1-12-2023

# Digital Twin of Atmospheric Environment: Sensory Data Fusion for High-Resolution PM2.5 Estimation and Action Policies Recommendation

Kudaibergen Abutalip

Anas Al-lahham

Abdulmotaleb Elsaddik

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2022, accepted 30 December 2022, date of publication 12 January 2023,
date of current version 15 February 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3236414*

RESEARCH ARTICLE

# Digital Twin of Atmospheric Environment: Sensory Data Fusion for High-Resolution $PM_{2.5}$ Estimation and Action Policies Recommendation

**KUDAIBERGEN ABUTALIP[1], ANAS AL-LAHHAM[1], AND ABDULMOTALEB EL SADDIK[1,2], (Fellow, IEEE)**

[1]Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates
[2]MCRLAB, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Abdulmotaleb El Saddik (elsaddik@uottawa.ca)

**ABSTRACT** Particulate matter smaller than 2.5 microns ($PM_{2.5}$) is one of the main pollutants that has considerable detrimental effects on human health. Estimating its concentration levels with ground monitors is inefficient for several reasons. In this study, we build a digital twin (DT) of an atmospheric environment by fusing remote sensing and observational data. An integral part of the DT pipeline is the presence of feedback that can influence future input data. Estimated values of $PM_{2.5}$ obtained from an ensemble of Random Forest and Gradient Boosting are used to provide recommendations for decreasing the agglomeration levels. We formulate a simple optimization problem for suggesting the recommendations and identify possible action policies, such as cloud seeding, scheduling of air filtering, and SMS notifications. The $PM_{2.5}$ estimation part of the proposed DT pipeline has achieved RMSE and $R^2$ of 38.94 and 0.728 (95%, CI 0.717-0.740). In addition, we investigate different approaches for quantitatively examining the contribution of each independent variable.

**INDEX TERMS** $PM_{2.5}$, digital twins, satellite data, health risk, air pollution.

## I. INTRODUCTION

The problem of air pollution remains one of the greatest environmental risks on Earth. According to the World Health Organization (WHO) statistics [1], 4.2 million deaths every year result from exposure to surrounding air contamination, and 9 out of 10 people globally live where air quality is significantly above WHO guideline limits. Nitrogen dioxide, particulate matter, sulfur dioxide, and carbon monoxide are among the biological or physical substances which alter the natural properties of the atmosphere. Particulate pollutants smaller than 2.5 micrometers are known to be one of the most hazardous contaminants. Being tiny allows them to be easily inhaled and stay inside the lungs, which usually causes various respiratory, and cardiovascular diseases, including mortal ones [2], [3]. Therefore, the estimation of $PM_{2.5}$ and the development of preventive policies are substantial research questions in this field.

The associate editor coordinating the review of this manuscript and approving it for publication was Yu Gu.

Existing ground monitors for $PM_{2.5}$ have many limitations as being absent in many countries and regions [4] and being expensive in terms of management and resources. They do not meet the requirements of widespread coverage and low latency, which hinders the possibility of taking precautionary measures. The development of models for forecasting $PM_{2.5}$ concentration levels based on remote sensing data and or other meteorological variables is a promising solution. Many attempts have been made to construct such approaches. According to Chu et al. [5], methods involving satellite data, developed in or before 2016, have mostly used the next models: Multiple Linear Regression (MLR), Mixed-Effect Model (MEM), Chemical Transport Model (CTM), and Geographically Weighted Regression (GWR). They concluded that each model had advantages and disadvantages based on usage scenarios. However, in terms of predictive performance MEM performed the best and can be suitable for scarce data, while MLR exhibited the worst results. CTM was better for global-scale predictions, and GWR showed better results at the local level. The authors also emphasize the importance of

using auxiliary meteorological variables with land use information for boosting models' forecasting capacity. With the rise of more sophisticated machine learning (ML) techniques, including deep learning (DL) models, trends for estimating $PM_{2.5}$ have shifted. Furthermore, boundaries for the practical application of $PM_{2.5}$ estimation methods are not concretized yet. The number of studies focusing on the modeling of the problem is large, whereas there are not many that focus on further integration of approaches into a broader framework. Recently, the concept of Digital Twins (DT) has attracted much more attention. A digital twin is a digital replica of a living or non-living object, structure, environment, or event with a requirement of bidirectional information exchange. In a DT pipeline, feedback from a model is used for making insightful, preventive, or foreseeing decisions. In the context of air quality control, and $PM_{2.5}$ specifically, questions regarding the usage of model predictions for taking precautionary actions, and automating them remain open.

**Our contributions:**

- We build a digital twin of a non-living entity (atmospheric environment).
- $PM_{2.5}$ levels estimation model for $5 \times 5$ km districts in three cities is proposed.
- We investigate action policies for decreasing pollutant levels.
- The most influential features of the given datasets are identified based on two algorithms.

## II. RELATED WORK

Details of relevant studies for pollutant level estimation and Digital Twins (DT) and their potential merits and drawbacks are analyzed in this section.

### A. $PM_{2.5}$ PREDICTION

#### 1) CLASSICAL ML METHODS

$PM_{2.5}$ concentration prediction has been investigated using a variety of effective machine learning methods. ML approaches, for which satellite-derived AOD data has been used for training, have achieved significant improvement in predictive performance compared to previous conventional methods. Zamani et al. [6] proposed a model for estimating $PM_{2.5}$ levels in Tehran, Iran based on a combination of Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Deep Neural Network (DNN). The authors investigate 23 different variables, including widely distributed ground-based $PM_{2.5}$ measurements, meteorological data, and remote sensing Aerosol Optical Depth (AOD) at 3 km and 10 km spatial resolution from 2015 to 2018. A considerable fraction of entries (96%) for AOD at 3 km resolution was missing and the authors excluded it for some experiments. This is the first study of its kind to look into the significance of the features that have the greatest impact on $PM_{2.5}$ prediction. The study reports that the XGBoost method has demonstrated the highest model performance ($R^2 = 0.8$, MAE $= 10.0\ \mu g/m^3$, and RMSE $= 13.62\ \mu g/m^3$). Based on different methods for evaluating feature importances, the authors conclude that historical observations of $PM_{2.5}$, wind speed, visibility, day of the year, altitude, and temperature are one of the most influential factors for forecasting air agglomeration levels for the given region.

Ma et al. [7], also utilize XGBoost for estimating the same pollutant mass concentrations in Shanghai, China. They use ground measurements of meteorological conditions and data from the operational air quality numerical prediction system (WRF-Chem) at the Shanghai Meteorological Service (SMS). Similarly, Masood et al. [8] use two different ML models to predict daily $PM_{2.5}$ concentration over a period of 2 years in Delhi, India. The authors propose Artificial Neural Network (ANN) and Support Vector Machine (SVM). This study used multiple meteorological parameters as input to their model with a total of 687 data points. The authors report that ANN is capable of outperforming other ML methods and achieves an MSE of 0.0191. However, comparative examination in this study is done with a small dataset.

In light of the approaches that focused on a single ML model for the final prediction, the employment of an ensemble approach that utilizes multiple methods to obtain a better final prediction is essential. Yazdi et al. [9] has introduced an ensemble ML approach for daily $PM_{2.5}$ prediction that included predictions from three different ML models: RF, a gradient boosting machine (GBM), and a k-nearest neighbor (KNN). The paper shows the results of extensive experiments ran separately on each approach. The individual machine learners are ensembled using a generalized additive model (GAM). A total of 27 covariates are used to train the models over a period of 9 years. The authors use ten-fold cross-validation for evaluation. The proposed model performs well, with $R^2$ of 0.828. However, for spatial variability prediction, $R^2$ drops to 0.396. The authors believe that it is attributable to the area's lower spatial variance in pollution levels.

#### 2) DEEP LEARNING METHODS

Original works for estimating $PM_{2.5}$ involving deep learning methods started appearing in 2016 and further escalation has been observed in the following years.

Li et al. [10] propose to combine Convolutional Neural Network with a Long Short-term Memory model (CNN-LSTM) for forecasting the pollutant concentration for the next 24 hours in Beijing. The authors use 1D convolutions for feature extraction and LSTM for modeling temporal interdependencies between meteorological variables, such as temperature, atmospheric pressure, wind direction, accumulated hours of rain and snow, wind speed, dew point, and $PM_{2.5}$ concentration with a total of 43800 records. They indicate the presence of a considerable number of missing values, and to mitigate the problem they fill in zero values. MAE of 13.96 and RMSE of 17.93 have been reported in this work. If the majority of the data points were missing, filling them with zeros might have affected inductive biases learned by the model. Therefore, more sophisticated

imputing strategies (e.g. interpolation, computations based on the sliding window) could be used to further improve the performance.

In a similar manner to the previous work and for the same region, Pak et al. [11] utilize the CNN-LSTM model with an additional spatio-temporal analysis based on Mutual Information (MI). The data is collected from 384 monitoring stations and contained 13152 records for each of them. Different from the previous work, a smaller number of features are used. modeling starts with computing a spatio-temporal feature vector (STFV), which is further fed into the network. The authors indicate that this step helps to increase the overall predictive performance of the model. RMSE of 2.997, MAE of 2.21, and MAPE of 0.039 have been reported. The authors conduct an extensive evaluation of the proposed model. Precomputing the STFV vector is an interesting addition to the pipeline, which lets the CNN-LSTM model to easier capture representative relationships between variables. Additionally, the authors provide insights on how air quality and meteorological data affect the prediction metrics separately and clearly show the importance of using both.

Chen et al. [12] emphasizes possible issues usually associated with datasets used for $PM_{2.5}$ estimation: ineffectiveness of conventional correlation analysis for high-dimensional meteorological data, and frequential fluctuations in it. To address these issues the authors have proposed a methodology based on a wavelet transform that decomposes different meteorological time series into multiple lower-resolution sub-time series with varying frequencies. They also developed an LSTM model modified with radial basis function (RBF-LSTM) for extracting the most important features. Feature selection based on the proposed method has been evaluated with various models, such as RF, Decision Tree (DT), Multilayer Perceptron (MLP), and RNN are shown to be effective. The qualitative and quantitative analysis provided in this study underline the importance of preprocessing and addressing issues inherent to this type of data. However, the methods they have used might be replaced for increasing the time efficiency of the proposed pipeline.

In another work [13], a recurrent neural network (RNN) has been used for improving the currently used Community Multiscale Air Quality (CMAQ) model for predicting $PM_{2.5}$ amounts in Seoul metropolitan area in Korea. Air quality data and surface meteorological variables, like surface pressure, air temperature, meridional wind, zonal wind, relative humidity, and dew-point temperature are considered for modeling pollutant concentrations. The RNN-CMAQ approach outperforms and corrects statistical biases present in CMAQ-only predictions, such as overestimations due to small or large peaks of particulate concentrations. Overall, the modified model exhibited 50% lower MAE than the original method. The main limitation of RNN is the absence of a straightforward option for computing each variable's relevance. Some works have been proposed to alleviate this issue [14], however, they usually come with large computational requirements as they are mainly based on experiments with the exclusion or inclusion of the specific variable and evaluation of corresponding results from such changes. Though pure RNN can capture temporal interdependencies more efficiently than classical approaches, it requires a more sophisticated training procedure and is more prone to the problem of vanishing gradients.

Overall, deep learning techniques have attracted more attention for $PM_{2.5}$ estimation and have shown promising results, but classical ML approaches still outperform DL models on time-series data [15].

### B. DIGITAL TWINS

Digital Twins (DT) have been first introduced in 2003 [16], [17]. The concept itself is defined in various ways by different authors. For example, El-Saddik et al. [18] defines the DT concept as a virtual representation of any living or non-living entity. In a similar manner, NASA [19] has proposed the following definition - "A Digital Twin is an integrated multiphysics, multiscale, probabilistic simulation of an as-built vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its corresponding flying twin".

A digital twin consists of three parts: physical objects, virtual objects, and connected data. An important part of the DT pipeline is interaction. Feedback or prediction from it can affect the future input data. As a complete system, in general, DT should have the following characteristics [18], [20]:

1) **Unique identifier.** To connect with their twin, digital twins should have unique identifiers.
2) **Trust.** Real twins must be able to trust their digital twin for them to conduct sensitive responsibilities.
3) **Privacy and security.** Ethical and privacy concerns should be taken into account while designing the DT.
4) **Real-time reflection.** It is important to combine different kinds of physical object data so that real-time mapping of physical things can continue.
5) **Sensors and actuators.** The actual living or non-living entity that is being twinned may have sensors that enable its digital copies to imitate their senses, including sight, hearing, taste, smell, and touch.
6) **Interaction and convergence.** It coexists with the entire lifecycle of physical objects and evolves alongside it.
7) **Evolution and iteration.** It is capable of not only describing but also optimizing physical objects using an iterative virtual model.
8) **Representation.** Depending on the application, digital twins could have a virtual representation such as a hologram, or even a humanoid social robot. They could also be software components with no real representation.

As discussed earlier, although the number of studies focusing on the modeling of the problem is large, there is a lack of focus on further integration of approaches into a broader framework. One potential direction is to integrate the system into a framework based on the DT paradigm.

Many studies have researched and analyzed digital twins up to date [16], [21]. The digital twin is currently utilized mostly in product design and service management, manufacturing, agriculture, and real-time equipment monitoring in the industry [22], [23], [24], [25]. The digital twin concept combined with a weather forecast system is not described in any research papers or technical reports.

## III. METHODOLOGY

Overall DT pipeline (Figure 1) involves data from multiple sources, which require specific preprocessing steps. Next, the merged data is used to estimate the pollutant concentrations. Agglomeration levels are further used to generate recommendations based on user input, and potential action policies are suggested. Details of each step are outlined in this section.

### A. STUDY AREA

Three urban areas that represent different parts of the world, namely South Coast Air Basin, Delhi, and Taipei, are included in this study. The number of city districts available for each of the three studied locations is 33, 14, and 7. The study period is from Feb. 2018 to Aug. 2021.

#### 1) LOS ANGELES (SOUTH COAST AIR BASIN)

The South Coast Air Basin (SCAB or SoCAB), California. One of California's recognized regional air basin areas. A large portion of the Greater Los Angeles Area, which has a population of over 18 million and generates close to a third of the state's total emissions of criterion pollutants, is located in the region, which has an approximate area of 17100 $km^2$. Located 43.6 meters above sea level, South Coast Air Basin. SCAB is supported by five Lead (Pb) air monitoring stations with single-pollutant source effects.

#### 2) DELHI

is a city and union territory in India that contains New Delhi, the nation's capital. It has a surface area of 1,484 $km^2$ [26]. The population of Delhi is over 11 million people. 13 of the 1600 cities reviewed by the World Health Organization (WHO) were among the top twenty most polluted cities, with Delhi leading the list for $PM_{2.5}$ concentrations with levels at least 10 times higher than Washington, DC, and three times higher than Beijing [27]. In Delhi, the residential, transportation, and industrial sectors are mostly responsible for $PM_{2.5}$ production.

#### 3) TAIPEI

The Taipei region is in eastern China, with rugged terrain and a temperate continental monsoon climate. It has become a hotspot for environmental research such as aerosol monitoring and air pollution assessment due to severe air pollution. As a result, the Taipei region was chosen as the focus of this study.

### B. SATELLITE DATA

#### 1) MULTI-ANGLE IMPLEMENTATION OF ATMOSPHERIC CORRECTION (MAIAC)[1]

Using data from the two Moderate Resolution Imaging Spectroradiometer (MODIS) satellite sensors, the MAIAC algorithm produces high-resolution aerosol and land surface reflectance reports. MAIAC has undergone significant changes for better cloud/snow identification, aerosol retrievals, and atmospheric correction of MODIS data.

#### 2) THE PHYSICAL CONCEPT OF MAIAC

To optimize all stages of MAIAC processing, it concentrates on a precise characterization of the surface background since, given the daily rate of global MODIS observations, the surface changes more slowly than aerosols and clouds. To use polar-orbiting data and monitor the same grid cell over time, MAIAC first grids MODIS measurements to a fixed grid with 1 km precision. Several HDF4-format atmospheric and surface products, including daily **MCD19A2** products, are provided by MAIAC.

The MCD19A2 Version 6 product contains daily AOD information at different bands at 1 kilometer (km) pixel resolution. The atmospheric properties and view geometry used for generating the MAIAC Land Surface Bidirectional Reflectance Factor (BRF) or surface reflectance, the MCD19A1 product, are provided by the MCD19A2 product. The MCD19A2 contains several scientific layers, such as fine mode fraction over water, column water vapor over land and clouds (in cm), smoke injection height (m above ground), AOD QA, and the AOD model at 1km, cosine of the solar zenith angle, the cosine of the view zenith angle, the relative azimuth angle, the scattering angle, and the glint angle at 5km. Figure 2 displays a low-resolution image of the AOD of the blue band at 0.55 m.

### C. METEOROLOGICAL DATA

The National Centers for Environmental Prediction (NCEP) developed the Global Forecast System (GFS), a weather forecasting model that produces data for numerous atmospheric and land-soil variables including temperatures, winds, precipitation, soil moisture, and atmospheric ozone concentration. The system incorporates four separate models (atmosphere, ocean, land/soil, and sea ice) to accurately explain meteorological conditions. GFS data[2] is distributed in grib2 format at 0.25° resolution.

Total cloud cover, Dewpoint Temperature, Pressure, Zonal Wind, Snow Depth, Temperature, Potential Evaporation Rate, Water Runoff, Relative Humidity, Winds, Total Precipitation, Ice Cover, Pressure reduced to Mean Sea Level (MSL), Meridional Wind, and Total Cloud Cover are among the dataset layers found in GFS.

[1] https://lpdaac.usgs.gov/products/mcd19a2v006/
[2] https://rda.ucar.edu/datasets/ds084.1/

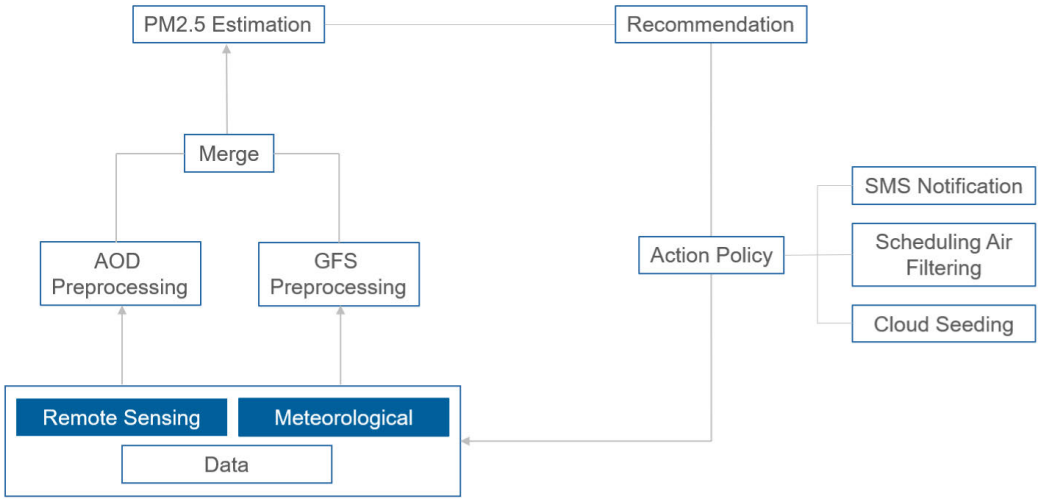


**FIGURE 1.** Schematic illustration of the DT pipeline.

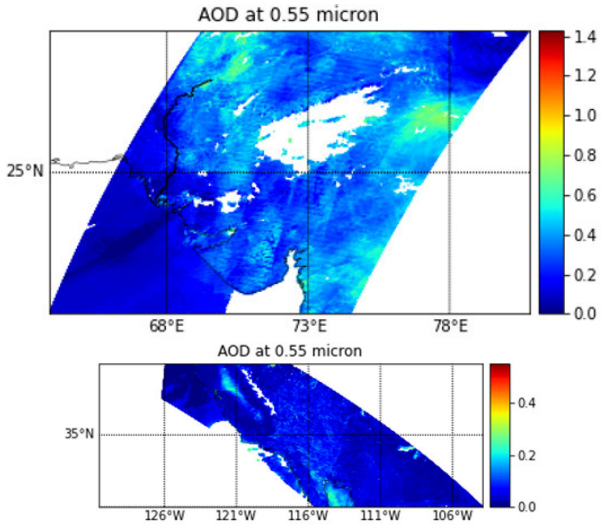


**FIGURE 2.** Example images from MCD19A2 product.

### D. DATA PREPROCESSING

#### 1) SATELLITE DATA

AOD data is in HDF format [28] and appropriate preprocessing steps are outlined further: 1) layer extraction - selection of the necessary variable information only, 2) offsetting and scaling 3) construction of the grid in WGS84 (EPSG:4326) coordinate system, 4) subsetting the geographical region of interest. As the data is in a raster format, mean, 95th percentile, min, max, standard deviation, and variance of AOD values were extracted. The main problem with AOD data is missing values (Table 1). We experiment with several approaches for time series imputation among which linear interpolation performed the best (Table 2). Data is interpolated for each city district separately while ensuring proper placement of data points (sorting by date).

#### 2) METEOROLOGICAL DATA

GFS data variables are selected based on the next investigations. Lit et al. [29] indicate that meteorological factors might considerably affect $PM_{2.5}$ concentration levels, and summarize correlation levels for rainfall, precipitation, wind speed, humidity, wind direction, atmospheric stability, relative humidity, daily average temperature, minimum temperature, and maximum temperature. More specific product descriptions are available in Table 3. GFS data is aggregated concerning the product specifications. Since satellite and GFS data have different resolutions, coordinates for each district in three locations are rounded to the nearest GFS coordinate to join with AOD values.

**TABLE 1.** Missing AOD data statistics.

| Variable | Missing % |
|---|---|
| Optical Depth 055 | 66.33 |
| Optical Depth 047 | 66.34 |
| Column WV | 35.29 |
| AOD Model | 66.34 |
| AOD QA | 21.98 |
| AOD Uncertainty | 35.057 |
| Injection Height | 96.72 |
| Fine Mode Fraction | 100 |

#### 3) FEATURE ENGINEERING

We experiment with additional date features, geographical coordinates as attributes, mean encoding for each grid id, and derive wind magnitude from GFS features ($u$ and $v$ components of wind).

In addition to the main date features, such as *day of the year, month, and year*, we add more: *week of the month, boolean indicators of a start or end of the month, quarter, year*. *Geographical coordinates* are incorporated to include information about regional differences. *Target encoding* has been performed on district ids. Lastly, *wind magnitude* has been computed from appropriate GFS variables in the

**TABLE 2.** Evaluation of various imputation techniques for the data from MCD19A2 product on the validation set with RF model. WS: window size. I: interpolation, K: Kalman, MA: moving average. Descriptions of the used methods can be found in imputeTS R package documentation [30].

| Method | RMSE | $R^2$ |
|---|---|---|
| Linear I | **52.85** | **0.68** |
| $3^{rd}$ order Spline I | 54.34 | 0.66 |
| K Structural (smooth) | 56.67 | 0.63 |
| K Structural (run) | 55.29 | 0.65 |
| K ARIMA State Space (smooth) | 54.14 | 0.66 |
| K ARIMA State Space (run) | 53.79 | 0.66 |
| Exponential MA (WS:7) | 55.79 | 0.64 |
| Exponential MA (WS:15) | 58.68 | 0.60 |
| Simple MA (WS:7) | 54.37 | 0.66 |
| Simple MA (WS:15) | 54.67 | 0.66 |
| Linear MA (WS:7) | 55.43 | 0.64 |
| Linear MA (WS:15) | 56.29 | 0.63 |
| Stine I | 53.98 | 0.66 |
| Mean | 56.65 | 0.63 |
| Mode | 54.56 | 0.65 |
| Median | 55.83 | 0.64 |

**TABLE 3.** Selected GFS products' specifications. Variables are requested for the period between 2015-2022. *: 4 for 24 hours, SFC: ground or water surface, HTGL: specified height above ground, EATM: entire atmosphere (considered as a single layer), A PCP: total precipitation, APTMP: apparent temperature, CRAIN: categorical rain (yes=1; no=0), GUST: wind speed (gust), HGT: geopotential height, PRATE: precipitation rate, PRES: pressure, SUNSD: sunshine duration, TOZNE: total ozone, U GRD: u-component of wind, V GRD: v-component of wind.

| Parameter | Product | Level | Units |
|---|---|---|---|
| A PCP | 6-h Accumulation* | SFC | $kg{\cdot}m^{-2}$ |
| APTMP | 24-h Forecast | HTGL: 2m | K |
| CRAIN | 6-h Average | SFC | - |
| GUST | 24-h Forecast | SFC | $m{\cdot}s^{-1}$ |
| HGT | 24-hour Forecast | SFC | gpm |
| PRATE | 6-h Average | SFC | $kg{\cdot}m^{-2}{\cdot}s^{-1}$ |
| PRES | 24-hour Forecast | SFC | Pa |
| SUNSD | 24-hour Forecast | SFC | s |
| TOZNE | 24-hour Forecast | EATM | Dobson |
| U GRD | 24-hour Forecast | HTGL: 10m | $m{\cdot}s^{-1}$ |
| V GRD | 24-hour Forecast | HTGL: 10m | $m{\cdot}s^{-1}$ |

following way:

$$w_{mag} = \sqrt{u_{wind}^2 + v_{wind}^2}$$

where $u_{wind}$ and $v_{wind}$ are horizontal and vertical components of wind respectively.

The final view of the merged data contains 27 descriptor variables, 34313, and 13505 entries for train and test sets accordingly. 20% of the training set is used for validation.

### E. MODELING

#### 1) LINEAR MODELS

To establish the baseline scores, simple Linear, Lasso, Ridge, Support Vector, and ElasticNet regression models are utilized. Lasso and Ridge are L1 and L2-regularized versions of the linear regression. ElasticNet combines both L1 and L2 priors as regularizers. Support Vector Regression is a counterpart of Support Vector Machines for regression problems.

#### 2) TREE-BASED MODELS

To increase the generalization and predictive performance of the overall pipeline, a combination (averaging of predictions) of Random Forest (RF) and Gradient Boosting (GB) models are used. The main reason behind this choice is that tree-based models remain more suitable for multivariate time-series data and achieve better results compared to other approaches, including DL models [15].

RF is an ensemble of Decision Trees, which are generated in large numbers during the training procedure. Bootstrapping the training dataset, integration of trees of different sizes, and averaging the results help to decrease the likelihood of overfitting. The final result is an average of outputs across all Decision Trees. For the final RF model, Optuna Hyperparameter Tuning Framework has been used for finding optimal parameters. GB [31] builds an ensemble of trees by the sequential process - at each iterative step, a new model is created which takes into account errors of the previous model. Such an approach allows for the production of a more robust decision surface. The final prediction is obtained by an additive strategy.

### F. FEEDBACK

To maintain the bidirectional flow of information we first estimate recommendations upon which further action policies are based.

#### 1) RECOMMENDATION

We use Optuna Hyperparameter Tuning Framework [32] and view AOD variables (influenceable to some extent by human actions) as hyperparameters to optimize for favorable $PM_{2.5}$ levels. The problem formulation can be defined as follows. The objective is to minimize the absolute difference between the prediction of the pollutant level ($y_p$) and favorable value ($y_f$) for the given day:

$$|y_p - y_f|$$

by tweaking AOD parameters ($A$) with reduction ratio ($r$):

$$A \cdot r$$

subject to:

$$0.1 < r < 0.9$$

#### 2) FEEDBACK

We describe three potential policies to follow after receiving a recommendation based on existing environmental studies on $PM_{2.5}$.

In an extensive environmental review on $PM_{2.5}$ [29], the positive impact of meteorological factors on $PM_{2.5}$ mass concentrations has been emphasized. The diffusion process of

**TABLE 4.** Results of $PM_{2.5}$ estimation experiments. AOD: inclusion of AOD variables, GFS: inclusion of GFS variables, OPT: tuned with Optuna Framework (for RF only), ADF: additional date features, GC: geographical coordinates as features, ME: mean encoding of grid ids, WM: wind magnitude. IMP: imputation strategy: D: dropping, LF: 1) last observation carried forward and then 2) next observation carried backward, L: linear interpolation (around 129 non-interpolated entries are dropped). * - AOD 047, # - some columns containing meta information are dropped for this and the next experiments, ↑ - trained on validation and training sets combined.

| Model | AOD | GFS | IMP | OPT | ADF | GC | ME | WM | Val RMSE | Val $R^2$ | Test RMSE | Test $R^2$ (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RF | ✓ | × | D | × | × | × | × | × | 87.70 | 0.376 | 81.01 | -0.175 (-0.192, -0.159) |
| RF | ✓* | × | D | × | × | × | × | × | 88.90 | 0.359 | - | - |
| RF | ✓ | ✓ | D | × | × | × | × | × | 73.21 | 0.570 | 46.76 | 0.608 (0.594, 0.623) |
| RF | ✓ | ✓ | LF | × | × | × | × | × | 54.90 | 0.654 | 44.33 | 0.648 (0.634, 0.662) |
| RF# | ✓ | ✓ | L | × | × | × | × | × | 51.53 | 0.696 | 42.92 | 0.670 (0.657, 0.683) |
| GB | ✓ | ✓ | L | × | × | × | × | × | 50.02 | 0.713 | - | - |
| RF + GB | ✓ | ✓ | L | ✓ | × | × | × | × | 49.34 | 0.721 | 40.82 | 0.702 (0.690, 0.714) |
| RF + GB | ✓ | ✓ | L | ✓ | ✓ | × | × | × | 47.65 | 0.7399 | 43.43 | 0.662 (0.649, 0.675) |
| RF | ✓ | ✓ | L | ✓ | ✓ | ✓ | × | × | 47.72 | 0.740 | 44.32 | 0.648 (0.635, 0.662) |
| GB | ✓ | ✓ | L | ✓ | ✓ | ✓ | × | × | 47.53 | 0.741 | 42.70 | 0.674 (0.660, 0.686) |
| RF | ✓ | ✓ | L | ✓ | ✓ | ✓ | ✓ | ✓ | 47.00 | 0.753 | 43.78 | 0.656 (0.643, 0.670) |
| RF | ✓ | ✓ | L | ✓ | × | ✓ | ✓ | ✓ | 49.20 | 0.722 | - | - |
| GB | ✓ | ✓ | L | ✓ | × | ✓ | ✓ | ✓ | 50.60 | 0.706 | - | - |
| RF + GB | ✓ | ✓ | L | ✓ | × | ✓ | ✓ | ✓ | 49.48 | 0.720 | 40.02 | 0.713 (0.702, 0.725) |
| RF↑ + GB↑ | ✓ | ✓ | L | ✓ | × | ✓ | ✓ | ✓ | - | - | 38.94 | 0.728 (0.717, 0.740) |

the particulates is correlated with high-temperature weather. Rainfall decreases mass concentrations by 56.3% on average, and the effect prolongs up to 72 hours, though within the first hour of rain concentration levels remain intact. The study also reports a negative relationship between precipitation and particulate levels in the air. Wind speed plays a crucial role in the aggravation of pollution levels. Overall, the aforementioned meteorological events substantially impact the diffusion and retention levels of $PM_{2.5}$. The possibility of taking advantage of natural events on days with the highest air pollution levels predicted by the model seems appealing. Climatic events can not be forced in general, however, some attempts have been made. Cloud seeding [33] is a method of artificially increasing condensation in clouds to encourage them to produce rain. The first policy is **planning of cloud seeding in accordance with pollutant level predictions from the model in regions where such technology is used.**

In the same study, [29], 1 or 2 hours delay between outdoor and indoor $PM_{2.5}$ has been indicated. If model predictions can be made available to a broader audience in an easier way with lower latency, this delay can play a crucial role in taking preventive measures. The second policy is **scheduling of specific air filters designed or improved for $PM_{2.5}$ ( [34], [35]) based on the predictions made by the forecasting model.** Proper scheduling can make the usage of such filters more cost-effective.

In general, $PM_{2.5}$ mainly originates from combustion: car engines, heating fuel, natural gas, or coal-fired power plants. The COVID-19 pandemic lockdown allowed us to quantize to some extent their broadly known effects on air quality. Shi et al. [36] reported 35% reduction in the surface $PM_{2.5}$ concentration in China, and also decrease of 31% for $PM_{10}$ in Barcelona (Spain) [37] and 43% for $PM_{2.5}$ in India [38] has been observed during the pandemic. Though lockdown can not be forced due to air pollution, techniques for mitigating hazardous effects based on $PM_{2.5}$ models' predictions can be built upon and their effects can be numerically approximated from these studies. The third policy is **to encourage citizens to switch to public transport via phone notifications on days with predicted highest pollutant agglomeration.**

### G. FEATURE IMPORTANCES

To establish a better understanding of the outputs of the model, different approaches for quantitatively examining the contribution of each independent variable are investigated.

#### 1) SHAP

Shapley Additive Explanations (SHAP) [14] is a method for interpreting predictions of any model based on Shapley values from game theory which reflects the contribution of each feature. The shapely values are determined through (1). Given a model where a group $N$ (with $n$ features) is used to predict an output ($N$), the contribution of each feature $i$ ($\phi_i$) on the model output, v($N$) is determined based on their marginal contribution. The main idea is to connect local explanations with optimal credit allocation. Intuitively idea can be explained as investigating each feature's effect in isolation.

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [\mathrm{v}(S \cup \{i\}) - \mathrm{v}(S)] \quad (1)$$

The additive feature attribution method is used to define a linear function of binary features $g$:

$$\mathrm{g}\left(z'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \quad (2)$$

where $z' \in \{0, 1\}^M$, equals 1 when a feature is observed, otherwise it equals 0, and $M$ is the number of input features.

#### 2) RF IMPORTANCES

Importance scores based on RF are computed as the accumulation (mean, std) of how much each feature has contributed to the decrease in variance at each split in each tree. In decision trees, each node represents an optimal data-splitting condition to decrease variance within a split group. This allows us to determine how significant each feature is for the given task and dataset.

### H. METRICS

Root-mean-square error (RMSE) and $R^2$ are used as the main evaluation metrics for forecasting accuracy in this work. RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}$$

where $y_i$ and $f_i$ represent ground truth values and predicted values respectively. $R^2$ indicates the proportion of variances in the dependent variable explained by independent variables and is defined as follows:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where RSS is the residual sum of squares and TSS is the total sum of squares.

## IV. RESULTS AND EVALUATION

Results of experiments for establishing baseline scores with the linear models are provided in Table 5. Maximum $R^2$ of 0.47 was achieved by SVR with radial basis function kernel.

Results of the main experiments are summarized in Table 4. Training the RF model with default hyperparameters with only AOD data, in which entries with any missing values are dropped, yields RMSE and $R^2$ of 87.70, 0.376, and 81.01, -0.175 on the validation set and test set respectively. Even though GFS data has a lower resolution, adding meteorological features significantly boosts the model's generalization capability resulting in the test RMSE and $R^2$ of 46.76 and 0.608 (95% CI, 0.594-0.623). Such impact underlines the importance of incorporating climatic descriptors in $PM_{2.5}$ estimation systems. Further, missing points were imputed with linear interpolation. Combining predictions of RF, tuned with Optuna framework, and GB results in lower RMSE (-6) and increased $R^2$ (+0.1) on the test set. Additional feature engineering helps to explain 72% of the variance in the test set and to decrease RMSE below 39.

The most significant variables (7) based on RF feature importances (Figure 3) and SHAP values (Figure 4) for the given data and task are grid ids encoded with global mean values, pressure, month in the year, precipitation rate per day, apparent temperature, sunshine duration and a categorical indication of rain for the given day. Mean encoding incorporates historic global values into the model's decision-making process, which might explain its highest contribution score.

**TABLE 5. Summary of results for linear models on the validation set.**

| Model | $R^2$ |
|---|---|
| LR | 0.384 |
| SVR (RBF kernel) | 0.476 |
| SVR (Linear kernel) | 0.406 |
| SVR (Polynomial kernel) | 0.415 |
| Lasso | -0.049 |
| Ridge | 0.337 |
| ElasticNet | -0.049 |

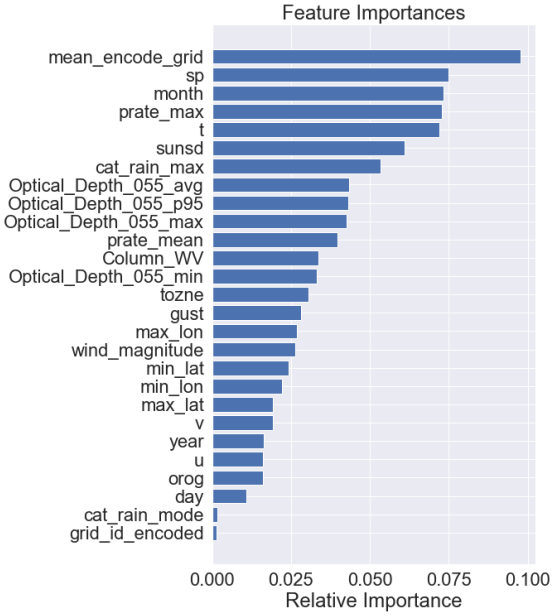


**FIGURE 3. Random Forest feature importances. Descriptions of feature names can be found in Table 6.**

For AOD variables, interpolation of most of their values might have decreased their importance scores.

As we do not construct criteria or protocols for assessing the effectiveness of the policies, quantitative evaluations for them are not provided. Explanations are further outlined in the discussion section of this work.

## V. DISCUSSION

In this section, we outline the limitations of this work and related future research directions.

The recommendation-feedback part of the overall pipeline requires means of assessment. Such a method based on simulation or any other technique was not implemented, which does not allow for objectively evaluating proposed action policies. In addition, making this part of the DT more intelligent by using data-driven approaches, or more complex models, such as Neural Networks seems like an interesting direction to investigate further. For example, MLP can be used to formulate the inverse problem for making more reliable recommendations. Besides, there is no computationally established relationship between recommendations and policy to follow.

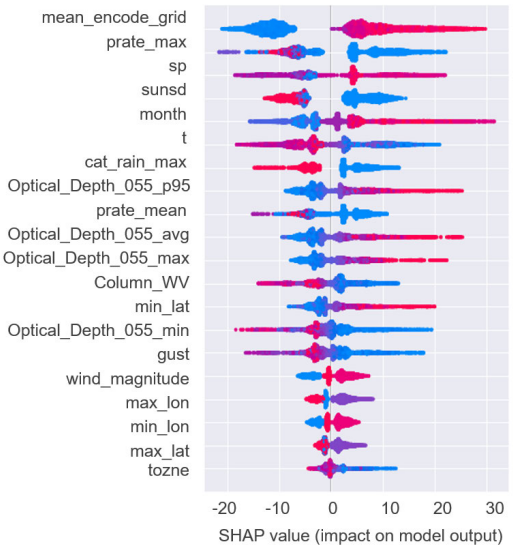


**FIGURE 4.** SHAP feature importances. Descriptions of feature names can be found in Table 6.

**TABLE 6.** Descriptions of feature names.

| Feature | Explanation (GFS variable if applicable) |
|---|---|
| mean_encode_grid | Encoded district id (mean of $PM_{2.5}$ levels) |
| prate_max | Maximum precipitation rate (PRATE) |
| sp | Pressure (PRES) |
| sunsd | Sunshine duration (SUNSD) |
| month | Month of the year |
| t | Temperature (APTMP) |
| cat_rain_max | Categorical indication of raining (CRAIN) |
| Optical_Depth_055_p95 | 95th percentile for AOD value |
| prate_mean | Average precipitation rate (PRATE) |
| Optical_Depth_055_avg | Mean AOD value |
| Optical_Depth_055_max | Maximum AOD value |
| Column_WV | Column water vapor |
| min_lat | South latitude |
| Optical_Depth_055_min | Minimum AOD value |
| gust | Speed of wind gusts (GUST) |
| wind_magnitude | Magnitude of wind |
| max_lon | East longitude |
| min_lon | West longitude |
| max_lat | North latitude |
| tozne | Total ozone (TOZNE) |

For the pollutant level estimation part, investigated deep learning methods (RNN, CNN-LSTM, Temporal Fusion Transformer [39]) did not produce results better than baseline linear models. Recent study [15] compares these architectures to tree-based models and reports possible reasons for their low performance. Based on their findings, DL models can be theoretically better adapted for the task considered in this work, however, it remains one of the promising research directions.

## VI. CONCLUSION

In this work, we build a digital twin of an atmospheric environment by fusing 1 km resolution satellite data (AOD) and 27 km (0.25 degree) resolution meteorological data (GFS). We use estimated values of $PM_{2.5}$ for 33, 14, and 7 $5 \times 5$ km districts in Los Angeles, Delhi, and Taipei to provide recommendations for decreasing them to more healthy levels by formulating a simple optimization problem. Further, possible policy actions to follow are investigated. The estimation part of the DT pipeline has achieved RMSE and $R^2$ of 38.94 and 0.728 (95% CI, 0.717-0.740). The most influential parameters for the given dataset are identified based on two relevant algorithms. Recommendation and feedback parts require additional investigation for constructing evaluation techniques and will be future research directions to focus on.

## ACKNOWLEDGMENT

*(Kudaibergen Abutalip and Anas Al-lahham contributed equally to this work.)*

## REFERENCES

[1] *Who Global Air Quality Guidelines: Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Accessed: Feb. 15, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240034228
[2] G. Hoek, R. M. Krishnan, R. Beelen, A. Peters, B. Ostro, B. Brunekreef, and J. D. Kaufman, "Long-term air pollution exposure and cardio—Respiratory mortality: A review," *Environ. Health*, vol. 12, no. 1, p. 43, Dec. 2013.
[3] R. D. Brook, S. Rajagopalan, C. A. Pope, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, A. Peters, D. Siscovick, S. C. Smith, L. Whitsel, and J. D. Kaufman, "Particulate matter air pollution and cardiovascular disease," *Circulation*, vol. 121, no. 21, pp. 2331–2378, 2010.
[4] Y.-L. Zhang and F. Cao, "Fine particulate matter ($PM_{2.5}$) in China at a city level," *Sci. Rep.*, vol. 5, no. 1, p. 14884, Oct. 2015.
[5] Y. Chu, Y. Liu, X. Li, Z. Liu, H. Lu, Y. Lu, Z. Mao, X. Chen, N. Li, M. Ren, F. Liu, L. Tian, Z. Zhu, and H. Xiang, "A review on predicting ground $PM_{2.5}$ concentration using satellite aerosol optical depth," *Atmosphere*, vol. 7, no. 10, p. 129, Oct. 2016. [Online]. Available: https://www.mdpi.com/2073-4433/7/10/129
[6] M. Z. Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "$PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data," *Atmosphere*, vol. 10, no. 7, p. 373, Jul. 2019.
[7] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the XGBoost machine learning method in $PM_{2.5}$ prediction: A case study of Shanghai," *Aerosol Air Quality Res.*, vol. 20, no. 1, pp. 128–138, 2020.
[8] A. Masood and K. Ahmad, "A model for particulate matter ($PM_{2.5}$) prediction for Delhi based on machine learning approaches," *Proc. Comput. Sci.*, vol. 167, pp. 2101–2110, Jan. 2020.
[9] M. D. Yazdi, Z. Kuang, K. Dimakopoulou, B. Barratt, E. Suel, H. Amini, A. Lyapustin, K. Katsouyanni, and J. Schwartz, "Predicting fine particulate matter ($PM_{2.5}$) in the greater London area: An ensemble approach using machine learning methods," *Remote Sens.*, vol. 12, no. 6, p. 914, 2020.
[10] T. Li, M. Hua, and X. Wu, "A hybrid CNN-LSTM model for forecasting particulate matter ($PM_{2.5}$)," *IEEE Access*, vol. 8, pp. 26933–26940, 2020.
[11] U. Pak, J. Ma, U. Ryu, K. Ryom, U. Juhyok, K. Pak, and C. Pak, "Deep learning-based $PM_{2.5}$ prediction considering the spatiotemporal correlations: A case study of Beijing, China," *Sci. Total Environ.*, vol. 699, Jan. 2020, Art. no. 133561. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969719334813
[12] Y.-C. Chen and D.-C. Li, "Selection of key features for $PM_{2.5}$ prediction using a wavelet model and RBF-LSTM," *Int. J. Speech Technol.*, vol. 51, no. 4, pp. 2534–2555, Apr. 2021.
[13] H. Chang-Hoi, I. Park, H.-R. Oh, H.-J. Gim, S.-K. Hur, J. Kim, and D.-R. Choi, "Development of a $PM_{2.5}$ prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, republic of Korea," *Atmos. Environ.*, vol. 245, Jan. 2021, Art. no. 118021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231020307548

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 4768–4777.

[15] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" 2022, *arXiv:2207.08815*.

[16] M. Grieves, "Digital twin: Manufacturing excellence through virtual factory replication—A whitepaper by dr. Michael grieves," Florida Inst. Technol., Melbourne, FL, USA, White Paper 1–7, 2015.

[17] M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in *Transdisciplinary Perspectives on Complex Systems*. Berlin, Germany: Springer, 2017, pp. 85–113.

[18] A. E. Saddik, "Digital twins: The convergence of multimedia technologies," *IEEE Multimed.*, vol. 25, no. 2, pp. 87–92, Apr./Jun. 2018.

[19] E. Glaessgen and D. Stargel, "The digital twin paradigm for future NASA and US air force vehicles," in *Proc. 53rd AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn. Mater. Conf. 20th AIAA/ASME/AHS Adapt. Struct. Conf.*, 2012, p. 1818, doi: 10.2514/6.2012-1818.

[20] Y. Liu, L. Zhang, Y. Yang, L. Zhou, L. Ren, F. Wang, R. Liu, Z. Pang, and M. J. Deen, "A novel cloud-based framework for the elderly healthcare services using digital twin," *IEEE Access*, vol. 7, pp. 49088–49101, 2019.

[21] K. Reifsnider and P. Majumdar, "Multiphysics stimulated simulation digital twin methods for fleet management," in *Proc. 54th AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn., Mater. Conf.*, Apr. 2013, p. 1578.

[22] C. Pylianidis, S. Osinga, and I. N. Athanasiadis, "Introducing digital twins to agriculture," *Comput. Electron. Agricult.*, vol. 184, May 2021, Art. no. 105942.

[23] B. Ketzler, V. Naserentin, F. Latino, C. Zangelidis, L. Thuvander, and A. Logg, "Digital twins for cities: A state of the art review," *Built Environ.*, vol. 46, no. 4, pp. 547–573, Dec. 2020.

[24] Y. Jiang, S. Yin, K. Li, H. Luo, and O. Kaynak, "Industrial applications of digital twins," *Philos. Trans. Roy. Soc. A*, vol. 379, no. 2207, 2021, Art. no. 20200360.

[25] R. Rosen, G. Von Wichert, G. Lo, and K. D. Bettenhausen, "About the importance of autonomy and digital twins for the future of manufacturing," *IFAC-Papers OnLine*, vol. 48, no. 3, pp. 567–572, 2015.

[26] A. K. Aap, M. S. Aap, R. Sabha, and L. Sabha, "Anil Baijal, IAS [4]," *Population*, p. 6, 2011.

[27] S. K. Sahu and S. H. Kota, "Significance of PM2.5 air quality at the Indian capital," *Aerosol Air Quality Res.*, vol. 17, no. 2, pp. 588–597, 2017.

[28] Q. Koziol, *HDF5*. Boston, MA, USA: Springer, 2011, pp. 827–833, doi: 10.1007/978-0-387-09766-4_44.

[29] *India Expo Centre & Mart Greater Noida: Information About Delhi*. UNCCD. Accessed: May 8, 2022. [Online]. Available: http://unccdcop14india.gov.in/about-delhi

[30] S. Moritz and T. Bartz-Beielstein, "ImputeTS: Time series missing value imputation in R," *R J.*, vol. 9, no. 1, pp. 207–218, 2017.

[31] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[32] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.

[33] T. Al Hosari, A. Al Mandous, Y. Wehbe, A. Shalaby, N. Al Shamsi, H. Al Naqbi, O. Al Yazeedi, A. Al Mazroui, and S. Farrah, "The UAE cloud seeding program: A statistical and physical evaluation," *Atmosphere*, vol. 12, no. 8, p. 1013, Aug. 2021. [Online]. Available: https://www.mdpi.com/2073-4433/12/8/1013

[34] S. Ma, M. Zhang, J. Nie, J. Tan, B. Yang, and S. Song, "Design of double-component metal–organic framework air filters with $PM_{2.5}$ capture, gas adsorption and antibacterial capacities," *Carbohydrate Polym.*, vol. 203, pp. 415–422, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0144861718310944

[35] Z. Wei, Q. Su, X. Wang, S. Long, G. Zhang, Q. Lin, and J. Yang, "Nanofiber air filters with high-temperature stability and superior chemical resistance for the high-efficiency $PM_{2.5}$ removal," *Ind. Eng. Chem. Res.*, vol. 60, no. 27, pp. 9971–9982, Jul. 2021, doi: 10.1021/acs.iecr.1c01821.

[36] X. Shi and G. P. Brasseur, "The response in air quality to the reduction of Chinese economic activities during the COVID-19 outbreak," *Geophys. Res. Lett.*, vol. 47, no. 11, Jun. 2020, Art. no. e2020GL088070. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL088070

[37] A. Tobías, C. Carnerero, C. Reche, J. Massagué, M. Via, M. C. Minguillón, A. Alastuey, and X. Querol, "Changes in air quality during the lockdown in Barcelona (Spain) one month into the SARS-CoV-2 epidemic," *Sci. Total Environ.*, vol. 726, Jul. 2020, Art. no. 138540. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969720320532

[38] S. Sharma, M. Zhang, J. Gao, H. Zhang, and S. H. Kota, "Effect of restricted emissions during COVID-19 on air quality in India," *Sci. Total Environ.*, vol. 728, Aug. 2020, Art. no. 138878. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969720323950

[39] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021000637

**KUDAIBERGEN ABUTALIP** was born in Zhetisay, Kazakhstan, in 1999. He received the bachelor's degree in mathematical and computer modeling from International Information Technology University, Almaty, Kazakhstan, in 2017. He is currently pursuing the master's degree in computer vision with the Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates. His research interests include digital twins and applications of computer vision in various domains, including medical imaging analysis.

**ANAS AL-LAHHAM** received the bachelor's degree from King Saud University, Saudi Arabia, with a thesis titled "Sky-Imager Based Forecast of Solar Irradiance Using Machine Learning." He is currently pursuing the M.Sc. degree in computer vision with the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates. His research interests include computer vision, signal and image processing, renewable energy, and digital twins. He received the Full Scholarship for his bachelor's degree.

**ABDULMOTALEB EL SADDIK** (Fellow, IEEE) is a Distinguished Professor with the School of Electrical Engineering and Computer Science, University of Ottawa, and a Professor with the Mohamed bin Zayed University of Artificial Intelligence. He has supervised more than 150 researchers. He has coauthored ten books and more than 550 publications and chaired more than 50 conferences and workshops. He received research grants and contracts totaling more than $22 million. His research interests include the establishment of digital twins to facilitate the wellbeing of citizens using AI, the IoT, AR/VR, and 5G to allow people to interact in real time with one another as well as with their smart digital representations in the metaverse. He is a fellow of Royal Society of Canada, an ACM Distinguished Scientist, and a fellow of the Engineering Institute of Canada and the Canadian Academy of Engineers. He received several international awards, such as the IEEE I&M Technical Achievement Award, the IEEE Canada C.C. Gotlieb (Computer) Medal, and the A.G.L. McNaughton Gold Medal for important contributions to the field of computer engineering and science.