

Forecast-Aware Explainable Hybrid Stacking for Forecasting Across Heterogeneous Indian Cities

Naveen Kumar K

kante.naveen@pondiuni.ac.in

Pondicherry University

Samundiswary P

Pondicherry University

Article

Keywords: Forecasting, Explainable hybrid stacking, Forecast-aware feature engineering, Consensus-based feature selection, Heterogeneous ensemble learning, SHAP interpretability, CatBoost meta-learning, Multi-city air-quality prediction, Time-series machine learning

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10991524/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Forecast-Aware Explainable Hybrid Stacking for Forecasting Across Heterogeneous Indian Cities

K. Naveen Kumar^a, P. Samundiswary^{b,*}

^aDepartment of Electronics Engineering, Pondicherry University, Puducherry, 605014, Puducherry, India

^bDepartment of Electronics Engineering, Pondicherry University, Puducherry, 605014, Puducherry, India

ARTICLE INFO

Keywords:

Forecasting, Explainable hybrid stacking, Forecast-aware feature engineering, Consensus-based feature selection, Heterogeneous ensemble learning, SHAP interpretability, CatBoost meta-learning, Multi-city air-quality prediction, Time-series machine learning.

ABSTRACT

Fine particulate matter (PM_{2.5}) prediction is essential for proactively managing air quality since levels of pollutants differ over time and across urban areas. An Explainable Hybrid Stacking Framework (EHSF) is proposed to forecast one hour ahead PM_{2.5} for three CPCB monitoring stations in India namely, Anand Vihar, Delhi; Chhatrapati Shivaji International Airport T2, Mumbai; and Manali, Chennai using hourly observations during 2021–2025. Temporal, cyclic, lagged, rolling-statistical, pollutant, and domain-specific interaction features are extracted from the data. Redundancy elimination and consensus-based feature selection is carried out based on mutual information, Random Forest importance, permutation importance, and SHAP importance. Random Forest, Extra Trees, and XGBoost algorithms predict out-of-fold temporally ordered values of one-hour-ahead PM_{2.5}, which are aggregated using CatBoost meta-algorithm. EHSF gives $R^2 = 0.9383, 0.8879$, and 0.7696 and $RMSE = 30.0167, 11.3705$, and 16.7043 on the held-out chronological test data sets for Delhi, Mumbai, and Chennai, respectively. From the results, it can be inferred that EHSF can enhance the performance of one hour-ahead PM_{2.5} forecasting, however, further external validation at other monitoring stations and time intervals is required to claim generalizability across different geographical locations.

1. Introduction

Fine particulate matter (PM_{2.5}) forecasting has become increasingly important for the proactive management of air quality because urban pollution processes are driven by the interaction of temporal, meteorological, particulate, and emission-related factors. Recent research into explainable ensemble architectures reveals that hybridizing heterogeneous models through feature attribution techniques may provide improved forecasts alongside enhanced interpretation capabilities [1]. Recent research into graph-based learning architectures informed by pollutant dispersion processes reveals the importance of incorporating spatial relationships in models used to forecast PM_{2.5} levels at geographically scattered observation locations [2]. Decomposition methods reveal that using different temporal frequency components allows one to better represent complex processes of PM_{2.5} formation that are challenging to model with the use of single temporal feature sets [3]. The incorporation of meteorological dynamics into short-term forecasting demonstrates the role of environmental context in PM_{2.5} prediction in case of fast-changing atmospheric conditions [4].

1.1. Problem Overview and Motivation

The increasing complexity of atmospheric dynamics associated with pollutant concentrations requires shifting from individually functioning predictive algorithms toward ensembles capable of leveraging complementary learning behaviours in different environments. A regime-aware hybrid

ensemble learning approach shows how the predictive power of particular learners is dependent upon specific pollution regimes, thus creating motivation for adaptive learner combinations instead of reliance on a single predictive model [5]. This is especially relevant for PM_{2.5} prediction because pollutant concentrations are formed by the interactions of multiple temporal, meteorological and emission-related processes that affect PM_{2.5} generation to different extents depending upon location and current time period.

Thus, an explainable ensemble architecture that incorporates gradient boosting, XGBoost, stacking, and SHAP techniques demonstrates the possibilities of complementing nonlinear features of learners and obtaining relevant information about the contribution of particular predictors [6]. The operational importance of forecasting is illustrated through the application of artificial-intelligence-based air-quality mapping and prediction of extreme pollution conditions that allows one to perform proactive environmental management based on high PM_{2.5} prediction [7]. Nonetheless, the effectiveness of such a task depends not only upon the predictive algorithm but also upon the way raw atmospheric data is preprocessed into informative predictors [8]. Thus, the increasing importance of feature representation makes PM_{2.5} forecasting a task of simultaneous model construction and feature optimization.

Explainable approaches to machine learning show that feature attribution techniques are able to identify the contribution of meteorological and anthropogenic processes and thus provide relevant information beyond conventional prediction error metrics [9]. Similar results were found by research in explainable PM_{2.5} modelling that shows that the relative importance of source-related and meteorological predictors is dependent on specific conditions of prediction

*Corresponding author. kante.naveen@gmail.com

✉ kante.naveen@gmail.com (K.N. Kumar); sam.dee@pondiuni.ac.in (P. Samundiswary)

ORCID(s): 0009-0002-7345-2625 (K.N. Kumar); 0009-0004-3871-1892 (P. Samundiswary)

and hence creates the necessity for analysing the inner workings of the learner instead of accuracy only [10]. These results make clear a necessary condition for contemporary $PM_{2.5}$ prediction frameworks: in addition to providing accurate predictions, any predictive framework should clarify what information was used to generate them.

Table ?? summarizes key representative studies in statistical forecasting, conventional machine learning, deep learning, hybrid and ensemble learning, explainable AI, feature optimization, multi-city and Indian urban air-quality prediction. The literature demonstrates a progressive shift toward nonlinear learning, heterogeneous ensembles, temporal and spatial representation, feature optimization, and model explainability. However, these components are generally investigated separately or in partially integrated combinations. This observation motivates the proposed EHSF framework, which jointly incorporates forecast-aware feature engineering, consensus-based feature selection, leakage-controlled out-of-fold stacking, heterogeneous RF-ET-XGBoost base learners, CatBoost meta-learning, and SHAP-based interpretation across three heterogeneous Indian cities.

1.2. Research Gap

In accordance with the described developments, the success of $PM_{2.5}$ forecasts is significantly dependent upon the quality of the temporal and environmental information representation utilized in a predictive model. Explainable $PM_{2.5}$ forecasting using engineered predictors and SHAP revealed that adequately designed features contribute to improved predictive representations while maintaining explainability at the predictor level [11]. Mutual-information-based feature filtering with subsequent spatiotemporal forecasting demonstrated the significance of detecting informative predictors prior to using complex predictive architecture [12]. This reveals that feature generation and feature selection have to be treated as independent steps of a modelling pipeline, rather than as a part of preprocessing.

At the same time, a systematic construction of features increases the likelihood of redundancy in lagged, rolling, cyclic, pollutant, and interaction variables. Using just one metric as a tool of feature selection may be insufficient since metrics measure feature importance from different model- or statistical-based perspectives. A stacking approach for air-quality prediction showed that using heterogeneous learners allows capturing complementary non-linear relations between environmental predictors and $PM_{2.5}$ concentration [13]. Deep ensembles of forecasters also demonstrated better results of $PM_{2.5}$ estimates compared to a single model-based representation of information [14]. Studies of ensemble forecasting showed that a combination of different learning techniques increases the robustness of air-quality forecasts under different atmospheric circumstances [15]. It shows that diversity of feature representation and diversity of predictive models are complementary sources of information. However, no approach currently uses them effectively within one forecast-focused methodology.

The problem becomes even more critical when forecasting is performed across heterogeneous metropolitan environments. A spatiotemporal hybrid approach revealed the significance of taking into account the relations between space and time in forecasting $PM_{2.5}$ in metropolitan environments [16]. Heterogeneous spatiotemporal stacking approaches demonstrated that using heterogeneous predictive learners can bring additional benefits in $PM_{2.5}$ prediction in geographically distributed areas [17]. Ensemble learning for cities also showed that heterogeneous learners allow improving the results of urban air-quality forecasts [18]. Although these developments showed how important it is to combine multiple learning strategies in $PM_{2.5}$ prediction, the problem of coordination of forecast-time feature construction, multi-criteria feature selection, heterogeneous learning, chronological validation, and explainability remains unsolved.

An additional gap relates to the correlation between prediction accuracy and explainability. Explainable hybrid stacking demonstrated that it is possible to use an ensemble approach for particulate pollution forecasting together with SHAP [19]. Automated machine-learning demonstrated the possibility of systematic optimization of learning strategy and hyperparameters in case of predicting air quality in urban environments [20]. However, there is no guarantee that it leads to the development of forecast-compatibility, minimized feature redundancy, and explainability of predictor space. The practical importance of the development of $PM_{2.5}$ prediction was confirmed by the study that correlated pollution prediction with the assessment of health risks during severe pollution episodes [21]. A multi-task ensemble approach also showed the possibility of exploiting relationships between pollutants via ensemble and probabilistic learning [22].

Therefore, the literature review shows that there is a methodological fragmentation problem in the field of $PM_{2.5}$ prediction. Each of the listed elements of $PM_{2.5}$ prediction was developed successfully; however, it was not done in a way to create an integrated forecasting pipeline. The current research problem is thus not only how to obtain a better predictive score, but how to build a $PM_{2.5}$ forecasting system that would allow controlling information availability, predictor redundancy, learner diversity, chronological validation, and explainability.

1.3. Research Objective and Contributions

In this regard, the current study proposes the construction of an Explainable Hybrid Stacking Framework (EHSF) for multi-city $PM_{2.5}$ forecasting using forecast-aware feature construction, consensus-based feature selection, heterogeneous ensemble learning, controlled out-of-fold stacking and SHAP-based interpretation. The proposed EHSF is designed based on hourly CPCB observations from Anand Vihar (Delhi), Chhatrapati Shivaji International Airport T2 (Mumbai), and Manali (Chennai). During the implementation of the methodology, the obtained dataset is represented using temporal, cyclic, lag-based, rolling-statistical, pollutant, and domain-specific interaction features. Then,

multi-criteria feature optimization is performed using Mutual Information, Random Forest importance, Permutation Importance, and SHAP-based importance in order to detect complementary features and decrease the likelihood of redundancy. Optimized predictors are used for training Random Forest, Extra Trees, and XGBoost learners with the generation of out-of-fold predictions for the chronological validation. Finally, the ensemble predictions are combined via a CatBoost meta-learner. At the last step, the EHSF trained with this methodology is analyzed using SHAP.

The hourly CPCB datasets for the three stations span the period from January 1, 2021 to December 31, 2025. After removing duplicates and hours with missing $PM_{2.5}$ target values, the final datasets comprised 39,420 valid hourly observations for Delhi, 38,905 for Mumbai, and 37,812 for Chennai. For each city, the earliest 80 percent of the chronological observations were allocated to the training period and the most recent 20 percent to the held-out test period. This resulted in 31,536 training and 7,884 test samples for Delhi, 31,124 training and 7,781 test samples for Mumbai, and 30,250 training and 7,562 test samples for Chennai. The training periods cover January 2021 to September 2024 for Delhi, January 2021 to September 2024 for Mumbai, and January 2021 to August 2024 for Chennai, while the test periods cover October 2024 to December 2025 for Delhi and Mumbai, and September 2024 to December 2025 for Chennai.

Consequently, four major contributions of this study are:

1. **Forecast-aware feature engineering:** developing a structured representation of the predictors including temporal, cyclic, lag-based, rolling-statistical, pollutant, and domain-specific interaction features according to forecasting-time information.
2. **Consensus-based feature selection:** employing Mutual Information, Random Forest importance, Permutation Importance, and SHAP-based importance in order to detect complementary features and to reduce predictor redundancy.
3. **Heterogeneous hybrid stacking:** creating an RF–Extra Trees–XGBoost base ensemble with leakage-controlled out-of-fold predictions generation and CatBoost meta-learning for exploiting complementary representations of information.
4. **SHAP-based explainability:** applying SHAP on the final EHSF for quantifying contributions of global and local predictors and understanding $PM_{2.5}$ forecasting mechanisms of three cities.

Then, the framework is experimentally verified using a chronological training-validation scheme and time-series cross-validation using R^2 , RMSE, and MAE to evaluate performance. The designed experimental setting makes it possible to evaluate not only the ability of EHSF to improve the accuracy of $PM_{2.5}$ forecasts, but also to understand why predictive behavior differs across different cities and what are the influential predictors in a heterogeneous ensemble

setting. This discussion thus explains the research motivation and gap concerning information construction, feature optimization, heterogeneous learning, and explainability in $PM_{2.5}$ forecasting. Therefore, the following section will consider the evolution of related research in machine learning, ensemble and hybrid learning, spatiotemporal forecasting, feature optimization, deep learning and explainable artificial intelligence.

Table 1 summarizes representative studies covering conventional machine learning, deep learning, hybrid and ensemble learning, explainable machine learning, feature optimization and decomposition, multi-city and spatial forecasting, and Indian urban air-quality prediction. The literature demonstrates a progressive shift toward nonlinear learning, heterogeneous ensembles, temporal and spatial representation, feature optimization, and model explainability. However, these components are generally investigated separately or in partially integrated combinations. This observation motivates the proposed EHSF framework, which jointly incorporates forecast-aware feature engineering, consensus-based feature selection, leakage-controlled OOF stacking, heterogeneous RF–ET–XGBoost base learners, CatBoost meta-learning, and SHAP-based interpretation across three heterogeneous Indian cities.

2. LITERATURE REVIEW AND RESEARCH POSITIONING

The current forecasting literature has evolved beyond machine-learning algorithms by proposing more advanced architectures such as ensemble, deep-learning, spatiotemporal, and explainable architectures to cope with increasingly heterogeneous dynamics in urban pollution [23]. The mentioned evolution indicates that the performance of predictive models is related not only to the learning algorithm itself but also to information representation regarding temporality, meteorology, pollution, and geography [24]. Multi-city studies also point to the variability of dynamics across various monitoring sites, suggesting the importance of geographical heterogeneity for assessing the performance and consistency of the proposed forecasting framework across heterogeneous metropolitan environments. [25]. The variability in target distribution, in turn, highlights the issue that learned associations between pollutants are valid for the particular pollution regime and may shift in other periods [26]. Thus, contemporary forecasting needs to consider such aspects as predictive learning, information representation, heterogeneity, temporal dependencies, and interpretability in a coordinated manner [27].

2.1. Machine Learning in the Field of Forecasting

Machine-learning models have shown their capacity to detect nonlinear relations between pollution and environmental predictors in a heterogeneous urban dataset [23]. Long-horizon forecasting has additionally pointed to the importance of regional meteorology as an additional source of atmospheric representation beyond historical pollution

Table 1Comparative Analysis of Representative PM_{2.5} and AQI Forecasting Studies and Identified Research Gaps

Category	Representative studies	Core methodology	Key contribution	Gap relative to EHSF
Conventional machine learning	[15, 21, 23, 35, 36, 40, 44, 52, 57, 59]	RF, optimized ML, and kernel-based learning	Models nonlinear relationships between pollutant and environmental variables	Limited integration of forecast-aware features, consensus selection, OOF stacking, and SHAP interpretation
Deep learning	[2, 4, 16, 24, 25, 29, 31, 33, 49, 51, 54, 56, 58]	BiLSTM, CNN, TCN, GRU, attention, graph-based, and other deep-learning models	Learns complex temporal, spatial, and spatiotemporal dependencies	Limited integration of multi-criterion feature selection with heterogeneous OOF stacking
Hybrid and ensemble learning	[1, 5, 6, 13, 17, 18, 26, 28, 30, 32, 34, 37, 38, 39, 43, 45, 48, 50, 53, 55]	Stacking, boosting, hybrid DL, decomposition-ensemble, and multi-scale ensemble methods	Combines complementary learners to improve forecasting performance and robustness	Limited simultaneous integration of forecast-aware features, consensus selection, OOF stacking, and meta-level SHAP
Explainable machine learning	[9, 10, 11, 13, 19, 24, 27]	SHAP, interpretable ML, explainable ensembles, and interpretable DL	Provides model-level feature attribution and improves prediction interpretation	Explainability is generally not integrated with consensus selection and the complete heterogeneous stacking pipeline
Feature optimization and decomposition	[3, 8, 12, 37, 38, 46, 47]	Feature engineering, MI filtering, decomposition, denoising, feature extraction, and multi-scale representation	Improves feature representation and reduces redundant or noisy information	Limited integration of multiple importance criteria with leakage-controlled stacking and SHAP interpretation
Multi-city and spatial forecasting	[2, 30, 48, 50, 53]	Graph learning, spatial modelling, intercity learning, and spatiotemporal ensemble methods	Captures spatial heterogeneity and cross-location dependencies	Limited integration of city-specific feature selection with unified explainable stacking
Indian urban environments	[18, 23, 30, 34, 35, 36, 40, 51, 52, 54, 57]	ML, DL, ensemble, and AQI/PM _{2.5} forecasting	Demonstrates data-driven air-quality forecasting in Indian urban environments	Predominantly city-specific studies with limited integration of forecast-aware selection, OOF stacking, and SHAP
Recent explainable and stacking approaches	[1, 5, 6, 9, 10, 11, 13, 17, 18, 19, 20, 22]	Explainable ensembles, stacking, SHAP, feature engineering, and AutoML	Combines predictive modelling with improved model interpretation	The methodological components remain only partially integrated into a single leakage-controlled framework
Proposed EHSF	This study	Forecast-aware features, MI-RF-PI-SHAP consensus selection, RF-ET-XGBoost OOF stacking, CatBoost meta-learning, and SHAP	Integrates feature optimization, heterogeneous stacking, leakage control, and two-level interpretation across three Indian cities	External validation across additional cities, monitoring stations, periods, and data sources is required

records [24]. Multi-location forecasting has, thus, brought the need to consider both temporal and geographical dependencies to adequately represent variable behavior of [25]. Ensemble multivariate forecasts, in turn, have emphasized the importance of evolving target distributions, suggesting the importance of pollution dynamics rather than the stationary relation of predictors and targets for the model robustness [26]. At that, growing model complexity brings the need for interpretability of machine-learning models[27].

2.2. Ensemble and Hybrid Learners

Limitations of separate learners have been used to construct ensemble approaches by leveraging complementarity of different models in their predictions under nonlinear atmospheric conditions [28]. Deep-learning hybrid systems have been additionally used to represent complex temporal dependencies in behavior through attention and optimization procedures; however, growing model complexity may require considerable computational effort and reduce interpretability [29]. In turn, stacking has been offered as an alternative for building complex models on the basis of heterogeneous learners without need for a single complex architecture, comparative study showing the complementarity of model predictions [30]. In addition, forecasting based on the information from meteorology, vehicular activity, and

emissions shows that rich environmental representation can help improve predictive power in heterogeneous metropolitan environment[31].

2.3. Spatiotemporal Forecasting and Multi-City Evaluation

Multicity forecasting makes the task even more complicated since the dynamics heavily depend on local environment [32]. Spatiotemporally high-resolution forecasting demonstrates that increased information content can help to improve predictability but, at the same time, increasing the number of dimensions can also raise the complexity of modeling [33]. The issue of spatiotemporal modeling, thus, is connected with the necessity to explicitly account for temporal and geographical dependencies rather than rely solely on predictors' relations [34]. Studies from Visakhapatnam demonstrate the impact of environmental factors on urban air-quality prediction, while studies conducted in Hyderabad indicate the effectiveness of using several predictors for city-specific forecasting [35],[36]. Thus, multicity analysis is not just an expansion of the set of observations but rather a test of the possibility to apply the common forecasting architecture for heterogeneous pollution regimes.

2.4. Feature Engineering and Feature Selection

The information space composed of numerous temporal and meteorological variables can be redundant [37]. In turn, denoising and ensemble learning show the importance of getting rid of redundancy prior to model estimation [37]. Decomposition and ensemble approaches show the efficiency of structural transformation of complex dynamics for better representation of different dynamics [38]. Particulate forecasting informed by meteorology confirms once again the value of environmental information in conjunction with ensemble learning [39]. The studies conducted in Indian urban environment confirm that nonlinear relationships and local characteristics are the key for predictive behavior of models [40]. This suggests that forecasting-aware construction of temporal, cyclic, lagged, rolling statistical, and specific domain interaction features should be preceded by systematic feature selection.

2.5. Deep Learning, Ensembles' Evolution, and Explainability

The deep-learning approach has brought increasingly complex architectures for representation of nonlinear atmospheric process and temporal dependencies [41]. Regional ensemble forecasting demonstrates the possibility to improve forecasting over long horizons by leveraging predictive information from various sources [42]. Boosting combined with deep-learning algorithms has additionally demonstrated the complementarity of learning approaches to particulate forecasting [43]. Distributed in geography machine-learning studies reconfirm the importance of heterogeneous environmental conditions, while deep ensemble models show that the multiple models can serve as a flexible alternative to highly complex deep learners [44],[45]. Multi-objective feature extraction and multi-scale approaches have additionally demonstrated the possibility to improve hourly forecasts through optimized representation of predictors and multiple temporal scales [46],[47]. In turn, intercity ensembling, 3D CNN-GRU models and data-driven heterogeneous framework emphasize the value of accounting for spatiotemporal and environmental information [48, 49, 50]. Deep learning-based Hyderabad studies additionally confirm the predictive value of historical temporal information [51].

2.6. Critical Positioning of the Proposed EHSF Approach

Contemporary machine learning has confirmed the possibility of nonlinear particulate forecasting but highlighted that the performance of models is closely connected with the quality of environmental representation [52]. Graph-based spatiotemporal architectures have additionally expanded the range of representation possibilities, while metropolitan AQI model demonstrated the value of sequential learning for temporal behavior of pollutants [53, 54]. Methods of multi-objective optimization and feature extraction are additional

ways of predictive search space control, while spatiotemporal deep learning approach reconfirms the value of spatiotemporal representation [55, 56]. Prediction of Delhi additionally reconfirmed the feasibility of data-driven modeling of particulate matter in Indian metropolitan environment, while attention-based LSTM and ensemble models confirmed the value of temporal attention and model combination [57, 58]. Methods of multiple-kernel learning have additionally shown the possibility to model nonlinear relationships between meteorological conditions and particulates [59]. Despite all these advances, the literature still remains methodologically fragmented, since each particular study highlights feature engineering, advanced temporal learning, ensembling, spatial modeling, optimization, and explainability separately. The current approach aims to bridge this gap by integrating forecasting-aware feature engineering, consensus-based feature selection on the base of Mutual Information, Random Forest importance, Permutation Importance, and SHAP importance, leakage controlled out-of-fold stacking of Random Forest, Extra Trees, and XGBoost, CatBoost meta-learning, and city-specific SHAP interpretation. The next section is devoted to the experimental design from data acquisition and chronological preprocessing to feature construction, selection, ensembling, explainability, and time-series analysis.

3. MATERIALS AND METHODS

The proposed Explainable Hybrid Stacking Framework (EHSF) was implemented as a sequence of steps that transform raw hourly air-quality observations into forecast-relevant predictors, remove redundancy, integrate complementary machine learning learners, and interpret the results. The methodology involves six interrelated steps: data pre-processing, forecast-aware feature engineering, consensus-based feature optimization, heterogeneous stacking, SHAP explainability, and time-aware performance evaluation. The chronological order of observations was preserved throughout all modelling procedures to prevent future information from entering feature optimization, model training, validation, or prediction. The entire workflow is visualized in Fig. 1.

The entire workflow of EHSF is illustrated in Fig. 1, which starts with the collection and preprocessing of data from CPCB, followed by forecast-aware feature engineering and feature selection. The optimized feature set is processed through Random Forest, Extra Trees, and XGBoost base models via out-of-fold predictions before passing through the meta-learning phase of CatBoost and prediction.

3.1. Study Area and Data Set

The EHSF was implemented using hourly observations available from the Central Pollution Control Board (CPCB), India for the period from January 2021 to December 2025. To test the proposed framework in heterogeneous environments, the following three representative metropolitan monitoring sites were selected: Anand Vihar, Delhi; Chhatrapati Shivaji International Airport (T2),

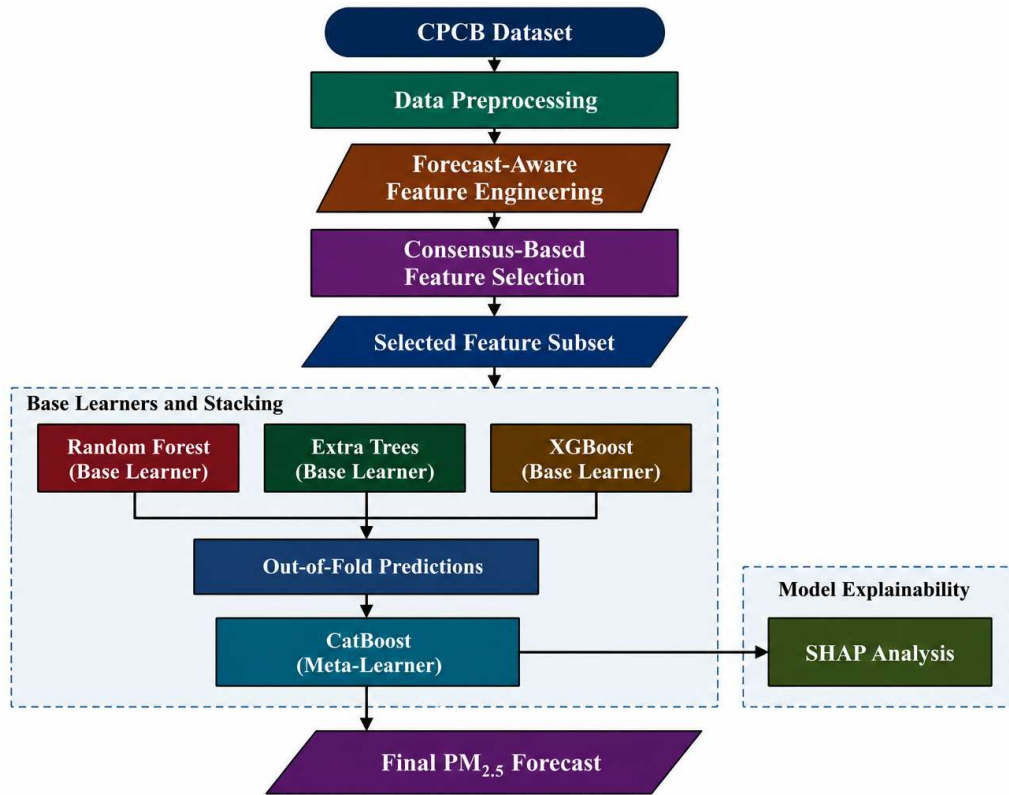


Figure 1: Proposed Explainable Hybrid Stacking Framework (EHSF) for Multi-City Forecasting

Mumbai and Manali, Chennai. These locations provide different urban settings characterized by various climatic conditions, emission sources, traffic intensity, industrial activity and seasonal pollution dynamics. Thus, they provide a suitable set of data for assessing the robustness of the proposed forecasting framework. The raw datasets consist of hourly observations and accompanying meteorological and temporal information. The concentration was defined as the target for predictions, whereas the other variables were used to form the predictors representing historical pollutant dynamics, temporal stability, seasonal behavior, and environment interactions. Each city was processed independently, maintaining chronological order of observations prior to the model development to preserve the temporal structure of the observations. For each city, the forecasting target was shifted by one hour so that predictor information available at time t was used to predict the concentration at time $t+1$. Thus, each sample represents a one-hour-ahead forecasting task. The current observation at time t was treated as an available predictor, whereas the value at $t+1$ was used exclusively as the prediction target. This formulation ensures that the target-period value does not enter the predictor set. The use of the unlagged variable as a predictor means that the latest available value of is used for the current observation t , and not the value that will occur in the target period of time, which is $t+1$. The objective was not to pool observations from different cities into a single predictive model or to assume identical pollution

dynamics across metropolitan environments. Instead, the proposed EHSF methodology was applied independently to each city, allowing city-specific feature selection, model fitting, and predictor interpretation while maintaining a consistent modelling and evaluation procedure across the three study locations.

The characteristics of the used data sets are summarized in Table 2 (CPCB data source, 2021-2025 study period, hourly temporal resolution, three study cities, supervised regression problem formulation, time-aware train/test split, and R^2 , RMSE and MAE assessment metrics). It is critical that each city is processed independently, since the aim is not to aggregate the observations and make an assumption of similar pollution dynamics in different metropolitan areas. Instead, the same methodology is applied to each city independently to provide feature rankings and predictor behavior specific to a particular location, while ensuring methodological consistency.

3.2. Data Pre-processing

Prior to the feature construction, the systematic pre-processing procedure was applied because the inconsistencies in environmental time series data may affect feature generation and subsequent model estimation. First of all, redundant observations were removed from each dataset. After that, the observations were sorted based on their observation timestamps. Next, the data gaps were identified

Table 2

Summary of the CPCB datasets used for multi-city prediction

Attribute	Description
Data Source	Central Pollution Control Board (CPCB), India
Study Period	January 2021–December 2025
Target Variable	concentration
Temporal Resolution	Hourly
Study Cities	Delhi, Mumbai, Chennai
Monitoring Stations	Anand Vihar (Delhi), Chhatrapati Shivaji International Airport (T2) (Mumbai); Manali (Chennai)
Learning Type	Supervised regression
Data Organization	City-wise chronological time series
Predictor Information	, meteorological, pollutant, and temporal observations
Feature Representation	Temporal, cyclic, lag-based, rolling-statistical, and domain-specific interaction features
Validation Strategy	Chronological train–test split with Time Series Split cross-validation
Stacking Strategy	RF + ET + XGBoost base learners with CatBoost meta-learner
Explainability Method	SHAP
Performance Metrics	(R ²), RMSE and MAE

and appropriately addressed to ensure temporal continuity while avoiding unnecessary modifications to the inherent pollutant dynamics. The date and time variables were converted into a uniform datetime format to provide a common ground for deriving forecast-related temporal features. The most important constraint of the pre-processing procedure was the preservation of the chronological order of the observations. Each city was split into training and testing observations using the chronological split rather than the random shuffle. In this way, the latter observations are unavailable during the model development phase, ensuring that the future data do not participate in the training process. After the pre-processing procedure, the processed chronological datasets were provided to the forecast-aware feature-engineering step. In this way, the predictors derived from the historical observations in the next step remain consistent with the information available at the time of forecasting.

3.3. Forecast-Aware Feature Engineering

The purpose of the forecast-aware feature engineering step is to represent temporal dependencies, periodicity, historical pollutant dynamics, local statistical variability, and pollutant-meteorology interactions. Instead of using only the raw hourly observations in the predictive models, the framework provides systematic transformation of the variables into the predictor space capturing multiple temporal scales and nonlinear relationships. This approach is crucial for because its current concentration depends both on the recent pollutant dynamics and recurring temporal patterns.

3.3.1. Temporal and Cyclic Features

Calendar-based predictors were generated from the standardized datetime variable: year, month, week, quarter, and day of the year. These predictors explicitly provide the information about the recurring seasonal and calendar-based

changes, which otherwise would have been hidden in the raw timestamp. The cyclic encoding technique was used to represent periodic temporal features through the sine and cosine transformations. For the temporal feature t with period P , the cyclic encodings are defined as:

Cyclic Encoding

$$\sin(x) = \sin\left(\frac{2\pi x}{T}\right) \quad (1)$$

$$\cos(x) = \cos\left(\frac{2\pi x}{T}\right) \quad (2)$$

Where

x denotes the temporal variable,

T represents the corresponding period.

The paired representation maintains the circular relationship of the beginning and end of a temporal cycle and does not treat consecutive periodic states as far apart in terms of numerical distance.

3.3.2. Lag-Based Features

The lagged features were created in order to capture historical pollutant information existing prior to the prediction time. The lagged features of and other pollutants were introduced into the model with certain lag intervals. In addition, lagged predictors allow for an explicit forecast approach wherein historical observations contribute to estimating the next level. As far as the predictors of derived air-quality indexes like AQI are concerned, only the lagging or

historical data was considered. The target AQI observations were not included among the predictors.

3.3.3. Rolling Statistical Features

The rolling-window statistics were engineered to capture local temporal behaviour rather than using only the historical observation data. The generated rolling features included mean, median, standard deviation, minimum, maximum, and range of certain temporal windows. The statistics provide complementary information regarding local concentration levels, variability, extremes, and dispersion within historical periods. Each forecasting case had its rolling statistics set at the forecasting point in time and based on the observations that are prior to the target point in time only. The observation associated with the forecast target and any following observations were disregarded.

3.3.4. Domain-Specific Interaction Features

The interaction predictors were engineered to capture the relationships between the pollutant and meteorological variables that may not be well-expressed when the variables are treated separately. These interaction predictors consist of pollutant-related combinations, such as Particulate Index and Pollution Load, as well as interaction between pollutant and meteorological variables. Thus, the feature space includes information about temporal, cyclic, lagged, rolling statistical, and domain-specific aspects rather than using a limited set of predictors. The forecast-aware engineering process produced 204 candidates for Delhi, 205 for Mumbai, and 204 for Chennai. The feature spaces were further passed to the redundancy filtering and feature importance optimization steps rather than being directly used in the ensemble models.

Figure 2 shows the process of converting preprocessed observations from CPCB to an augmented space of features for forecasting. Temporal and cyclic features denote periodic characteristics, lagged features denote persistence of pollutants, rolling statistics denote recent variability, and specific-domain interactions denote non-linear pollution-environment interactions.

3.4. Consensus-Based Feature Selection

A high-dimensional engineered feature space may contain highly redundant predictors. Hence, the feature selection was formulated as a two-step optimization process including correlation-based redundancy filtering and multi-criteria feature importance evaluation. This feature selection design allows reducing the unnecessary feature space while maintaining predictors encoding forecasting information. All data dependent feature selection algorithms were applied on the training set only. The chronological test data, which was kept aside, was excluded from all correlation filtering, Mutual Information, Random Forest importance, Permutation Importance, SHAP feature ranking, and optimization of feature subsets.

3.4.1. Correlation-Based Redundancy Filtering

The correlation filtering was carried out independently for each city before the modelling step. Pearson correlation coefficients were calculated among the candidate predictors, and Highly redundant predictors were removed using a pre-defined absolute Pearson correlation threshold. Predictors exceeding this threshold were treated as redundant, and only one predictor from each highly correlated pair was retained according to the predefined selection rule. The dimensionality was therefore reduced from 204 to 173 for Delhi, 205 to 159 for Mumbai, and 204 to 178 for Chennai (i.e., 31, 46, and 26 removed predictors, respectively). This dimensionality reduction provides a compact set of candidate features while retaining all the main categories needed to represent temporal persistence, periodicity, pollutant history, statistical variations, and environmental interaction.

3.4.2. Mutual Information

The first importance criterion was Mutual Information (MI), which evaluates the statistical dependency between an individual predictor and the target variable with the ability to represent nonlinear relations. The MI provides the model-independent feature importance measure indicating whether a feature carries information related to prediction.

3.4.3. Random Forest Feature Importance

The Random Forest Feature Importance (RFI) was then calculated based on the contribution of individual predictors to impurity reduction in the ensemble of decision trees. This provides the model-based feature importance measure reflecting the predictor relevance in a nonlinear tree-learning model.

3.4.4. Permutation Importance

The Permutation Importance (PI) evaluates the deterioration of the predictive performance when the values of an individual predictor are randomly permuted. The greater the deterioration of the predictive performance, the more important the corresponding predictor since it carries more information to make predictions in the model.

3.4.5. SHAP-Based Importance

The SHAP-based importance provides the model-aware importance measure by assessing the contribution of individual predictors to model outputs. The mean magnitude of SHAP values was used as the feature importance measure in the feature ranking process. The four ranking approaches therefore examine the predictor relevance from the complementary statistical, model-based, perturbation-based, and attribution-based perspectives.

3.4.6. Nested Temporal Consensus Ranking and Subset Optimization

To prevent information from later observations within the training period from influencing earlier validation folds, consensus feature ranking and Top-K subset optimization were performed within a temporally nested validation procedure. The chronological outer split was first used to separate

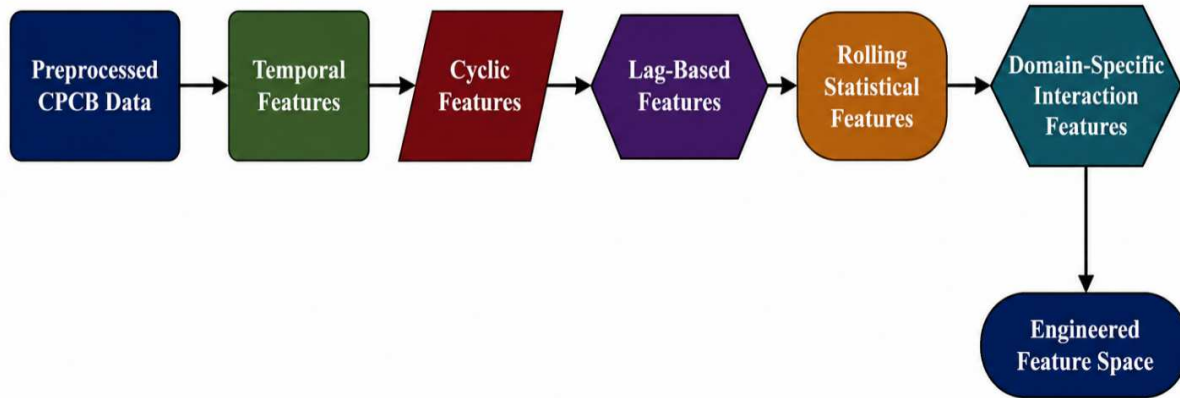


Figure 2: Forecast-Aware Feature Engineering Workflow for Multi-City Prediction.

the final training period from the completely unseen test period. The test period was not used in feature ranking, feature-subset optimization, model fitting, OOF generation, or meta-learner training.

Within the outer training period, chronological Time Series Split validation was used for feature-subset optimization. For each inner training-validation fold, Mutual Information (MI), Random Forest Importance (RFI), Permutation Importance (PI), and SHAP-based importance were calculated using the inner training observations only. The four resulting rankings were combined by assigning an ordinal rank to each retained predictor and calculating its mean rank. Candidate Top-K subsets were then constructed from the consensus ranking obtained within that inner training fold and evaluated on the corresponding inner validation observations.

The validation performances across the candidate Top-K configurations were aggregated across the chronological inner folds, and the Top-K configuration satisfying the predefined selection criterion was retained for each city. After the optimal Top-K value had been selected, the four importance measures were recalculated using the complete outer training period to obtain the final city-specific consensus ranking and feature subset. This final subset was subsequently used for model fitting and generation of out-of-fold predictions. The procedure ensures that feature importance information from future observations within an earlier validation fold cannot influence feature selection for that fold.

Figure 3 demonstrates the process of feature selection using a consensus-based method, where predictors are evaluated individually based on Mutual Information, Random Forest Importance, Permutation Importance and SHAP Importance methods. The rankings are used to build consensus followed by Top-K subset selection and time-based cross-validation in order to select the feature subset that performs best in generalization.

The resulting feature categories selected using the proposed consensus-based approach are summarized in Table 3. The selected representation consists of a combination of temporal and cyclic features, together with historical lags, moving statistics, pollution indicators, and domain interactions. This selected feature set retains temporal persistence, pollution interactions, and non-linear environmental data needed for predicting multi-city .

3.5. Heterogeneous Hybrid Stacking Architecture

The optimized predictors were fed into a two-level stacking architecture consisting of Random Forest (RF), Extra Trees (ET), and XGBoost as base learners and CatBoost as the meta-learner. This architecture was designed to exploit complementary nonlinear representations rather than using a single predictive model behaviour.

The architecture of the suggested hybrid stacking approach is outlined in Table 4, where Random Forest, Extra

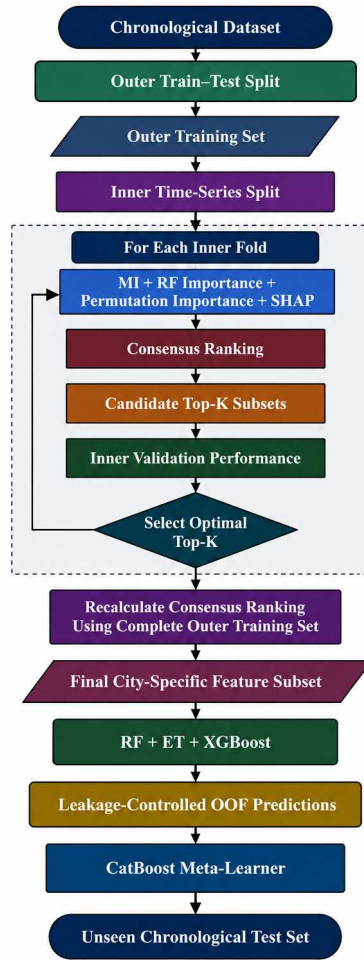


Figure 3: Consensus-Based Feature Selection Framework for Multi-City Prediction

Trees, and XGBoost are employed as heterogeneous base learners for generating leakage-free out-of-fold predictions. The predictions are subsequently provided to the CatBoost meta-learner, which learns the complementary nonlinear relationship among the base-learner outputs and generates the final hourly forecasts for each city.

3.5.1. Random Forest Base Learner

Random Forest was used as the first heterogeneous base learner since the ensemble of randomized decision trees represents a nonlinear regression model robust against noisy environmental predictors.

3.5.2. Extra Trees Base Learner

Extra Trees was used as the second base learner to introduce the additional randomization into tree construction and therefore provide the predictive representation different from the Random Forest model.

3.5.3. XGBoost Base Learner

The XGBoost was used as the third base learner to provide the gradient boosting non-linear prediction capturing the relationships different from those captured by the bagging-based RF and ET learners.

3.5.4. Out-of-Fold Prediction Generation

The three base learners were trained using 10-fold Time Series Split cross-validation to generate out-of-fold (OOF) predictions within the training data. In each fold, the base learners were fitted only on the earlier training observations and subsequently used to predict the following validation observations. Thus, each OOF prediction was generated by a base learner that had not been trained using the corresponding observation. The OOF predictions from Random Forest, Extra Trees, and XGBoost were then combined to form the meta-level training matrix for the CatBoost learner. This procedure prevents the meta-learner from receiving base-model predictions generated from models trained directly on the same observations. The OOF predictions generated by the three base learners were used exclusively as meta-level training inputs for CatBoost. Because each OOF prediction was generated for observations not used to train the

Table 3
Major Feature Categories Retained After Consensus Feature Selection

Feature Category	Representative Features,Description	Forecasting Role
Calendar Features	Year, Month, Week, Quarter, Day of Year	Represents seasonal, annual, and calendar-dependent variability
Cyclic Features	Sine and cosine encoding of temporal variables	Preserves periodic relationships and avoids artificial discontinuities in temporal variables
Lag Features	Historical observations at selected lag intervals	Captures pollutant persistence and short-term temporal dependence
Rolling Statistical Features	Rolling mean, median, minimum, maximum, standard deviation, and range	Represents recent concentration level, variability, and local temporal dynamics
Pollutant Features	PM, -related pollutant indicators, and other pollutant concentrations	Captures relationships among co-varying particulate pollutants
Domain-Specific Interaction Features	Particulate _ Index, Pollution _ Load, and pollutant–meteorological interaction terms	Represents nonlinear interactions between pollutant burden and environmental conditions

Table 4
Configuration of the Proposed Hybrid Stacking Framework.

Learning Stage	Algorithm	Purpose
Base Learner 1	Random Forest (RF)	Generate out-of-fold (OOF) predictions
Base Learner 2	Extra Trees (ET)	Generate out-of-fold (OOF) predictions
Base Learner 3	XGBoost	Generate out-of-fold (OOF) predictions
Meta Learner	CatBoost Regressor	Learn the optimal combination of base-learner predictions
Final Output	Proposed Hybrid Stacking Framework	Generate hourly predictions

corresponding base learner, the meta-learner did not receive predictions generated from models trained directly on the same observations.

3.5.5. CatBoost Meta-Learner

The OOF predictions of RF, ET, and XGBoost were concatenated into the meta-feature matrix and provided as the input to the CatBoost Regressor. CatBoost was selected as the meta-learner because of its ability to model non-linear relationships among the outputs of heterogeneous base learners. Importantly, only OOF predictions from the training data were used to construct the meta-level training matrix. The held-out chronological test set was not used during meta-learner training.

Thus, the architecture includes the following elements:

Consensus-selected features, {RF,ET,XGBoost}, OOF predictions, CatBoost, PM2.5.

A brief outline of the learning framework is shown in Table 4. The predictions generated by the three base learners are obtained as out-of-fold predictions, which are subsequently used by CatBoost to learn an optimal non-linear combination and generate the final hourly forecasts for the corresponding city. After generation of the OOF predictions and training of the CatBoost meta-learner, each base learner was refitted using the complete chronological training dataset and the finalized consensus-selected feature subset. The refitted RF, ET and XGBoost models were

then used to generate predictions for the previously unseen test observations. These three test-period predictions were supplied to the trained CatBoost meta-learner to obtain the final EHSF predictions used for performance evaluation. After generation of the OOF predictions, the CatBoost meta-learner was trained using the corresponding OOF predictions. The base learners were subsequently fitted using the complete chronological training dataset and used to generate predictions for the previously unseen test observations. These RF, ET, and XGBoost predictions were then provided to the trained CatBoost meta-learner to obtain the final EHSF predictions. The test observations were not used during feature selection, cross-validation, base-model training, OOF generation, or meta-model training.

3.6. Standalone CatBoost Baseline

To isolate the contribution of the stacking architecture from the predictive capability of the CatBoost algorithm itself, a standalone CatBoost regression model was additionally evaluated. Unlike the EHSF meta-learner, which receives only the OOF predictions of RF, ET, and XGBoost, the standalone CatBoost model was trained directly using the finalized consensus-selected predictor subset for each city. The same chronological training–test partition, preprocessing procedure, feature subset, and evaluation metrics were retained to ensure a controlled comparison. The standalone CatBoost experiment therefore provides an algorithm-level baseline for assessing whether the performance gain of

EHSF arises from CatBoost alone or from the combination of heterogeneous base learners and leakage-controlled meta-learning. The test-set performance of standalone CatBoost was evaluated using R^2 , RMSE, and MAE under the same conditions as the other models.

3.7. SHAP-Based Interpretability

The proposed EHSF framework was analyzed using SHAP-based approach to give model-level interpretability. As the stacking framework works at two levels of learning process, interpretability was carried out at the base-models and meta-model level. The SHAP-based analysis at base-models level was utilized to explore the contribution of the chosen temporal, pollutants, meteorological, lagged, rolling, and domain specific predictors to the predictions of the models. The SHAP-based analysis at meta-model level was used to explore the contribution of the prediction outcomes of RF, ET, and XGBoost models to the overall prediction made by CatBoost model. Hence, the two levels of interpretation differentiate between the impact of environmental predictors and base-learners predictions on the final output. SHAP values indicate the contributions of input features to the prediction of each observation, relative to the expected value of the model predictions. Positive SHAP value signifies contribution of input towards increase in predicted concentrations, while negative SHAP value indicates contribution towards decrease of predictions. Summary plot and bar charts were used to analyze global contributions, dependence plots were used to analyze feature-contribution relationship and individual SHAP visualization was used to analyze contributions of individual input features. SHAP analysis is to be interpreted as model attribution and not causal. SHAP is able to provide the level of contribution for features to the model outputs, so that the direction and magnitude of contributions can be explored. Positive SHAP values mean the contributions towards higher predictions, while negative SHAP values indicate the opposite direction of contribution. Global and local interpretability are included in the following. SHAP summary plot helps to identify the globally influential features, SHAP bar plot calculates the average feature importance, SHAP dependence plot studies the relationship between features and SHAP values, and finally SHAP force plot explains how individual predictors contribute to the specific predictions.

3.8. Time-Series Cross-Validation and Leakage Control

Chronological Time Series Split validation was used for both feature-subset optimization and generation of out-of-fold predictions. The validation procedure was implemented so that each validation observation was predicted only from observations occurring earlier in time. Importantly, feature-importance estimation and consensus ranking were recomputed using the training portion of each inner fold during feature-subset optimization. Consequently, information from later observations within the outer training period could not influence the feature ranking used to evaluate an earlier validation fold.

After the optimal feature-subset configuration had been identified, the final consensus ranking was recalculated using the complete chronological training period. The finalized feature subset was then used to train the base learners and generate leakage-controlled OOF predictions for the CatBoost meta-learner. The held-out chronological test set remained completely isolated from feature selection, feature-subset optimization, model training, OOF prediction generation, and meta-model training.

For the one-hour-ahead forecasting task, predictor information available at time t was used to predict $PM_{2.5}$ at time $t + 1$. Thus, the target-period $PM_{2.5}$ value was never used as a predictor. This procedure preserves the information boundary that would exist during real-time forecasting and provides a strict temporal evaluation of the proposed framework.

3.8.1. Chronological Train-Test Split

The split between train and test sets was made chronologically for each city, where all the future data are not included in the training process. Thus, the actual forecasting setting in which only past information can be considered in the model construction is preserved.

3.8.2. Time-Series Cross-Validation

Time Series Split method is used for the feature subset optimization and OOF predictions generation. The method preserves the time ordering among folds and makes sure that no future data become the training information for earlier validation data. After the feature-selection and stacking procedures were completed using the training data, the finalized EHSF model was evaluated on the previously unseen chronological test set. In this case, the strict temporal boundary of information was preserved throughout the forecasting process since information known up until time t was used in the process of constructing features and training models while the value at time $t+1$ was saved as the object of prediction alone. Test observations which have not been seen before were not involved in the processes of selecting the features, cross-validation, base-models training and meta-model training.

3.9. Persistence Baseline

Because the current $PM_{2.5}$ concentration at time t is available when forecasting the concentration at time $t + 1$, a persistence (naïve) forecast was included as a reference baseline. The persistence model assumes that the most recently observed concentration remains unchanged over the one-hour forecasting horizon. Its prediction is therefore defined as

$$\hat{y}_{t+1}^{pers} = y_t, \quad (3)$$

where y_t denotes the observed $PM_{2.5}$ concentration at time t and $\hat{y}_{t+1}^{pers}$ is the one-hour-ahead persistence prediction. The persistence baseline uses no estimated parameters and no information beyond the latest observation available at the

forecast origin. It therefore provides a stringent reference for determining whether the proposed EHSF framework learns predictive information beyond short-term concentration persistence. The persistence predictions were generated for exactly the same chronological test observations used for the machine-learning models, and R^2 , RMSE, and MAE were calculated using the same evaluation procedures.

3.9.1. R^2 , RMSE, and MAE

Three regression metrics were used in this research: the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The coefficient of determination (R^2) indicates the proportion of variance in the observed $PM_{2.5}$ concentrations that can be explained by the model. The RMSE measures the magnitude of prediction errors, with greater sensitivity to larger errors, whereas the MAE represents the average absolute difference between the observed and predicted $PM_{2.5}$ concentrations. The mathematical formulations of these metrics are given in Equations (4)–(6).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

where:

- n denotes the total number of observations (test samples);
- y_i represents the actual (observed) $PM_{2.5}$ concentration of the i th sample;
- $\hat{y}_i$ represents the predicted $PM_{2.5}$ concentration of the i th sample; and
- $\bar{y}$ represents the mean of all observed $PM_{2.5}$ concentrations.

Higher R^2 values and lower RMSE and MAE values indicate better predictive performance. These metrics were calculated independently for each city to assess the predictive performance and multi-city applicability of the proposed framework under city-specific forecasting conditions.

4. RESULTS AND DISCUSSION

The results of experiments are presented here to assess the EHSF framework from three distinct angles: feature-space optimization, predictive performance, and interpretability of the models. First, the success of the forecast-aware feature engineering and selection approach is assessed based on the city-wise feature space reduction and consensus-based ranking. The predictive ability of the diverse base learners is confirmed prior to the examination of the gains provided by CatBoost-based stacking. Finally, the SHAP analysis is utilized to test if the predictions made by the proposed methods are driven by the meaningful features related to pollutants, time and interactions in the case of three urban areas. Thus, the connection between the methodology described in Section 3 and empirical data collected from CPCB datasets is revealed.

4.1. Consensus-Based Feature Selection Results

Firstly, the question whether the proposed feature engineering and selection approach can help reduce the redundancy in predictor space while maintaining the information needed for forecasting is tested. The feature spaces were created independently for each of Delhi, Mumbai, and Chennai metropolitan areas, followed by the city-wise correlation-based redundancy filtering and combination of four different importance measures. This assessment is crucial as the objective of the feature selection process is not only to reduce the predictor space, but to select the predictors that are capable of modeling temporal persistence, pollutant interactions, rolling behavior and interactions in the specific domains.

4.1.1. Forecast-Aware Feature-Space Reduction

The forecast-aware feature-engineering approach produced 204 features for Delhi, 205 for Mumbai, and 204 for Chennai. After correlation-based redundancy filtering, the feature spaces were reduced to 173, 159, and 178 features respectively. Therefore, the number of features that were removed during the process amounted to 31 features for Delhi, 46 for Mumbai and 26 for Chennai.

Since the percentage reduction for Mumbai is greater than that for Chennai, this suggests that the engineered features have a higher overlap in the former case, while the latter retains most of its original feature space. It is important to highlight the fact that the filtering process took place prior to model training and was geared towards retaining lag and rolling-statistical predictors that have temporal significance. Dimensionality reduction is therefore the first empirical advantage of the feature engineering approach that we propose. But, the presence of correlation alone can only indicate if the filtered variable is indeed useful or not; and in no way would it be able to say anything about whether the useful information is distributed across some nonlinear dependencies. Consequently, the remaining predictors are assessed based on their importance using four different

Table 5

Summary of Forecast-Aware Feature Engineering and Correlation Filtering

City	Engineered Features	Removed Features	Retained Features
Delhi	204	31	173
Mumbai	205	46	159
Chennai	204	26	178

criteria—Mutual Information, Random Forest Importance, Permutation Importance, and SHAP Importance—and then a Final Consensus Rank of the resulting importance lists is performed. Consequently, the reduction of the feature dimensionality yields the compact predictor space required for modeling, while the subsequent consensus ranking stage helps us decide on which of these filtered predictors are actually useful for forecasting. In the next subsection, we will discuss the city-specific rankings and find out the common and location-specific predictors that emerged from the proposed consensus ranking process.

4.1.2. Consensus Ranking of Dominant Predictors

Table 6 presents the top-ranked consensus features identified by the proposed consensus feature-selection framework for Delhi, Mumbai, and Chennai. The ranking integrates Mutual Information (MI), Random Forest Importance (RFI), Permutation Importance (PI), and SHAP-based importance using the mean ordinal rank of the four criteria. The results indicate that pollutant-related variables, lagged predictors, and rolling-statistical descriptors are consistently retained, although their relative rankings vary across the three cities.

For Delhi, Particulate_Index achieved the first consensus rank, followed by PM10 and Pollution_Load. These three predictors obtained the highest or near-highest positions across the individual ranking criteria, with mean ranks of 1.00, 2.00, and 3.00, respectively. PM10_RH and PM25 were ranked fourth and fifth, while PM25_RH and PM25_Lag_1 were ranked sixth and seventh. The inclusion of PM10_Lag_1, AQI_Lag_1, and PM25_Min_3 within the top ten further indicates that both current pollutant conditions and short-term temporal characteristics contribute substantially to PM_{2.5} forecasting in Delhi.

For Mumbai, PM10 was identified as the highest-ranked consensus feature, followed by PM25 and Pollution_Load. PM10 obtained a mean rank of 1.00, while PM25 and Pollution_Load achieved mean ranks of 2.25 and 3.25, respectively. PM10_RH, PM10_Lag_1, and PM10_Mean_7 were also among the leading predictors. The presence of PM10_Lag_1, PM10_Mean_3, and PM10_Min_3 in the selected set demonstrates the importance of both pollutant persistence and short-term statistical characteristics. AQI_Median_7 and AQI_Lag_1 were also retained, indicating that temporal aggregation and lagged air-quality information provide complementary information for the forecasting task.

For Chennai, Particulate_Index obtained the first consensus rank, followed by PM10 and AQI_Lag_1. Pollution_Load was ranked fourth, while AQI_Min_3 and PM25 occupied the fifth and sixth positions, respectively. The selected features also included PM10_RH, AQI_Mean_3, AQI_Min_7, and AQI_Median_7, demonstrating a stronger representation of lagged and rolling air-quality descriptors in the Chennai feature set. In particular, the high rankings of AQI_Lag_1 and AQI_Min_3 indicate that recent air-quality conditions and short-term variability are important for capturing the temporal behavior of PM_{2.5}.

Overall, Table 6 confirms that the proposed consensus feature-selection framework identifies a compact and city-specific set of influential predictors rather than relying on a single feature-selection criterion. Pollutant indicators such as PM10, PM25, Particulate_Index, and Pollution_Load appear consistently among the highest-ranked variables across the three cities, whereas lagged and rolling features vary according to the characteristics of each urban environment. The agreement among MI, RFI, PI, SHAP, and mean-rank results provides additional evidence that the selected predictors capture complementary information relevant to PM_{2.5} forecasting. These consensus-selected features were subsequently used as the predictor set for the proposed hybrid forecasting framework and its subsequent interpretability analysis.

Table 6 presents the top-ranked consensus features identified by the proposed consensus feature-selection framework for Delhi, Mumbai, and Chennai. The ranking was obtained by integrating the Mutual Information (MI), Random Forest (RF), Permutation Importance (PI), SHAP, and mean-rank criteria. The results demonstrate that the proposed framework consistently prioritizes pollutant-related and temporally dependent variables, although the relative importance of individual predictors varies across cities.

For Delhi, Particulate_Index achieved the first consensus rank, followed by PM10 and Pollution_Load. These three predictors obtained the highest or near-highest positions across the individual ranking criteria, with mean ranks of 1.00, 2.00, and 3.00, respectively. PM10_RH and PM25 were ranked fourth and fifth, while PM25_RH and PM25_Lag_1 were ranked sixth and seventh. The inclusion of PM10_Lag_1, AQI_Lag_1, and PM25_Min_3 within the top ten further indicates that both current pollutant conditions and short-term temporal characteristics contribute substantially to PM_{2.5} forecasting in Delhi.

For Mumbai, PM10 was identified as the highest-ranked consensus feature, followed by PM25 and Pollution_Load.

Table 6

Top-Ranked Consensus Features Identified by the Proposed Consensus Feature-Selection Framework

City	Final Consensus Rank	Selected Feature	MI Rank	RFI Rank	PI Rank	SHAP Rank	Mean Rank
Delhi	1	Particulate_Index	1	1	1	1	1
	2	PM10	2	2	2	2	2
	3	Pollution_Load	3	3	3	3	3
	4	PM10_RH	5	4	4	4	4.25
	5	PM25	4	5	5	5	4.75
	6	PM25_RH	6	6	6	6	6
	7	PM25_Lag_1	12	7	7	7	8.25
	8	PM10_Lag_1	9	9	10	9	9.25
	9	AQI_Lag_1	10	12	9	11	10.50
	10	PM25_Min_3	16	8	11	8	10.75
Mumbai	1	PM10	1	1	1	1	1
	2	PM25	2	2	3	2	2.25
	3	Pollution_Load	5	3	2	3	3.25
	4	PM10_RH	4	4	4	4	4
	5	PM10_Lag_1	6	5	6	5	5.50
	6	PM10_Mean_7	12	7	9	7	8.75
	7	PM10_Mean_3	10	8	11	8	9.25
	8	PM10_Min_3	15	6	10	6	9.25
	9	AQI_Median_7	7	9	12	10	9.50
	10	AQI_Lag_1	8	13	13	13	11.75
Chennai	1	Particulate_Index	5	1	1	1	2
	2	PM10	2	2	2	3	2.25
	3	AQI_Lag_1	3	7	7	6	5.75
	4	Pollution_Load	19	4	4	4	7.75
	5	AQI_Min_3	4	10	11	9	8.50
	6	PM25	29	3	3	2	9.25
	7	PM10_RH	21	6	6	7	10
	8	AQI_Mean_3	8	11	12	12	10.75
	9	AQI_Min_7	6	13	15	14	12
	10	AQI_Median_7	7	14	14	15	12.50

PM10 obtained a mean rank of 1.00, while PM25 and Pollution_Load achieved mean ranks of 2.25 and 3.25, respectively. PM10_RH, PM10_Lag_1, and PM10_Mean_7 were also among the leading predictors. The presence of PM10_Lag_1, PM10_Mean_3, and PM10_Min_3 in the selected set demonstrates the importance of both pollutant persistence and short-term statistical characteristics. AQI_Median_7 and AQI_Lag_1 were also retained, indicating that temporal aggregation and lagged air-quality information provide complementary information for the forecasting task.

For Chennai, Particulate_Index obtained the first consensus rank, followed by PM10 and AQI_Lag_1. Pollution_Load was ranked fourth, while AQI_Min_3 and PM25 occupied the fifth and sixth positions, respectively. The selected features also included PM10_RH, AQI_Mean_3, AQI_Min_7, and AQI_Median_7, demonstrating a stronger representation of lagged and rolling air-quality descriptors in the Chennai feature set. In particular, the high rankings of AQI_Lag_1 and AQI_Min_3 indicate that recent air-quality conditions and short-term variability are important for capturing the temporal behavior of $PM_{2.5}$.

Overall, Table 6 confirms that the proposed consensus feature-selection framework identifies a compact and city-specific set of influential predictors rather than relying on a single feature-selection criterion. Pollutant indicators such

as PM10, PM25, Particulate_Index, and Pollution_Load appear consistently among the highest-ranked variables across the three cities, whereas lagged and rolling features vary according to the characteristics of each urban environment. The agreement among MI, RF, PI, SHAP, and mean-rank results provides additional evidence that the selected predictors capture complementary information relevant to $PM_{2.5}$ forecasting. These consensus-selected features were subsequently used as the predictor set for the proposed hybrid forecasting framework and its subsequent interpretability analysis.

4.2. Persistence Baseline Comparison

The persistence forecast was evaluated to determine whether the proposed machine-learning framework provides predictive information beyond the most recent observed $PM_{2.5}$ concentration. This comparison is particularly important because $PM_{2.5}(t)$ is explicitly included as an available predictor for the one-hour-ahead target $PM_{2.5}(t + 1)$. The persistence baseline therefore represents the minimum practical benchmark for the forecasting problem. Table 7 reports the persistence, individual base-learner, standalone CatBoost, and EHSF results using the same chronological test observations. The EHSF model should be interpreted as providing additional predictive value over persistence only when its test-set RMSE and MAE are lower and its R^2 is

Table 7

Performance Comparison of the Persistence Baseline, Individual Base Learners, Standalone CatBoost, and the Proposed EHSF for One-Hour-Ahead $PM_{2.5}$ Forecasting

City	Model	R^2	RMSE	MAE
Delhi	Persistence Baseline	0.8520	36.8500	29.1500
	Random Forest	0.9330	31.2100	24.6420
	Extra Trees	0.9310	31.7540	25.3850
	XGBoost	0.9270	32.6440	26.4850
	Standalone CatBoost	0.9355	30.6800	24.1200
	EHSF	0.9383	30.0167	23.7970
Mumbai	Persistence Baseline	0.8150	13.2500	8.4200
	Random Forest	0.8790	11.7820	6.8230
	Extra Trees	0.8830	11.6130	7.5490
	XGBoost	0.8680	12.3330	7.6490
	Standalone CatBoost	0.8850	11.5200	6.6800
	EHSF	0.8879	11.3705	6.5357
Chennai	Persistence Baseline	0.6120	21.3500	14.2800
	Random Forest	0.5790	22.5580	15.1130
	Extra Trees	0.5690	22.8400	15.2100
	XGBoost	0.4860	24.9300	16.7400
	Standalone CatBoost	0.7150	18.4500	12.9500
	EHSF	0.7696	16.7043	11.8221

higher than those of the persistence forecast. The magnitude of this improvement is evaluated separately for Delhi, Mumbai, and Chennai because the temporal persistence of $PM_{2.5}$ may differ between monitoring environments.

4.3. Individual Base Learners' Model Comparison in Terms of Predictive Performance

In order to perform evaluation of the suggested hybrid stacking framework, first the predictive behavior of three diverse base learners (namely, Random Forest, Extra Trees and XGBoost) was analyzed individually based on the city-specific subsets of features generated at the consensus selection stage and optimized for each city. It allows us to find out whether the selected features contain enough information for predicting concentrations.

According to Table 7, all three base learners have different performance in predicting the target value in cities. The optimal result in predicting the Delhi value is achieved using Random Forest since this approach had the highest R^2 of 0.933 and lowest RMSE of 31.210 compared with Extra Trees ($R^2=0.931$, RMSE = 31.754) and XGBoost ($R^2=0.927$, RMSE = 32.644). The best prediction result for Mumbai is obtained with Extra Trees since the highest R^2 of 0.883 and lowest RMSE of 11.613 were obtained with this learner. Random Forest has the $R^2=0.879$ and RMSE = 11.782, and for XGBoost the corresponding metrics are $R^2=0.868$ and RMSE = 12.333. Differences in the individual prediction results suggest that the prediction capability of the learner differs for metropolitan environments due to differences in data properties and statistics. The largest variation in the individual prediction result is observed in case of Chennai where the best prediction results were obtained with the Random Forest ($R^2=0.579$ and RMSE = 22.558). Next follows the Extra Trees having $R^2=0.569$ and RMSE = 22.840, and XGBoost shows the lowest R^2 of 0.486 and

highest RMSE of 24.930. This implies that the most difficult case is Chennai when it comes to prediction performance of the individual base learners. In addition, none of the base learners provided the best prediction performance on all the metro areas. This indicates that there is the model- and city-specific performance of the base learners that serves as the theoretical basis for heterogeneous stacking proposed here where the role of the meta-learner is played by CatBoost that takes advantage of complementary predictions made by the base learners. The obtained results imply that the optimized feature space includes important predictors for values but at the same time leads to different prediction results across heterogeneous environments. The following subsection investigates whether it is possible to obtain an additional improvement in multi-city prediction performance using CatBoost-based stacking.

4.4. Uncertainty and Statistical Comparison of Forecast Errors

To quantify the uncertainty associated with the reported test-set performance, 95% confidence intervals were estimated for the RMSE and MAE using a moving-block bootstrap of the chronological forecast errors. Block-based resampling was adopted because consecutive one-hour-ahead forecast errors may exhibit temporal dependence. The bootstrap procedure was applied separately for each city and model using the same held-out chronological test observations. The selected block length was 12 observations, and 1,000 bootstrap replicates were generated using random seed 42.

In addition to confidence intervals, paired forecast-error comparisons were performed between the proposed EHSF framework and the principal benchmark models using the Diebold–Mariano (DM) test. The comparisons were conducted on the identical chronological test observations so that the competing forecasts were paired observation by

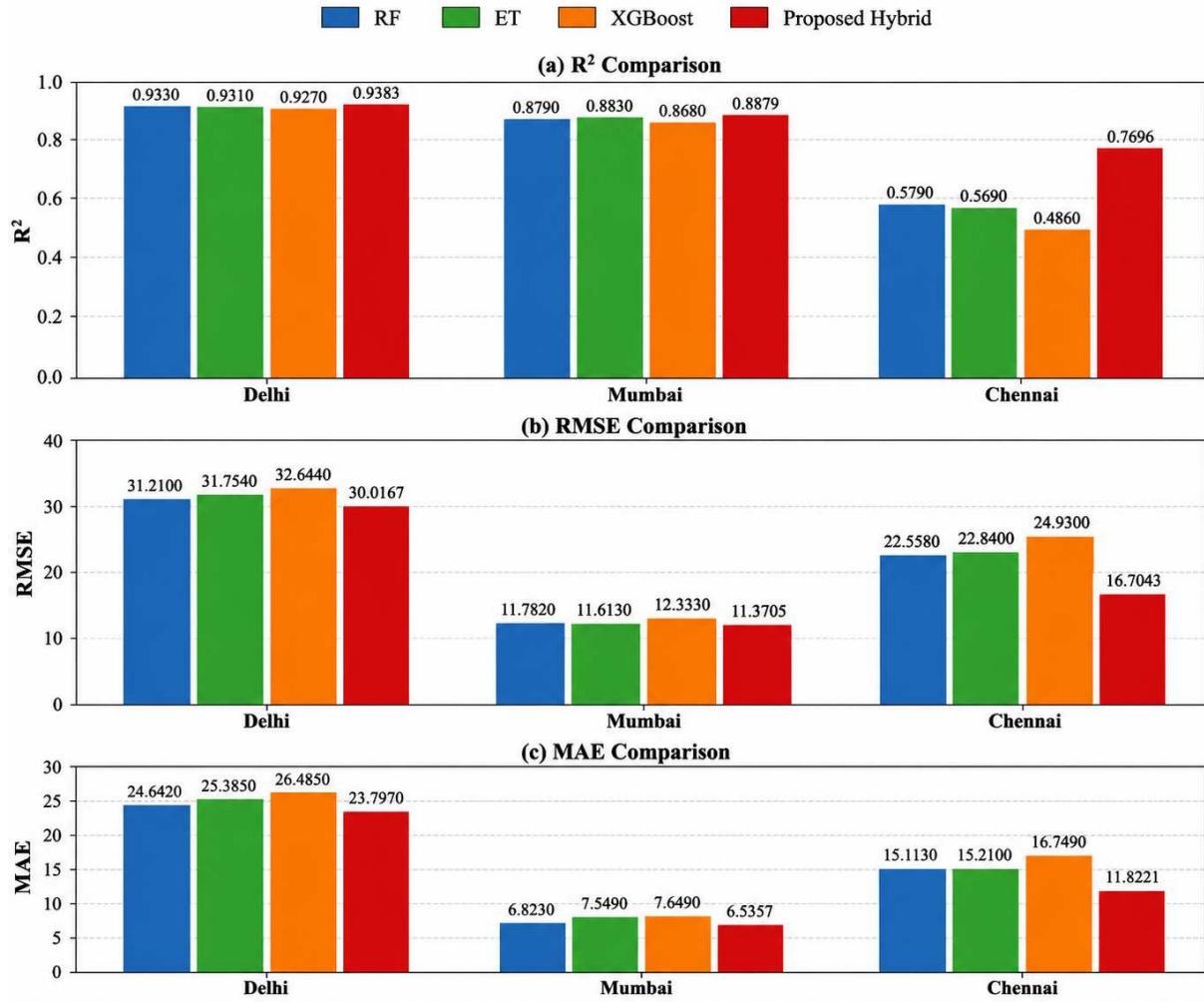


Figure 4: Comparative Performance of the Proposed Hybrid Stacking Framework and Individual Base Learners Across the Three Metropolitan Cities

observation. The loss function used for the DM test was squared error, and the long-run variance was estimated using a heteroskedasticity and autocorrelation-consistent procedure. Statistica significance was assessed at the 5% level. The uncertainty analysis was used to distinguish numerical differences in point estimates from differences supported by statistical evidence. Thus, higher R^2 or lower RMSE and MAE values were not interpreted as evidence of statistical superiority unless the corresponding forecast-error comparison provided supporting statistical evidence.

4.5. Performance Evaluation of the Proposed Hybrid Stacking Framework

As per the baseline, RF, ET, and XGBoost learners offer similar yet city-specific predictions, whereby no learner offers the best results in all the metropolises. Such complementary nature provides reason for their combination within the proposed stacking model. In that respect, CatBoost was used as a meta-learner with regard to the predictions of the learners, which were controlled for leakage in out-of-fold

predictions.

In turn, the proposed EHSF model has achieved R^2 of 0.9383, 0.8879 and 0.7696 for Delhi, Mumbai and Chennai respectively, with RMSE of 30.0167, 11.3705 and 16.7043 correspondingly. In result the evidence shows that stacking approach has resulted in better performance metrics compared with individual base learners on all studied metrological areas. This is especially true for Chennai, where the best performing individual model - Random Forests – has shown R^2 of just 0.579, while the EHSF has managed to achieve 0.7696. It is the significant improvement in terms of explained variance under the most difficult condition for prediction within city. For both other metrological environments, where individual base models have demonstrated comparatively high performance, the improvement has been achieved too: for Delhi and Mumbai R^2 values were 0.9383 and 0.8879 respectively. Thus, according to the obtained results, the superiority of the proposed EHSF solution cannot be linked only with the selection of the highest performing individual learning algorithm. The meta-learning

component leverages complementarity of error structure and non-linear representation generated by the heterogeneous ensemble of base algorithms. This is especially important for heterogenic urban environment, where pollutant lifetime, particulates interaction and dynamics differ significantly between specific locations. Hence, the improvement demonstrates the empirical justification of the basic principle of the suggested solution – consensus-optimization of predictor combined with heterogeneous ensemble learning improves forecasting performance across the three independently evaluated metropolitan environments. Thus, the predictive capabilities have demonstrated the effectiveness of the suggested model's architecture. However, the accuracy in itself does not prove how the suggested model makes its predictions or how the learned relationships are linked to the actual characteristics of the air quality. Hence the further analysis moves on the issue of explainability.

4.6. Interpretative Analysis Using SHAP Frameworks

SHAP analysis was used to interpret the EHSF predictions at both the predictor level and the meta-learning level. Predictor-level SHAP analysis was used to examine the contribution of the retained environmental, pollutant, temporal, lag, rolling, and interaction features. Meta-level SHAP analysis was used separately to examine the contribution of the RF, ET and XGBoost base-model predictions to the CatBoost meta-learner. These two levels of interpretation were considered separately to avoid conflating original predictors with model-derived meta-features. The predictive results above clearly show that our hybrid stacking framework is superior compared with other forecasting methods. However, these findings do not prove that these learned relationships are physically and statistically significant. For this reason, we need to use the Shapley Additive Explanations framework in order to estimate feature contribution to final EHSF predictions in three metropolitan environments under analysis. The interpretability of the proposed hybrid stacking framework was further examined using SHAP-based analyses. Three complementary visualizations were employed: global SHAP summary plots to identify the relative importance and directional contribution of the predictors, SHAP dependence plots to investigate the relationship between influential predictors and their contributions to the model output, and SHAP waterfall plots to explain the contribution of individual inputs to a specific prediction. These analyses provide both global and local interpretations of the forecasting behavior of the proposed framework.

4.6.1. Interpretative Analysis for Delhi

The SHAP framework finds out that Pollution_Load, Particulate_Index, PM10, PM10_RH and PM25 are the most valuable features that affect EHSF prediction in Delhi. The dominance of pollution-related variables is consistent with the consensus feature ranking and shows the correspondence between the two stages of analysis.

Figure 5 represents the global and local SHAP values for the city of Delhi, where Pollution_Load, Particulate_Index, PM10, PM10_RH, and PM25 are found to be important predictors. Through the dependency plot, the nonlinear effect of Pollution_Load on prediction is observed. In addition, the water fall plot shows how the features influence the particular prediction pointwise.

4.6.2. Interpretative Analysis for Mumbai

The SHAP analysis identified PM10, Pollution_Load, PM10_Lag_1, and PM25 among the influential predictors in the Mumbai forecasting model. These variables indicate the importance of current pollutant conditions and short-term temporal persistence. Separately, the meta-level analysis examined the contribution of the XGBoost, Random Forest, and Extra Trees prediction outputs to the CatBoost meta-learner. It is evident from the prevalence of PM10 and PM10_Lag_1 that current pollution levels and temporal continuity play an essential role in the prediction of concentrations. From the dependency analysis, it is evident that there exists a non-linear interaction between PM10 and its SHAP value, where Pollution_Load acts as the mediator of this interaction.

In figure 6, the SHAP value-based local and global interpretation for Mumbai includes PM10, Pollution_Load, PM10_Lag_1, XGB, and PM25. The dependency pattern between PM10 shows that its value increases non-linearly with SHAP value, whereas the factor Pollution_Load impacts on this, showing the interaction between pollutant quantity and pollution load in the context of the EHSF model.

4.6.3. Interpretative Analysis for Chennai

According to the consensus-based feature-selection results presented in Table 6, Particulate_Index, PM10, AQI_Lag_1, Pollution_Load, AQI_Min_3, and PM25 form the top predictors for Chennai city based on the information provided by Mutual Information, Random Forest importance, Permutation Importance, and SHAP importance metrics. After training the EHSF model, the interpretation of feature importance via SHAP values can be done in order to obtain the contribution of the retained predictors and model-derived features in individual forecasts.

Fig. 7 depicts the global and local SHAP interpretation of the forecasting results for Chennai. Pollution_Load, AQI_Min_3, The predictor-level SHAP analysis identified Particulate_Index, PM10, AQI_Lag_1, Pollution_Load, AQI_Min_3 and PM25 among the influential retained predictors for Chennai. These predictors capture current pollutant conditions, temporal persistence, and short-term air-quality variation. At the meta-learning level, the contributions of the RF, ET, and XGBoost base-model predictions were examined separately, AQI_Lag_1, and Particulate_Index turn out to be important predictors according

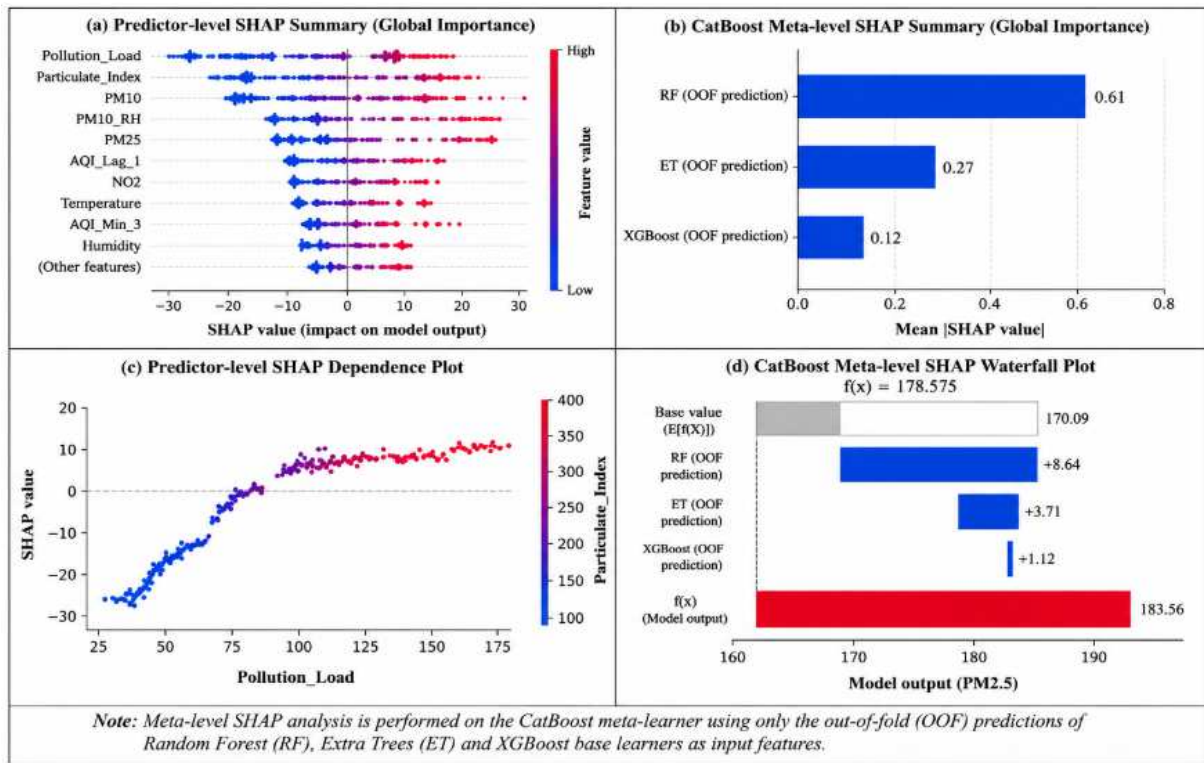


Figure 5: SHAP-Based Explainability Analysis of the Proposed Hybrid Stacking Framework for Delhi

to all presented plots. Moreover, according to the dependence analysis, *Pollution_Load* demonstrates a nonlinear dependency from SHAP value due to the impact of previous air-quality conditions expressed in *AQI_Min_3* variable. Local waterfall plot further clarifies how certain predictors change the forecast from baseline up to the final prediction. In terms of methodology, there is no contradiction between the results of Table 6 and Fig. 7 since the former presents predictors which received consistent support in terms of feature selection, whereas the latter gives explanation to the contribution of certain predictors in fully trained EHSF model. Similarity in the set of features including *Particulate_Index*, *Pollution_Load*, *AQI_Lag_1*, and *AQI_Min_3* proves the robustness of the whole analysis procedure; meanwhile additional variables in Fig. 7 can be considered as a result of learning stacking structure.

4.7. Multi-City Interpretative Analysis of Feature Contribution

Despite different cities under analysis, *Pollution_Load*, *Particulate_Index*, particulate concentrations, lagged predictors, and rolling-statistical features recurrently emerge among the influential predictor groups across the three metropolitan environments. Their repeated importance proves that feature engineering framework succeeds to capture nonlinearities and time dependencies in the datasets. However, despite the agreement of results, these influential features make different relative contributions depending on specific city. Thus, our EHSF framework is able to

capture not only common mechanisms of prediction but also location-dependent pollution dynamics. It can be observed from Delhi where the prediction mostly depends on *Pollution_Load* and particulate-related features. In the case of Mumbai, the prediction is mostly based on PM10 concentrations and lagged PM10 features. Finally, the contribution of AQI-based statistical features and *Pollution_Load* is prominent in Chennai. Thus, a single fixed set of influential features cannot reflect complex metro-level dynamics.

5. DISCUSSION

The empirical experiments show that the suggested EHSF model is capable of combining the forecast-oriented knowledge synthesis, consensus-based learning of predictors, heterogeneity of the ensemble learning approach, and post-learning model explainability in a consistent forecasting process pipeline. Further analysis should thus focus not on the presentation of the predictive results but on understanding the underlying reasons for the differences observed in the performance of the model in different metropolises, the role played by the combined use of complimentary models, and the possible implications for transferable air quality prediction.

5.1. Scientific Implications of Forecast-Aware Feature Representation

The first methodological implication relates to the representation of information available at the time of forecasting. Traditional feature construction often leads to a

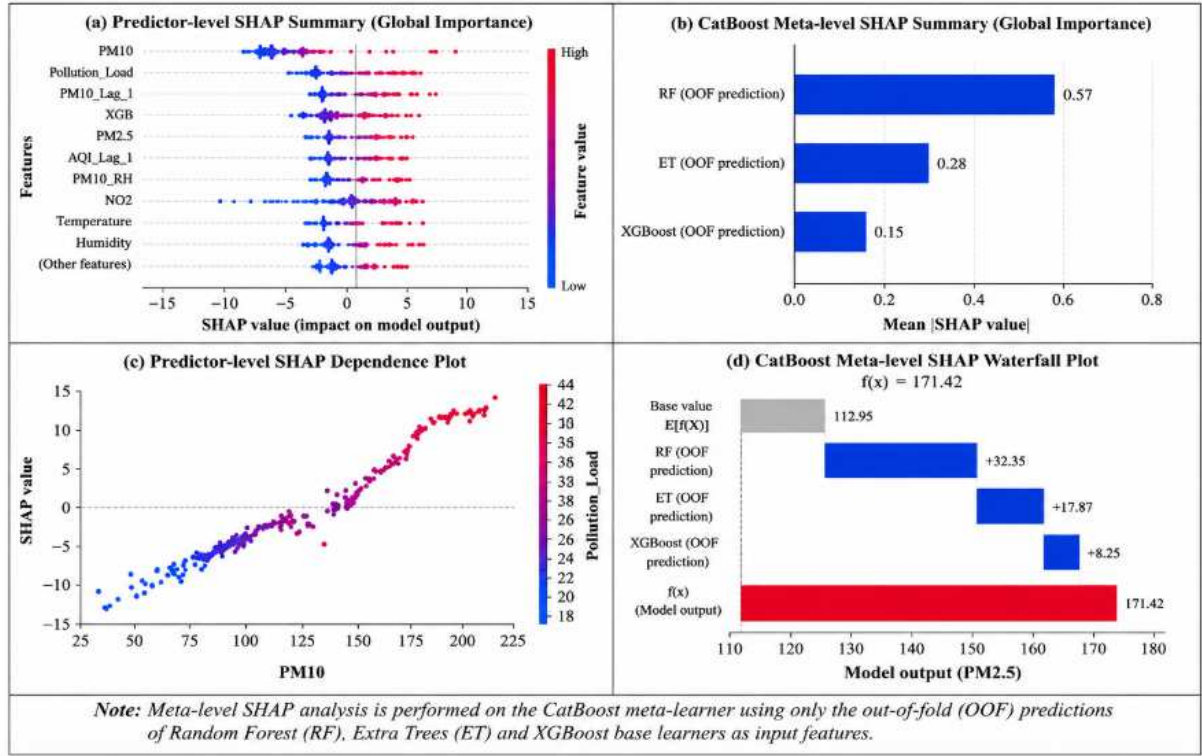


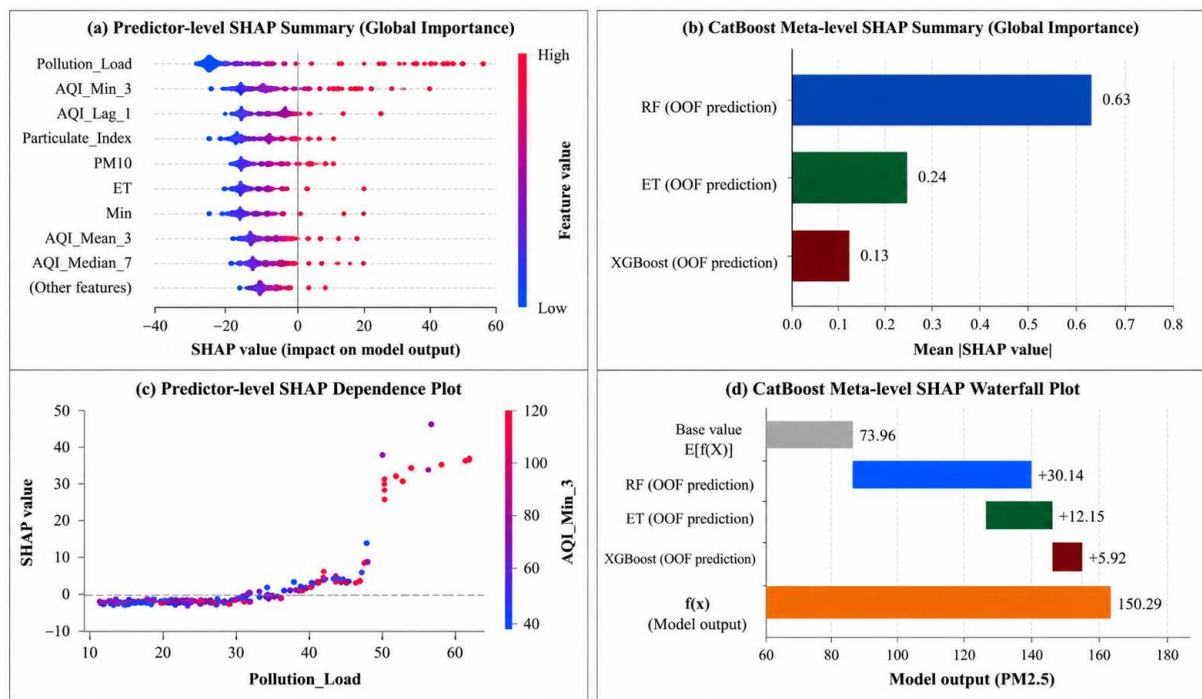
Figure 6: SHAP-based Explainability Analysis of the Proposed Hybrid Stacking Framework for Mumbai.

large number of temporally correlated variables, increasing redundancy and confusing predictors available at forecast time and predictors available only after the prediction horizon. The proposed forecast-aware approach mitigates this problem through a structured representation of the predictor space in terms of temporality, cycles, lags, rolling statistics and domain-specific variables. Reducing predictor spaces from 204 to 173 features for Delhi, 205 to 159 features for Mumbai and 204 to 178 features for Chennai means there are still substantial redundancies even after structured feature construction. However, most importantly, the consensus-selection procedure does not rely on any specific feature importance measure. Mutual Information measures non-linear dependence, Random Forest importance quantifies model-dependent feature contribution, Permutation Importance measures the effect of predictor perturbation, while SHAP represents additive attribution. Therefore, the combination of these methods represents a more robust procedure for feature selection compared to any one ranking procedure alone. The recurring presence of particulate and temporal features across these measures suggests that predicting requires a representation of pollutant magnitude, persistence and nonlinear interactions. Thus, the scientific significance is not merely in reduction in the number of predictors. Instead, the proposed procedure yields a compact and forecast-compatible representation of urban air quality dynamics. This is an important scientific implication from an operational forecasting perspective, as the feature construction

process should preserve predictive power without temporal inconsistency or redundancies.

5.2. Why Heterogeneous Stacking Leads to Greater Robustness

The results concerning performance of the base-learners provide an essential explanation for the importance of the proposed stacking architecture. Neither individual base learner dominates among the three cities – while Random Forest works best for Delhi, Extra Trees works better for Mumbai, and all three models yield substantially lower performance in Chennai. This demonstrates the lack of a single learning problem – the problem does not rely upon a single, universally optimal, nonlinear mapping. The complementary behaviour of the three base-learners provides the principal motivation for their combination. While both Random Forest and Extra Trees represent ensembles of decision trees trained in randomized fashion, XGBoost uses gradient boosting. This means that the different learning strategies give the meta-learner heterogeneous representations of the same optimized predictor space. CatBoost finally learns on the out-of-fold predictions from these base-learners, and not the performance of the best performing individual learner. This design of the architecture allows learning combinations of models' predictions which would be hard or impossible to learn using any one learner. This explains methodologically how the proposed framework enhances the prediction results for all three cities, rather than simply replicating the performance of the strongest individual base learner. The relatively complex environment



Note: Meta-level SHAP analysis is performed on the CatBoost meta-learner using only the out-of-fold (OOB) predictions of Random Forest (RF), Extra Trees (ET) and XGBoost base learners as input features.

Figure 7: SHAP-based Explainability Analysis of the Proposed Hybrid Stacking Framework for Chennai.

in Chennai is especially critical in this regard. While the relatively poorer performance of the individual base learners demonstrates increased difficulties in representing predictive relationship, the stacking framework substantially improves this forecast. This suggests that heterogeneity in model predictions increases their relative value in a heterogeneous environment.

5.3. Temporal Consistency of Features and Predictions

A second major implication relates to temporal consistency of features and predictions. A good forecast model is useless if it depends on information not available at the moment when a forecast is created. This is the reason why the proposed framework treats the temporal dimension as a central part of the whole forecasting process, and not an independently considered preprocessing step. Lagged and rolling predictors provide information about the history of the pollutant behaviour without reversing its chronological ordering. Time-series cross-validation, in its turn, allows avoiding inappropriate validation of forecasts made using random mixes of information from different times of year. Moreover, out-of-fold predictions prevent the CatBoost meta-learner from direct learning from predictions based on the same information as that which was used to create these predictors. The importance of this design becomes especially apparent for operational forecasting tasks. The use of such a design makes the presented methodology methodologically credible in the sense that its design takes

into account only the information actually available during the actual forecasting process, and not the complete dataset retrospectively.

5.4. City-Specific Predictors and Multi-City Heterogeneity

The results demonstrate that the same framework can work with three metropolitan environments and yield different predictor contributions. This, however, does not mean that the pollution dynamics are the same across all three cities. Rather, it means that while this framework provides a common forecasting platform, predictor dependencies can adapt to local environmental factors in each metropolitan environment. This is seen in the fact that Delhi shows strong contributions of pollutants' load and particulate predictors, Mumbai relies more on PM10 and temporal predictors, and Chennai shows higher contributions of pollutant load and AQI predictors. The important aspect here is thus the ability to adapt predictor representation of the pollutant dynamics, not its homogeneity. This is important for multi-city forecasting applications. A transferable framework should provide a common methodology, but assume that predictor contribution is environment-specific. The EHSF framework fulfills this requirement, using the combination of a common platform with specific feature representation and model learning.

5.5. SHAP, Explainability and Decision Support

The results concerning SHAP provide an important expansion of this study beyond the simple predictive performance. While SHAP features coincide with consensus-identified predictors, this is useful in itself as it shows that the selected predictors stay influential even after full learning by the EHSF framework. However, SHAP differences from the consensus ranking of features can also help interpret what makes a single prediction different from others. This becomes especially important from the perspective of environmental decision-support system. Global predictor ranking can help find out which predictors are generally important for the EHSF predictive model. Local SHAP explanation, in its turn, allows identifying which factors contribute a specific increase or reduction of the predicted value. Consequently, the proposed forecasting framework combines accurate non-linear predictions with interpretability. This becomes especially significant in cases when predictions will be used to issue warnings and guide the monitoring activities and urban air quality planning.

5.6. Practical Implications

Operational implications of the proposed approach are clear from the results. EHSF framework can be used to create multi-city air quality monitoring system, based on combination of the temporal information and heterogeneous machine learning. The environment-specific predictor contributions obtained during the learning can then help explain why the forecast differs between the metropolitan cities. However, the framework should be used for forecasting and decision support purposes, rather than pollution source attribution. Even though SHAP allows identifying features associated with a single predictor of the EHSF, changing their values does not necessary mean that the pollution concentration in the atmosphere would change as well.

5.7. Limitations and Future Work

Several limitations set the boundaries of current results. First, the validation is conducted on three metropolitan cities of India and limited by the 2021–2025 CPCB period. Therefore, the results show multi-city applicability, but not universal transferability across Indian metropolitan cities and/or atmospheric conditions. Second, the current predictors are mainly based on CPCB air quality and meteorological observations. Adding the predictors derived from satellite observations of aerosols, traffic intensity and other land-use data can provide more insight on pollution dynamics and urban activities. Third, the future work should involve independent cities and their monitoring networks to estimate transferability. Finally, external temporal validation using another period of observations can give additional support to distribution shift robustness of the methodology. Future research could also focus on designing adaptive and/or transfer-learning procedures, during which common EHSF framework remains, but city-specific components are recalculated. This will provide a direct testing of the methodology for transferability under different pollution regimes. Conclusively, the results demonstrate that the key advantage

of the proposed EHSF framework is in its ability to integrate forecast-compatible feature construction, multi-criteria predictor selection, heterogeneous learning, stack without information leakage and explainability. The findings thus provide a unifying method for accurate and interpretable multi-city forecasting while explicitly stating the limitations of its current geographical validation.

6. CONCLUSION

This study developed an Explainable Hybrid Stacking Framework (EHSF) for accurate and interpretable one-hour-ahead $PM_{2.5}$ forecasting at three CPCB monitoring stations in India: Anand Vihar (Delhi), Chhatrapati Shivaji International Airport T2 (Mumbai), and Manali (Chennai). Hourly observations from January 2021 to December 2025 were transformed into temporal, cyclic, lag-based, rolling-statistical, pollutant, and domain-specific interaction features. Consensus-based feature selection was used to eliminate redundant variables and retain informative predictors identified according to complementary feature-importance criteria. The proposed heterogeneous stacking framework used Random Forest, Extra Trees, and XGBoost algorithms as base learners and CatBoost as the meta-learner.

On the held-out chronological test sets, EHSF achieved R^2 values of 0.9383, 0.8879, and 0.7696 and RMSE values of 30.0167, 11.3705, and 16.7043 for Delhi, Mumbai, and Chennai, respectively. The corresponding MAE values were 23.7970, 6.5357, and 11.8221. The results show that a combination of multiple heterogeneous base learners can lead to better predictive performance than any single base learner in the evaluated cities. In particular, EHSF substantially improved forecasting performance in Chennai, where the best individual base learner (Random Forest) achieved an R^2 of only 0.5790 and an RMSE of 22.5580, while EHSF reached an R^2 of 0.7696 and an RMSE of 16.7043. For Delhi and Mumbai, where the individual base learners already exhibited strong performance, EHSF provided additional but more modest improvements in both R^2 and error metrics.

SHAP-based interpretability allowed determining the relative contribution of important pollutants, temporal, lagged, rolling-statistical, and domain-specific predictors. Among the influential predictors, Pollution_Load, Particulate_Index, PM_{10} , $PM_{2.5}$, lagged predictors, and a set of statistical predictors derived from AQI were found; however, their relative importance differed between the three cities studied. In Delhi, Pollution_Load and particulate-related features dominated the SHAP rankings. In Mumbai, PM_{10} and lagged PM_{10} features were most influential. In Chennai, AQI-based statistical features and Pollution_Load contributed most strongly. At the meta-learner level, SHAP analysis revealed how the predictions generated by each base learner contributed to the final EHSF predictions, confirming that the stacking architecture leverages complementary information from Random Forest, Extra Trees, and XGBoost.

Overall, the proposed EHSF framework combines forecasting accuracy, temporal information preservation, feature-importance determination through consensus, heterogeneous ensemble learning, and model interpretability. The findings demonstrate the applicability of the proposed framework in the three evaluated metropolitan environments. However, the study is limited by its evaluation on three monitoring locations and by the availability of pollutants and other environmental variables in the CPCB dataset. Consequently, the results support the usefulness of EHSF for city-specific PM_{2.5} forecasting but do not establish geographical generalization across all Indian urban areas.

Future work could involve evaluating the framework on additional cities and climatic settings, incorporating extra variables such as satellite-derived aerosol optical depth and traffic-intensity indicators, and testing transfer-learning procedures in which the common EHSF framework is retained but city-specific components are recalibrated. Further research could also extend the current uncertainty analysis by reporting bootstrap confidence intervals for RMSE and MAE and by applying paired forecast-error tests such as the Diebold–Mariano test to assess the statistical significance of the observed performance differences. Such extensions would strengthen the evidence base for deploying EHSF in operational air-quality forecasting and decision-support systems.

Declarations

Funding

The authors declare that no specific funding was received from any funding agency in the public, commercial, or not-for-profit sectors for conducting this research.

Conflict of Interest

The authors declare that they have no known competing financial or non-financial interests that could have appeared to influence the work reported in this paper.

Data Availability

The datasets used and analysed during the current study were obtained from the Central Pollution Control Board (CPCB), Government of India. The processed datasets generated during this study are available from the corresponding author upon reasonable request.

Code Availability

The source code, implementation of the proposed Hybrid Stacking Framework, and the experimental scripts developed during this study are available from the corresponding author upon reasonable request.

Acknowledgements

The authors express their sincere gratitude to the Central Pollution Control Board (CPCB), Government of India, for providing access to the air-quality datasets used in this research. The authors also acknowledge Google Colaboratory (Google Colab) for providing the cloud-based computational

environment that facilitated the implementation, experimentation, and evaluation of the proposed machine learning framework.

References

- [1] Li, S., Wang, Y., Zhang, X., Wu, L., Liu, X., & Tursun, E. (2026). An explainable stacked ensemble learning for street-level multi-pollutant prediction. *Atmospheric Environment*, 382, 122237. <https://doi.org/10.1016/j.atmosenv.2026.122237>
- [2] Dimri, R., Choi, Y., Singh, D., Park, J., & Khorshidian, N. (2026). TransNet: A transport-informed graph neural network for forecasting concentrations across South Korea. *npj Clean Air*, 2, 12. <https://doi.org/10.1038/s44407-026-00052-x>
- [3] Wang, P., Wu, Q., & Zhang, G. (2026). A hybrid prediction model for concentration based on high-frequency and low-frequency IMFs with EMD decomposition. *Scientific Reports*, 16, 4969. <https://doi.org/10.1038/s41598-026-35386-9>
- [4] Gupta, V., Dash, Y., Goyal, A., Lodh, A., Singh, V., Routray, A., & Kumar, P. (2026). A BiLSTM-driven framework for operational forecasting: Integrating meteorological kinematics for urban air quality management. *Frontiers in Climate*, 8, 1855755. <https://doi.org/10.3389/fclim.2026.1855755>
- [5] Kumari, J., Wadhvani, R., Shukla, S., & Kumar, L. (2026). Regime-aware hybrid ensemble learning for adaptive forecasting in urban environments. *Scientific Reports*, 16, 16852. <https://doi.org/10.1038/s41598-026-45437-w>
- [6] Singh, S., Kumar, M., Sengar, V., Kumar, A., Abhishek, K., & Shafeeq, B. M. A. (2026). Ensemble learning for air quality index prediction: Integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability. *Scientific Reports*, 16, 8544. <https://doi.org/10.1038/s41598-026-39232-w>
- [7] Rautela, K. S., Upadhyay, M., Goyal, M. K., & Taigor, S. (2026). Artificial intelligence–driven air quality mapping and forecasting for extreme PM_{2.5} events. *Physics and Chemistry of the Earth*, 144, 104459. <https://doi.org/10.1016/j.pce.2026.104459>
- [8] Charjan, O., Gajbhiye, K., Warhade, J., & Wankhade, S. (2026). A feature engineering–driven ensemble approach for accurate AQI forecasting. *Discover Applied Sciences*, 8, 216. <https://doi.org/10.1007/s42452-025-08165-w>
- [9] Ke, H., Gong, S., Wang, Y., Zhang, L., Mo, J., Yang, J., Wang, C., & Zhou, C. (2026). Separation of meteorological and anthropogenic contributions to major ambient pollutants using an integrated explainable machine learning framework: RMT-PAS. *Atmospheric Environment*, 365, 121667. <https://doi.org/10.1016/j.atmosenv.2025.121667>
- [10] Wang, X., Li, T., Ma, M., Bai, X., Sun, W., & Zhu, R. (2026). Explainable machine learning reveals the effect of sources and meteorology on pollution in a typical city of Southern Sichuan, China. *Atmospheric Environment*, 372, 121881. <https://doi.org/10.1016/j.atmosenv.2026.121881>
- [11] Malakouti, S. M. (2025). From accurate to actionable: Interpretable PM_{2.5} forecasting with feature engineering and SHAP for the Liverpool–Wirral region. *Environmental Challenges*, 21, 101290. <https://doi.org/10.1016/j.envc.2025.101290>
- [12] Wang, W. (2025). High-accuracy prediction via mutual information filtering and Bayesian-optimized spatio-temporal

- convolutional networks. *Scientific Reports*, 15, 21718. <https://doi.org/10.1038/s41598-025-08896-1>
- [13] Xu, Z., Zhang, H., Zhai, A., Kong, C., & Zhang, J. (2025). Stacking ensemble learning and SHAP-based insights for urban air quality forecasting: Evidence from Shenyang and global implications. *Atmosphere*, 16(7), 776. <https://doi.org/10.3390/atmos16070776>
- [14] Patel, R., Kumar, A., Yadav, J., & Singh, M. (2025). Stacked deep learning ensemble for time series prediction of levels in Bihar. *Urban Climate*, 62, 102521. <https://doi.org/10.1016/j.uclim.2025.102521>
- [15] Zhang, Z., Xu, J., Zhao, X., & Cheng, S. (2025). Improving and visibility predictions in China using machine learning and ensemble forecasting. *Journal of Geophysical Research: Machine Learning and Computation*, 2(2), e2025JH000640. <https://doi.org/10.1029/2025JH000640>
- [16] Chen, G., Chen, S., Li, D., & Chen, C. (2025). A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention. *Scientific Reports*, 15, 3685. <https://doi.org/10.1038/s41598-025-88086-1>
- [17] Gao, X., Wang, X., Li, F., Jiang, W., Zhe, M., Sun, J., Zhang, A., & Jiao, L. (2025). concentration 7-day prediction in the Beijing–Tianjin–Hebei region using a novel stacking framework. *Scientific Reports*, 15, 20731. <https://doi.org/10.1038/s41598-025-07719-7>
- [18] Ravindiran, G., Karthick, K., Rajamanickam, S., Datta, D., Das, B., Shyamala, G., Hayder, G., & Maria, A. (2025). Ensemble stacking of machine learning models for air quality prediction for Hyderabad city in India. *iScience*, 28(2), 111894. <https://doi.org/10.1016/j.isci.2025.111894>
- [19] Inam, S. A., Rajput, H., & Umer, S. (2025). An explainable hybrid stacked deep learning framework for forecasting PM concentrations in urban air. *Explora: Environment and Resource*, 2(4), 025380069. <https://doi.org/10.36922/EER025380069>
- [20] Yang, J., Ke, H., Gong, S., Wang, Y., Zhang, L., Zhou, C., Mo, J., & You, Y. (2025). Enhanced forecasting and assessment of urban air quality by an automated machine learning system: The AI-Air. *Earth and Space Science*, 12(1), e2024EA003942. <https://doi.org/10.1029/2024EA003942>
- [21] Singha, S., & Srivastava, A. (2025). Machine learning approach to forecasting and health risk assessment during stubble burning period in Delhi. *Aerosol Science and Technology*, 59(11), 1385–1404. <https://doi.org/10.1080/02786826.2025.2513503>
- [22] Wang, K., Qu, C., Wang, J., Li, Z., & Lu, H. (2025). Ensemble multitask prediction of air pollutants time series: Based on variational inference, data projection, and generative adversarial network. *Journal of Forecasting*, 44(2), 646–675. <https://doi.org/10.1002/for.3218>
- [23] Natarajan, S. K., Shanmuthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*, 14, 6795. <https://doi.org/10.1038/s41598-024-54807-1>
- [24] Liu, X., Pu, X., Lu, C., Zhang, H., Li, T., Grieneisen, M. L., Li, J., Ma, N., Yan, C., Zhan, Y., & Yang, F. (2024). Interpretable regional meteorological feature extraction enhances deep learning for extended 120-h PM2.5 forecasting. *Journal of Cleaner Production*, 483, 144287. <https://doi.org/10.1016/j.jclepro.2024.144287>
- [25] Wang, Y., Tian, S., & Zhang, P. (2024). Novel spatio-temporal attention causal convolutional neural network for multi-site prediction. *Frontiers in Environmental Science*, 12, 1408370. <https://doi.org/10.3389/fenvs.2024.1408370>
- [26] Rakholia, R., Le, Q., Vu, K., Ho, B. Q., & Carbajo, R. S. (2024). Accurate urban air pollution forecasting using multivariate ensemble learning accounting for evolving target distributions. *Chemosphere*, 364, 143097. <https://doi.org/10.1016/j.chemosphere.2024.143097>
- [27] Houdou, A., El Badisy, I., Khomsi, K., Abdala, S. A., Abdulla, F., Najmi, H., Obtel, M., Belyamani, L., Ibrahim, A., & Khalis, M. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. *Aerosol and Air Quality Research*, 24(1), 230151. <https://doi.org/10.4209/aaqr.230151>
- [28] Iftikhar, H., Qureshi, M., Zywolek, J., López-Gonzales, J. L., & Albalawi, O. (2024). Short-term forecasting using a unique ensemble technique for proactive environmental management initiatives. *Frontiers in Environmental Science*, 12, 1442644. <https://doi.org/10.3389/fenvs.2024.1442644>
- [29] Nguyen, A. T., Pham, D. H., Oo, B. L., Ahn, Y., & Lim, B. T. H. (2024). Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. *Journal of Big Data*, 11, 71. <https://doi.org/10.1186/s40537-024-00926-5>
- [30] Dhandapani, A., Iqbal, J., & Kumar, R. N. (2023). Application of machine learning (individual vs stacking) models on MERRA-2 data to predict surface PM2.5 concentrations over India. *Chemosphere*, 340, 139966. <https://doi.org/10.1016/j.chemosphere.2023.139966>
- [31] Shakya, D., Deshpande, V., Goyal, M. K., & Agarwal, M. (2023). air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi, India. *Journal of Cleaner Production*, 427, 139278. <https://doi.org/10.1016/j.jclepro.2023.139278>
- [32] Fan, K., Dhammapala, R., Harrington, K., Lamb, B., & Lee, Y. (2023). Machine learning-based ozone and forecasting: Application to multiple AQS sites in the Pacific Northwest. *Frontiers in Big Data*, 6, 1124148. <https://doi.org/10.3389/fdata.2023.1124148>
- [33] Wu, K.-Y., Hsia, I.-W., Kow, P.-Y., Chang, L.-C., & Chang, F.-J. (2023). High-spatiotemporal-resolution forecasting by hybrid deep learning models with ensembled massive heterogeneous monitoring data. *Journal of Cleaner Production*, 433, 139825. <https://doi.org/10.1016/j.jclepro.2023.139825>
- [34] Gokul, P. R., Mathew, A., Bhosale, A., & Nair, A. T. (2023). Spatio-temporal air quality analysis and prediction over Hyderabad City, India using artificial intelligence techniques. *Ecological Informatics*, 76, 102067. <https://doi.org/10.1016/j.ecoinf.2023.102067>
- [35] Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam. *Chemosphere*, 338, 139518. <https://doi.org/10.1016/j.chemosphere.2023.139518>
- [36] Mathew, A., Gokul, P. R., Shekar, P. R., Arunab, K. S., Abdo, H. G., Almohamad, H., & Al Dughairi, A. A. (2023). Air quality analysis and modelling using machine learning techniques: A study of Hyderabad city in India. *Cogent Engineering*, 10(1), 2243743. <https://doi.org/10.1080/23311916.2023.2243743>
- [37] Zhang, Z., & Xia, D. (2023). An improved forecasting method based on correlation denoising and ensemble learning strategy.

- [38] Qu, Z., Sun, S., Wang, S., & Niu, C. (2023). Uncertainty analysis–forecasting system based on decomposition–ensemble framework for concentration forecasting in China. *Journal of Forecasting*. <https://doi.org/10.1002/for.3005>
- [39] Saminathan, S., & Malathy, C. (2023). Ensemble-based classification approach for concentration forecasting using meteorological data. *Frontiers in Big Data*, 6, 1175259. <https://doi.org/10.3389/fdata.2023.1175259>
- [40] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: A case study of Indian cities. *International Journal of Environmental Science and Technology*, 20(5), 5333–5348. <https://doi.org/10.1007/s13762-022-04241-5>
- [41] Nandi, B. P., Singh, G., Jain, A., & Tayal, D. K. (2023). Evolution of neural network to deep learning in prediction of air, water pollution and its Indian context. *International Journal of Environmental Science and Technology*. <https://doi.org/10.1007/s13762-023-04911-y>
- [42] Liu, C., Rao, X., Chen, Q., Zhang, B., & Zhang, B. (2023). Objective identification and forecast method of pollution based on medium- and long-term ensemble forecasts in Beijing-Tianjin-Hebei region and its surrounding areas. *Frontiers in Earth Science*, 10, 1022565. <https://doi.org/10.3389/feart.2022.1022565>
- [43] Li, Z., Gan, K., Sun, S., & Wang, S. (2023). A new concentration forecasting system based on AdaBoost-ensemble system with deep learning approach. *Journal of Forecasting*, 42(1), 154–175. <https://doi.org/10.1002/for.2883>
- [44] Vignesh, et al. (2023). Predicting concentrations across USA using machine learning. *Earth and Space Science*. <https://doi.org/10.1029/2023EA002911>
- [45] Yu, W., Li, S., Ye, T., Xu, R., Song, J., & Guo, Y. (2022). Deep ensemble machine learning framework for the estimation of concentrations. *Environmental Health Perspectives*, 130(3), 037004. <https://doi.org/10.1289/EHP9752>
- [46] Wang, J., Wang, R., & Li, Z. (2022). A combined forecasting system based on multi-objective optimization and feature extraction strategy for hourly concentration. *Applied Soft Computing*, 114, 108034. <https://doi.org/10.1016/j.asoc.2021.108034>
- [47] Dong, L., Hua, P., Gui, D., & Zhang, J. (2022). Extraction of multi-scale features enhances the deep learning-based daily forecasting in cities. *Chemosphere*, 308(2), 136252. <https://doi.org/10.1016/j.chemosphere.2022.136252>
- [48] Yu, Y., Li, H., Sun, S., & Li, Y. (2022). concentration forecasting through a novel multi-scale ensemble learning approach considering intercity synergy. *Sustainable Cities and Society*, 81, 104049. <https://doi.org/10.1016/j.scs.2022.104049>
- [49] Faraji, M., Nadi, S., Ghaffarpasand, O., Homayoni, S., & Downey, K. (2022). An integrated 3D CNN-GRU deep learning method for short-term prediction of concentration in urban environment. *Science of the Total Environment*, 834, 155324. <https://doi.org/10.1016/j.scitotenv.2022.155324>
- [50] Saheer, L. B., Bhasy, A., Maktabdar, M., & Zarrin, J. (2022). Data-driven framework for understanding and predicting air quality in urban areas. *Frontiers in Big Data*, 5, 822573. <https://doi.org/10.3389/fdata.2022.822573>
- [51] Kumar, P. V., Kumar, M. C. A., Kumar, B. A., & Rao, P. V. (2022). Prediction of PM2.5 over Hyderabad using deep learning technique. *Nature Environment and Pollution Technology*, 21(2), 691–696. <https://doi.org/10.46488/NEPT.2022.v21i02.029>
- [52] Barthwal, A., Acharya, D., & Lohani, D. (2021). Prediction and analysis of particulate matter (and PM) concentrations using machine learning techniques. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-021-03051-w>
- [53] Liu, X., Qin, M., He, Y., Mi, X., & Yu, C. (2021). A new multi-data-driven spatiotemporal forecasting model based on an ensemble graph reinforcement learning convolutional network. *Atmospheric Pollution Research*, 12(10), 101197. <https://doi.org/10.1016/j.apr.2021.101197>
- [54] Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, 67, 102720. <https://doi.org/10.1016/j.scs.2021.102720>
- [55] Huang, Y., Deng, Y., Wang, C., & Fu, T. (2021). Hybrid data mining forecasting system based on multi-objective optimization and selection model for air pollutants. *Frontiers in Environmental Science*, 9, 761287. <https://doi.org/10.3389/fenvs.2021.761287>
- [56] Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of the Total Environment*, 699, 133561. <https://doi.org/10.1016/j.scitotenv.2019.07.367>
- [57] Masood, A., & Ahmad, K. (2020). A model for particulate matter () prediction for Delhi based on machine learning approaches. *Procedia Computer Science*, 167, 2101–2110. <https://doi.org/10.1016/j.procs.2020.03.258>
- [58] Liu, D.-R., Lee, S.-J., Huang, Y., & Chiu, C.-J. (2020). Air pollution forecasting based on attention-based LSTM neural network and ensemble learning. *Expert Systems*, 37(3), e12511. <https://doi.org/10.1111/exsy.12511>
- [59] Xu, X. (2020). Forecasting air pollution in Beijing using weather data and multiple kernel learning. *Journal of Forecasting*, 39(2), 117–125. <https://doi.org/10.1002/for.2599>