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

by Abhinandan Roul

Submission date: 08-Jun-2023 07:19PM (UTC+0530)
Submission ID: 2111772715
File name: SDP_Report_June_8th_v1.2_without_cover_page_etc_abhinandan.docx (1.67M)
Word count: 5594
Character count: 32706

PREFACE

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. Multiple models each for AQI and temperature is trained to determine the best possible model, with evaluation. 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.

INDIVIDUAL CONTRIBUTIONS

Abhinandan Roul	Problem formulation, Experimentation, Result Analysis, Documentation
Shubhaprasad Padhy	Solution design, Experimentation, Documentation
Sambit Kumar Sahoo	Literature survey, Result Analysis, and documentation
Ayush Pattanayak	Literature survey, Documentation, Data Collection and Processing

Chapter 1: INTRODUCTION

1.1 Project Overview

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 affectair quality and heatwave occurrence.

The use of multivariate machine learning approaches for dynamic prediction of air quality andestimating 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 astemperature, 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 andsupport decision-making related to these issues [4].

1.2 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 can 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 toheat wave mitigation?
- RQ2: What are the different factor that affect air quality and heatwave occurrence andsupport 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?

1.3 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.

1.4 Uniqueness of the work

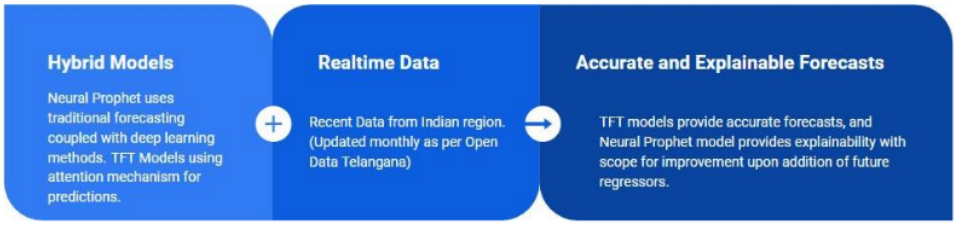


Fig 1: Key features of proposed work

- Use of Hybrid methods (DL and ML) for prediction.
- Usage of diverse and recent data (from India region).
- Application of attention based time series model (TFT) for predictions.
- For short to medium-term forecasts, NeuralProphet improves forecast accuracy by 55 to92 percent.
- Explainable model with individually interpretable components.

Chapter 2: LITERATURE SURVEY

2.1 Existing System

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 a 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.

2.2 Problem Identification

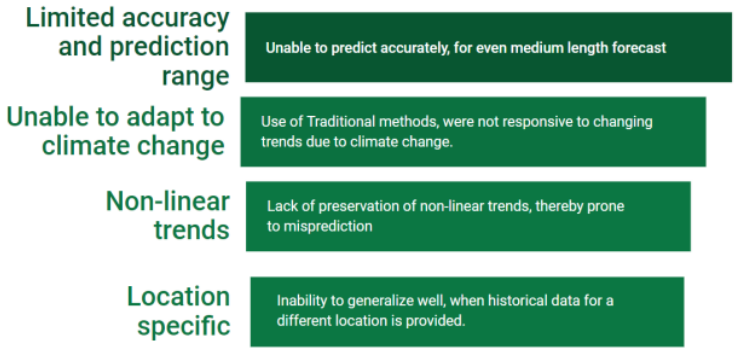


Fig 2 Problem identification

- Unable to predict with accuracy in advance for medium to longer time steps.
- Use of traditional methods, were not responsive to changing trends due to climate change:Traditional methods used for predicting time series data are usually based on historical patterns, which may not accurately capture the changing trends due to climate change. Forinstance, traditional methods for forecasting the weather may not consider the effects of global warming, leading to inaccurate predictions.
- Non-linear features in trends were not preserved: Traditional models often make assumptions about the linear nature of the underlying data, which may not be appropriate for many real-world time series datasets. Non-linear features, such as trends or seasonality, may be missed in these models, leading to inaccuracies in the predictions.
- Inability to generalize well, when historical data for a different location is provided: Traditional time series models may struggle to generalize them This is because different locations may have different patterns and trends in their time series data, and traditional models may not be able to capture these differences effectively.

Chapter 3: MATERIALS AND METHODS

3.1 Dataset description

Fig 3 Indian Methodological Department and Telengana Pollution Control Board

3.1.2 Indian Meteorological Department (IMD):

- Daily AQI Data from 15 cities across India. (2016-2020) with parameters ($PM_{2.5}$, PM_{10} , NO, NO_2, NO_X , NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene)

3.1.1 Telangana Pollution control board (T-PCB):

- Ambient air quality data (2015-2022) with parameters (SO_2, NO_X, PM_{10}, $PM_{2.5}$, NH_3,Ozone, CO, Benzene) with more than 25 sampling locations across Telangana.
- Weather data (2017-2022) from T-PCB: Daily data for 33 locations across Telangana,with parameters as rainfall, min and max temperature, humidity and wind speed.

3.2 Schematic layout/ Model diagram

- Data Collection: From provided sources as mentioned in previous slide.
- Pre-processing: Feature selection done using Correlation analysis. Min-Max normalization is done for AQI, and temperature.
- Modelling: Trained using Neural Prophet, Temporal Fusion Transformer.
- Hyperparameter tuning: For determining best model.

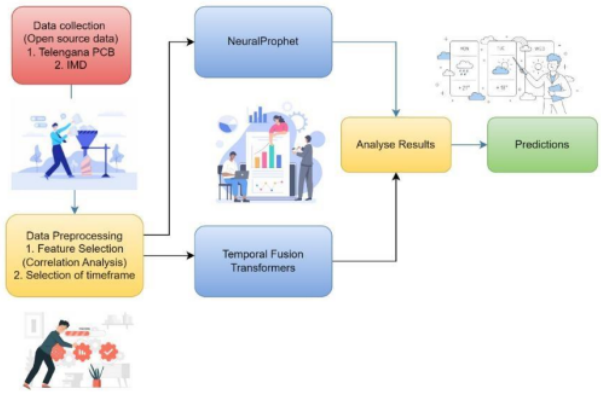


Fig 4: Process flow for methodology

3.3 Methods Used

3.3.1 Neural Prophet

Fig 5. Neural Prophet Library

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 multivariateapproach.

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 ayearly 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 regressionsis 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 asthe 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). An 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 $\theta_i$determines the size and direction of the impact of a certain historical value on the projection.

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

3.3.2 Temporal Fusion Transformers

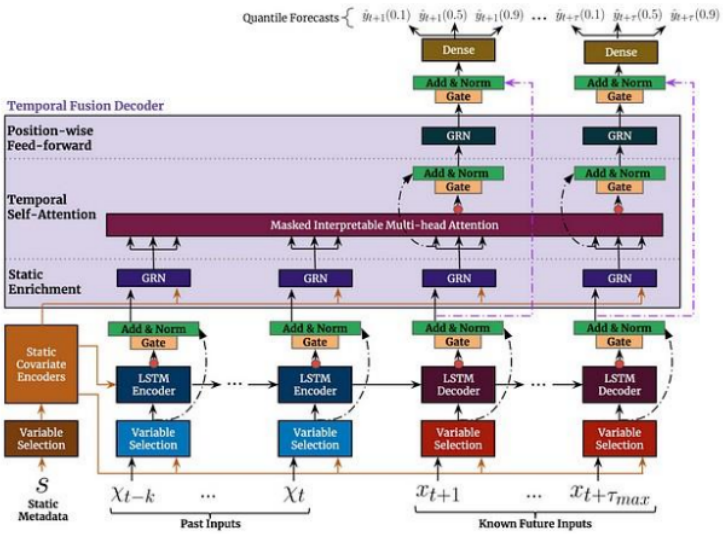


Fig 6. Temporal Fusion Transformer Architecture

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 Tools Used

For AQI:

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 normalizedusing 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=Trueto 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 (10times the training batch size) for the validation data loader. The PyTorch Data Loader class isused 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. Thetrainer 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 initializesthe model with the same categorical encoding and normalization parameters as the training dataset. The model architecture is specified with hyperparameters such as the learning rate, hidden size, attention head size, dropout rate, and output size (which determines the number ofquantiles to predict). 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 reduce_on_plateau_patience parameter isused to reduce the learning rate by a factor of 10 if the validation loss does not improve after 4epochs.

For 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.

3.5 Evaluation Measures Used

3.5.1 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

3.5.2 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.

Chapter 4: RESULTS / OUTPUTS

4.1 System Specification

Fig 7: Pictorial representation of system componentsThe experiments were carried out using the following specifications.

- Operating System: Windows 11 64-bit
- Processor: Core i5 11th generation
- RAM: 16GB
- Data Preprocessing:
  - Local Storage: Utilized for efficient data pre-processing.
- Data Storage:
  - Cloud Storage: Processed data securely uploaded for accessibility and dataintegrity.
- Training Environment:
  - Google Colab and Kaggle: Leveraged for training machine learning models.
- Benefits of System Configuration:
  - Robust and Efficient: The system configuration facilitated a reliable andefficient experimental framework.

Powerful Computing Resources: Google Colab and Kaggle provided computational resourcesfor implementing and evaluating sophisticated models.

4.2 Parameters Used

TFT Architecture:

(Creation of TimeSeriesDataSet for training purpose)

- Max Encoder Length=730
- Max Prediction Length=180
- LSTM Layers=2
- Time Varying Reals Encoder= humidity, max temp, min temp, max and min windspeed, rain
- Time Varying Reals Decoder=humidity, max temp, min temp, max and min windspeed, rain
- Learning Rate=0.03
- Hidden Size=16
- Attention Head Size=3
- Dropout=0.1
- Hidden Continuous Size=8
- Output Size=7
- Loss= Quantile Loss

Neural Prophet:

- Yearly Seasonality= True
- Weekly Seasonality= False
- Daily seasonality= False
- Batch Size=10
- Trend Regularizer=4

4.3 Experimental Outcomes

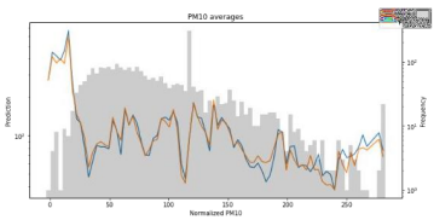


Fig 8: Prediction of PM_{10}

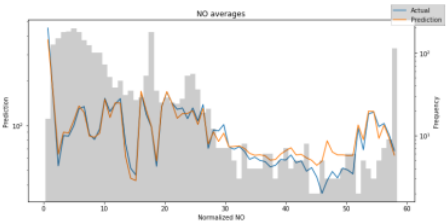


Fig 9: Prediction of NO

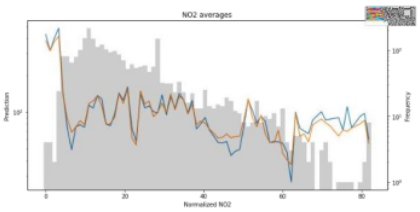


Fig 10: Prediction of NO_2

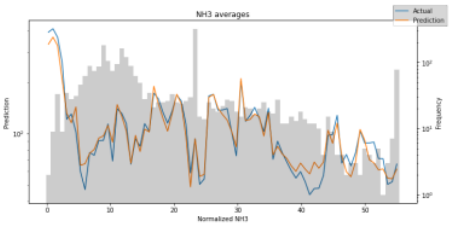


Fig 11: Prediction of NH_3

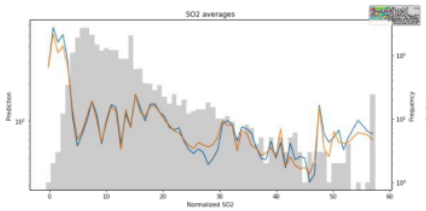


Fig 12: Prediction of SO_2

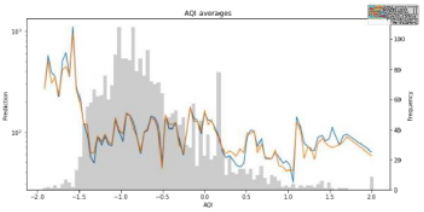


Fig 13: Prediction of AQI

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

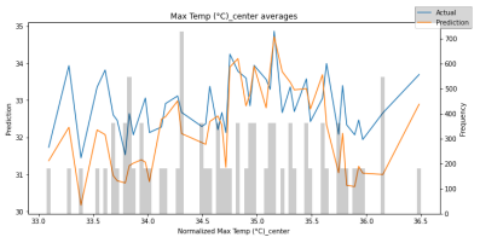


Fig 14: Prediction of Max Temp

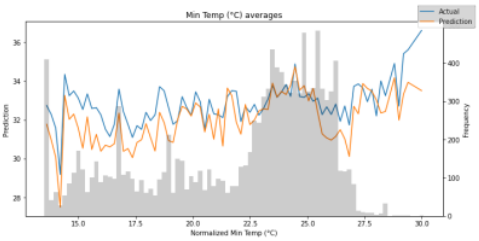
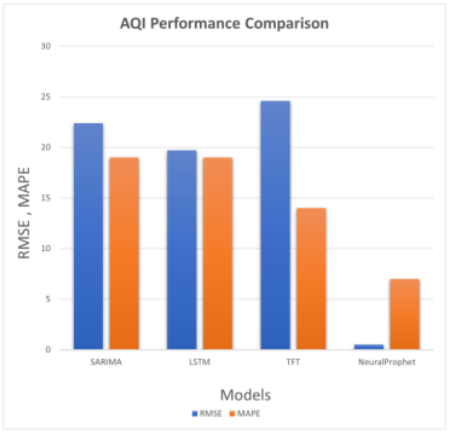


Fig 15: Prediction of Min Temp

On experimentation with different models and hyperparameters it is found that multivariate forecasting using Neural Prophet model with additive future regressors performed best for AQIprediction. 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 of4%.

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	



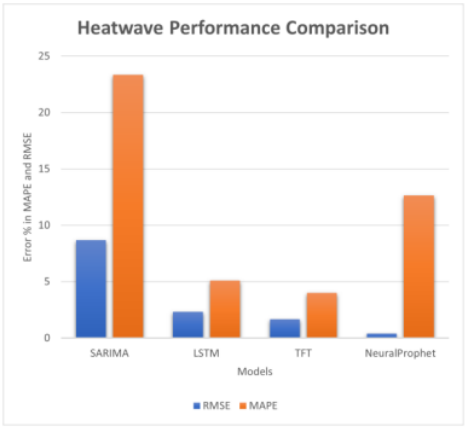


Fig. 16. AQI Error comparison

Fig. 17. Heatwave Error comparison

The fig 16, 17 show the error metrics for the comparison of AQI and Heatwave prediction across different models. It is found that NeuralProphet provides better performance in AQI analysis and TFT provides better performance in Heatwave prediction.

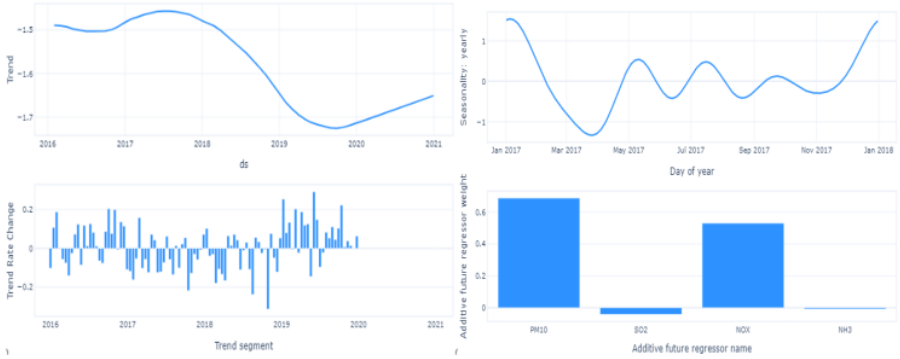
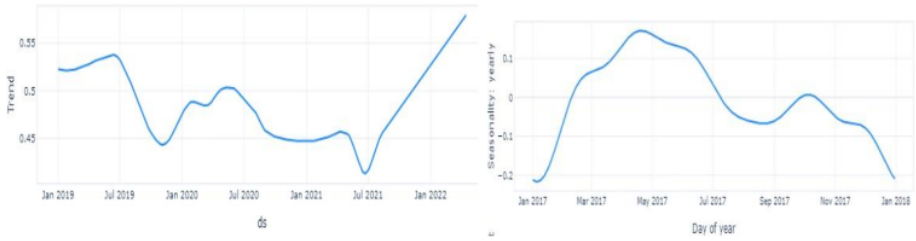


Fig. 18. Neural Prophet Model metrics for Multivariate AQI



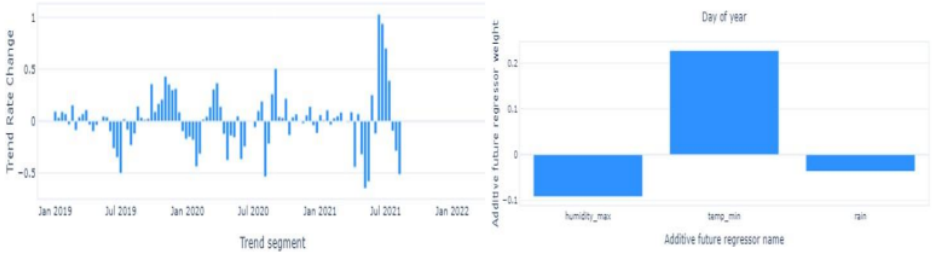


Fig.19 Neural Prophet Model metrics for Temperature prediction (for Heatwave)

The visualization of the trend of data with seasonality is depicted in fig. 18, and fig. 19 for AQI and heatwave prediction. It is seen that for AQI, PM10 and NOx values hold the highest weightas 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 theday provided a significant role in prediction.

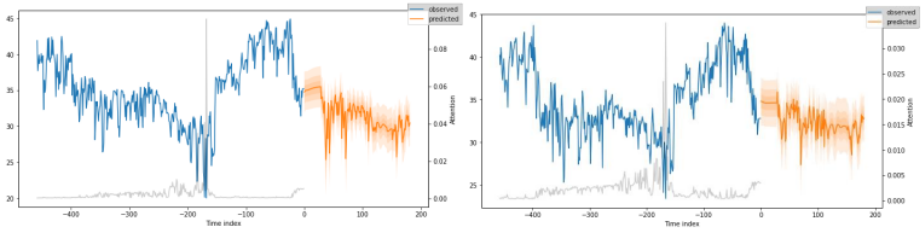


Fig 20. Max Temp Prediction using TFT (for future timeline) - Adilabad Region

Fig.21. Max Temp Prediction using TFT (for future timeline) - Karim Nagar Region

In fig 20 and 21 it is indicated the max temp prediction for the heatwave analysis done by the TFT model for the region of Adilabad and Karimnagar.

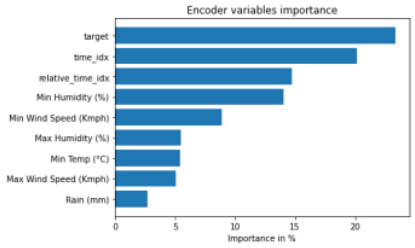


Fig 22. encoder variable importance as per TFT model

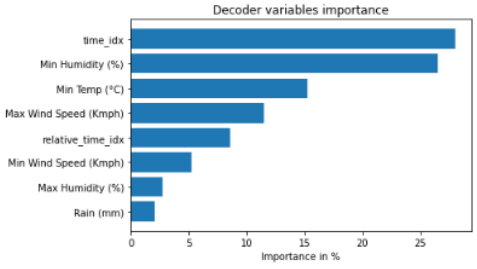


Fig. 23. decoder variable importance as per TFT model

The fig 22 and 23 indicate the importance of variables in training the TFT model and forecasting/prediction on given data.

Chapter 5: CONCLUSIONS

5.1: Key Findings

The study on AQI and Heatwave prediction is an accurate and reliable way to forecastand 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 univariateand 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 selectionof variables as additive regressors for future forecast. The use of NeuralProphet, Temporal Fusion Transformer(TFT) with extensive hyperparameter tuning provided uswith best results,i.e., MAPE of 7% on AQI (univariate) and MAPE of 4% on TFT (multivariate) for heatwave (Max Temp) prediction.

5.2: Future scope of Proposed System

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 interpretthe results of the model and consider the underlying data generating process when usingthe AFR in neural prophet model.

Chapter 6: REFERENCES

1. Kiyan, A., Gheibi, M., Akrami, M., Moezzi, R., Behzadian, and K.: AComprehensive Platform for Air Pollution Control System Operation in SmartCities of Developing Countries: A Case Study of Tehran. Environmental IndustryLetters. 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-correlationof climate change, air pollution and urban sustainability using novel machinelearning 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 airpollution, 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 networkpredictions of pollutant emissions from open burning of crop residues: Applicationto 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 predictionusing an effective hybrid deep learning model, (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 indexbased 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 basedon 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 usingdeep 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 InterpretabilityFiltering 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).

Chapter 7: APPENDICES

Appendix: Experimental Details

In this appendix, we give thorough information about the experimental setup and technique used in the study to forecast the Air Quality Index (AQI) and temperature in this appendix. The goal was to create reliable prediction models utilizing AR-Net and the Temporal Fusion Transformer (TFT) based on a multivariate approach. The models were trained and tested using an open-source dataset of Indian air quality..

Dataset:

The dataset used in this study consists of historical air quality measurements and meteorological data from various locations in India. The dataset includes variables such as temperature, humidity, wind speed, and pollutant concentrations, including PM2.5, PM10, SO2, and NO2. The data spanned a specific time period, and it was preprocessed to ensure data quality and consistency.

Data Preprocessing:

Prior to model training, several preprocessing steps were performed on the dataset:

Missing data imputation: Missing values in the dataset were filled using appropriate techniques such as interpolation or mean imputation.

Outlier detection: Outliers in the data were identified and handled using robust statistical methods or by removing the outliers.

Normalization: The data was normalized to bring all variables to a comparable scale, ensuring that no variable dominated the others.

Model Architecture:

Two different models were developed—one for predicting AQI and another for predicting temperature. The models used a multivariate approach, taking into account various atmospheric and meteorological variables. The AR-Net(NeuralProphet) and TFT architectures were implemented to capture the temporal dependencies and interactions between different variables.

Training and Evaluation:

The dataset was split into training and testing sets using a specified time cutoff. The training set was used to train the models, and the testing set was used for evaluation. The models were evaluated using several performance metrics, including Mean Absolute Percentage Error (MAPE).

Experimental Setup:

The experiments were conducted on a high-performance computing cluster to leverage parallel processing capabilities. The models were trained using appropriate libraries and frameworks for deep learning, such as TensorFlow or PyTorch.

Future Scope:

The experimental findings show that the suggested multivariate technique based on AR-Net and TFT forecasts AQI and temperature effectively. Accurate forecasts were obtained by combining atmospheric and meteorological factors as input characteristics, such as temperature, humidity, wind speed, and pollution concentrations. The accuracy of the models trained on the open-source dataset of air quality in India was impressive, demonstrating their potential for real-world applications.

Note: The full code implementation and detailed experimental results are available upon request.

Chapter 8: REFLECTION OF THE TEAM MEMBERS ONTHE PROJECT

- Write what you learned as a team:

As a team, we learned that multivariate machine learning approaches can effectively predict air quality dynamics and estimate heatwave occurrences. By analyzing variousenvironmental variables, such as temperature, humidity, wind speed, rainfall for heatwave prediction and pollutant concentrations for AQI prediction, we achieved accurate and timely predictions, contributing to improved air quality management andheatwave preparedness strategies.

- Write what you learned as a member:

As a member of the team, I learned that multivariate machine learning approaches are powerful tools for predicting air quality dynamics and estimating heatwave occurrences. By leveraging a wide range of environmental factors and applying advanced algorithms, we were able to achieve accurate and reliable predictions, whichcan significantly contribute to effective air quality management and heatwave preparedness strategies.

- Write a thoughtful paragraph on strengths and weaknesses of your designprocess:

One strength of our design process in utilizing multivariate machine learning approaches for air quality prediction and heatwave estimation is the comprehensive analysis of various environmental variables, leading to a holistic understanding of thesystem. However, a weakness lies in the potential complexity of the models, which may pose challenges in terms of interpretability and scalability. Continuous refinement and optimization of the models can address this limitation, ensuring a balance between accuracy and simplicity.

Chapter 9: SIMILARITY REPORT

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

ORIGINALITY REPORT

12%	10%	8%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	3%
2	www.bongos.net.au Internet Source	1%
3	deepai.org Internet Source	1%
4	www.frontiersin.org Internet Source	1%
5	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%
6	link.springer.com Internet Source	1%
7	Xu Feng, Tzung-May Fu, Hansen Cao, Heng Tian, Qi Fan, Xiaoyang Chen. "Neural network predictions of pollutant emissions from open	1%

burning of crop residues: application to air quality forecasts in Southern China", Atmospheric Environment, 2019
Publication

8	Submitted to Napier University Student Paper	1 %
9	github.com Internet Source	<1 %
10	www.mdpi.com Internet Source	<1 %
11	Submitted to University of Bolton Student Paper	<1 %
12	ciit.finki.ukim.mk Internet Source	<1 %
13	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 %
14	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 %

15	repositorio.ufc.br Internet Source	<1 %
16	whp66.ok-em.com Internet Source	<1 %
17	Shankar Subramaniam, Naveenkumar Raju, Abbas Ganesan, Nithyaprakash Rajavel et al. "Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review", Sustainability, 2022 Publication	<1 %
18	Submitted to University of Bath Student Paper	<1 %
19	lppm.ub.ac.id Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		