Multivariate Machine Learning Approaches for Dynamic Prediction of Air Quality and Estimating Heatwave Occurrence

by Abhinandan Roul

Submission date: 08-Jun-2023 12:05PM (UTC+0530)
Submission ID: 2111581499
File name: SDP_Paper_June_8th_v1.7.docx (1.17M)
Word count: 4987
Character count: 28908

Multivariate Machine Learning Approaches for Dynamic Prediction of Air Quality and Estimating Heatwave Occurrence

Abhinandan Roul[1], Shubhaprasad Padhy[2], Sambit Kumar Sahoo[3], Ayush Pattanayak[4] Manoranjan Parhi[5]

[1,2,3,4,5]Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India

Abstract. Pollutants in the air lead to degradation in the air quality which leads to global warming and unpredictable heatwave occurrence. Air pollution has been a leading cause of respiratory diseases and premature deaths. To forecast AQI and temperature, a multivariate approach based on AR-Net and Temporal Fusion Transformer (TFT) is proposed. A combination of atmospheric and meteorological variables as input features to train the model, including temperature, humidity, wind speed, and pollutant concentrations (PM2.5, PM10, SO2, NO2). Open-source dataset of Air quality in India, is used to train the models and evaluate the experiments. With an MAPE of 7% for AQI and MAPE of 4%for heatwave prediction, it is concluded that the results demonstrate appreciable accuracy in predicting AQI and the occurrence of heatwaves.

Keywords: Air Quality, Time Series Forecasting, Deep Learning, Climate Change, Heatwaves

1 Introduction

Air quality is a critical factor that affects public health and the environment. Accurate prediction of air quality is essential in providing timely warnings and supporting decision-making. In addition, the occurrence of heat waves can exacerbate the impact of poor air quality on human health [1]. Traditional methods of air quality prediction rely on statistical models that use data from air quality monitoring stations. However, these models are limited by the availability and quality of data, as well as the complexity of the underlying processes that affect air quality and heatwave occurrence.

The use of multivariate machine learning approaches for dynamic prediction of air quality and estimating heatwave occurrence has significant implications for public health, urban planning, and environmental policy [2]. Accurate prediction of air quality and temperature can support decision-making related to health advisories, transportation planning, and energy consumption, while the estimation of heatwave occurrence can support emergency response planning and heat mitigation strategies.

In recent years, there has been a growing interest in the use of machine learning techniques to improve air quality prediction and estimate heatwave occurrence. This paper presents a study on the application of multivariate machine learning approaches, specifically Neural Prophet and Temporal Fusion Transformer, to predict air quality dynamics and estimate heatwave occurrence in a metropolitan area.

The process of predicting air quality index (AQI) and temperature involves several steps, including data collection, pre-processing, and feature extraction. First, data such as air quality measurements, meteorological data, and demographic data are collected from multiple sources, such as air quality monitoring stations and weather stations. The data is then pre-processed to remove outliers and missing values and to normalize the data for consistency [3]. Feature extraction involves selecting relevant variables that affect air quality and temperature, such as temperature, humidity, wind speed, and pollutant concentrations. These variables are used to train a machine learning model, such as Neural Prophet or Temporal Fusion Transformer, to predict AQI and temperature over time. The model is evaluated using metrics such as mean absolute error and root mean square error to assess its accuracy and effectiveness. Overall, the multivariate machine learning approach to AQI and temperature prediction involves a complex process of data collection, pre-processing, and feature extraction, followed by model training and evaluation to produce accurate and reliable predictions.

The study utilizes a combination of air quality, meteorological, and demographic data to develop a model that can accurately predict air quality levels and estimate the occurrence of heat waves. The results of the study demonstrate the potential of multivariate machine learning approaches in improving air quality prediction and estimating heatwave occurrence, with the ability to capture complex patterns and relationships in data.

This paper contributes to the field of air quality prediction and heatwave estimation by introducing a novel approach that utilizes multivariate machine learning techniques. The results of the study have implications for urban planning, public health, and environmental policy, as they can provide insights into the factors that affect air quality and heatwave occurrence and support decision-making related to these issues [4].

1.1 Research Gap and Research Questions

A variety of research has been carried out using the multivariate machine learning approaches for dynamic prediction of air quality and heat wave occurrence that had significant implications for public health , urban planning and environmental policy decision-making related to health advisories. It should be noted that these studies are meant for the transportation planning, energy consumption, while the estimation of heatwave occurrence in support emergency response planning and heat mitigation strategies.
The following research questions (RQs) were taken into account for this investigation:

- RQ1: What is the main goal behind air quality index prediction and how it is related to heat wave mitigation?
- RQ2: What are the different factor that affect air quality and heatwave occurrence and support decision-making related to these issues?
- RQ3: What are the benefits of measuring the air quality index in time series?
- RQ4: What are the preventive measures against the change in AQI and heat wave?
- RQ5: What are the major paybacks faced during the prediction of particular time series?

1.2 Motivations and Objectives

The motivation behind using multivariate machine learning approaches for dynamic prediction of air quality and estimating heatwave occurrence is to improve the accuracy and timeliness of air quality predictions and heatwave occurrence estimates. This is important because poor air quality and heatwaves can have significant negative impacts on human health, the environment, and the economy. Multivariate machine learning approaches can use multiple variables and data sources to build models that can better capture the complex relationships between air quality, weather patterns, and other environmental factors. By incorporating real-time data, these models can provide more accurate and timely predictions, enabling authorities to take action to mitigate the impact of poor air quality and heatwaves.

The objective of using these approaches is to develop models that can accurately predict air quality and estimate the occurrence of heatwaves, with the aim of improving public health outcomes and reducing the economic and environmental costs associated with poor air quality and heatwaves. These models can also help policymakers to develop more effective strategies for managing air quality and mitigating the impacts of heatwaves.

1.3 Contributions

The significant contribution of the work can be noted as the following.

- Identification of future air quality and temperature prediction with ease.
- Comparison of various deep learning techniques to select the best model for each use case, i.e., AQI and heatwave prediction.
- Collection of new data from Telangana State Pollution control board, for fine tuning to India region.
- Provides an easily implementable methodology for prediction, which is cost effective.

1.4 Paper Organization-

The remainder of the paper is organized as the following. Section 2 explains the recent research in this area. Section 3 explains the methodology and modeling techniques. This section is followed by the results and discussion section, where we compare the performances of various time series models. Section 5 mentions the conclusion, future scope and areas of improvement of the paper.

2 Related work

With the help of combined daily meteorological observation and captured daily fire pixel data from the Moderate Resolution Imaging Spectroradiometer (MODIS), Feng et al.'s work [5] focused to train their models using BPNN ensembles. In order to estimate daily fire pixel counts, the study used the climatological biomass combustion residues data from the Fire Inventory from NCAR (FINN). These estimates were then used to power the WRF-Chem regional air quality model Significant improvements in the precision of daily PM2.5 concentration estimates in Southern China were made by integrating the BPNN-ensemble-forecasted everyday biomass fire pollutant particulates. A decrease in average inaccuracy of modeled surface PM2.5 values from -9.1% to -1.2% mirrored this improvement.

Using machine learning (ML) technologies, such as Support Vector Machines (SVM), random forests, and artificial neural networks, Khan et al.'s [6] study was aimed at creating a heatwave prediction model that is resilient to climate change. The study investigated the association between various ocean-atmospheric features and heatwave days (HWDs), stressing the necessity for a rolling approach in creating a robust climate forecasting model. With an a%NRMSE of 36, an R2 of 0.87, a md score of 0.76, and a rSD of 0.88 throughout the validation period, SVM outperformed the other ML algorithms in terms of predicting HWDs. These findings highlight the potential of the SVM model as a trustworthy tool for heatwave forecasting in the context of climate change.

According to a study by Asadollah et al. [7], a novel hybrid approach called decision tree (ABR-DT) and Ada-Boost Regression may be used to estimate annual heatwave days (HWDs) in Iran using synoptic predictors. The most effective structure was created by reducing the many predictors and their properties using the principal component analysis. Using just the particular humidity and wind components as predictors, the grid-point-specific performance evaluation demonstrated the superiority of ABR-DT, which displayed a correlation coefficient (CC) of 0.860 and a mean absolute error (MAE) of 6.929 as its metrics. The geographical performance indicators throughout Iran's eight distinct climate areas also demonstrated ABR-DT's superior performance, increased 185 and 19%, respectively, to the CC and MAE of its two alternatives.

This review describes the ways for handling model uncertainty in accordance with the phases for developing ANN models, conducted by Cabaneros et al. [8]. The review, which was based on 128 studies published between 2000 and 2022, shows that input uncertainty received more attention than structural, parameter, and output uncertainties. Neuro-fuzzy networks have been used the most, then ensemble techniques. The application of techniques that may quantify uncertainty, such as bootstrapping, Monte Carlo simulation and Bayesian analysis, was also constrained. This review recommends the creation and use of methodologies that can manage and quantify the unpredictability related to the creation of ANN models.

This work suggests an air quality prediction model based on the improved VLSTM with multichannel input and multi route output (IVLSTM-MCMR), according to research by Zhu et al. [9]. The IVLSTM and MCMR modules are part of the suggested model. The suggested IVLSTM module is created by strengthening the VLSTM inner structure in order to minimise the amount of factors that contribute to the convergence's acceleration. A multichannel data input model (MC) with better linear similarity dynamic time wrapping is added in the MCMR module to choose the appropriate data for the IVLSTM input. A multiroute output model (MR), which outputs the results of several target stations with various attributes by various routes, is created to incorporate the findings from MC..

According to a study by Zeng et al. [10], deep learning (DL) algorithms for time series data forecasting, such as the long short-term memory (LSTM) neural networks and recurrent neural network (RNN) have garnered lot of attention recently and have been used to forecast air quality indexes (AQIs). In this paper, a novel forecasting model that combines the Nested Long Short-Term Memory (NLSTM) neural network and Extended Stationary Wavelet Transform (ESWT) and is presented for forecasting PM2.5 air quality. In terms of several error metrics, including, MAE, MAPE, RMSE absolute error and R2 the findings demonstrate that the suggested technique surpasses state-of-the-art forecasting methods and recently published studies.

Sarkar et al. [11] researched on a study which mentioned, various error-prone approaches, including R-

Squared (R2), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) methods, are included in order to forecast the AQI value for Particulate Matter (PM2.5) m at a certain location in Delhi. This method combines the Gated Recurrent Unit (GRU) and Long-Short-Term Memory (LSTM) deep learning models to predict the AQI of the environment. Several machine learning (ML) and deep learning (DL) models, including LSTM, K-nearest neighbour (KNN), linear regression (LR), GRU and support vector machine (SVM), are also trained on the same dataset to compare their performances with the proposed hybrid (LSTM-GRU) model. With an MAE value of 36.11 and an R2 value of 0.84, it is discovered that the suggested hybrid model performs superiorly.

A climate model was suggested in a study by Dumas et al. [12] as a different method of predicting the occurrence of intense, long-lasting heatwaves. This new method will be helpful for a number of important scientific objectives, such as the analysis of climate model statistics, the development of a quantitative proxy for resampling uncommon occurrences in climate models, the investigation of the effects of climate change, and eventually, forecasting. used large-class under sampling, transfer learning, and 1,000 years' worth of climate model data to train a convolutional neural network. The trained network performs much better than the untrained network in predicting the presence of prolonged, intense heatwaves using the observed snapshots of the surface temperature and the 500hPa geopotential height fields.

Work	Author/ Year	Methodology	Findings	Advantages	Limitations
5	Feng et al.	• MODIS-recorded fire occurrences • BPNN ensembles-predicted fire occurrences	• Applied to the normalized mean bias between the simulated (PM2.5) concentrations is the NCAR Fire Inventory (FINN).	• The BPNN ensembles effectively predicted the fire pixel's daily fluctuation	• The accuracy of the BPNN model was just 70%.
6	Khan et al	•Elimination method was used to identify the input variables by using recursive feature which was based on SVM. •Developed using window technology over a 5-year time step.	With an %N-RMSE of 36, R2 of 0.87, the SVM's greater ability to forecast HWDs was exhibited.	The forward rolling model based on SVM performed better in predicting heatwaves.	To make better decisions on the likelihood of heatwaves, forecast uncertainty may be evaluated.
7	Asadollah et al.	Decision tree and Ada-Boost Regression- A hybrid technique for heatwave forecast	ABR-DT performed better, proven from Grid point based evaluation. Correlation is 0.860, and MAE=6.929, which used wind and specific humidity as input features.	Hybrid models performed better.	Inaccurate results due to alteration of predictor variables
8	Cabaneros et al	• Bootstrap method • Bayesian method • Fuzzy method • Sensitivity simulation • Genetic algorithm • Monte Carlo analysis • Ensemble technique	The emergence of ANN models has made it possible for researchers to produce precise AP forecasts without the theoretical knowledge necessary for conventional physics-based models.	The creation and use of techniques that can address and quantify the uncertainty that surrounds the creation of ANN models.	It is necessary to conduct more study on how to better report the accuracy and uncertainty of ANN model findings.
9	Fang et al	A deep learning network module with fewer parameters and greater use of previous data is initially presented as an IVLSTM structure.	The suggested IVLSTM-MCMR performs well in terms of precision and efficiency because it is able to react fast to variations in time series.	The more precise air quality forecast models FFA and STE both perform better than conventional methods.	The ability to understand periodicity and enhance forecast accuracy is more challenging.

10	Chen et al.	A framework for AQI forecast using deep learning, with wavelet transform and zero mean normalization.	Reduction in absolute error and R2 index, indicates that model fits the data accurately.	Exhibits better performance than others. ESWT NLSTM, framework is better in real life.	• Inconsistent data decomposition. •Unable to fit properly.
11	Sarkar et al	• Data collection • Data pre-processing which includes Data Inputation, Data Aggregation, Data Normalization, and Feature Selection	•For forecast the AQI of additional cities suggested method can be expanded. •The MAE score of 36.11 and the R2 value of 0.84 demonstrate the hybrid model's superior performance.	To compare model performances with the suggested hybrid (LSTM-GRU) model to provide more precise information, models are developed on an identical dataset.	Exploring and using model hyper-parameter optimisation is possible.
12	Dumas et al.	Large-class undersampling, transfer learning, and 1,000 years' worth of climate model outputs used in CNN	TPR= 56.3 (12.9), FPR=2.1 (without Transfer learning)	Predicted for three distinct degrees of rigour, beginning as soon as 15 days before the event (30 days before the event's conclusion)	Less positive occurrences are included in the datasets, which makes training's learning task more challenging.

3 Methodology

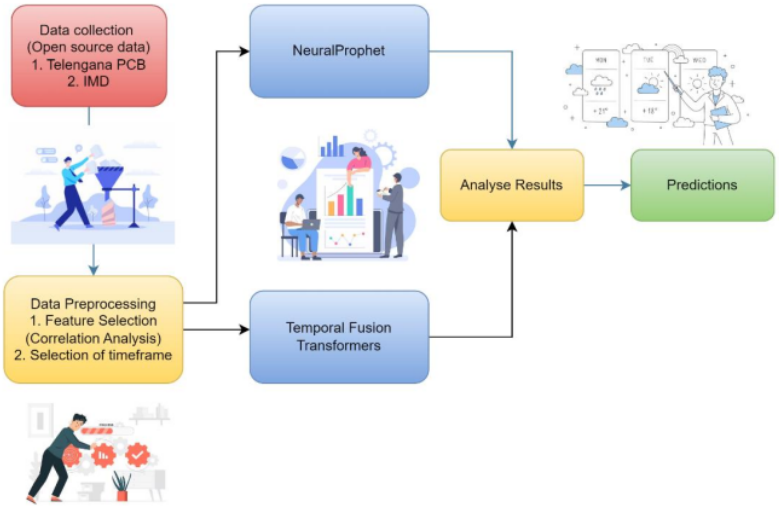


Fig 1. Process flow for methodology

3.1 Preprocessing

Data is collected from Telangana pollution control board, IMD and Kaggle dataset (open source). The collected data is then compiled into preferred csv file for easy manipulation. This is followed by selection of time duration for analysis. A correlation analysis is done to find out which features have higher impact on AQI prediction.

	City	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	Air_quality
City	1.000000	0.062446	-0.021395	0.047182	-0.078589	0.024948	-0.182565	-0.255886	-0.101327	0.001530	0.002070	-0.062730	-0.047806	-0.121778	-0.053770
PM2.5	0.062446	1.000000	0.520767	0.435245	0.358287	0.375938	0.164615	0.093695	0.120451	0.166064	0.030539	0.132609	0.087813	0.633654	0.408614
PM10	-0.021395	0.520767	1.000000	0.433864	0.352396	0.416523	0.200548	0.023932	0.165468	0.208398	0.027522	0.126400	0.030257	0.447686	0.208080
NO	0.047182	0.435245	0.433864	1.000000	0.465848	0.747409	0.155804	0.221445	0.174881	0.020788	0.045202	0.151930	0.084393	0.438039	0.259407
NO2	-0.078589	0.358287	0.352396	0.465848	1.000000	0.581938	0.151731	0.361750	0.388413	0.289435	0.042179	0.297913	0.188965	0.531952	0.259199
NOx	0.024948	0.375938	0.416523	0.747409	0.581938	1.000000	0.129433	0.231878	0.205912	0.089095	0.049004	0.184764	0.085447	0.436282	0.212592
NH3	-0.182565	0.164615	0.200548	0.155804	0.151731	0.129433	1.000000	-0.014252	-0.048679	0.076295	0.004093	0.020040	-0.030328	0.097916	0.077250
CO	-0.255886	0.093695	0.023932	0.221445	0.361750	0.231878	-0.014252	1.000000	0.476784	0.043311	0.066022	0.285559	0.181215	0.650475	0.215848
SO2	-0.101327	0.120451	0.165468	0.174881	0.388413	0.205912	-0.048679	0.476784	1.000000	0.160489	0.038955	0.273314	0.249808	0.454182	0.170843
O3	0.001530	0.166064	0.208398	0.020788	0.289435	0.089095	0.076295	0.043311	0.160489	1.000000	0.023982	0.135498	0.084323	0.199100	0.030637
Benzene	0.002070	0.030539	0.027522	0.045202	0.042179	0.049004	0.004093	0.066022	0.038955	0.023982	1.000000	0.693310	0.094190	0.052027	0.037358
Toluene	-0.062730	0.132609	0.126400	0.151930	0.297913	0.184764	0.020040	0.285559	0.273314	0.135498	0.693310	1.000000	0.289731	0.288749	0.160761
Xylene	-0.047806	0.087813	0.030257	0.084393	0.188965	0.085447	-0.030328	0.181215	0.249808	0.084323	0.094190	0.289731	1.000000	0.186991	0.092216
AQI	-0.121778	0.633654	0.447686	0.438039	0.531952	0.436282	0.097916	0.650475	0.454182	0.199100	0.052027	0.288749	0.186991	1.000000	0.467065
Air_quality	-0.053770	0.408614	0.208080	0.259407	0.259199	0.212592	0.077250	0.215848	0.170843	0.030637	0.037358	0.160761	0.092216	0.467065	1.000000

Fig 2. Correlation analysis

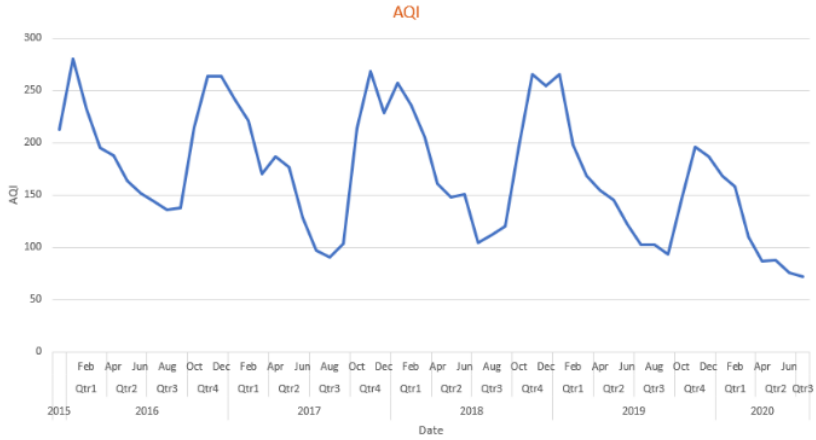


Fig 3. AQI shows a repeating pattern.

From the table we can notice that PM2.5, PM10, NOx, CO show positive correlation coefficient.

3.2 Neural Prophet

We propose a methodology to use time series forecasting with all air quality parameters such as SO2, NO2, PM10, PM2.5, NOx, CO, Benzene, Toluene and Xylene to model a multivariate approach.

First, the modelling was done with a univariate approach using, Meta's Neural Prophet framework. The model components include seasonality, trend, events, regressors, auto regression, covariates and global modelling. It is the intersection of traditional methods and recent deep learning methods. We fine-tuned the model with a trend regularization of 2, with a yearly seasonality.

The system learns to identify change points, or dates when a distinct deviation in trend happens. These dots are evenly initialized along the time axis. The cost of each of these linear regressions is then added to the model's total loss. Gradient descent reduces loss and hence improves regression. Specifically, the optimal parameter of linear regression is the slope, often known as the growth value.

Here, the use of yearly seasonality has been set to True, which enable the model to learn patterns which repeat in every 365 days. From EDA, it was clear that the AQI follows a yearly seasonality. Here, the model assumes that target (AQI) is a periodic and continuous function, which can be expressed as a Fourier series.

$$S_P(t) = \sum_{j=1}^{k} \left(a_j \cdot cos\left(\frac{2\pi jt}{p}\right) + b_j \cdot sin\left(\frac{2\pi jt}{p}\right) \right)$$

Where k indicates the number of Fourier terms for the seasonality with periodicity (p).
In cases with multiple seasonality, n values differ, for each periodicity.

The practice of regressing a variable's future value versus its past values is known as auto-regression (AR). A important component of many forecasting applications is auto-regression. The number of prior values contained is typically referred to as the AR (p) model's order p. Hence, a coefficient θ_i fits each historical value. Each coefficient θ_i determines the size and direction of the impact of a certain historical value on the projection.

$$\boldsymbol{y_t = c + \sum_{i=1}^{i=p} \theta_i \cdot y_{t-i} + e_t}$$

3.3 Temporal Fusion Transformers

The Temporal Fusion Transformer (TFT), a novel attention-based structure, integrates multi-horizon forecasting with excellent accuracy and comprehensible temporal dynamics. To learn temporal correlations at different scales, TFT uses interpretable self-attention layers for long-term reliance and recurrent layers for local processing. TFT uses specialized components to select critical features and a series of gating layers to suppress redundant components, enabling greater performance in a range of conditions.
The multivariate modelling was done using TFT architecture with using PyTorch. The time varying known values were set as the features and AQI was set as unknown values.

3.4 AQI Prediction-

We defined a time series forecasting problem with a maximum prediction length of 150 days (5 months), using the Pytorch Forecasting library. The data is assumed to be stored in a pandas DataFrame with a column "time_idx" representing the time index, and a column "AQI" representing the target variable to be predicted. There is also a categorical variable "City" which identifies the city to which each time series belongs.
Then a training dataset was created using the TimeSeriesDataSet class from the Pytorch Forecasting library. The training dataset is defined as a subset of the original data, with a training cutoff that is 150 days before the end of the time index. The encoder length is set to 720, which is twice the maximum prediction length, to allow for a long enough history to capture seasonal patterns. The dataset is grouped by city, and the target variable is normalized using the soft plus function and normalized by group. The time index, as well as several real-valued features related to air quality, are included as time-varying known reals.
A validation dataset is also created using the same TimeSeriesDataSet class, with predict=True to indicate that it should be used for predicting the last 150 days of each time series. The validation dataset is created from the training dataset to ensure that the categorical encoding and normalization parameters are consistent.
Finally, the code creates data loaders for the model using the 'to_dataloader' method of the TimeSeriesDataSet class. The batch size is set to 128 for the training data loader and 1280 (10 times the training batch size) for the validation data loader. The PyTorch Data Loader class is used to load data in batches for efficient training of the model.
Next, the code sets up a PyTorch Lightning Trainer object to train and evaluate the model. The trainer is configured to run for a maximum of 30 epochs and log results to a TensorBoardLogger. Early stopping is used to stop training if the validation loss does not improve by at least 1e-4 for 10 consecutive epochs. The learning rate is also monitored and logged during training.
The TFT model is defined using the TemporalFusionTransformer class. This method initializes the model with the same categorical encoding and normalization parameters as the training dataset. Hyperparameters such as the learning rate, hidden size, attention head size, dropout rate, and output size (which determines the number of quantiles to predict) are used to train the model. The loss function is set to the QuantileLoss, which minimizes the quantile loss between the predicted and actual quantiles. The reduce_on_plateau_patience parameter reduces the learning rate by a factor of 10 if the validation loss does not improve after 4 epochs.

Heatwave Prediction-
The heatwave prediction is done, using TFT architecture, implemented using PyTorch forecasting. The collected data is preprocessed and trained for 30 epochs with 3 attention heads, learning rate of 0.03.

4 Results and Discussion

4.1 MAPE (Mean Absolute Percentage Error)

MAPE (Mean Absolute Percentage Error) is a frequently used statistic to assess the precision of predictions provided by regression models. It calculates the percentage difference between a variable's expected value and its actual value.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where, n = iteration of the summation function, A_t= Actual Value, F_t= Predicted value

RMSE (Root Mean Squared Error)

Root Mean Squared Error (RMSE) is a statistic used to assess the accuracy of regression model predictions. It is calculated by taking the square root of the average of the squared discrepancies between the expected and actual values.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

These two metrics being very useful for evaluation regression workloads, we efficiently tried to use this.
Different experiments were conducted using NeuralProphet, TFTs and NBeats models to determine the efficacy of each on AQI Prediction and heatwave prediction. The primary datasets used for this work is sourced from Telengana state pollution control board.

Prediction of each feature

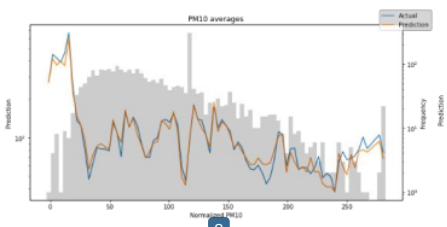


Fig 4

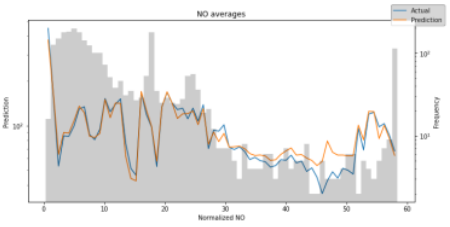
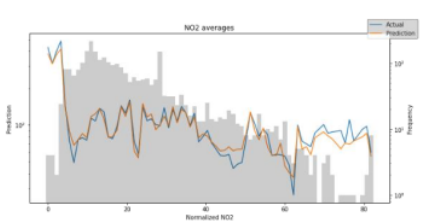


Fig 5



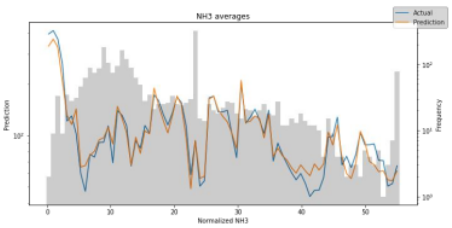


Fig 6

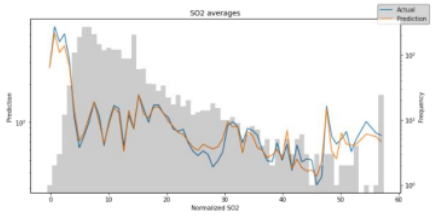


Fig 7

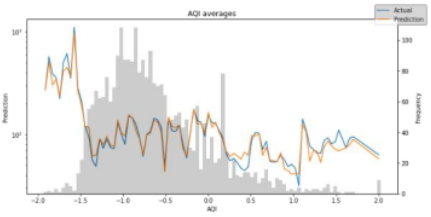
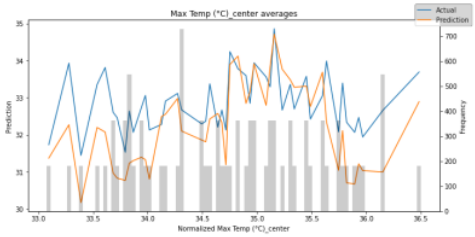


Fig 8

Fig 9

The above figures i.e Fig. 4 to Fig. 9 show that it is possible to accurately predict the features using TFT and it provides a feasible technique to predict the AQI.



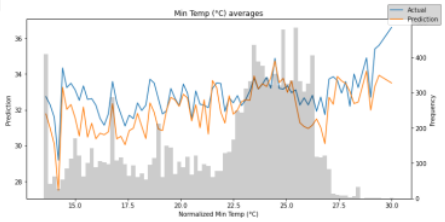


Fig 10

Fig 11

Fig. 4-11 Prediction vs Actual Values with TFT model

On experimentation with different models and hyperparameters it is found that multivariate forecasting using NeuralProphet model with additive future regressors performed best for AQI prediction. SO_2, PM_{10} were used as the future regressors. The mean absolute percentage error (MAPE) is found to be 7%.

For Heatwave determination, the best result is obtained by using TFT Model with a MAPE of 4 %.

Table 1. Comparison of results from various models

Model	AQI	Heatwave (Max Temp)
SARIMA	MAPE=19.38% RMSE=22.38	MAPE=23.33%, RMSE= 8.67%
LSTM	MAPE= 18.99%, RMSE= 19.72	MAPE=5.1%, **RMSE=2.32**
Neural-Prophet (Univariate)	MAPE=0.17, (17%), RMSE= 0.3	MAPE= 0.05 (5%), RMSE= 0.09
NeuralProphet (Multivariate)	Future Regressors as SO_2, PM_{10} MAPE= 0.07 (7%) RMSE= 0.05	Future Regressors as max. Humidity, min. temp and precipitation MAPE= 0.1263 (12.63%) RMSE=0.084
Temporal Fusion Transformer (TFT) (Multivariate)	MAPE=0.14, (14%) RMSE= 24.6	MAPE=0.04, (4%) RMSE=1.66
NBeats	MAPE=0.24 (24%), RMSE= 0.036	

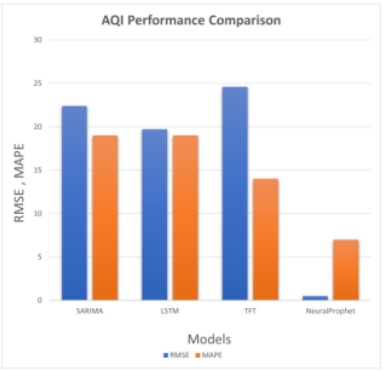


Fig. 12. AQI Error comparison

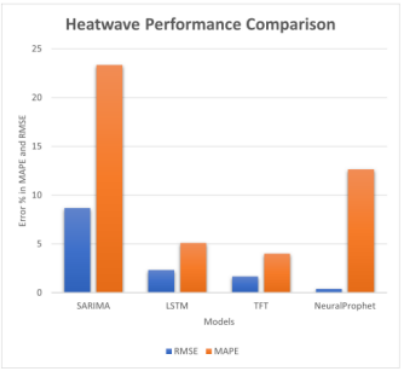


Fig. 13. Heatwave Error comparison

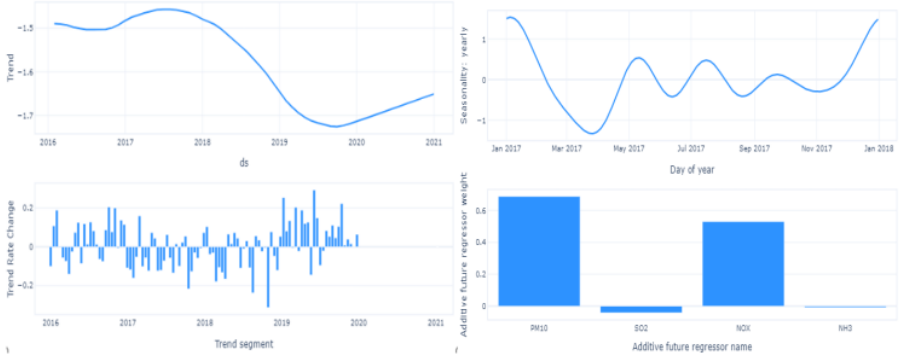


Fig. 14. Neural Prophet Model metrics for Multivariate AQI

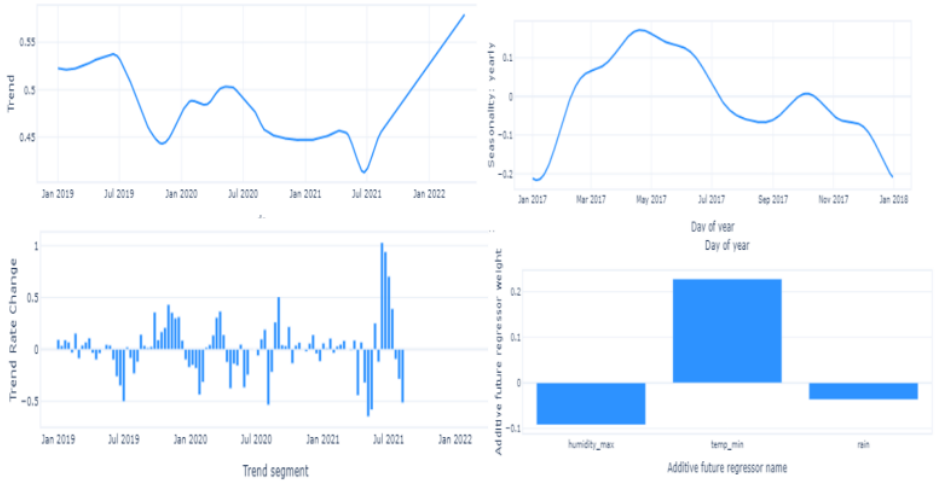


Fig.15 Neural Prophet Model metrics for Temperature prediction (for Heatwave)

The visualization of the trend of data with seasonality is depicted in fig. 14, and fig. 15 for AQI and heatwave prediction. It is seen that for AQI, PM10 and NOx values hold the highest weight as additive future regressor. A high weight for the variables indicates that the additional time series data provided is highly correlated with the time series being forecasted. This suggests that the exogenous variable included in the data may have a significant impact on the time series being forecasted and can help improve the accuracy of the forecast. Similarly for heatwave prediction, it is found that the maximum humidity and minimum temperature of the day provided a significant role in prediction.

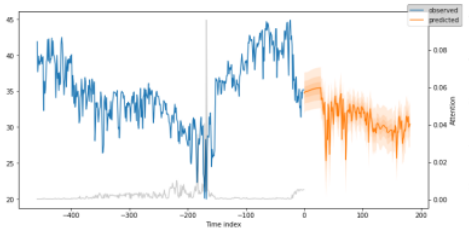


Fig 16. Max Temp Prediction using TFT (for future timeline) - Adilabad Region

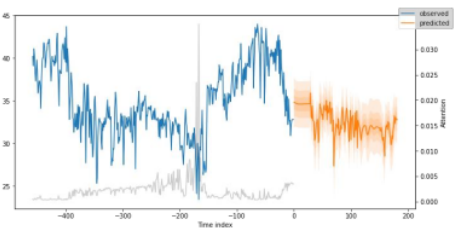


Fig.17. Max Temp Prediction using TFT (for future timeline) - Karim Nagar Region

In the fig 16 and 17 it is indicated the max temp prediction for the heatwave analysis done by the TFT model for the region of Adilabad and Karimnagar.

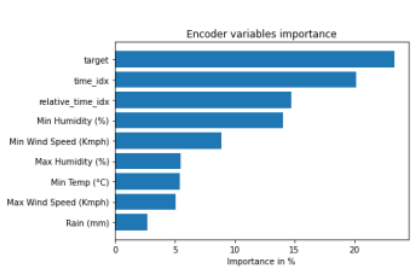


Fig 18. encoder variable importance as per TFT model

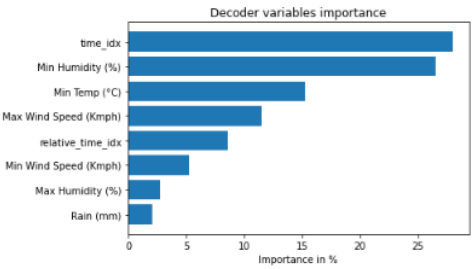


Fig. 19. decoder variable importance as per TFT model

The fig 18 and 19 indicate the importance of variables in training the TFT model and forecasting/prediction on given data.

5 Conclusion and Future Scope

The study on AQI and Heatwave prediction is an accurate and reliable way to forecast and predict weather events. The research conducted on opensource data from IMD as well as Telangana PSB, focused on the use of the latest DL architectures for univariate and multivariate forecasting techniques. The correlation analysis determined that PM10, PM2.5, SO_2 and NOx were correlated with the AQI and minimum temperature and humidity correlated with the temperature. This provided us an insight into selection of variables as additive regressors for future forecast. The use of NeuralProphet, Temporal Fusion Transformer(TFT) with extensive hyperparameter tuning provided us with best results,i.e., MAPE of 7% on AQI (univariate) and MAPE of 4% on TFT (multivariate) for heatwave (Max Temp) prediction.

There are minor issues which need to be clarified for further studies such as a high weight for the Additive Future Regressor (AFR) variable does not necessarily imply causation. It is possible that the high correlation between the AFR and the time series being forecasted is simply a coincidence. Therefore, it is important to carefully interpret the results of the model and consider the underlying data generating process when using the AFR in neural prophet model.

6 Reference

1. Kiyan, A., Gheibi, M., Akrami, M., Moezzi, R., Behzadian, and K.: A Comprehensive Platform for Air Pollution Control System Operation in Smart Cities of Developing Countries: A Case Study of Tehran. Environmental Industry Letters. Vol. 1 No. 1 (2023): Environmental Industry Letters (EIL), https://doi.org/10.15157/EIL.2023.1.1.10-27, (2023).

2. Balogun, A.-L., Tella, A., Baloo, L., Adebisi, N.: A review of the inter-correlation of climate change, air pollution and urban sustainability using novel machine learning algorithms and spatial information science, http://dx.doi.org/10.1016/j.uclim.2021.100989, (2021).

3. Kalajdjieski, J., Zdravevski, E., Corizzo, R., Lameski, P., Kalajdziski, S., Pires, I.M., Garcia, N.M., Trajkovik, V.: Air Pollution Prediction with Multi-Modal Data and Deep Neural Networks, http://dx.doi.org/10.3390/rs12244142, (2020).

4. Ravindra, K. et al.: Generalized additive models: Building evidence of air pollution, climate change and human health, http://dx.doi.org/10.1016/j.envint.2019.104987, (2019).

5. Feng, X., Fu, T.-M., Cao, H., Tian, H., Fan, Q., Chen, X.: Neural network predictions of pollutant emissions from open burning of crop residues: Application to air quality forecasts in southern China, http://dx.doi.org/10.1016/j.atmosenv.2019.02.002, (2019).

6. Khan, N., Shahid, S., Ismail, T.B. et al. Prediction of heat waves over Pakistan using support vector machine algorithm in the context of climate change. Stoch Environ Res Risk Assess 35, 1335–1353 (2021). https://doi.org/10.1007/s00477-020-01963-1

7. Asadollah, S.B.H.S., Khan, N., Sharafati, A., Shahid, S., Chung, E.-S., Wang, X.-J.: Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models, http://dx.doi.org/10.1007/s00477-021-02103-z, (2021).

8. Cabaneros, S.M., Hughes, B.: Methods used for handling and quantifying model uncertainty of artificial neural network models for air pollution forecasting, http://dx.doi.org/10.1016/j.envsoft.2022.105529, (2022).

9. Fang, W., Zhu, R., Lin, and J.C.-W.: An air quality prediction model based on improved Vanilla LSTM with multichannel input and multiroute output, http://dx.doi.org/10.1016/j.eswa.2022.118422, (2023).

10. Zeng, Y., Chen, J., Jin, N., Jin, X., Du, Y.: Air quality forecasting with hybrid LSTM and extended stationary wavelet transform,

http://dx.doi.org/10.1016/j.buildenv.2022.108822, (2022).

11. Sarkar, N., Gupta, R., Keserwani, P.K., Govil, M.C.: Air Quality Index prediction using an effective hybrid deep learning model, http://dx.doi.org/10.1016/j.envpol.2022.120404, (2022).

12. Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., Bouchet, F.: Deep Learning-Based Extreme Heatwave Forecast, http://dx.doi.org/10.3389/fclim.2022.789641, (2022).

13. Li, G., Tang, Y., Yang, H.: A new hybrid prediction model of air quality index based on secondary decomposition and improved kernel extreme learning machine, http://dx.doi.org/10.1016/j.chemosphere.2022.135348, (2022).

14. Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., Li, F.: Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering, http://dx.doi.org/10.1016/j.eswa.2020.114513, (2021).

15. Ren, Z., Ji, X.: On prediction of air pollutants with Takagi-Sugeno models based on a hierarchical clustering identification method, http://dx.doi.org/10.1016/j.apr.2023.101731, (2023).

16. Aragão, D.P., Oliveira, E.V., Bezerra, A.A., dos Santos, D.H., da Silva Junior, A.G., Pereira, I.G., Piscitelli, P., Miani, A., Distante, C., Cuno, J.S., Conci, A., Gonçalves, L.M.G.: Multivariate data driven prediction of COVID-19 dynamics: Towards new results with temperature, humidity and air quality data, http://dx.doi.org/10.1016/j.envres.2021.112348, (2022).

17. Gao, M., Yang, H., Xiao, Q., Goh, M.: COVID-19 lockdowns and air quality: Evidence from grey spatiotemporal forecasts, http://dx.doi.org/10.1016/j.seps.2022.101228, (2022).

18. Das, B., Dursun, Ö.O., Toraman, S.: Prediction of air pollutants for air quality using deep learning methods in a metropolitan city, http://dx.doi.org/10.1016/j.uclim.2022.101291, (2022).

19. Jin, X.-B., Wang, Z.-Y., Gong, W.-T., Kong, J.-L., Bai, Y.-T., Su, T.-L., Ma, H.-J., Chakrabarti, P.: Variational Bayesian Network with Information Interpretability Filtering for Air Quality Forecasting, http://dx.doi.org/10.3390/math11040837, (2023).

20. Janarthanan, R., Partheeban, P., Somasundaram, K., Navin Elamparithi, P.: A deep learning approach for prediction of air quality index in a metropolitan city, http://dx.doi.org/10.1016/j.scs.2021.102720, (2021).

Multivariate Machine Learning Approaches for Dynamic Prediction of Air Quality and Estimating Heatwave Occurrence

ORIGINALITY REPORT

11%	10%	10%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	4%
2	www.mdpi.com Internet Source	1%
3	www.frontiersin.org Internet Source	1%
4	Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, Tomas Pfister. "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting", International Journal of Forecasting, 2021 Publication	1%
5	Xu Feng, Tzung-May Fu, Hansen Cao, Heng Tian, Qi Fan, Xiaoyang Chen. "Neural network predictions of pollutant emissions from open burning of crop residues: application to air quality forecasts in Southern China", Atmospheric Environment, 2019 Publication	1%

6	link.springer.com Internet Source	1 %
7	Submitted to University of Edinburgh Student Paper	<1 %
8	Nairita Sarkar, Rajan Gupta, Pankaj Kumar Keserwani, Mahesh Chandra Govil. "Air Quality Index prediction using an effective hybrid deep learning model", Environmental Pollution, 2022 Publication	<1 %
9	Yongkang Zeng, Jingjing Chen, Ning Jin, Xiaoping Jin, Yang Du. "Air quality forecasting with hybrid LSTM and extended stationary wavelet transform", Building and Environment, 2022 Publication	<1 %
10	www.ncbi.nlm.nih.gov Internet Source	<1 %
11	Submitted to Bournemouth University Student Paper	<1 %
12	whp66.ok-em.com Internet Source	<1 %
13	lppm.ub.ac.id Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		