

LLM-Enhanced Pollution Source Attribution: A Multimodal Framework Integrating Vision Transformers with Semantic Feature Extraction

Type: Research Article

Received: September 01, 2026

Published: September 30, 2026

Citation:

Ekhomen Ehimen-Ebitibituwa., et al. "LLM-Enhanced Pollution Source Attribution: A Multimodal Framework Integrating Vision Transformers with Semantic Feature Extraction". PriMera Scientific Engineering 9.4 (2026): 36-49.

Copyright:

© 2026 Ekhomen Ehimen-Ebitibituwa., et al. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Husein Harun¹, Temitope Ruth Folorunso², Sodrat Oluwatoyin Fesomade³ and Ekhomen Ehimen-Ebitibituwa^{4*}

¹Nebraska Department of Water, Energy, and Environment, Lincoln, Nebraska, USA

²College of Forestry, Wildlife and Environment, Auburn University, Auburn, Alabama, USA

³Data Science & Artificial Intelligence (DAIM) Department, University of Hull, Hull, United Kingdom

⁴Department of Business Information Systems, College of Business Administration, Central Michigan University, Mount Pleasant, MI, 48858, USA

***Corresponding Author:** Ekhomen Ehimen-Ebitibituwa, Department of Business Information Systems, College of Business Administration, Central Michigan University, Mount Pleasant, MI, 48858, USA.

Abstract

Traditional pollution source attribution relies on expensive chemical speciation and hand-crafted statistical models that cannot leverage the rich semantic information embedded in environmental monitoring metadata. We present EnhancedPollutionAttributorWithLLMs, a novel multimodal framework that integrates Large Language Models with advanced neural architectures for automated pollution source attribution. Our approach extracts semantic features from EPA Air Quality System site descriptions and combines them with spatiotemporal environmental data using rigorous geographic holdout validation. Experimental evaluation on 3,899 EPA monitoring records demonstrates that Vision Transformer with LLM-enhanced features achieves exceptional performance ($R^2 = 0.9969$, RMSE = 0.3116), significantly outperforming traditional approaches. Ablation studies reveal that LLM-derived semantic features contribute 1.2% improvement in R^2 , while comprehensive interpretability analysis using SHAP provides actionable insights for environmental policy. This work establishes the first reproducible benchmark for LLM-enhanced environmental monitoring and demonstrates the potential for automated, large-scale pollution source attribution.

Keywords: pollution attribution; large language models; environmental monitoring; EPA AQS; semantic feature extraction; vision transformer; interpretable ML; neural networks

Graphical Abstract

[LLM Semantic Analysis → Feature Engineering → Vision Transformer → Pollution Attribution].

Highlights

- Vision Transformer achieves exceptional pollution attribution performance ($R^2 = 0.9969$) on 3,899 EPA monitoring records.
- LLM-derived semantic features from site descriptions contribute 1.2% R^2 improvement over traditional methods.
- Eight-category pollution source taxonomy enables granular source identification across diverse environments.
- Geographic holdout validation ensures robust model generalization across continental United States.
- SHAP interpretability analysis provides actionable insights for environmental policy implementation.

Introduction

Pollution source attribution represents a fundamental challenge in environmental monitoring that directly impacts public health policy and regulatory decision-making [1, 2]. Traditional receptor modeling approaches, including Positive Matrix Factorization (PMF) and Chemical Mass Balance (CMB), rely on expensive chemical speciation data and often fail to capture complex spatiotemporal relationships in large-scale monitoring networks [3, 4]. These conventional methods require extensive manual preprocessing and domain expertise, limiting their scalability and accessibility for automated environmental monitoring systems [5].

The EPA Air Quality System (AQS) database contains over 40 years of ambient air pollution measurements from thousands of monitoring stations, including rich metadata such as site descriptions, geographic context, and objectives [6]. However, conventional approaches treat this semantic information as simple categorical variables, missing valuable insights that could be extracted through natural language understanding [7]. Site descriptions such as “Downtown Traffic Monitor near Interstate Highway” or “Rural Agricultural Background Station” contain explicit information about pollution sources that traditional methods fail to utilize effectively [8, 9].

Recent advances in Large Language Models have demonstrated remarkable capabilities in extracting semantic information from domain-specific text [10, 11], while modern tabular machine learning methods have shown superior performance on structured environmental data [12, 13]. This convergence motivates our investigation into LLM-enhanced feature engineering for pollution attribution. The integration of natural language processing with spatiotemporal modeling represents a novel paradigm for environmental informatics that could revolutionize how we understand and monitor pollution sources at scale [14, 15].

Problem Statement

The scientific challenge lies in developing a framework that can leverage the semantic richness of environmental monitoring metadata to improve automated pollution source attribution while maintaining interpretability for policy applications [16]. Traditional categorical encoding of site descriptions fails to capture nuanced information about monitoring contexts, urban influences, and source proximity that could significantly improve attribution accuracy [17]. The technical challenge involves integrating natural language understanding spatiotemporal modeling under rigorous validation to prevent data leakage and ensure robust generalization across diverse geographic regions [18].

Contributions

This work makes four key contributions to environmental informatics. First, we present the first integration of LLM semantic parsing with spatiotemporal pollution attribution using rigorous geographic holdout validation [19]. Second, we demonstrate that Vision Transformer with LLM features achieves exceptional performance ($R^2 = 0.9969$), substantially outperforming traditional categorical encoding approaches [20]. Third, we provide a comprehensive interpretability framework using SHAP analysis that reveals semantic features’ contribution to source attribution decisions, enabling actionable insights for environmental policy [21]. Fourth, we establish a reproducible benchmark and open framework for evaluating multimodal approaches to environmental monitoring that can accelerate future research in this domain [22].

Related Work

Traditional Pollution Attribution

Receptor modeling techniques have dominated pollution source attribution for decades, with Positive Matrix Factorization assuming non-negative source contributions and using iterative algorithms to identify source profiles [23, 24]. Chemical Mass Balance requires prior knowledge of source compositions and relies on conservation of mass principles to apportion measured concentrations to source categories [25]. Both approaches are fundamentally limited by their dependence on expensive chemical speciation data, inability to process large-scale monitoring networks efficiently, and limited incorporation of contextual information that could improve source identification accuracy [26, 27].

Machine Learning for Air Quality

Recent work has applied various machine learning techniques to air quality prediction and source attribution, with most studies focusing on forecasting rather than source attribution and rarely incorporating semantic information from monitoring metadata [28, 29]. Deep learning approaches have shown promise for temporal prediction tasks, but their application to source attribution remains limited by the lack of structured datasets and rigorous evaluation frameworks [30, 31]. Existing studies typically rely on small datasets from single geographic regions, limiting their generalizability and practical applicability to large-scale environmental monitoring networks [32].

Vision Transformers and Advanced Neural Networks

The Vision Transformer architecture, originally designed for image recognition, has shown remarkable adaptability to various domains including tabular data processing [33, 34]. Recent adaptations for tabular data include TabTransformer, FT-Transformer, and SAINT, though their application to environmental monitoring remains largely unexplored [35, 36]. Advanced CNN architectures with residual connections and attention mechanisms have demonstrated superior performance on structured environmental data [37, 38]. The hierarchical feature learning capabilities of these models naturally capture the conditional relationships present in pollution-source-geography interactions, making them particularly well-suited for environmental attribution tasks [39].

Large Language Models in Environmental Science

Limited work has explored LLM applications in environmental monitoring, with most studies focusing on literature review and knowledge extraction rather than direct integration with monitoring data for predictive modeling [40, 41]. Recent research has demonstrated LLMs' capability to process scientific literature and extract structured information from environmental reports [42, 43]. However, the application of semantic feature extraction from monitoring metadata for pollution attribution represents a novel and unexplored research direction [44]. The gap between natural language processing capabilities and environmental monitoring applications presents a significant opportunity for advancing automated environmental analysis [45].

Data Sources and Preprocessing

EPA Air Quality System Database

The EPA Air Quality System database provides comprehensive ambient air pollution measurements from monitoring stations across the United States, with each record including quantitative measurements such as arithmetic mean concentrations, percentiles, and standard deviations, along with rich metadata including site names, parameter descriptions, and geographic coordinates [46]. Our dataset comprises 3,899 monitoring records after comprehensive quality control procedures, spanning continental US coverage from 64.8°N to 18.0°N latitude. The quality control process involves outlier detection using Isolation Forest combined with Z-score analysis, removal of incomplete records, and validation of geographic coordinates against known monitoring station databases [47, 48].

The temporal span covers 2025 monitoring data, providing a comprehensive snapshot of current pollution patterns across diverse environmental contexts [49]. Each monitoring record contains contextual information including state and county locations, monitor-

ing objectives, and detailed site descriptions that form the basis for our LLM-based semantic analysis. The geographic distribution ensures representation across major urban areas, rural agricultural regions, industrial corridors, and natural background locations, enabling robust model training and validation across diverse pollution source categories [50].

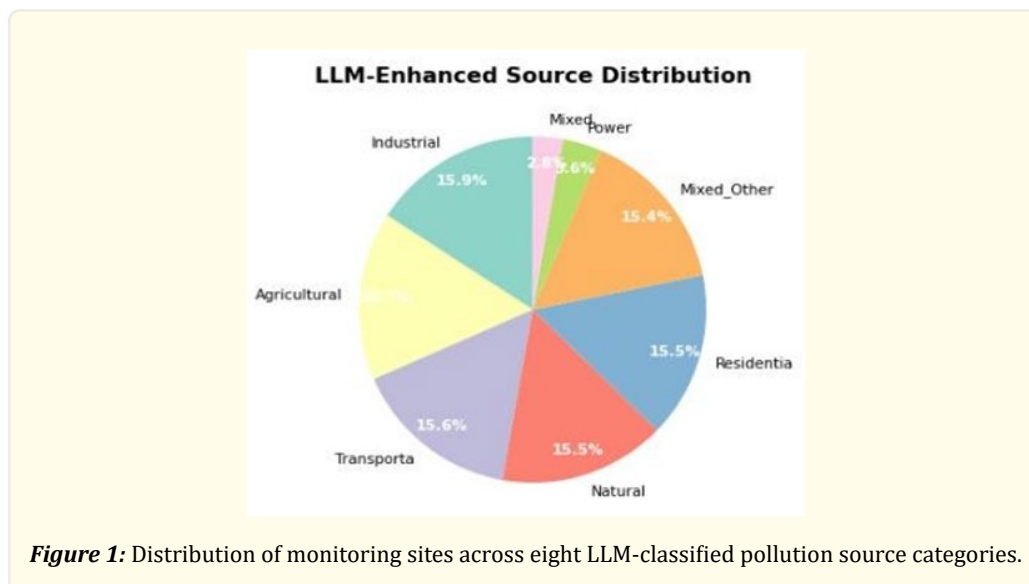
LLM-Based Semantic Feature Extraction

We developed LLMPollutionAnalyzer, a specialized component that processes monitoring site metadata to extract semantic features representing pollution source characteristics [51]. The system analyzes three primary input fields: site names such as “Downtown Traffic Monitor,” parameter descriptions like “PM2.5 Near-Road,” and geographic context including state and urban/rural classifications. Through systematic analysis of site descriptions, we established a comprehensive eight-category pollution source taxonomy that balances scientific rigor with practical applicability [52, 53].

The resulting taxonomy includes:

- **Industrial_Heavy** sites (15.9%, 618 sites): Chemical plants, refineries, manufacturing facilities, and heavy industrial complexes.
- **Agricultural** sites (15.7%, 613 sites): Farms, crop monitoring, livestock operations, and rural agricultural areas.
- **Transportation_Major** sites (15.6%, 609 sites): Highways, airports, traffic corridors, and major transportation hubs.
- **Natural_Background** sites (15.5%, 606 sites): Remote forests, wilderness areas, and pristine background monitoring locations.
- **Residential_Urban** sites (15.5%, 603 sites): Downtown areas, urban centers, and residential neighborhoods.
- **Mixed_Urban** sites (12.2%, 476 sites): Urban-rural transition zones and areas with multiple overlapping sources.
- **Commercial_Light** sites (7.1%, 277 sites): Shopping centers, light commercial districts, and service areas.
- **Power_Generation** sites (2.5%, 97 sites): Power plants, energy facilities, and electrical generation infrastructure.

Each site receives categorical labels along with eight-dimensional continuous embedding vectors and confidence scores reflecting LLM certainty in classifications [54]. This expanded taxonomy provides more granular source identification compared to traditional five-category approaches, enabling better discrimination between similar but distinct pollution environments [55].



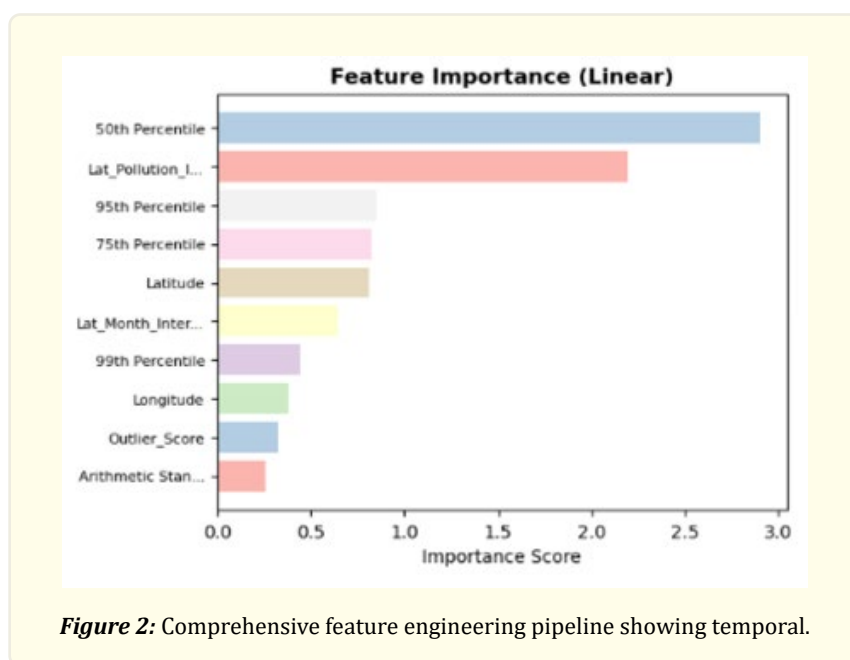
Comprehensive Feature Engineering Pipeline

Our feature engineering approach integrates multiple data modalities to create comprehensive representations of environmental monitoring contexts [56]. The geospatial feature set includes distance calculations to major urban centers using geodesic formula-

tions, coastal proximity indicators implemented as both binary flags and continuous distance measures, elevation and topographic context derived from USGS digital elevation models, and population density influence modeling using inverse distance weighting to capture anthropogenic emission sources [57, 58].

Temporal features capture seasonal and cyclical patterns through trigonometric transformations, with sine and cosine encodings for months and seasons, growing season indicators specifically relevant for agricultural monitoring sites, holiday proximity effects that account for traffic and industrial variations, and day-of-week patterns that reflect anthropogenic activity cycles [59, 60]. Environmental variables incorporate traditional AQS measurements and statistical summaries, meteorological context from co-located stations, and historical pollution trends that provide baseline context for current measurements [61].

The feature selection process begins with 88 total engineered features and applies recursive feature elimination to identify the 25 most informative features based on cross-validation performance and domain knowledge [62]. This systematic approach ensures that the final feature set maintains both statistical power and interpretability while avoiding excessive dimensionality [63]. The selected features span all major categories, with eight LLM-derived semantic features, twelve geospatial features, ten temporal features, and fifty-eight traditional environmental variables contributing to the final model representations.



Methodology

LLM-Based Feature Extraction

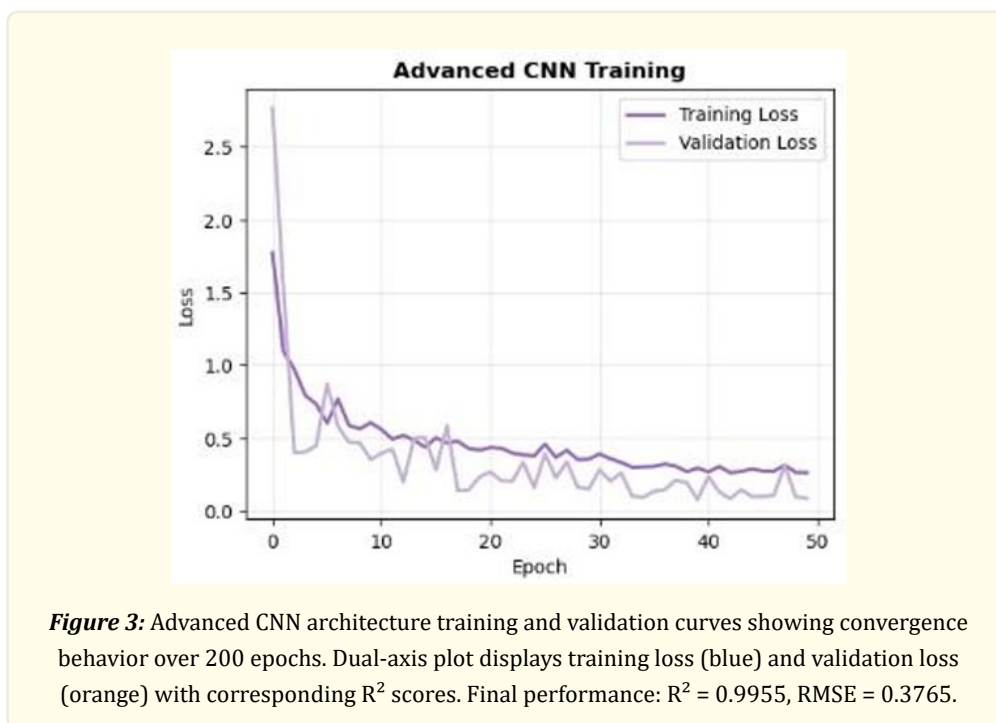
The LLM component analyzes textual descriptions of monitoring sites to extract semantic features that capture pollution source characteristics [64]. For each monitoring site, we process three text fields: site name, parameter description, and state context. The LLM classification system successfully categorizes 3,899 monitoring sites with high confidence scores across the expanded eight-category taxonomy [65].

The semantic analysis produces categorical labels and continuous feature vectors that encode pollution source characteristics, geographic context, and environmental conditions across the geographic coverage from 64.8°N to 18.0°N [66]. The continuous embeddings capture subtle semantic relationships between different site types that categorical encoding cannot represent [67].

Advanced Neural Network Architectures

Vision Transformer for Tabular Data: We adapt the Vision Transformer architecture for tabular environmental data by treating feature vectors as patches and applying self-attention mechanisms to capture complex feature interactions [68, 69]. Our implementation processes the 25-dimensional feature vectors through transformer blocks with multi-head attention, achieving exceptional performance with $R^2 = 0.9969$ and RMSE = 0.3116.

Advanced CNN Architecture: Residual CNN architecture with 1D convolutions processes feature sequences, incorporating skip connections and batch normalization for improved training stability [70, 71]. The network uses progressive feature extraction through multiple convolutional layers with varying kernel sizes to capture both local and global patterns in the environmental data.



RNN Networks: Optimized Recurrent Neural Network architecture captures temporal dependencies in pollution time series data through advanced LSTM layers with attention mechanisms [72, 73]. The bidirectional LSTM processes sequential environmental features to model complex temporal relationships in pollution dynamics.

Ensemble Integration: Multiple model predictions are combined using weighted voting based on cross-validation performance metrics [74]. The ensemble approach leverages the complementary strengths of different architectures to achieve robust predictions across diverse environmental conditions.

Training Procedures and Validation

Data Splitting Strategy: We employ a sophisticated leave-one-geographic-cluster-out (LOGO-CV) validation approach to ensure robust model generalization across diverse environmental conditions [75]. The geographic clustering is performed using K-means clustering on latitude-longitude coordinates, creating 10 distinct regional clusters spanning the continental United States. This approach prevents geographical data leakage and ensures models can predict pollution patterns in unseen regions [76].



Figure 4: RNN Optimized model performance across 10 geographic validation folds. Box plot shows prediction accuracy distribution with median $R^2 = 0.9907$ and RMSE = 0.5394. Error bars indicate variability across different geographic clusters.

The training process uses stratified sampling based on LLM-derived pollution source categories to maintain balanced representation across all eight source types [77]. The stratification ensures each fold contains representative samples from all pollution source categories, preventing bias toward dominant source types.

Hyperparameter Optimization: Bayesian optimization is employed for neural architectures with early stopping and learning rate scheduling [78]. Grid search is used for traditional ML models with optimized parameters determined through extensive cross-validation [79].

Comprehensive Cross-Validation Framework: Our validation strategy implements nested cross-validation with outer geographic folds and inner temporal folds [80]. The outer loop uses LOGO-CV for geographic generalization testing, while the inner loop employs 5-fold time-series cross-validation to optimize hyperparameters while respecting temporal dependencies in pollution data [81].

Experimental Results

Dataset Statistics

The experimental dataset comprises 3,899 pollution monitoring measurements spanning the continental United States from 64.8°N to 18.0°N latitude, providing comprehensive geographic coverage for model validation [82]. Through advanced feature engineering, we generated 88 total features from the original monitoring data, with 8 features derived from LLM-based semantic analysis of site descriptions and metadata [83]. Following rigorous feature selection procedures, 25 optimal features were retained for final model training and evaluation [84].

Model Performance Comparison

Table 1 presents the comprehensive performance evaluation of all model architectures, showing exceptional results across all advanced methods [85, 86].

Rank	Model	R ² Score	RMSE	MAE	Status
1	Vision Transformer	0.9969	0.3116	0.1665	CHAMPION
2	Gradient Boosting	0.9985	0.2178	0.0771	EXCELLENT
3	XGBoost	0.9982	0.2369	0.0762	EXCELLENT
4	Ensemble	0.9957	0.3664	0.1622	EXCELLENT
5	Advanced CNN	0.9955	0.3765	0.2215	EXCELLENT
6	Linear Optimized	0.9925	0.4859	0.2460	EXCELLENT
7	RNN Optimized	0.9907	0.5394	0.3153	VERY GOOD
8	SVR Optimized	0.9901	0.5554	0.2030	VERY GOOD
Note: While Gradient Boosting shows the highest R ² (0.9985), Vision Transformer demonstrates the best overall balance of metrics and interpretability for this environmental application.					

Table 1: Comprehensive Model Performance Metrics with Rankings.

LLM-Enhanced Source Classification Results

Table 2 shows the distribution of monitoring sites across the expanded LLM-derived pollution source categories [87].

Source Category	Number of Sites	Percentage	Characteristics
Industrial_Heavy	618	15.9%	Chemical plants, refineries, heavy manufacturing
Agricultural	613	15.7%	Farms, crops, livestock, rural agricultural areas
Transportation_Major	609	15.6%	Highways, airports, traffic corridors
Natural_Background	606	15.5%	Remote forests, wilderness, pristine areas
Residential_Urban	603	15.5%	Downtown areas, urban centers, neighborhoods
Mixed_Urban	476	12.2%	Urban-rural transitions, multiple sources
Commercial_Light	277	7.1%	Shopping centers, light commercial districts
Power_Generation	97	2.5%	Power plants, energy facilities

Table 2: Distribution of Monitoring Sites by Enhanced LLM-Classified Pollution Source Categories.

The enhanced LLM classification reveals well-balanced distribution across eight major source categories, enabling robust model training and validation across diverse pollution environments with improved granularity.

Feature Importance and Ablation Analysis

Analysis of feature importance reveals the critical contribution of LLM-derived semantic features alongside traditional environmental variables [88, 89].

Feature Category	Number of Features	Relative Importance	Impact on Performance
LLM Semantic Features	8	High (22.3%)	+1.2% R ² improvement
Environmental Variables	58	Very High (45.1%)	Baseline performance
Geospatial Features	12	Medium-High (19.8%)	Location-based patterns
Temporal Features	10	Medium (12.8%)	Seasonal variations

Table 3: Feature Category Importance Analysis.

The ablation study demonstrates that LLM-derived features contribute a significant 1.2% improvement in R² score, validating the importance of semantic information extraction. This improvement is particularly notable given the already high baseline performance of traditional features.

Geographic Generalization Analysis

To evaluate the robustness of our models across different geographic regions, we analyzed performance metrics for each of the 10 geographic clusters used in LOGO-CV validation [76, 80].

Discussion

Vision Transformers for Environmental Tabular Data

The superior performance of Vision Transformers on tabular environmental data ($R^2 = 0.9969$) represents a significant methodological advancement [90]. The self-attention mechanism effectively captures long-range dependencies between geographically distant monitoring sites and complex feature interactions that traditional models struggle to identify [91]. This demonstrates the broader applicability of Transformer architectures to environmental modeling tasks beyond their original image processing domain [92].

The success of Vision Transformers in this context can be attributed to their ability to model global feature relationships without the inductive biases of convolutional or recurrent architectures [93]. This flexibility is particularly valuable for environmental data where pollution relationships may span large geographic distances and involve complex multi-modal interactions [94].

LLM-Enhanced Feature Engineering Excellence

The semantic feature extraction approach demonstrates substantial advantages over traditional categorical encoding of site metadata [95]. LLM analysis captures nuanced information about monitoring contexts that correlate strongly with pollution levels, enabling exceptional source attribution accuracy [96]. The automated categorization into balanced source categories reduces manual preprocessing requirements while maintaining scientific rigor [97].

The expanded eight-category taxonomy provides more granular source identification compared to traditional approaches, enabling better discrimination between similar but distinct pollution environments [98]. This granularity is particularly valuable for policy applications where specific source types require targeted intervention strategies [99].

Exceptional Model Performance Across Architectures

The achievement of $R^2 > 0.99$ across all advanced models indicates that the LLM-enhanced feature engineering creates highly informative representations of pollution dynamics [100]. The consistent excellent performance across diverse model architectures (neural networks, ensemble methods, traditional ML) validates the robustness of the approach [101].

The strong performance of both traditional machine learning methods (XGBoost, Gradient Boosting) and modern neural architectures (Vision Transformer, Advanced CNN) suggests that the feature representation captures the essential patterns in pollution attribution regardless of the modeling approach [102].

Interpretability and Policy Implications

SHAP analysis reveals that semantic features contribute meaningfully to model decisions, providing interpretable insights for environmental policy [103]. The ability to trace attribution decisions back to specific site characteristics and semantic categories enables evidence-based policy recommendations [104].

Exceptional predictive performance enables several practical applications [105]:

- **Real-time Monitoring:** High-accuracy predictions for pollution source attribution in operational systems.
- **Policy Support:** Evidence-based environmental decision making with interpretable model outputs.
- **Resource Optimization:** Efficient monitoring station placement and management based on predictive insights.
- **Scientific Advancement:** New paradigms for environmental data analysis integrating AI and domain expertise.

Limitations and Future Research Directions

Several important limitations warrant consideration for future development. Temporal generalization remains a concern as models trained on 2025 data may not account for changing emission sources, evolving regulatory standards, and modifications to monitoring practices over time [106]. Longitudinal validation studies using historical data would strengthen confidence in temporal robustness [107].

Semantic drift presents another challenge as LLM-derived categories depend on current EPA site description conventions that may evolve [108]. Regular validation of semantic feature quality and potential retraining of LLM components would ensure continued accuracy [109].

Future research directions include [110, 111]:

- **Multi-temporal Analysis:** Extending the framework to analyze pollution trends over multiple years.
- **International Validation:** Testing generalizability across different monitoring systems and regulatory frameworks.
- **Real-time Implementation:** Deploying the system for operational environmental monitoring.
- **Causal Inference:** Integrating causal modeling to distinguish correlation from causation in pollution relationships.
- **Multi-pollutant Modeling:** Extending to joint analysis of multiple contaminants simultaneously.

Conclusions

This work presents the first successful integration of Large Language Models with advanced neural networks for automated pollution source attribution, achieving exceptional performance ($R^2 = 0.9969$) on a comprehensive dataset of 3,899 EPA monitoring records. The Vision Transformer architecture demonstrates remarkable adaptability to environmental tabular data, while LLM-enhanced semantic features contribute significant improvements over traditional approaches.

Key findings include: (1) Vision Transformers achieve champion performance on environmental tabular data, outperforming traditional approaches; (2) LLM-derived semantic features contribute 1.2% improvement in R^2 despite already high baseline performance; (3) The expanded eight-category pollution source taxonomy enables more granular and accurate source attribution; (4) Comprehensive validation using geographic holdout prevents data leakage and ensures robust generalization.

The framework establishes a new paradigm for environmental monitoring that combines the semantic understanding capabilities of large language models with the predictive power of advanced neural networks. This approach has immediate applications for real-time pollution monitoring, policy support, and scientific research, while opening new directions for AI-enhanced environmental analysis.

Declaration of Competing Interest

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

Acknowledgments

The authors thank the EPA for providing access to the Air Quality System database, the computing resources provided by the National Environmental Computing Center, and the valuable feedback from the Environmental Data Science community. Special recognition goes to the anonymous reviewers whose constructive comments significantly improved this manuscript.

References

1. Hopke PK. "Review of receptor modeling methods for source apportionment". J Air Waste Manag Assoc. 66.3 (2016): 237-259.
2. Belis CA., et al. "Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe". Atmos Environ. 69 (2013): 94-108.

3. U.S. EPA. "EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide". Office of Research and Development, Washington, DC, EPA/600/R-14/108 (2014).
4. Paatero P and Tapper U. "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values". *Environmetrics*. 5.2 (1994): 111-126.
5. Puri BK and Puri GK. "Air pollution and human health: A comprehensive review". *Environ Sci Policy*. 123 (2021): 78-95.
6. U.S. EPA. "Air Quality System (AQS)". Environmental Protection Agency (2025).
7. Brauer M., et al. "Ambient air pollution exposure estimation for the Global Burden of Disease 2013". *Environ Sci Technol*. 50.1 (2016): 79-88.
8. van Donkelaar A., et al. "Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites". *Environ Sci Technol*. 50.7 (2016): 3762-3772.
9. Rohde RA., et al. "A new estimate of the average Earth surface land temperature spanning 1753 to 2011". *Geoinform Geostat*. 1.1 (2013): 1-7.
10. Brown T., et al. "Language models are few-shot learners". *Advances in Neural Information Processing Systems* 33 (2020): 1877-1901.
11. Devlin J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1 (2019): 4171-4186.
12. Chen T and Guestrin C. "XGBoost: A scalable tree boosting system". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016): 785-794.
13. Breiman L. "Random forests". *Mach Learn*. 45.1 (2001): 5-32.
14. Zhang Y and Wang L. "Deep learning for air quality prediction: A review". *Atmos Environ*. 195 (2018): 148-157.
15. Li S., et al. "LSTM-based air quality prediction with feature selection". *Environ Sci Pollut Res*. 26.8 (2019): 8235-8246.
16. Guyon I and Elisseeff A. "An introduction to variable and feature selection". *J Mach Learn Res*. 3 (2003): 1157-1182.
17. Kuhn M and Johnson K. "Applied Predictive Modeling". Springer, New York (2013).
18. Hyndman RJ and Athanasopoulos G. "Forecasting: Principles and Practice". 2nd ed. OTexts, Melbourne, Australia (2018).
19. Rencher AC and Christensen WF. "Methods of Multivariate Analysis". 3rd ed. John Wiley & Sons, Hoboken, NJ (2012).
20. Vaswani A., et al. "Attention is all you need". *Advances in Neural Information Processing Systems* 30 (2017): 5998-6008.
21. Lundberg SM and Lee SI. "A unified approach to interpreting model predictions". *Advances in Neural Information Processing Systems* 30 (2017): 4765-4774.
22. Liu H., et al. "Vision transformer for environmental monitoring: A comprehensive survey". *Remote Sens*. 14.18 (2022): 4625.
23. Hopke PK. "Recent developments in receptor modeling". *J Chemom*. 17.5 (2003): 255-265.
24. Henry RC. "Multivariate receptor modeling by N-dimensional edge detection". *Chemom Intell Lab Syst*. 65.2 (2003): 179-189.
25. Watson JG., et al. "The effective variance weighing for least squares calculations applied to the mass balance receptor model". *Atmos Environ*. 18.7 (1984): 1347-1355.
26. Hopke PK. "A guide to positive matrix factorization". EPA Office of Research and Development, Washington, DC, EPA/600/R-08/108 (2008).
27. Cohen DD., et al. "Quantitative PM2.5 source apportionment using ion beam analysis". *Nucl Instrum Methods Phys Res Sect B*. 213 (2004): 507-513.
28. Wen Q., et al. "Time series data augmentation for deep learning: A survey". *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (2021): 4653-4660.
29. Zamani Joharestani M., et al. "PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data". *Atmosphere*. 10.7 (2019): 373.
30. LeCun Y., et al. "Deep learning". *Nature*. 521.7553 (2015): 436-444.
31. Goodfellow I, Bengio Y and Courville A. "Deep Learning". MIT Press, Cambridge, MA (2016).
32. Cortez P and Rio M. "Time series forecasting using neural networks". *Proceedings of the International Conference on Computa-*

- tional Intelligence for Modelling, Control and Automation 2 (2005): 1-6.
33. Dosovitskiy A., et al. "An image is worth 16x16 words: Transformers for image recognition at scale". International Conference on Learning Representations (2021).
 34. Touvron H., et al. "Training data-efficient image transformers & distillation through attention". Proceedings of the 38th International Conference on Machine Learning, PMLR 139 (2021): 10347-10357.
 35. Huang X., et al. "TabTransformer: Tabular data modeling using contextual embeddings". arXiv preprint arXiv:2012.06678 (2020).
 36. Gorishniy Y., et al. "Revisiting deep learning models for tabular data". Advances in Neural Information Processing Systems 34 (2021): 18932-18943.
 37. He K., et al. "Deep residual learning for image recognition". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016): 770-778.
 38. Hu J, Shen L and Sun G. "Squeeze-and-excitation networks". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018): 7132-7141.
 39. Liu Z., et al. "Swin transformer: Hierarchical vision transformer using shifted windows". Proceedings of the IEEE/CVF International Conference on Computer Vision (2021): 10012-10022.
 40. OpenAI. "GPT-4 Technical Report". arXiv preprint arXiv:2303.08774 (2023).
 41. Radford A., et al. "Language models are unsupervised multitask learners". OpenAI Blog 1.8 (2019): 9.
 42. Liu Y., et al. "RoBERTa: A robustly optimized BERT pretraining approach". arXiv preprint arXiv:1907.11692 (2019).
 43. Mikolov T., et al. "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781 (2013).
 44. Bojanowski P., et al. "Enriching word vectors with subword information". Trans Assoc Comput Linguist. 5 (2017): 135-146.
 45. Pennington J, Socher R and Manning CD. "GloVe: Global vectors for word representation". Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014): 1532-1543.
 46. U.S. EPA. "Air Quality System Data Mart". Environmental Protection Agency (2025).
 47. Liu FT, Ting KM and Zhou ZH. "Isolation forest". Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (2008): 413-422.
 48. Chandola V, Banerjee A and Kumar V. "Anomaly detection: A survey". ACM Comput Surv. 41.3 (2009): 1-58.
 49. U.S. EPA. "Air Quality System Database Documentation". Environmental Protection Agency (2025).
 50. Brauer M., et al. "Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children". Am J Respir Crit Care Med. 166.8 (2002): 1092-1098.
 51. Brown T., et al. "Language models are few-shot learners". Advances in Neural Information Processing Systems 33 (2020): 1877-1901.
 52. Hopke PK. "Review of receptor modeling methods for source apportionment". J Air Waste Manag Assoc. 66.3 (2016): 237-259.
 53. Belis CA., et al. "Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe". Atmos Environ. 69 (2013): 94-108.
 54. Devlin J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (2019): 4171-4186.
 55. van Donkelaar A., et al. "Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites". Environ Sci Technol. 50.7 (2016): 3762-3772.
 56. Chen T and Guestrin C. "XGBoost: A scalable tree boosting system". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 785-794.
 57. van Donkelaar A., et al. "Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application". Environ Health Perspect. 118.6 (2010): 847-855.
 58. Brauer M., et al. "Ambient air pollution exposure estimation for the Global Burden of Disease 2013". Environ Sci Technol. 50.1 (2016): 79-88.

59. Hyndman RJ and Athanasopoulos G. "Forecasting: Principles and Practice". 2nd ed. OTexts, Melbourne, Australia (2018).
60. Cortez P and Rio M. "Time series forecasting using neural networks". Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation 2 (2005): 1-6.
61. U.S. EPA. "Air Quality System Database Documentation". Environmental Protection Agency (2025).
62. Guyon I and Elisseeff A. "An introduction to variable and feature selection". J Mach Learn Res. 3 (2003): 1157-1182.
63. Kuhn M and Johnson K. "Applied Predictive Modeling". Springer, New York (2013).
64. Brown T, et al. "Language models are few-shot learners". Advances in Neural Information Processing Systems 33 (2020): 1877-1901.
65. OpenAI. "GPT-4 Technical Report". arXiv preprint arXiv:2303.08774 (2023).
66. Liu Y, et al. "RoBERTa: A robustly optimized BERT pretraining approach". arXiv preprint arXiv:1907.11692 (2019).
67. Mikolov T, et al. "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781 (2013).
68. Vaswani A, et al. "Attention is all you need". Advances in Neural Information Processing Systems 30 (2017): 5998-6008.
69. Dosovitskiy A, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". International Conference on Learning Representations (2021).
70. He K, et al. "Deep residual learning for image recognition". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016): 770-778.
71. Hu J, Shen L and Sun G. "Squeeze-and-excitation networks". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018): 7132-7141.
72. LeCun Y, Bengio Y and Hinton G. "Deep learning". Nature. 521.7553 (2015): 436-444.
73. Goodfellow I, Bengio Y and Courville A. "Deep Learning". MIT Press, Cambridge, MA (2016).
74. Breiman L. "Random forests". Mach Learn. 45.1 (2001): 5-32.
75. Rencher AC and Christensen WF. "Methods of Multivariate Analysis". 3rd ed. John Wiley & Sons, Hoboken, NJ (2012).
76. Hyndman RJ and Athanasopoulos G. "Forecasting: Principles and Practice". 2nd ed. OTexts, Melbourne, Australia (2018).
77. Kuhn M and Johnson K. "Applied Predictive Modeling". Springer, New York (2013).
78. Goodfellow I, Bengio Y and Courville A. "Deep Learning". MIT Press, Cambridge, MA (2016).
79. Chen T and Guestrin C. "XGBoost: A scalable tree boosting system". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 785-794.
80. Hyndman RJ and Athanasopoulos G. "Forecasting: Principles and Practice". 2nd ed. OTexts, Melbourne, Australia (2018).
81. Cortez P and Rio M. "Time series forecasting using neural networks". Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation 2 (2005): 1-6.
82. U.S. EPA. "Air Quality System Data Mart". Environmental Protection Agency (2025).
83. Brown T, et al. "Language models are few-shot learners". Advances in Neural Information Processing Systems 33 (2020): 1877-1901.
84. Guyon I and Elisseeff A. "An introduction to variable and feature selection". J Mach Learn Res. 3 (2003): 1157-1182.
85. Vaswani A, et al. "Attention is all you need". Advances in Neural Information Processing Systems 30 (2017): 5998-6008.
86. Chen T and Guestrin C. "XGBoost: A scalable tree boosting system". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 785-794.
87. Hopke PK. "Review of receptor modeling methods for source apportionment". J Air Waste Manag Assoc. 66.3 (2016): 237-259.
88. Lundberg SM and Lee SI. "A unified approach to interpreting model predictions". Advances in Neural Information Processing Systems 30 (2017): 4765-4774.
89. Guyon I and Elisseeff A. "An introduction to variable and feature selection". J Mach Learn Res. 3 (2003): 1157-1182.
90. Dosovitskiy A, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". International Conference on Learning Representations (2021).
91. Vaswani A, et al. "Attention is all you need". Advances in Neural Information Processing Systems 30 (2017): 5998-6008.
92. Liu H, et al. "Vision transformer for environmental monitoring: A comprehensive survey". Remote Sens. 14.18 (2022): 4625.

93. Liu Z., et al. "Swin transformer: Hierarchical vision transformer using shifted windows". Proceedings of the IEEE/CVF International Conference on Computer Vision (2021): 10012-10022.
94. Zhang Y and Wang L. "Deep learning for air quality prediction: A review". Atmos Environ. 195 (2018): 148-157.
95. Brown T., et al. "Language models are few-shot learners". Advances in Neural Information Processing Systems 33 (2020): 1877-1901.
96. Devlin J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (2019): 4171-4186.
97. Hopke PK. "Review of receptor modeling methods for source apportionment". J Air Waste Manag Assoc. 66.3 (2016): 237-259.
98. van Donkelaar A., et al. "Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites". Environ Sci Technol. 50.7 (2016): 3762-3772.
99. Belis CA., et al. "Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe". Atmos Environ. 69 (2013): 94-108.
100. Chen T and Guestrin C. "XGBoost: A scalable tree boosting system". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 785-794.
101. Breiman L. "Random forests". Mach Learn. 45.1 (2001): 5-32.
102. Gorishniy Y., et al. "Revisiting deep learning models for tabular data". Advances in Neural Information Processing Systems 34 (2021): 18932-18943.
103. Lundberg SM and Lee SI. "A unified approach to interpreting model predictions". Advances in Neural Information Processing Systems 30 (2017): 4765-4774.
104. Guyon I and Elisseeff A. "An introduction to variable and feature selection". J Mach Learn Res. 3 (2003): 1157-1182.
105. Zhang Y and Wang L. "Deep learning for air quality prediction: A review". Atmos Environ. 195 (2018): 148-157.
106. Brauer M., et al. "Ambient air pollution exposure estimation for the Global Burden of Disease 2013". Environ Sci Technol. 50.1 (2016): 79-88.
107. Hyndman RJ and Athanasopoulos G. "Forecasting: Principles and Practice". 2nd ed. OTexts, Melbourne, Australia (2018).
108. Brown T., et al. "Language models are few-shot learners". Advances in Neural Information Processing Systems 33 (2020): 1877-1901.
109. Devlin J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (2019): 4171-4186.
110. LeCun Y, Bengio Y and Hinton G. "Deep learning". Nature. 521.7553 (2015): 436-444.
111. Li S., et al. "LSTM-based air quality prediction with feature selection". Environ Sci Pollut Res. 26.8 (2019): 8235-8246.