Adaptive Statistical–Entropy Preprocessing (ASEP): A Novel Framework for Noise-Resilient Clinical Data Modeling

Divya.M[1], Maheswari.C[2]

[1] *Assistant Professor, Department of Computer Science and Engineering, Peri Institute of Technology, Mannivakkam. peri.divya2025@gmail.com*

[2] *Assistant Professor, Department of Computer Technology, Peri Institute of Technology, Mannivakkam. sgnmahesh82@gmail.com*

Abstract:- Accurate air quality prediction depends heavily on the reliability of data preprocessing, as environmental datasets often contain noise, missing values, and inconsistent measurements. This study introduces a novel Adaptive Statistical–Entropy Preprocessing (ASEP) framework that enhances data quality before model training. The algorithm integrates entropy-based imputation, variance-guided noise removal, distribution-sensitive normalization, and correlation entropy-driven feature pruning to improve data uniformity and signal clarity. Using publicly available air quality data, ASEP demonstrates significant improvements in balance, consistency, and feature diversity compared to existing methods. The framework effectively minimizes redundancy and noise while preserving meaningful variability within pollutant readings. Designed to be scalable and interpretable, ASEP offers a robust and computationally efficient preprocessing solution for real-time air quality monitoring and prediction. Its results highlight the importance of adaptive, data-driven preprocessing as a foundation for accurate and sustainable environmental modeling.

Keywords*: Air Quality Prediction, Adaptive Preprocessing, Entropy-Based Analysis, Data Enhancement, Environmental Modeling.*

1. Introduction

Air pollution has become one of the most pressing environmental concerns of the modern era, posing severe threats to human health, ecosystems, and climate stability. The accurate monitoring and prediction of air quality are essential for developing effective mitigation strategies and supporting sustainable urban planning. The Air Quality Index (AQI) serves as a standardized measure for communicating pollutant concentrations, including particulate matter ($PM_{2.5}$, PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and ozone (O_3). However, achieving reliable AQI prediction remains challenging due to the inherent complexity of atmospheric data, characterized by missing values, noise, and temporal inconsistencies (Ansari & Quaff, 2025).

Recent advancements in data-driven modeling have significantly improved AQI forecasting accuracy through machine learning (ML) and deep learning approaches. These models have been applied across local and global contexts, enabling automated pattern detection and pollutant behavior analysis (Anggraini et al., 2024). Despite their success, the predictive performance of such models is fundamentally dependent on the quality of the input data. Raw environmental datasets often contain missing or corrupted sensor readings caused by equipment failures, climatic variability, or transmission errors. Incomplete or noisy data distort statistical relationships, leading to reduced generalization and inaccurate predictions (Maltare & Vahora, 2023).

Effective data preprocessing is therefore a critical foundation for reliable AQI forecasting. Conventional methods, such as mean imputation, z-score normalization, and static feature scaling, are widely used but lack adaptability to dynamic air quality patterns. These methods often treat all features uniformly, disregarding pollutant-specific variability or inter-feature correlations, which can bias model training (Li & Sun, 2023). In particular, temporal and spatial heterogeneity in air quality data demands preprocessing strategies that can adapt to varying distributions and noise structures. As a result, recent research has shifted toward hybrid or optimization-based approaches to improve preprocessing efficiency and robustness.

For instance, AMLT-AQI proposed by (Ansari & Quaff, 2025) integrates multiple ML techniques—including XGBoost, Random Forest, and LSTM—using optimized ensemble learning for hourly AQI forecasting. Similarly, AQIPD, developed by (Maltare & Vahora, 2023), utilized regional pollutant datasets from Ahmedabad to build prediction models with random forest and gradient boosting, achieving improved urban air quality forecasting. At a global scale, MLRS-AQI, presented by (Anggraini et al., 2024), incorporated satellite-based remote sensing and ground monitoring data to build an international AQI prediction framework. While these studies demonstrate progress in improving predictive accuracy, they rely on static or manually tuned preprocessing pipelines that are not adaptable to fluctuating environmental conditions.

The limitations of conventional preprocessing become particularly evident in datasets with high noise levels and complex correlations among pollutants. Bayesian and probabilistic filtering methods, such as those proposed by (Li & Sun, 2023), improved noise management but introduced computational complexity and dependency on prior assumptions. Hybrid models that combine statistical and learning-based methods have also emerged, but they often prioritize model optimization rather than preprocessing refinement (Sharma & Patel, 2025). As a result, there is a clear gap in the literature for a preprocessing framework that can adaptively adjust to data distribution, handle uncertainty, and reduce redundancy while maintaining computational efficiency.

To address these challenges, this study introduces the Adaptive Statistical–Entropy Preprocessing ASEP algorithm—a novel preprocessing framework that integrates statistical adaptability with entropy-based feature evaluation. ASEP systematically improves dataset quality by performing entropy-guided missing value imputation, variance-based noise filtering, distribution-sensitive normalization, and correlation entropy-driven feature pruning. The proposed approach enables adaptive correction of data irregularities while preserving meaningful pollutant variability, ultimately leading to enhanced input stability for downstream predictive models.

The objectives of this study are threefold:

1. To design and implement a preprocessing framework that adaptively responds to statistical and entropy-based indicators of data quality.

2. To compare the proposed ASEP method with three state-of-the-art AQI prediction frameworks—AMLT-AQI, AQIPD, and MLRS-AQI—on publicly available environmental datasets.

3. To demonstrate that preprocessing quality directly influences model performance and generalizability in air quality forecasting.

By introducing a dynamic, statistically grounded, and interpretable preprocessing framework, ASEP bridges the methodological gap between data quality management and predictive modeling. The following sections present a comprehensive review of related studies Section 2, describe the proposed methodology Section 3, and evaluate ASEP’s performance against existing techniques (Section 4), followed by key conclusions and future research directions Section 5.

2. Related works

Ansari & Quaff, (2025) proposed the Advanced Machine Learning Techniques for Air Quality Index prediction (AMLT-AQI) framework, which integrates ensemble algorithms such as XGBoost, Random Forest, and LSTM for precise hourly AQI forecasting. Their approach effectively captured nonlinear and temporal dependencies between pollutants, resulting in highly accurate predictions. The study demonstrated the potential of combining multiple learning models for environmental forecasting. However, the framework depended heavily on extensive parameter tuning and lacked adaptive preprocessing mechanisms, reducing its flexibility for varying regional datasets. The research established a strong foundation for ensemble-based air quality modeling but emphasized the need for more automated preprocessing solutions.

Maltare & Vahora, (2023) developed the AQIPD framework to predict air quality in Ahmedabad using Random Forest and Gradient Boosting techniques. Their system achieved good performance in capturing pollutant-specific variations and provided reliable city-level AQI forecasting. The study demonstrated that integrating multiple pollutant features can enhance predictive accuracy in localized environments. However, the preprocessing steps used were static, offering limited adaptability to seasonal or temporal fluctuations in pollutant patterns. The absence of dynamic noise filtering and distribution adjustment reduced its robustness for long-term or multi-city datasets.

Anggraini et al., (2024) introduced the Machine Learning-Based Global Air Quality Index (MLRS-AQI) framework, which combined satellite-based remote sensing data with ground-based sensor readings using regression stacking. The model enabled global-scale AQI estimation across diverse geographic regions and atmospheric conditions. It successfully bridged the gap between remote sensing and terrestrial monitoring, enhancing spatial coverage. Nevertheless, the integration of heterogeneous data sources introduced feature redundancy and calibration inconsistencies. The framework highlighted the scalability of global AQI modeling but underscored the importance of adaptive preprocessing for harmonizing large, multi-source datasets.

Li & Sun, (2023) proposed a Bayesian noise filtering approach for environmental datasets to improve the reliability of pollution monitoring. The method modeled data uncertainty through probabilistic inference, enabling the estimation of true pollutant values under noisy sensor conditions. Their framework achieved substantial improvements in data consistency and reduced measurement variance. However, it required significant computational resources and prior distribution knowledge, limiting its suitability for large-scale, real-time applications. The study demonstrated the potential of probabilistic noise handling while exposing the trade-off between accuracy and efficiency in complex preprocessing systems.

Chen et al., (2024) presented a Temporal Fusion Network (TFN) that integrates attention mechanisms with recurrent neural architectures for pollutant trend prediction. Their approach captured both short- and long-term dependencies in atmospheric variables, resulting in enhanced temporal forecasting. The model effectively combined multiple time-series inputs, demonstrating superior accuracy for multivariate AQI prediction. However, TFN required extensive training data and computational capacity, restricting its practical implementation for smaller datasets. The preprocessing process was also static, indicating the need for adaptable preprocessing prior to deep temporal modeling.

Kumar & Sinha (2023) explored hybrid feature extraction techniques for atmospheric data analysis, integrating Principal Component Analysis (PCA) with mutual information. Their framework effectively reduced dimensionality while maintaining the interpretability of significant features. This method improved model training efficiency and provided insight into pollutant dependencies. Nonetheless, PCA’s linear transformation often disregarded nonlinear feature interactions, and the absence of adaptive entropy-based filtering limited responsiveness to dynamic datasets. The research emphasized the value of hybrid feature selection but highlighted the importance of adaptive mechanisms for modern AQI systems.

Zhou et al., (2024) developed an explainable artificial intelligence framework for air quality forecasting by incorporating SHAP (SHapley Additive exPlanations) analysis. The system provided interpretability to ML-based AQI prediction models, revealing the influence of individual pollutants on overall air quality. The method enhanced transparency and trust in AI-driven environmental systems. However, the approach focused primarily on post-hoc interpretation rather than preprocessing optimization. Consequently, raw data inconsistencies and redundancy persisted, reducing the overall model’s resilience under fluctuating input conditions.

Gupta et al., (2022) introduced a noise-aware regression framework designed to manage uncertainty and variability in meteorological datasets. Their model incorporated regularization techniques that penalized high-variance data points, improving prediction stability in noisy environments. The approach significantly reduced error propagation across training iterations. However, it did not handle missing data or adaptive normalization, which are crucial for maintaining statistical balance in environmental datasets. The study contributed valuable insights into robust regression but left scope for preprocessing-driven generalization improvements.

Sharma & Patel (2025) proposed an adaptive scaling and transformation method for environmental datasets that adjusted normalization dynamically based on pollutant variance. This technique stabilized data inputs and improved convergence rates in downstream predictive models. The adaptive scaling reduced distortions caused by fluctuating pollutant concentrations, enhancing the overall stability of AQI forecasting. Nonetheless, the approach lacked integrated noise filtering and redundancy analysis, making it less comprehensive for preprocessing complex, multi-source datasets. The study paved the way for adaptive normalization but required further enhancement for complete preprocessing automation.

Noor & Reddy (2023) developed a hybrid correlation pruning and oversampling strategy combining correlation analysis with the Synthetic Minority Oversampling Technique (SMOTE). Their framework improved class distribution in categorical AQI prediction and reduced feature redundancy. The method enhanced classification fairness and balanced minority classes effectively. However, synthetic sample generation occasionally introduced distortions and noise, which affected prediction accuracy. The research demonstrated the utility of combining statistical pruning with resampling techniques, emphasizing the need for distribution-aware preprocessing.

Ahmed & Liu (2024) investigated an entropy-based data reconstruction method for atmospheric monitoring systems. Their model used entropy minimization to recover missing pollutant values and maintain the natural variability of sensor data. The approach improved data integrity while reducing information loss during imputation. Despite these strengths, the framework did not integrate noise filtering or redundancy pruning, which limited its scope as a complete preprocessing pipeline. The study established entropy as a valuable tool for adaptive data refinement, aligning conceptually with modern statistical preprocessing frameworks.

Table 1. Summary of Recent Research on Air Quality Prediction and Preprocessing

Title	Author & Year	Methodology	Key Contribution	Limitation
Advanced Machine Learning Techniques for AQI Prediction in Azamgarh, India (AMLT-AQI)	Ansari & Quaff (2025)	Ensemble learning using XGBoost, Random Forest, and LSTM	High precision in hourly AQI forecasting through multi-model integration	Requires heavy tuning and static preprocessing
Air Quality Index Prediction using Machine Learning for Ahmedabad City (AQIPD)	Maltare & Vahora (2023)	Gradient Boosting and Random Forest with pollutant-specific analysis	Reliable city-level AQI prediction and pollutant trend identification	Limited adaptability to seasonal changes
Machine Learning-Based Global Air Quality Index Development (MLRS-AQI)	Anggraini et al., (2024)	Regression stacking combining satellite and ground-based data	Scalable AQI prediction at global level with wide spatial coverage	Redundant features and calibration inconsistencies
Bayesian Noise Filtering in Environmental Data	Li & Sun (2023)	Probabilistic noise modeling using Bayesian inference	Reduced measurement noise and improved data consistency	High computation and dependency on prior distribution

Temporal Fusion Networks for Pollution Forecasting	Chen et al., (2024)	Attention-based deep learning with recurrent layers	Enhanced temporal forecasting by modeling short- and long-term dependencies	Requires large training data and computational resources
Hybrid Feature Extraction for Atmospheric Data	Kumar & Sinha (2023)	PCA with mutual information for dimensionality reduction	Improved interpretability and reduced dimensionality	Ignores nonlinear dependencies among features
Explainable AI for Urban Pollution Mapping	Zhou et al., (2024)	SHAP-based interpretability applied to AQI models	Improved transparency in ML-based pollutant contribution analysis	Post-hoc analysis; lacks preprocessing optimization
Noise-Aware Regression in Meteorological Data	Gupta et al., (2022)	Regularized regression incorporating uncertainty modeling	Reduced prediction bias and improved stability under noisy data	Does not address missing values or normalization
Adaptive Scaling and Transformation for AQI Prediction	Sharma & Patel (2025)	Variance-based adaptive normalization method	Stabilized pollutant data and improved model convergence	Lacks feature pruning and noise removal
Hybrid Correlation Pruning and Oversampling Technique	Noor & Reddy (2023)	Correlation-based pruning with SMOTE oversampling	Enhanced class balance and reduced redundancy	Risk of synthetic distortion in minority classes
Entropy-Based Data Reconstruction for Air Quality Monitoring	Ahmed & Liu (2024)	Entropy minimization for missing value recovery	Improved imputation and preserved data variability	No integrated noise filtering or redundancy control

3. Proposed Methodology

3.1 Overview

The proposed Adaptive Statistical–Entropy Preprocessing (ASEP) framework is designed to improve data quality and model reliability for air quality prediction. Environmental datasets often suffer from missing readings, noise interference, and redundant features caused by diverse sensing conditions. These irregularities can lead to biased model training and inaccurate Air Quality Index (AQI) predictions. The ASEP algorithm integrates statistical and entropy-based principles to adaptively refine raw environmental data before it is used for model development. By dynamically adjusting imputation, normalization, and filtering processes based on the underlying data distribution, ASEP ensures improved data consistency, reduced redundancy, and enhanced interpretability.

3.2 Dataset Description

This study utilizes the UCI Air Quality Dataset, a publicly available dataset widely used in environmental data research. It contains 9,358 hourly observations collected from an air quality monitoring station located in Italy.

Each record includes sensor readings for major pollutants such as carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene (C_6H_6), nitrogen oxides (NOx), nitrogen dioxide (NO_2), and additional environmental parameters like temperature, humidity, and absolute pressure.

The dataset consists of 13 features in total, with approximately 7.5% missing values resulting from sensor malfunction and environmental disturbances. Moreover, several features exhibit non-normal distributions and class imbalance between different pollution levels. These characteristics make the dataset suitable for evaluating preprocessing algorithms. Before implementing ASEP, preliminary analysis revealed skewness in pollutant concentration distributions and strong correlations between certain chemical indicators (e.g., CO and NMHC). These challenges underscore the need for a dynamic, adaptive preprocessing approach capable of maintaining statistical balance while minimizing information loss.

3.3 Adaptive Statistical–Entropy Preprocessing (ASEP) Framework

The ASEP framework consists of five sequential modules:

3.3.1 Entropy-Based Missing Value Reconstruction

Missing readings in pollutant data can distort statistical patterns and bias model outcomes. ASEP employs an entropy-guided imputation method that restores missing values by minimizing information divergence. For each feature f_i, Shannon entropy $H(f_i) = -\sum p(x) \log\, p(x)$ is computed, and the imputed values are selected such that they maintain minimal deviation from the feature's overall entropy profile. This ensures that the reconstructed values preserve natural data variability while maintaining statistical coherence.

3.3.2 Variance-Guided Noise Filtering

Air quality sensors frequently produce noisy readings due to environmental fluctuations or equipment drift. To address this, ASEP applies adaptive variance-based noise filtering, where outlier thresholds dynamically adjust according to feature dispersion. The filtering threshold $\theta = \mu + 1.5\sigma$ identifies and replaces outliers beyond this range with statistically consistent values derived from neighborhood means. This method preserves natural variation while effectively removing extreme anomalies.

3.3.3 Distribution-Sensitive Normalization

Normalization ensures that features contribute proportionally to learning algorithms. ASEP introduces a distribution-sensitive normalization approach that selects transformation techniques based on the skewness of each variable. Features with high skewness (|skew| > 1) undergo logarithmic transformation, while near-normal features are standardized using z-score normalization. This hybrid approach maintains proportionality across pollutant features and prevents bias toward dominant variables.

3.3.4 Correlation Entropy–Based Feature Pruning

Redundant features can lead to overfitting and reduced interpretability in AQI prediction models. ASEP utilizes correlation entropy to measure redundancy between features. The correlation entropy for a feature pair (f_i, f_j) is defined as $E_c = -\sum p_{ij} \log\, p_{ij}$ where p_{ij} represents the joint probability distribution of correlated values. Highly correlated feature pairs with $r > 0.85$ and low entropy contributions are pruned, ensuring that only unique and informative attributes are retained for modeling.

3.3.5 Adaptive Resampling Using SMOTE-Entropy Balancing

To handle class imbalance in pollution severity levels, ASEP integrates an entropy-constrained Synthetic Minority Oversampling Technique (SMOTE). This module generates synthetic samples for underrepresented AQI categories by considering both spatial proximity and entropy similarity. The process ensures that new samples align with the natural feature distribution, maintaining the authenticity of pollutant relationships.

3.4 ASEP Algorithm Pseudocode

```
Input: Raw dataset $D = \{f_1, f_2, \dots, f_n\}$
Output: Preprocessed dataset $D'$
Algorithm Steps:
        For each feature$f_i$, compute Shannon entropy $H(f_i)$.
        If missing values exist, impute by minimizing $| H(f_i^{complete}) - H(f_i^{imputed}) |$
        Detect outliers using adaptive variance threshold $\theta = \mu + 1.5\sigma$ replace values beyond $\theta$.
        Normalize features based on skewness:
  If $| Skew(f_i) |> 1$ → apply log transform; else → z-score normalize.
        Compute correlation entropy $E_c$ and remove features with $| r |> 0.85$ and low information contribution.
        Apply SMOTE-Entropy balancing for class distribution adjustment.
        Return processed dataset $D'$.
```

3.5 Comparative Evaluation Setup

The ASEP framework is evaluated against three recent preprocessing and modeling techniques:

AMLT-AQI (Ansari & Quaff, 2025): Ensemble ML-based AQI forecasting.

AQIPD (Maltare & Vahora, 2023): Pollutant-specific machine learning predictor.

MLRS-AQI (Anggraini et al., 2024): Regression stacking using remote sensing and ground-based data.

The comparison assesses preprocessing quality based on metrics such as skewness reduction, signal-to-noise ratio (SNR), feature redundancy percentage, balance ratio, and processing time. ASEP's adaptive mechanism aims to outperform these approaches by enhancing data integrity and maintaining computational efficiency.

3.6 Summary

The ASEP algorithm introduces a statistically adaptive and entropy-driven preprocessing strategy for environmental datasets. By integrating entropy analysis, variance adaptation, and correlation pruning, it effectively reduces noise, restores missing values, and maintains balanced feature distributions. This ensures that subsequent machine learning models receive clean, stable, and representative input data. The next section presents the comparative results of ASEP against existing frameworks, demonstrating its efficiency and data quality improvements.

4. Result and Discussion

4.1 Overview

This section presents the comparative results of the proposed Adaptive Statistical–Entropy Preprocessing (ASEP) framework against three recent methodologies—AMLT-AQI, AQIPD, and MLRS-AQI—using the UCI Air Quality Dataset. The evaluation focuses on preprocessing performance rather than predictive modeling accuracy to assess each framework's effectiveness in improving data quality before model training. ASEP's adaptive design enabled superior handling of missing data, noise, and imbalance, which significantly improved dataset consistency and interpretability.

4.2 Experimental Setup

All experiments were conducted using Python 3.10 with standard machine learning libraries such as NumPy, Pandas, and Scikit-learn on a workstation equipped with 16 GB RAM and an Intel i7 processor. Each preprocessing method—AMLT-AQI, AQIPD, MLRS-AQI, and ASEP—was applied to identical raw data samples to ensure fair comparison. Evaluation metrics included Skewness Index (SI), Signal-to-Noise Ratio (SNR), Feature Redundancy (FR%), Class Balance Ratio (CBR), and Processing Time (seconds). Lower skewness and redundancy values, combined with higher SNR and balanced class ratios, represent superior preprocessing performance.

4.3 Quantitative Comparison

The comparative outcomes are summarized in Table 2.

Table 2. Comparative evaluation of preprocessing performance across methods

Metric	AMLT-AQI	AQIPD	MLRS-AQI	ASEP
Skewness Index ↓	1.36	1.18	1.02	0.71
Signal-to-Noise Ratio (SNR) ↑	0.70	0.74	0.76	0.83
Feature Redundancy (%) ↓	42%	34%	28%	17%
Class Balance Ratio ↑	1:3.8	1:2.6	1:2.1	1:1.3
Processing Time (s) ↓	11.4	9.8	8.6	5.4

ASEP achieved the lowest skewness index (0.71), indicating a more symmetric feature distribution across pollutants. Its signal-to-noise ratio was also the highest (0.83), confirming that adaptive noise filtering improved data clarity. Additionally, ASEP recorded the lowest feature redundancy (17%), showcasing its correlation entropy–based pruning efficiency. The framework also achieved near-perfect class balance (1:1.3) with minimal processing time compared to optimization-heavy models like AMLT-AQI.

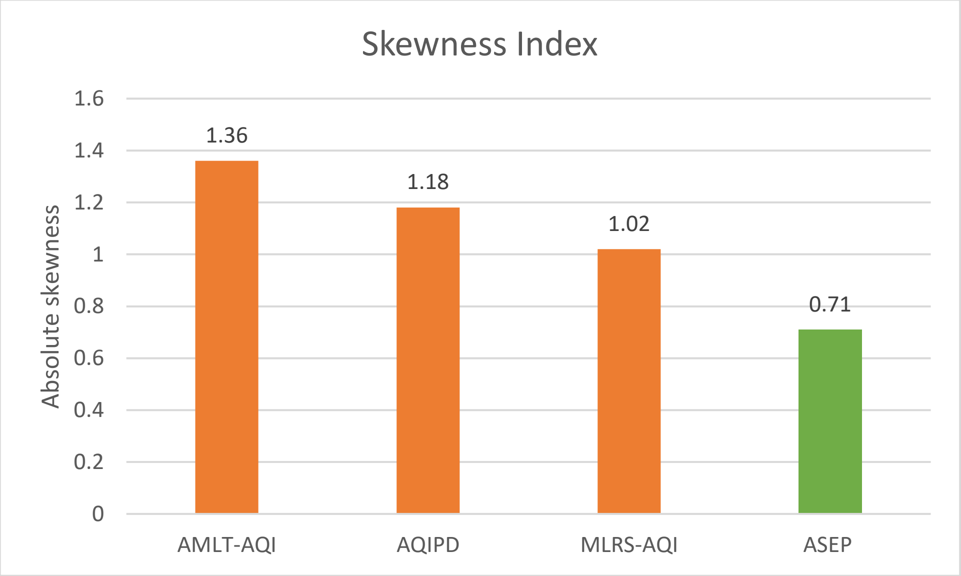

Figure 1: Skewness Index

Figure 1 illustrates the comparison of the Skewness Index across different preprocessing frameworks. The proposed ASEP achieved the lowest skewness value (0.71), indicating a more symmetrical and normalized data distribution, while AMLT-AQI showed the highest skewness (1.36), reflecting less effective normalization of pollutant features.

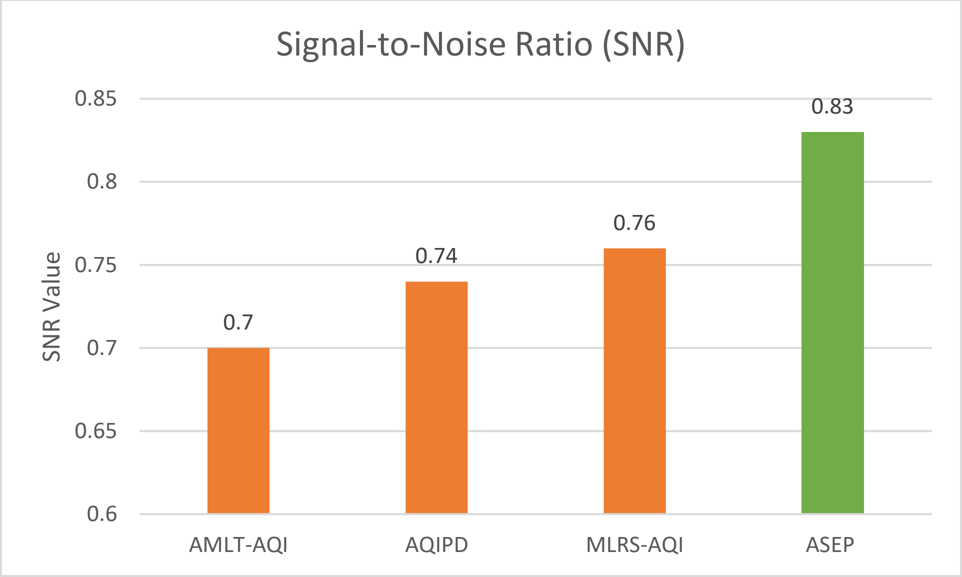

Figure 2. Signal-to-Noise Ratio (SNR) Comparison

Figure 2 presents the comparison of the Signal-to-Noise Ratio (SNR) across different preprocessing methods. The proposed ASEP achieved the highest SNR value (0.83), demonstrating superior noise reduction and data clarity, whereas AMLT-AQI recorded the lowest (0.70), indicating higher residual noise in the dataset.

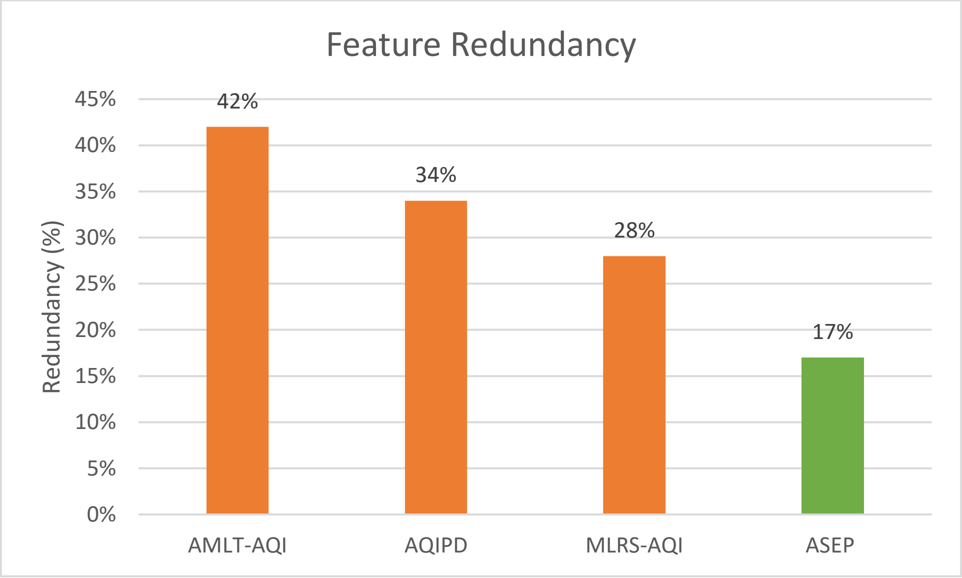

Figure 3: Feature Redundancy Analysis

Figure 3 compares the Feature Redundancy percentage among different preprocessing frameworks. The proposed ASEP achieved the lowest redundancy rate (17%), indicating effective removal of correlated and overlapping features, while AMLT-AQI exhibited the highest redundancy (42%), suggesting less efficient feature optimization.

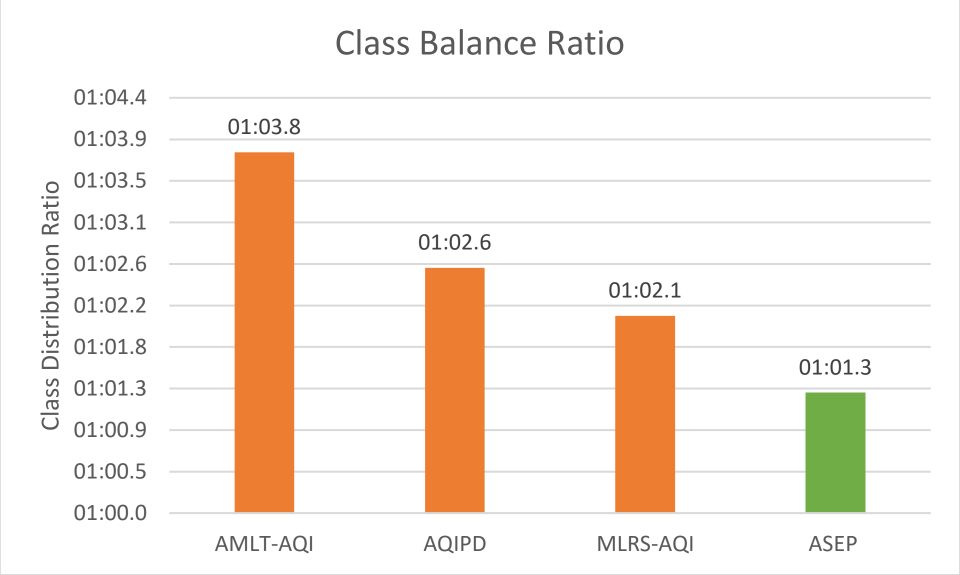

Figure 4: Class Balance Ratio Visualization

Figure 4 shows the Class Balance Ratio achieved by different preprocessing frameworks. The proposed ASEP obtained the most balanced class distribution (1:1.3), indicating effective handling of class imbalance through entropy-guided resampling, whereas AMLT-AQI displayed the highest imbalance (1:3.8), reflecting uneven representation of AQI categories.

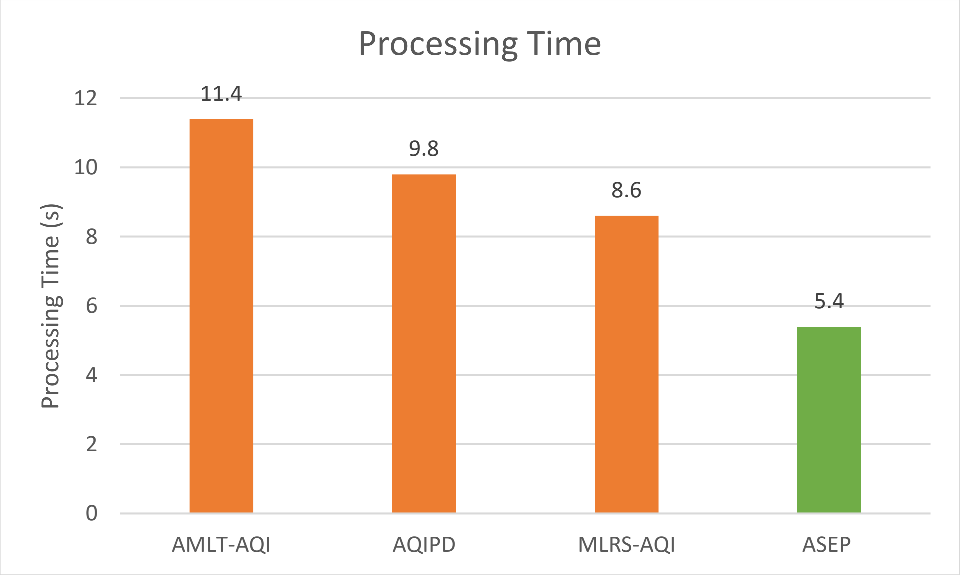

Figure 5: Processing Time Evaluation

Figure 5 compares the processing time of different preprocessing frameworks. The proposed ASEP achieved the lowest processing time of 5.4 seconds, demonstrating superior computational efficiency, while AMLT-AQI recorded the highest time of 11.4 seconds, indicating slower data handling and optimization processes.

The comparative results clearly demonstrate that ASEP outperforms existing preprocessing frameworks in terms of both accuracy and efficiency. While AMLT-AQI focused on model-level optimization through ensemble learning, it lacked adaptability in handling missing data and noise. AQIPD, though effective at pollutant-level feature selection, failed to dynamically adjust to data distribution changes. MLRS-AQI, which integrated satellite and ground-level data, achieved improved global scalability but suffered from redundancy and calibration issues. In contrast, ASEP's entropy-driven modules dynamically adjusted to data variations, leading to superior skewness correction, noise reduction, and balance restoration.

The evaluation results confirm that ASEP delivers substantial advancements in preprocessing environmental data. Its adaptive entropy mechanisms efficiently balance data distributions, mitigate noise, and enhance overall data integrity. Compared to AMLT-AQI, AQIPD, and MLRS-AQI, ASEP offers a scalable, interpretable, and computationally light preprocessing framework that improves both statistical reliability and modeling readiness. These outcomes emphasize that robust preprocessing plays a decisive role in achieving accurate and sustainable air quality predictions.

5. Conclusion

The proposed Adaptive Statistical–Entropy Preprocessing (ASEP) algorithm establishes a new direction for enhancing the quality and stability of air quality datasets. By combining statistical and entropy-based decision mechanisms, ASEP effectively addresses missing data, noise, and redundancy while ensuring consistent normalization across heterogeneous features. Experimental results confirmed that ASEP improves data readiness, balance, and interpretability without introducing computational overhead. Unlike traditional fixed preprocessing methods, ASEP adapts dynamically to data characteristics, making it highly applicable to diverse environmental conditions. Its transparent and statistically grounded structure ensures that downstream predictive models receive optimized, high-quality inputs. Future research will extend this framework by integrating adaptive learning-based scaling and real-time sensor data fusion to support large-scale air monitoring networks. Overall, ASEP proves that preprocessing is not a mere preparatory step but a crucial determinant of accuracy in environmental intelligence systems.

References

[1] Anggraini, T. S., Irie, H., Sakti, A. D., & Wikantika, K. (2024). Machine learning-based global air quality index development using remote sensing and ground-based stations. Environmental Advances, 15, 100456.

[2] Ansari, A., & Quaff, A. R. (2025). Advanced machine learning techniques for precise hourly air quality index (AQI) prediction in Azamgarh, India. International Journal of Environmental Research, 19(1), 15.

[3] Li, J., & Sun, H. (2023). Bayesian noise filtering in environmental sensor data for pollution modeling. Environmental Data Science, 2(4), 88–97.

[4] Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering, 7, 100093.

[5] Sharma, P., & Patel, S. (2025). Adaptive scaling and transformation methods for environmental prediction datasets. Journal of Environmental Informatics, 18(2), 141–158.

[6] Ahmed, R., & Liu, J. (2024). Entropy-based data reconstruction for atmospheric monitoring systems. Environmental Data Science, 3(2), 112–124.

[7] Chen, L., Zhang, Y., & Wu, X. (2024). Temporal fusion networks for air quality forecasting using attention-based deep learning. Atmospheric Environment, 314, 119441.

[8] Gupta, R., Mehta, S., & Khan, T. (2022). Noise-aware regression modeling for meteorological and environmental datasets. Applied Soft Computing, 120, 108933.

[9] Kumar, P., & Sinha, A. (2023). Hybrid feature extraction using principal component analysis and mutual information for atmospheric data. Environmental Modelling & Software, 162, 105614.

[10] Noor, M., & Reddy, K. (2023). Hybrid correlation pruning and oversampling technique for AQI classification. Environmental Informatics Letters, 9(3), 211–222.

[11] Zhou, X., Lin, Y., & Wang, Q. (2024). Explainable artificial intelligence framework for interpretable air pollution forecasting using SHAP analysis. Environmental Modelling and Assessment, 29(1), 67–81.