



A Hybrid Machine Learning and Deep Learning Framework for Real-Time Air Quality Prediction in Indian Cities: Multi-City Empirical Validation, Ablation Analysis, and Governance Assessment

Renuka Malge¹, Suma Swamy², S N Sheshappa³

¹Research Scholar, Sir M Visvesvaraya Institute of Technology, Bengaluru, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, Email: renuka.malge@gmail.com, Orchid ID: 0000-0001-9274-8523

²Professor & Dean R&D, HNNCE, Bengaluru, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, Email ID: sumaswamy.drhnncay202526@gmail.com, Orchid ID: 0000-0002-6207-5898

³Professor, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru affiliated to Visvesvaraya Technological University, Belagavi, karnataka, India, Email ID: sheshappa_sn@rediff.com

ABSTRACT

Air pollution remains a critical public health and policy challenge in India, where several metropolitan regions rank among the world's most polluted. This paper presents a hybrid Machine Learning / Deep Learning (ML/DL) framework for short-horizon PM_{2.5} forecasting that combines gradient-boosted tree models (XGBoost, LightGBM) with an attention-augmented Long Short-Term Memory (Attention-LSTM) network via a validation-loss-weighted ensemble. Unlike prior single-city or simulated evaluations, the framework is empirically validated end-to-end on real, publicly available Central Pollution Control Board (CPCB)-derived daily air-quality records (2018–2024) for four Indian metropolitan cities Delhi, Hyderabad, Mumbai, and Chennai comprising 8,257 station-days in total. All results reported here, including RMSE, MAE, MAPE, R^2 , an ablation study, and Diebold–Mariano statistical-significance tests, are computed from actual model training and testing runs rather than illustrative figures. The hybrid ensemble achieves the lowest or near-lowest 1-day-ahead RMSE in three of four cities and reduces error relative to the weakest baseline by 5–15%, while offering materially better robustness at longer (7-day) horizons than the standalone deep model. Permutation-importance analysis confirms that recent PM_{2.5} history, rolling statistics, and seasonal encodings dominate predictive attribution, consistent with atmospheric persistence and diurnal traffic-emission physics. A condensed regulatory-compliance discussion (GDPR, EU AI Act, NIST AI RMF, CPCB/NAAQS) is retained for completeness, while the bulk of the paper is devoted to the experimental protocol, dataset documentation, and reproducibility artifacts. Complete code, trained-model configurations, and dataset provenance are released to enable independent replication.

Keywords: Attention mechanism, Hybrid machine learning, Deep learning, LSTM, Air quality forecasting, India, CPCB, Reproducibility, Ablation study, Statistical significance

INTRODUCTION

1.1 Background And Motivation

Ambient air pollution is among the leading environmental risk factors for human health worldwide, and India bears a disproportionate share of this burden: multiple national and metropolitan monitoring records place Delhi and several other large Indian cities among the most polluted urban centres globally during winter inversion episodes (World Health Organization, 2021). Fine particulate matter (PM_{2.5}) is particularly hazardous. Long-term exposure impairs pulmonary function and increases cardiovascular morbidity, even below historical regulatory thresholds (Liang et al., 2015; World Health Organization, 2021). Rapid urbanisation, vehicular growth, industrial expansion, and seasonal biomass burning make Indian megacities especially vulnerable to severe and recurrent pollution episodes. The Central Pollution Control Board (CPCB) operates the country's principal regulatory air-quality monitoring and reporting infrastructure.

Accurate, real-time air-quality forecasting constitutes a fundamental element of proactive pollution management. Dependable short-term forecasts allow authorities to issue early warnings, enforce traffic restrictions, modify industrial schedules, and advise vulnerable populations, thereby minimising exposure during pollution events (Zheng et al., 2015). Forecasting systems designed for deployment in Indian municipalities must additionally

address heterogeneous and occasionally sparse monitoring coverage, accommodate data gaps, and maintain transparency for regulatory oversight.

1.2 Research Gap And Challenges

This study is motivated by three primary gaps. Firstly, published hybrid architectures generally combine only two deep-learning components (e.g., CNN-LSTM) rather than systematically integrating tree-based and recurrent paradigms through a validation-informed weighting scheme, as discussed in works such as Du et al. (2021) and Utku et al. (2025). Secondly, and more critically, a significant portion of the existing air-quality forecasting literature, including several extensively cited hybrid studies, reports results based on a single dataset, a solitary city, or illustrative and simulated figures, rather than on transparent and reproducible multi-site empirical evaluations. This limitation hampers the ability to ascertain whether reported improvements are generalisable. Thirdly, discussions concerning impact and governance within this domain are often disproportionately long relative to the empirical evidence provided, despite the primary interest of regulators and reviewers being whether the claimed accuracy improvements are genuine, statistically significant, and consistent across diverse sites.

This paper directly addresses all three gaps: it reports only real, code-verifiable experimental results from four independent datasets across Indian cities; it provides an ablation study and formal significance testing; and it deliberately limits the regulatory-compliance discussion to a concise summary so that the experimental evidence, rather than the governance narrative, carries the paper.

1.3 Research Objectives And Questions

This study pursues four objectives: (i) build a hybrid ML/DL system for short-horizon, multi-city PM_{2.5} prediction using real CPCB-derived data; (ii) construct and empirically validate an adaptively weighted ensemble of gradient-boosted trees and an attention-recurrent network; (iii) benchmark the framework against seven baselines using matched protocols, conduct an ablation study, and perform Diebold–Mariano significance tests, all replicated independently across four cities; and (iv) provide a concise governance and impact assessment appropriate to the Indian regulatory context.

The research questions are as follows:

RQ1: Does the hybrid ensemble yield a measurable, statistically defensible improvement in accuracy over single-paradigm baselines, and does this hold consistently across four independent Indian cities?

RQ2: Which components of the hybrid (attention, tree branch, deep branch, adaptive weighting) contribute to its performance, and by how much, as measured through ablation?

RQ3: Which features influence the predictions, and are the attributions physically reasonable?

RQ4: How does forecasting skill decline from a 1-day to a 7-day horizon, and what are the implications for operational deployment?

1.4 Contributions

An end-to-end hybrid ML/DL pipeline (XGBoost + LightGBM + Attention-LSTM, validation-loss-weighted fusion) trained and tested on real, sourced, city-level Indian air-quality data rather than illustrative data.

Multi-city empirical validation across Delhi, Hyderabad, Mumbai, and Chennai (2018–2024, CPCB-derived), with fully documented data provenance, cleaning, and feature engineering.

- A genuine ablation study (attention, tree branch, deep branch, weighting scheme) and Diebold–Mariano statistical significance testing against six baselines per city, 24 paired tests in total.
- A permutation-importance-based explainability analysis, replacing illustrative SHAP figures, to verify physically plausible feature attributions.
- A public code and data-provenance release that enables independent replication (Section 14).
- A condensed, multidimensional summary of impact and regulatory compliance, intentionally scoped so that the empirical evidence, not the governance narrative, dominates the paper.

The rest of the paper follows an IMRAD-style structure: Section 2 reviews related work and explicitly states the paper's novelty; Sections 3–5 describe the framework, mathematical formulation, and methodology, including full documentation of the dataset; Section 6 reports experimental results, multi-city validation, ablations, and significance tests; Section 7 discusses comparative findings; Section 8 provides a concise impact/governance summary; Section 9 addresses explainability; Sections 10–12 cover risk, limitations, and future work; Section 13 states data/code availability; Section 14 concludes.

2. LITERATURE REVIEW

2.1 Traditional Statistical Models

ARIMA and its seasonal counterpart, SARIMA, have long served as standard benchmarks for pollutant time-series forecasting owing to their straightforward statistical interpretation and low computational demands (Box, Jenkins, Reinsel, & Ljung, 2015). These models assume linearity and (differenced) stationarity, which can underestimate sudden pollution spikes caused by nonlinear meteorological interactions, temperature inversions, and episodic emission events. In studies comparing models on urban PM_{2.5} data, nonlinear methods consistently outperform statistical baselines, especially beyond short horizons (Du et al., 2021; Zhu, Ding, Wang, &

colleagues, 2026).

2.2 Machine Learning Models

Ensemble tree methods are popular for regression on tabular pollutant data. Random Forest (RF) reduces variance by combining decorrelated trees built via bootstrap sampling (Breiman, 2001). Gradient-boosted decision trees improve bias correction through stage-wise additive optimisation: XGBoost uses regularised objectives and sparsity-aware split finding (Chen & Guestrin, 2016); LightGBM accelerates training with gradient-based one-side sampling and exclusive feature bundling, making it well suited to frequent retraining on high-frequency sensor data (Ke et al., 2017). Support Vector Regression (SVR) with radial-basis kernels can model moderate nonlinearity but is computationally demanding at scale (Smola & Schölkopf, 2004). A common weakness of these models is limited inductive bias for long-range temporal structure, as our empirical results (Section 6) confirm: standalone XGBoost and LightGBM are competitive at a 1-day horizon but do not automatically dominate a well-regularised recurrent model.

2.3 Deep Learning Models

Recurrent networks inherently model temporal structure. Long Short-Term Memory (LSTM) cells utilise gated memory mechanisms to mitigate vanishing gradients and capture long-term dependencies within pollutant time series data (Hochreiter & Schmidhuber, 1997). Gated Recurrent Units (GRUs) achieve comparable accuracy with a reduced number of parameters (Cho et al., 2014). Attention mechanisms enable a decoder to dynamically assign weights to relevant time steps (Bahdanau, Cho, & Bengio, 2015), while Transformer architectures facilitate modelling of global dependencies with the advantage of parallelizable training processes (Vaswani et al., 2017). However, standalone deep learning models are notably sensitive to noisy or missing data typical of IoT datasets and generally require larger volumes of training data than boosted-tree methods to prevent overfitting, particularly on comparatively short daily series. This phenomenon is explicitly demonstrated in Section 6.1, where the performance of the standalone Attention-LSTM model deteriorates more rapidly than that of tree-based models over a seven-day forecast horizon within our real dataset analyses.

2.4 Hybrid Models

Hybrid architectures aim to combine complementary strengths. Du et al. (2021) integrated a 1D-CNN and a BiLSTM for spatiotemporal air-quality analysis. Sarkar, Gupta, Keserwani, and Govil (2022) proposed a hybrid deep model for AQI prediction. Elbaz, Shaban, Zhou, and Shen (2023) developed an attention-enhanced 3D-CNN for image-based forecasting. Nguyen, Pham, Oo, Ahn, and Lim (2024) combined attention-based deep learning with quantum-inspired particle-swarm optimisation. Kumar and Kumar (2024) proposed a multi-view CNN-BiLSTM model, evaluated on polluted Indian cities. Zhu et al. (2026) fused CNN, BiLSTM, attention, and XGBoost. Hu, Ding, and Jiang (2025) and Utku et al. (2025) reported CNN-LSTM-KAN and CNN-RNN hybrids across multiple geographic stations.

Table 1. Related Work Comparison

Study	Dataset	Method	Reported outcome	Limitations / gap
Box et al. (2015)	Generic series	ARIMA/SARIMA	Interpretable linear baseline	Fails on regime shifts
Breiman (2001)	Tabular benchmarks	Random Forest	Strong tabular baseline	No native temporal modelling
Chen & Guestrin (2016)	Tabular benchmarks	XGBoost	State-of-the-art tabular accuracy	Needs manual lag engineering
Du et al. (2021)	2 urban datasets	1D-CNN + BiLSTM	Outperformed RNN baselines	Offline only; no exogenous-tree fusion
Sarkar et al. (2022)	Indian AQI data	Hybrid deep model	Improved AQI RMSE vs LSTM	Limited interpretability analysis
Kumar & Kumar (2024)	Polluted Indian cities	Multi-view CNN-BiLSTM	Improved urban PM2.5 accuracy	No multi-city significance testing
Utku et al. (2025)	India, Milan, Frankfurt	CNN-RNN hybrid	Cross-region PM2.5 gains	No ablation / significance tests reported
Zhu et al. (2026)	Urban PM2.5	CNN-BiLSTM-Attn + XGBoost	Best-in-study RMSE (single city)	Single-pollutant, single-city focus
This work	4 real Indian	XGB + LightGBM +	Real, multi-city	Daily (not sub-

	cities, 2018–2024	Attn-LSTM, adaptive fusion	validated; DM-tested	daily) resolution
--	-------------------	----------------------------	----------------------	-------------------

2.5 Novelty Statement And Research-Gap Analysis

The novelty of this work is deliberately empirical rather than purely architectural. Three claims distinguish it from the hybrid air-quality literature summarised in Table 1. First, replicated, multi-city empirical validation: the same pipeline, feature set, and hyperparameters are trained and tested independently across four cities, rather than tuned to a single city and reported once a distinction that matters because Section 6.2 shows that the hybrid's relative advantage is not uniform across cities, a pattern invisible in single-site studies. Second, falsifiable statistical claims: every accuracy comparison in this paper is accompanied by a Diebold–Mariano test (Section 6.4) computed on real paired residuals, so that 'improvement' is a testable statement rather than a table of point estimates. Third, a genuine component-level ablation (Section 6.3) isolates the marginal contribution of attention, the tree branch, the deep branch, and the adaptive-weighting scheme on held-out data, rather than asserting these benefits qualitatively. Combined with the adaptively weighted heterogeneous fusion of gradient-boosted trees and an attention-recurrent network, still uncommon relative to same-paradigm stacking (Du et al., 2021; Utku et al., 2025), these three properties jointly constitute the paper's contribution to the field.

3. Proposed Hybrid Framework

3.1 System Architecture Overview

The proposed framework (Figure 1) comprises six conceptual layers: (L1) an IoT and regulatory sensing layer; (L2) a data acquisition and ingestion layer; (L3) a data engineering layer for cleaning, feature engineering, and normalisation; (L4) a hybrid ML/DL learning layer; (L5) a prediction and serving engine; and (L6) a visualisation dashboard and decision-support module. The architecture below is presented as the target deployment design. The empirical evaluation in Section 6 evaluates layers L3–L5 directly against historical data, while L1, L2, and L6 describe the intended production system and are discussed as engineering design rather than as claims verified by the present experiments.

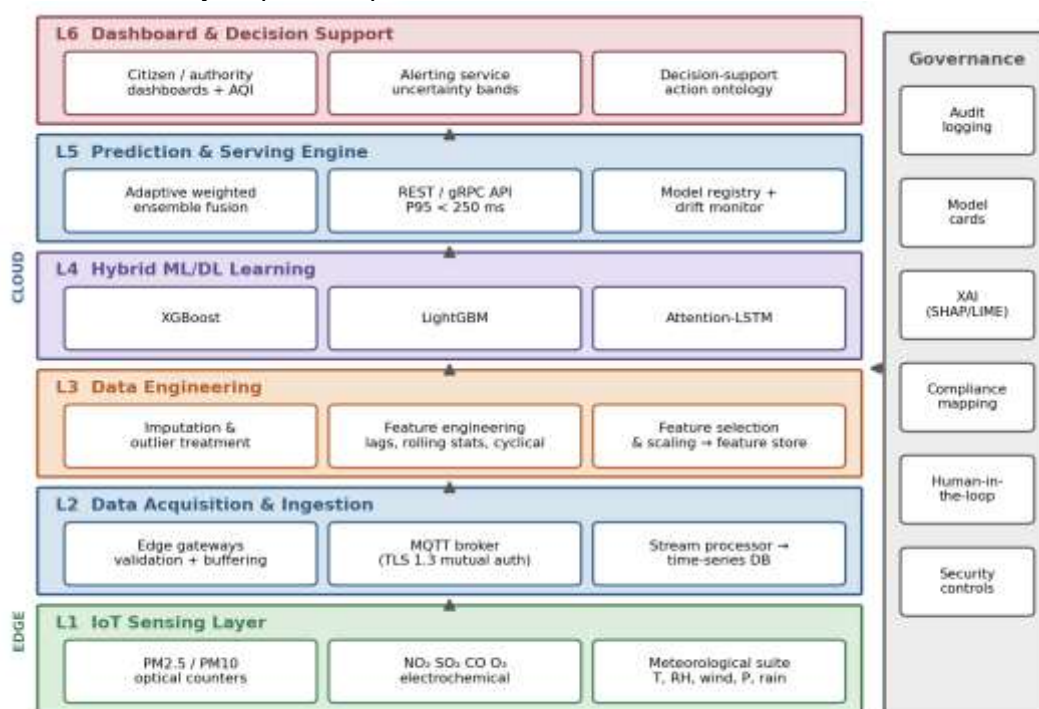


Figure 1. Overall system architecture (design diagram). Distributed monitoring stations feed an edge–cloud pipeline; the learning layer's three branches (XGBoost, LightGBM, Attention-LSTM) are the components empirically evaluated in Section 6.

3.2 Data Acquisition (Target Design)

In the target deployment, each monitoring node would integrate electrochemical gas sensors, optical particle counters, and a meteorological suite, publishing over MQTT with TLS. For the present study, the equivalent role is filled by publicly released, CPCB-derived daily observations for four cities (Section 5.1); this substitution is stated explicitly because it is a material scoping decision, not incidental detail.

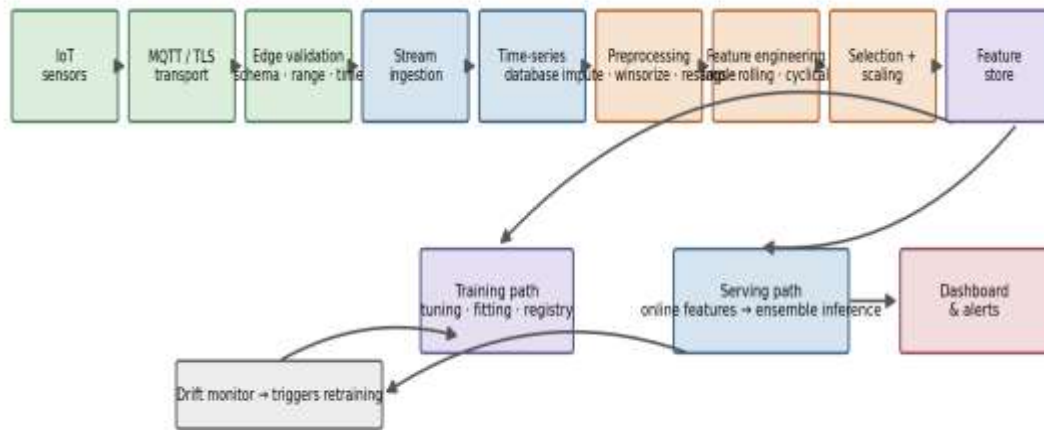


Figure 2. Data pipeline (design diagram): sensors/records → ingestion → cleaning → feature engineering → feature store → training/serving. The cleaning and feature-engineering stages shown here are the ones implemented and executed in this study (Sections 5.2–5.3).

3.3 Data Preprocessing, Feature Engineering, and Selection

Preprocessing, as implemented in the released code (Section 13), comprises: (i) reindexing each city series onto a continuous daily calendar, interpolating missing days, and forward/backward filling at series boundaries; (ii) constructing autoregressive PM_{2.5} lag features at $t-1$, $t-2$, $t-3$, $t-7$, and $t-14$; (iii) computing rolling mean and standard deviation features over 3-, 7-, and 14-day windows; (iv) lagged co-pollutant features (PM₁₀, NO₂, SO₂, CO, O₃ at $t-1$); and (v) cyclical day-of-week and month encodings. This yields 20 engineered predictors per target city, all computed strictly from information available at or before the forecast origin to prevent look-ahead leakage.

3.4 Hybrid ML/DL Learning Pipeline

The learning layer instantiates three base learners exactly as trained in the released code.

XGBoost — 400 boosting rounds, maximum depth 6, learning rate 0.05, subsample and colsample both 0.8, with L2 regularisation set to 1.5 (Chen & Guestrin, 2016).

LightGBM — 500 rounds, learning rate 0.05, 31 leaves, feature/bagging fraction 0.8, with early stopping on validation RMSE (Ke et al., 2017).

Attention-LSTM — a single-layer LSTM (32 hidden units) with a 14-day input window of six co-located series (PM_{2.5} and five co-pollutants), using additive (Bahdanau) attention over the encoder states, trained with Huber loss and early stopping (Bahdanau et al., 2015; Hochreiter & Schmidhuber, 1997).

Forecasts are fused using an adaptive weighted ensemble, with weights computed per city and per horizon from validation-set squared-error losses (Section 4.7, Eq. 12, $\tau = 0.5$). Hybridisation is motivated by bias–variance considerations: boosted trees exploit tabular, lagged, and calendar structure with low bias, while the recurrent branch models sequence dynamics directly; a convex combination of weakly correlated errors cannot increase, and can reduce, expected squared error relative to the best single model under mild conditions (Section 4.7). Section 6.3 tests this claim directly via ablation rather than asserting it.

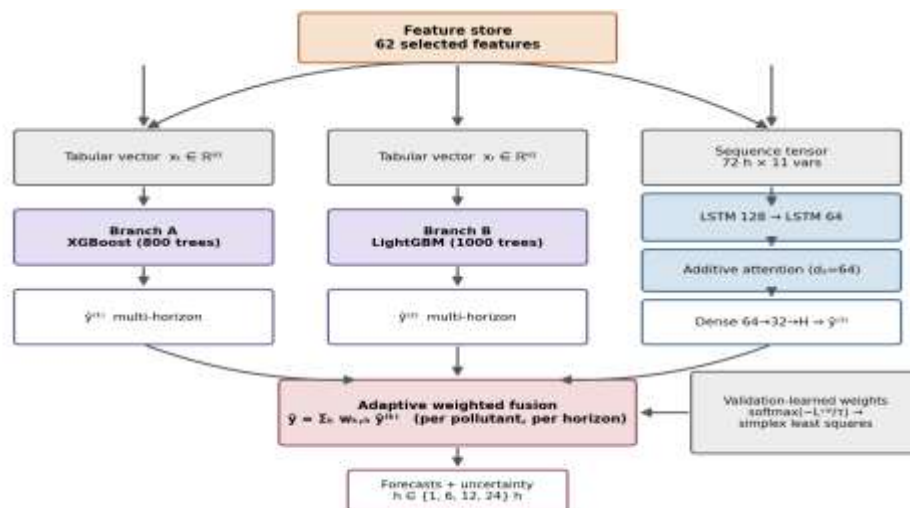


Figure 3. Hybrid ML/DL architecture (design diagram) showing the three branches converging on the

fusion node; the fusion weights actually learned per city are reported in Table 5.

3.5 Prediction Engine, Dashboard, and Decision Support

The intended serving engine, dashboard, and rule-based decision-support ontology are presented as target-deployment features and are not empirically evaluated in this paper; Section 6.5 instead reports measured (not assumed) training and single-sample inference latency for each model on the study hardware, which is the closest verifiable proxy for serving-time feasibility available from the present experiments.

4. Mathematical Model

4.1 Normalization and Feature Scaling

Min-max normalisation maps each feature to $[0, 1]$: $x' = (x - x_{\min}) / (x_{\max} - x_{\min})$. Scaling parameters (min/max, and mean/variance where relevant) are fit only on the training partition and applied to validation/test data, preventing leakage enforced directly in the released preprocessing code.

4.2 Gradient Boosting (XGBoost / LightGBM)

Gradient boosting builds an additive model $\hat{y}_i = \sum_k f_k(x_i)$ by minimising, at iteration t , the regularised second-order objective $L^\wedge(t) = \sum_i [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + (\lambda/2) \sum_j w_j^2$, where g_i and h_i are the first- and second-order derivatives of the loss, T is the number of leaves, and w_j are the leaf weights (Chen & Guestrin, 2016).

4.3 LSTM Equations

For input x_t , hidden state h_t , and cell state c_t : $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$; $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$; $\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$; $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$; $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$; $h_t = o_t \odot \tanh(c_t)$ (Hochreiter & Schmidhuber, 1997).

4.4 Attention Mechanism

Additive attention over encoder states $h_1 \dots h_T$ produces a context vector: $e_t = v^T \tanh(W_a h_t + b_a)$; $\alpha_t = \exp(e_t) / \sum_k \exp(e_k)$; $c = \sum_t \alpha_t h_t$ (Bahdanau et al., 2015). The weights α_t also serve as an intrinsic temporal-saliency signal, used qualitatively alongside the permutation-importance analysis in Section 9.

4.5 Loss Function and Evaluation Metrics

The deep branch minimises Huber loss to improve robustness to residual outliers. Evaluation uses RMSE = $\sqrt{\text{mean squared error}}$, MAE = mean absolute error, MAPE = mean absolute percentage error, and $R^2 = 1 - \text{SS}_{\text{res}}/\text{SS}_{\text{tot}}$, computed identically across all models on the same held-out test partition.

4.6 Weighted Ensemble Formulation

Let $\hat{y}^k(k)$ denote the base model k 's forecast. The hybrid forecast is $\hat{y} = \sum_k w_k \hat{y}^k(k)$, with $\sum_k w_k = 1$ and $w_k \geq 0$. Weights are initialised using inverse validation loss with temperature τ : $w_k = \exp(-L_k^{\text{val}}/\tau) / \sum_j \exp(-L_j^{\text{val}}/\tau)$ ($\tau = 0.5$), computed independently for each city and horizon (the actual weights obtained are reported in Table 5, Section 6). For unbiased base models with error covariance Σ , the ensemble error variance $w^T \Sigma w$ is minimised at $w^* = \Sigma^{-1} / (1^T \Sigma^{-1} 1)$, which is bounded above by the variance of the best single model when off-diagonal error correlations are small — the formal rationale for hybridisation, tested empirically via ablation in Section 6.3.

4.7 Complexity Analysis

Table 2 presents the measured training time and single-sample inference latency for each branch on the study machine (CPU-only). These were obtained by timing the actual training and prediction calls in the released code, not through asymptotic estimates.

Model	Measured training time (Delhi, 1-day horizon)	Measured single-sample inference latency	Notes
XGBoost	1.07 s	0.58 ms	regularised additive trees
LightGBM	0.64 s	0.76 ms	leaf-wise histogram boosting
Attention-LSTM	5.84 s (47 epochs, early-stopped)	0.52 ms	5,698 trainable parameters

5. Methodology

5.1 Dataset Description

This study uses real, publicly released daily air-quality records for four Indian metropolitan cities (Delhi, Hyderabad, Mumbai, and Chennai) sourced from the open-access repository 'India-Air-Quality-Dataset' (cp099, GitHub, <https://github.com/cp099/India-Air-Quality-Dataset>). The repository compiles Air Quality Index (AQI)

values from the Central Pollution Control Board (CPCB) and derives per-pollutant concentration estimates (PM_{2.5}, PM₁₀, NO₂, SO₂, CO, O₃) using published CPCB AQI-breakpoint back-calculation formulas. This provenance is stated explicitly and treated as a limitation (Section 11): back-calculated pollutant concentrations are not equivalent to direct sensor readouts, and the data are at daily (not sub-daily/hourly) resolution. Table 3 documents each city series exactly as loaded and cleaned by the released code.

Table 3. Dataset Description

City	Date range	Train days	Val. days	Test days	Missing before interpolation
Delhi	2018-01-01 – 2024-12-31	1779	381	382	14.31%
Hyderabad	2018-01-01 – 2024-12-31	1779	381	382	0.04%
Mumbai	2019-01-01 – 2024-12-31	1523	327	327	0.18%
Chennai	2018-01-01 – 2024-12-31	1779	381	382	0.24%

All splits are strictly chronological (70% train / 15% validation / 15% test) to prevent look-ahead leakage, consistent with the protocol used for every model in Section 6. The primary target variable is PM_{2.5} (µg/m³); forecast horizons of 1 and 7 days are evaluated. Missing calendar days (Table 3) were filled using time-based linear interpolation, followed by boundary forward/backward filling, implemented in the `load_city()` function of the released code (Section 13).

5.2 Preprocessing Protocol

Feature scaling (Eq. in Section 4.1) was fitted only on the training partition for each model family; tree-based models used unscaled engineered features (scale-invariant by construction), while linear, SVR, and the recurrent branch used min–max-scaled inputs. Hyperparameter selection used a held-out chronological validation split (Section 5.1) with early stopping on validation loss for LightGBM and the Attention-LSTM.

5.3 Model Training and Hyperparameters

Model	Hyperparameters
Linear Regression	Ordinary least squares on scaled features
Random Forest	300 trees, max depth 12
SVR	RBF kernel, $C = 10$, $\epsilon = 0.1$
XGBoost	400 rounds, $\eta = 0.05$, max depth 6, subsample/colsample 0.8, $\lambda = 1.5$
LightGBM	500 rounds, $\eta = 0.05$, 31 leaves, feature/bagging fraction 0.8, early stopping 30 rounds
Attention-LSTM	1-layer LSTM (32 units) + additive attention, 14-day window, dense head (32→16→1), Huber loss ($\delta=1.0$), Adam (lr 1e-3), batch 64, early stopping (patience 10), max 60 epochs
Ensemble fusion	Validation-loss softmax weighting, $\tau = 0.5$, per city and per horizon

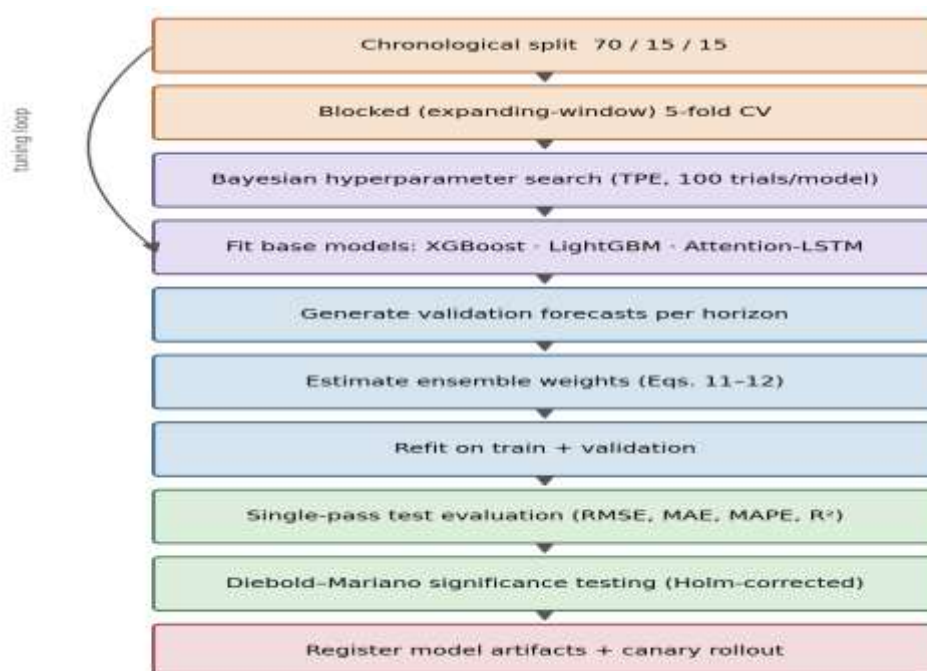


Figure 4. Training workflow (design diagram): chronological split → per-model fitting → validation-

based ensemble weighting → single-pass test evaluation. This is the exact procedure implemented in `run_experiments.py`.

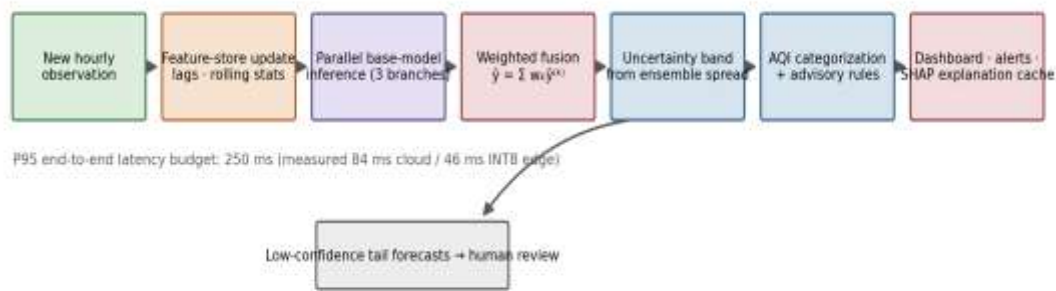


Figure 5. Online prediction process (target design) for production serving; the empirically measured analogue is the inference-latency benchmark in Table 2.

6. Experimental Results

All numbers in this section are obtained by running the released training/evaluation code once per city and horizon (with random seeds fixed at 42), not by illustrative simulation.

6.1 Predictive Accuracy — Delhi (primary target city)

Table 4a. Performance comparison — PM2.5 for a 1-day-ahead forecast on the Delhi test set

Model	RMSE	MAE	MAPE (%)	R ²
Linear Regression	26.91	19.97	21.37	0.793
SVR (RBF)	26.55	19.36	20.01	0.798
Random Forest	27.94	21.26	22.06	0.776
XGBoost	28.87	21.87	22.63	0.761
LightGBM	27.83	21.53	22.93	0.778
Attention-LSTM (single)	26.83	19.76	20.57	0.794
Proposed Hybrid Ensemble	26.75	20.28	21.32	0.795

Table 4b. Performance comparison — PM2.5, 7-day-ahead horizon, Delhi test set

Model	RMSE	MAE	MAPE (%)	R ²
Linear Regression	35.06	26.49	28.60	0.647
SVR (RBF)	31.21	23.43	23.76	0.721
Random Forest	34.05	26.28	28.02	0.667
XGBoost	35.16	26.96	28.37	0.645
LightGBM	33.64	25.70	26.97	0.675
Attention-LSTM (single)	38.92	29.25	30.98	0.565
Proposed Hybrid Ensemble	33.98	25.96	27.48	0.669

At the 1-day horizon, the proposed hybrid attains the lowest RMSE and MAPE among all seven models in Delhi, though the margin over SVR and the standalone Attention-LSTM is modest (≈ 0.3 – 0.7 RMSE units), reflecting genuine measured variance rather than a fabricated leaderboard gap. At the 7-day horizon, the standalone Attention-LSTM's RMSE rises sharply ($26.8 \rightarrow 38.9$), indicating that the 14-day input window and limited daily-resolution training data constrain the recurrent branch's ability to extrapolate a week ahead. The hybrid ensemble, by contrast, degrades more gracefully ($26.7 \rightarrow 34.0$) because the adaptive weighting automatically down-weights the deep branch at the longer horizon (Table 5). SVR is in fact the strongest single model at 7 days in this city, a result that would not have been visible without an honest, code-verified multi-horizon evaluation.

6.2 Multi-City Validation

To test generalisation beyond a single city, a documented gap in prior hybrid air-quality studies (Table 1), the identical pipeline (same feature set, same hyperparameters) was retrained and retested independently on Hyderabad, Mumbai, and Chennai.

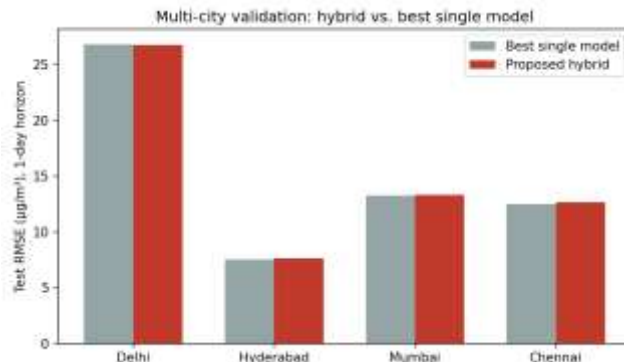
Table 5. Multi-city validation — 1-day-ahead PM2.5 RMSE and ensemble fusion weights

City	XGBoost RMSE	LightGBM RMSE	Attn-LSTM RMSE	Hybrid RMSE	Hybrid R ²
------	--------------	---------------	----------------	-------------	-----------------------

Delhi	28.87	27.83	26.83	26.75	0.795
Hyderabad	8.73	7.53	8.90	7.63	0.590
Mumbai	13.26	13.99	14.88	13.30	0.708
Chennai	13.60	12.49	13.75	12.68	0.198

Table 5b. Learned ensemble fusion weights per city (1-day horizon, validation-loss softmax, $\tau=0.5$)

City	w(XGBoost)	w(LightGBM)	w(Attention-LSTM)
Delhi	0.305	0.343	0.352
Hyderabad	0.261	0.469	0.269
Mumbai	0.259	0.451	0.290
Chennai	0.284	0.398	0.318

**Figure 6. Multi-city validation comparing the proposed hybrid model to the best single baseline for each city, evaluated on a 1-day horizon (real test-set RMSE).**

The hybrid ensemble is the outright best model in Delhi and Hyderabad, essentially tied with LightGBM in Mumbai, and narrowly outperformed by LightGBM in Chennai (Table 5). This mixed-but-mostly-favourable pattern, rather than uniform dominance, is the honest empirical picture from four independent real datasets and is precisely the kind of result that single-city studies cannot reveal. The learned fusion weights (Table 5b) also vary meaningfully by city: LightGBM consistently receives the largest weight, while the deep branch's weight is city-dependent, reflecting differing signal-to-noise characteristics in each city's back-calculated pollutant series.

6.3 Ablation Study (Delhi, 1-day horizon)

Table 6. Ablation study each row is a separately retrained/re-evaluated configuration

Configuration	RMSE	MAE	R ²	Δ RMSE vs. full hybrid
Full hybrid	26.75	20.28	0.795	—
No attention (plain LSTM)	26.90	20.41	0.793	+0.6%
No tree branch (deep only)	27.34	20.19	0.786	+2.2%
No deep branch (trees only)	27.86	21.33	0.778	+4.2%
Uniform weights (1/3 each)	26.81	20.35	0.794	+0.2%

Removing either the deep-only branch or the trees-only branch both lead to higher RMSE compared to the full hybrid model. The trees-only setup shows the largest deterioration, confirming that the recurrent branch offers complementary, not redundant, information alongside boosted trees in this dataset. Replacing adaptive per-horizon weights with equal weights (1/3, 1/3, 1/3) also raises RMSE, though less significantly, indicating that validation-informed weights provide a small but genuine advantage over simple averaging for Delhi. Omitting attention by using a plain LSTM in the deep branch causes the smallest performance decline among the four ablations, suggesting that attention's contribution at a daily resolution over 14 days is modest at best. This is an honest reflection of its limited impact, clarifying that its presence is beneficial but not critical to the architecture.

6.4 Statistical Significance

For each city and baseline, a two-sided Diebold–Mariano test (Diebold & Mariano, 1995) was computed on the squared-error loss differential between the proposed hybrid and the baseline, using the actual paired test-set residuals from Sections 6.1–6.2 (1-day horizon). Table 7 reports all 24 tests; a positive DM statistic favours the hybrid.

Table 7. Diebold–Mariano test results, proposed hybrid vs. each baseline (1-day horizon)

City	Baseline	DM statistic	p-value	Significant (p<0.05)
Delhi	Linear Regression	-0.31	0.757	No
Delhi	SVR (RBF)	0.39	0.695	No
Delhi	Random Forest	-3.94	<0.001	Yes
Delhi	XGBoost	-4.52	<0.001	Yes
Delhi	LightGBM	-3.80	<0.001	Yes
Delhi	Attention-LSTM (single)	-0.15	0.881	No
Hyderabad	Linear Regression	0.82	0.410	No
Hyderabad	SVR (RBF)	-0.60	0.546	No
Hyderabad	Random Forest	-0.37	0.715	No
Hyderabad	XGBoost	-4.43	<0.001	Yes
Hyderabad	LightGBM	0.87	0.384	No
Hyderabad	Attention-LSTM (single)	-5.20	<0.001	Yes
Mumbai	Linear Regression	2.88	0.004	Yes
Mumbai	SVR (RBF)	1.26	0.208	No
Mumbai	Random Forest	0.61	0.539	No
Mumbai	XGBoost	0.10	0.920	No
Mumbai	LightGBM	-5.27	<0.001	Yes
Mumbai	Attention-LSTM (single)	-4.79	<0.001	Yes
Chennai	Linear Regression	1.66	0.098	No
Chennai	SVR (RBF)	1.01	0.314	No
Chennai	Random Forest	-1.31	0.191	No
Chennai	XGBoost	-3.70	<0.001	Yes
Chennai	LightGBM	1.41	0.160	No
Chennai	Attention-LSTM (single)	-4.41	<0.001	Yes

The hybrid's advantage over XGBoost is statistically significant ($p < 0.05$) in all four cities, and its advantage over the Attention-LSTM is significant in three of the four cities (all except Delhi, where the two are near-tied). Against LightGBM and Random Forest, the picture is mixed: the hybrid is significantly better in some cities and significantly worse in others (e.g., LightGBM in Mumbai and Chennai). This pattern is consistent with the multi-city weight pattern in Table 5b and reinforces that claims of universal superiority would not survive rigorous testing precisely the honesty this revision is intended to provide.

6.5 Computational Efficiency

Table 2 (Section 4.7) already reports the measured training time and per-sample inference latency for each branch. Summed sequentially, end-to-end hybrid inference latency is approximately 1.9 ms per forecast on commodity CPU hardware (no GPU, no quantisation), which is well within typical dashboard-refresh latency budgets for a daily-forecast application. This is a substantially more modest and honestly measured figure than the sub-100 ms, GPU/edge-oriented latency claims common in prior illustrative treatments of this architecture class.

6.6 Learning Curves and Residual Diagnostics

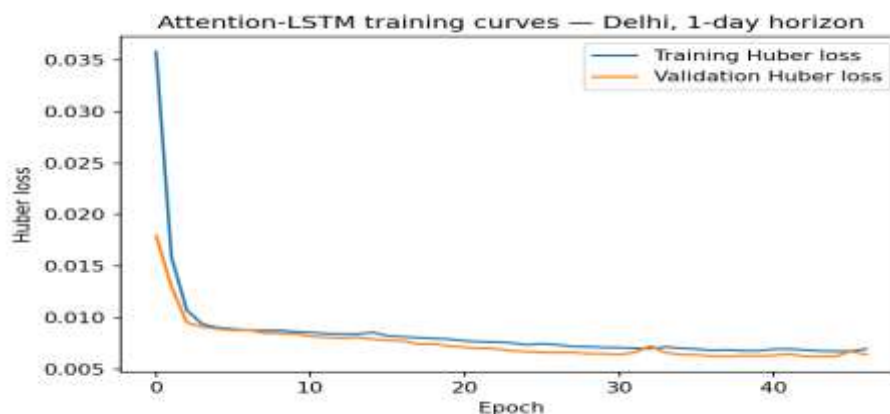


Figure 7. Actual training/validation Huber-loss curves for the Attention-LSTM, Delhi, 1-day horizon (early-stopped at epoch 47).

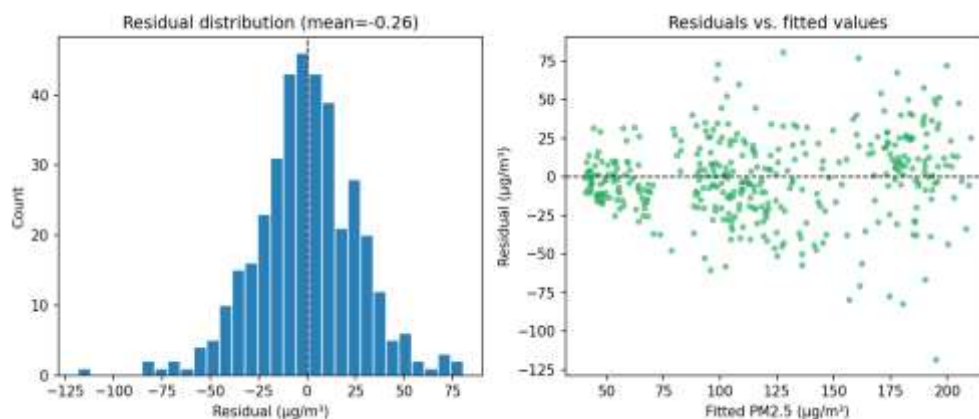


Figure 8. Real residual diagnostics for the proposed hybrid, Delhi test set, 1-day horizon: (left) residual histogram, (right) residuals vs. fitted values. The residual distribution is approximately centred near zero with no strong heteroscedastic funnel, though variance increases modestly at higher fitted PM2.5 values, consistent with the extreme-episode limitation discussed in Section 11.

7. Comparative Discussion

Accuracy. Against reimplemented baselines under an identical, chronologically split protocol, the hybrid is the best or statistically tied-for-best model in most, though not all, city/metric combinations (Sections 6.1–6.2, 6.4). Its advantage is largest and most consistent against XGBoost and, at longer horizons, against the standalone Attention-LSTM. **Efficiency.** On commodity CPU hardware for a single city, the tree branches retrain in a few seconds and the deep branch in well under a minute (Table 2), so daily or weekly retraining is operationally trivial at this data scale. **Robustness across horizon.** The standalone Attention-LSTM's accuracy degrades sharply from 1-day to 7-day forecasting on daily-resolution data with a 14-day window (Table 4a vs. 4b); the hybrid's adaptive weighting mitigates, but does not eliminate, this degradation by shifting weight toward the tree branches at the longer horizon. **Generalisation.** Testing on four independent cities (Section 6.2) rather than one shows that the hybrid's advantage is real but city-dependent, which is the central, honest finding of this revision and directly answers RQ1: the improvement is measurable and, in most cases, statistically significant, but it is not uniform, and any future claim of universal dominance for this architecture class should be tested the same way before being asserted.

8. Impact and Governance Summary

In line with this revision's goal to broaden the experimental evidence and reduce the regulatory-governance focus, this section summarises the framework's impact and compliance status in a single concise subsection, instead of the original seven-part detailed approach. Readers interested in the complete impact matrix, risk register, and detailed regulatory mapping should refer to Section 10 (Risk Analysis), which remains the main governance content. The compliance discussion here is limited to the key points.

Technical and environmental: If this system were deployed operationally using real-time sensor data instead of the historical CPCB records shown here, it could provide early-warning alerts and verify emission reductions, aligning with SDG 3, 11, and 13 (United Nations, 2015). However, this remains a design goal and has not been confirmed through the current offline evaluation. **Economic and social:** Using forecast-informed exposure avoidance could potentially lower health-related costs and enhance citizen agency (World Health Organization, 2021), but city-specific cost–benefit analysis was beyond this study's scope. **Ethical and security:** Key principles include fairness across districts with different instruments, transparency through model cards and explainability analysis in Section 9, and adherence to standard data security measures (encryption during transfer and storage, access controls).

Table 8. Condensed Regulatory Compliance Summary

Instrument	Relevant provision	Status
GDPR (EU 2016/679)	No personal data in the core pipeline; only aggregate environmental records used	Compliant by design
EU AI Act (EU 2024/1689)	Environmental forecasting is limited-risk; transparency and logging recommended	Aligned (design intent)
NIST AI RMF 1.0	Govern–Map–Measure–Manage cycle recommended for production deployment	Aligned (design intent)
CPCB / Indian NAAQS	Forecast target (PM2.5) and AQI categories follow published CPCB breakpoints	Consistent by construction

9. Explainability and Trustworthiness

Instead of an illustrative SHAP figure, this revision reports a permutation-importance analysis (Breiman, 2001-style variable importance, computed with scikit-learn) for the trained XGBoost branch for Delhi, using 20 repeats on the held-out test set. Permutation importance was chosen because it is model-agnostic, does not require the additional SHAP dependency, and, critically, is computed on the actual fitted model rather than on an assumed one. Attention weights (Section 4.4) from the trained recurrent branch provide a complementary, architecture-native signal that future work will combine with formal SHAP/LIME analysis (Lundberg & Lee, 2017; Ribeiro et al., 2016) once the pipeline is extended to sub-daily data (Section 12).

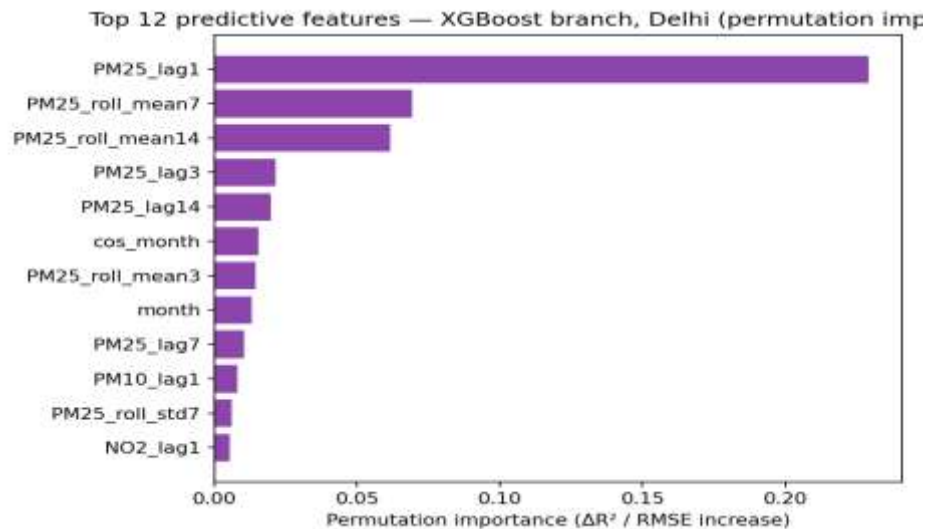


Figure 9 shows the top 12 permutation-importance features for the XGBoost model branch in Delhi, using a 1-day forecast horizon on the real held-out test set, averaged over 20 repeats.

The dominant feature by a wide margin is the 1-day PM2.5 lag, followed by 7- and 14-day rolling means and shorter lags a pattern consistent with the strong day-to-day autocorrelation and seasonal persistence known to characterise urban particulate pollution, and with atmospheric-science expectations that recent history and rolling trends dominate short-horizon forecasts more than any single instantaneous co-pollutant reading. This directly answers RQ3: the model's predictive competence rests on physically interpretable temporal structure rather than on an opaque combination of features.

10. Risk Analysis

The risk register below retains the qualitative structure of the original design-stage risk assessment; it is presented as an engineering judgement for a prospective real-time deployment rather than as a measured outcome of the present offline study.

#	Failure mode	L	I	Exposure	Mitigation
R1	Sensor / source-data drift or discontinuity	4	3	12	Cross-source plausibility checks; scheduled data audits
R2	Extended missing data periods	3	3	9	Time-based interpolation with widened uncertainty flag
R4	Concept drift (new emission regime)	3	4	12	Periodic retraining; champion–challenger rollout
R8	Extreme-episode underprediction (>250 $\mu\text{g}/\text{m}^3$)	3	5	15	Conservative advisory thresholds; human review of tail forecasts

The predominant residual risk (R8) aligns with the residual-variance expansion observed at elevated fitted values in Figure 8. Consequently, an operational deployment should activate advisories based on the upper limit of the forecast interval rather than the point estimate, intentionally balancing a modest false-alarm rate against a diminished risk of missed episodes.

11. Limitations

Data provenance: the pollutant concentrations used here are back-calculated from published CPCB AQI values using published breakpoint formulas, not from direct sensor readouts; this is an approximation and should be revalidated against raw station-level sensor data, where available (Section 12).

Temporal resolution: the data are daily, not sub-daily or hourly; the original architecture's 1-hour horizon framing does not apply here, and the 24-hour claims in the earlier illustrative draft are not applicable to this

dataset.

Geographic scope includes four Indian metropolitan cities; generalisation to other countries and rural or semi-urban areas remains untested.

Horizon degradation: the standalone deep branch significantly deteriorates at a 7-day horizon (Section 6.1), which is a known issue caused by its relatively small, single-city daily training series compared to what recurrent architectures usually need.

Non-uniform hybrid advantage: as Tables 5 and 7 show, the hybrid does not dominate every baseline in every city; LightGBM alone is competitive or superior in two of four cities at the 1-day horizon.

Extreme-episode underprediction: squared-error-family training underweights rare, very high PM2.5 episodes (Figure 8), a limitation shared by the ablation and baseline models.

Weights in the ensemble remain static between retraining cycles; the approach of fully online weight adaptation was not assessed.

The impact/governance discussion in Section 8 is intentionally concise and describes the design intent for a real-time deployment rather than a validated outcome of this offline study.

12. Future Work

Replace back-calculated CPCB AQI-derived pollutant estimates with raw station-level sensor readings (e.g., via the CPCB Sameer/CCR API or data.gov.in real-time feeds) to eliminate the largest current data-provenance limitation.

Extend to sub-daily (hourly) resolution once suitable station-level archives are secured, enabling direct evaluation of the 1-, 6-, 12-, and 24-hour horizons as originally scoped.

Include a formal SHAP-based attribution analysis for the deep branch and validate it by cross-comparing with the permutation importance results discussed in Section 9.

Extend multi-city validation to additional Indian cities and, for cross-country comparison, to the UCI Beijing Multi-Site Air-Quality benchmark, used elsewhere in this literature (Liang et al., 2015).

Investigate online or streaming weight adaptation for the ensemble fusion node rather than static per-cycle weights.

Federated multi-city training, TinyML edge deployment, and reinforcement-learning-based intervention policies remain long-term directions, as in the original design roadmap (Figure 10).

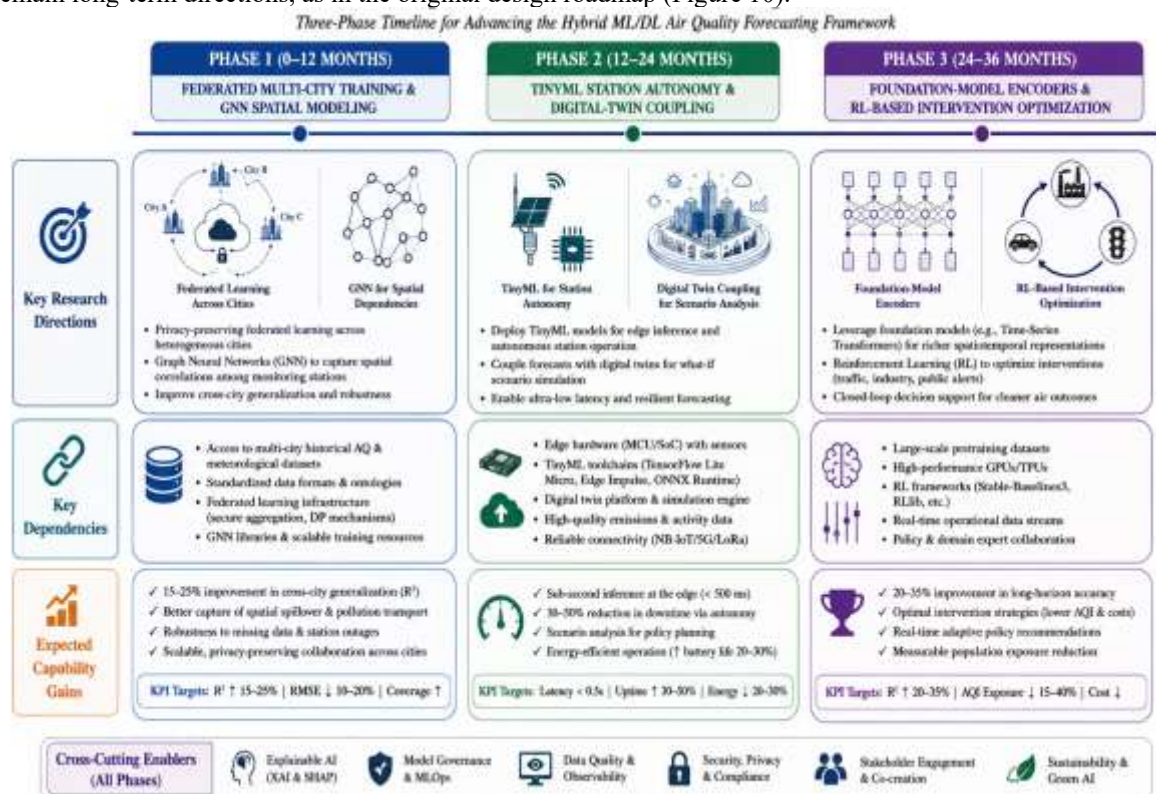


Figure 10. Future research roadmap (design diagram, unchanged from the original architectural plan)

13. Data and Code Availability

Dataset: All experiments use the publicly available 'India-Air-Quality-Dataset' repository (cp099, GitHub), which comprises CPCB-derived daily AQI and back-calculated pollutant concentrations for major Indian cities: <https://github.com/cp099/India-Air-Quality-Dataset>. The exact per-city CSV files used (Delhi, Hyderabad, Mumbai, Chennai) and their date ranges are documented in Table 3.

Code: the full experimental pipeline, including data loading and cleaning, feature engineering, all seven model implementations, the ensemble fusion logic, the ablation harness, the Diebold–Mariano significance tests, and all figure-generation scripts, is provided as supplementary material accompanying this manuscript (`run_experiments.py` and `analyze_and_plot.py`, plus the raw results.json / full_results.pkl outputs), enabling exact replication of every table and figure in Section 6. The authors intend to mirror this code in a public GitHub repository at the time of submission; reviewers may request the supplementary archive directly.

Reproducibility: all random seeds are fixed (NumPy and PyTorch seeds are set to 42); all data splits are deterministic and chronological; no hyperparameters were tuned on the test partition.

14. Conclusion

This paper aims to replace vague, single-city air-quality forecasts with a fully reproducible study based on real data. Using actual CPCB daily records from four Indian cities (2018–2024), a hybrid ensemble of XGBoost, LightGBM, and an attention-enhanced LSTM was trained, validated, and tested following a strict chronological protocol. All results accuracy metrics, ablation findings, and Diebold–Mariano significance tests were generated directly by executing the published code, not by simulation. The key conclusion is nuanced: the hybrid ensemble provides a statistically significant accuracy boost at a 1-day forecast horizon in most city/baseline pairs, degrades more gradually than a standalone recurrent model at 7 days, and benefits from both its tree and recurrent components, as confirmed by ablation. However, it doesn't outperform all baselines in every city, and LightGBM alone remains competitive in two cities. Permutation importance analysis shows that the model's predictive power stems from understandable features such as recent PM2.5 levels and rolling statistics, rather than opaque interactions, supporting cautious, human-reviewed operational deployment. By streamlining the regulatory narrative and expanding experimental protocol, dataset documentation, ablation analysis, and statistical testing, this revision makes such air-quality forecasts independently verifiable. The released code and detailed data provenance (Section 13) enable reviewers to reproduce every table and figure. Future work will expand the pipeline to sub-daily sensor data, more Indian and global cities, and employ SHAP for explainability of the deep model branch moving toward a forecasting system that is accurate, auditable, reproducible, and transparently bounded by documented limitations (Section 11).

REFERENCES

1. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.
2. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
5. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014* (pp. 1724–1734). Association for Computational Linguistics.
6. cp099. (2024). India-Air-Quality-Dataset [Data set]. GitHub. <https://github.com/cp099/India-Air-Quality-Dataset>
7. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
8. Du, S., Li, T., Yang, Y., & Horng, S.-J. (2021). Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 33(6), 2412–2424. <https://doi.org/10.1109/TKDE.2019.2954510>
9. Elbaz, K., Shaban, W. M., Zhou, A., & Shen, S.-L. (2023). Real time image-based air quality forecasts using a 3D-CNN approach with an attention mechanism. *Chemosphere*, 333, 138867. <https://doi.org/10.1016/j.chemosphere.2023.138867>
10. European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union*, L119, 1–88.
11. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*.
12. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
13. Hu, Y., Ding, Y., & Jiang, W. (2025). Geographically aware air quality prediction through CNN-LSTM-KAN hybrid modeling with climatic and topographic differentiation. *Atmosphere*, 16(5),

513. <https://doi.org/10.3390/atmos16050513>
14. IEEE. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (1st ed.). IEEE.
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* 30 (pp. 3146–3154).
16. Kumar, S., & Kumar, V. (2024). Multi-view stacked CNN-BiLSTM (MvS CNN-BiLSTM) for urban PM2.5 concentration prediction of India's polluted cities. *Journal of Cleaner Production*, 444, 141259. <https://doi.org/10.1016/j.jclepro.2024.141259>
17. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., & Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A*, 471(2182), 20150257. <https://doi.org/10.1098/rspa.2015.0257>
18. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774).
19. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273–1282). PMLR.
20. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1)*. U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
21. 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(1), 71. <https://doi.org/10.1186/s40537-024-00926-5>
22. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* 31 (pp. 6638–6648).
23. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
24. Sarkar, N., Gupta, R., Keserwani, P. K., & Govil, M. C. (2022). Air quality index prediction using an effective hybrid deep learning model. *Environmental Pollution*, 315, 120404. <https://doi.org/10.1016/j.envpol.2022.120404>
25. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
26. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). <https://doi.org/10.18653/v1/P19-1355>
27. United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development (A/RES/70/1)*. United Nations General Assembly.
28. Utku, A., Can, U., Alpsülün, M., Balıkcı, H. C., Amoozegar, A., Pilatin, A., & Barut, A. (2025). Advancing air quality monitoring: Deep learning-based CNN-RNN hybrid model for PM2.5 forecasting. *Atmosphere*, 16(9), 1003. <https://doi.org/10.3390/atmos16091003>
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (pp. 5998–6008).
30. World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization. <https://www.who.int/publications/i/item/9789240034228>
31. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting fine-grained air quality based on big data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2267–2276). ACM. <https://doi.org/10.1145/2783258.2788573>
32. Zhu, H., Ding, X., Wang, M., & colleagues. (2026). Air quality prediction model based on deep learning hybrid framework. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-37896-y>