# A Comparative Review of Attribute Selection Techniques for PM2.5 Prediction Using Machine Learning Models

**Dr. Sachin Arun Thanekar**

**Associate Professor, Computer Science and Engineering Department, MITADT University, Loni Kalbhor, Pune, India**

**DOI: https://dx.doi.org/10.51244/IJRSI.2025.1210000212**

**Received: 07 October 2025; Accepted: 14 October 2025; Published: 15 November 2025**

## ABSTRACT

Accurate prediction of fine particulate matter (PM2.5) is vital for understanding and mitigating air pollution impacts on public health. With the rise of machine learning (ML) in environmental forecasting, selecting the most influential features remains a critical preprocessing step. This review paper evaluates the effectiveness of various attribute selection techniques applied to PM2.5 prediction, including filter, wrapper, and embedded methods. We compare the results from Random Forest, LASSO regression, Recursive Feature Elimination (RFE), and correlation analysis. Our comparative analysis reveals that Random Forest consistently highlights meteorological variables such as temperature and wind speed as top contributors, whereas LASSO reduces model complexity by focusing on core pollutants. The paper provides insights for researchers aiming to develop robust and computationally efficient models for real-time PM2.5 forecasting.

**Keywords:** PM2.5, Feature Selection, Random Forest, LASSO, Machine Learning, Air Quality Prediction, Attribute Selection

## INTRODUCTION

A large set of features can be considered for use by machine learning models for PM2.5 prediction. These typically include meteorological parameters (temperature, relative humidity, wind speed, atmospheric pressure), concentration of other pollutants ($NO_2$, $SO_2$, CO, $O_3$), time-related variables (season, month, hour), location, and perhaps socio-economic indicators. However, using all available features does not guarantee better model performance. Irrelevant or redundant features may only add noise, increase the time cost due to overfitting, or even require additional computing resources. That is where attribute (feature) selection approaches become relevant. These methods try to select a smaller subset of features that are considered most informative and relevant, which improves interpretability, reduces dimensionality, and increases the prediction performance of the models. Proper feature selection not only streamlines the modeling process but also leads to the identification of drivers responsible for PM2.5 changes, which may be helpful information from a scientific and policy perspective. Chen, J., Li, S., Huang, G., & Liu, X. (2023), Wang, Y., Liu, Y., Wu, Y., & Zhang, L. (2022).

## REVIEW METHODOLOGY

This review followed a structured literature review approach inspired by the PRISMA guidelines for systematic reviews, adapted to the scope of a comparative methodological review. The steps were as follows:

**Database Selection:** We searched major academic databases, including IEEE Xplore, ScienceDirect, SpringerLink, and Google Scholar.

**Search Strategy:** Keywords such as “PM2.5 prediction,” “feature selection,” “attribute selection,” “machine learning,” and “air quality forecasting” were used in various Boolean combinations.

**Inclusion Criteria:** Studies were included if they (a) focused on PM2.5 or air quality prediction, (b) applied or compared feature/attribute selection techniques, and (c) reported quantitative performance metrics.

**Exclusion Criteria:** Papers were excluded if they (a) did not involve machine learning methods, (b) were purely theoretical without experiments, or (c) duplicated findings already covered in earlier reviews.

**Screening Process:** An initial set of papers was identified. After title/abstract screening and full-text evaluation, some papers were retained for final analysis.

**Data Extraction and Analysis:** For each selected study, details such as feature selection technique, dataset used, machine learning models applied, and performance metrics ($R^2$, RMSE, MAE, etc.) were extracted and compared.

**Importance of Attribute Selection**

Feature selection serves as an important preprocessing procedure of a machine learning pipeline that endeavors to identify and keep only the variables considered most useful and informative for the betterment of predictive models. The purpose of feature selection is to cast away irrelevant, redundant, or noisy attributes that do not contribute meaningfully to the target variable, here PM2.5 concentration levels, thus improving model efficiency, interpretability, and predictive accuracy. Sharma, A., & Saini, L. M. (2021).

Potential input variables considered for PM2.5 forecasting are indeed rather broad and often very high dimensional. Common inputs may include concentrations of pollutants such as $NO_2$, $SO_2$, CO, $O_3$, and PM10, meteorological parameters such as temperature, humidity, wind speed, wind direction, rainfall, and atmospheric pressure, temporal indicators such as time of day, day of the week, season, and month, as well as spatial information including geographic coordinates, altitude, land use patterns, and proximity to pollution sources. This heterogeneity of features represents the complexity and multifactorial nature of air pollution that gets influenced by both anthropogenic and natural causes. Ahmed, R. M., & Badr, A. (2021), Zhang, Y., Wang, J., & Wang, S. (2020).

However, not all these features contribute equally to the PM2.5 levels. Some carry redundant information (for example: PM10 and PM2.5 tend to be strongly correlated to each other), whereas some would introduce irrelevant noise or spurious relationships that would hamper the performance of a model. Greater inclusion of such features can be detrimental as the model learns to fit the noise in the training data, gives lower performance on new data, and is thus, overfitting in nature. Conversely, optimal feature selection has been known to help improve the generalizability of models, reduce computational cost, as well as the time necessary for training and inference-speed-two great issues in real-time air quality monitoring systems.

Considering how multidimensional and often sputter-co-the dimension does air quality data sets, feature selection could wait not only be beneficial but necessary. Such high-dimensional data could almost serve to mask the very patterns one would want to detect, complicating analysis and validation strategies. That conundrum worsens, as different regions and climates might demand different subsets of features for their best performance, which highlights the necessity for dynamic, context-aware feature selection mechanisms.

According to Guyon and Elisseeff (2003), feature selection is essential to improve learning algorithms, especially when datasets are such that the number of features exceeds the number of observations-many scenarios typical of environmental datasets that cover many seasons or multiple locations. They classify feature selection methods into three broad categories: Filter methods (independent of any learning algorithm), Wrapper methods (which use a model performance measure to evaluate the quality of a feature subset), and Embedded methods (where feature selection is performed as part of model construction).

Chandrashekar and Sahin (2014) went further to say that feature selection helps in reducing the dimensionality of data, thereby simplifying models and improving their interpretability, which is an exigency in policy-related contexts like the management of air quality. Their studies indicate that through the right selection of features, domain specialists are capable of better understanding the relationships linking environmental factors and pollution levels, leading to more informed decision-making.

A sound PM2.5 prediction model will weigh its feature selection with care. Feature selection creates a compromise between complexity and performance, making sure that the model is workable and prolific. With a growing number of machine learning techniques becoming an integral part of environmental monitoring systems, the significance of an intelligent and well-justified assessment of feature selection shall multiply, resulting in more intelligent, faster, and transparent forecasts of air quality.

### Commonly Used Attribute Selection Techniques

### Filter Methods

Filter methods assess the relevance of features by examining their statistical relationship with the target variable.

**Correlation Analysis**: Correlation Analysis is a statistical technique that considers the strength and direction of a linear relationship between two continuous variables. In air quality prediction, the Pearson correlation is mostly applied in determining how climate variables and pollutants are varied in relation to certain target variables, such as PM2.5 concentrations. This classification method is very powerful in the initial stage of feature selection because it eliminates those variables that are irrelevant or weakly correlated and thus simplifies the modeling process. For instance, one study utilized the Pearson correlation analysis technique to identify meaningful positive and negative correlations between PM2.5 levels and other variables such as $NO_2$, temperature, and wind speed. Consequently, these environmental factors should be deemed important for particulate matter concentration modeling (Hu et al., 2013).

**Chi-square Test:** The Chi-square Test is a statistical test for assessing the associations between two categorical variables by comparing the observed and expected frequency distribution in a contingency table and is thus useful in feature selection for testing a given feature's independence from the target variable, e.g., to evaluate whether a certain feature is informative for a classification assignment. However, with most features in air quality datasets being continuous variables, such as pollutant concentrations and meteorological indices (e.g., PM2.5 concentration, temperature, humidity), the Chi-square Test becomes less applicable. In order to make use of this methodology for AQI-related studies, all continuous variables in the study would require conversion into categorical bins, which will pose threats to the retention of useful information and further jeopardize predictive performance. Thus, this technique has little relevance in the case of continuous AQI data (Hu et al., 2013).

### Wrapper Methods

Wrapper techniques evaluate subsets of features based on model performance.

**Recursive Feature Elimination (RFE):** Recursive Feature Elimination is a wrapper-type feature selection tool that recursively fits a model and removes the least important features until the desired number of features to select is finally reached, thereby stripping away irrelevant or redundant inputs. Because these approaches can easily rank the features by their importance, RFE is very useful for linear modeling (SVM, logistic regression) and tree-based models (decision trees, random forests). By gradually pruning out the weaker predictors, RFE reduces overfitting and improves generalization and accuracy of a model by removing noise from the dataset. In the air quality prediction framework, RFE filters out some meteorological and pollutant variables with higher influence and performance tied to the predictive model (Guyon et al., 2002).

**Sequential Feature Selection:** Sequential Feature Selection builds an optimal feature list by incrementally adding or removing one feature at a time and evaluating the resulting feature subset using a learning algorithm. Forward selection would start with no features and keep adding whichever feature most improves the model at each step, whereas backward elimination would proceed in the opposite manner by iteratively throwing away the least impactful features. It is a model-dependent technique; thus, it discriminates between feature subsets by way of a given learning algorithm and allows optimization suited to the model's reaction to various combinations. Given that feature numbers are not too large,Sequential Feature Selection should prove fruitful in recognizing those variables most relevant to augmenting accuracy, robustness, and interpretability for regression modeling. Such an exhaustive feature study reduces model complexity such that predictive strength is either maintained or enhanced.

### Embedded Methods

Embedded methods integrate feature selection as part of the model training process.

**Random Forest Feature Importance:** Random Forest Feature Importance is considered an embedded method since it is executed during the Random Forest training stage for ranking the relevance of input features based on how they contribute to predicting accuracy. The Random Forest, a tree ensemble model, builds several decision trees whose outputs are amalgamated to yield a more robust result and guard against overfitting. During training, the algorithm sums each feature's capacity to reduce impurity (Gini impurity or entropy, for example) across all trees, assigning greater importance to features that help shrink the prediction error more significantly. For air quality modeling purposes, Random Forest has been especially efficient in distinguishing between key meteorological variables like temperature and humidity, which substantially affect PM2.5 concentration levels. This user-friendly feature importance ranking built into Random Forest makes it a great medium for both feature selection and environmental data prediction modeling (Liaw & Wiener, 2002).

**LASSO Regression:** The name LASSO stands for Least Absolute Shrinkage and Selection Operator and is embedded feature selection with an intervention introducing L1 regularization into the regression model that penalizes the absolute magnitude of the regression coefficient. Through the regularization, the coefficients are forced to become zero for variables deemed relatively unimportant or redundant, resulting in a selection of features during the very training of the model. With high-dimensional data sets and many variables, the output generated by LASSO can be very helpful, deciding to keep only those predictors which are most relevant, thus making interpretation easier and alleviating the problem of overfitting. LASSO is used in air quality prediction to discard less important pollutant variables while preserving the model's performance, hence making it the appropriate means for developing very sparse but still acceptably accurate predictions (Tibshirani, 1996).

Table 1: Classification Chart: Attribute Selection Techniques

| Method Type | Technique | Approach | Data Type Suitability | Example Usage |
|---|---|---|---|---|
| **Filter Methods** | Pearson Correlation | Statistical correlation with target | Continuous | Identifying linear relationships (e.g., PM2.5 vs. $NO_2$, temperature, wind speed) |
| | Chi-square Test | Chi-square score for categorical relevance | Categorical | Feature relevance in classification (less relevant for AQI) |
| **Wrapper Methods** | Recursive Feature Elimination (RFE) | Iteratively removes features and evaluates | Continuous/Categorical | Enhancing model accuracy using SVM, Random Forest |
| | Sequential Feature Selection | Adds/removes features step-by-step | Continuous/Categorical | Forward/Backward selection using custom scoring metrics |
| **Embedded Methods** | Random Forest Feature Importance | Built-in importance scoring during training | Mostly continuous | Ranking meteorological variables like humidity, temperature |
| | LASSO Regression | L1 regularization penalizes weak features | Continuous | Eliminating redundant variables while maintaining model performance |

## Comparative Evaluation

The ability of feature selection methods was assessed in terms of being able to interpret the model, predictive abilities measured by R2-RMSE, and computational costs. Results indicate:

Random Forest generates increased robustness with regard to feature importance, exploiting the ensemble learning structure that sets up a plethora of decision trees and amalgamates their outputs to be able to predict with a higher degree of stability. This process excels in managing the complexities of nonlinear interactions among features; hence, it would be a top choice for datasets where relationships between variables are seldom linear, such as the cases of air quality prediction. By looking at the role each feature plays in reducing impurity on all trees and ranking them on the basis of this, Random Forest can distinguish between meteorological and pollutant factors that have the greatest effect on the outcome and, thus, assist with selections to improve the model and interpretability. Its analysis of high-dimensional data and uncovering of hidden patterns makes Random Forest a distinct contender among other feature selection methods of environmental modeling criteria (Liaw & Wiener, 2002).

LASSO regression simplified the models by accomplishing retardation of coefficients through its L1 regularization technique that penalizes the real-valued present form of regression coefficients. Actual penalization forcibly drives the values of regression coefficients of trivial and highly correlated variables to zero, producing a sparse model that keeps really relevant features. In doing so, it excludes all inputs that may redundantly explain the variance in a dependant variable, PM2.5 concentration being one of those in surface-air quality forecasting. This dimensionality reduction technique enhances model interpretability and computational efficiency and lowers the chances of overfitting, especially for high-dimensional datasets. More importantly, LASSO does this without notably detracting from the predictive performance; hence the two most suitable properties for lightweight and robust modeling approaches to be used in real-time or resource-limited environments (Tibshirani, 1996).

The Recursive Feature Elimination technique increased model predictability moderately, systematically selecting features that rank least important in an iterative manner, wherein one iteration involves training the model, evaluating those features for importance, and braking the least supportive contributor for it to possibly improve the accuracy while decreasing overfitting due to the quality of being highly informative. Its computational intensity rendered it paradoxical to use for real-time or large-scale applications since it required a higher number of rounds for training and evaluation to settle on an optimum feature list, whereas RFE proved to be useful in pruning the final feature sets, therefore increasing the generalization of the models (Guyon et al., 2002).

## Quantitative Synthesis of Comparative Studies

To substantiate the comparative analysis, we compiled results from representative prior studies that applied feature selection techniques for PM2.5 prediction. The synthesis focuses on (a) datasets used, (b) feature selection methods applied, (c) machine learning models tested, and (d) performance metrics such as $R^2$, RMSE, and MAE.

Table 2: Datasets and Feature Selection Techniques in Prior PM2.5 Prediction Studies

| Study | Dataset (Region/Size) | Feature Selection Method | Model(s) Used | Key Features Selected |
|---|---|---|---|---|
| Chen et al. (2023) | China, multi-city air quality dataset | Hybrid FS (Correlation + RFE) | Random Forest, XGBoost | Temperature, humidity, $NO_2$, $SO_2$ |
| Wang et al. (2022) | Beijing AQI (5 years, hourly) | Random Forest Importance | RF, LSTM | Meteorological + pollutant mix |

| | | | | |
|---|---|---|---|---|
| Sharma & Saini (2021) | Delhi AQI (3 years) | LASSO Regression | SVR, ANN | $NO_2$, CO, humidity |
| Ahmed & Badr (2021) | Cairo (hourly, 2 years) | Wrapper (Sequential FS) | RF, SVM | PM10, wind speed, $O_3$ |
| Zhang et al. (2020) | China (multi-city, daily) | Hybrid Feature Selection (SSA + FS) | BiGRU, RF | Pollutant concentrations + temperature |

Table 3: Performance Comparison of Feature Selection Techniques

| **Study** | **Method** | **Model** | **$R^2$** | **RMSE** | **MAE** | **Key Findings** |
|---|---|---|---|---|---|---|
| Chen et al. (2023) | Correlation + RFE | RF | 0.87 | 12.4 | 9.6 | RFE improved robustness over baseline |
| Wang et al. (2022) | RF Importance | LSTM | 0.91 | 10.2 | 7.8 | RF Importance improved temporal prediction |
| Sharma & Saini (2021) | LASSO | SVR | 0.84 | 14.1 | 10.9 | LASSO reduced overfitting in high-dim data |
| Ahmed & Badr (2021) | Sequential FS | RF | 0.79 | 15.7 | 11.5 | Sequential FS enhanced interpretability |
| Zhang et al. (2020) | Hybrid FS | BiGRU | 0.93 | 9.5 | 6.7 | Hybrid FS gave best accuracy, esp. with deep learning |

### Critical Discussion of Feature Selection Techniques

While feature selection methods such as Random Forest, LASSO, and RFE are widely applied, their relative strengths and weaknesses depend strongly on dataset characteristics, forecasting objectives, and computational constraints.

### Random Forest Feature Importance

Random Forest excels in capturing **nonlinear interactions** and **complex variable dependencies**, which are common in air quality data where meteorological and pollutant factors interact in nonlinear ways (Wang et al., 2022). Empirical evidence shows that Random Forest consistently identifies meteorological variables (e.g., humidity, temperature, wind speed) as key predictors. However, the method may **overemphasize correlated features**, inflating their importance scores. Computationally, Random Forest is more demanding than filter-based methods but remains tractable for medium-to-large datasets, making it well-suited for **regional forecasting tasks** where interpretability and robustness are critical.

### LASSO Regression

LASSO is particularly effective when the dataset is **high-dimensional with many redundant or irrelevant variables**. By shrinking coefficients of weak predictors to zero, it produces **sparse, interpretable models** (Sharma & Saini, 2021). Unlike Random Forest, it handles correlated predictors by selecting only one of them, reducing redundancy. However, it assumes **linear relationships**, which may limit its predictive performance when pollutant–meteorological interactions are nonlinear. LASSO is computationally lightweight, making it suitable for **real-time forecasting on constrained devices**.

### Recursive Feature Elimination (RFE)

RFE systematically removes less important features and can improve generalization, but it is **computationally expensive** due to repeated model training (Guyon et al., 2002). This makes it impractical for **large-scale or real-time systems**, though it remains valuable for **offline model refinement** and when the dataset size is modest.

### Hybrid / Sequential Feature Selection

Empirical studies (e.g., Chen et al., 2023; Zhang et al., 2020) demonstrate that hybrid approaches combining filter and embedded methods often outperform single-method approaches. These methods balance interpretability and accuracy, particularly when applied with deep learning models such as BiGRU or LSTM. The trade-off is increased **algorithmic complexity and implementation cost**, which may not be practical in low-resource monitoring systems.

### Comparative Insights:

Random Forest outperforms LASSO when **nonlinearities and feature interactions dominate**, such as in multi-city or seasonal datasets.

LASSO outperforms Random Forest when the **goal is simplicity, interpretability, or real-time deployment** under resource constraints.

RFE is useful for **smaller datasets** where computational cost is manageable, but less scalable to big-data scenarios.

Hybrid methods are most effective in **research settings** aiming for the highest predictive accuracy, but they require more computational resources.

### Discussion of Quantitative Results

Random Forest and LASSO consistently balance predictive power and interpretability.

RFE improves model robustness but is computationally heavy.

Hybrid feature selection (statistical + model-based) often outperforms single-method approaches, especially with deep learning models.

Across studies, RMSE reductions of 10–20% were reported when applying feature selection compared to using raw features.

### Challenges and Limitations

Despite the progress in applying feature selection techniques to PM2.5 prediction, several challenges and limitations remain evident across the reviewed literature:

### Dataset Heterogeneity and Limited Generalizability

Most studies rely on datasets from specific regions (e.g., Beijing, Delhi, Cairo), with distinct climatic and pollutant profiles. As a result, feature subsets identified as important in one region may not generalize to another, particularly across different climates or levels of industrialization. This geographic dependency limits the transferability of feature selection outcomes.

### Assumptions of Static Feature Importance

Many approaches assume that the relative importance of features remains constant over time. However, empirical findings indicate that meteorological drivers (e.g., humidity, wind speed) and pollutant interactions vary across seasons and long-term policy interventions. Static selection may therefore fail to capture evolving dynamics, leading to reduced accuracy in long-term forecasting.

### Computational Costs and Real-Time Constraints

Techniques such as Recursive Feature Elimination and hybrid feature selection often require repeated training cycles, making them computationally expensive. While feasible for offline experiments, these methods are difficult to deploy in real-time air quality monitoring systems that demand rapid inference and resource efficiency.

### Model-Specific Biases

Embedded methods such as Random Forest and LASSO inherently reflect the biases of their underlying models. Random Forest may overemphasize correlated variables, while LASSO favors sparsity but assumes primarily linear relationships. This model-dependency can skew feature selection outcomes, potentially overlooking relevant predictors in complex nonlinear contexts.

### Lack of Standardized Benchmarks

Performance evaluation across studies is often inconsistent, with varied metrics ($R^2$, RMSE, MAE) and non-uniform datasets. The absence of benchmark datasets and standardized protocols hinders direct comparison between techniques and weakens the evidence base for determining best practices.

### Limited Integration with Deep Learning and Spatiotemporal Models

Although deep learning architectures (e.g., BiGRU, LSTM) are increasingly used for PM2.5 prediction, most feature selection studies are confined to traditional ML models. The challenge lies in adapting feature selection frameworks to high-dimensional, spatiotemporal inputs that characterize modern environmental datasets.

## Conclusion and Future Work:

This review highlights that while feature selection significantly improves PM2.5 prediction, most existing studies operate under implicit assumptions that limit generalizability. For instance, many works assume that **the same feature subset remains optimal across all regions and seasons**, whereas empirical evidence suggests that meteorological drivers vary with geography and time. Similarly, several studies assume **linear relationships** between pollutants and PM2.5, which may overlook nonlinear dynamics observed in urban air quality.

### Concrete gaps include:

Limited exploration of **cross-regional transferability** of selected features.

Insufficient attention to **computational trade-offs**, with many studies prioritizing accuracy while ignoring deployment constraints in real-time systems.

A lack of **dynamic or adaptive feature selection mechanisms** that update feature importance as environmental conditions shift (e.g., seasonal changes, policy interventions reducing pollutant levels).

### Future research should therefore:

Develop **dynamic, context-aware feature selection algorithms** that adapt feature sets based on spatiotemporal variability.

Investigate **lightweight hybrid approaches** that balance predictive accuracy with computational efficiency for real-time monitoring.

Explore **transfer learning and domain adaptation** to test whether feature sets selected in one region can generalize to others.

Establish **benchmark datasets and standardized evaluation metrics** to enable fairer comparisons across studies.

By addressing these gaps, future work can move toward intelligent, robust, and scalable feature selection frameworks that better support real-world air quality forecasting.

## REFERENCES

1. Chen, J., Li, S., Huang, G., & Liu, X. (2023). A robust hybrid feature selection method for high-dimensional environmental data analysis. Environmental Modelling & Software, 160, 105620.
2. Wang, Y., Liu, Y., Wu, Y., & Zhang, L. (2022). Feature selection and ensemble learning for air quality prediction: A case study in China. Ecological Indicators, 139, 108930.
3. Sharma, A., & Saini, L. M. (2021). Air quality prediction using optimized feature selection and machine learning algorithms. Applied Soft Computing, 113, 107872.
4. Ahmed, R. M., & Badr, A. (2021). FSFC-AQ: Feature selection for forecasting city-level air quality using hybrid approaches. Environmental Science and Pollution Research, 28, 28837–28852.
5. Zhang, Y., Wang, J., & Wang, S. (2020). Air quality prediction based on SSA optimized BiGRU neural network with hybrid feature selection. Science of The Total Environment, 721, 137763.
6. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.
7. Gaurav, R., & Behera, H. S. (2021). Hybrid feature selection model for predicting air quality index. Environmental Processes, 8(2), 883–902.
8. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157–1182.
9. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.
10. Hu, X. M., Klein, P. M., & Xue, M. (2013). Evaluation of the updated YSU planetary boundary layer scheme within WRF for air quality simulations in the Houston–Galveston area. Atmospheric Environment, 92, 274–283.
11. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.
12. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.