tanh	0.0063	0.0579
2	112	Tanh		
3	208	Tanh		

TABLE III: GRU MODEL AFTER 11 TRIALS OF TRAINING IN KERAS HYPER-TUNER

GRU model				
Hidden Layer	No. of Units	Activation Function	Val-Loss -MSE	Val-MAE
1	16	Relu	0.0056	0.0528
2	240	Relu		
3	432	Relu		

TABLE IV: TRAINING AND TESTING MAPE OF THE OPTIMIZED MODELS

Model	MAPE training	MAPE testing	Learning rate
SRNN	18.41	14.35	0.001
LSTM	2.40	7.63	0.01
GRU	2.89	7.08	0.001

TABLE V: TRAINING MSE AND MAE OF THE OPTIMIZED MODELS

Model	Training MSE	Training MAE
SRNN	0.000529	0.0165
LSTM	0.0005	0.0159
GRU	0.00087	0.0209

In fig 3, the next 10 hours' temperature was predicted using last 50 hours' pollutant concentration data using SRNN model obtained from the Keras hyper tuner.

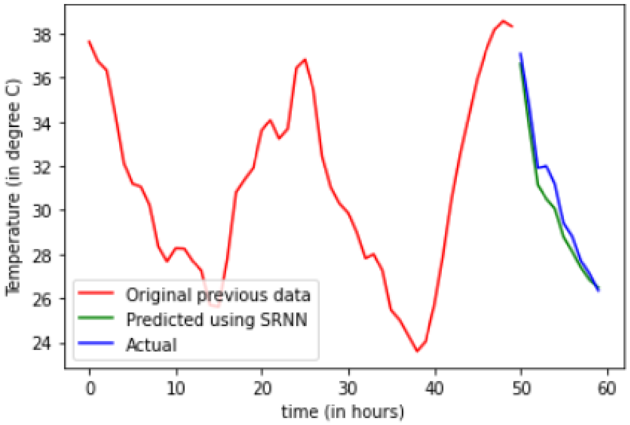

Fig 3: Temperature data predicted from SRNN model in table 1 and its comparison with original data (in blue)

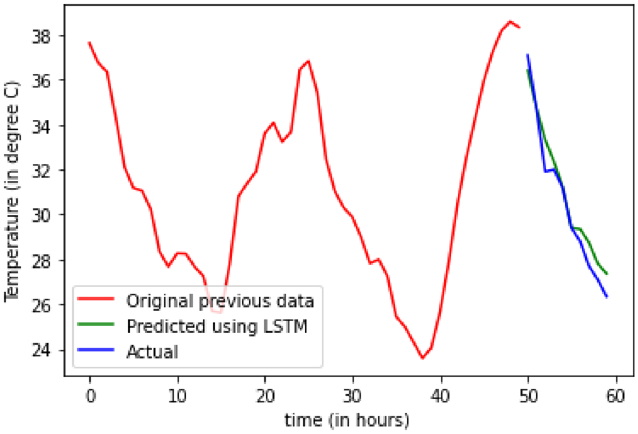

Fig. 4 Temperature data predicted from LSTM model in table 2 and its comparison with original data (in blue)

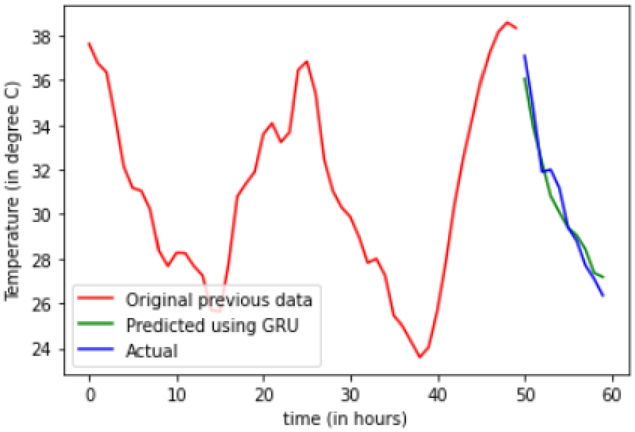

Fig 5: Temperature data predicted from GRU model in table 3 and its comparison with original data (in blue)

Similarly, in figure 4, the next 10 hours' temperature was predicted using the last 50 hours' pollutant concentration data using the LSTM model obtained from the Keras hyper tuner.

And in figure 5, the next 10 hours' temperature was predicted using the last 50 hours' pollutant concentration data using the GRU model obtained from the Keras hyper tuner.

It is generally observed that accurate forecasts have MAPE less than 20 %. From table IV it can be seen that both training and testing MAPE for each model was less than 20 %, hinting that the forecast made are good and accurate.

VII. Conclusion

In this paper, the temperature's dependency on air quality was analyzed. The next 10 hour's temperature was predicted by giving input of the past 50 hour's air pollutant concentration data. For this, the optimized network architecture was found by hyper-parameter tuning for each type of model i.e., GRU, LSTM and SRNN. Then we obtained different error functions – MSE, MAE, MAPE for testing and training set for each model. It was found that the testing and training MAPE were, 14.35% and 18.41% for SRNN, 7.63% and 2.40% for LSTM, and 7.08% and 2.89% for GRU.

The optimum structure of each type of model was obtained, for SRNN it was a 2 hidden layered network, consisting of 304 and 368 units in respective layer. For LSTM, it consisted of 48, 112, and 208 LSTM units in each hidden layer. The GRU model consisted of 16, 240, and 432 units in respective hidden layers.

VIII. References

[1] Y.S. Park, S. Lek, "Artificial Neural Network: Multi Perceptron for Ecological Modeling", ch.7, Development in Environmental Modelling, Sven Erik Jorgensen, Elsevier, vol. 28, pp. 123-140, 2016.

[2] Nithyashree K R, S Bhumika, Sahana R, Ranjitha V, "Air Quality Index Prediction using LSTM", International research journal of engineering and technology, Volume 07, Issue 6, June 2020, Pages 4848-4851.

[3] Temesegan Walelign Ayele, Rutvik Mehta, "Air pollution monitoring and prediction using IoT", IEEE Conference 2018.

[4] Peijiang Zhao, Koji Zettsu," Convolution Recurrent Neural Network Based Dynamic Transboundary Air Pollution Prediction",IEEE Conference 2019.

[5] Inyoung Park, Hyun Soo Kim et. al., "Temperature Prediction Using the Missing Data Refinement Model Based on a Long Short-Term Memory Neural Network", *Atmosphere,* vol.10(11), 2019, doi: https://doi.org/10.3390/atmos10110718

[6] Y. Radhika & M. Shashi, "Atmosphere Temperature Prediction using Support Vector Machines", *International Journal of Computer Theory and Engineering,* vol. 1, no.1, pp. 1793-8201, April 2009

[7] CPCB (Center for pollution control board, India), DPCC (Delhi Pollution Control Committee),HSPCB (Haryana State Pollution Control Board), UPPCB (Uttar Pradesh Pollution Control Board) & SAFAR(System of Air Quality and Weather Forecasting and Research), "Central Control Room for Air Quality management – All India", https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data