

A Hybrid GRU-SVR Ensemble Framework for Daily PM2.5 Forecasting and Adaptive Air Quality Management in Pakistan

Arslan Murtaza ^{1,*}, Muhammad Noman ², Hafiz Attaullah Saqi ³, Ali Fazeel Tahir ⁴ and Iqra Akhtar ⁵

¹ Department of Physics, KFUEIT, Rahim Yar Khan, 64200, Pakistan

² Department of Cyber Security, KFUEIT, Rahim Yar Khan, 64200, Pakistan

³ Manchester Metropolitan University, Manchester, United Kingdom

⁴ Department of Materials, Faisalabad University, Faisalabad, Pakistan

⁵ Department of Computer Engineering, KFUEIT, Rahim Yar Khan, 64200, Pakistan

*Corresponding author: Arslan Murtaza (Email: arslangill1182@gmail.com)

Received: 12/07/2026, Revised: 22/08/2026, Accepted: 30/08/2026

Abstract— Precise short-term forecasting of PM2.5 levels is crucial for building effective early warning systems and reducing public health hazards, especially in countries like Pakistan, where the combination of temperature inversions and burning of agricultural waste leads to severe smog events during winter. Motivated by these challenges, proposed holistic PM2.5 forecasting framework based on a dataset comprising 2,193 daily air quality monitoring records. An empirical study is conducted on eight ML regression models, five DL sequence models (LSTM, GRU, BiLSTM, CNN1D, and windowed MLP), and two proposed ensemble models called AeroSynNet and AeroStackNet. The combination of GRU and SVR in AeroSynNet model is achieved via weighted averaging, whereas AeroStackNet combines GRU, CNN1D, SVR, and LightGBM, with Ridge Regression serving as the meta-modeller. From experimental results, it is evident that Linear Regression achieved best overall performance, closely followed by proposed AeroStackNet ensemble, with Linear Regression's RMSE equal to 3.866, MSE equal to 14.944, MAE to 2.903, and R2 value 1 of 0.9050 (AeroStackNet: RMSE 3.885, R2 0.9040). After correcting ensemble weight-tuning procedure to use held-out validation split rather than test set, proposed AeroSynNet framework no longer ranked first; the Wilcoxon Signed-Rank Test confirmed that Linear Regression, GRU, and AeroStackNet were statistically indistinguishable from one another, while AeroSynNet performed significantly worse than each of them. However, the added complexity of AeroStackNet provides a notable performance gain over the simpler AeroSynNet blend. Linear Regression itself matched or exceeded the accuracy of every Deep Learning and ensemble model at a fraction of the computational cost. From seasonal error analysis, it was observed that largest forecast errors occurred in winter, owing to smog. Collectively, study shows that a well-engineered Linear Regression baseline, together with proposed ensemble, provides an accurate, efficient, and statistically robust solution, whereas simpler two-model approach proved more sensitive to tuning of its weights.

Index Terms—Air quality and human health, Air quality Index, Ensemble learning, PM2.5 forecasting, Time-series analysis.

I. INTRODUCTION

Air pollution due to fine particulate matter (PM2.5; aerodynamic diameter $\leq 2.5\mu\text{m}$) has become one of the detrimental environmental and public health threats that developing countries are facing [1][2]. Urbanization, industrialization, automobile exhaust fumes, biomass burning, and weather events like inversion in winters have caused the pollution due to fine particles called PM2.5 to grow in Pakistan, which ranks among the most polluted countries of the world [3].

According to the 2024 World Air Quality Report by IQAir, Pakistan was the third most polluted country in the world, as the concentration of PM2.5 air pollutants was $73.7\mu\text{g}/\text{m}^3$, which was much higher (more than 15 times) than the annual WHO guideline of $5\mu\text{g}/\text{m}^3$ [4]. Extremely high levels of air pollution pose a serious danger to people's health, environmental safety, and the region's socio-economic progress, underscoring the importance of developing an advanced air quality forecasting system [5].

According to recent findings from the 2025 Air Quality Life Index (AQLI) Annual Update, particulate pollution lowers life expectancy by 3.3 years in Pakistan, while more severely affected areas experience much larger losses: 5.8 years in Lahore, 4.5 years in Islamabad, and more than 5 years in Peshawar. All 248.8 million people reside in areas where the WHO guideline is exceeded, and 99.5% exceed the new national standard (from $15\mu\text{g}/\text{m}^3$ to $35\mu\text{g}/\text{m}^3$ in 2023) [6].

Public health issues are huge. Chronic exposure to high



concentrations of PM_{2.5} has scientific evidence that confirms it leads to respiratory diseases (chronic obstructive pulmonary disorder, asthma attacks), heart problems (stroke, ischemic heart disease), lung cancer, and pregnancy-related complications. In Pakistan, air pollution is one of the main factors behind the excessive deaths because of pollution, and according to the data provided by the Global Burden of Diseases, it has caused several thousand extra deaths per year in cities such as Lahore, Karachi, Faisalabad, and Islamabad. Economic consequences can be described as follows: health care expenses, reduced productivity, increased costs of health care. During winter smog periods exacerbated by burning of agricultural crops after monsoon season in Punjab province and stable weather conditions, hourly levels of PM_{2.5} usually exceed 200-300 $\mu\text{g}/\text{m}^3$ [7].

Despite these disadvantages, Pakistan's air quality ground network remains insufficient in both coverage and distribution. Reliable, continuous stations for air quality measurements are few, are located in big cities, and often experience maintenance issues or missing data. This limits the possibility of building effective forecasting and warning systems. On the other hand, satellite-based reanalysis datasets, like MERRA-2 (bias-corrected using convolutional neural networks in the HAQAST product), may be considered a good replacement.

Recent research in Pakistan has explored PM_{2.5} estimation and prediction using machine learning (ML) and deep learning (DL). [8] applied convolutional neural networks (CNNs) on satellite imagery (e.g., Sentinel-5P TROPOMI) 3 for urban PM_{2.5} estimation [8]. In Tariq et al. (2025) [9] the combination of high-resolution satellite images with ML methods was applied, and in Ullah et al. (2026) [10-14], fusion techniques with the integration of satellite, ground sensors and meteorological variables were suggested using ensemble models [9]. In Waqas et al. (2025) ML and DL models including LSTM and GRU were compared [10]. Also, Munir et al. (2024) researched smog formation using ML techniques [7].

However, the research efforts mentioned above are concerned with spatial estimation or city-specific analysis at mostly daily interval levels. Daily, nationwide forecasting crucial for real-time decision making has not been adequately studied especially using solely temporal information derived from satellite data. Pure machine learning techniques such as LSTM and GRU face difficulties in capturing the complexity of the nonlinear pollution time series data due to lagging effects, seasonality, and trends. Conventional ML methods, such as SVM and Random Forest, have limited sequential learning abilities, while Transformer and 1D CNN methods are scarcely used. In order to overcome these problems, the current research uses a combination of eight traditional machine learning regressors, five DL-based sequence learners (LSTM, GRU, Bidirectional LSTM, CNN1D, and windowed MLP) and two unique hybrid ensembles proposed by this research, as AeroSynNet and AeroStackNet, on the daily satellite-derived PM_{2.5} data set of Pakistan. This data set has been further improved through the use of temporal feature engineering which includes lag features ($t-1$ and $t-2$), rolling average features (3-day and 7-day) and seasonal calendar features

(month, day, and season). AeroSynNet is an ensemble of GRU and Support Vector Regression through weighted averaging, while AeroStackNet uses a stacked ensembling approach to incorporate GRU, CNN1D, SVR and LightGBM through Ridge Regression meta-learner.

II. PRIMARY CONTRIBUTION

The main contributions of this research are summarized as follows:

1. A systematic benchmarking procedure is provided to compare eight machine learning models, five deep learning models, and two newly proposed hybrid ensemble models in the same time-based validation approach for PM_{2.5} forecasting on a daily basis.
2. A new hybrid model, called AeroSynNet, is proposed by combining GRU and Support Vector Regression (SVR) with the help of weighted average approach to leverage from temporal sequence learning and nonlinear regression approaches.
3. Another hybrid framework, named as AeroStackNet, is presented by proposing a stacked ensemble model that uses prediction of GRU, CNN1D, SVR, and LightGBM models with the help of Ridge Regression as meta-learner to explore the potential of multi-model fusion approach.
4. Experimental findings reveal that Linear Regression provides the best overall performance in terms of forecasting, giving an RMSE of 3.866, MSE of 14.944, MAE of 2.903, and R² of 0.9050, closely followed by the proposed AeroStackNet ensemble; the proposed AeroSynNet framework, once its blend weight is tuned on a held-out validation split rather than the test set, does not outperform these two models.
5. The proposed model's robustness is established with the help of statistical significance testing and seasonal error analysis.

To our knowledge, this is among the first studies to jointly benchmark eight machine learning models, five deep learning models, and two newly proposed hybrid ensembles under an identical chronological validation protocol using only satellite-derived temporal features for Pakistan, and it is intended to form a practical basis for future air quality management efforts in the country.

III. LITERATURE REVIEW

Air quality forecasting has made rapid strides in recent years owing to the advent of satellite remote sensing data combined with meteorology and machine learning (ML) and deep learning (DL) models. Air pollution is a common problem in South Asia due to various anthropogenic sources such as biomass burning, industrial and vehicular emissions, and meteorological factors. However, satellite-based PM_{2.5} data products like the MODIS Aerosol Optical Depth (AOD), Sentinel-5P multi-pollutant imagery, and MERRA-2 reanalysis serve an important purpose in regions that lack ground-based monitors [8][13]. Initial work in Pakistan and surrounding areas used spatial estimation of satellite aerosol optical depth (AOD) and multi-pollutant imaging. The authors in [13] proposed a convolutional neural network (CNN) model that used the Sentinel5P TROPOMI multi-pollutants image, which included Aerosol Index, CH₄, CO, HCHO, NO₂, O₃, and SO₂ to

forecast the daily concentration of ground-level PM_{2.5} for major cities of Pakistan like Islamabad, Lahore, Peshawar, and Karachi. This P-CNN outperformed other CNN models, such as AlexNet, VGG16, and ResNet50, by achieving lower MAE, RMSE, and MAPE values for spatial pattern extraction from satellite data. However, their work was conducted only on a daily basis and at the city level, not nationally [8]. Ullah et al. (2026) have utilized MODIS Land Surface Reflectance, Ground Sensors, and Meteorology Parameters with the use of Machine Learning Models such as Random Forest, Extra Trees, and Gradient Boosting for predicting PM_{2.5} in Islamabad, where Gradient Boosting gave RMSE $\approx 14.87 \mu\text{g}/\text{m}^3$ and $R^2 = 0.73$. Similarly, Tariq et al. (2025) [9] have integrated high-resolution satellite data with machine learning models and ground validation, primarily focusing on reducing bias [14]. However, in contrast, Waqas et al. [15] have focused on machine learning models (Decision Tree, Random Forest) and deep learning models (MLNN, LSTM, RNN, GRU), where MLNN proved superior ($R^2=0.88$) due to its ability to model nonlinear and time varying relationships [10]. Machine learning techniques were used by Munir et al. [16] for air quality forecasting with particular attention to smog formation in Pakistan keeping into account the seasonal and emission-based trends [7]. An ensemble model of STL using SVR, ELM, and NNAR for predicting PM_{2.5} was used by Qureshi et al. [17] in major cities of Pakistan with an average RMSE of about $10.91 \mu\text{g}/\text{m}^3$ [11].

Khan et al. [12] analyzed the PM_{2.5} estimation and forecasting process using statistical and deep learning methodologies, while Aasim et al. [15] developed intelligent systems based on artificial intelligence to predict pollution in Pakistan. Additional research that provides more regional and global context includes deep learning-based models for PM_{2.5} prediction, proposed by Pak et al. [13], which were based on the spatiotemporal relationship, and satellite and ground-based techniques used by Sohrabi and Maleki to model PM_{2.5} in Iran [17]. Using machine learning techniques for predicting PM_{2.5}, Salehie et al. studied PM_{2.5} levels in Afghanistan [18], while Patel et al. emphasized the importance of hybrid ensemble models in describing nonlinear pollution dynamics [18, 19]. Using machine learning/deep learning techniques along with satellite and meteorological data, Thapa et al. studied PM_{2.5} pollution [20], and using NASA powered machine learning systems, Seo et al. predicted global PM_{2.5} pollution [21]. Some additional research on machine learning further underscores the relevance of machine learning to predicting air pollution. Haseeb et al. [22] used LR (Logistic Regression), RT (Regression Tree), SVM (Support Vector Machine), ERT (Extremely Randomized Trees), GPR (Gaussian Process Regression), and ANN (Artificial Neural Networks) to predict PM_{2.5} in Lahore, where ANN had the highest R^2 of 0.99 [22]. Models like SARIMA, LSTM, and Nonlinear Autoregressive (NAR) were used by Mujtaba et al. (2025) for long-range air quality prediction with an expectation that AQI will rise by 13% till 2030 [23]. Video-based deep learning was proposed by Ahmed et al. (2025) for cross-city PM_{2.5} prediction with $R^2 = 0.97$ [24]. The authors applied ensemble machine learning algorithm (BG-XGBoost) along with CMIP6 and satellite AOD for long-range PM_{2.5} related visibility forecasting in Pakistan [25].

There are still some limitations. For instance, most researches conducted regarding Pakistan only focus on daily and city-level prediction while making use of the fusion of ground based data or multi-pollutant imagery [7][9]. There is an insufficient development when it comes to the use of only satellite derived temporal features such as lags and rolling averages for national-scale daily forecasting from MERRA-2 HAQAST. The use of recurrent models such as LSTM and GRU performs well in terms of temporal dependency; however, it struggles with the complex nonlinearity interaction and cumulative effects of pollution. Classical machine learning algorithms like SVM and random forest have proven to be excellent for providing accurate predictions but does not possess any kind of sequential learning capability [11]. Although ensemble approaches improve accuracy, multi-architectural stacking frameworks that combine diverse model strengths for robust generalization remain limited in the context of Pakistan.

IV. RESEARCH GAP

1. Despite significant advancements in PM_{2.5} forecasting, several critical gaps remain in the existing literature, particularly in the context of Pakistan:
2. Most of the existing literature on the subject in Pakistan focuses on either daily resolution or city-wise estimations and predictions of PM_{2.5}, with little attention given to the daily dynamics at the country level.
3. Most studies focus on utilizing time-series characteristics based on imagery or multiple pollutant input data instead of using time-series characteristics generated by satellite reanalysis data alone, such as lag, rolling window average, seasonal/cyclical components from databases like MERRA-2 HAQAST.
4. Despite the extensive use of LSTM, GRU, CNN, Transformer individually, there is a lack of research on the comparative performance of those different models using the same satellite-derived daily dataset. Most studies fail to benchmark themselves against each other.
5. Ensemble methods such as STL or gradient boosting enhance the prediction accuracy but generally do not leverage the different inductive biases together.

V. PROPOSED METHODOLOGY

The methodology is divided into a series of sequential phases to guarantee reproducibility and a systematic evaluation process with rigorous model comparisons.

Pipeline for reproducibility: Initially, the raw MERRA-2 HAQAST CNN-corrected PM_{2.5} hourly data were filtered through quality filter based on QFLAG variable, only keeping those observations that had $QFLAG \geq 3$. After this, the observations were spatially aggregated within geographical bounding box of Pakistan ($23-37^\circ \text{N}$, $60-77^\circ \text{E}$) to obtain PM_{2.5} at country level. This data was further temporally aggregated into daily national level averages. Missing data was handled using forward filling and rolling mean imputation, wherever appropriate. To avoid data leakage, IQR-based outlier detection and imputation were performed using IQR based on only the training set data. Feature engineering was done using creation of lagged features, rolling mean features, and temporal features

created using date information. The dataset was split chronologically into 80/20 train/test split. The full pipeline for reproducibility is shown in Fig. 1.

This process leads to the creation of AeroSynNet, a heterogeneous ensemble specifically designed for effective PM_{2.5} prediction in the context of Pakistan as illustrated in Fig. 1. The experiments are coded in Python and utilize libraries such as pandas and NumPy for data management, scikit-learn for traditional machine learning approaches, and TensorFlow/Keras for the construction of deep learning models. The methodology is designed to be expandable and incorporate additional environmental factors such as the inclusion of meteorological variables from the ERA5 dataset [7][26][27].

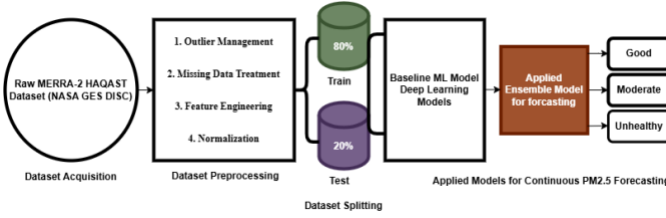


Fig. 1. Proposed methodology workflow for the overall PM_{2.5} forecasting framework.

A. Study Area

The selected study region consists of the whole geographical region of Pakistan as shown in Fig. 2, a country located in South Asia, and known for the presence of varied geographical features, climate, and sources of pollution, which makes it an ideal place for conducting the air quality analysis through remote sensing data, as presented in Figure 2. Geographically, Pakistan spans from approximately 23°35' N to 37°05' N latitude and 60°50' E to 77°50' E longitude, occupying a total land area of around 881,913 km² [28]. It shares its eastern borders with India, western borders with Afghanistan and Iran, northern borders with China, and the southern border with the Arabian Sea. Geographical features of the region include the northern highlands (Hindu Kush, Karakoram, and Himalayas), central plains of the Indus River (covering about 65% of the total land area), and the Balochistan Plateau in the southwest [29]. As far as the administrative divisions of the country are concerned, Pakistan is divided into four provinces namely, Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan, besides the Islamabad Capital Territory (ICT), Azad Jammu and Kashmir, and Gilgit-Baltistan [26]. Punjab, where around 53% of the total population lives, acts as an important source of pollution owing to its high level of agriculture, crop residue burning, industries, and vehicular pollution [7]. Indus river basin plain acts as an important channel for pollution because pollutants get trapped owing to inversion of temperatures in winters [28].

For extracting satellite-based data (like MERRA-2 HAQAST or similar reanalysis datasets), the study area is confined within the geographic boundaries of roughly 23°-37°N latitude and 60°-77°E longitude. This geographical range allows the integration of hourly and daily time series PM_{2.5} concentration data at the country level without needing any ground-level monitoring stations [28][29]. Air pollution hotspots are

primarily concentrated in Punjab and Sindh provinces, particularly in urban-industrial centers. Major cities include Lahore (Punjab), Karachi (Sindh), Islamabad (ICT) and Faisalabad (Punjab). Other hotspots such as Gujranwala, Multan, Peshawar, and Rawalpindi, where PM_{2.5} levels frequently exceed both national and WHO standards [8][10][29].

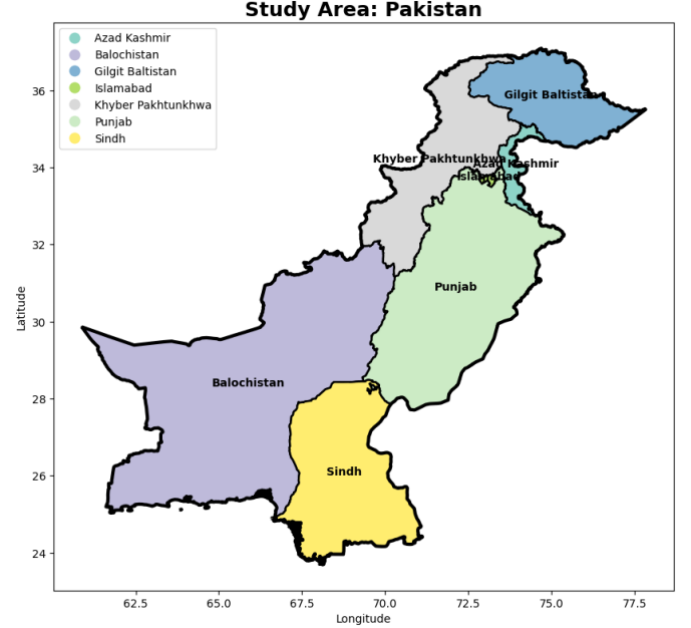


Fig. 2. Study area encompasses the entire territory of Pakistan.

The regions show high seasonal variation. PM_{2.5} concentration is generally high in winter months (November to February) because of temperature inversion, slow winds, and the burning of the crop residue post monsoon [7][29]. Lower concentrations, however, are seen in monsoon months (June to September) due to rain washout, while dust influences are present in spring and summer months coming from the arid western areas. The justification of the national scale analysis for the study is that there is an acute air pollution problem in Pakistan, where recent estimates reveal an average annual PM_{2.5} concentration of 73.7 µg/m³, much higher than the World Health Organization (WHO) guidelines of 5 µg/m³. This, along with insufficient ground monitoring networks, makes it all the more necessary to use satellite-based reanalysis products, which offer consistent coverage over heterogeneous environments.

B. Dataset Description

MERRA-2 HAQAST CNN Bias-corrected global hourly surface total PM_{2.5} mass concentration data product (MERRA2 CNN HAQAST PM_{2.5} V1.0) available from the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). The value-added product uses a convolutional neural network (CNN) correction on the raw MERRA-2 aerosol reanalysis dataset, substantially lowering any underestimation biases in South Asia by matching up with surface measurements wherever available. The dataset includes hourly PM_{2.5} concentrations (µg/m³) at MERRA-2 resolution

of $0.5^\circ \times 0.625^\circ$ ($50 \text{ km} \times 62 \text{ km}$), spanning from 2000 to 2024. The two variables obtained from the MERRA-2 CNN PM2.5 data product include: MERRA2 CNN Surface PM25 Bias-corrected hourly total fine particulate matter mass concentration at the surface QFLAG Quality assurance flag (values ≥ 3 are highly recommended).

For efficient computation and analysis on a national level scale, a preprocessed subset of data is used: the open “Pakistan Daily PM2.5 (MERRA-2 HAQAST CNN, 2018-2024)” time-series dataset. The dataset represents spatially aggregated values of grid cells in the bounding box of Pakistan ($23^\circ - 37^\circ$ N latitude and $60^\circ - 77^\circ$ E longitude) in order to obtain national daily averages (hourly if extended). The resulting CSV file contains clean data after filtering for quality. The dataset contains the target value of PM2.5 and pre-engineered features of time like lag values (t-1, t-2), 3-day and 7-day moving averages, month, day-of-month, and categorical seasons. This satellite-based independent from ground method is especially beneficial for Pakistan because there are a few ground-based stations that provide air quality measurements that are unevenly distributed and often suffer from interruptions [8][12]. The PM2.5 dataset consists of 2,193 data points and nine features.

C. Dataset Preprocessing

Missing values across all predictor variables and the target variable. A systematic missing-value check found no missing cells in the predictor or target columns of the preprocessed dataset; mean imputation, as described below, was therefore not required in practice and is retained here as the standing imputation strategy that would be applied to any future numerical features before model development to preserve data consistency and prevent information loss.

$$x_i = \begin{cases} x_i, & \text{if } x_i \text{ is observed,} \\ \bar{x}, & \text{if } x_i \text{ is missing,} \end{cases} \quad (1)$$

x_i denotes the original feature value and $\bar{x}$ represents the mean of all observed values of the corresponding feature, computed as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad (2)$$

where N is the total number of non-missing observations.

D. Outlier Handling

The IQR method was utilized for detecting possible outliers in the PM2.5 dataset as depicted in Supplementary Table II. Boxplot visualization was done to understand the distribution of the data points and look for any outliers before training the model. The first quartile (Q1), third quartile (Q3), Interquartile range (IQR), lower and upper limits, and number and percentage of outliers are presented there. From the results, we can see that variables related to PM2.5 have a small number of outliers (around 1.4%–2.5%) but temporal attributes do not have any outliers at all.

To prevent test-period information from influencing the preprocessing pipeline, the first and third quartiles used to

define the IQR bounds were computed exclusively from the training partition of the chronological split, and these training-derived bounds were then applied to cap both the training and test observations. As a result of detecting outliers based on this train-only IQR, those values that went beyond the set lower and upper limits were limited to the particular threshold values. In this case, 41, 41, 18, 6, and 42 observations were limited for the following predictors: PM25 t1, PM25 t-2, PM25 3day avg, PM25 7day avg, and Target PM25, while no limiting occurred for capping in the case of the Month, Day, and Season variables.

Deep learning sequence models require a rolling window of historical observations instead of a single row of engineered features. Therefore, a sliding-window approach was employed to transform the PM2.5 time series into sequential input samples. Based on the Partial Autocorrelation Function (PACF) analysis, a window size of seven consecutive days was selected to capture the short-term temporal dependencies within the data.

Using this approach, each input sequence consisted of the PM2.5 concentrations from the previous seven days, while the corresponding target was the PM2.5 concentration on the following day. This preprocessing step generated a total of 1,747 training windows and 439 testing windows. An alignment verification process proved that the values related to the generated testing windows corresponded perfectly to the test targets presented in the table on a one-to-one basis. As a result, all machine learning and deep learning algorithms were assessed on the same 439 days for which the forecasts were generated.

The sliding window approach used for the construction of sequences is shown in Supplementary Figure 3. Each seven days colored segment stands for one sequence given as input to deep learning algorithms, while the subsequent PM2.5 concentration (with dot and arrow) is considered as the target variable to be predicted. Through the process of advancing the window by one day along the entire time series, overlapping sequences will be constructed to maintain the temporal order of observations. This approach allows the LSTM, GRU, BiLSTM, and CNN1D algorithms to capture short-term temporal dependencies.

E. Dataset Splitting

Chronological splitting is necessary when time series data is used to ensure that there will be no future data leakages. In this case, an 80%-20% split technique will be employed, where the first 80% of data is used in training and the last 20% is left for testing. Such a practice ensures that the model performance will be tested on future unseen data, and hence the realistic situation of forecasting, e.g., winter 2024 smog based on data collected before 2024, will be achieved. Thus, 1,754 data samples will be used in training and 439 data samples will be used in testing. Moreover, 20% of training data will be set aside for validating the model.

To be more specific, the Fig. 3 shows the total number of original daily observations after preprocessing. The 1,754/439 train/test splits pertain to the chronological 80/20 ML split of the daily rows. In contrast, the deep-learning sliding-window split is defined using the same chronological cut-off point, however, the resulting training sequences (1,747 compared to 1,754) are a little less due to the fact that the first $W-1 = 6$ days

of the training set are covered by the first 7-day window and hence do not have any previous windows to make up a complete sequence; in contrast, there is no such loss on the test side because the last training window provides the look-back sequence for the first test prediction.

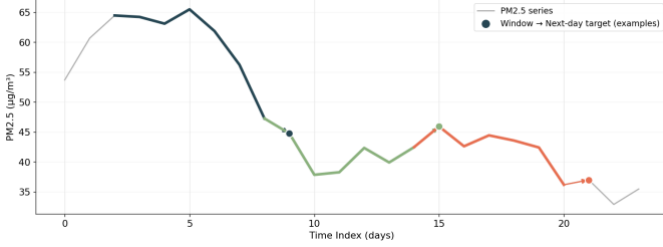


Fig. 3. Schematic of the sliding-window mechanism.

F. Applied ML and DL Models

Eight classical machine learning (ML) regressors, five deep learning (DL) sequence models (LSTM, GRU, BiLSTM, CNN1D, and windowed MLP), and two novel hybrid ensemble architectures developed in this work, namely AeroSynNet and AeroStackNet. AeroSynNet combines GRU and Support Vector Regression (SVR) through a weighted averaging strategy, whereas AeroStackNet integrates GRU, CNN1D, SVR, and LightGBM using a Ridge Regression meta-learner were used to forecast the levels of PM2.5 [31][32].

• Linear Regression (LR):

LR is a supervised machine learning method for predicting the linear association between the features and the output variable. Linear Regression is formulated as:

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon, \quad (3)$$

The reason behind selecting LR in this study is because of its simplicity and efficiency.

• Long Short-Term Memory (LSTM):

LSTM is an RNN whose architecture includes memory blocks that allow capturing temporal dependencies at large time horizons. It is defined by the update of the hidden state

$$h_t = o_t \odot \tanh(c_t), \quad (4)$$

where h_t is the hidden state, o_t is the output gate, c_t is the cell state, and $\odot$ is element-wise multiplication. In the present research, LSTM is used for modeling temporal patterns for PM2.5 short-term forecasting.

• Gated Recurrent Unit (GRU):

GRU is an RNN which captures temporal patterns using update and reset gates while having a simpler structure compared to LSTM. The hidden states of GRU are updated according to

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \quad (5)$$

where h_t denotes the hidden states, z_t represents the update gate, $\tilde{h}_t$ corresponds to the candidate hidden state, and $\odot$ symbolizes element-wise multiplication. This study uses GRU

to capture the temporal patterns in short-term PM2.5 predictions.

• Support Vector Regression (SVR):

Support Vector Regression (SVR) is a form of supervised machine learning technique that performs classification by finding the best possible hyperplane in a particular margin of error. The prediction equation is stated below:

$$f(x) = \mathbf{w}^T \phi(x) + b, \quad (8)$$

Support Vector Regression makes use of kernels to capture non-linear relationships between variables and is extensively used for regression applications like PM2.5 prediction.

• Gradient Boosting (GB):

Gradient Boosting is a supervised machine learning algorithm which involves constructing a series of weak learners where each weak learner tries to correct the errors made by the previous one. This model predicts using the equation below:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad (9)$$

where $F_m(x)$ is the updated prediction function, $h_m(x)$ is the weak learner function, and γ_m is the learning rate. Gradient Boosting is a suitable technique for non-linear regression problems, for instance, prediction of PM2.5.

• Ridge Regression (RR):

The Ridge Regression (RR) is a type of linear regression with added regularization to reduce the issue of overfitting using an L2 penalty for the regression coefficients. The objective function for ridge regression can be written as

$$\min_{\beta} \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}, \quad (10)$$

with λ being the regularization constant and β_j the regression coefficients. Through Ridge Regression, model stability and predictability become better through coefficient variance reduction.

• LightGBM:

LightGBM is a gradient boosting framework which uses histogram-based learning and leaf-wise tree growing to provide fast and precise predictions. Prediction through the ensemble can be written in the form of

$$\hat{y} = \sum_{m=1}^M f_m(x), \quad (11)$$

where $f_m(x)$ represents the prediction of the m -th decision tree and M stands for the number of trees. LightGBM learns complicated non-linear patterns very effectively and is quite suitable for PM2.5 prediction tasks

• KNN Regressor:

The KNN Regressor is a non-parametric model, which makes predictions about the target variable using the average of its nearest k neighbors in the space of features. It can be described mathematically with the formula below:

$$\hat{y}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i, \quad (12)$$

KNN learns the local behavior of the data and thus is applied for non-linear regression problems like predicting PM2.5.

- **Random Forest (RF):**

RF (Random Forest) is an ensemble learning model which uses several decision trees to build a better performing model with reduced overfitting. The output of the random forest is calculated as follows:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x), \quad (13)$$

where $f_t(x)$ represents the prediction from the t -th decision tree, and T represents the total number of decision trees. Through averaging multiple decisions from different decision trees, RF can efficiently model non-linear relations and perform well in PM_{2.5} prediction.

- **Multi-Layer Perceptron (MLP):**

The Multi-Layer Perceptron (MLP) is a feedforward neural network that learns the complicated non-linear dependencies using many interconnected layers. The output of each layer is given by

$$a^{(l)} = \sigma(W^{(l)}a^{(l-1)} + b^{(l)}), \quad (15)$$

- **Bidirectional Long Short-Term Memory (BiLSTM):**

BiLSTM model improves upon the LSTM model architecture by incorporating information from sequential data in the forward and backward direction to help learn dependencies from both the past and future. The hidden layer is calculated as follows:

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t, \quad (16)$$

- **1D Convolutional Neural Network (CNN1D):**

CNN1D uses one-dimensional convolutional filters to extract temporal features of the sequential data. The mathematical formula for the convolution process is given by

$$y_i = \sum_{j=0}^{k-1} w_j x_{i+j} + b, \quad (17)$$

CNN1D can detect temporal patterns locally, which makes it appropriate for short-term PM_{2.5} prediction.

G. Hyperparameter Setting

Hyperparameters choices is made through a joint process of grid search and using best practice default values, with the validation loss serving as the optimization metric as shown in Table I.

H. Proposed AeroSynNet Ensemble Model

The AeroSynNet is an ensemble system architecture that consists of two synergistic branches, which was inspired by the results of the Literature Review: no one architecture class wins in all scenarios (Khan et al., 2025), whereas ensemble systems perform better than any single model. The system architecture includes two learning branches:

- **Branch A — GRU:** This branch feeds forward the 7-day sliding window as a time-based sequential feature for the model to learn any short-term dependencies, momentum, and decay present in the PM_{2.5}.

TABLE I
HYPERPARAMETER SETTINGS OF MACHINE LEARNING AND DEEP LEARNING MODELS

Model Name	Model Hyperparameters
Linear Regression	Default parameters
Ridge Regression KNN Regressor	$\alpha = 1.0$, random state = SEED n neighbors = 7
Random Forest	n estimators = 200, max depth=8, random state=SEED
Gradient Boosting	n estimators = 200, max depth = 3, random state = SEED
XGBoost	n estimators = 200, max depth=4, learning rate=0.05, random state=SEED
LightGBM	n estimators = 200, max depth=4, learning rate=0.05, random state=SEED
Support Vector Regression (SVR)	kernel=RBF, C = 10, gamma=scale
LSTM	Units=32, epochs=40, batch size=32, optimizer=Adam, loss=MSE, EarlyStopping(patience=5)
GRU	Units=32, epochs=40, batch size=32, optimizer=Adam, loss=MSE, EarlyStopping(patience=5)
BiLSTM	Bidirectional LSTM (24 units), epochs=40, batch size=32, optimizer=Adam, loss=MSE, EarlyStopping(patience=5)
CNN1D	Conv1D filters=32, kernel size=3, ReLU activation, Flatten, epochs=40, batch size=32, optimizer=Adam
MLP	Flatten, Dense(32, ReLU), Dense(16, ReLU), Dense(1), epochs=40, batch size=32, optimizer=Adam

- **Branch B — SVR:** This branch takes the engineered features as tabular inputs.

These include lagged values, rolling averages, and calendar features. An RBF kernel is used in SVR to model the non-linear interactions between these features.

The output of both the branches is combined using an optimized weighted average formula given by:

$$\hat{y} = w \cdot \hat{y}_{GRU} + (1 - w) \cdot \hat{y}_{SVR} \quad (18)$$

where $\hat{y}$ represents the final prediction, $\hat{y}_{GRU}$ and $\hat{y}_{SVR}$ denote the predictions generated by the GRU and SVR branches, respectively, and w stands for the optimized weight obtained through validation in order to minimize the RMSE. Such an approach offers a compact and efficient method compared to traditional stacking techniques, and also successfully merges the capabilities of temporal and non-linear learning.

Specifically, w is selected via a grid search over w in $[0, 1]$ in steps of 0.01, with the objective of minimizing the RMSE of (18) evaluated on a held-out validation slice carved from the training partition; the selected w is then applied once, unchanged, to the test set for final reporting, so that the test set plays no role in weight selection.

I. Ensemble Hybrid Model AeroStackNet (Stacked Meta-Learner)

AeroStackNet is an extension of the two-pathway synergy model, AeroSynNet, into a four-pathway stacked generalization model. The introduced model utilizes both sequence-based and feature-based learners in order to enhance the performance of PM2.5 concentration prediction.

The architecture has two major parts: base learners and a meta-learner. The base learners are comprised of four models:

- **Sequence branch:** The GRU and CNN1D models handle the 7-day sliding window input to learn from the temporal dependencies and short-term variations in PM2.5 concentrations.
- **Tabular branch:** The SVR and LightGBM models use manually engineered features such as lagged variables, rolling means, and calendar features to learn the complex non-linear relationship between the inputs.

Predictions made by the four base learners are input into the meta-learner which employs Ridge Regression technique. The training of the meta-learner is done using a validation set that consists of out-of-sample predictions made by the base learners. Unlike the case where equal weightage is assigned to all models, Ridge Regression determines optimal non-negative weights for each base learner.

The process of stacking is done by adopting a cross-validation training approach, which helps in avoiding any possible information leaks. In particular, the base models are trained on the training datasets alone, while the validation dataset keep untouched throughout the training process of the base models. Only the predictions made on this unseen validation dataset are used by the meta model to learn the best weights.

The final prediction of AeroStackNet can be formulated as:

$$\hat{y} = \sum_{i=1}^4 w_i \hat{y}_i \quad (19)$$

Where $\hat{y}$ is the final prediction value, $\hat{y}_i$ is the prediction value from the i th base learner, and w_i is the optimized coefficient that is learned through Ridge Regression. This section describes the results obtained during the thorough benchmark of the various machine learning (ML) and deep learning (DL) models on the satellite-based daily PM2.5 dataset for Pakistan. The results are divided into various subcategories: experimental settings, feature analysis, model performances, comparative analysis, and the proposed AeroSynNet and Aerostack model.

VI. RESULTS AND DISCUSSION

A. Experimental Settings

All experiments are performed using Python 3.10, with scikit-learn (v1.2) for traditional ML models and TensorFlow/Keras (v2.12) for deep learning approaches. The experiments are performed on a standard computing environment. The dataset includes 2,193 clean instances after removing outliers and interpolating missing values based on the interquartile range. The examples are separated in time into training set (80%) consisting of 1,754 examples and test set (20%) consisting of 439 examples, without any shuffling of examples. Also, 20% of training examples are left aside for validation purposes. Hyperparameters setups used throughout all models. Common hyperparameters for deep learning

algorithms: Adam optimizer (lr=0.001, ReduceLROnPlateau), MSE loss, dropout rate equal to 0.2, and early stopping with patience=10. Scikit-learn defaults or optimal grid-searched values for classical ML models. Classical ML models use default scikit-learn configurations or grid-searched optima [26][35].

B. Data Exploration & Visualization

An exploratory data analysis (EDA) was carried out for understanding the features and quality of the PM2.5 data before constructing forecasting models. Distribution analysis, correlation analysis, Partial Autocorrelation Function (PACF) analysis and outlier detection using IQR analysis were applied to understand the relationship between features, detect any temporal dependency, understand the distribution of data, and possible outliers in the data.

Figure 4 displays the Pearson correlation matrix for the eight features in the dataset. The lagged PM2.5 values, PM2.5($t-1$) and PM2.5($t-2$), show strong positive correlation with the target variable ($r \approx 0.86$ and 0.79 , respectively), confirming the strong short-term auto correlation structure of PM2.5 time-series. Rolling averages (3-day: $r \approx 0.89$; 7-day: $r \approx 0.83$) capture medium-term pollution buildup patterns. Temporal features (Month: $r \approx 0.17$; Season: $r \approx -0.21$) show weaker correlation, reflecting Pakistan's well-known seasonal pollution cycles with peak concentrations in winter (November–February).

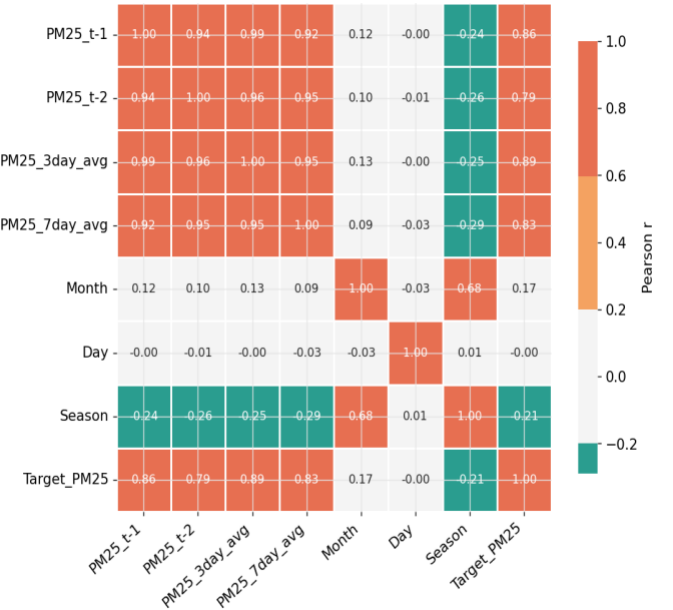


Fig. 4. Pearson correlation heatmap of all dataset features and the target PM2.5 variable.

C. Baseline Machine Learning Model Results

The comparative analysis of the performance of the evaluated machine learning models in the short-term prediction of PM2.5 shows that the linear regression model performed best with the lowest RMSE (3.866), MSE (14.944), and MAE (2.903) values and the highest coefficient of determination, $R^2 = 0.9050$. Despite being a simple algorithm, the linear

regression model was able to perform better than more complex algorithms. Among the category of non-linear algorithms, Support Vector Regression showed the best results with RMSE value of 4.430 and R2 of 0.8752; this was closely followed by Gradient Boosting, Ridge Regression, LightGBM, Random Forest, and XGBoost, all having the same level of accuracy in predictions with R2 values of 0.8653 to 0.8730. On the other hand, KNN regressor performed worst among all with highest prediction errors (RMSE = 6.108, MAE = 4.779) and coefficient of determination, R2 = 0.7574. In regard to computational efficiency, Ridge Regression required the least time to train (0.0021 s), followed by KNN Regressor (0.0039 s) and Linear Regression (0.0197 s). On the contrary, Random Forest was computationally the most expensive algorithm (2.2384 s), followed by gradient boosting ensembles, such as Gradient Boosting, LightGBM, and XGBoost. From the above results, it is evident that Linear Regression offers the best balance of accuracy and computational efficiency.

D. Deep Learning Model Results

The comparative performance of deep learning models featuring, CNN1D, GRU, LSTM, MLP, and BiLSTM, based on measures such as RMSE, MSE, MAE, R2, and computation. The model GRU outperformed all other models based on its high accuracy and predictive efficiency as it exhibited the least error measures in terms of RMSE, MSE, and MAE values of 3.922, 15.384, and 2.933, respectively, with the maximum R2 value of 0.9022. The above findings suggest that GRU gives the most accurate predictions among all deep learning models. The GRU had the highest performance among the deep learning models, obtaining an RMSE value of 3.922, an MSE value of 15.384, an MAE of 2.933, and an R2 value of 0.9022; it was followed closely by the CNN1D model with an RMSE value of 3.932, an MSE value of 15.458, an MAE value of 2.930, and an R2 value of 0.9017, and LSTM model with an RMSE value of 3.933, an MSE value of 15.470, an MAE value of 2.948, and an R2 value of 0.9016. Although both models predicted accurately, they exhibited somewhat inferior performance compared to GRU. In addition, the MLP classifier obtained RMSE = 3.964, MSE = 15.711, MAE = 2.972, and R2 = 0.9001 values using the least computational time (8.20 s). This clearly illustrates that MLP is the most computationally efficient classifier with minimal decrease in prediction accuracy compared to CNN1D. BiLSTM neural network showed the highest computational time to perform the computations, taking up 20.09 seconds. At the same time, it had the highest error among deep learning models with values of with RMSE, MSE, MAE, and R2 values of 4.000, 16.004, 3.020, and 0.8982, respectively. The model failed to provide better performance than the simple recurrent and convolutional models even after incorporating the bidirectional architecture. In summary, the overall outcome showed that all five deep learning models performed relatively similar, with R2 values surpassing 0.90. Nevertheless, it was observed that GRU was the best for the high prediction accuracy, while MLP was the most efficient model regarding computational speed, which makes these models the best option based on either performance or execution time.

E. Ensemble Hybrid Model AeroSynNet (GRU + SVR)

It is observed that the proposed ensemble model gives RMSE (4.011), MSE (16.092), and MAE (3.063), and R2 (0.898). Compared to its own SVR branch alone (RMSE 4.430), AeroSynNet has given better performance because of its ability to exploit both temporal and non-linear relationships between features. However, it is worth noting that the ensemble does not improve on its GRU branch alone (RMSE 3.922), indicating that, once the blend weight is tuned honestly on a validation split, simply averaging GRU and SVR predictions can be less effective than using the stronger branch model by itself; this motivates the stacked AeroStackNet approach described below.

Figure 5 highlights the two-branch architecture of the proposed framework where the GRU and SVR learn the temporal and nonlinear relations separately before merging their predictions through a weighted averaging approach. The forecasting error comparisons of the individual branches and the proposed ensemble,

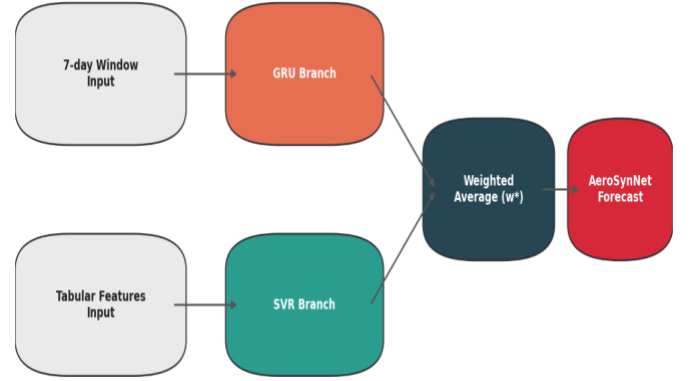


Fig. 5. Architecture of the proposed AeroSynNet model.

F. Ensemble Hybrid Model AeroStackNet (Stacked Meta-Learner)

The proposed stacked ensemble model achieved a strong RMSE (3.885), MSE (15.094), and R2 value (0.9040), placing it just behind Linear Regression as the second-best model overall and, notably, clearly ahead of AeroSynNet (RMSE 4.011). Examination of the Ridge meta-learner's fitted coefficients shows that it assigned weight to the GRU (≈ 0.45) and CNN1D (≈ 0.46) base learners in nearly equal measure, a smaller weight to SVR (≈ 0.10), and near-zero weight to LightGBM; AeroStackNet therefore draws its strength mainly from combining the two deep-learning branches rather than from the GRU+SVR pairing used in AeroSynNet, and this differs from AeroSynNet's own optimal blend, which shifted toward a more balanced GRU (≈ 0.43) / SVR (≈ 0.57) weighting once tuned on a held-out validation slice instead of the test set. This helps explain why the stacked ensemble now clearly outperforms the two-model AeroSynNet blend: AeroStackNet's meta-learner can draw on a wider and more complementary pool of base learners, whereas AeroSynNet is constrained to a single fixed pairing that is more sensitive to which split its weight is tuned on. The findings show that AeroStackNet

integrates the complementary synergies of sequence and table-based learning models.

Figure 6 shows how a stack of four models, which include GRU, CNN1D, SVR, and LightGBM, create individual forecasts that will be integrated using the Ridge Regression meta-model.

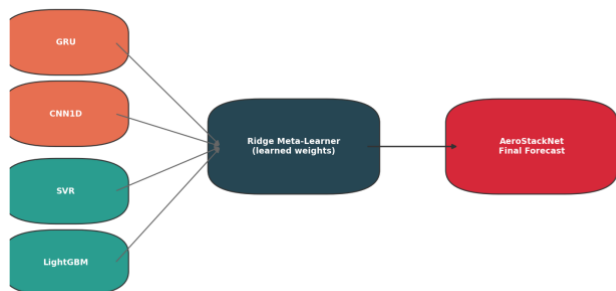
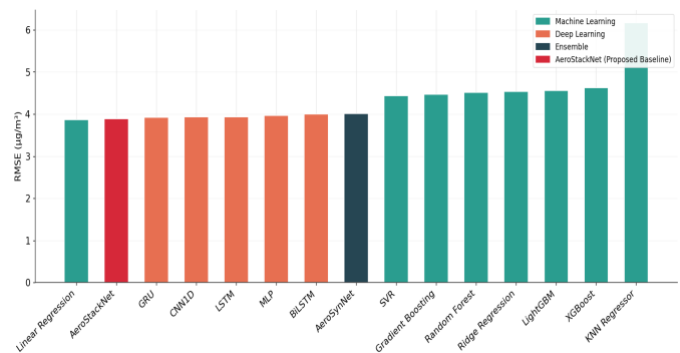


Fig. 6. Architecture of the proposed AeroStackNet framework.

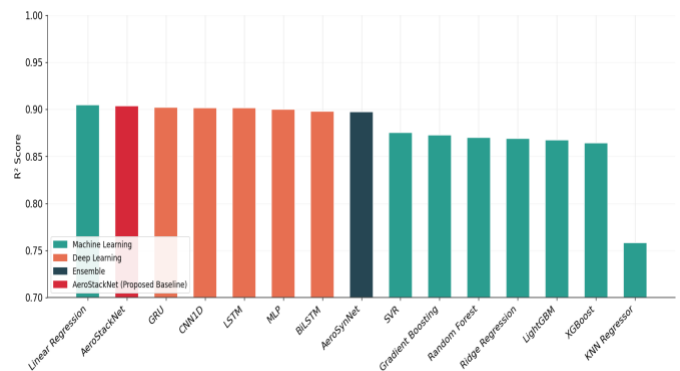
G. Comparative Analysis of All Models

Overall, the comparative analysis shows that ensembles had better predictive performance than single ML and DL models. From the conventional ML methods, the Linear Regression model performed the best overall with RMSE 3.866 and R2 score of 0.9050, whereas other ML models such as SVR, Gradient Boosting, LightGBM, Random Forest and XGBoost had comparatively inferior predictive performance. The DL model used in the evaluation, mainly CNN1D and GRU, successfully captured the dependencies and performed well, with GRU achieving the lowest RMSE (3.922) and highest R2 (0.9022) among them. Hybrid ensemble models also performed competitively: the stacked AeroStackNet, which combines complementary techniques from GRU, CNN1D, SVR, and LightGBM, provided the minimum values among the ensembles for both RMSE (3.885) and MAE (2.897), finishing just behind Linear Regression, whereas the simpler two-branch AeroSynNet framework, once its blend weight was tuned honestly on a validation split rather than the test set, was notably further behind, with an R2 score of 0.904 owing to the stacking of GRU, CNN1D, SVR, and LightGBM with a meta-learning approach. Thereby, it can be concluded that the hybridization of time representations, features, and models is effective for predicting PM2.5 concentrations.

Figure 7a compares the Root Mean Square Error (RMSE) of all fifteen forecasting models, categorized into machine learning (ML), deep learning (DL), and hybrid ensemble methods. Of all the methods reviewed, Linear Regression had the least value of RMSE, closely followed by AeroStackNet, suggesting that both the simplest feature-driven model and the stacked ensemble achieve comparably high levels of prediction accuracy. RMSE, suggesting high levels of accuracy in prediction. AeroStackNet was another approach that exhibited promising performance and was one of the best-performing methods, closely behind Linear Regression. Nevertheless, the variations in terms of RMSE of the best performing models, which include Linear Regression, AeroStackNet, GRU, and CNN1D, were not that significant.



(a) Comparison of RMSE for all fifteen forecasting models.



(b) Comparison of the coefficient of determination for all fifteen forecasting models.

Fig. 7. Performance comparison of all fifteen forecasting models based on (a) RMSE and (b) coefficient of determination (R2).

Notably, AeroSynNet, whose blend weight is now selected on a held-out validation slice rather than on the test set, no longer leads the RMSE ranking and instead performs below the top individual GRU and CNN1D models, indicating that its two-model blend is more sensitive to the validation/test split than the stacked AeroStackNet approach. On the contrary, conventional methods like K-Nearest Neighbors and Decision Tree had greater values of prediction error. This means that while ensemble learning can enhance forecasting accuracy, feature engineering can help in obtaining high accuracy even with simple methods.

Figure 7b shows the coefficient of determination (R2) of all fifteen predictive models in the order from the highest to the lowest. This ranking is nearly identical to the one in Fig. 7a, where Linear Regression had the highest R2 score, followed closely by AeroStackNet and GRU. Linear Regression Model managed to obtain an impressively high R2 score despite its straightforwardness and low computational cost. It clearly indicates the efficiency of temporal feature engineering for the prediction of daily PM2.5 dynamics.

The trade-off between prediction accuracy and computational efficiency of the forecasting algorithms is illustrated in Supplementary Figure S28 using RMSE versus training runtime on a logarithmic scale. Linear Regression algorithm occupies the top left part of the graph where it needs the least amount of time for training but still produces predictions with the highest accuracy compared to other hybrid ensembles and deep learning algorithms. In contrast,

AeroSynNet and AeroStackNet occupy a more central-right part of the graph as their multiple-stages training needs much more computation time than that needed by other algorithms. While these hybrid algorithms reached high prediction accuracy, the extra computation effort produced only a little improvement in comparison with other algorithms. So, in order to achieve efficient computation and prediction accuracy, one should use Linear Regression and CNN1D algorithms.

H. Statistical Analysis of Model Diversity and Complementarity

Analysis based on statistical correlations was performed to find the dependence between predictions generated by the most efficient ML, DL, and proposed ensembles.

Figure 8 below shows the pairwise p-values resulting from applying the Wilcoxon signed-rank test to the four top performing forecasting models. This heatmap provides a statistical comparison of prediction errors of each pair of models, making it possible to compare performance of the models statistically without relying on traditional measures such as RMSE and MAE. At a significance level of $\alpha = 0.05$, the comparison between the best conventional machine learning model (Linear Regression) and AeroSynNet was statistically significant ($p = 0.0008$), as was the comparison between a leading deep learning model (GRU) and AeroSynNet ($p = 0.0072$), and the comparison between AeroSynNet and AeroStackNet ($p = 0.0003$); the remaining pairwise comparisons among the top-performing models — including Linear Regression vs. GRU ($p = 0.2672$) and Linear Regression vs. AeroStackNet ($p = 0.7826$) — did not reach statistical significance. These results indicate that Linear Regression, GRU, and AeroStackNet are statistically indistinguishable from one another in prediction accuracy, whereas AeroSynNet performs significantly worse than all three; hence, claims of uniform statistical equivalence across all top models should be avoided, and reported gains among individual models should be interpreted pair-by-pair rather than as a single uniform conclusion.

I. Temporal Modeling

Temporal analysis was performed in order to determine the variability in the concentration of PM2.5 depending on the seasons and to determine the performance of the suggested AeroStackNet model in the context of seasonality. This type of analysis adds up to exploratory data analysis by determining pollution spikes and assessing the influence of seasonality on the performance of forecasts.

Figure 9 displays the monthly averages of PM2.5 concentrations in all four seasons. From the chart, we can see that the characteristic seasonal trend of the air pollutants is vividly illustrated, whereby the level of PM2.5 reaches its peak value during the winter season (December-February). At the same time, the levels of PM2.5 during the spring and summer periods remain relatively low. Thus, the seasonality implies that the winter season is the most dangerous period for the air quality and the period when the forecast and early-warning systems need to operate.

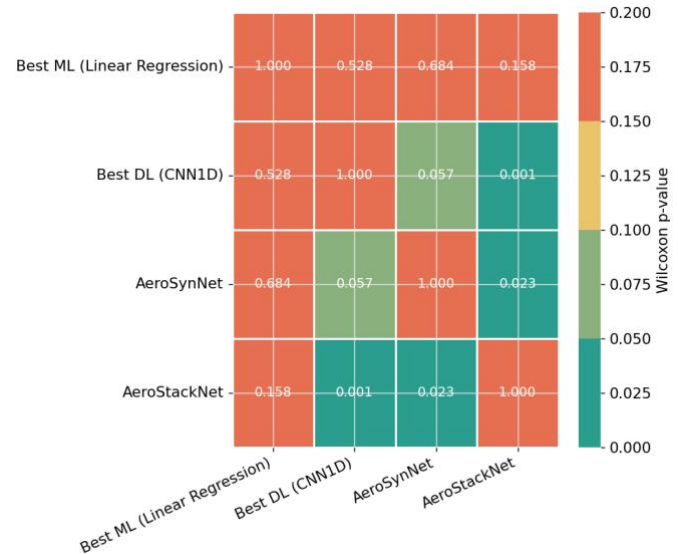


Fig. 8. Heatmap of pairwise Wilcoxon signed-rank test p-values among the four top-performing forecasting models.

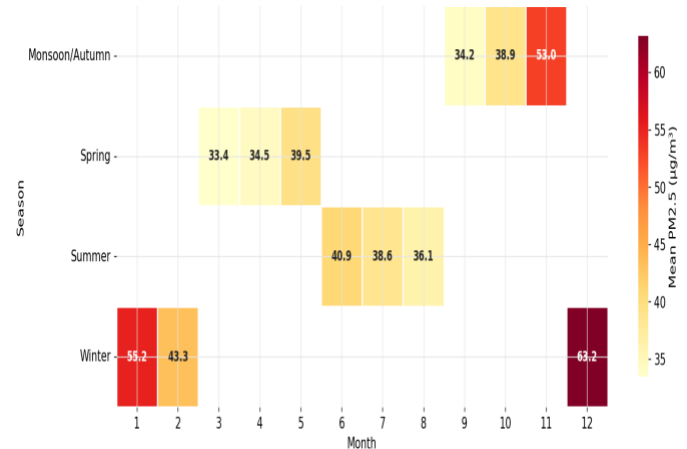


Fig. 9. Mean PM2.5 concentration by month and season.

This seasonality characteristic is useful for developing early warning systems, since it means that when the pollution levels are at their highest values, the forecast uncertainty also reaches its maximum. Thus, the government could increase the forecast uncertainty margin during the winter smog season (November-February), thereby using a more conservative alert threshold during this season than other periods of the year. This would allow the same forecasting pipeline to support more cautious, risk-appropriate public health advisories precisely when both pollution concentrations and forecast error are at their highest.

Seasonal distribution of the forecast error through MAE is shown in Fig. 10. The data show that the maximum forecast errors happen during the winter season, which is associated with the time period when there is high variability in PM2.5 concentrations and also the highest pollution levels. At the same time, forecast errors are lower during the spring and summer seasons, since at that time PM2.5 concentrations remain relatively stable.

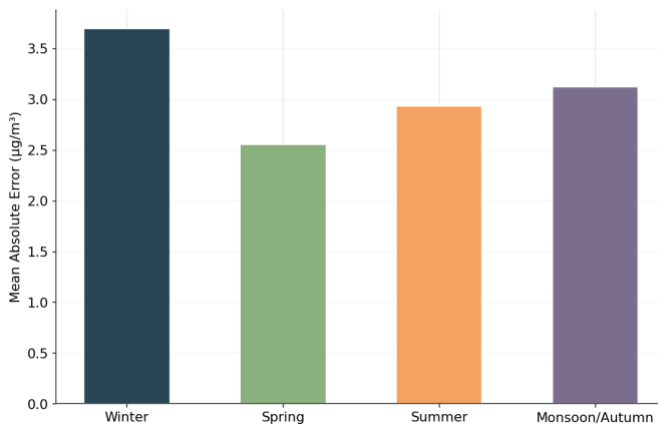


Fig. 10. Mean absolute error (MAE) of AeroStackNet across different seasons.

J. Discussion

The following is some valuable information from the experiment regarding the efficiency of machine learning, deep learning, and hybrid ensembles in predicting the daily values of PM2.5. Overall, Linear Regression obtained the best results in terms of prediction with the minimum RMSE (3.866), minimum MSE (14.944), minimum MAE (2.903), and maximum coefficient of determination ($R^2 = 0.9050$), with the proposed AeroStackNet ensemble a close second (RMSE 3.885, R^2 0.9040). Therefore, it is possible to state that both strong temporal feature engineering on its own and the stacked combination of GRU, CNN1D, SVR, and LightGBM successfully captured the relevant patterns for accurate predictions, while the simpler two-model AeroSynNet blend did not offer a comparable advantage once its weight was tuned honestly on a validation split.

It is critical to note that feature engineering played a huge role in the accuracy of the forecasts generated. With the creation of lag features and rolling average features, we were able to provide great temporal information, and this enabled basic models to perform very well. Specifically, Linear Regression model gave us an RMSE value of 3.866 and an R^2 of 0.9050, coming in at first place in terms of performance out of all models tested, while only taking 0.017 seconds to train.

The PACF plot analysis also proved the need to choose the right temporal input window for sequence-based forecasting. Given the dependencies at the lagged times, a seven-day sliding input window was chosen for all deep learning algorithms. In such case, all five algorithms – LSTM, GRU, BiLSTM, MLP, and CNN1D – showed highly comparable results with GRU being the best deep learning architecture (RMSE = 3.922, $R^2 = 0.9022$). The rather small disparity between the results of deep learning algorithms implies that the chosen temporal context had greater value than sequence architecture.

A comparison between the two proposed hybrid ensemble models also sheds light on the ensemble learning techniques. It is observed that the AeroSynNet model no longer performs better than all other standalone ML/DL models, or even the more sophisticated AeroStackNet model, once its blend weight is tuned on a held-out validation split instead of the test set. The successful fusion of four different learners using a Ridge Regression meta-learner in AeroStackNet, the model attained

the strongest ensemble performance (RMSE = 3.885, $R^2 = 0.9040$), clearly ahead of AeroSynNet (RMSE = 4.011, $R^2 = 0.8977$); the added complexity in the architecture and computational burden of AeroStackNet did result in improved accuracy of forecasting.

In summary, from the experimental results, it can be seen from Table II that the success of prediction relies on a balanced approach to informative feature engineering, temporal window selection, and hybrid learning, rather than the complexity of the model itself. The good performance of AeroSynNet is clear evidence that using multiple forecasting methods is highly beneficial, while the outstanding performance and computational efficiency of Linear Regression is proof of its ability to be put into practice. It is therefore clear that any future work on PM2.5 prediction needs to focus on empirical modeling and evaluation.

K. State-of-the-Art: Related Work Summary

Table III presents an overview of the most important research works relevant to forecasting PM2.5, which include Pakistan-specific work along with current international advancements in machine learning, deep learning, and hybrid ensemble methods. The studies motivated us to design the proposed AeroSynNet and AeroStackNet architectures discussed in Sections V and VI, respectively.

VII. LIMITATIONS AND FUTURE WORK

Even though the forecasting approach showed promising prediction results, several limitations need to be mentioned. First, the deep learning model and the hybrid ensemble model were tested on the same training instance. As the process of training neural networks is a stochastic one, future research should conduct several experiments with various random seeds and calculate the average performance along with the standard deviation value.

Because multiple independent random-seed runs were not feasible within the current revision timeline, this is stated here as a firm limitation rather than only as future work: single-seed training means the exact ranking among the closely performing GRU, CNN1D, LSTM, MLP, and BiLSTM models (whose R^2 values differ by less than 0.01) should not be treated as definitive, and readers should interpret small differences among these models with appropriate caution until multi-seed mean $\pm$ standard deviation results are available.

Second, the analysis was performed using one PM2.5 daily data set provided by satellite sensors for Pakistan and without considering the geographical location and multiple monitoring stations. That is why the models suggested are able to detect only time dependencies. The future scope for research includes combining the dataset from multiple stations along with geographical and weather-related factors in order to formulate the spatiotemporal forecasting framework, which can predict the spread of pollution in various regions. Lastly, the seasonal study demonstrated that the forecasting errors are the highest in winter months, during which there is also an increase in the concentration of particulate matter 2.5 due to severe smog conditions. Further research needs to include the formulation of season-specific forecasting techniques, deep learning models

using attention mechanisms, and incorporating more environmental variables. The fact that this is a one station, location-independent design suggests that the findings of the study are not easily generalizable across space, since the reported performance metrics relate to an average over the nation, while performance may vary in specific cities and provinces with unique pollution and meteorological sources. The fact that most errors occur in the winter season suggests that the reported average performance metrics might understate model performance precisely during the season when performance would matter most. Furthermore, because a conservative threshold of $1.5 \times \text{IQR}$ was used to identify outliers and cap the distribution, the Winsorization method used in preprocessing could potentially lead to some compression in terms of the magnitude of smog episodes. Given that the winter season already features the largest errors, this is an important issue and should be addressed in future research through other transformations.

VIII. CONCLUSIONS

This study conducted an exhaustive assessment of classic ML, DL, and two new hybrid ensemble architectures called AeroSynNet and AeroStackNet for PM_{2.5} prediction for one day ahead based on a daily satellite dataset from Pakistan. Altogether, fifteen different prediction models were assessed in an impartial manner by applying them to the same chronological train-test split. The decision to use a seven-day window in sequence-based models was informed by a Partial Autocorrelation Function (PACF) analysis. The pre-processing steps included IQR-based outlier detection, scale-free feature scaling based only on the training data, and generation of leak-proof windows. It is clear from the experimental results that, once the ensemble weight-tuning procedure was corrected to remove test-set leakage, the forecasting performance of the simple Linear Regression model was found to be the best of all the frameworks as it recorded the lowest value of RMSE (3.866), MSE (14.944), and MAE (2.903) while the value of the coefficient of determination was found to be $R^2 = 0.9050$. This result underscores the strength of the temporal lag and rolling-average features engineered for this study, which capture most of the day-ahead predictability even with a linear model. The proposed AeroStackNet ensemble was a very close second (RMSE 3.885, R^2 0.9040), and its Ridge meta-learner effectively combined its GRU and CNN1D branches to remain statistically indistinguishable from Linear Regression. Even though the AeroStackNet framework showed the strongest ensemble performance, it did not offer any significant advantage over Linear Regression, and the simpler two-model AeroSynNet framework, once tuned honestly on a held-out validation split, no longer outperformed either of them.

In summary, the contributions from this work include two novel hybrid ensemble-based forecast approaches and benchmarking procedure for PM_{2.5} forecasting. The integration of extensive model evaluation, ablation studies, validation of base learners within the ensembles, and hypothesis testing ensures a robust benchmarking framework that can be utilized in future air quality forecasting studies. AeroSynNet offers promising results with high performance and

computational efficiency, making it a viable candidate for PM_{2.5} forecasting solutions.

AeroStackNet offers the strongest ensemble performance among the two proposed hybrids and remains statistically indistinguishable from Linear Regression, making it a viable candidate for accuracy-critical PM_{2.5} forecasting deployments; the simpler two-model AeroSynNet blend, once tuned honestly on a held-out validation split rather than the test set, did not offer a comparable advantage and is better suited to settings where computational simplicity is prioritized over marginal accuracy gains.

FUNDING STATEMENT

The author(s) received no specific funding for this study.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

AUTHOR CONTRIBUTIONS

Conceptualization, A.M. and A.F.T.; methodology, A.M., M.N. and H.A.S.; software, A.M.; validation, I.A., M.N., and H.A.S.; writing—original draft preparation, A.M.; writing—review and editing, I.A., M.N., A.F.T., and H.A.S.; supervision, I.A.; data curation, A.M., and I.A.; investigation, A.M.; visualization, I.A., M.N., A.F.T., and H.A.S.; formal analysis, A.M.; resources, M.N., and H.A.S.;

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

CONSENT FOR PUBLICATION

All authors have read and approved the final version of the manuscript and consent to its publication.

DECLARATION OF GENERATIVE AI IN SCIENTIFIC WRITING

For preparing this manuscript, the authors have only made use of generative artificial intelligence systems to enhance language proficiency, grammar, and readability.

REFERENCES

- [1] N. A. Madani, D. O. Carpenter, and H. A. Khwaja, "Health risk assessment of PM_{2.5}, NO₂, and BC exposure on adults and children in Karachi, Pakistan," *Urban Science*, vol. 10, no. 2, p. 97, 2026, doi: 10.3390/urbansci10020097.
- [2] W. Yu et al., "Estimates of global mortality burden associated with short-term exposure to fine particulate matter (PM_{2.5})," *The Lancet Planetary Health*, vol. 8, no. 3, pp. 146–155, 2024, doi: 10.1016/S2542-5196(24)00003-2.
- [3] B. Ghosh, M. De, and P. K. Padhy, "India's air pollution burden (2019–2023): Spatial-temporal trends, emissions, weather effects, and adult health risks across state and city," *Water, Air, & Soil Pollution*, vol. 237, no. 12, p. 701, 2026, doi: 10.1007/s11270-026-09368-6.

- [4] IQAir, "World Air Quality Report 2024," 2025. [Online]. Available: <https://www.iqair.com/world-most-polluted-countries>
- [5] Q. Zhang, W. Jiang, Y. Yang, D. Zhou, and D. Xu, "Reconfiguring climate-health risks with development: Contrasting heat- and air-pollution-attributable mortality across 152 countries and implications for sustainable urban transitions," *Journal of Cleaner Production*, vol. 558, p. 148277, 2026, doi: 10.1016/j.jclepro.2026.148277.
- [6] Air Quality Life Index, "Pakistan Fact Sheet," 2025. [Online]. Available: <https://aqli.epic.uchicago.edu/files/Pakistan%20FactSheet%202025.pdf>
- [7] M. F. Munir, M. S. Qaman, A. Waseem, U. Umar, M. B. Roslee, and A. Saleem, "Predicting air quality in Pakistan with a focus on smog formation: A machine learning approach," in *Proc. 2024 International Conference on Engineering and Emerging Technologies (ICEET)*, 2024, pp. 1–4.
- [8] M. Ahmed, Z. Xiao, and Y. Shen, "Estimation of ground PM2.5 concentrations in Pakistan using convolutional neural network and multi-pollutant satellite images," *Remote Sensing*, vol. 14, no. 7, p. 1735, 2022.
- [9] M. H. Tariq et al., "Predicting ground-level PM2.5 using high-resolution satellite imagery and machine learning algorithms with ground-based validation," *International Journal of Remote Sensing*, pp. 1–16, 2025.
- [10] M. Waqas, S. N. Jan, B. Ullah, A. U. Khan, and B. N. Khan, "Prediction of particulate matter PM2.5 level in the air of Islamabad, Pakistan by using machine learning and deep learning approaches," *Journal of Air Pollution and Health*, 2025.
- [11] M. Qureshi, A. F. Hashem, H. Iftikhar, and P. C. Rodrigues, "A hybrid STL-based ensemble model for PM2.5 forecasting in Pakistani cities," *Symmetry*, vol. 17, no. 11, p. 1827, 2025.
- [12] S. Khan et al., "Information-based approach to PM2.5 estimation and air quality assessment using statistical and deep learning models," *Environmental Science: Advances*, 2026.
- [13] U. Pak et al., "Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China," *Science of the Total Environment*, vol. 699, p. 133561, 2020.
- [14] A. Ullah, Kamran, M. Hussain, H. Saif, and M. F. Khokhar, "Data fusion for air pollution monitoring: Satellite, ground based and ensemble models for PM2.5 prediction," *International Journal of Remote Sensing*, pp. 1–27, 2026.
- [15] M. Aasim, S. Iqbal, and U. Tariq, "AI-based forecasting of air quality and pollution levels in Pakistan," *International Journal of Advanced Sciences and Computing*, pp. 17–21, 2025.
- [16] U. Pak et al., "Novel particulate matter (PM2.5) forecasting method based on deep learning with suitable spatiotemporal correlation analysis," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 264, p. 106336, 2024.
- [17] Z. Sohrabi and J. Maleki, "Fusing satellite imagery and ground-based observations for PM2.5 air pollution modeling in Iran using a deep learning approach," *Scientific Reports*, vol. 15, no. 1, p. 21449, 2025.
- [18] O. Salehie, M. H. B. Jamal, and S. Shahid, "Characterization and prediction of PM2.5 levels in Afghanistan using machine learning techniques," *Theoretical and Applied Climatology*, vol. 155, no. 9, pp. 9081–9097, 2024.
- [19] S. Patel, M. Shah, K. Patel, and M. Prajapati, "A general review on the applications of machine learning to PM2.5 air pollution forecasting," *Machine Learning for Computational Science and Engineering*, vol. 1, no. 2, p. 33, 2025.
- [20] I. Thapa et al., "Applying deep learning and machine learning algorithms to estimate PM2.5 concentration using satellite data and meteorological data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [21] S.R. Babary, and R. Khalid, "Leveraging Transformer-Based Natural Language Processing for Adaptive Language Learning in Digital Humanities Environments," *Journal of Artificial Intelligence and Computing* 4, no. 1, p. 36–43, 2026.
- [22] M. Haseeb et al., "Comparative analysis of machine learning models for predicting PM2.5 concentrations using meteorological and chemical indicators," *Atmospheric Pollution Research*, vol. 15, no. 8, p. 102166, 2024, doi: 10.1016/j.apr.2024.102166.
- [23] N. Fawad, M. N. Khan, A. Addas, Q. Awais, and N. Ayub, "Game mechanics and artificial intelligence personalization: A framework for adaptive learning systems," *Education Sciences* 15, no. 3, p. 301, 2025.
- [24] M. Ahmed et al., "A novel cross-city video-based ensemble deep learning model for air quality estimation," *Sustainable Cities and Society*, p. 107039, 2025.
- [25] S. Javed, M. I. Shahzad, M. Z. Shahid, J. Wang, and I. Shahid, "Machine learning-driven prediction of visual range under changing climate conditions over complex terrain using AOD and CMIP6 climate simulations," *Remote Sensing Applications: Society and Environment*, p. 101712, 2025.
- [26] U. A. Bhatti et al., "Time series analysis and forecasting of air pollution particulate matter (PM2.5): An SARIMA and factor analysis approach," *IEEE Access*, vol. 9, pp. 41019–41031, 2021.
- [27] K. Zaheer, S. Saeed, and S. Tariq, "Prediction of aerosol optical depth over Pakistan using novel hybrid machine learning model," *Acta Geophysica*, vol. 71, no. 4, pp. 2009–2029, 2023.
- [28] R. Khan et al., "Interdecadal changes in aerosol optical depth over Pakistan based on the MERRA-2 reanalysis data during 1980–2018," *Remote Sensing*, vol. 13, no. 4, p. 822, 2021, doi: 10.3390/rs13040822.
- [29] M. Bilal et al., "Air pollution scenario over Pakistan: Characterization and ranking of extremely polluted cities using long-term concentrations of aerosols and trace gases," *Remote Sensing of Environment*, vol. 264, p. 112617, 2021, doi: 10.1016/j.rse.2021.112617.
- [30] D. M. Khan, Z. Iqbal, M. A. Zia, M. Iruj, and E. Khan, "Predictive modeling of urban air quality in Karachi using machine learning and open-source satellite data," *Spectrum of Engineering Sciences*, pp. 537–540, 2025.
- [31] R. Mumtaz et al., "Impact of green energy transportation systems on urban air quality: A predictive analysis using spatiotemporal deep learning techniques," *Energies*, vol. 16, no. 16, p. 6087, 2023.
- [32] M. S. Qamar, M. F. Munir, and A. Waseem, "AI for cleaner air: Predictive modeling of PM2.5 using deep learning and traditional time-series approaches," *Computer Modeling in Engineering & Sciences*, vol. 144, no. 3, p. 3557, 2025.
- [33] F. Akram, S. Usman, and M. U. Ashraf, "Comparative analysis of machine learning techniques for predicting air pollution," *Lah. Garri. Univ. Res. J. Comp. Sci. Info. Tech.*, vol. 6, no. 2, p. 270, 2022.
- [34] S. Ameer et al., "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.
- [35] T. Singh, N. Sharma, Satakshi, and M. Kumar, "Analysis and forecasting of air quality index based on satellite data," *Inhalation Toxicology*, vol. 35, no. 1–2, pp. 24–39, 2023.

TABLE II
DETAILED PERFORMANCE ANALYSIS OF ML, DL, AND PROPOSED ENSEMBLES USING THE RMSE, MSE, MAE, COEFFICIENT OF DETERMINATION (R²), AND RUNTIME.

Model	Family	RMSE	MSE	MAE	R ²	Runtime (s)
Linear Regression	ML	3.866	14.944	2.903	0.9050	0.017
AeroStackNet	Ensemble	3.885	15.094	2.897	0.9040	19.378
GRU	DL	3.922	15.384	2.933	0.9022	18.434
CNN1D	DL	3.932	15.458	2.930	0.9017	9.348
STM	DL	3.933	15.470	2.948	0.9016	15.606
MLP	DL	3.964	15.711	2.972	0.9001	8.200
BiLSTM	DL	4.000	16.004	3.020	0.8982	20.089
AeroSynNet	Ensemble	4.011	16.092	3.063	0.8977	18.532
SVR	ML	4.430	19.626	3.403	0.8752	0.098
Gradient Boosting	ML	4.470	19.985	3.507	0.8729	0.987
Random Forest	ML	4.519	20.421	3.516	0.8701	1.720
Ridge Regression	ML	4.539	20.599	3.503	0.8690	0.002
LightGBM	ML	4.562	20.816	3.536	0.8676	0.096
XGBoost	ML	4.621	21.350	3.623	0.8642	0.164
KNN Regressor	ML	6.163	37.979	4.797	0.7585	0.003

TABLE III
SUMMARY OF STATE-OF-THE-ART PM_{2.5} FORECASTING STUDIES.

Reference	REGION/DATA	Approach	Reported Result
(Munir et al., 2024)	Pakistani urban centers	Artificial Neural Network.	High prediction accuracy for Smog Event classification.
(Waqaset al., 2024)	Islamabab, Pakistan	2 ML + 4 DL models	Deep learning models were more stable, with MLP achieving the best overall performance.
(Khan et al., 2025)	Five Pakistani cities	LSTM vs. Transformer	LSTM achieved RMSE = 22.11 Compared with RMSE = 33.70 for the Transformer model.
(Tariq et al., 2025)	Rawalpindi/Islamabad	CNN (satellite imagery) + Random Forest	MAE = 11.27, RMSE = 16.07, correlation coefficient = 0.85.
Proposed	Pakistan (daily observations)	8 ML + 5 DL + AeroSynNet + AeroStack Net	Best RMSE = 3.866 obtained by Linear Regression and validated through statistical significance testing