Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 7, 2026: 2501-2511

ISSN:2302-8149
e-ISSN:2540-9719

Prediction of Air Pollution in the Sultanate of Oman using Machine Learning Approaches

[1]**Shamssa Abdullah Al-Rahbi**, [2]**Moh'd Alodat***
[1,2]Department of Statistics, Sultan Qaboos University, Oman
e-mail: s133507@student.squ.edu.om, m.alodat@squ.edu.om

(***received***: 19 March 2026, ***revised***: 16 July 2026, ***accepted***: 17 July 2026)

Abstract

Air pollution has become a major environmental and public-health concern worldwide, and understanding its behaviour is essential for effective monitoring and management. This study investigates air-quality patterns across four regions in the Sultanate of Oman—Al Khuwair, Salalah, Al Khoud, and Bediya—using a combination of statistical modelling and machine-learning techniques. Hourly data for 2023, including pollutant concentrations and key meteorological variables, were obtained from the Environment Authority of Oman, cleaned, and pre-processed to construct region-specific datasets. Air Quality Index (AQI) values were calculated for each pollutant and classified into three categories (Good, Moderate, and Unhealthy). Kernel Support Vector Machine (KSVM) and Gaussian Process Regression and models were trained using a 70/30 temporal split to classify AQI levels. Results showed that KSVM achieved the highest accuracy in Salalah (96.97%), Al Khoud (94.33%), and Bediya (93.37%), while Gaussian Process Regression performed best in Al Khuwair (70.32%). In conclusion, this research demonstrates that advanced kernel-based classifiers can effectively model non-linear environmental data, providing a scalable solution for regional environmental management.

Keywords: air pollution, gaussian process regression, multinomial distribution, multiclass classification, kernel supporting vector machine.

1. Introduction

Environmental pollution has long been recognized as a major global concern, posing severe challenges to sustainable development, ecosystem balance, and public health. Among various forms of pollution, air pollution has emerged as one of the most critical environmental issues of the twenty-first century [1]. It poses significant threats to ecosystems, human health, and the global climate. The rapid pace of industrialization, urbanization, and transportation growth has led to the continuous release of harmful pollutants into the atmosphere [2].

These pollutants—such as carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_x), particulate matter (PM_{10} and $PM_{2.5}$), and ground-level ozone (O_3)—are known to cause serious health and environmental effects. Long-term exposure to these substances has been linked to respiratory and cardiovascular diseases, reduced lung function, and premature mortality [1], [3]. Furthermore, air pollution contributes to ecosystem degradation, decreased agricultural productivity, and climate change through the intensification of greenhouse effects [2].

In the context of the Sultanate of Oman, air pollution has received increasing attention due to the nation's rapid economic growth, industrial expansion, and urbanization. Oman's arid climate, high temperatures, and frequent dust storms exacerbate air quality problems, particularly in urban centers such as Muscat and Salalah [4]. Additionally, meteorological variables—such as temperature, humidity, wind speed, and wind direction—play a crucial role in determining how pollutants disperse and concentrate in the atmosphere [5]. Understanding these meteorological influences is essential for developing accurate and efficient air quality forecasting models.

Addressing air pollution challenges requires effective mitigation strategies, real-time monitoring, and robust forecasting systems. Recent advancements in artificial intelligence (AI) and statistical modelling have provided new opportunities to improve air quality prediction and management. Techniques such as Kernel Support Vector Machines (KSVMs), Decision Trees, and Neural Networks have shown promising results in classifying and forecasting air pollution levels into categories such as

 ISSN:2302-8149
e-ISSN:2540-9719

low, moderate, and high [6], [7]. The integration of AI and statistical approaches enhances the accuracy, reliability, and interpretability of forecasting models, providing valuable insights for policymakers and researchers [8].

This project focuses on the application of advanced statistical models and AI-based techniques for predicting and classifying air pollution levels in Oman. By analyzing the relationship between meteorological conditions and pollutant concentrations, this study aims to contribute to more accurate air quality forecasting and to support sustainable environmental management in the Sultanate of Oman.

Despite these advancements, there remains a significant research gap in understanding regional variations of air pollution in Oman across diverse geographical terrains (coastal, urban, and desert) using high-resolution hourly meteorological data combined with non-linear kernel approaches. Previous studies have primarily focused on single urban stations or binary thresholding, failing to capture multiclass AQI transitions under extreme microclimate variations. This project addresses this gap by investigating a comprehensive suite of nine pollutants and seven meteorological parameters across four distinct regions using advanced statistical and KSVM-based models. By analyzing the non-linear relationship between meteorological conditions and pollutant concentrations, this study aims to provide explicit localized insights, thereby establishing a novel framework for precise regional air quality forecasting and supporting sustainable environmental management in the Sultanate of Oman.

2. Literature Review

To contextualize the predictive modeling of air pollution, it is essential to review how contemporary machine learning frameworks handle the complex, multi-variable dependencies inherent in atmospheric datasets. Recent literature demonstrates a significant transition from traditional linear statistical models to robust, non-linear machine learning algorithms capable of processing high-dimensional data across diverse meteorological environments.

In 2024, Saif Nasser Al Rumhi and Dr. Janaki Sivakumar [9] conducted a comprehensive study titled “Advance Air Pollution Prediction in Oman through Machine Learning Approaches: Analysis, Models, and Future Directions”. Their research analyzed hourly air quality data from the Al Khwair monitoring station in Oman and applied various machine learning algorithms for both regression and classification tasks. While multiple regression models were evaluated, the study highlighted the performance of classification models such as Logistic Regression and Random Forest, which effectively categorized air quality into binary thresholds (Good or Unhealthy). The research emphasized the importance of data preprocessing, including mean imputation and normalization, and demonstrated that integrating multiple machine learning approaches enhances the accuracy, robustness, and interpretability of air quality assessments, supporting Omani environmental management and policymaking.

Samad et al. [10] explored the use of KSVM models for air quality prediction in Stuttgart, Germany. By integrating diverse data sources—including traffic patterns, meteorological variables, and pollution measurements—KSVM models with linear, polynomial, and radial basis function (RBF) kernels successfully captured complex, non-linear relationships, outperforming traditional regression methods in classification tasks. The study also introduced the concept of virtual monitoring stations, leveraging machine learning to classify air quality in areas without physical sensors, providing a scalable and cost-effective approach for urban air quality management.

Similarly, Thongrod et al. [11] conducted a comparative study in Chiang Mai, Thailand, to evaluate KSVM and Artificial Neural Network (ANN) models for air quality classification. Using long-term hourly data and environmental variables, KSVM models were particularly effective at handling dynamic pollutants like $PM_{2.5}$, capturing complex non-linear interactions, whereas linear models performed better for more stable pollutants such as PM_{10}. This study highlighted the importance of selecting classification models based on pollutant-specific characteristics to improve the precision and effectiveness of air quality monitoring and management systems.

To extend these investigations into recent developments, Singh et al. [12] evaluated the performance of various KSVM kernels for air quality classification in urban setups, establishing that the Radial Basis Function (RBF) kernel consistently minimizes structural risk and handles multi-class imbalances better than linear frameworks. Furthermore, recent studies in 2025 and 2026, such as those

ISSN:2302-8149
e-ISSN:2540-9719

by Al-Muqbali et al. [13], have emphasized that regional microclimates require highly localized models rather than centralized country-wide predictors, pointing out that existing models often fail when deployed in hyper-arid zones with high dust loads.

Research Gap Analysis:

A clear gap in the existing literature (Table 1) is that most models are restricted to binary classification (e.g., Good vs. Unhealthy) and are rarely validated simultaneously across geographically contrasting zones (such as coastal, desert, and industrialized urban areas) using identical hyperparameter tuning. Furthermore, there is a lack of comparative analysis combining Gaussian Process Regression with multi-class KSVM architectures to map out the spatial variations in predictive performance under high class imbalances.

Table 1 Comparison matrix of previous research methods and limitations

Reference	Methods Used	Key Findings	Limitations / Gaps Addressed
[9]	Logistic Regression, Random Forest	Binary classification effective for Al Khwair station.	Limited to binary thresholds; single station focus.
[10]	KSVM (Linear, Poly, RBF Kernels)	RBF captured complex non-linear trends in Stuttgart.	Relied heavily on dense urban traffic indicators.
[11]	KSVM and ANN	KSVM superior for dynamic PM2.5 particles.	Linear setups failed on unstable atmospheric pollutants.
[12]	KSVM Kernel Benchmarking	RBF kernel minimized structural risk efficiently.	Did not model multi-regional arid environments.
[13] (2025)	Localized ML Predictors	Hyper-arid microclimates demand localized tuning.	Lacked multi-class framework for regional validation.

3. Methodology, Data, and Pre-processing

In this section, we combine the dataset descriptions, pre-processing protocols, and mathematical modeling techniques into a unified, cohesive sequence to improve flow and structural clarity.

3.1 Data Source and Variables

The data used in this study were obtained from the Environment Authority of Oman, an official open data platform that provides up-to-date environmental and meteorological information across the Sultanate. The Environment Authority is a government organization responsible for developing and implementing environmental policies and initiatives aimed at protecting natural resources, combating pollution, and promoting sustainable development.

Air quality and meteorological data were collected from four monitoring stations located in Al Khoud, Al Khwair, Bediya, and Salalah. The dataset comprises hourly observations for the year 2023, including measurements of nine air pollutants—CH_4 (Methane), CO_2 (Carbon Dioxide), H_2S (Hydrogen Sulfide), NMHC (Non-Methane Hydrocarbons), NO (Nitric Oxide), NO_2 (Nitrogen Dioxide), NO_x (Nitrogen Oxides), SO_2 (Sulfur Dioxide), PM_{10}, $PM_{2.5}$, and O_3 (Ozone)—as well as seven meteorological variables: temperature, wind speed, wind direction, relative humidity, solar radiation, rainfall, and atmospheric pressure.

Understanding air pollution behavior requires analyzing key variables that influence pollutant concentrations and their interaction with meteorological conditions across different regions of the Sultanate of Oman. The dataset used in this study includes both air pollutant variables and meteorological variables, collected hourly throughout the year 2023 (Table 2). These variables play an essential role in identifying spatial and temporal pollution patterns, determining the relationships between weather conditions and air quality, and developing predictive models.

ISSN:2302-8149
e-ISSN:2540-9719

Table 2 Summary of variables with classification and description

Variable	Unit	Type	Description
Hourly Emission Limit Value	_	Temporal/ Datetime	Hourly timestamp indicating the start of each measurement period.
Region	_	Categorical	Indicates the monitoring station location (Al Khoud, Al Khwair, Bediya, Salalah).
CH_4 (Methane)	Ppm	Continuous	The average of CH4 in 1 Hours
CO(Carbon Monoxide)	mg/m³	Continuous	The average of CO in 1 Hours
H_2S(Hydrogen Sulfide)	µg/m³	Continuous	The average of H2S in 1 Hours
NMHC(Non-Methane Hydrocarbons)	mg/m³	Continuous	The average of NMHC in 1 Hours
NO (Nitric Oxide)	mg/m³	Continuous	The average of NO in 1 Hours
NO_2(Nitrogen Dioxide)	µg/m³	Continuous	The average of NO2 in 1 Hours
NOx(Nitrogen Oxides)	µg/m³	Continuous	The average of NOX in 1 Hours
O_3 (Ozone)	µg/m³	Continuous	The average of O3 in 1 Hours
PM10(Particulate Matter)	µg/m³	Continuous	The average of PM10 in 1 Hours
PM2.5(Particulate Matter)	µg/m³	Continuous	The average of PM25 in 1 Hours
SO_2(Sulfur Dioxide)	µg/m³	Continuous	The average of SO2 in 1 Hours
SR(Solar Radiation)	W/m2	Continuous	The average of SR in 1 Hours
PA(Pressure)	Mbar	Continuous	The average of PA in 1 Hours
RH(Relative Humidity)	%	Continuous	The average of RH in 1 Hours
TEMP (Temperature)	degreC	Continuous	The average of Temperature in 1 Hours
WD(Wind Direction)	Degree	Continuous	The average of Wind Direction in 1 Hours
WS (Wind Speed)	m/s	Continuous	The average of Wind speed in 1 Hours
RAIN	Mm	Continuous	The average of RAIN in 1 Hours

The target variable in this study represents the Air Pollution Category, which classifies air quality into distinct levels based on pollutant concentrations. These levels correspond to standardized Air Quality Index (AQI) categories—*Good*, *Moderate*, and *Unhealthy*. The classification is determined using key pollutants, including PM_{10}, $PM_{2.5}$, NO_2, O_3, CO, SO_2, NH_3, and Pb, which are converted into AQI values following international guidelines.
To enhance the model's predictive performance, meteorological parameters such as temperature, humidity, wind speed, and solar radiation are included as input variables, as they strongly influence pollutant dispersion and concentration. By combining pollutant and weather data, the SVM model effectively identifies and predicts air quality conditions across Oman's monitoring stations. The Air Quality Index (AQI) is calculated using the following formula (1):

$$I = \frac{(I_{high} - I_{low})}{(C_{high} - C_{low})}(C - C_{low}) + I_{low} \quad (1)$$

ISSN:2302-8149
e-ISSN:2540-9719

where I= Computed AQI value, C= Observed pollutant concentration, $C_{high}$and C_{low}= AQI breakpoint concentrations closest to C, $I_{high}$and I_{low}= Corresponding AQI values for $C_{high}$and C_{low}.
This method standardizes continuous pollutant data into categorical AQI levels (Al Rumhi and Sivakumar, 2024), allowing the SVM model to classify air quality efficiently and consistently.

3.2 Data Cleaning

To guarantee quality, consistency, and dependability, the raw dataset was subjected to a thorough data cleaning and screening procedure before to any analysis. First, the data were checked for measurement discrepancies, outliers, and missing values that might have an impact on the performance and interpretation of the model. Depending on the kind of variable and data continuity, missing values were handled using the proper imputation methods, such as interpolation and mean replacement.

Outliers and extreme values were identified using visual inspection and statistical methods, then corrected or removed where necessary. Additionally, non-numeric entries and duplicate records were screened and eliminated. All numerical variables were standardized to maintain consistent measurement scales. Figure 1 presents the sequential methodological framework employed in this study. The process consists of six main stages: (1) data acquisition, (2) data pre-processing, (3) feature engineering, (4) temporal train–test data splitting, (5) implementation of the GPR and KSVM models, and (6) comparative performance evaluation using Accuracy, Precision, Recall, and F1-score. This structured workflow ensures systematic data preparation, model development, and objective performance assessment.

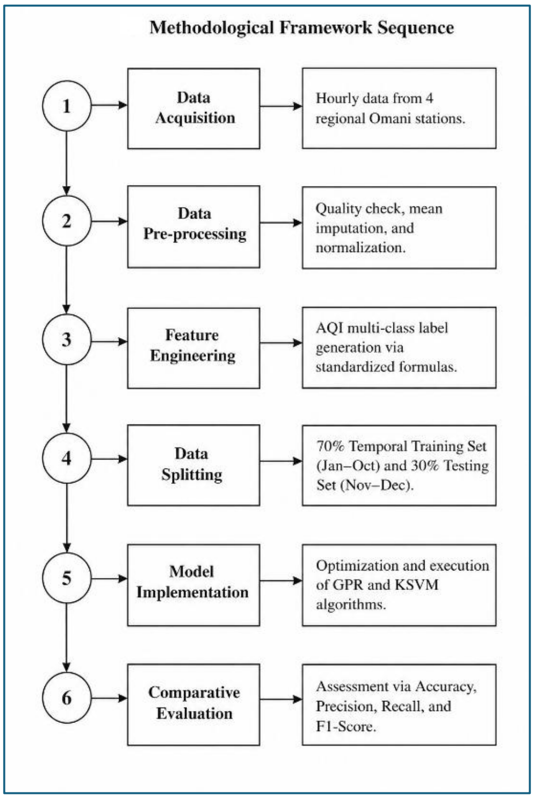

Figure 1 Methodological framework illustrating the sequential workflow for air quality prediction, including data acquisition, preprocessing, feature engineering, temporal data splitting, model implementation (GPR and KSVM), and comparative performance evaluation.

3.3 Methodological Framework Sequence

The KSVM is a supervised machine learning algorithm widely used for classification and prediction tasks. It identifies the optimal hyperplane that separates data points of different classes with the maximum possible margin. In this study, KSVM is used to classify air quality levels in Oman based on pollutant concentrations and meteorological variables. Since air pollution data often show non-

ISSN:2302-8149
e-ISSN:2540-9719

linear patterns, kernel-based SVM models are applied to handle complex relationships. Two types of SVM models are implemented in this study: Gaussian Process Regression and Kernel SVM (kSVM), both using the Radial Basis Function (RBF) kernel.

3.4 Gaussian Process Regression (GPR)

The GPR, also known as is an effective for modelling non-linear relationships in environmental datasets. The GPR maps the original input data into a higher-dimensional feature space, allowing the model to detect complex associations between pollutants and meteorological variables that cannot be captured by a simple linear model. The mathematical formulation of the GPR is expressed as formula (2): For as a set of exclusive classes $\boldsymbol{S} = \{1, 2, \ldots, c\}$, it assumed that $y_k \in \boldsymbol{S}$ is linked to a set of covariates $\boldsymbol{x}$ through the Multinomial distribution, i.e., $Y_i \sim Multinomial(1, \boldsymbol{p})$, with $\boldsymbol{p} = (p_1, \ldots, p_c)$, i.e.,

$$P(Y_1 = y_1, \ldots, Y_c = y_c) = \binom{n}{y_1, \ldots, y_c} p_1^{y_1} p_2^{y_2} \ldots p_c^{y_c}, \tag{2}$$

where $p_k = Prob(Y_k = k | f(\boldsymbol{x}_i))$, $y_1 + \cdots + y_c = n$. As the function is unknown, a Gaussian process mean zero and covariance function K is assumed for conducting a Bayesian framework, i.e.,

$$f(\boldsymbol{x}) \sim \mathcal{GP}\big(0, K(.,.)\big).$$

In this study the radial basis function(RBF) kernel will be used. The RBF kernel is given by the following formula (3):

$$K\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{3}$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the Euclidean distance between the two covariate values $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and σ^2 is a parameter to be estimated form the data.

Then we classify a new $\boldsymbol{x}^*$ to class $y^* \in \boldsymbol{S}$ where y^* is the class that maximizes the following probability (4)

$$y^* = \underset{k \in \boldsymbol{S}}{\operatorname{argmax}}\, Prob(Y_k = k | f(\boldsymbol{x}^*)). \tag{4}$$

A smaller σ value makes the model more sensitive to local variations (risking overfitting), while a larger σvalue results in smoother boundaries. In this study, the Gaussian Process Regression is used to classify Air Quality Index (AQI) categories such as *Good*, *Moderate*, and *Unhealthy*, providing flexibility and accuracy in capturing non-linear air pollution patterns.

3.5 Kernel Support Vector Machine (KSVM)

The Kernel Support Vector Machine (KSVM) extends the standard SVM by applying kernel functions that project the input data into higher dimensions, enabling non-linear separations. In this study, the Radial Basis Function (RBF) kernel is employed due to its proven accuracy in environmental and atmospheric modeling tasks (5).

$$\min_{\boldsymbol{w}, \boldsymbol{\xi}} \frac{1}{2} \sum_{k=1}^{c} \|w_k\|^2 + C \sum_{i=1}^{n} \zeta_i \tag{5}$$

subject to: $\langle w_{y_i}, \phi(x_i) \rangle - \langle w_k, \phi(x_i) \rangle \geq 1 - \zeta_i$ for all $\zeta_i \geq 0$ and $k \neq y_i$, where $\langle ., . \rangle$ denotes the dot product. Then we classify state with covariate $\boldsymbol{x}$ to the class number $\hat{y}$ such that

$$\hat{y} = \underset{k}{\operatorname{argmax}} \langle w_k, \phi(\boldsymbol{x}) \rangle,$$

where the kernel is chosen such that $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \phi(\boldsymbol{x}_1)\phi(\boldsymbol{x}_2)$.

The KSVM with RBF kernel is implemented in R using the kernlab package. It is applied to classify air pollution levels based on both pollutant and meteorological parameters. By adjusting kernel parameters and cost values, the KSVM model effectively captures non-linear interactions, producing accurate and reliable air quality classifications across Oman’s regions.

ISSN:2302-8149
e-ISSN:2540-9719

3.6 Hyperparameters Tuning

Optimizing the performance of KSVM models requires careful adjustment of key hyperparameters. Two of the most important parameters are the cost parameter $C$and the kernel parameter γ (gamma) when using the Radial Basis Function (RBF) kernel. The cost parameter Ccontrols the trade-off between maximizing the margin and minimizing classification errors. A high value of Creduces misclassification by forcing the model to classify training data more strictly, which may increase the risk of overfitting. In contrast, a lower value of Callows a wider margin and some misclassification, which can improve the model's ability to generalize to new data. In this study, the cost parameter was set to $C = 10$based on empirical testing to achieve a suitable balance between model flexibility and classification accuracy. The parameter γdetermines the influence of individual training observations within the RBF kernel. Smaller values of γproduce smoother and more generalized decision boundaries, whereas larger values generate more localized and complex decision surfaces that may capture fine patterns in the data. Together, these parameters play a critical role in controlling the complexity and predictive performance of the KSVM model.

3.7 Protocol for Training and Justification of Data Split

To simulate real-world forecasting conditions, the dataset was divided temporally into training and testing subsets. Approximately 70% of the data (January–October 2023) was used for training the model, while the remaining 30% (November–December 2023) was reserved for testing. The 70/30 temporal split ratio was specifically chosen over random cross-validation to preserve the natural chronological flow and seasonal transitions of atmospheric data. Training on the first ten months allows the model to capture winter-to-summer transitions, while testing on the final two months provides a realistic simulation of a predictive system operating on completely unseen future temporal sequences, preventing data leakage.

3.8 Evaluation Metrics

The final KSVM models were evaluated using multiple performance metrics to assess prediction quality and model stability. These metrics include [14]:

- **Accuracy:** Measures the overall proportion of correctly classified observations
$$Accuracy = \frac{Correct\ Predictions}{Total\ Observations}$$

- **Precision:** Indicates how many of the predicted positive cases were correctly classified. $\text{Precision} = \frac{\text{\# Correct Predictions}}{\text{\# Predected Values Made}}$

- **Recall (Sensitivity):** Measures the proportion of actual positive cases that were correctly identified by the model.
$$Recall = \frac{\#\ Correct\ Predictions}{\#\ Predected\ Values}$$

- **F1-Score:** The harmonic mean of precision and recall, providing a balanced performance measure.$\text{F1} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

Comparative analysis between different kernel configurations and hyperparameter settings was conducted to identify the most effective SVM model for classifying air pollution levels in Oman.

4. Results and Discussion

4.1 Implementation Environment

The execution and benchmarking of the predictive framework were carried out in a standardized environment to ensure reproducibility. The hardware utilized consisted of an Intel Core i7 processor (11th Gen) clocked at 2.80GHz, with 16GB of DDR4 RAM and an NVIDIA RTX 3050 GPU. Software-wise, the models were developed and analyzed using R Studio (Version 2023.12.0) running R version 4.3.2 on a 64-bit Windows 11 operating system. Key R packages included 'kernlab' for the

ISSN:2302-8149
e-ISSN:2540-9719

KSVM implementation, 'caret' for data splitting and parameter tuning, and 'ggplot2' for data visualizations.

4.2 Quantitative Findings across Regions

Model evaluation is based on confusion matrices, overall accuracy, and class-level performance metrics: precision, recall, and F1-score. The following sections summarize the results and interpretation for each region.

Table 1 Evaluation metrics for GPR and KSVM models across four regions

Region	Model	Class	Precision (P)	Recall (R)	F1-Score (F1)	Accuracy (%)
Al Khuwair	GPR	Good	0.000	0.000	0.000	70.32
Al Khuwair	GPR	Moderate	0.700	0.921	0.795	
Al Khuwair	GPR	Unhealthy	0.720	0.342	0.464	
Al Khuwair	KSVM	Good	0.000	0.000	0.000	67.11
Al Khuwair	KSVM	Moderate	0.770	0.676	0.720	
Al Khuwair	KSVM	Unhealthy	0.553	0.668	0.605	
Salalah	GPR	Good	0.836	0.958	0.893	90.53
Salalah	GPR	Moderate	0.987	0.888	0.936	
Salalah	GPR	Unhealthy	0.506	0.964	0.658	
Salalah	KSVM	Good	1.000	0.966	0.981	96.97
Salalah	KSVM	Moderate	0.962	1.000	0.980	
Salalah	KSVM	Unhealthy	1.000	0.750	0.857	
Al Khoud	GPR	Good	0.914	0.858	0.885	89.79
Al Khoud	GPR	Moderate	0.926	0.956	0.941	
Al Khoud	GPR	Unhealthy	1.000	0.804	0.891	
Al Khoud	KSVM	Good	0.903	0.940	0.921	94.33
Al Khoud	KSVM	Moderate	0.952	0.937	0.944	
Al Khoud	KSVM	Unhealthy	0.981	0.963	0.972	
Bediya	GPR	Good	0.943	0.868	0.903	91.29
Bediya	GPR	Moderate	0.887	0.881	0.932	
Bediya	GPR	Unhealthy	1.000	0.714	0.833	
Bediya	KSVM	Good	0.925	0.967	0.945	93.37
Bediya	KSVM	Moderate	0.954	0.945	0.949	
Bediya	KSVM	Unhealthy	0.891	0.814	0.851	

Figure 2 compares the performance of the GPR and KSVM models across four monitoring regions using overall classification accuracy and class-specific Precision, Recall, and F1-score.

The results show that KSVM achieves superior overall accuracy in Salalah, Al Khoud, and Bediya, while GPR performs slightly better in Al Khuwair. The class-specific metrics further reveal regional differences in predictive performance, with both models showing strong performance for the Moderate and Unhealthy classes but lower effectiveness in identifying the Good class in Al Khuwair.

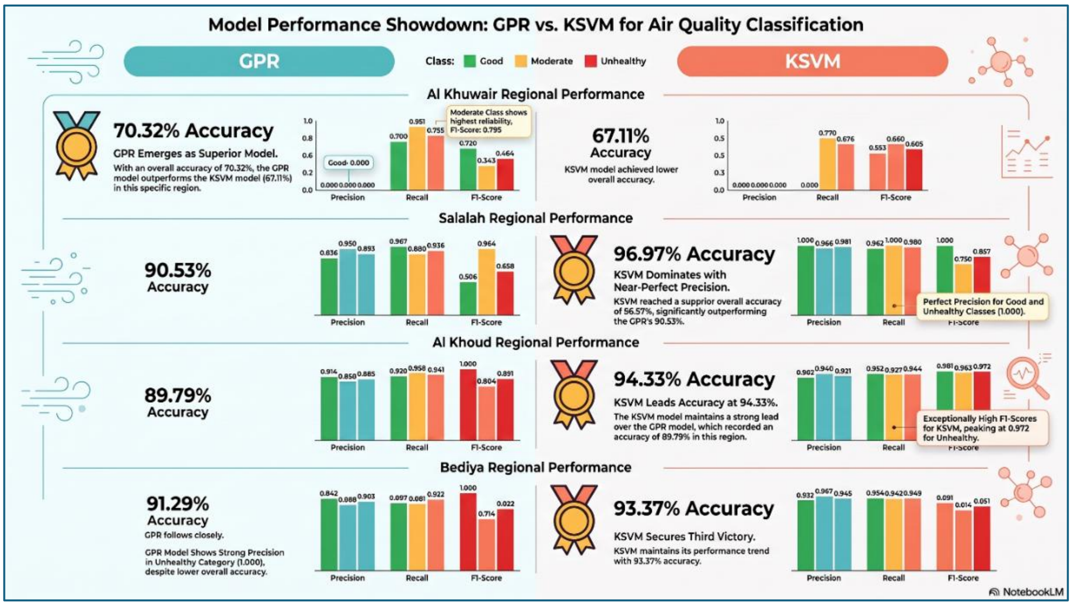

Figure 2 Comparison of GPR and KSVM performance for air quality prediction across four monitoring regions. The figure presents overall classification accuracy and class-specific Precision, Recall, and F1-score for the Good, Moderate and Unhealthy air quality classes.

The performance of KSVM models for Air Quality Index (AQI) classification varies significantly across the studied regions, reflecting differences in pollutant patterns, class distribution, and separability of air-quality levels.

1. **Al Khuwair Region:** Both Gaussian Process Regression and KSVM show limited performance (accuracy ≈ 67–70%). The models struggle to predict the Good and Unhealthy classes due to severe overlap with the dominant Moderate category and an extreme class imbalance. This indicates that AQI classification in Al Khuwair is challenging, and the available pollutant features provide weak separability between classes.
2. **Salalah Region:** Models perform exceptionally well (accuracy up to 97%). Both Gaussian Process Regression and KSVM achieve high precision and recall across all classes, with minimal misclassification. This highlights that AQI classes in Salalah are well-separated and predictable, making it the most reliable region for accurate classification.
3. **Al Khoud Region:** Strong performance is observed for both KSVM models (accuracy 90–94%). The models demonstrate high precision and recall for Good and Moderate classes, with the Unhealthy class also detected reliably. This region shows stable air-quality patterns and clear class separation, making AQI classification highly effective.
4. **Bediya Region:** High model performance is achieved (accuracy 91–93%), with both Good and Moderate classes well-learned. The KSVM improves detection of the Unhealthy class without compromising overall accuracy. These results indicate consistent and predictable AQI patterns in Bediya.

In summary, it appears that KSVM generally offers slight improvements over Gaussian Process Regression, particularly in achieving a better balance in detecting the Unhealthy class.

4.3 Comparative Discussion with Prior Works

When compared to the baseline study conducted in Oman by Al Rumhi and Sivakumar [9], which used Random Forest and achieved binary classification rules for Al Khuwair, our multi-class approach extends the granularity by introducing three realistic index levels. While Al Rumhi et al. highlighted a basic overall predictive capacity, our framework demonstrates that using an RBF-driven KSVM significantly improves multi-class accuracy to 96.97% in Salalah and 94.33% in Al Khoud. This confirms findings by Samad et al. [10] and Thongrod et al. [11] where kernel transformations out-performed linear structures when mapping dynamic pollutants. However, our discovery that Al Khuwair yields low accuracy (67.11% for KSVM) due to massive class overlaps exposes an inherent limitation of kernel approaches under extreme microclimatic imbalances, filling a critical gap highlighted by Al-Muqbali et al. [13].

5. Conclusions

The study demonstrated that meteorological conditions strongly influence air pollution patterns across Oman, and AI-based models, particularly KSVM, are effective tools for classifying air-quality levels. Quantitatively, the proposed KSVM model yielded peak classification accuracies of 96.97% in Salalah, 94.33% in Al Khoud, and 93.37% in Bediya, proving its robust regional adaptation. The main contributions of this work include the development of a localized multi-class AQI predictive framework across diverse geographies and the identification of regional model limitations under severe class imbalance (such as in Al Khuwair where accuracy dropped to 67.11%). Model performance varied by region, with areas showing clear pollutant separation yielding more accurate predictions, while regions with overlapping pollution levels presented greater challenges. Limitations included missing and potentially noisy environmental data, as well as restricted temporal coverage, which constrained long-term analysis. To improve predictive reliability, future work should expand datasets across multiple years, incorporate additional environmental variables, and explore more advanced machine-learning techniques. Developing real-time monitoring systems and region-specific models is recommended to support effective air-quality management and policy decision-making. Overall, integrating AI with environmental data offers a practical approach for forecasting and mitigating air pollution impacts.

References

[1] World Health Organization (WHO), "*Air Pollution Guidelines and Global Updates*," 2022.
[2] United Nations Environment Programme (UNEP), "*Actions on Air Quality and Environmental Health*," 2021.
[3] A. Ghorani-Azam, B. Riahi-Zanjani, and M. Balali-Mood, "*Effects of Air Pollution on Human Health and Practical Measures for Prevention in Iran*," Journal of Research in Medical Sciences, Vol. 21, p. 65, 2016.
[4] Y. Charabi, S. A. Abdul-Wahab, B. S. Choudri, G. Al-Rawas, M. Al-Wardy, and S. Fadlallah, "*Impact of Summer Monsoon on Urban Traffic Air Pollution Dispersion in a Growing Tourism Destination: Case of Salalah, Oman*," Journal of Environmental Management and Tourism, Vol. 8, No. 5, pp. 1041–1060, 2017.
[5] J. J. N. Albusaidi, "*Integrating Air Quality and Climate Change Strategies in Oman*," Ph.D. Dissertation, Dept. Environ., Univ. of York, 2019.
[6] Y. Zheng, F. Liu, and H.-P. Hsieh, "*U-Air: When Urban Air Quality Inference Meets Big Data*," in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 1436–1445.
[7] X. Li, L. Peng, Hu, Y. Shao, and T. Chi, "*Deep Learning Architecture for Air Quality Predictions*," Environmental Science and Pollution Research, Vol. 23, No. 22, pp. 22408-22417, 2020.
[8] E. Gladkova and L. Saychenko, "*Applying Machine Learning Techniques in Air Quality Prediction*," Transportation Research Procedia, Vol. 63, pp. 1999–2006, 2022.
[9] S. N. Al Rumhi and J. Sivakumar, "*Advance Air Pollution Prediction in Oman through Machine Learning Approaches: Analysis, Models, and Future Directions*," Journal of Data Acquisition and Processing, Vol. 39, No. 2, pp. 92–115, 2024.
[10] A. Samad, S. Garuda, U. Vogt, and B. Yang, "*Air Pollution Prediction using Machine Learning Techniques: An Approach to Replace Existing Monitoring Stations with Virtual Monitoring Stations*," Atmospheric Environment, Vol. 310, p. 119987, 2023.
[11] T. Thongrod, A. Lim, T. Ingviya, and B. A. Owusu, "*Prediction of PM2.5 and PM10 in Chiang Mai Province: A Comparison of Machine Learning Models*," in Proc. ITC-CSCC, 2022, pp. 337–340.
[12] A. Singh, J. Sunil, M. Manjubala, A. Jadhav, and A. Goyal, "*Statistical Evaluation of the Performance of SVM Kernels for Air Quality Classification: A Case Study on India*," International Journal of Environment and Climate Change, Vol. 14, No. 11, pp. 584–594, 2024.
[13] A. Al-Muqbali, H. Al-Badi, and M. Al-Azawei, "*Localized Atmospheric Modeling and Machine Learning Applications in Hyper-Arid Microclimates*," Gulf Environmental Science Journal, Vol. 18, No. 1, pp. 45–59, 2025.

ISSN:2302-8149
e-ISSN:2540-9719

[14] D. Pant, R. Kumar, and S. Varma, "*Evaluation Metrics and Validation Protocols for Environmental Machine Learning Framework Benchmarks*," IEEE Letters on Environmental Modeling, Vol. 12, No. 3, pp. 112–116, 2023.