Addressing the Coupled Optimization of Feature Selection and Hyperparameter Tuning Using a TPE-Driven XGBoost-RFE Framework

N. Mohamed Abdul Kader Jailani
School of Computer Science & Applications, REVA University, Bangalore, India
jailani.msa@gmail.com (corresponding author)

Geeta C. Mara
School of Computing & Information Technology, REVA University, Bangalore, India
geetac.mara@reva.edu.in

Received: 22 September 2025 | Revised: 24 December 2025 | Accepted: 3 January 2026

 DOI: https://doi.org/10.48084/etasr.15024

ABSTRACT

This study presents a methodological advancement for machine learning by developing a framework that solves the coupled problem of feature selection and hyperparameter optimization. The proposed TPE-XGBoost-RFE algorithm integrates a sequential model-based optimization technique, the Tree-structured Parzen Estimator (TPE), with a wrapper feature selection method. This approach concurrently searches for a globally optimal combination of predictive features and model hyperparameters. The efficacy of the framework is demonstrated in the task of predicting long-term tropospheric ozone concentrations. This integrated process identifies an optimal 22-feature subset, reducing dimensionality by 37% while simultaneously tuning nine key XGBoost hyperparameters. The robustness of this subset is validated across multiple machine learning models, all exhibiting superior predictive performance with lower error metrics compared to those trained on the full feature set or through a simpler filter-based method. This study demonstrates that a unified optimization strategy is critical for developing high-performing predictive models.

Keywords-feature selection; hyperparameter optimization; XGBoost; recursive feature elimination (RFE); tree-structured parzen estimator (TPE); ozone prediction; optuna

I. INTRODUCTION

Accurate prediction of long-term tropospheric ozone (O_3) concentration is a fundamental challenge in environmental science, critical to assessing impacts on public health and ecosystems [1]. While early predictive efforts utilized linear models, the complex, non-linear relationships governing ozone formation necessitated the adoption of sophisticated data-driven paradigms [2-11]. The literature records a clear methodological evolution: Support Vector Machines (SVM) were initially adopted for their ability to handle high-dimensional spaces [12, 13], followed by ensemble methods such as Random Forests and Gradient Boosting, which demonstrated robust resistance to overfitting [14, 15]. Recently, tree-based boosting algorithms, particularly XGBoost, have emerged as the dominant standard due to their exceptional predictive accuracy and computational efficiency [2, 16-18].

Despite these advances, the efficacy of these models is often constrained by the high dimensionality of environmental datasets, introducing noise and computational inefficiency [10, 13, 18].To mitigate this, feature selection is a critical preprocessing step [5, 19], and similar principles of relevance identification and information reduction have been explored in other data-intensive learning domains [20, 21]. The existing literature categorizes these strategies into filters, wrappers, and embedded techniques [19, 22]. Filter methods, such as Pearson correlation, are computationally efficient but limited by their failure to account for feature interdependencies [19, 23]. In contrast, wrapper methods like Recursive Feature Elimination (RFE) evaluate subsets based on model performance, offering superior capture of variable interactions [24]. Recent hybrid approaches have combined RFE with metaheuristic algorithms such as Grey Wolf Optimization (GWO) [25] and Ant Colony Optimization (ACO) [26] coupled with SVM to enhance search efficiency in multi-target ozone prediction tasks. However, RFE is computationally intensive, and its success is inextricably linked to the underlying estimator's configuration; an arbitrary or poorly tuned model can cause the algorithm to converge on a suboptimal feature subset [4, 24].

Complex models, such as XGBoost, require precise hyperparameter tuning to achieve optimal performance [27, 28]. Traditional optimization methods, such as grid and random search, suffer from the "curse of dimensionality" and are often

intractable for navigating large search spaces [28, 29]. Consequently, the field has shifted to Sequential Model-Based Optimization (SMBO) techniques, such as Bayesian Optimization [16]. The Tree-structured Parzen Estimator (TPE) algorithm has proven particularly effective in this domain, offering a probabilistic approach to intelligently explore hyperparameter distributions [17, 28]. However, although recent studies have successfully applied TPE in biomedical domains [17], its application to coupled environmental modeling challenges remains underexplored.

A critical review of the current landscape reveals a significant methodological gap: feature selection and hyperparameter optimization are predominantly treated as separate sequential tasks [28, 29]. Common workflows tune hyperparameters on full feature sets before selection, or vice versa, creating a decoupled approach that is fundamentally suboptimal [16, 27]. The optimal configuration for a full feature set is rarely optimal for a reduced subset, and performing RFE with a poorly tuned model may erroneously eliminate important features [24, 28].

This study addresses this interdependency by proposing a unified framework that leverages TPE to concurrently optimize XGBoost hyperparameters while guiding the RFE process. By treating these tasks as a single, coupled optimization problem, this method effectively navigates the complex search space to identify a globally optimal solution [6, 7]. The key contributions of this study are:

- Unified optimization framework: The development of a TPE-optimized XGBoost-RFE algorithm solves the coupled problem of feature selection and hyperparameter tuning using the Optuna framework [6].
- Robust configuration: The successful optimization of nine key XGBoost hyperparameters provides a robust model configuration that simultaneously searches for the optimal feature subset [8].
- Dimensionality reduction: An empirical demonstration shows that the proposed method yields a 37% reduction in dimensionality (35 → 22 features) while significantly improving predictive performance compared to baselines [8, 9].
- Cross-model validation: The robustness of the selected subset is validated across multiple machine learning algorithms, proving superior accuracy over models trained on full or Pearson-correlated feature sets [3, 4, 9].

II. METHODOLOGY

A systematic coupled optimization framework was employed with three experimental configurations: Full feature, XGBoost+TPE tuning with all 35 input features, and TPE optimization via the Optuna framework to tune nine XGBoost hyperparameters. The baseline isolates the contribution of hyperparameter optimization without feature selection. XGBoost-RFECV with TPE tuning followed a decoupled approach, where RFE with Cross-Validation (RFECV) was first performed using a default XGBoost configuration that reduced the feature set to 29 variables, followed by TPE hyperparameter optimization on the reduced subset. Optuna TPE-XGBoost-RFECV was implemented as a truly coupled optimization strategy, where the TPE optimizer guided the RFE process by providing tuned hyperparameters at each iteration. In this nested framework, the outer TPE loop explored the hyperparameter space, while the inner RFE loop identified the optimal feature subset for each proposed configuration, ultimately selecting 22 features that jointly minimize the cross-validated Mean Absolute Error (MAE). This design ensures that feature importance evaluations during RFE are conducted with a well-configured model, preventing the erroneous elimination of informative predictors that may occur when using arbitrary default parameters.

A. Exploratory Data Analysis -AQ-Bench Dataset

The study utilized the AQ-Bench dataset [30] as the foundation for the Multi-Target Regression (MTR) framework [31], selected for its ability to predict diverse ozone concentration metrics. The dataset comprised approximately 5,500 samples that integrated heterogeneous data sources, including topographic, climatic, and anthropogenic features. The dataset successfully captured complex environmental interactions, establishing Nitrogen Dioxide levels, human activity proxies, and altitude as the primary drivers of ozone formation. It also ensured that the above-mentioned correlation was able to uncover the physical mechanisms governing air quality. To ensure integrity, a rigorous preprocessing pipeline was implemented to address critical data gaps. Missing target values, encoded as -999, were identified as a significant risk for skewing Mean Squared Error (MSE) and distorting feature weights. Consequently, these missing values were masked during the loss calculation.

A preliminary feature reduction was performed to eliminate irrelevant and redundant variables. This included the removal of constant and quasi-constant features (*VarianceThreshold* = 0.01), duplicated features, and the non-informative *id* column. A filter-based reduction using Pearson correlation was also applied, where features with a correlation less than 0.1 to the target variable were discarded, and from pairs of features with a correlation greater than 0.9, one was removed to mitigate multicollinearity. This preprocessing stage resulted in a final dataset of 35 input features, which was partitioned into training (70%), testing (20%), and validation (10%) sets. A copy of the full 35-feature dataset was retained for baseline model comparison.

B. The Mathematical Formulation

The core of the feature selection process relies on the XGBoost algorithm. The objective function for XGBoost at a given iteration t is defined as:

$$\mathcal{O}^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \quad (1)$$

where n is the number of samples, L is a differentiable convex loss function that measures the difference between the true target y_i and the predicted target $\hat{y}_i$, and Ω is a regularization term that penalizes the complexity of the model's trees. The prediction $\hat{y}_i$ is an additive combination of t regression trees f_k. The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \quad (2)$$

where T is the number of leaves in the tree, w is the vector of scores on the leaves, and γ and λ are regularization parameters controlling the penalty for tree complexity and leaf weights, respectively. This objective function is optimized to guide the feature importance calculation in each step of the RFE process.

C. The Integrated Optimization Algorithm

Algorithm 1 details the procedural implementation of the TPE-XGBoost-RFE framework. This algorithm operationalizes the nested-loop structure visualized in Figure 2, where an outer TPE loop guides an inner RFE loop. The outer loop, managed by the TPE optimizer, is responsible for intelligently exploring the hyperparameter space. The inner loop performs a complete RFE cycle (Figure 1) for each set of hyperparameters proposed by the TPE.

```
Algorithm 1: TPE_XGBOOST_RFE
procedure TPE_XGBOOST_RFE(D_train, H, N)
  study ←
  CreateOptunaStudy(sampler=TPESampler)
  for trial from 1 to N do
    h_current ←
    study.suggest_parameters(H)
    S_full ← GetAllFeatures(X)
    S_current ← S_full
    results_rfe ← []
    while |S_current| > 0 do
    //Start inner RFE loop
      model ← TrainXGBoost(S_current, y,
      h_current)
      mae_cv ← Get5FoldCrossValMAE(model,
      S_current, y)
      Append {mae_cv, S_current} to
      results_rfe
      importances ←
      GetFeatureImportances(model)
      feature_to_remove ←
      GetLeastImportantFeature(
      importances)
      S_current ← RemoveFeature(S_current,
      feature_to_remove)
    end while
    best_run_mae, best_run_subset ←
    FindMinMAE(results_rfe)
    study.report(best_run_mae,
    trial)
    // Return objective score to TPE
    study.set_user_attr("best_subset",
    best_run_subset)
  end for
  h* ← study.best_params
  S* ← study.best_trial.user_attrs[
  "best_subset"]
  return h*, S*
end procedure
```

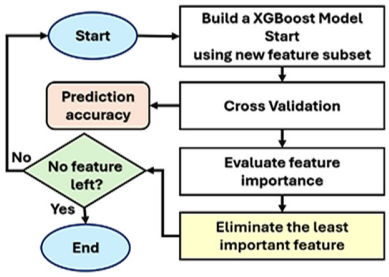

Fig. 1. Procedural flow of the inner RFE cycle.

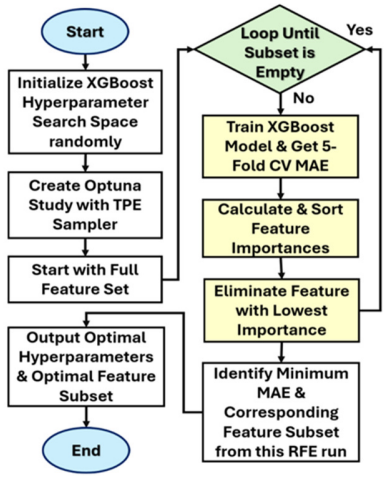

Fig. 2. The architecture of the nested TPE-XGBoost-RFE algorithm.

The critical link between the two loops is the objective function score. Upon completion of an inner RFE cycle, the minimum cross-validated MAE achieved during that cycle was returned to the TPE optimizer. This allows the TPE algorithm to learn which hyperparameter regions lead to better feature selection outcomes. Figure 3 shows the iterative search for this minimum, demonstrating how the TPE algorithm explores the search space to find better-performing hyperparameter combinations. This entire nested process is repeated for a predefined number of iterations to find the globally optimal hyperparameter set and its corresponding optimal feature subset.

- Input: Training data $D_{train} = \{X, y\}$, hyperparameter search space H, Number of TPE trials N
- Output: Optimal hyperparameters h^*, optimal feature subset S^*.

D. Model Evaluation and Benchmarking

Following the identification of the optimal 22-feature subset by the TPE-XGBoost-RFE framework, its generalizability and effectiveness were rigorously evaluated. Six distinct machine learning algorithms were trained on this reduced feature set: XGBoost, Random Forest (RF), LightGBM (LGBM), Gradient Boosting (GBM), Support Vector Regression (SVR), and K-Nearest Neighbors (KNN). To establish a performance baseline, the same set of models was also trained on the full 35-feature dataset.

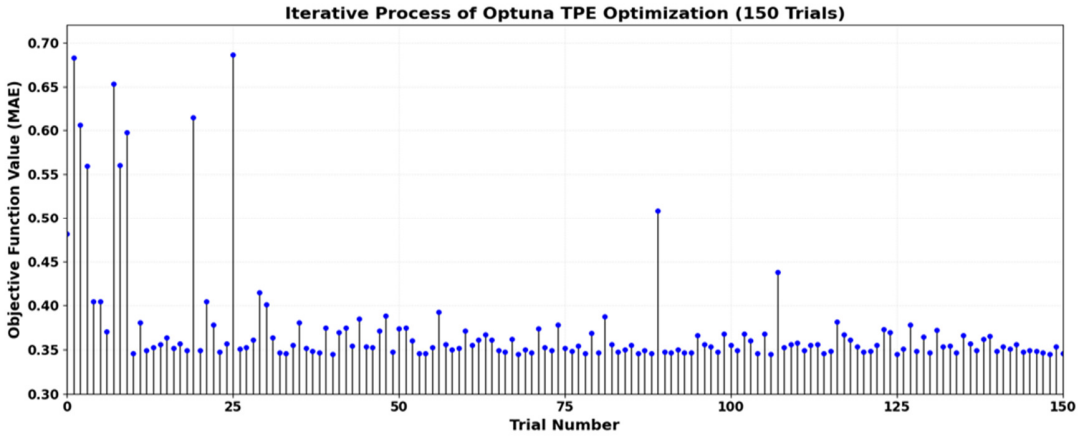

Fig. 3. Convergence of the Objective Function (MAE) across TPE Optimization Trials

The predictive performance of all models was quantified using three standard regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). Comparing the metrics from models trained on the selected subset against the baseline models allowed for a comprehensive assessment of the feature selection framework's impact on predictive accuracy and model parsimony.

III. RESULTS

This section presents the empirical outcomes of applying the TPE-XGBoost-RFE framework for ozone concentration prediction. The findings demonstrate the efficacy of the hyperparameter optimization, the subsequent feature selection, and the final predictive performance of various machine learning models.

A. Hyperparameter Optimization

The initial phase of this study focused on identifying the optimal hyperparameter configuration for the XGBoost model to guide the RFE process. Using the TPE algorithm within the Optuna framework, an optimized set of nine key hyperparameters was determined. Table I details the search range for each parameter and the final optimized values derived from the optimization process. Notably, the algorithm converged on a learning rate of 0.0066, n_estimators of 1800, and a max depth of 11, indicating a preference for a relatively deep model trained with a small learning rate over many iterations to achieve fine-grained performance.

TABLE I. XGBOOST HYPERPARAMETER SEARCH SPACE AND OPTIMIZED VALUES IDENTIFIED BY TPE

Hyperparameter	Range	Optimized value
Learning_rate	(0.0001–0.1)	0.0066
N_estimators	(100–3000)	1800
Max_depth	(3–15)	11
Min_child_weight	(1–10)	4
Gamma	(1e-7–1.0)	6.7E-07
Reg_alpha	(1e-8–1.0)	1.06E-07
Reg_lambda	(0.01–1.0)	0.5545
Colsample_bytree	(0.5–1.0)	0.6697
Subsample	(0.5–1.0)	0.612

B. Feature Selection and Dimensionality Reduction

The core contribution of the TPE-XGBoost-RFE method was its ability to concurrently perform feature selection. The process began with an initial set of 35 features. As shown in Table II, the integrated optimization framework identified a highly predictive subset consisting of only 22 features. This represents a substantial 37% reduction in feature dimensionality.

The selection process was guided by minimizing the MAE during cross-validation. This optimization yielded a dramatic improvement in model precision: the proposed framework achieved a cross-validation MAE of 2.372, compared to the Baseline (Full Features) model, which resulted in an MAE of 2.478. While the computational cost for the proposed method was higher (73.7 minutes) due to the extensive search space, the significant reduction in error justifies the investment.

TABLE II. IMPACT OF TPE OPTIMIZATION ON FEATURE DIMENSIONALITY AND CROSS-VALIDATION ERROR

Model	MAE	GPU time (minutes)	Feature number
Optuna TPE-XGBoost-RFECV (proposed)	2.396	282.9	22
XGBoost-RFECV +TPE tuning	2.422	2.7	29
All features XGBOOST + TPE tuning	2.502	2.3	35
BO-XGBoost-RFE [32]	2.410	-	22
XGBoost-RFE [32]	2.516	-	29

C. Comparative Model Performance

To validate the robustness and effectiveness of the selected 22-feature subset, its performance was evaluated across six distinct machine learning algorithms. The results were compared against two baseline scenarios: models trained on the original 35-feature set (All features) and models trained on features selected via a standard filter method ("After Pearson FS"). The results in Table III demonstrate a consistent and significant improvement in predictive accuracy for all six models trained on the feature subset identified by the TPE-XGBoost-RFE framework.

TABLE III. IMPACT OF TPE-XGBOOST-RFE ON THE PREDICTIVE PERFORMANCE OF VARIOUS MACHINE LEARNING MODELS

Model	Optuna TPE-XGBoost-RFECV (proposed)				BO-XGBoost-RFE			After Pearson FS			All Features (Baseline)		
	MAE	RMSE	R^2	GPU time (s)	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
XGB	2.372	3.232	0.721	21.8	2.386	3.281	0.718	2.59	3.462	0.675	2.478	3.368	0.698
RF	2.583	3.52	0.669	32.0	2.374	3.206	0.72	2.5	3.380	0.69	2.5	3.266	0.71
LGBM	2.618	3.568	0.66	14.8	-	-	-	2.747	3.693	0.63	2.637	3.537	0.66
GBM	2.656	3.62	0.65	52	-	-	-	2.784	3.743	0.62	2.695	3.614	0.645
SVR	2.882	3.927	0.588	8.2	2.676	3.631	0.659	2.912	3.871	0.583	2.677	3.62	0.636
KNN	2.958	4.031	0.566	0.4	2.801	3.808	0.606	2.873	3.837	0.601	2.846	3.834	0.601

The proposed method outperformed the baselines by a wide margin. Specifically, the XGBoost model trained on the optimized subset achieved an MAE of 2.372, an RMSE of 3.232, and an R^2 of 0.721. In contrast, the baseline XGBoost model (All features) yielded an MAE of 2.478, an RMSE of 3.368, and an R^2 of 0.698. The superiority of the TPE-XGBoost-RFE method is further highlighted when comparing the error magnitudes. The proposed model's MAE (2.372) is approximately 4.3% lower than that of the baseline model (2.478). The Pearson Feature Selection (FS) method failed to provide competitive results, yielding an MAE of 2.59 for XGBoost—worse than the unoptimized baseline. These findings empirically validate that the integrated optimization framework produces a significantly more parsimonious model (22 features vs. 35) that is universally more accurate.

IV. DISCUSSION

This study addressed the coupled challenge of feature selection and hyperparameter optimization for predicting long-term tropospheric ozone concentrations. The proposed TPE-XGBoost-RFE framework provides a superior solution, leading to highly parsimonious models with unambiguous and significant performance enhancements across all algorithms tested. The performance evaluation reveals a critical insight into the value of an integrated optimization strategy. The reduction of the feature space from 35 to 22 without loss of predictive power suggests that the original dataset contained a high degree of redundancy and noise. The proposed model's MAE of 2.372 compared to the baseline's 2.478 indicates that the selected subset captures the primary drivers of ozone formation much more effectively than the full feature set. This can be attributed to the wrapper-based approach, which evaluates features within the context of model performance, thereby capturing complex interactions that static correlation measures inherently miss. The failure of Pearson FS (MAE > 2.5) reinforces this, showing that linear filter methods are insufficient for this complex environmental modeling task.

A noteworthy observation is the relationship between error reduction and explained variance. Although the baseline model achieved a respectable R2 of 0.698, its high MAE (2.478) suggests that it was prone to significant local errors. The proposed framework improved R^2 to 0.721 while simultaneously reducing MAE to 2.372. This implies that the TPE-XGBoost-RFE framework did not simply prune variables but actively identified a subset that allows the model to generalize with exceptional precision. The simultaneous tuning of the hyperparameters ensured that this feature evaluation was guided by a well-configured model, preventing the erroneous elimination of important features.

These findings advocate for a deliberate and integrated approach to model development. The proposed Optuna TPE-XGBoost-RFECV framework was evaluated through three progressive optimization pipelines, which processed 15 targets individually and constituted the majority of the optimization time. This iterative, per-target optimization enabled target-specific feature selection and hyperparameter tuning, improving predictive performance. The search for optimal hyperparameter tuning of TPE coupled with XGBOOST-RFECV required 4.7 hours, and the final operational XGBoost model, which used the optimally tuned hyperparameters obtained from TPE-XGBOOST-RFECV, required only 21.8 s to retrain with new data. This demonstrates that the proposed Optuna TPE-XGBoost-RFECV framework is practically viable for operational deployment, offering state-of-the-art predictive performance with reasonable computational requirements for both the initial optimization and the operational model.

V. CONCLUSION

This research addresses a critical methodological gap in ML applications: the interdependent nature of feature selection and hyperparameter optimization. The novel TPE-XGBoost-RFE framework formalizes a solution by treating these two tasks as a single, coupled optimization problem. By nesting an RDE process within a TPE optimization loop, the proposed algorithm systematically navigates the high-dimensional search space to identify a globally optimal combination of features and hyperparameters.

The efficacy of the framework was empirically validated, confirming the central hypothesis that decoupled or overly simplistic approaches yield suboptimal solutions. The findings conclusively demonstrate that this unified strategy is superior, producing a more compact and performant feature set that benefits every algorithm tested.

The primary contribution of this study is a methodological advancement that establishes a new benchmark to build parsimonious but powerful predictive models. Building on this work, future directions should prioritize a deeper domain-specific analysis of the selected features. Investigating meteorological significance will bridge the gap between data-driven models and atmospheric science, improving interpretability. Furthermore, future research could explore alternative optimization algorithms to improve computational efficiency, solidifying this integrated paradigm for robust data-driven discovery.

REFERENCES

[1] E. K. Juarez and M. R. Petersen, "A Comparison of Machine Learning Methods to Forecast Tropospheric Ozone Levels in Delhi," *Atmosphere*, vol. 13, no. 1, Dec. 2021,, https://doi.org/10.3390/atmos13010046.

[2] M. A. M. Bhuiyan, R. K. Sahi, M. R. Islam, and S. Mahmud, "Machine Learning Techniques Applied to Predict Tropospheric Ozone in a Semi-Arid Climate Region," *Mathematics*, vol. 9, no. 22, Jan. 2021, Art. no. 2901, https://doi.org/10.3390/math9222901.

[3] M. J. Jiménez-Navarro, M. Martínez-Ballesteros, F. Martínez-Álvarez, and G. Asencio-Cortés, "Explaining deep learning models for ozone pollution prediction via embedded feature selection," *Applied Soft Computing*, vol. 157, May 2024, Art. no. 111504, https://doi.org/10.1016/j.asoc.2024.111504.

[4] L. Zhang *et al.*, "Explainable ensemble machine learning revealing the effect of meteorology and sources on ozone formation in megacity Hangzhou, China," *Science of The Total Environment*, vol. 922, Apr. 2024, Art. no. 171295, https://doi.org/10.1016/j.scitotenv.2024.171295.

[5] Z. Li, Y. Wang, J. Liu, and J. Xian, "Using machine learning to unravel chemical and meteorological effects on ground-level ozone: Insights for ozone-climate control strategies," *Environment International*, vol. 201, July 2025, Art. no. 109567, https://doi.org/10.1016/j.envint.2025.109567.

[6] Z. Liu *et al.*, "Comparison of machine learning methods for predicting ground-level ozone pollution in Beijing," *Frontiers in Environmental Science*, vol. 13, Apr. 2025, Art. no. 1561794, https://doi.org/10.3389/fenvs.2025.1561794.

[7] Q. Pan, F. Harrou, and Y. Sun, "A comparison of machine learning methods for ozone pollution prediction," *Journal of Big Data*, vol. 10, no. 1, May 2023, Art. no. 63, https://doi.org/10.1186/s40537-023-00748-x.

[8] Z. Xiao, Y. Lu, and G. Xiu, "Multi-Machine Learning Approaches to Modeling Small-Scale Source Attribution of Ozone Formation." Gases/Machine Learning/Troposphere/Chemistry (chemical composition and reactions), Mar. 05, 2025, https://doi.org/10.5194/egusphere-2025-160.

[9] K. Do, M. Mahish, A. K. Yeganeh, Z. Gao, C. L. Blanchard, and C. E. Ivey, "Emerging investigator series: a machine learning approach to quantify the impact of meteorology on tropospheric ozone in the inland southern California," *Environmental Science: Atmospheres*, vol. 3, no. 8, pp. 1159–1173, 2023, https://doi.org/10.1039/D2EA00077F.

[10] N. E. Selin *et al.*, "Global health and economic impacts of future ozone pollution," *Environmental Research Letters*, vol. 4, no. 4, Oct. 2009, Art. no. 044014, https://doi.org/10.1088/1748-9326/4/4/044014.

[11] S. Räss and M. C. Leuenberger, "Analysis and prediction of atmospheric ozone concentrations using machine learning," *Frontiers in Big Data*, vol. 7, Jan. 2025, https://doi.org/10.3389/fdata.2024.1469809.

[12] N. M. A. K. Jailani and G. C. Mara, "Ozone Concentration Forecasting: Assessing the Efficacy of MLP, DNN, and XGBoost in Environmental Bench-AQ Dataset," in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, Chikkaballapur, India, Apr. 2024, pp. 1–5, https://doi.org/10.1109/ICKECS61492.2024.10616879.

[13] C. Betancourt *et al.*, "Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties," *Geoscientific Model Development*, vol. 15, no. 11, pp. 4331–4354, June 2022, https://doi.org/10.5194/gmd-15-4331-2022.

[14] L. Castro-Martín, M. del Mar Rueda, R. Ferri-García, and C. Hernando-Tamayo, "On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures," *Mathematics*, vol. 9, no. 23, Jan. 2021, Art. no. 2991, https://doi.org/10.3390/math9232991.

[15] T. L. He *et al.*, "Deep Learning to Evaluate US NOx Emissions Using Surface Ozone Predictions," *Journal of Geophysical Research: Atmospheres*, vol. 127, no. 4, 2022, Art. no. e2021JD035597, https://doi.org/10.1029/2021JD035597.

[16] Y. Wang and X. S. Ni, "A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization." arXiv, Jan. 24, 2019, https://doi.org/10.48550/arXiv.1901.08433.

[17] D. Akritidis *et al.*, "A deep stratosphere-to-troposphere ozone transport event over Europe simulated in CAMS global and regional forecast systems: analysis and evaluation," *Atmospheric Chemistry and Physics*, vol. 18, no. 20, pp. 15515–15534, Oct. 2018, https://doi.org/10.5194/acp-18-15515-2018.

[18] D. Effrosynidis and A. Arampatzis, "An evaluation of feature selection methods for environmental data," *Ecological Informatics*, vol. 61, Mar. 2021, Art. no. 101224, https://doi.org/10.1016/j.ecoinf.2021.101224.

[19] C. Ferhatoglu and B. A. Miller, "Choosing feature selection methods for spatial modeling of soil fertility properties at the field scale," in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, Seattle, WA, USA, Nov. 2022, https://doi.org/10.1145/3557915.3565531.

[20] A. M. A. Zeyad and A. Biradar, "A-Hybrid-Text-Summarization-Approach-Using-Neural-Networks-and-Metaheuristic-Algorithms.pdf," *International Journal of Safety and Security Engineering*, vol. 13, no. 3, pp. 479-489, 2023.

[21] A. M. A. Zeyad and A. Biradar, "Abstractive Text Summarization: A Hybrid Evaluation of Integrating Flan-T5 (Dual Framework) with Pegasus Reveals Conciseness Advantages across Diverse Datasets," *International Journal of Computer Network and Information Security*, vol. 17, no. 6, pp. 98–115, Dec. 2025, https://doi.org/10.5815/ijcnis.2025.06.07.

[22] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[23] L. Kovács, "Feature selection algorithms in generalized additive models under concurvity," *Computational Statistics*, vol. 39, no. 2, pp. 461–493, Apr. 2024, https://doi.org/10.1007/s00180-022-01292-7.

[24] J. P. Chaudhari *et al.*, "Recursive Feature Elimination and Optimized Hybrid Ensemble Approach for Early Heart Disease Prediction," *Advances in Technology Innovation*, vol. 10, no. 1, pp. 58–71, Jan. 2025, https://doi.org/10.46604/aiti.2024.13825.

[25] K. R. Swetha and M. A. K. N. Jailani, "Multi-Target Ozone Prediction Using Hybrid GWO+SVM-RFE Feature Selection," in *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*, Bengaluru, India, Dec. 2025, pp. 1–5, https://doi.org/10.1109/NMITCON65824.2025.11188068.

[26] K. R. S. Kumar and M. A. K. N. Jailani, "Hybrid Feature Selection Using ACO+SVM-RFE for Multi-Target Regression in Ozone Modeling," in *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*, Bengaluru, India, Dec. 2025, https://doi.org/10.1109/NMITCON65824.2025.11188161.

[27] J. Adkins, M. Bowling, and A. White, "A Method for Evaluating Hyperparameter Sensitivity in Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 124820–124842, Dec. 2024, https://doi.org/10.52202/079017-3964.

[28] M. Rezaali, M. S. Jahangir, R. Fouladi-Fard, and D. Keellings, "An ensemble deep learning approach to spatiotemporal tropospheric ozone forecasting: A case study of Tehran, Iran," *Urban Climate*, vol. 55, May 2024, Art. no. 101950, https://doi.org/10.1016/j.uclim.2024.101950.

[29] L. Li, "Towards Efficient Automated Machine Learning." 2020

[30] C. Betancourt, T. Stomberg, R. Roscher, M. G. Schultz, and S. Stadtler, "AQ-Bench: a benchmark dataset for machine learning on global air quality metrics," *Earth System Science Data*, vol. 13, no. 6, pp. 3013–3033, June 2021, https://doi.org/10.5194/essd-13-3013-2021.

[31] M. A. K. Jailani N and G. C. Mara, "Feature Selection in Ozone Feature Space Impacts Performance in Gradient Boosting, Random Forest, Xgboost and Adaptive Boosting Regressors," in *2024 International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bengaluru, India, Feb. 2024, pp. 1–6, https://doi.org/10.1109/ICCTAC61556.2024.10581262.

[32] B. Zhang, Y. Zhang, and X. Jiang, "Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm," *Scientific Reports*, vol. 12, no. 1, June 2022, Art. no. 9244, https://doi.org/10.1038/s41598-022-13498-2.