PM2.5 PREDICTION USING FACEBOOK PROPHET, TEMPORAL FUSION TRANSFORMER AND SARIMA MODEL

PHURINAT PIPATTANAJAROENKUL

Graduate School Srinakharinwirot University

2024

การพยากรณ์ค่าฝุ่นละออง PM2.5 ด้วย Facebook Prophet, Temporal Fusion Transformer และ SARIMA Model

มหาวิทยาลัยศรีนครินทรวิโรฒ

ภูริณัฐ พิพัฒนเจริญกุล

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567

ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

PM2.5 PREDICTION USING FACEBOOK PROPHET, TEMPORAL FUSION TRANSFORMER AND SARIMA MODEL

PHURINAT PIPATTANAJAROENKUL

An Master’s Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2024

Copyright of Srinakharinwirot University

THE MASTER'S PROJECT TITLED

PM2.5 PREDICTION USING FACEBOOK PROPHET, TEMPORAL FUSION TRANSFORMER AND SARIMA MODEL

BY

PHURINAT PIPATTANAJAROENKUL

HAS BEEN APPROVED BY THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER OF SCIENCE IN DATA SCIENCE AT SRINAKHARINWIROT UNIVERSITY

(Assoc. Prof. Dr. Chatchai Ekpanyaskul, MD.)

Dean of Graduate School

ORAL DEFENSE COMMITTEE

.... Major-advisor	.... Chair
(Asst. Prof. Dr.Sirisup Laohakiat)	(Asst. Prof. Dr.Chantri Polprasert)
	.... Committee
	(Asst. Prof. Dr.Waraporn Viyanon)

Title	PM2.5 PREDICTION USING FACEBOOK PROPHET, TEMPORAL FUSION TRANSFORMER AND SARIMA MODEL
Author	PHURINAT PIPATTANAJAROENKUL
Degree	MASTER OF SCIENCE
Academic Year	2024
Thesis Advisor	Assistant Professor Dr. Sirisup Laohakiat

Particulate matter 2.5 (PM2.5) has been considered as the main air pollutant in recent years. It not only harms the environment but also causes severe impacts on human health. There are several measures to handle this toxic substance. One of these methods is to predict its concentration in advance using machining models. This technique plays an important role in assisting government agencies and people to prepare for the incoming high air pollution period. This study aimed to investigate the performance of machine learning models in PM2.5 prediction. The air quality dataset of India and Pakistan were used to predict PM2.5 levels using Facebook Prophet, Temporal Fusion Transformer (TFT) and SARIMA model. The results showed that only Facebook Prophet could yield high performance in PM2.5 prediction in the India dataset. Furthermore, it was also indicated that India's holidays did not influence the concentration of PM2.5 in the India dataset. For Pakistan dataset, it was found that the Temporal Fusion Transformer (TFT) demonstrated the lowest prediction errors over other machine learning models. There was a little overfitting sign during its training process due to the limitation of dataset's size. Still, the TFT performed good outcomes, thanks to its ability to leverage other factors to improve prediction efficiency. The results also suggested that SARIMA model was inappropriate for modeling data with high granular seasonality (such as daily data). In conclusion, the Facebook Prophet was the most suitable model for PM2.5 prediction in this study.

Keyword : PM2.5 PREDICTION / FACEBOOK PROPHET / TEMPORAL FUSION TRANSFORMER / SARIMA MODEL

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my advisor, Asst. Prof. Dr. Sirisup Laohakiat for his suggestion, encouragement and support for this research. Furthermore, I also would like to express my appreciation to Asst. Prof. Dr. Chantri Polprasert from Asian Institute of Technology (AIT) for serving as a chairman. In addition, I am thankful to Asst. Prof. Dr. Waraporn Viyanon for serving as a committee. This research could not be finished without their valuable recommendations.

I am thankful to the Department of Computer Science, Faculty of Science, Srinakharinwirot University for the financial support.

PHURINAT PIPATTANAJAROENKUL

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Due to the growth of urbanization and industrialization through the world in recent years, the air pollution has been regarded as a threat causing severe effects on both environment and human life (Jia et al., 2023). According to the World Health Organization (WHO), it has been found that around seven million people got health problems or even died in 2016 because of toxic air exposure (Thangavel et al., 2022). Furthermore, it has been expected that the relationship between air pollution and human illness and death rate might have a continuously increasing tendency in the future (Gu et al., 2023). Among the various harmful components of the polluted air, the particulate matter 2.5 (PM2.5) is the most concern and used as one of the air quality indicators. As a result of its tiny size (less than 2.5 μm), the PM2.5 can reach to the deepest area of human's respiratory tract and it can be transported over the human body through bloodstream (Basith et al., 2022; Li et al., 2016). Moreover, it has been reported that PM2.5 can trigger serious health symptoms for instance, respiratory inflammation, cardiovascular disease, lung cancer, etc. (Miller & Xu, 2018; Nowak et al., 2013). One of several methods to handle the PM2.5 situation is to predict its concentration in advance by efficient mathematical models. This might help the related organizations to initiate the proper protocol or the public to prepare themselves for the upcoming PM2.5 high season (Pant et al., 2017; Y. Zhang et al., 2024).

In this study, air quality of India during November 2017 to June 2022 and Gujrat, Pakistan during February 2022 to June 2024 were used to predict the PM2.5 concentration by using Facebook Prophet, Temporal Fusion Transformer (TFT) and SARIMA model.

1.2 Objectives

1. To collect and preprocess relevant PM2.5 concentration data from various sources, ensuring completeness and accuracy for effective model training.

2. To implement and compare multiple machine learning models for PM2.5 concentration prediction.

1.3 Hypotheses

1. Machine learning models can effectively predict PM2.5 concentrations based on historical environmental, meteorological, and other relevant data, with certain models outperforming others in terms of accuracy, generalizability, and computational efficiency.

2. Hyperparameter tuning and feature selection will lead to substantial performance gains, particularly for models with higher complexity.

1.4 Scope of the Study

1. Air quality dataset of India during November 2017 to June 2022 and Gujrat, Pakistan during February 2022 to June 2024 will be used to predict the PM2.5 concentration.

2. Various machine learning models will be used and compared to find the most suitable method for the datasets.

1.5 Expected Outcome

1. The results from the most suitable model provide useful insights of the future air quality condition.

CHAPTER 2

THEORETICAL BACKGROUND AND LITERATURE REVIEW

2.1 Particulate matter 2.5 (PM2.5)

Air pollution has been considered as a global issue for many years. There is a lot of research concluded that it not only causes environmental problems but also affects the human health (Chirasophon & Pochanart, 2020). In addition, it was found that air pollution is one of the main factors causing public illness around the world (Jia et al., 2017). The polluted air is composed of several toxic substances such as carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NOx), volatile organic compounds (VOCs), ozone (O_3), heavy metals and particulate matter also known as PM (Kampa & Castanas, 2008; Li et al., 2022).

Based on the particle size, the particulate matter can be categorized into three main types namely PM10 ($\leq$ 10 μm) also called coarse particle, PM2.5 ($\leq$ 2.5 μm) also called fine particle and PM0.1 ($\leq$ 0.1 μm) also called ultrafine particle, respectively. Especially PM2.5, which has been used as one of the air quality indicators, seems to be the major problem at the present time. The reasons why PM2.5 has been more concerned than the other particulate matters can be described as follows 1) it is abundant in the air. 2) it has long residence time in the atmosphere and 3) it can be transported over long distances (Basith et al., 2022; Li et al., 2022). The source of PM2.5 can be both naturogenic and anthropogenic origins. For naturogenic origin, it has been found that PM2.5 can be produced by dust storms, sea spray aerosols, forest fires, volcanic activities, etc. For anthropogenic origin, PM2.5 can be emitted to the atmosphere by several human activities for example, transportation, fossil fuel combustion, industrial processes, residential cooking, cigarette smoking, etc. (Garcia et al., 2023; Kollanus et al., 2017).

As mentioned previously, PM2.5 has many impacts on human life. There was a study in island of Taihu Lake, in Suzhou City, China aimed to find the relationship between PM2.5 and PM10 exposure and daily traffic accidents by using four different mathematical models. The results showed that as the concentrations of PM2.5 and PM10 increased, the traffic accident rate also increased. These pollutants might limit the visibility on the road and had some effects on the driver’s decision abilities (Wan et al., 2020). Furthermore, it has been reported that the particulate matter problem in the northern part of Thailand, particularly in Chiang Mai which has affected the number of tourist arrivals. Because Thailand is the popular destination of tourists around the world, the decreasing of tourists and tourism businesses has inevitably influenced on Thailand’s economy (Srinamphon et al., 2022). In terms of public health, it has been observed that PM2.5 can deeply accumulate in human’ lung resulting from its miniature size. Moreover, it can be carried to other organs including brain, heart, kidney, liver, etc. via blood vessels (Garcia et al., 2023; Jia et al., 2017). Many studies have shown that the exposure of PM2.5 can cause several serious symptoms or increase the risk of some diseases for instance, respiratory tract inflammation, asthma, lung cancer, heart disease, Alzheimer’s disease, chronic kidney disease, etc. Besides, there was evidence that pregnant women who have exposed PM2.5 have a tendency of abnormal childbirth such as preterm birth or low birth weight (Li et al., 2022; Thangavel et al., 2022).

Due to the countless impacts which PM2.5 can cause, the advanced PM2.5 concentration forecasting is important. It has several benefits for example, the government or policymaker can propose suitable mitigation measures and people can prepare themselves for high PM2.5 concentration seasons which help to reduce the risk of PM2.5 exposure (Geng et al., 2021; Y. Zhang et al., 2024). However, the prediction of PM2.5 concentration in advance might be a difficult task because the concentration of PM2.5 has a characteristic called seasonal variation (Ao et al., 2019). It has been reported that the PM2.5 in New Dehi was abundant during winter and monsoon days, while it was higher than the standard during January to March in Thailand (Chirasophon & Pochanart, 2020; Sahu & Kota, 2017). Therefore, the mathematical models which have ability to handle this seasonal behavior such as Facebook Prophet, Temporal Fusion Transformer (TFT) and Seasonal Autoregressive Integrated Moving Average model (SARIMA) might play an important role in PM2.5 prediction (Wang & Guo, 2009).

2.2 Facebook Prophet

Facebook Prophet is a model specially designed for time series forecasting by Facebook data science department. It is a generalized additive model which aims to break down time series data into three main components for example, trend, seasonality and holidays. Each separated part is modelled individually, after that, they are combined together to perform the final prediction (Taylor & Letham, 2017). It is suitable for the time series showing strong seasonal behavior. In addition, it does not need a large amount of data to understand the data's characteristics for prediction as other time series forecasting techniques. The Prophet model is capable of handling missing data, outliers and trend variability without requiring exogenous interaction from user. With all the advantages mentioned above make the Facebook Prophet's fitting process fast and attributes the model very user-friendly (Cai, 2023; Zhao et al., 2018).

The general equation of Facebook Prophet can be explained by a below equation (Duarte & Faerman, 2019).

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \qquad (1)$$

$y(t)$ = The forecasted value at time *t*

$g(t)$ = The trend function which analyzes non-periodic change in time series, the trend can be captured in two ways: linear trend (Piece-wise linear model) and non-linear trend (Logistic growth model).

$s(t)$ = The seasonal function which considers periodic change in time series such as daily, weekly, yearly, etc. by leveraging the Fourier' series.

$h(t)$ = The effect of holidays or the special events which take place in an irregular schedule.

$\varepsilon(t)$ = The error term which is anticipated to display the normal distribution.

It was found that the Prophet outperformed SARIMA model in air pollution forecasting of Bhubaneswar city, India. The result demonstrated that the accuracy metrics (MSE, RMSE) of the Prophet on predictions of various pollutants such as SPM, NO_2, SO_2, RSPM, etc. were lower than SARIMA's. This indicated that the Prophet had the better performance in air pollution prediction (Samal et al., 2019). Moreover, it has been reported that the Prophet was used to predict the temperature in Myintkyina, Myanmar. The daily temperature data during 2010-2017 was brought to fit the model and make the prediction. The result showed that both predicted value and actual value were close to each other and the RMSE is only 5.7573. This outcome emphasized that the model could capture the patterns of the data and provide good prediction. Therefore, the Prophet model could be considered as an another option for weather forecasting (Oo & Phyu, 2020).

2.3 Temporal Fusion Transformer (TFT)

Temporal Fusion Transformer is a novel mathematical model for time series forecasting. By leveraging the concepts of RNN's recurrence and transformer's attention mechanism together, it can capture both short term and long-term dependencies of time series data resulting in high accuracy prediction outcomes. The general equation of Temporal Fusion Transformer can be explained by a below equation (Lim et al., 2021; Liu et al., 2024).

$$\hat{y}_{i(q,t,\tau)} = f_{q}(\tau, y_{i,t-k:t}, z_{i,t-k:t}, x_{i,t-k:t+\tau}, s_i) \quad (2)$$

$\hat{y}_{i(q,t,\tau)}$ = The predicted target value for the i-th time series at time *t* for quantile *q* and forecast horizon τ

f_q = The function used to compute the quantile forecast

q = The quantile

t = The current timestep

τ = The forecast horizon

k = The lookback window (historical window size)

$y_{i,t-k:t}$ = The historical target values for the time series from *t-k* to time *t*

$z_{i,t-k:t}$ = The unknown inputs (past inputs/past covariates) for the i-th time series from *t-k* to time *t*

$x_{i,t-k:t+\tau}$ = The known inputs (future inputs/future covariates) for the i-th time series from *t-k* to time *t* + τ

s_i = The static covariates for the i-th time series

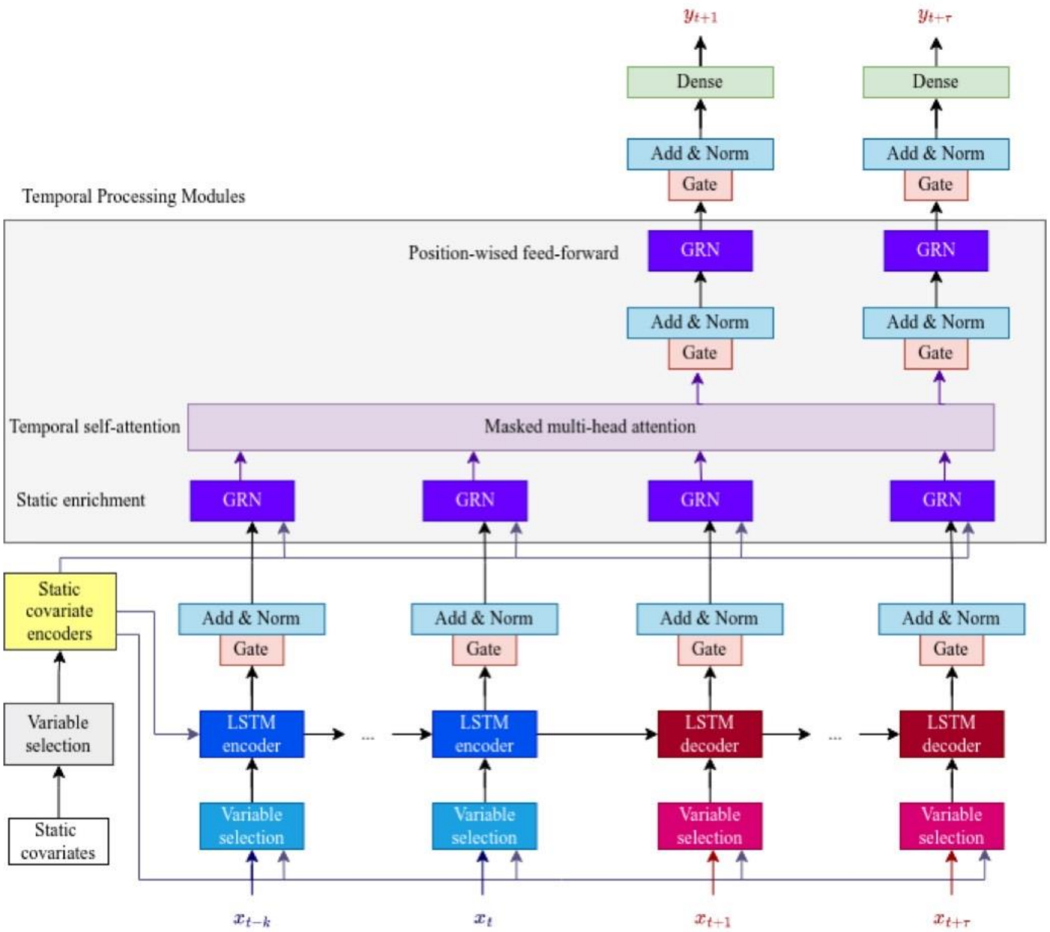

Figure 1. The architecture of Temporal Fusion Transformer (Fayer et al., 2023)

From the architecture of Temporal Fusion Transformer shown in Fig. 1., the main components are described upwardly. Firstly, the input variables including unknown inputs (past inputs / past covariates), known inputs (future inputs / future covariates) and static covariates are passed through the Variable Selection Network (VSN). This step plays an important role in selecting only important features and filter out features which are unnecessary to the prediction. Secondly, the selected unknown inputs and known inputs go to the LSTM's encoder and decoder respectively. The encoder aims to learn the temporal relationships and patterns of the unknown inputs while the decoder focuses on the target prediction by considering the context from the encoder along with the known inputs. The static covariates go to the static covariate encoder which provides the static context. This context helps the Temporal Fusion Transformer to yield the predicted outputs with high accuracy by leveraging both dynamic and stable features. Thirdly, all processed information is fed to the Gated Residual Network (GRN). With various mechanisms such as gating mechanism, residual connection and normalization, this network ensures that only the most important information can pass through its layers. Furthermore, this process helps the model to increase its learning ability and stabilize the training operation. Fourthly, the information then flows to the Temporal Multi-head Attention layers. This part is the crucial key of the Temporal Fusion Transformer which helps the model capture the long-term dependencies and complex patterns of the time series data. This technique also deals with the weakness of the LSTM which cannot understand the long-term relationship of long sequence data. Therefore, with the combination of attention mechanism and LSTM to focus on the long-term and short-term dependencies respectively, making the Temporal Fusion Transformer is an effective tool in the time series analysis. Finally, the outputs from attention layers reach the dense layer to perform the quantile forecasts. The quantile forecasts predict the various quantiles as a range of possible future target values. This is helpful to understand the uncertainty and variability of the predictions (Deforce et al., 2022; Fayer et al., 2023; Rasiya Koya & Roy, 2024; Stefenon et al., 2024).

It has been reported that the airport delay of top 30 airports in the US was analyzed by using Temporal Fusion Transformer. In this study, airport ID, month, day of the weeks were us as static covariates. Historical information such as count of actual arrival and departure, averaged arrival and departure delays, cumulative of arrival and departure delays and count of on-time arrival and departure were used as past covariates, while scheduled airport demand and capacity along with the weather and air traffic conditions were used as future covariates. The prediction result is satisfied with low MAE value. In addition, the model also highlighted the important features influencing the prediction which helped for better understanding of model's interpretability (Liu et al., 2024). Another example of Temporal Fusion Transformer usage was the prediction of dam level in hydroelectric power plants. Many machine learning techniques for instance, Adaptive Neuro-Fuzzy Inference System (ANFIS), Long-Short Term Memory (LSTM), Ensemble method and Temporal Fusion Transformer were used to analyze time series data and compare their forecasting performance. The result revealed that Temporal Fusion Transformer gave the lowest RMSE among all machine learning models. From the outcome, it seemed that Temporal Fusion Transformer might be a good assistant in operational management of hydroelectric power plants (Stefenon et al., 2024).

2.4 SARIMA Model

SARIMA model is an extension of ARIMA model which is designed to deal with seasonal time series. Although it is a modified version, it still has the same core concept as ARIMA model's which is forecasting the future values based on the patterns including trend or seasonality of historical data with the expectation that these characteristics might happen in upcoming time periods. Furthermore, its concept is based on some important assumptions for instance, the time series needs to be stationary and the input data must be a univariate series (Box et al., 2016; Doreswamy et al., 2020).

The general equation of SARIMA model can be explained by a below equation (Momani, 2009).

$$(1 - \phi_p B)(1 - \Phi_P B^s)(1 - B)^d (1 - B^s)^D X_t = (1 - \theta_q B)(1 - \gamma_Q B^s)\epsilon_t \quad (3)$$

φ_p = Nonseasonal autoregressive coefficient of order p.

Φ_P = Seasonal autoregressive coefficient of order P with the length of seasonal period s.

θ_q = Nonseasonal moving average coefficient of order P.

γ_Q = Seasonal moving average coefficient of order Q with the length of seasonal period s.

B = Backshift operator ($BX_t = X_{t-1}$).

B^s = Seasonal backshift operator ($B^sX_t = X_{t-s}$).

X_t = The predicted value at current time period.

d = Nonseasonal differencing order.

D = Seasonal differencing order.

ϵ_t = The error term.

s = The length of seasonal period.

From the equation, it seems the SARIMA model is composed of both nonseasonal and seasonal components. Therefore, it is usually easily expressed as ARIMA (p, d, q) (P, D, Q) s. The p, d and q refer to nonseasonal autoregressive (AR), differencing and moving average (MA) orders, respectively while, the P, D and Q refer to seasonal autoregressive (AR), differencing and moving average (MA) orders with the length of seasonal period s, respectively (Bhatti et al., 2021).

There are a lot of applications of SARIMA model in various fields. For example, it has been reported that SARIMA model was used to forecast the monthly flows in Waterval river, South Africa. The best model was ARIMA (3,0,2) $(3,1,3)_{12}$ and its result was similar to the measured monthly flows. This result was helpful information in water resource management (Tadesse et al., 2023). Moreover, it was found that the SARIMA model could predict the precise number of dengue cases in Brazil by using the previous data of one, two and twelve months before. The result might give the better view of the disease mechanism and help to establish the appropriate measures (Martinez et al., 2011). In 2009, the PM2.5 data in Los Angeles from January 1999 – December 2006 was analyzed by SARIMA model and forecast PM2.5 concentration in the upcoming years. The SARIMA model's result was close to the original data. Therefore, it was concluded that SARIMA model gave good performance in PM2.5 prediction in this study (Wang & Guo, 2009).

CHAPTER 3

METHODOLOGY

3.1 Experimental Framework

The experimental framework was shown in Fig. 2. In this study, the experiment could be divided into two main parts: 1. Data acquisition and data preprocessing 2. Model implementation. All details were described in the following sections.

มหาวิทยาลัยศรีนครินทรวิโรฒ

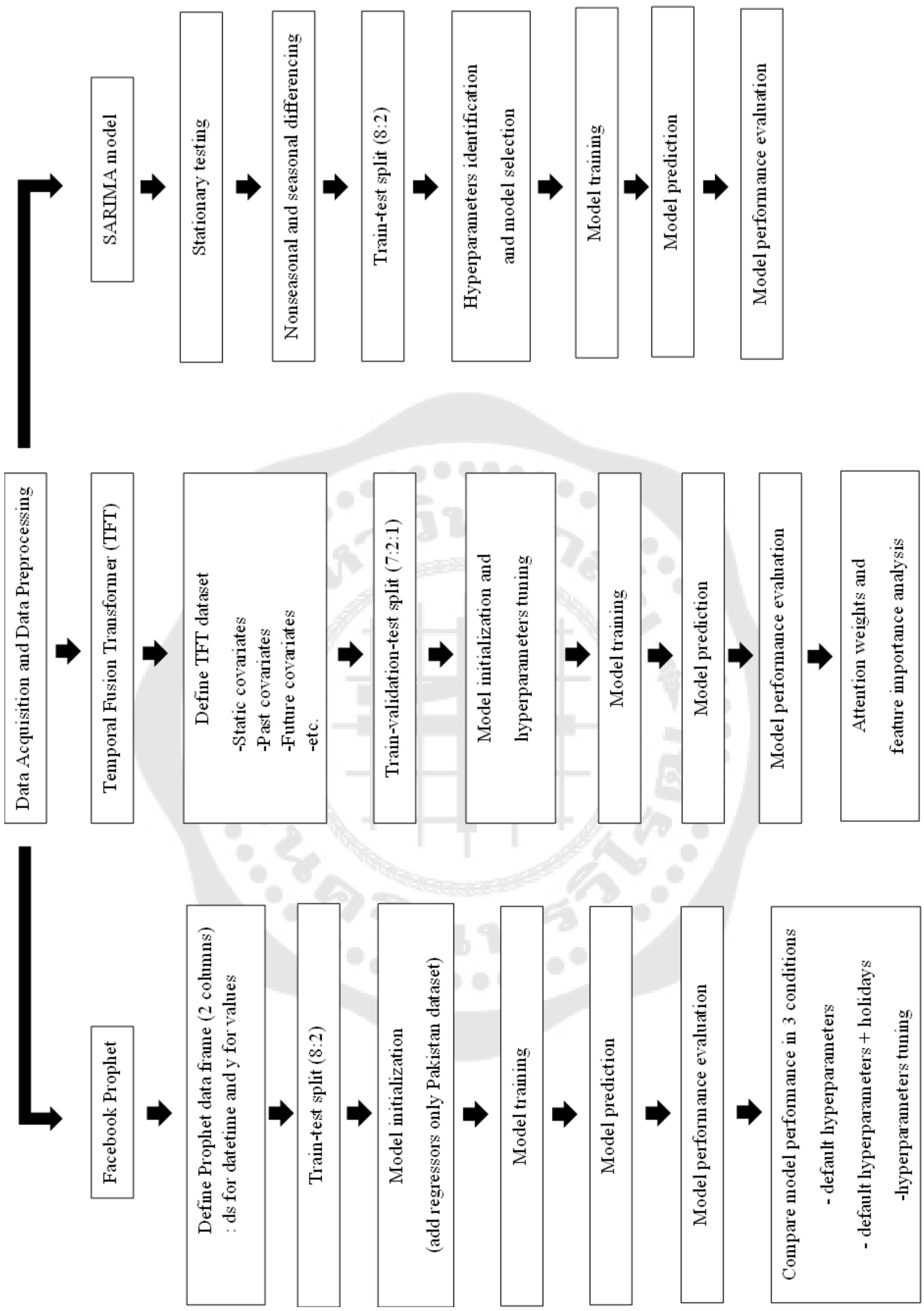

Figure 2. Experimental framework

3.2 Data acquisition and Data preprocessing

There were two datasets used in this study: 1) the air quality dataset of India during November 2017 to June 2022. 2) the air quality dataset of Gujrat, Pakistan during February 2022 to June 2024. They were downloaded from www.kaggle.com. Before any analytical procedures, both datasets were preprocessed through an Explanatory Data Analysis (EDA) such as missing values investigation, down-sampling (hourly to daily), variables' relationships and data's statistical properties investigation etc.

3.3 Model Implementation

3.3.1 Facebook Prophet

For PM2.5 prediction using the Facebook Prophet, the time series data needed to be prepared as two-column data frame ('ds' for datetime and 'y' for target value) which aligned to the model's requiring format. After that, the data were allocated into training set and test set in an 8:2 proportion. The Prophet model was initialized and fitted with the training data. Then, cross validation was performed to assess how well the model learned. Finally, the test set was used to evaluate the model's generalization ability via metrics: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

The PM2.5 prediction with Facebook Prophet was examined in three conditions. Condition 1: Prophet model with default hyperparameters, Condition 2: Prophet model with default hyperparameters and holidays components and Condition 3: Prophet model with hyperparameters tuning using Optuna, a hyperparameters optimization framework.

3.3.2 Temporal Fusion Transformer

In Temporal Fusion Transformer (TFT), the time series data were split into training set, validation set and test set in a ratio of 7:2:1, respectively. All required information such as static covariates, past covariates and future covariates, etc. were identified. The training data were fitted to the model and all hyperparameters were finely tuned using Optuna. After that, the model's learning performance was determined by cross validation technique. Then, the test set was compared with the predicted results to estimate model's efficiency via Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). In addition, the attention weights and feature importances were analyzed to understand how the TFT makes predictions of PM2.5.

3.3.3 SARIMA model

The time series data were tested their stationarities by augmented Dickey-Fuller test (ADF test) prior SARIMA model establishment. To acquire the stationarity purpose, the nonseasonal differencing and seasonal differencing techniques were applied. Once the data were already stationary, the train-test split was performed with 80% for training set and 20% for test set. The SARIMA model's parameters: p, d, q, P, D, Q and s were identified. The Akaike Information Criterion (AIC) was used to determine the best SARIMA model. The performance of the best SARIMA model was measured by Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE).

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data description

4.1.1 India dataset

The India dataset was a daily average PM2.5 concentration. It was composed of 1,616 rows and 2 columns such as date and avg_PM2.5. The data approximately covered around 4 years and 7 months and started from 7th November 2017 to 4th June 2022. According to statistical information, the maximum, minimum and average values of PM2.5 concentration were 120.35, 12.64 and 48.82 µg/m^3, respectively. There were not any missing values in the dataset.

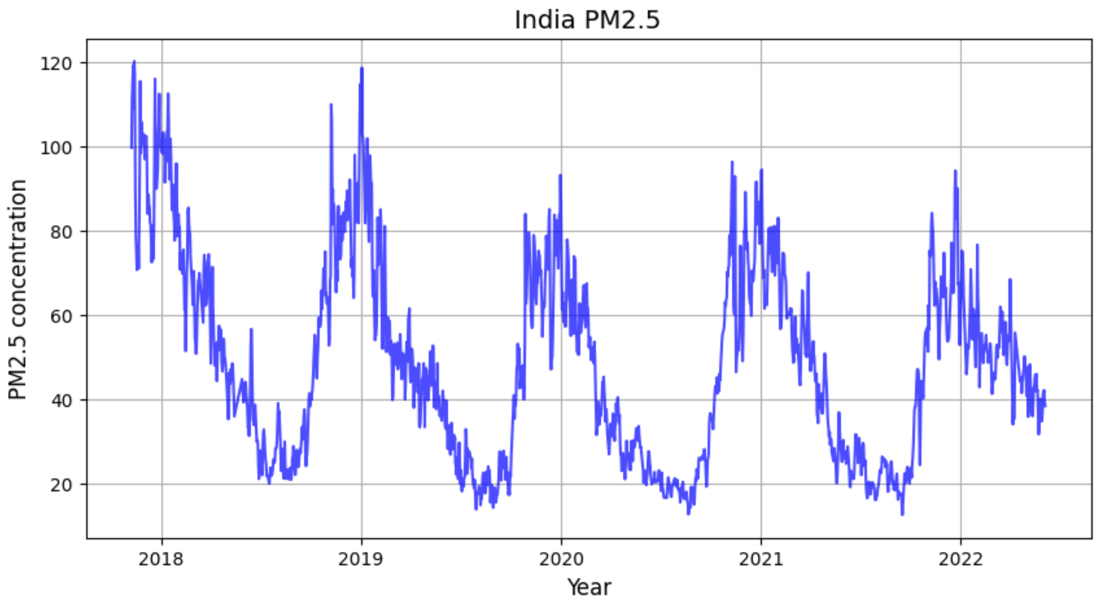

Figure 3. Daily average PM2.5 concentration of India

From the above diagram (Fig. 3.), It seemed the PM 2.5 levels rose from the middle to the end of one year, then fell from the beginning to the middle of the subsequent year. This seasonality relates to environmental and human activities which occur in the same period each year. For example, the monsoon season in the middle of the year brings strong winds that disperse air pollutants, leading to lower PM2.5

concentrations. At the end of the year, the winter season sets in. Due to low temperatures and reduced wind speed, air pollutants become trapped near the Earth's surface, leading to elevated PM2.5 levels. Additionally, this period coincides with the post-harvest burning of agricultural residue, which further contributes to increased PM2.5 concentrations. According to the diagram, PM2.5 levels appeared to follow a consistent, repeating pattern each year and are likely to continue this trend in the future. Therefore, the PM2.5 of India was considered as a representative of time series with strong and clear seasonality pattern in this study.

4.1.2 Pakistan dataset

The Pakistan dataset was also the daily average PM2.5 concentration. This dataset contained 874 rows of data with 9 columns including date, avg_temp, avg_aqi, avg_CO, avg_NO_2, avg_O_3, avg_PM 10, avg_SO_2 and avg_PM2.5. The date of sampling started from 1st February 2022 to 23rd June 2024 which covered around 2 years and 4 months. The maximum, minimum and average PM2.5 concentration were 355.89, 15.96 and 75.69 μg/m^3, respectively. The missing values were not found in the dataset.

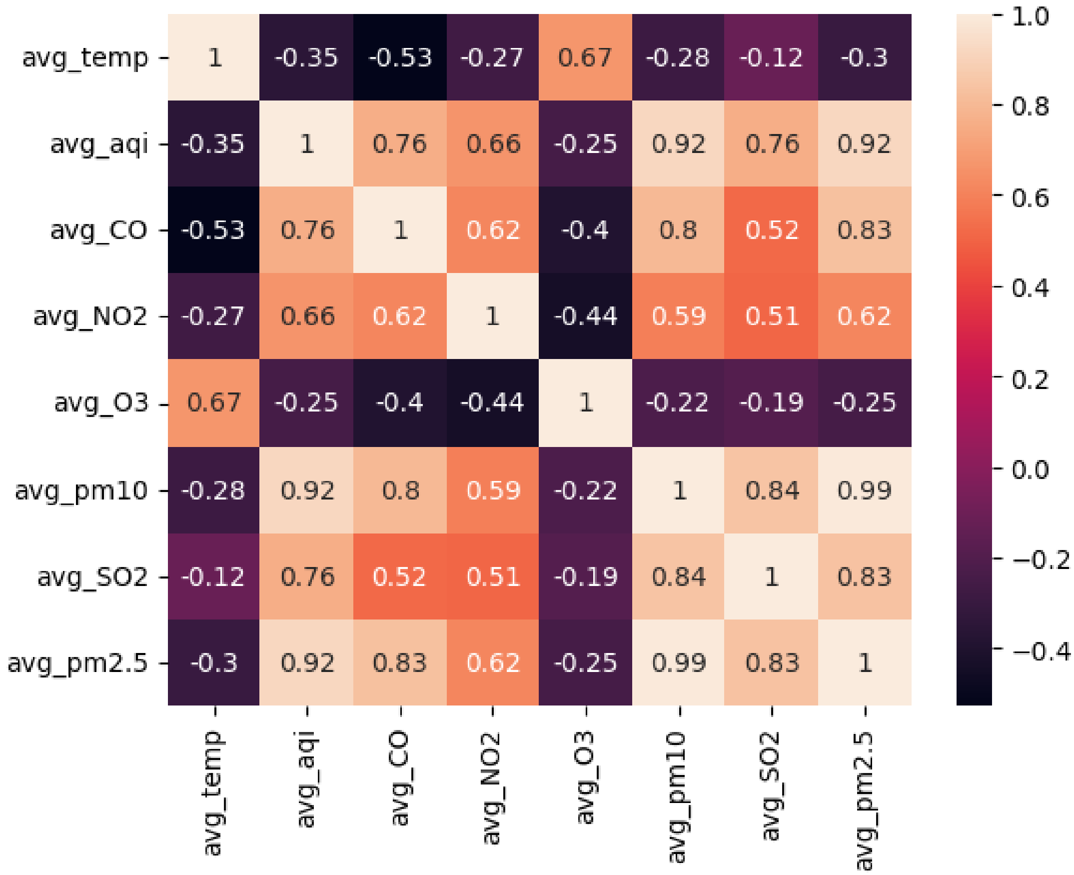

Figure 4. Heat map of Pakistan dataset' variables

The heat map was created to explore the relationships between PM2.5 concentration and other variables (Fig. 4.). According to the values of Pearson' correlation coefficients (Basumatary et al., 2023), it was found that avg_aqi, avg_CO, avg_PM10 and avg_SO_2 showed the very strong positive relationship with avg_PM2.5. The avg_NO_2 also strongly correlated with avg_PM2.5. On the contrary, avg_temp and avg_O_3 demonstrated the weak negative relationship with avg_PM2.5. Compared with the real-world air quality patterns can help better understand the interactions between PM2.5 and these environmental factors. PM2.5 in the atmosphere can be divided into primary and secondary. The primary PM2.5 is directly released from sources such as vehicle exhaust, industrial emission, biomass burning and dust. Moreover, PM 10, SO_2, NO_2 and CO are often emitted alongside PM2.5. While the secondary PM2.5 is generated through the chemical reaction of the primary precursors particularly SO_2, NO_2 and CO. This might be a reason that the concentrations of PM2.5 and these pollutants positively relate to each other (Basith et al., 2022; Xie et al., 2015). Temperature expressed a negative trend with PM2.5. Warm temperature affects the air to circulate which results the decrease of PM2.5 concentration. Conversely, low temperature contributes to air stagnation, weak wind speed, trapping pollutants in the atmosphere and worsening air quality (Vaishali et al., 2023; Yang et al., 2017). For the Ozone (O_3), the inverse relationship between O_3 and PM2.5 might be described as the small O_3 formation caused by PM2.5. With high amount of PM2.5 in the atmosphere, it can absorb or block the sunlight which plays an important role in O_3 generation (Zhu et al., 2019).

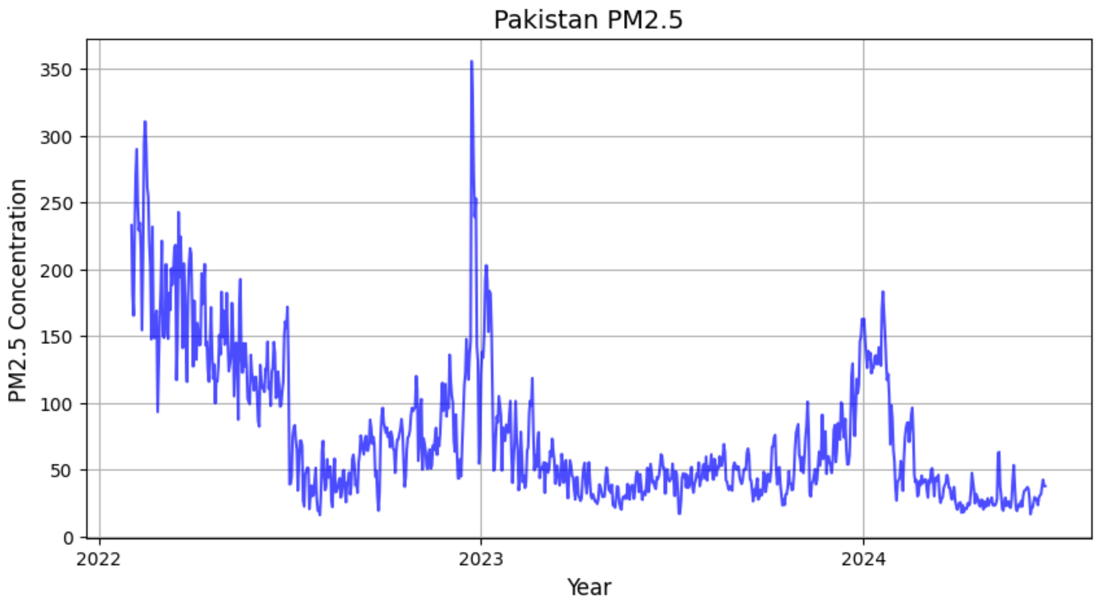

Figure 5. Daily average PM2.5 concentration of Pakistan

Similar as India dataset, the Pakistan' PM2.5 concentration also showed the seasonal pattern influenced by the real-world factors. However, the PM2.5 level of Pakistan dataset did not display the uniform pattern. For instance, the characteristics of the graph from 2022 to 2023 and from 2023 to 2024 were obviously different from each other. Hence, the PM2.5 of Pakistan was considered as a representative of time series with complex and unclear seasonality pattern in this study.

4.2 Facebook Prophet with India dataset

The Prophet model was tested on India dataset with three conditions as mentioned previously to find the best condition for this dataset. The details of hyperparameters of each condition were shown in Table 1. Furthermore, the holidays (special dates or annual events) of India which might influence the trend and seasonality of PM2.5 concentration were added to the model represented as condition 2: default model with holidays components (Daraghmeh et al., 2021). The India holidays were displayed in Table 2.

Table 1. Hyperparameters' details of each condition on India dataset

Hyperparameter	Default model (condition 1)	Default model with holiday components (condition 2)	Tuned model (condition 3)
changepoint_prior_scale	0.05	0.05	0.13
seasonality_prior_scale	10	10	0.18
seasonality_mode	additive	additive	multiplicative

Table 2. India holidays

Date	Holidays
January 26 (fixed every year)	Republic day
August 15 (fixed every year)	Independence Day
October 2 (fixed every year)	Gandhi Jayanti
Varies each year	Diwali
Varies each year	Holi
Varies each year	Eid al-Fitr
Varies each year	Eid al-Adha
December 25 (fixed every year)	Christmas Day
Varies each year	Maha Shivaratri
Varies each year	Good Friday

All Prophet models were initially trained on 180 days of the dataset. Every 90 days, the forecast evaluation was performed by predicting the next 30 days. This cross-validation technique helps in model's generalization ability assessment to new or unseen data over different time periods and can aid in overfitting detection (Lico et al., 2021). The Prediction results of Prophet models on India dataset under all conditions were illustrated in Fig. 6., Fig. 7., Fig. 8. and the models' performance metrics were demonstrated in Table 3.

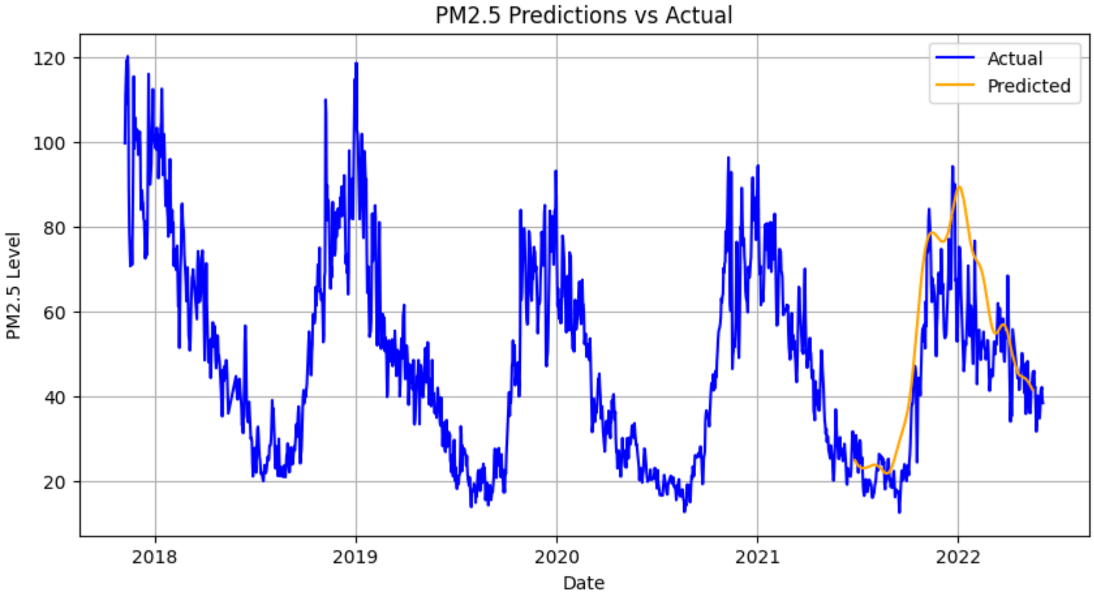

Figure 6. Prediction result of Prophet model on India dataset with default hyperparameters (condition 1)

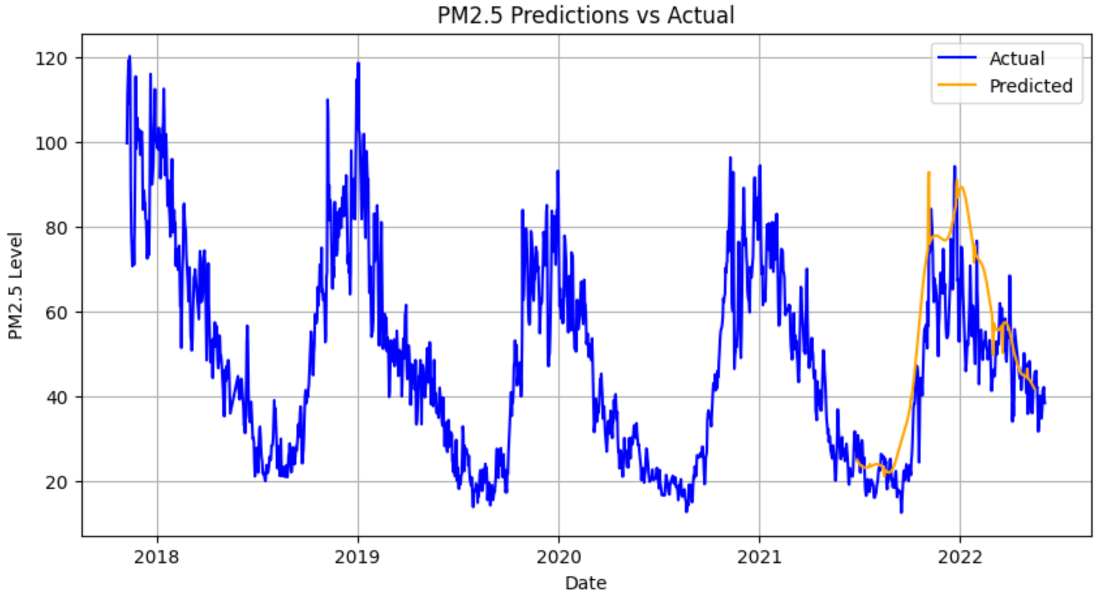

Figure 7. Prediction result of Prophet model on India dataset with default hyperparameters and holidays components (condition 2)

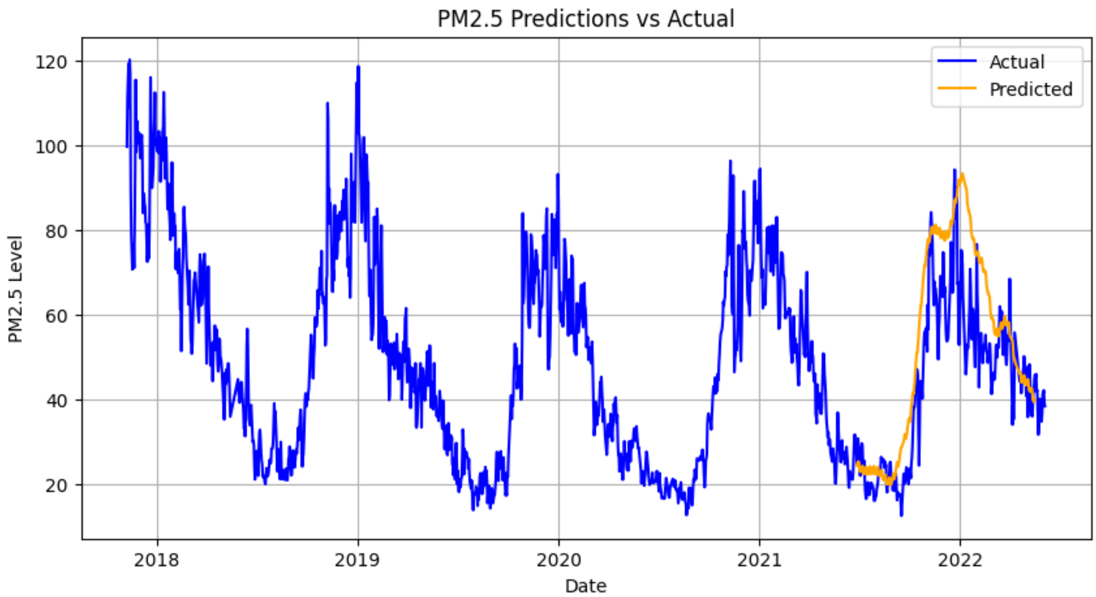

Figure 8. Prediction result of Prophet model on India dataset with hyperparameters tuning using Optuna (condition 3)

Table 3. Performance metrics of Prophet models on India dataset under all conditions

Metrics	Default model (condition 1)	Default model with holiday components (condition 2)	Tuned model (condition 3)
MAE	10.09	10.21	10.81
MSE	168.58	173.05	205.59
RMSE	12.98	13.15	14.34
MAPE	25.03	25.37	25.51

After analyzing prediction plots and performance metrics, it was found that the Prophet model with default hyperparameters (condition 1) yielded the best performance with the lowest values of MAE, MSE and RMSE followed by the prophet model with default hyperparameters and holiday components (condition 2) and the Prophet model with hyperparameters parameters tuning with Optuna (condition 3), respectively. The metrics values of all conditions remained at an acceptable level which were consistent with the alignment of predicted and actual values of PM2.5 concentrations in the diagrams. The lower values indicates higher accuracy in model prediction (Khanal et al., 2023). Furthermore, the similarity of metrics values between condition 1 and condition 2 implied that the India's holidays did not have any significant influence on trend or seasonality of the PM2.5 concentration in this dataset (Cai, 2023; Samal et al., 2019). Therefore, it could be concluded that the Prophet model with default hyperparameters (condition 1) was the best condition for the India dataset.

4.3 Facebook Prophet with Pakistan dataset

The Prophet model was also used to test on Pakistan dataset with three conditions as those in the India dataset. Since the Pakistan dataset contained many columns apart from date and avg_PM2.5 which were main required inputs, other variables such as avg_temp, avg_aqi, avg_CO, avg_NO_2, avg_O_3, avg_PM10 and avg_SO_2 were used as additional regressors or extra information that might improve the prediction performance of the Prophet (Huang, 2022). The hyperparameters' details and Pakistan holidays were described in Tables 4. and Table 5.

Table 4. Hyperparameters' details of each condition on Pakistan dataset

Hyperparameter	Default model (condition 1)	Default model with holiday components (condition 2)	Tuned model (condition 3)
changepoint_prior_scale	0.05	0.05	0.004
seasonality_prior_scale	10	10	0.01
seasonality_mode	additive	additive	multiplicative

Table 5. Pakistan holidays

Date	Holidays
February 5 (fixed every year)	Kashmir Solidarity Day
March 23 (fixed every year)	Pakistan Day
May 1 (fixed every year)	Labor Day
Varies each year	Eid ul-Fitr
Varies each year	Eid ul-Adha
August 14 (fixed every year)	Independence Day
Varies each year	Ashura
Varies each year	Eid Milad-un-Nabi
December 25 (fixed every year)	Quaid-e-Azam Day

The cross-validation procedure was also performed to the Pakistan dataset with the same configurations as those used in the India dataset to evaluate the learning performance of Prophet model in all three conditions. The predicted PM2.5 concentrations of the Prophet models compared with the actual values of the Pakistan dataset were visualized by Fig 9., Fig 10., Fig 11. and the models' performance metrics were shown in Table 6.

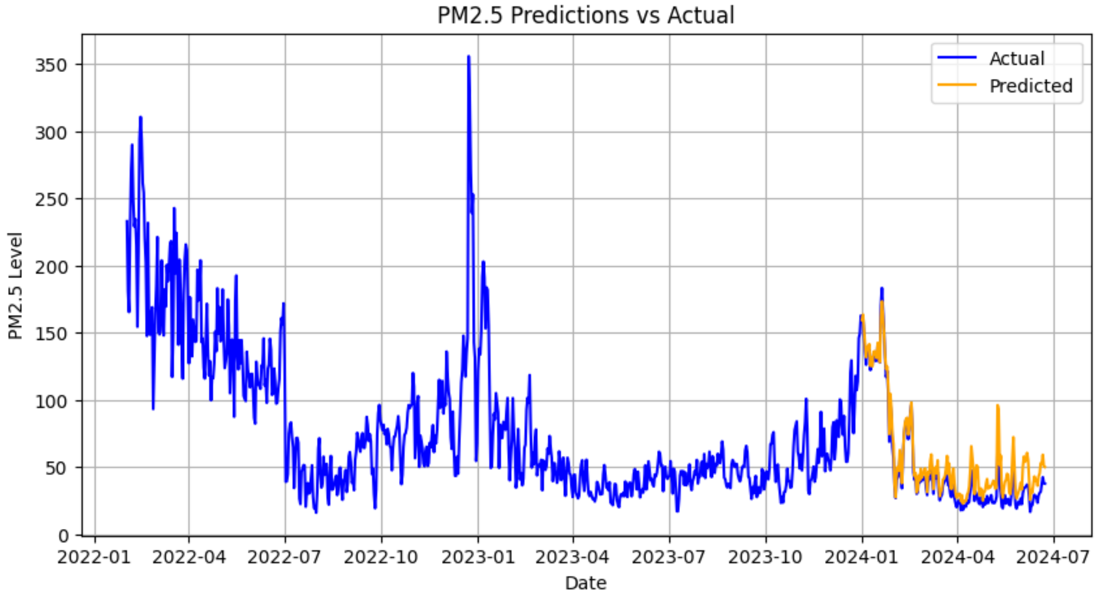

Figure 9. Prediction result of Prophet model on Pakistan dataset with default hyperparameters (condition 1)

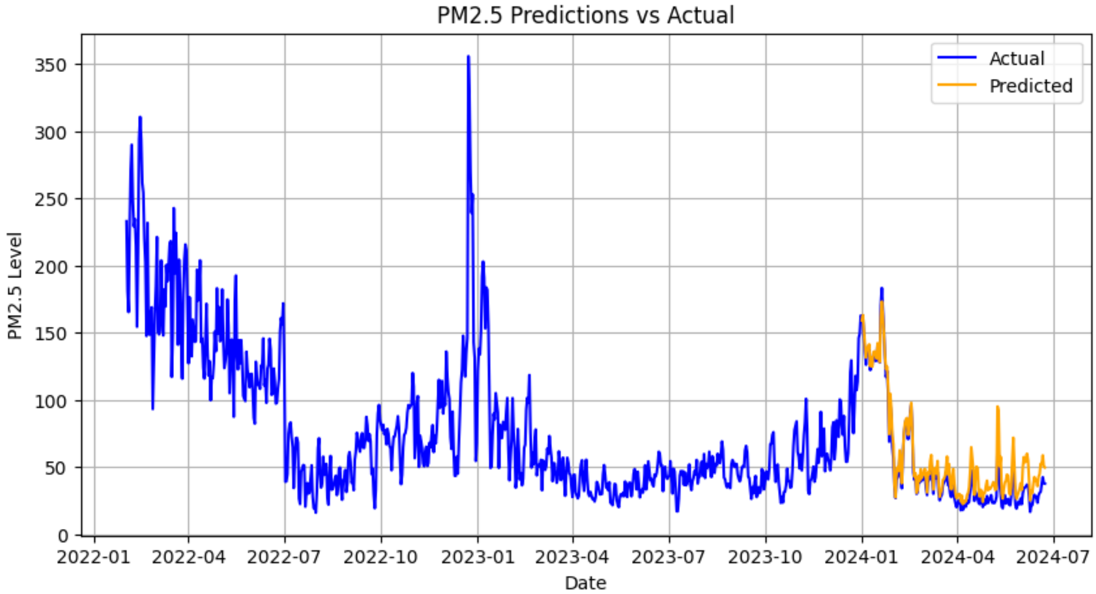

Figure 10. Prediction result of Prophet model on Pakistan dataset with default hyperparameters and holidays components (condition 2)

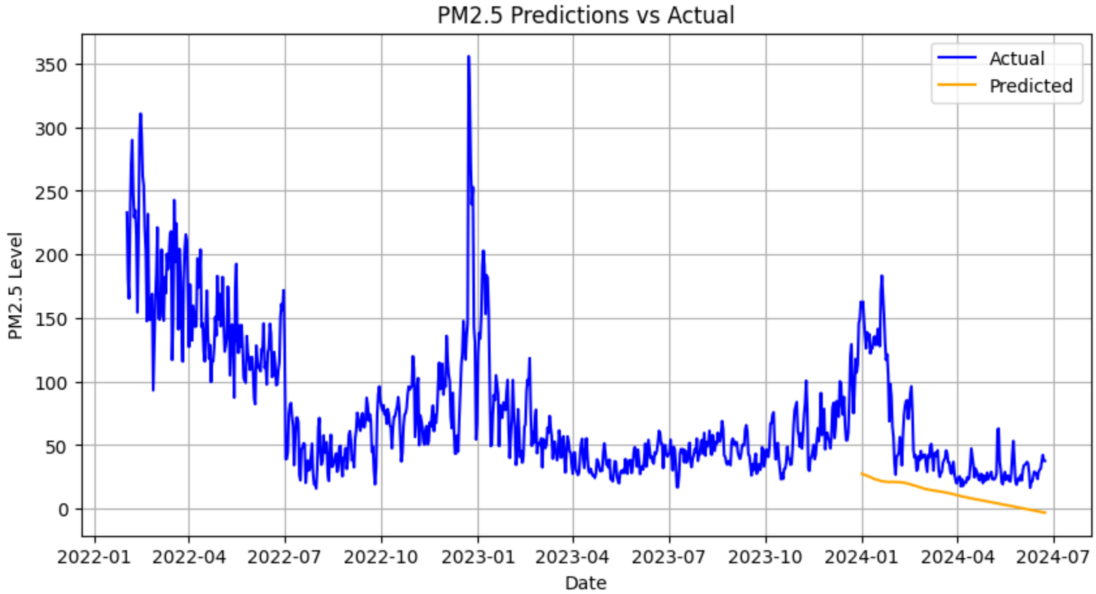

Figure 11. Prediction result of Prophet model on Pakistan dataset with hyperparameters tuning using Optuna (condition 3)

Table 6. Performance metrics of Prophet models on Pakistan dataset under all conditions

Metrics	Default model (condition 1)	Default model with holiday components (condition 2)	Tuned model (condition 3)
MAE	8.26	7.86	40.25
MSE	107.01	98.71	2803.27
RMSE	10.35	9.94	52.95
MAPE	25.69	24.34	76.02

Based on the results, the Prophet model with default hyperparameters and holidays components (condition 2) exhibited the higher accuracy in predicting PM2.5 concentrations than other two conditions considered by the lowest values of all metrics. However, with slightly better MAE, MSE, and RMSE compared to those on Condition 1, this suggested that the Pakistan’s holidays might have some little effects to the PM2.5

concentrations in this dataset. For the third condition, it seemed the tuned hyperparameters might not be appropriate to the dataset reflected by higher values of all metrics. The defined search space may have been too broad or too limited, preventing the model from finding optimal values and leading to poor results (Setianingrum et al., 2022). Therefore, it could be concluded that the Prophet model with default hyperparameters and holidays components (condition 2) was the best condition for the Pakistan dataset.

4.4 Temporal Fusion Transformer (TFT) with India dataset

In Temporal Fusion Transformer, the input data were separated into three categories for instance, 1) static covariates (static features / static inputs / features that do not change overtime) 2) past covariates (past inputs / past observed inputs / historical inputs / features that are only known up to the current time) 3) future covariates (future input / known future input / a-priori known inputs / features whose values are known in advance for future timestamps) (Giacomazzi et al., 2023; Liu et al., 2024). The input variables were shown in Table 7.

Table 7. Input variables description of India dataset

Variable Types	Variables
Static covariates	-
Future covariates	time_idx, year, month, day
Past covariates	avg_PM2.5 (target)

About 90 percent of total data were used to train and validate the TFT model with the split of 70/20 respectively. The last 10 percent of the data were stored as a test set to evaluate the model's performance. The max encoder length and max prediction length were set to 180 and 30 days, respectively, which meant the model used the past 180 days to perform each prediction and forecasts 30 days into the future. Hyperparameters including batch size, hidden size, attention head size, dropout, learning rate, hidden continuous size, LSTM layers and weight decay were finely tuned using Optuna, a hyperparameters optimization framework. The tuned hyperparameters of the TFT on India dataset were displayed in Table 8. The training and validation loss curve was visualized on Fig. 12.

Table 8. Tuned hyperparameters of TFT on India dataset

Hyperparameters	Values
Batch size	64
Hidden size	64
Attention head size	9
Dropout	0.44
Learning rate	0.0003
Hidden continuous size	12
LSTM layers	1
Weight decay	0.0001

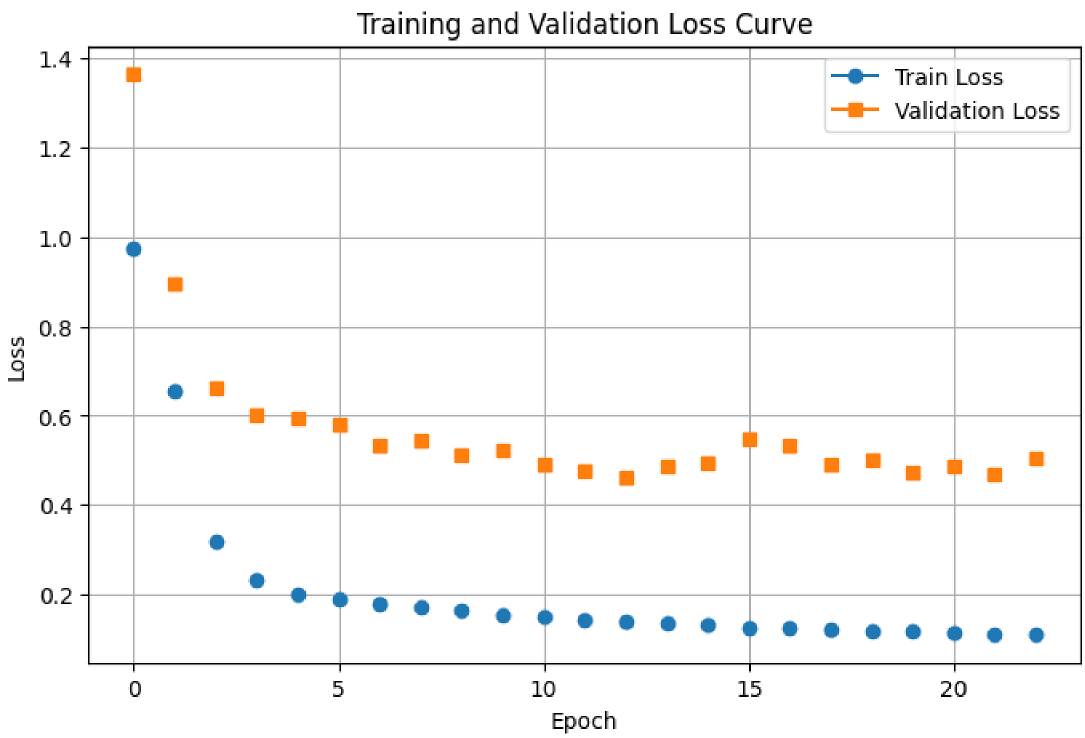

Figure 12. TFT's training and validation loss curve (India dataset)

From the training and validation loss curve, both training and validation loss obviously decreased between epochs 0 and 5 implied that the model had effective learning in the early training stage. After that, the training loss continuously reduced until the end of the process reflecting the model fitted the training data well. However, the validation loss started to increase and fluctuate after epoch 12, this behavior seemed to continue in the later epochs. This could be explained that the model was losing its generalization ability. The divergence of training and validation loss signified the model's overfitting.

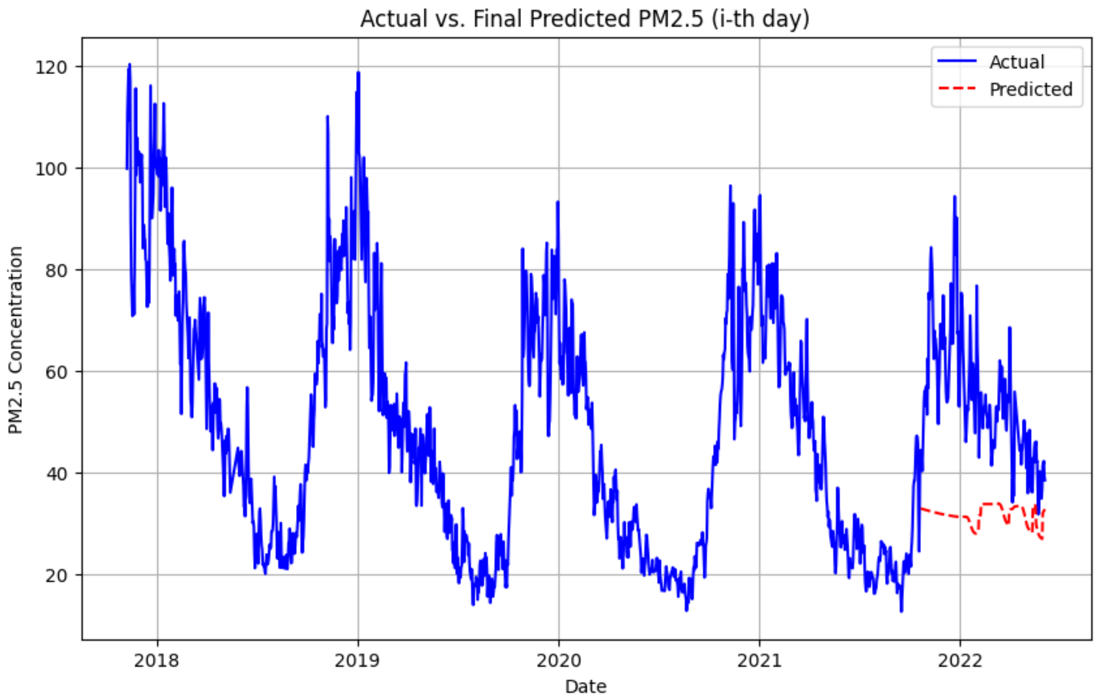

Figure 13. Prediction result of TFT model on India dataset

Table 9. Performance metrics of TFT model on India dataset

Metrics	Values
MAE	23.38
MSE	699.98
RMSE	26.46
MAPE	39.77

The above diagram (Fig. 13.) presented the plot of predicted PM2.5 concentration by the model against the real values. This suggested that the model's generalization was poor when the prediction was proceeded on the unseen data (test set). Besides, the high values of all metrics (Table 9.) also emphasized the unsatisfactory performance of the model. The reason causing the TFT model generated bad prediction results (overfitting) might be the size of India dataset. Only 1,616 datapoints might be insufficient to the TFT model for effective learning and data's pattern capturing. As a deep learning technique, the Temporal Fusion Transformer (TFT) needs a large amount of data to deliver the best prediction performance (Yuxuan, 2023).

4.5 Temporal Fusion Transformer (TFT) with Pakistan dataset

The input variables of Pakistan dataset were explained in Table 10. All settings of train-validation-test splitting, hyperparameters tuning, model training and model prediction were configured as same as those in India dataset. The tuned hyperparameters of the TFT on Pakistan dataset were expressed in Table 11. The training and validation loss curve was illustrated by Fig. 14.

Table 10. Input variables description of Pakistan dataset

Variable Types	Variables
Static covariates	-
Future covariates	time_idx, year, month, day
Past covariates	avg_temp, avg_aqi, avg_CO, avg_NO_2, avg_O_3, avg_PM10, avg_SO_2 and avg_PM2.5 (target)

Table 11. Tuned hyperparameters of TFT on Pakistan dataset

Hyperparameters	Values
Batch size	32
Hidden size	96
Attention head size	6
Dropout	0.22
Learning rate	0.00005
Hidden continuous size	14
LSTM layers	1
Weight decay	0.00008

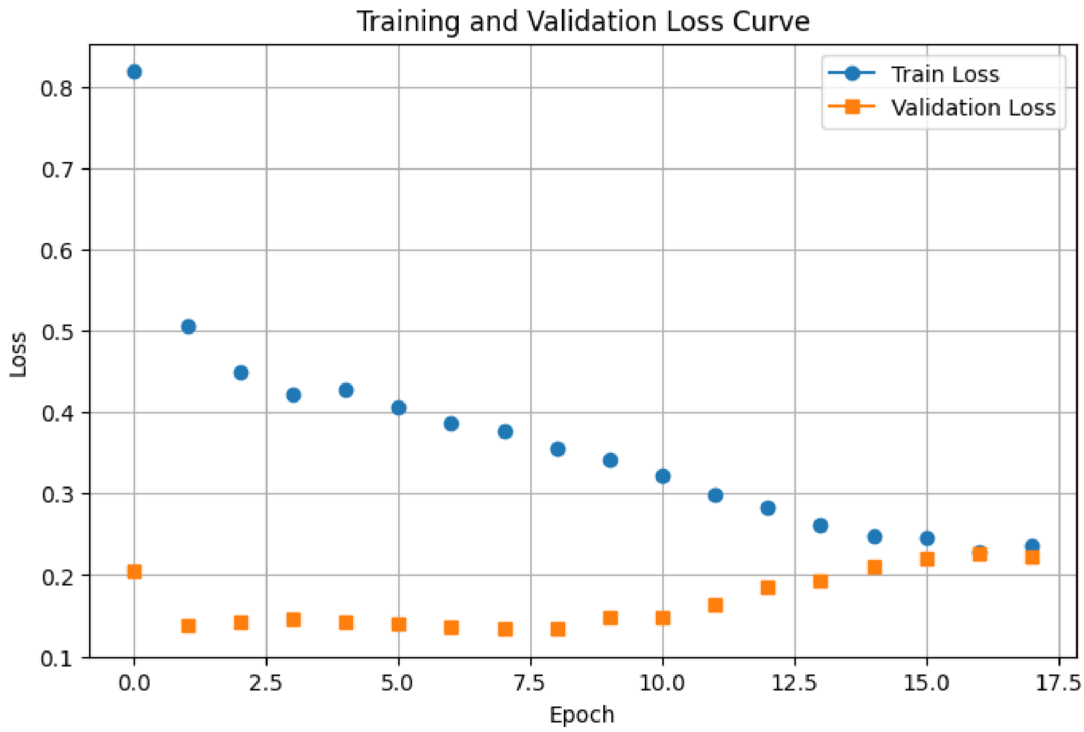

Figure 14. TFT's training and validation loss curve (Pakistan dataset)

From the training and validation loss curve, the training loss curve kept declining through the training process. This indicated that the model was learning from the training data. Meanwhile, the validation loss curve dropped at the early epochs then stayed steady until epoch 10. After that the validation loss curve began to climb up specifying the overfitting issue of the model. This meant the model tended to memorize the training data instead of learning from it. However, both training and validation loss curve seemed to stay close to each other. Thereby, this overfitting incident was considered as mild level.

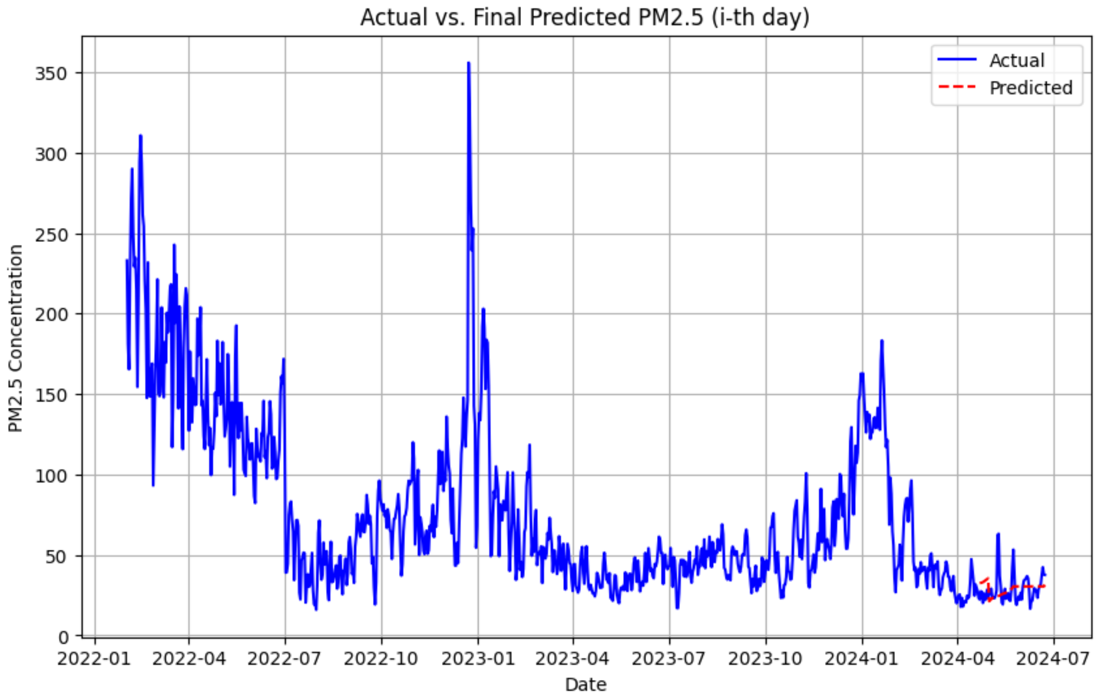

Figure 15. Prediction result of TFT model on Pakistan dataset

Table 12. Performance metrics of TFT model on Pakistan dataset

Metrics	Values
MAE	7.05
MSE	99.87
RMSE	9.99
MAPE	24.15

The prediction plot (Fig. 15.) and performance metrics (Tables 12.) of TFT model on Pakistan dataset implied good model's performance in PM2.5 prediction compared with the actual values. Even though the overfitting took place during the model training process, the model could still learn and capture data's pattern resulting in the acceptable generalization outcome. This might be described by the following reason. With only 874 datapoints of Pakistan dataset, it seemed that the amount of data was quite not enough for complex model like the TFT to yield fine prediction results. Nonetheless, the Pakistan dataset also owned other variables apart from avg_PM2.5 like avg_temp, avg_aqi, avg_CO, avg_NO_2, avg_O_3, avg_PM10 and avg_SO_2. These exogeneous variables might play an important role in model's performance improvement. By feeding these additional features along with avg_PM2.5 to the TFT, the model gains richer context for its learning. Hence, in Pakistan dataset, the model does not only learn from avg_PM2.5 as in the India dataset, but it can also see the multi-dimensional relationships between these external information and avg_PM2.5 which contributes to the better quality prediction results (C. Zhang et al., 2024). Moreover, the Temporal Fusion Transformer also proposes the attention score of each time index and feature importance to better understand which timesteps or features influenced to the prediction outputs (Liu et al., 2024).

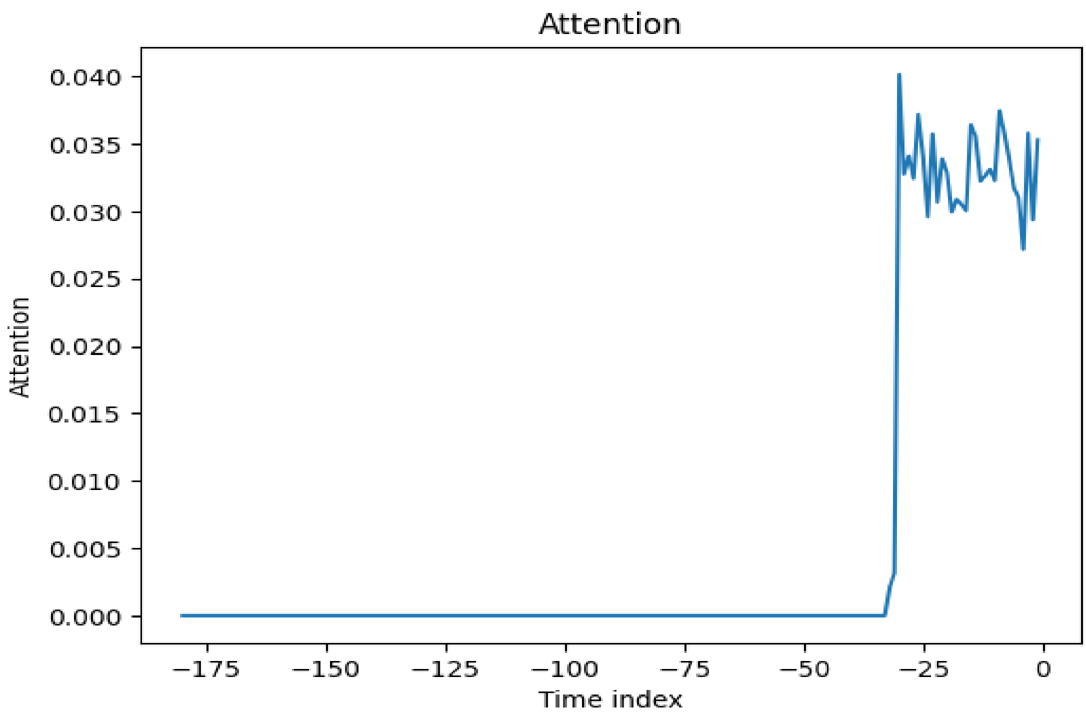

Figure 16. Attention score by time index of Pakistan dataset

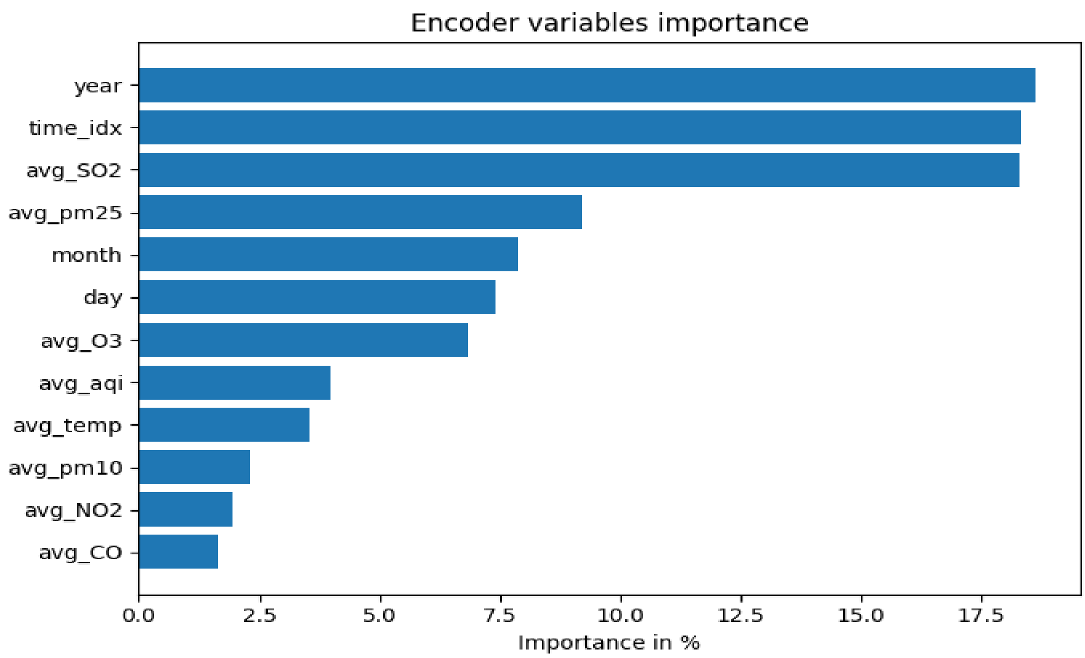

Figure 17. Encoder variables importance of Pakistan dataset

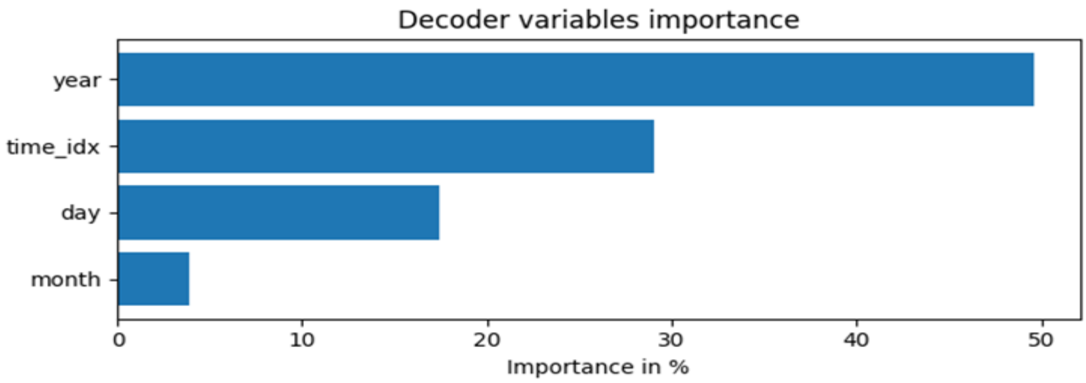

Figure 18. Decoder variables importance of Pakistan dataset

For the attention score (Fig. 16.), a higher attention score, a bigger influence on prediction outcomes (Liu et al., 2024). It was found that the time indices around -30 to 0 contributed to the model's prediction. In easy words, The TFT model leveraged the data from past 30 days to make prediction for PM2.5 levels. For the variables importance (Fig. 17. and Fig. 18.), year, time_idx and avg_SO_2 were the top three variables of the encoder. While year and time_idx were the important features for the decoder.

4.6 SARIMA model with India dataset and Pakistan dataset

Before conducting PM2.5 concentration prediction with SARIMA model, the time series was tested their stationarity with augmented Dickey-Fuller test (ADF-test). The results of ADF test showed that both time series were non stationary, thus the nonseasonal and seasonal differencing were implemented to eliminate trend and seasonality of the data (Ajewole et al., 2020; Roza et al., 2022). The SARIMA model's hyperparameters namely p, d, q, P, D, Q and s were identified and the Akaike Information Criterion (AIC) was used to determine the best model (Suleiman et al., 2023). However, it seemed that the hyperparameter 's' (seasonal period) could not be set as 365, as requiring the model to look back over daily data for a whole year causing the training process extremely slow and unresponsive. This reflected that SARIMA model is not suitable for highly granular seasonality scale modelling such as in daily data (Adams et al., 2019; Doreswamy et al., 2020; Makoni et al., 2018). To solve this issue, the Auto Regressive Integrated Moving-Average model (ARIMA) was used to make the prediction of PM2.5 concentration in both datasets instead. Nevertheless, ARIMA model was not designed to deal with seasonal time series, data's trend and seasonality were recovered to the prediction results. This made the outputs still had the same characteristics as the original data before compared with the test set. The details of ARIMA model used in PM2.5 prediction on India and Pakistan dataset were explained in Table 13.

Table 13. Details of ARIMA model

Dataset	ARIMA models	AIC values
India	ARIMA (1,1,1)	6378.99
Pakistan	ARIMA (1,0,0)	6324.49

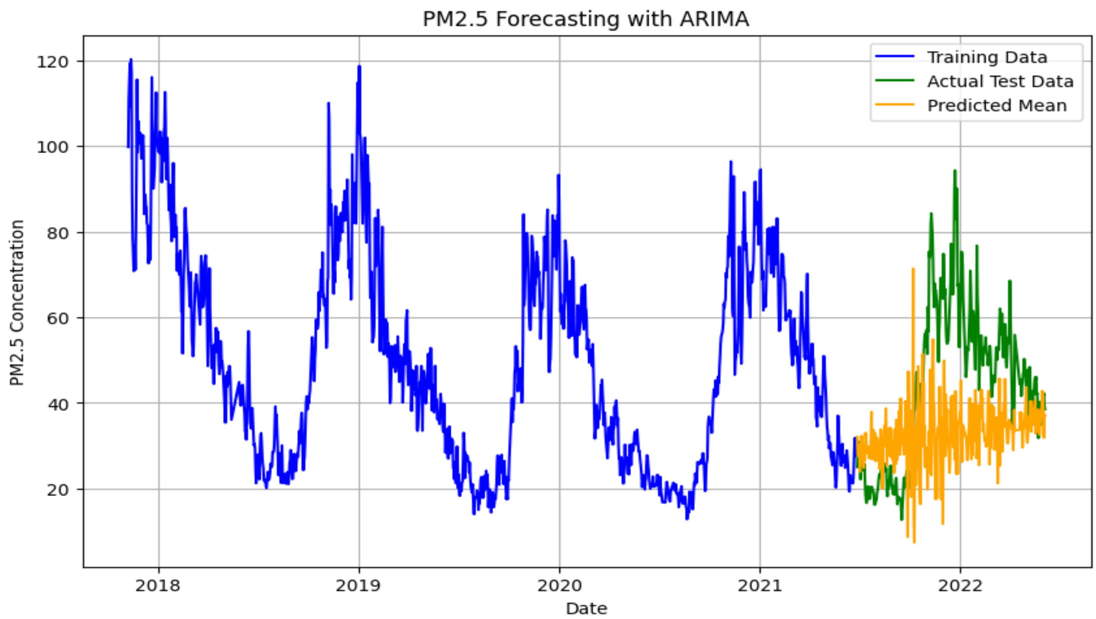

Figure 19. Prediction result of ARIMA model on India dataset

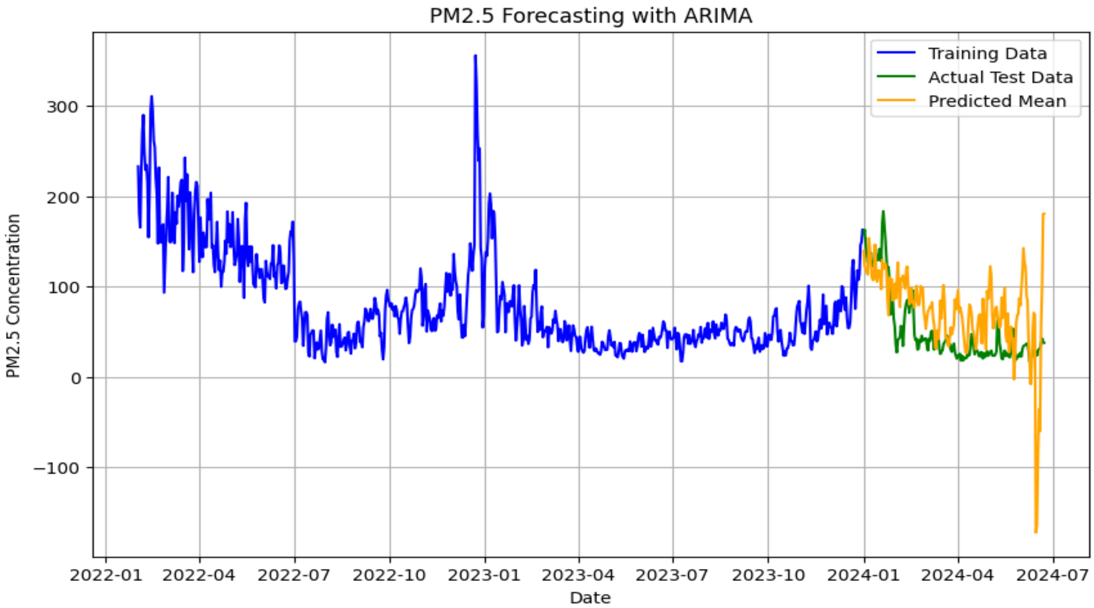

Figure 20. Prediction result of ARIMA model on Pakistan dataset

Table 14. Performance metrics of ARIMA model on India and Pakistan dataset

Metrics	India	Pakistan
MAE	17.23	38.46
MSE	474.60	2468.04
RMSE	21.79	49.68

It was found that both ARIMA models of India and Pakistan dataset were very complex based on their extremely high AIC values (Suleiman et al., 2023). Furthermore, the prediction plots (Fig. 19. And Fig. 20.) and models' performance metrics (Table 14.) also emphasized that the models could not generalize well with the unseen data resulting in significantly high prediction errors. Therefore, it might be confirmed that the implementation of SARIMA model on India and Pakistan dataset was not a good choice.

CHAPTER 5

CONCLUSIONS AND RECOMMMENDATIONS

5.1 Conclusions

In this study, the PM2.5 concentration of India and Pakistan dataset were predicted using three machine learning model namely Facebook Prophet, Temporal Fusion Transformer (TFT) and SARIMA model. The India dataset was an example of time series with strong and clear seasonality pattern, while the Pakistan dataset represented time series with unclear and complex seasonality pattern. The comparison of all models' performances in PM2.5 prediction was summarized in Table 15.

Table 15. The comparison of all models' performances in PM2.5 prediction

Dataset	Dataset's characteristics	Facebook Prophet	Temporal Fusion Transformer (TFT)	SARIMA model
India	Strong and clear seasonality	✓	✗	✗
Pakistan	Complex and unclear seasonality	✓	✓	✗

For India dataset, only Facebook Prophet could achieve the acceptable prediction performance. The Prophet with default hyperparameters (condition 1) was the best condition to this dataset. However, the performance of the Prophet with default hyperparameters and holidays components (condition 2) was very close to the condition

1's. This indicated that India's holidays did not have significant impact on PM2.5 concentration in the India dataset.

For Pakistan dataset, Both Prophet and TFT showed satisfied results. In this scenario, the Prophet with default hyperparameters and holidays components (condition 2) showed a little better performance than the Prophet with default hyperparameters (condition 1) which implied that Pakistan's holidays slightly influenced on the level of PM2.5 in this dataset. Furthermore, it was found that the TFT outperformed all models in Pakistan's PM2.5 prediction even if there was a small signal of overfitting during its training process. The overfitting might be due to the limited size of the dataset. Nevertheless, the TFT still yields good performance despite this constraint. This could be explained by leveraging other variables along with the target to improve prediction efficiency.

In addition, the experiments also indicated that SARIMA model was not suitable for fine grained seasonality modeling (such as daily data) emphasized by very high prediction errors.

In summary, it was suggested that the Facebook Prophet was the most suitable model to predict the PM2.5 level in this study.

5.2. Future work and recommendations

The results from this study demonstrated that the Prophet model possessed the best performance in PM2.5 prediction in both datasets. However, larger datasets are still necessary to ensure models' prediction ability, especially the TFT, the deep learning technique, which needs massive amount of data in training procedure to capture data's pattern. With its powerful computational power, it is expected that the TFT will outperform other models in PM2.5 prediction as long as there is not the data inadequacy issue.

REFERENCES

Adams, S., Mustapha, B., & Alumbugu, A. (2019). Seasonal Autoregressive Integrated Moving Average (SARIMA) Model for the Analysis of Frequency of Monthly Rainfall in Osun State, Nigeria. *Physical Science International Journal*, *22*, 1-9. https://doi.org/10.9734/PSIJ/2019/v22i430139

Ajewole, K., samuel oluwaseun, A., & Jemilohun, V. (2020). Test for Stationarity on Inflation Rates in Nigeria using Augmented Dickey Fuller Test and Phillips-Persons Test. *Vol. 16*, 11-14. https://doi.org/10.9790/5728-1603031114

Ao, X., Yuan, H., & Zhang, D. (2019). PM2.5 Analysis and Prediction Based on Seasonal Time Series Model. *IOP Conference Series: Earth and Environmental Science*, *371*(5), 052006. https://doi.org/10.1088/1755-1315/371/5/052006

Basith, S., Manavalan, B., Shin, T. H., Park, C. B., Lee, W. S., Kim, J., & Lee, G. (2022). The Impact of Fine Particulate Matter 2.5 on the Cardiovascular System: A Review of the Invisible Killer. *Nanomaterials (Basel)*, *12*(15). https://doi.org/10.3390/nano12152656

Basumatary, B., Dwivedi, M., & Verma, N. (2023). Social Attention Sparked by Highly Cited Research Publications on Women's Empowerment: An Assessment Based on Altmetric Tools. In (pp. 1-529).

Bhatti, U. A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., Yu, Z., & Yuan, L. (2021). Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach. *IEEE Access*, *9*, 41019-41031. https://doi.org/10.1109/ACCESS.2021.3060744

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis: Forecasting and control*. John Wiley & Sons, Ltd.

Cai, F. (2023). The Prediction of the Air Quality based on the Prophet Algorithm. *Highlights in Science, Engineering and Technology*, *39*, 1056-1060. https://doi.org/10.54097/hset.v39i.6705

Chirasophon, S., & Pochanart, P. (2020). The Long-term Characteristics of PM10 and

PM2.5 in Bangkok, Thailand. *Asian Journal of Atmospheric Environment*, *14*, 73-83. https://doi.org/10.5572/ajae.2020.14.1.073

Daraghmeh, M., Agarwal, A., Manzano, R., & Zaman, M. (2021). *Time Series Forecasting using Facebook Prophet for Cloud Resource Management*. https://doi.org/10.1109/ICCWorkshops50388.2021.9473607

Deforce, B., Baesens, B., Diels, J., & Serral Asensio, E. (2022, 2022). Forecasting Sensor-data in Smart Agriculture with Temporal Fusion Transformers.

Doreswamy, H., Kumar, K., & Gad, I. (2020). Time Series Analysis for Prediction of PM 2.5 Using Seasonal Autoregressive Integrated Moving Average (SARIMA) Model on Taiwan Air Quality Monitoring Network Data. *Journal of Computational and Theoretical Nanoscience*, *17*, 3964-3969. https://doi.org/10.1166/jctn.2020.8997

Duarte, D., & Faerman, J. (2019). *Comparison of Time Series Prediction of Healthcare Emergency Department Indicators with ARIMA and Prophet*. https://doi.org/10.5121/csit.2019.91810

Fayer, G., Lima, L., Miranda, F., Santos, J., Campos, R., Bignoto, V., Andrade, M., Moraes, M., Ribeiro, C., Capriles, P., & Goliatt, L. (2023). A Temporal Fusion Transformer Deep Learning Model for Long-Term Streamflow Forecasting: A Case Study in the Funil Reservoir, Southeast Brazil. *Knowledge-Based Engineering and Sciences*, *4*(2), 73-88. https://kbes.journals.publicknowledgeproject.org/index.php/kbes/article/view/7983

Garcia, A., Santa-Helena, E., De Falco, A., de Paula Ribeiro, J., Gioda, A., & Gioda, C. R. (2023). Toxicological Effects of Fine Particulate Matter (PM2.5): Health Risks and Associated Systemic Injuries—Systematic Review. *Water, Air, & Soil Pollution*, *234*(6), 346. https://doi.org/10.1007/s11270-023-06278-9

Geng, G., Zheng, Y., Zhang, Q., Xue, T., Zhao, H., Tong, D., Zheng, B., Li, M., Liu, F., Hong, C., He, K., & Davis, S. J. (2021). Drivers of PM2.5 air pollution deaths in China 2002–2017. *Nature Geoscience*, *14*(9), 645-650. https://doi.org/10.1038/s41561-021-00792-3

Giacomazzi, E., Haag, F., & Hopf, K. (2023). *Short-Term Electricity Load Forecasting*

Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources. https://doi.org/10.1145/3575813.3597345

Gu, Y., Henze, D. K., Nawaz, M. O., Cao, H., & Wagner, U. J. (2023). Sources of PM(2.5)-Associated Health Risks in Europe and Corresponding Emission-Induced Changes During 2005-2015. *Geohealth*, *7*(3), e2022GH000767. https://doi.org/10.1029/2022gh000767

Huang, Q. (2022). Forecasting stock prices using multi-macroeconomic regressors based on the facebook prophet model. *BCP Business & Management*, *25*, 231-242.

Jia, N., Li, Y., Chen, R., & Yang, H. (2023). A Review of Global PM2.5 Exposure Research Trends from 1992 to 2022. *Sustainability*, *15*(13).

Jia, Y.-Y., Wang, Q., & Liu, T. (2017). Toxicity Research of PM2.5 Compositions In Vitro. *International Journal of Environmental Research and Public Health*, *14*(3).

Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution*, *151*(2), 362-367. https://doi.org/https://doi.org/10.1016/j.envpol.2007.06.012

Khanal, I., Prakash Dhakal, O., Kumar Guragain, M., Karki, P., & Pandey, B. (2023). *Comparative Analysis of Time Series Forecasting Models for Predicting PM2.5 level in Kathmandu : SARIMA, Prophet and XGBoost* Proceedings of 14th IOE Graduate Conference, https://conference.ioe.edu.np/publications/ioegc14/IOEGC-14-170-PS2-001-138.pdf

Kollanus, V., Prank, M., Gens, A., Soares, J., Vira, J., Kukkonen, J., Sofiev, M., Salonen, R. O., & Lanki, T. (2017). Mortality due to Vegetation Fire-Originated PM2.5 Exposure in Europe-Assessment for the Years 2005 and 2008. *Environ Health Perspect*, *125*(1), 30-37. https://doi.org/10.1289/ehp194

Li, H., Fan, H., & Mao, F. (2016). A Visualization Approach to Air Pollution Data Exploration—A Case Study of Air Quality Index (PM2.5) in Beijing, China. *Atmosphere*, *7*(3).

Li, W., Lin, G., Xiao, Z., Zhang, Y., Li, B., Zhou, Y., Ma, Y., & Chai, E. (2022). A review of respirable fine particulate matter (PM(2.5))-induced brain damage. *Front Mol*

Neurosci, *15*, 967174. https://doi.org/10.3389/fnmol.2022.967174

Lico, L., Enesi, I., & Jaiswal, H. (2021). Predicting Customer Behavior Using Prophet Algorithm In A Real Time Series Dataset. *European Scientific Journal, ESJ*, *17*, 10. https://doi.org/10.19044/esj.2021.v17n25p10

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748-1764. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu, K., Ding, K., Cheng, X., Chen, J., Feng, S., Lin, H., Song, J., & Zhu, C. (2024). *Airport Delay Prediction with Temporal Fusion Transformers*. https://doi.org/10.48550/arxiv.2405.08293

Makoni, T., Chikobvu, D., & Maphosa, B. (2018). Forecasting Tourists Arrivals at Chinhoyi Caves Recreational Park in Zimbabwe Using SARIMA Model.

Martinez, E. Z., Silva, E. A., & Fabbro, A. L. (2011). A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Rev Soc Bras Med Trop*, *44*(4), 436-440. https://doi.org/10.1590/s0037-86822011000400007

Miller, L., & Xu, X. (2018). Ambient PM2.5 Human Health Effects—Findings in China and Research Directions. *Atmosphere*, *9*(11).

Momani, P. E. (2009). Time Series Analysis Model for Rainfall Data in Jordan: Case Study for Using Time Series Analysis. *American Journal of Environmental Sciences*, *5*. https://doi.org/10.3844/ajessp.2009.599.604

Nowak, D. J., Hirabayashi, S., Bodine, A., & Hoehn, R. (2013). Modeled PM2.5 removal by trees in ten U.S. cities and associated health effects. *Environmental Pollution*, *178*, 395-402. https://doi.org/https://doi.org/10.1016/j.envpol.2013.03.050

Oo, Z., & Phyu, S. (2020). Time Series Prediction Based on Facebook Prophet: A Case Study, Temperature Forecasting in Myintkyina. *International Journal of Applied Mathematics Electronics and Computers*, *8*, 263-267. https://doi.org/10.18100/ijamec.816894

Pant, P., Habib, G., Marshall, J. D., & Peltier, R. E. (2017). PM2.5 exposure in highly polluted cities: A case study from New Delhi, India. *Environmental Research*, *156*, 167-174. https://doi.org/https://doi.org/10.1016/j.envres.2017.03.024

Rasiya Koya, S., & Roy, T. (2024). Temporal Fusion Transformers for streamflow Prediction: Value of combining attention with recurrence. *Journal of Hydrology*, *637*, 131301. https://doi.org/https://doi.org/10.1016/j.jhydrol.2024.131301

Roza, A., Violita, E., & Aktivani, S. (2022). Study of Inflation using Stationary Test with Augmented Dickey Fuller & Phillips-Peron Unit Root Test (Case in Bukittinggi City Inflation for 2014-2019). *EKSAKTA: Berkala Ilmiah Bidang MIPA*, *23*, 106-116. https://doi.org/10.24036/eksakta/vol23-iss02/303

Sahu, S. K., & Kota, S. H. (2017). Significance of PM2.5 Air Quality at the Indian Capital. *Aerosol and Air Quality Research*, *17*(2), 588-597. https://doi.org/10.4209/aaqr.2016.06.0262

Samal, K. K., Babu, K., Das, S., & Acharya, A. (2019). *Time Series based Air Pollution Forecasting using SARIMA and Prophet Model*. https://doi.org/10.1145/3355402.3355417

Setianingrum, A. H., Anggraini, N., & Ikram, M. F. D. (2022, 20-21 Sept. 2022). Prophet Model Performance Analysis for Jakarta Air Quality Forecasting. 2022 10th International Conference on Cyber and IT Service Management (CITSM),

Srinamphon, P., Chernbumroong, S., & Tippayawong, K. Y. (2022). The Effect of Small Particulate Matter on Tourism and Related SMEs in Chiang Mai, Thailand. *Sustainability*, *14*(13).

Stefenon, S. F., Seman, L. O., da Silva, L. S. A., Mariani, V. C., & Coelho, L. d. S. (2024). Hypertuned temporal fusion transformer for multi-horizon time series forecasting of dam level in hydroelectric power plants. *International Journal of Electrical Power & Energy Systems*, *157*, 109876. https://doi.org/https://doi.org/10.1016/j.ijepes.2024.109876

Suleiman, S., Alabi, M., Usman, U., & Umar, M. (2023). Modeling and Forecasting the Crude Oil Price in Nigeria. *2*, 1-13.

Tadesse, K., Abcdef, & Dinka, M. (2023). Application of SARIMA model to forecastingmonthly flows in Waterval River, South Africa.

Taylor, S., & Letham, B. (2017). *Forecasting at scale*. https://doi.org/10.7287/peerj.preprints.3190v2

Thangavel, P., Park, D., & Lee, Y. C. (2022). Recent Insights into Particulate Matter (PM(2.5))-Mediated Toxicity in Humans: An Overview. *Int J Environ Res Public Health*, *19*(12). https://doi.org/10.3390/ijerph19127511

Vaishali, Verma, G., & Das, R. M. (2023). Influence of Temperature and Relative Humidity on PM2.5 Concentration over Delhi. *MAPAN*, *38*(3), 759-769. https://doi.org/10.1007/s12647-023-00656-8

Wan, Y., Li, Y., Liu, C., & Li, Z. (2020). Is traffic accident related to air pollution? A case report from an island of Taihu Lake, China. *Atmospheric Pollution Research*, *11*(5), 1028-1033. https://doi.org/https://doi.org/10.1016/j.apr.2020.02.018

Wang, W., & Guo, Y. (2009, 16-18 Oct. 2009). Air Pollution PM2.5 Data Analysis in Los Angeles Long Beach with Seasonal ARIMA Model. 2009 International Conference on Energy and Environment Technology,

Xie, Y., Zhao, B., Zhang, L., & Luo, R. (2015). Spatiotemporal variations of PM2.5 and PM10 concentrations between 31 Chinese cities and their relationships with SO2, NO2, CO and O3. *Particuology*, *20*, 141-149. https://doi.org/https://doi.org/10.1016/j.partic.2015.01.003

Yang, Q., Yuan, Q., Li, T., Shen, H., & Zhang, L. (2017). The Relationships between PM(2.5) and Meteorological Factors in China: Seasonal and Regional Variations. *Int J Environ Res Public Health*, *14*(12). https://doi.org/10.3390/ijerph14121510

Yuxuan, Z. (2023, 2023/01/06). Temporal Fusion Transformers Model for Traffic Flow Prediction. Proceedings of the BDEDM 2023, Changsha, People's Republic of China.

Zhang, C., Pan, G., Muresan, M., Lu, Z., & Fu, L. (2024). *Temporal Fusion Transformer for Real-Time Intersection Turning Movement Flow Forecasting Incorporating Exogenous Factors*. https://doi.org/10.1109/ITSC58415.2024.10919975

Zhang, Y., Sun, Q., Liu, J., & Petrosian, O. (2024). Long-Term Forecasting of Air Pollution Particulate Matter (PM2.5) and Analysis of Influencing Factors. *Sustainability*, *16*(1). https://mdpi-res.com/d_attachment/sustainability/sustainability-16-00019/article_deploy/sustainability-16-00019.pdf?version=1702976741

Zhao, N., Liu, Y., Vanos, J. K., & Cao, G. (2018). Day-of-week and seasonal patterns of PM2.5 concentrations over the United States: Time-series analyses using the Prophet procedure. *Atmospheric Environment*, *192*, 116-127. https://doi.org/https://doi.org/10.1016/j.atmosenv.2018.08.050

Zhu, J., Chen, L., Liao, H., & Dang, R. (2019). Correlations between PM2.5 and Ozone over China and Associated Underlying Reasons. *Atmosphere*, *10*(7).

APPENDIX

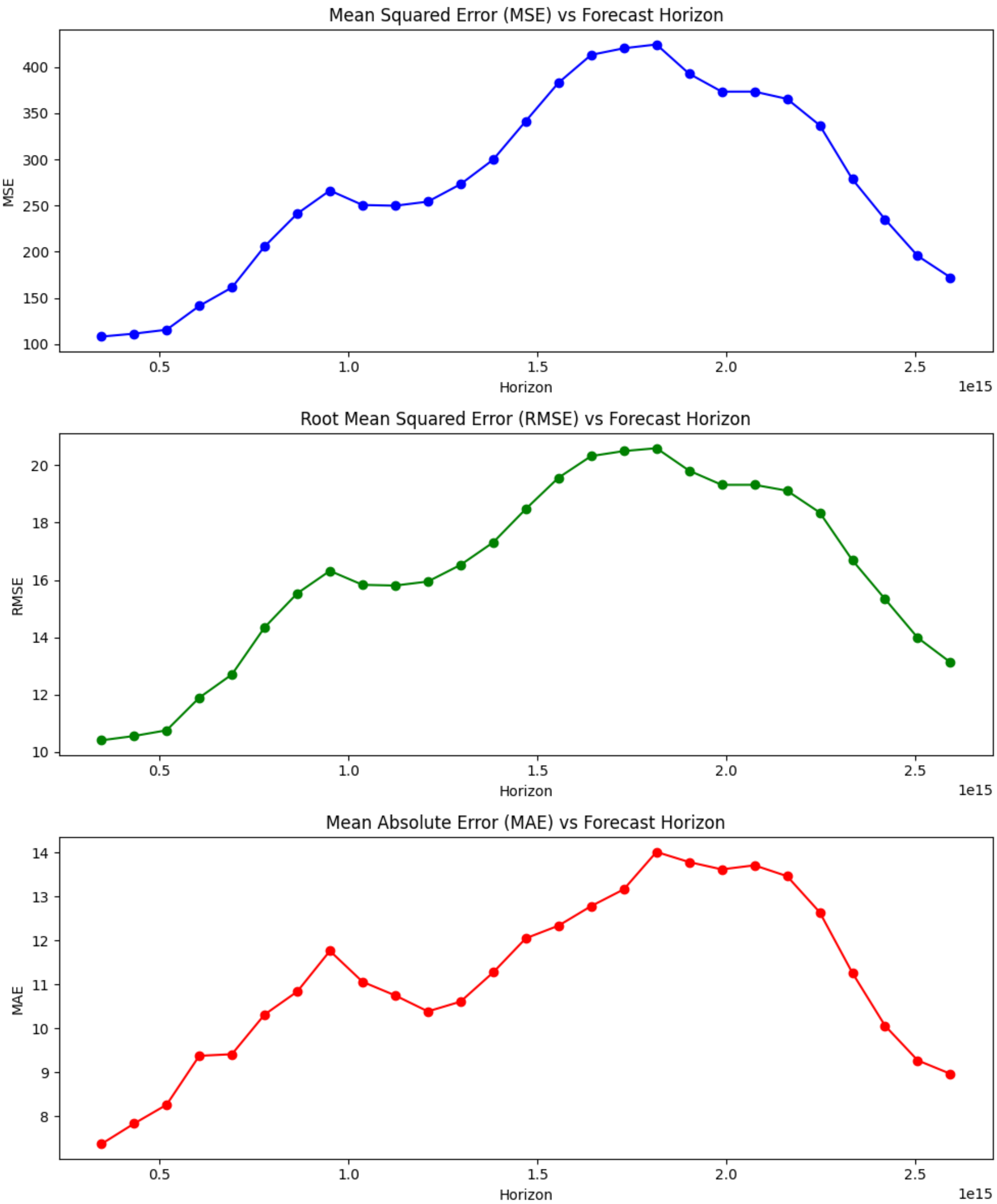

Figure 21. Cross validation results of Prophet model on India dataset with default hyperparameters (condition 1)

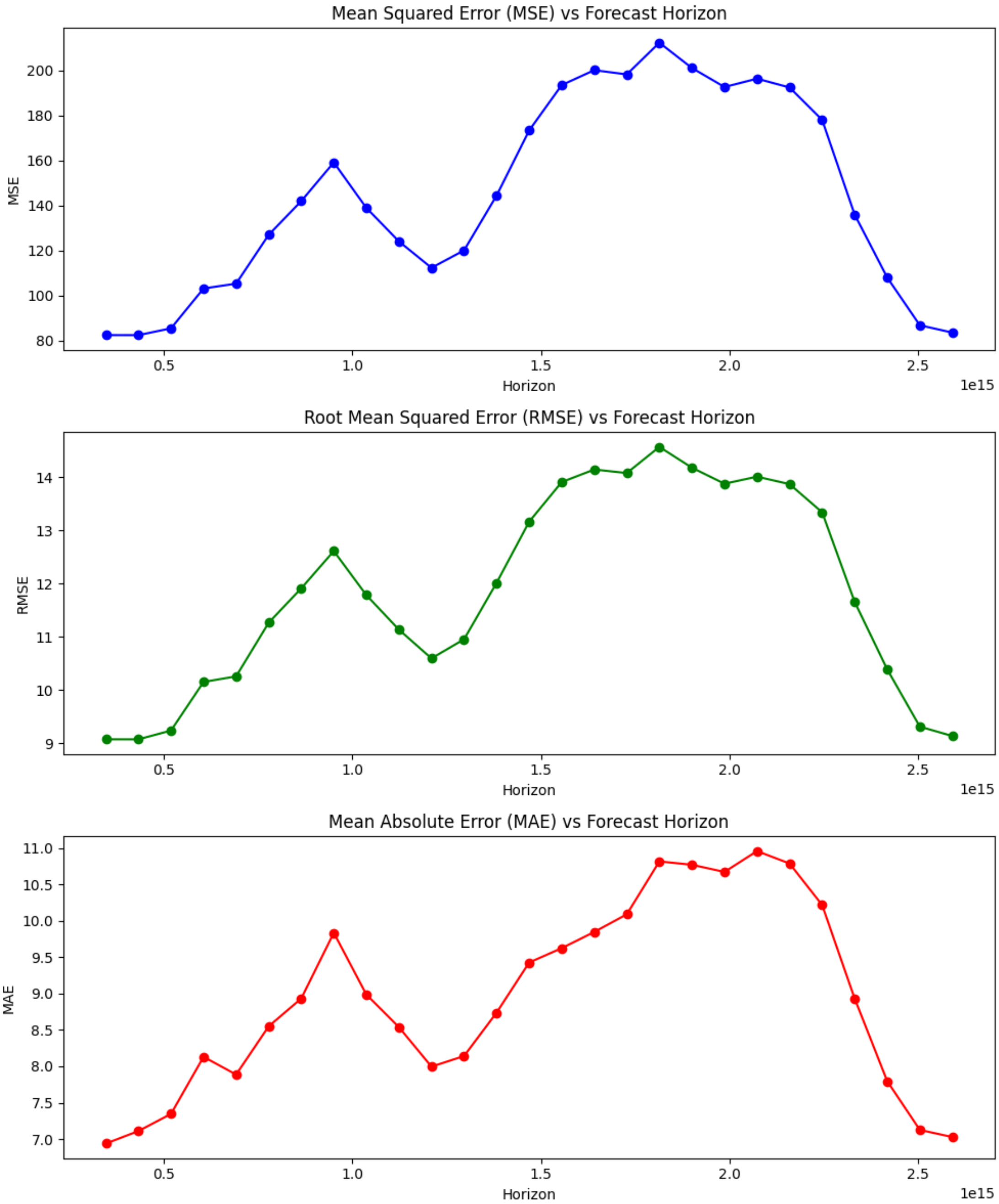

Figure 22. Cross validation results of Prophet model on India dataset with default hyperparameters and holiday components (condition 2)

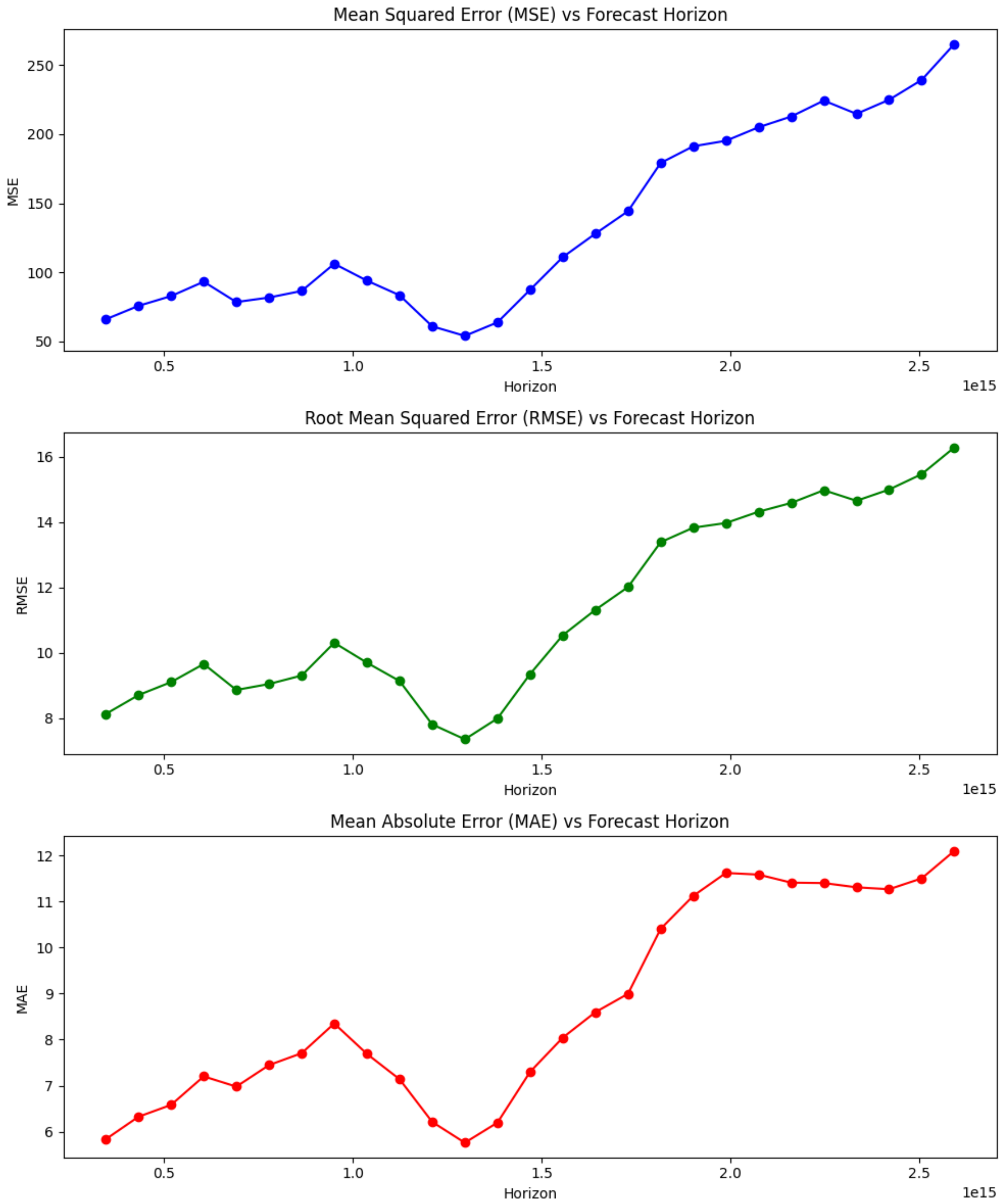

Figure 23. Cross validation results of Prophet model on India dataset with hyperparameters tuning using Optuna (condition 3)

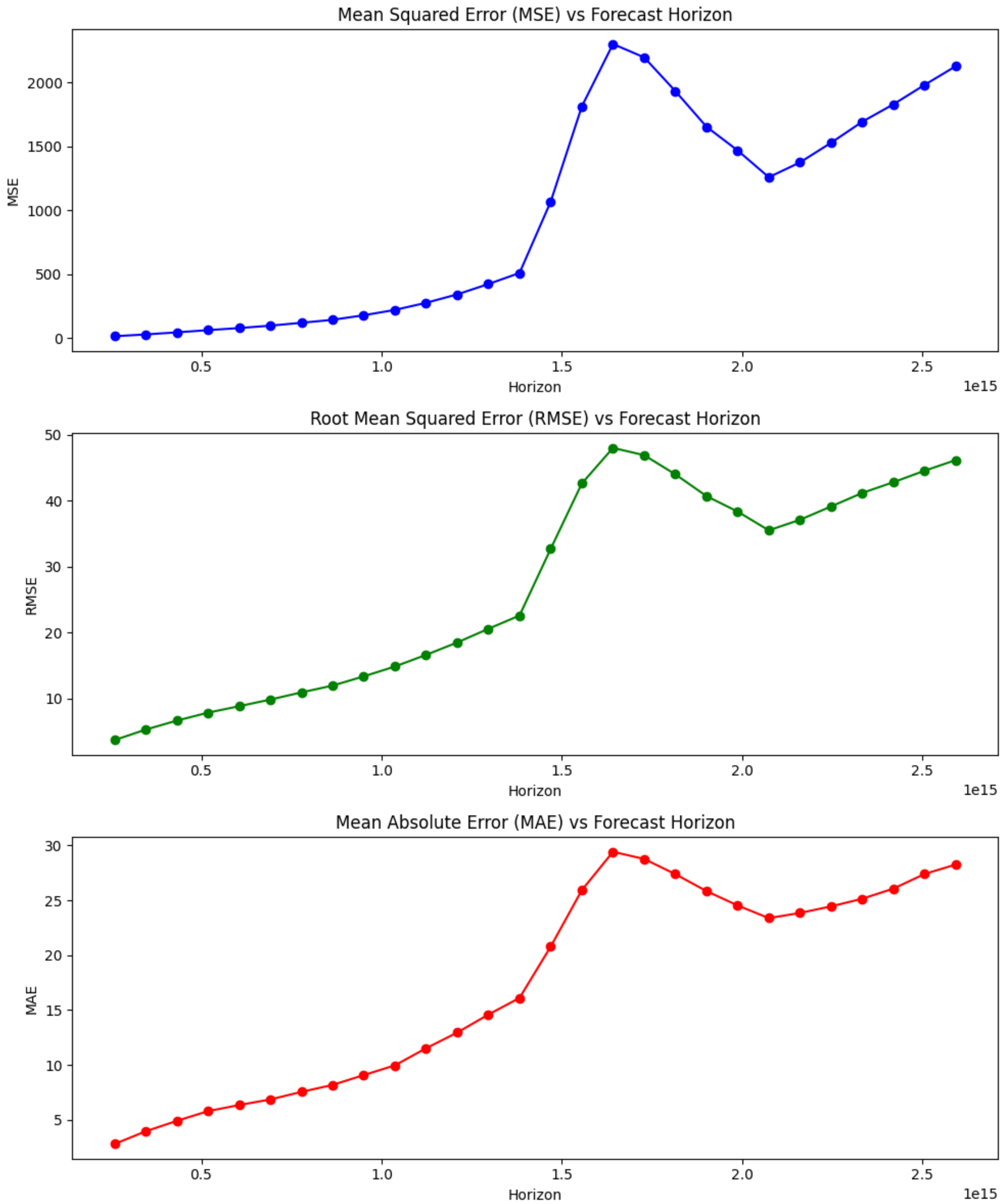

Figure 24. Cross validation results of Prophet model on Pakistan dataset with default hyperparameters (condition 1)

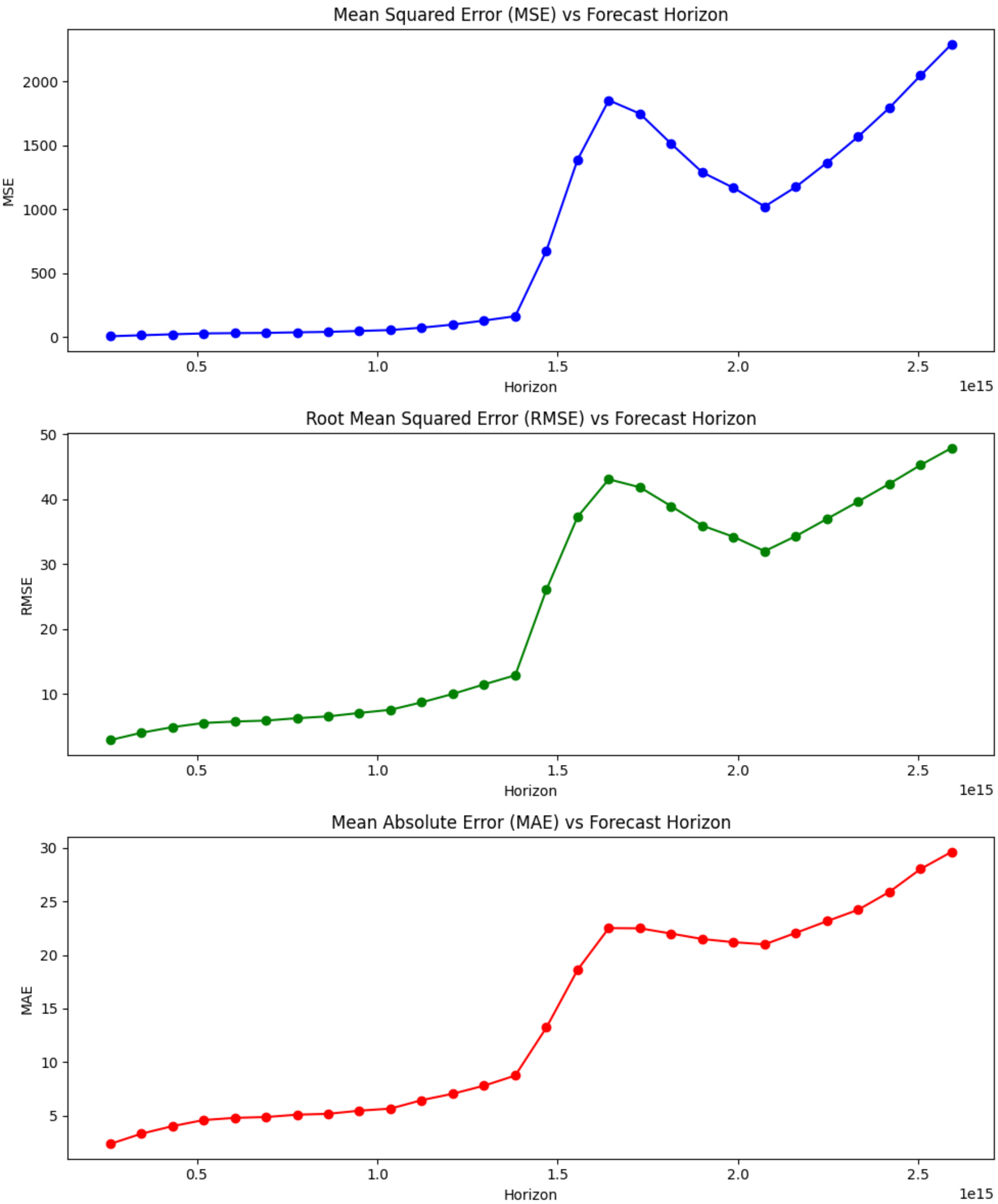

Figure 25. Cross validation results of Prophet model on Pakistan dataset with default hyperparameters and holiday components (condition 2)

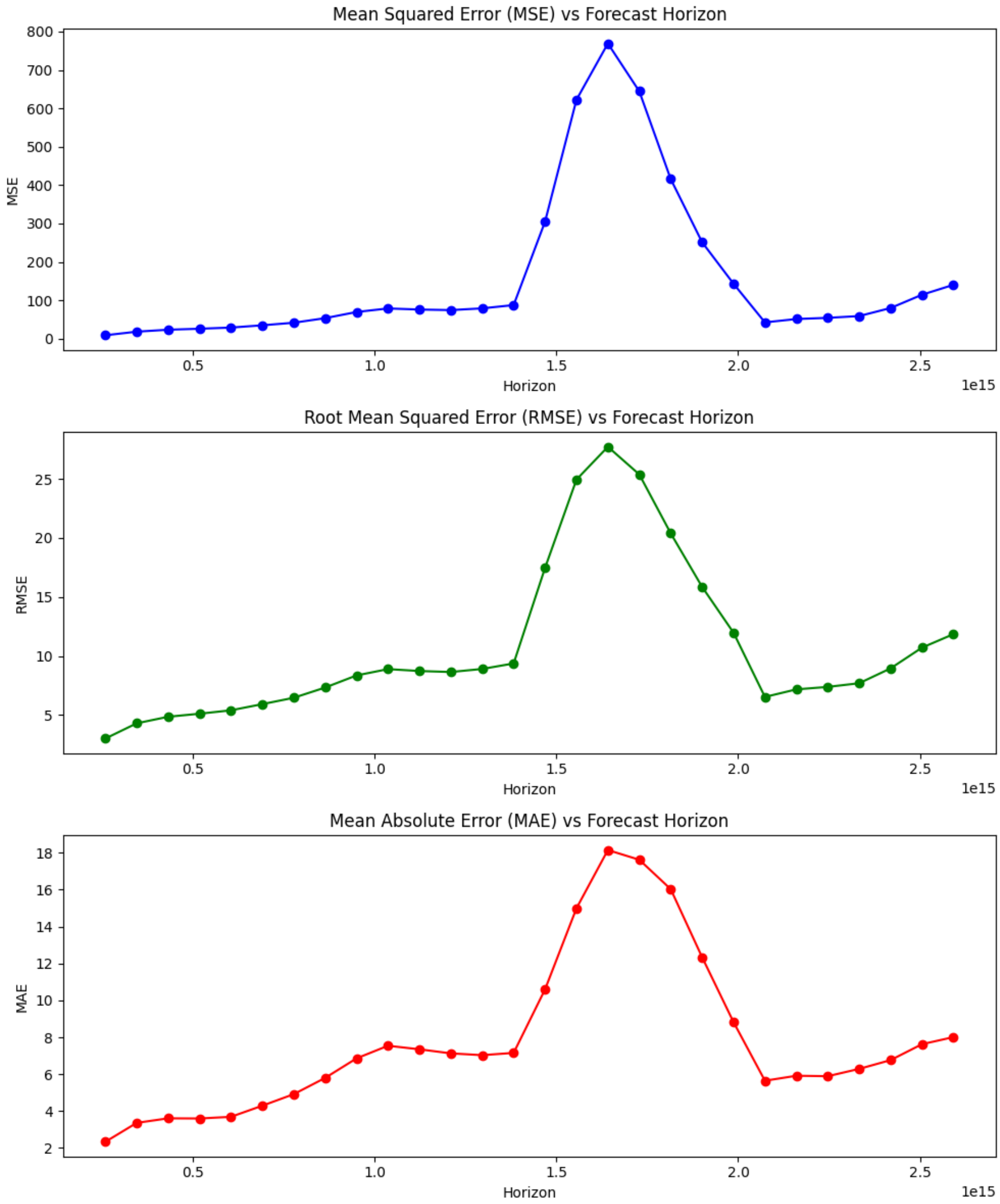

Figure 26. Cross validation results of Prophet model on Pakistan dataset with hyperparameters tuning using Optuna (condition 3)

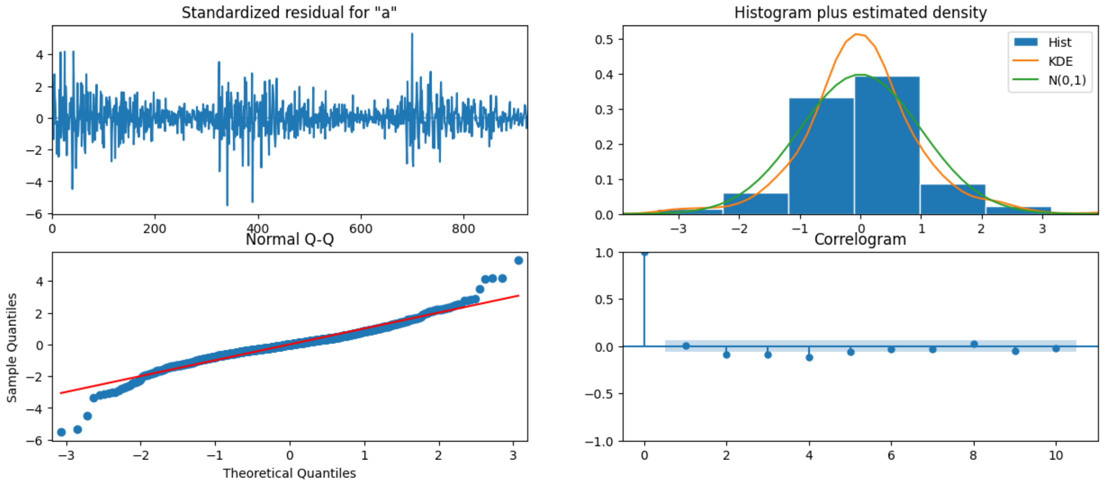

Figure 27. Diagnostic checking of ARIMA model on India dataset

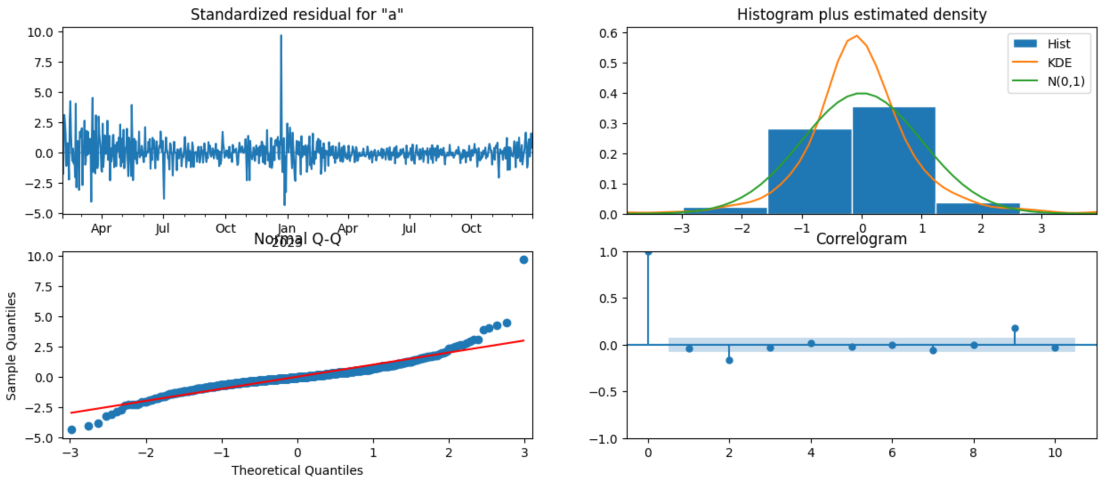

Figure 28. Diagnostic checking of ARIMA model on Pakistan dataset

VITA