

THE WAY WE APPROCH  
ON  
HOW WE ARE CLASSIFYING OUR  
DATA  
FOR  
HEATWAVES AND AQI

# INDEX

<b>1.</b>	<b>A RNN with LSTM and GRU based on the work done from Research paper</b>
<b>2.</b>	<b>GRU</b>
<b>3.</b>	<b>LSTM</b>
<b>4.</b>	<b>A neural network based on Neural Prophet</b>
<b>5.</b>	<b>Heat Wave Forecasting</b>
<b>6.</b>	<b>DATASET</b>
<b>7.</b>	<b>Usage</b>
<b>8.</b>	<b>AQI</b>
<b>9.</b>	<b>References</b>

We intend to use the following two methods for both questions—heat wave prediction and air quality index.

## A RNN with LSTM and GRU based on the work done from Research paper.

The two most sophisticated time series predictors, LSTM and GRU, were combined by the authors of the cited paper to produce results that were superior to those of the methods used separately. Modern algorithms for predicting time series data, like ours, include LSTM and GRU. In order to forecast time series data, we want to combine the strengths of LSTM and GRU.

### **LSTM:**

Recurrent neural networks (RNNs) of a LSTM (Long Short-Term Memory) variety are frequently employed for processing sequential data.

### **WHY WE USE LSTM:**

**1. Handling long-term dependencies:** Because LSTMs are made to record long-term dependencies in consecutive data, they are ideal for time series data, where patterns and trends can endure for a very long time.

**2. Handling disappearing gradients:** When handling lengthy sequences, conventional RNNs frequently experience the vanishing gradient problem. LSTMs are prepared to manage this problem. As a result, LSTMs work well with time series data, where patterns and trends can last for long time.

**3. Flexibility:** LSTMs can manage a broader variety of complex relationships in the data because they have a more complex structure than GRUs. Due to the complexity of the connections between variables and trends in the data, LSTMs are well adapted for time series data.

**4. performance:** For a variety of time series prediction tasks, LSTMs have been proven to outperform other types of RNNs, including GRUs, making them a popular option for many practitioners.

**5. Widely utilised:** LSTMs have been used and studied extensively for more than 20 years, and a substantial body of research and best practises have been formed around their use, making them an established and well-understood method for handling time series data.

## GRU:

### WHY WE USE GRUs:

**1. Handling long-term dependencies:** GRUs are made to capture long-term dependencies in sequential data, much as LSTM. This makes them a strong fit for time series data, where patterns and trends can endure for a very long time.

**2. More effective computation:** Compared to LSTMs, GRUs are more effective at processing big time series datasets, which can be computationally demanding.

**3. Simple to train:** GRUs are simpler to train and less prone to overfitting than LSTMs since they contain fewer parameters.

**4. Reliable performance:** GRUs have been demonstrated to be effective in a variety of time series prediction applications, such as predicting stock prices, sales, and energy usage. As a result, many practitioners favour using GRUs.

**5. Interpretability:** When working with time series data, where the interactions between variables and patterns in the data might be complex, GRUs' simpler structure than LSTMs can make them easier to analyse and comprehend.

The model will be created in Python using TensorFlow, and any necessary adjustments will be made to hyperparameters like the number of neurons and dense layers.

## **A neural network based on Neural Prophet**

### **WHY WE USE:**

Facebook Prophet was overtaken by NeuralProphet, which became the de facto benchmark for comprehensible, expandable, and user-friendly forecasting frameworks. Explainable forecasting is still a difficult task for commercial and operational decision making given the abundance of time series data. To fill the gap between interpretable classical approaches and scalable deep learning models, hybrid solutions are required.

A hybrid forecasting framework built on PyTorch and taught using common deep learning techniques is called NeuralProphet. Auto-regression and covariate modules, which can be set up as conventional linear regression or as neural networks, introduce local context. Other than that, NeuralProphet maintains Prophet's design ethos and offers the same fundamental model building blocks.

Our findings show that NeuralProphet, when applied to a set of generated time series, generates interpretable forecast components of equal or better quality than Prophet. For a variety of real-world datasets, NeuralProphet outperforms Prophet. NeuralProphet increases forecast accuracy for short- to medium-term forecasts by 55 to 92 percent.

## **Heat Wave Forecasting:**

### **Feature Extraction**

Studying the research on the factors of heat waves highlights a number of variables, as was mentioned previously in the introduction. Some of these factors, like humidity, precipitation, and windspeed, are well known locally. However, other, less well known locally, factors, like surface soil moisture and a combination of cloud cover and daylight hours, show promising results when used as components of larger statistical models, like the proposed SEMB model (Soil Energy and Moisture Budget). These elements will be the stimulus for our potent neural networks.

Since the Earth is a continuous system, phenomena occurring in one area, particularly the seas, have a major influence on a station's weather. Thus, as shown by Rohini P. et al. [4], variables like the ONI (Oceanographic Nino Index) play a significant part in deciding the climactic circumstances of a station in any given year. MJO (Madden-Julian Oscillation) amplitude and SST (Surface Sea Temperatures) from the world's main seas (Indian, Pacific, and Atlantic) at different pressure levels will also be taken into consideration, as shown in the work of Guigma K.H. et al. [5].

We intend to use a variety of methods, including but not limited to data scraping with bots, querying open APIs, and reverse engineering open government APIs, to gather data every day for the features listed below.

## DATASET:

FEATURES	DATASET
Cumulative Rainfall	Telangana Daily Weather Data 2018 - 2022([Telangana Weather Data 2022   Telangana Open Data Portal] ( <a href="https://data.telangana.gov.in/dataset/telangana-weather-data2022">https://data.telangana.gov.in/dataset/telangana-weather-data2022</a> ))
Temp Max	
Humidity Max	
Humidity Min	
Wind Speed Max	
Wind Speed Min	
Cloud Cover	National Information System for Climate and Environmental Studies (Satellite Data from various satellites such as CARTOSAT and INSAT)
Surface soil moisture	
Daylight Hours (Number of hours between sunrise and sunset)	Using python library suntime
ONI (Oceanographic Nino Index)	NOAA Oceanic Nino Index (ONI) [Climate Prediction Center - ONI ( <a href="https://noaa.gov">noaa.gov</a> )]
MJO amplitude (RMM1 + RMM2)	NOAA MJO Index Data 1971-2022 [NOAA][ <a href="https://www.psl.noaa.gov/mjo/mjoindex/">https://www.psl.noaa.gov/mjo/mjoindex/</a> ]
SST Tropical Pacific at various pressure levels	[International Comprehensive Ocean-Atmosphere Data Set (ICOADS)] [( <a href="https://icoads.noaa.gov/products.html">https://icoads.noaa.gov/products.html</a> )]
SST Tropical Atlantic at various pressure levels	
SST Indian Ocean at various pressure levels	

The data will be cleansed after it has been gathered. Encoding is not necessary because we have not examined categorical data; however, each datapoint will be scaled using standard scaling using sklearn's Standard Scaler.

## Usage

For each of the five areas, we intend to use two different models and the aforementioned methods.

The best model will be picked for each location after the findings are compared. The comparative measures will be the common MAE, MAPE, and R2 score metrics. Data from the first four years will serve as training data, and data from the final year (i.e. 2022) will serve as assessment data.

The forecast from the algorithms will then be saved as a CSV or Parquet file for later use

## AQI:

Different models for time series data for each of the pollutants For time frame 2018 - 2022

- PM 2.5
- PM 10
- NO2
- SO2
- CO
- O3

Then, as per India AQI rules, the highest scaled pollutant will determine the overall AQI.

## Features Extracted:

From literature review of the problem of Air Quality prediction, we decided on the factors for which data should be collected. Again, some were prominent and obvious factors affecting the air quality such as precipitation, number of vehicles and forest cover. This is also backed by relevant studies such as work done by Dongsheng Zhen et.al. [7]. However, there were surprising results in the literature such as the strong positive correlation of Per Capita Income [6] with AQI, i.e., the more the average income of a place, worse is the air quality. According to the work done in the same publication, Construction has been determined as a leading cause for worsening AQI. Thus, SGDP attributed to construction is also chosen as a feature [6]. Another such revelation was that Urbanization Level is not a significant contributor to the Air Quality of a place. Daily temperature is not considered as it is not a causal factor, rather it is an effect of the Air Quality.

## Usage

Similar to the Heat Wave models, we create a model for each pollutant for each of the 5 districts. The predicted concentration for each pollutant is then scaled to the AQI indices then choosing the worst pollutant, which is considered to be the prevailing AQI.

## Libraries to be used:

- PyTorch
- TensorFlow
- Suntime
- Python Base packages
- Matplotlib
- Seaborn
- Sklearn
- Flask
- Vue3

## References:

1. Hossain, E., Shariff, M.A.U., Hossain, M.S., Andersson, K. (2021). A Novel Deep Learning Approach to Predict Air Quality Index. In: Kaiser, M.S., Bandyopadhyay, A., Mahmud, M., Ray, K. (eds) Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Advances in Intelligent Systems and Computing, vol 1309. Springer, Singapore. [https://doi.org/10.1007/978-981-33-4673-4\\_29](https://doi.org/10.1007/978-981-33-4673-4_29)
2. NeuralProphet: Explainable Forecasting at Scale (arXiv:2111.15397)
3. Zeppetello, Lucas R. Vargas, David S. Battisti, and Marcia B. Baker. "The Physics of Heat Waves: What Causes Extremely High Summertime Temperatures?", Journal of Climate 35, 7 (2022): 2231-2251, accessed Feb 2, 2023, <https://doi.org/10.1175/JCLI-D21-0236.1>
4. Rohini, P., Rajeev an, M. & Srivastava, A. On the Variability and Increasing Trends of Heat Waves over India. Sci Rep 6, 26153 (2016). <https://doi.org/10.1038/srep26153>
5. Guigma, K.H., Guichard, F., Todd, M. et al. Atmospheric tropical modes are important drivers of Sahelian springtime heatwaves. Clim Dyn 56, 1967–1987 (2021). <https://doi.org/10.1007/s00382-020-05569-9>
6. Jun Ma, Yuexiong Ding, Jack C.P. Cheng, Feifeng Jiang, Yi Tan, Vincent J.L. Gan, Zhiwei Wan, Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, Volume 244, 2020, 118955, ISSN 0959-6526, <https://doi.org/10.1016/j.jclepro.2019.118955>.
7. Dongsheng Zhan, Mei-Po Kwan, Wenzhong Zhang, Xia fen Yu, Bin Meng, Qianqian Liu, The driving factors of air quality index in China, Journal of Cleaner Production (2018), doi: 10.1016/j.jclepro.2018.06.108 <https://proceedings.neurips.cc/paper/2017/file/a6db4ed04f1621a119799fd3d7545d3d-Paper.pdf>