

Integrating GIS and machine learning for pollution vulnerability mapping in South Africa's Kruger National Park

Mohlehli George Mohlehli

mohlemg@unisa.ac.za

University of South Africa, UNISA 0003

Kgabo Humphrey Thamaga

University of South Africa, UNISA 0003

Godwell Nhamo

University of South Africa, UNISA 0003

Research Article

Keywords: Pollution vulnerability, SVM, RF, GBM, remote sensing; spatial mapping; air quality

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11124282/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Integrating GIS and machine learning for pollution vulnerability mapping in South Africa's Kruger National Park

Mohlehli George Mohlehli^{1*}, Kgabo Humphrey Thamaga², and Godwell Nhamo¹

Institute for Sustainability and Corporate Citizenship (ISCC), University of South Africa, P.O. Box 392, UNISA 0003, Pretoria, South Africa¹.

Department of Geography, University of South Africa, Florida, South Africa, University of South Africa, P.O. Box 392, UNISA 0003, Florida, South Africa².

*Corresponding author: mohlemg@unisa.ac.za/mohlehlmohlehli@gmail.com

Abstract

Pollution in protected areas can affect ecosystem integrity, biodiversity conservation, and the long-term sustainability of natural resource management. This study developed an integrated GIS and machine learning framework to assess and map pollution vulnerability in South Africa's Kruger National Park (KNP). Environmental and atmospheric variables, including the Normalised Difference Vegetation Index (NDVI), nitrogen dioxide (NO₂), sulphur dioxide (SO₂), carbon monoxide (CO), methane (CH₄), particulate matter (PM_{2.5}), rainfall, and temperature, were integrated within Google Earth Engine and Python-based analytical environments. Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM) algorithms were applied to classify spatial patterns of pollution vulnerability and identify the variables contributing most strongly to the observed patterns. The results showed that GBM achieved an overall classification accuracy of 0.76, while RF produced an area under the curve (AUC) of 0.96 and SVM achieved an AUC of 0.90. SO₂ and NO₂ were among the most influential variables in the modelling framework, indicating the importance of atmospheric pollution indicators in explaining spatial variations in vulnerability. The resulting vulnerability maps identified areas with moderate to high pollution vulnerability that can support targeted environmental monitoring and conservation planning. By integrating remote sensing, GIS, and machine learning, the study provides a spatial decision-support framework for sustainable environmental management in a large, protected area. The findings contribute particularly to Sustainable Development Goal (SDG) 15 (Life on Land) through improved protection and management of terrestrial ecosystems, while also supporting SDG 13 (Climate Action) through climate-sensitive environmental planning and contributing to SDG 3 (Good Health and Well-being) through improved management of environmental pollution.

Keywords: Pollution vulnerability, SVM, RF, GBM, remote sensing; spatial mapping; air quality

1. Introduction

Air pollution has emerged as one of the most vital universal environmental challenges, posing a silent public health threat and a rapid need that threatens human health, wildlife survival, and ecosystem stability (Kamaludeen et al., 2025; Ravindiran et al., 2023; Yadav et al., 2026). While traditionally linked with rapid industrialisation and urban centres, the presence of hazardous particles such as particulate matter (PM_{2.5}), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂) can cause lasting damage even at low levels in sensitive ecosystems. Protected areas such as Kruger National Park (KNP) are not immune to these hazards, as pollutants can be transported over long distances from distant industrial or agricultural sources, affecting both flora and fauna species (Aram et al., 2024, 2024; Kamaludeen et al., 2025; Morapedi & Obagbuwa, 2023). To ensure strong protection for these sensitive regions, regular monitoring and the development of accurate vulnerability maps can identify at-risk areas and guide sustainable strategies.

Historically, air quality forecasting has relied heavily on deterministic and statistical approaches, which comprise linear regression and ARIMA methods. However, these traditional approaches have inherent limitations, especially when handling complex, non-linear relationships and high-dimensional variance inherent in atmospheric data (Liu et al., 2021; Morapedi & Obagbuwa, 2023; Ravindiran et al., 2023; Yadav et al., 2026). As a result, machine learning (ML) has gained significant popularity due to its strength in identifying patterns in large datasets and handling incomplete or inconsistent information. In recent years, researchers have successfully applied various ML models to predict the Air Quality Index (AQI) and map pollution hotspots across the globe (Kamaludeen et al., 2025; Liu et al., 2021; Morapedi & Obagbuwa, 2023). Studies in India and China have extensively used models like Random Forest (RF) and Support Vector Machines (SVM) to forecast pollutant concentrations in densely populated cities. In South Africa, recent research has begun utilising ML to predict PM_{2.5} levels in urban centres such as Pretoria, Johannesburg, and Durban, highlighting the growing necessity for localised air quality models (Aram et al., 2024; Ketu, 2022; Morapedi & Obagbuwa, 2023; Tırink, 2025). However, most existing studies focus on urban and/or industrial environments, leaving a significant research gap in the application of these technologies to large-scale protected wilderness areas.

Pollution management in protected areas forms an important component of sustainable development because environmental degradation can undermine ecosystem functioning, biodiversity conservation, and the long-term capacity of protected areas to provide ecological services (McNeely, 1994; Mcneely, 2020). The integration of spatial environmental information with machine learning provides an opportunity to identify areas that may require targeted monitoring and management interventions, thereby strengthening evidence-based environmental governance (Ncube & Ngulube, 2024; Partha Protim Roy et al., 2024). Within the United Nations 2030 Agenda for Sustainable Development, the relevance of this study is most directly associated with SDG 15 (Life on Land), which promotes the protection, restoration, and sustainable use of terrestrial ecosystems and the conservation of biodiversity (Huan & Zhu, 2023; Pokharel, 2026). By identifying spatial patterns of pollution vulnerability within KNP, the study offers information that can contribute to ecosystem protection and sustainable protected-area management.

The study also relates to SDG 13 (Climate Action) because it incorporates atmospheric and climatic variables into the assessment of environmental vulnerability (Gaddam et al., 2026; Hamer et al., 2025). Although the study does not directly quantify climate-change impacts, the resulting spatial information can support climate-sensitive planning and environmental monitoring. A further indirect connection exists with SDG 3 (Good Health and Well-being), particularly Target 3.9, which addresses the reduction of illness and mortality associated with hazardous chemicals and air, water, and soil pollution (Martins et al., 2024; Ogwu & Izah, 2024). However, this study does not measure human exposure or health outcomes. Therefore, its contribution to SDG 3 is limited to providing environmental information that supports pollution management. Thus, these linkages position pollution vulnerability assessment within a broader sustainable-development framework in which spatial environmental intelligence can support ecosystem conservation, climate resilience, and responsible environmental management (Al-Raei, 2024; Hossain, 2026).

The aim of this study was to develop an integrated GIS and machine learning framework for mapping pollution vulnerability in South Africa's Kruger National Park and to identify the environmental and atmospheric variables associated with spatial variations in vulnerability. The study further seeks to demonstrate how spatially explicit pollution information can support sustainable environmental management, protected-area conservation, and progress towards the Sustainable Development Goals, particularly SDG 15, SDG 13, and, indirectly, SDG 3. The work further seeks to implement and optimise RF, SVM, and Gradient Boosting Machine (GBM) algorithms to model complex non-linear relationships between environmental covariates and pollution vulnerability. The study also evaluates and compares the predictive performance of these models using statistical metrics such as the Receiver Operating Characteristic (ROC) for model performance testing, generates high-resolution GIS-based pollution vulnerability maps identifying spatial hotspots across the park, and identifies and quantifies the dominant environmental drivers influencing pollution susceptibility using feature importance analysis. Our study is guided by the hypothesis that joint ML approaches, particularly RF and GBM, will outperform kernel-based approaches such as SVM in predicting pollution vulnerability due to their superior ability to capture complex interactions among atmospheric, climatic, and vegetation-related variables. In addition, it is hypothesised that atmospheric pollutants, especially SO₂ and NO₂, together with climatic variability and vegetation dynamics, significantly influence the spatial distribution of pollution hotspots within KNP.

The study's strength lies in its integration of remote sensing, atmospheric monitoring, GIS, and comparative ML modelling within a protected savanna ecosystem that remains underrepresented in pollution vulnerability research. Relative to prior research that has concentrated on urban air quality or isolated pollutant monitoring using ML algorithms solely, this study provides the first integrated application of three distinct ML algorithms highlighted earlier as specifically tailored for pollution vulnerability mapping in a large African savanna protected area. The research uniquely combines multi-source satellite-derived variables (NDVI, NO₂, SO₂, CO, CH₄, PM_{2.5}, rainfall, and temperature) within a single spatiotemporal framework. It translates model outputs into high-resolution, management-ready vulnerability maps for KNP. This integrated approach not only advances methodological boundaries in conservation-oriented air quality assessment but also delivers actionable spatial intelligence for park managers. The study therefore advances both methodological and applied environmental

research by producing interpretable pollution susceptibility maps that can support ecosystem management, wildfire preparedness, biodiversity conservation, and long-term environmental planning within KNP.

2. Materials and Methods

2.1 Study Area

The KNP covers approximately 19,485 km² (van Wilgen et al., 2022), stretching around 360 km from north to south and 65 km from east to west (Venter et al, 2008). It is located in northeastern South Africa, primarily within the provinces of Limpopo and Mpumalanga. The park forms a major part of the Great Limpopo Transfrontier Park, bordering Mozambique to the east and Zimbabwe to the north. It is one of Africa's largest game reserves and a significant conservation area. The terrain is largely flat, with scattered low-lying hills, and is intersected by major river systems, including the Sabie, Olifants, Letaba, and Crocodile Rivers, which provide critical hydrological support to the park's ecological systems (Christie, 2024; Smit, 2013). The park's extensive and diverse ecosystems, encompassing savannas, grasslands, and riparian environments, make it an ideal setting for environmental monitoring (van Wilgen et al., 2022). The park's vegetation comprises mopane forests in both granite and basaltic regions farther north, open wooded grasslands on basalt in the southeast, and woodlands on granite in the southwest, with 147 mammal species, and the KNP is widely known for its abundant biodiversity (van Wilgen et al., 2022). These include grazers like zebras and wildebeest, browsers like kudus, and huge herbivores like elephants, hippopotami, buffaloes, and rhinos (Coetsee et al., 2012). The KNP is unique among savannas in Southern Africa in that it still has a full range of large herbivores and carnivores. In KNP, the average yearly rainfall falls between 350 and 750 mm. Seasonality and rainfall patterns are important ecological variables that impact vegetation phenology and the buildup of fire fuel (van Wilgen et al., 2022; Venter et al, 2008). The park's southern region experiences more rainfall than its northern one. Winters in the KNP are dry. Droughts are a significant problem as they affect vegetation and call for flexible fire management techniques (van Wilgen et al., 2022). Figure 1 presents the study location.

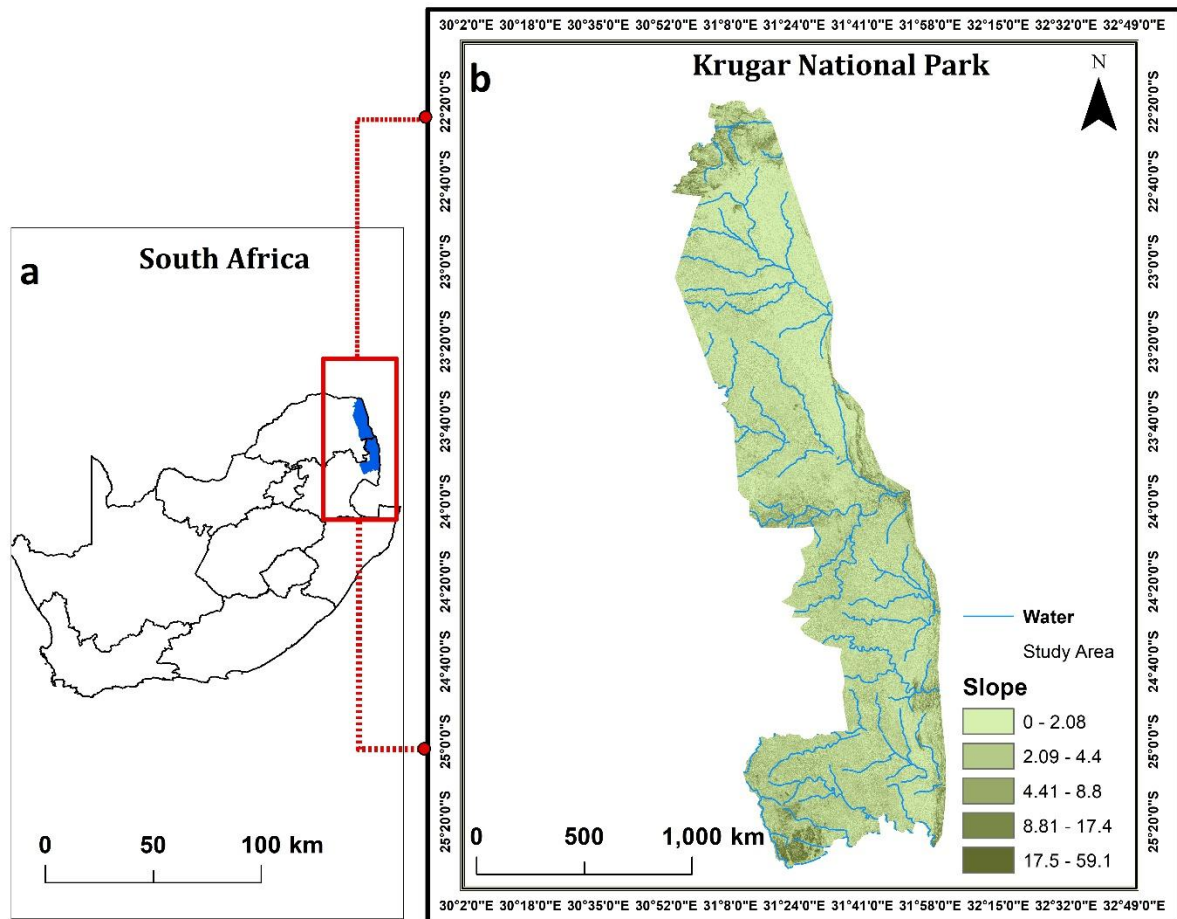


Figure 1. The spatial extent and geographical context of KNP

2.2 Research design

All geospatial data processing and initial data extraction were performed using the Google Earth Engine (GEE) cloud-computing platform, followed by statistical analysis and machine learning model implementation in Python (Figure 2).

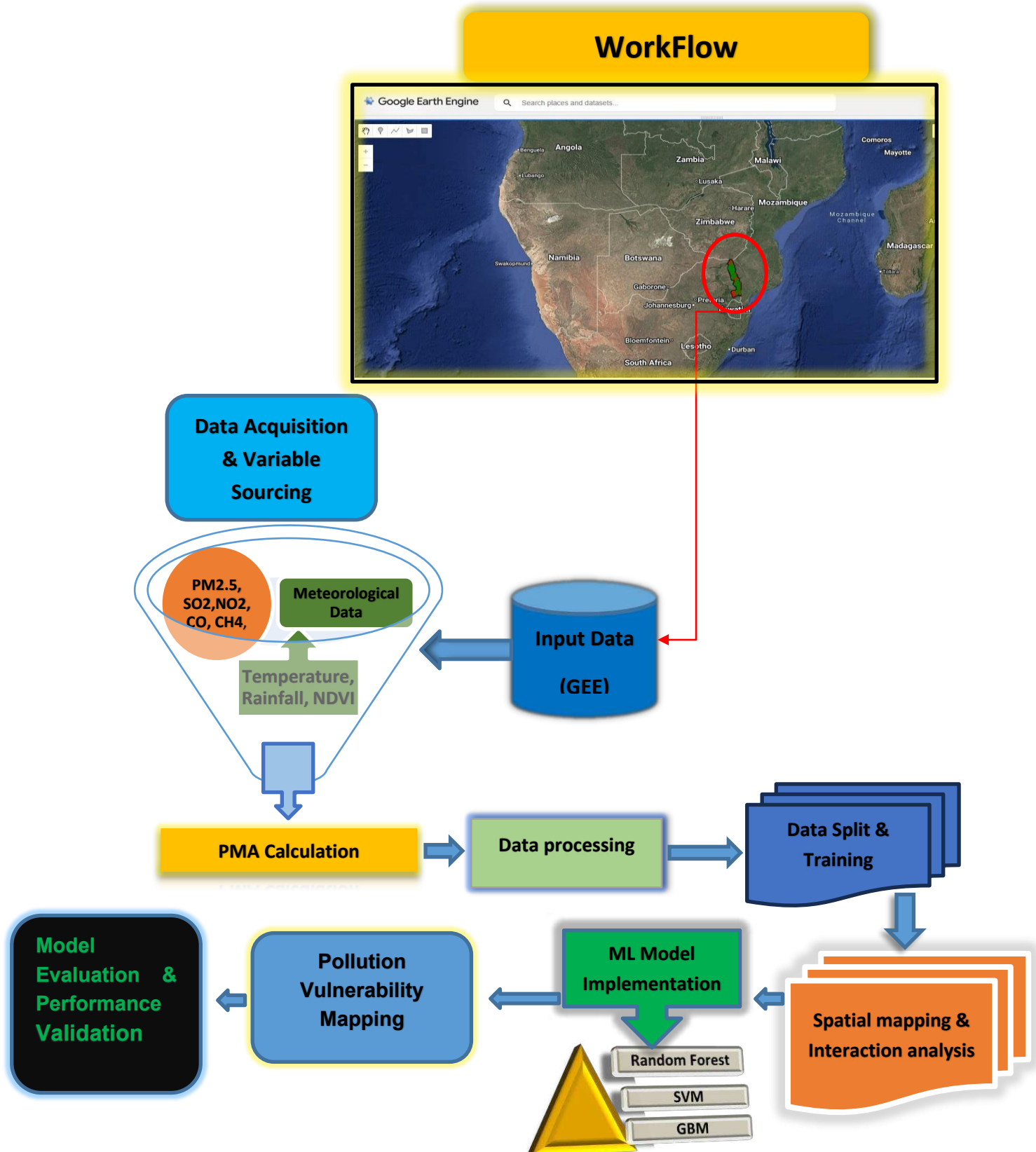


Figure 2. Position of the study area in the GEE image and flowchart of the methodology.

2.3 Data Acquisition and Pre-processing

All raw satellite imagery and atmospheric product collections were accessed and processed directly within the Google Earth Engine environment. For each image collection, data were filtered to the geographical bounds of the KNP (.filterBounds(AOI)) and the specified date range (.filterDate(start, end)). Mean composites were then generated across the entire study period (.mean()) and clipped to the KNP boundary (.clip(AOI)) to ensure that all analyses focused solely on the park area. Satellite data advancements not only enable the detection of air pollutants but also offer the advantage of covering large geographical areas over extended periods of time (Songsom et al., 2025).

2.4 Air Quality Data and Meteorological Data

The identification of pollution hotspots in the study area relied on an integrated analysis of atmospheric, environmental, and meteorological variables that characterise both pollutant behaviour and ecosystem sensitivity (Table 1). PM_{2.5} is a primary variable of concern because its tiny aerodynamic diameter allows it to remain in the atmosphere for long periods and penetrate deep into the respiratory systems of both humans and wildlife (Liu et al., 2021). Gaseous pollutants such as NO₂, SO₂, and CO serve as critical indicators of anthropogenic activity, such as vehicular traffic, industrial processes, and biomass burning and their high concentrations are linked to severe ecological damage, like acid rain and significant health risks (Kamaludeen et al., 2025; Ketu, 2022; Özüpak et al., 2025; Ravindiran et al., 2023). CH₄ (methane) is likewise monitored as a primary pollutant directly emitted from industrial or agricultural sources, contributing to overall atmospheric contamination and global warming (Kamaludeen et al., 2025).

We further utilised NDVI as a vital environmental proxy, where it indicates vegetation density, which can mitigate pollution levels through absorption or by providing shade that reduces land surface temperatures (Yadav et al., 2026). Meteorological covariates are essential for modelling the transport and settlement of these pollutants: rainfall acts as a natural controller that reduces pollutant intensity through wet deposition (the wash-out effect), while temperature drives the photochemical reactions necessary for ozone formation and influences whether pollutants are dispersed or trapped near the surface by seasonal inversion effects (Ravindiran et al., 2023). Together, these variables allow for the identification of specific geographic areas where high emission levels converge with unfavourable meteorological conditions and low environmental resilience, effectively pinpointing pollution hotspots (Kamaludeen et al., 2025).

The use of satellite-derived data is a critical methodological choice, as researchers like Kamaludeen et al. (2025) highlight that large-scale protected areas in the African context often suffer from a scarcity of reliable ground-based air quality sensors. By leveraging the Google Earth Engine (GEE) platform, the study harmonises disparate data types ranging from high-resolution 10-meter vegetation indices to 5.5-kilometre rainfall data into a unified spatiotemporal database. The atmospheric component relies heavily on the Copernicus Sentinel-5P mission, which Songsom et al. (2025) describes it as a powerful tool for monitoring trace gases, such as, and at a spatial resolution of 1113.2 meters. Particulate matter is estimated using Aerosol Optical Depth (AOD) from MODIS, a standard scientific proxy utilised because AOD represents the columnar density of aerosols that directly correlates with ground-level

particulate concentrations. To account for the environmental and meteorological drivers of pollution, the study incorporates NDVI from Sentinel-2 as an indicator of vegetation density, which can act as a bioindicator for pollution absorption. Furthermore, meteorological covariates such as rainfall (CHIRPS) and temperature (Landsat 8) are included to model the "wash-out" effect and heat-driven dispersion dynamics that significantly influence pollutants' settling or transport across the park's diverse landscape.

Table 1. Summary of the processed variables, their providers, and spatial resolutions

Variable	Dataset ID (Earth Engine)	Source	Spatial Resolution
NDVI	COPERNICUS/S2_SR_HARMONIZED	Sentinel-2	10 m (bands B4, B8)
NO₂	COPERNICUS/S5P/NRTI/L3_NO2	ESA Sentinel-5P	1113.2 m
SO₂	COPERNICUS/S5P/NRTI/L3_SO2	ESA Sentinel-5P (TROPOMI)	1113.2m
CO	COPERNICUS/S5P/NRTI/L3_CO	ESA Sentinel-5P (TROPOMI)	1113.2m
Rainfall	UCSB-CHG/CHIRPS/DAILY	UCSB Climate Hazards Group	5,566 m
Temperature	ECMWF/ERA5_LAND/HOURLY	Landsat 8	30 m
PM2.5	Aerosol Optical Depth (AOD) data from MODIS (Terra and Aqua satellites)	Landsat 8	1113.2m
CH4	COPERNICUS/S5P/NRTI/L3	Sentinel-5P	1113.2m

2.5 Machine Learning Tools

This study utilised three supervised learning classifiers to capture non-linear pollution environment relationships, namely RF, SVM, and GBM. Each of these ML tools will now be briefly considered.

2.5.1 Random Forest (RF)

RF is described as a collective method that builds multiple independent decision trees through bagging and random feature selection. It is highly resistant to overfitting and provides a clear ranking of feature importance (Salman et al., 2024; Yadav et al., 2026). In this study of Kruger National Park, RF serves as a strong ensemble supervised learning algorithm that builds multiple independent decision trees through bootstrapping and aggregation (bagging). Originally proposed by Goyal & Goyal (2024) and AlThuwaynee et al. (2021), this nonparametric method is particularly valuable for handling the complex, non-linear relationships and high-dimensional variance inherent in atmospheric and environmental

datasets. A primary strength of RF in mapping pollution vulnerability is its high resistance to overfitting by averaging the results of an entire forest of trees, the model reduces overall variance and ensures that predictions generalise accurately across the park's vast and diverse landscape (Wahba, 2024). This allows you to rank which variables, such as NDVI, rainfall, or temperature, are the most influential drivers of pollution hotspots in the park. Additionally, RF is renowned for its strength against noise and its ability to handle missing values automatically, which is essential when integrating multi-source satellite data. Recent literature consistently identifies RF as a top-performing model in air quality index (AQI) forecasting, frequently demonstrating superior predictive accuracy and low error rates compared to other machine learning techniques (Sharma et al., 2024).

2.5.2 Support Vector Machine (SVM)

This supervised learning method uses the kernel trick (e.g., Radial Basis Function) to map data into a high-dimensional space, finding the optimal hyperplane that classifies areas into different vulnerability levels (Sharma et al., 2024). In the study of Kruger National Park, SVM, including its application as Support Vector Regression (SVR), is a crucial methodological component due to its exceptional performance in environmental classification and regression tasks. A primary strength of SVM is its ability to handle complex, non-linear relationships and high-dimensional atmospheric datasets, which often present challenges for traditional linear models. This capability is powered by the kernel trick, specifically the Radial Basis Function (RBF) kernel, which transforms input variables into a higher-dimensional space to find an optimal separating hyperplane. Unlike models that only minimise empirical error, SVM aims to minimise structural risk, which significantly enhances its generalisation ability and prevents the model from overfitting to specific historical data (Sudjianto & Zhang, 2024). By identifying support vectors, the critical data points closest to the decision boundary, SVM achieves a strong classification of pollution-vulnerable areas, making it highly effective for identifying hotspots and generating the precise spatial maps needed to safeguard the park's fragile ecosystems.

2.5.3 Gradient Boosting Machine (GBM)

The Gradient Boosting Machine (GBM) approach is an ensemble learning approach that integrates multiple weak algorithms to generate a strong and precise model. It operates progressively, bringing new models into the collection gradually, with each new model focused on rectifying the mistakes of the prior one (Nadkarni et al., 2023; Songsom et al., 2025). While classical boosting algorithms offered theoretical precision and practical strength for early sensor analytics, their scalability and strength persist in being limited for modern high-volume sensing systems. Recent gradient boosting decision tree (GBDT) frameworks, for example, XGBoost, LightGBM, and CatBoost, extend Friedman's foundational GBM with substantial algorithmic, computational, and regularisation innovations (Nadkarni et al., 2023; Ramdani & Furqon, 2022). These frameworks now dominate predictive modelling in healthcare, environmental, energy, and industrial sensing, including semiconductor VM pipelines that increasingly integrate boosting with sparse feature selection methods (Xie & Tummala, 2025). This section reviews their theoretical foundations, system-level improvements, and representative sensor applications, showing how they address the challenges identified in Section 2 and overcome the limitations of classical boosting in Section 3. Unlike RF, these models build trees sequentially, where each new tree corrects the errors of the previous one.

This boosting approach is often superior in predictive accuracy and handling large-scale datasets.

2.6 Model Training, Validation, and Assessment

2.6.1 Data Splitting

The dataset was split into a training set that comprised 70-80%, and a testing set that constituted 20-30% of the data. This was done to assess how well the models generalise to unseen data (Gupta et al., 2023; Ravindiran et al., 2023; Tırink, 2025). Hyperparameter Optimisation utilise Grid Search or Bayesian Optimisation to fine-tune model parameters (e.g., number of trees in RF or the learning rate in GBM) to achieve the highest possible accuracy (Malakouti et al., 2024).

2.6.2 Spatial Vulnerability Mapping and Analysis

The final step is to translate the model outputs into actionable management tools. Vulnerability mapping, where we combine the model predictions to create a Susceptibility/Vulnerability Index, classifies regions of Kruger National Park such as “Very low”, “Low”, “Moderate”, “High”, and “Very high” vulnerability. GIS visualisation, use GIS software (ArcGIS) to produce high-resolution maps that visualise these hotspots, allowing park managers to identify specific zones where wildlife or ecosystems are most at risk from atmospheric pollution.

2.6.3 Validation of Model Performance

For validating machine learning models in environmental studies, heat maps and ROC curves provide essential insights into data relationships and classification accuracy (Fazil et al., 2023). Heat maps are primarily utilised for exploratory data analysis to visualise the correlation matrix between variables, where they use colour-coded intensities to represent the magnitude (ranging from +1 to -1) and direction of relationships between pollutants and environmental drivers (Ravindiran et al., 2023). The Receiver Operating Characteristic (ROC) curve serves as a graphical representation of a model's validity by plotting its sensitivity (True Positive Rate) on the Y-axis against its specificity (False Positive Rate) on the X-axis across various thresholds. To gain a deeper understanding of the spatial characteristics of the pollution classes as interpreted by each machine learning model, KDE plots were generated. The figure visualises the spatial density of the predicted classes for each of the three models (RF, GBM, and SVM).

2.6.4 Literature-based reference inventory and validation

An independent literature-based reference inventory was established to assess the spatial plausibility of the SVM-derived pollution/degradation patterns. Reference locations were selected from documented water-quality and ecological monitoring sites across the Crocodile, Sabie, Olifants and Luvuvhu river systems of the Kruger National Park (KNP). Riddell et al. (2019) provide geographically referenced monitoring sites and demonstrate that these river systems are subject to varying water-quality pressures associated with upstream anthropogenic activities. The selected locations were not used for SVM model training. Instead, the SVM class at each reference location was extracted and compared with the documented environmental and pollution conditions reported in the literature. This approach provided an independent spatial reference for assessing model credibility, particularly because previous monitoring has identified river-specific pollution and water-quality impacts, including agricultural, mining and other anthropogenic influences. Importantly, these sites were treated

as literature-based reference/validation locations rather than direct ground-truth observations, since the published study does not provide the same five pollution/degradation classes used by the SVM model. This distinction ensures that the validation procedure does not overstate model accuracy while providing an independent ecological basis for evaluating the mapped spatial patterns.

3. Results

3.1 Integration of Multi-Source Environmental and Atmospheric Variables

Figure 3 shows the temporal patterns of the environmental variables we analysed. The figure presents a panel of eight spatial distribution maps, each corresponding to one of the environmental and atmospheric variables analysed in this study: (a) Nitrogen Dioxide (NO_2), (b) Sulfur Dioxide (SO_2), (c) Carbon Monoxide (CO), (d) Methane (CH_4), (e) Particulate Matter ($\text{PM}_{2.5}$), (f) Rainfall, (g) Temperature, and (h) Normalised Difference Vegetation Index (NDVI).

Each map displays the spatial variability of the respective variable across the approximately 19,485 km² extent of Kruger National Park, with colour gradients ranging from low (cool colours: blue to green) to high (warm colours: yellow to red) values. For the pollutant gases, NO_2 (a) and SO_2 (b) exhibit elevated concentrations in the southern and south-central regions of the park, particularly in proximity to the Crocodile and Sabie River catchments, with values progressively decreasing towards the northern extents. CO (c) demonstrates a more spatially uniform distribution across the entire park, showing only minor variation between eastern and western boundaries. CH_4 (d) displays a distinct spatial pattern, with higher concentrations observed in the central and northern basalt plains, while lower values characterise the southern granite-dominated landscapes. $\text{PM}_{2.5}$ (e) shows a patchy distribution, with isolated hotspots scattered throughout the park rather than a clear north–south gradient. Rainfall (f) exhibits a clear south-to-north decreasing gradient, with the southern regions receiving notably higher annual precipitation compared to the arid northern sections. Temperature (g) displays the inverse pattern, with higher temperatures dominating the northern and central low-altitude areas, while relatively cooler conditions prevail in the southern high-rainfall zone.

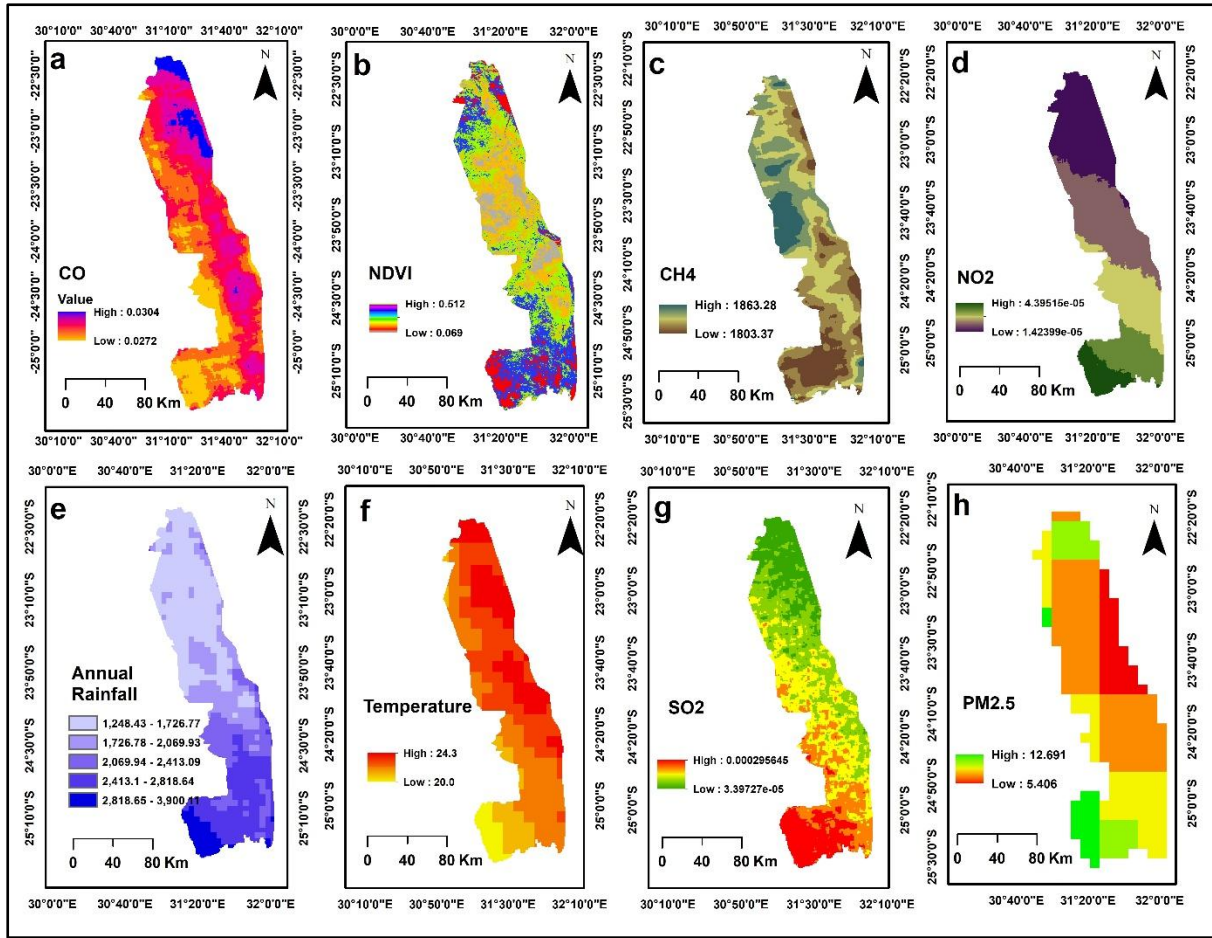


Figure 3. Spatial distribution of air quality variables

3.2 Pollution Variables Interactions

A heatmap was generated to understand the interaction of the variables applied in the pollution prediction models (Figure 4). The Correlation Matrix of Pollution Variables shows the linear relationships between all the pollution and environmental variables in your dataset. The values range from -1 to 1, where values close to 1 suggest a strong positive linear correlation (as one variable increases, the other tends to increase). Values closer to -1 suggest a strong negative linear correlation (as one variable increases, the other tends to decrease), and Values close to 0 indicate a weak or no linear correlation. Looking at the matrix, it can be observed that SO₂ and NO₂ show a very strong positive correlation of approximately 0.95. This suggests they often vary together, which is common as both are combustion byproducts. NDVI (Normalised Difference Vegetation Index) shows a strong positive correlation with NO₂ (0.67) and SO₂ (0.60), and a strong negative correlation with CH₄ (-0.67). This is interesting and might imply that areas with higher vegetation density (NDVI) tend to have higher NO₂ and SO₂ (perhaps due to industrial activity or urban areas with some green spaces) but lower CH₄. Moreover, temperature and rainfall exhibit a strong negative correlation (-0.74), which is expected as these are often inversely related across seasons or geographic zones, and CH₄ has a distinct negative correlation with NO₂ (-0.61) and SO₂ (-0.50).

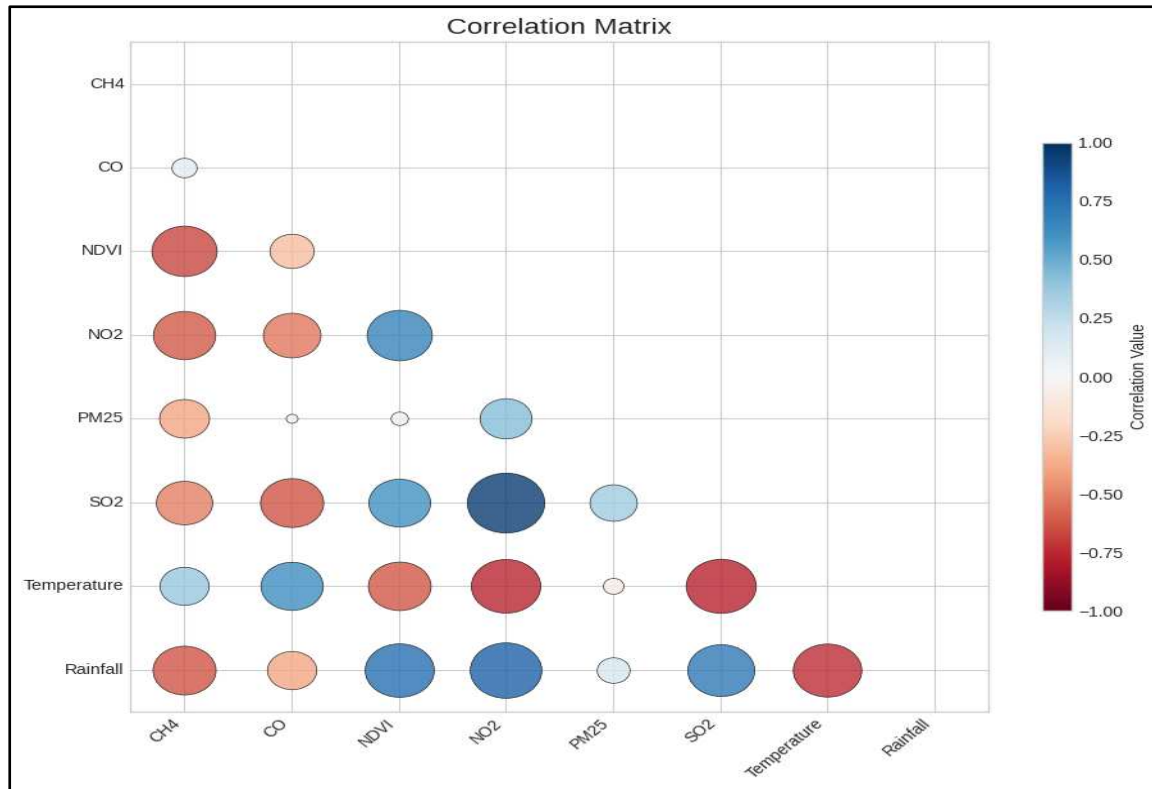


Figure 4. Heatmap correlations showing the relationships between investigated variables

3.3 Feature Importance for the Predictive Models

To delve into enhancing model interpretability, performance and efficiency, we performed a feature importance analysis for the three machine learning models using Python (Figure 5). The Random Forest model (a) identifies SO₂ as the most important feature, followed by NO₂, CH₄, Temperature, and Rainfall. CO, NDVI, and PM_{2.5} have comparatively lower importance. The accuracy of 0.74 suggests an acceptable output performance. The GBM model also ranks SO₂ as the most significant feature, with CH₄ and Rainfall following. Its feature importance distribution is somewhat similar to Random Forest, but with slightly different emphasis on the middle-ranked features. The accuracy performance of GBM suggests a good performance of 0.76, which is relatively high. For SVM, using permutation importance, SO₂ again emerges as the most important feature. CH₄ and NO₂ are also highly ranked. It's worth observing that SVM's accuracy is the lowest relative to RF and GBM, which can be reflected in its importance values.

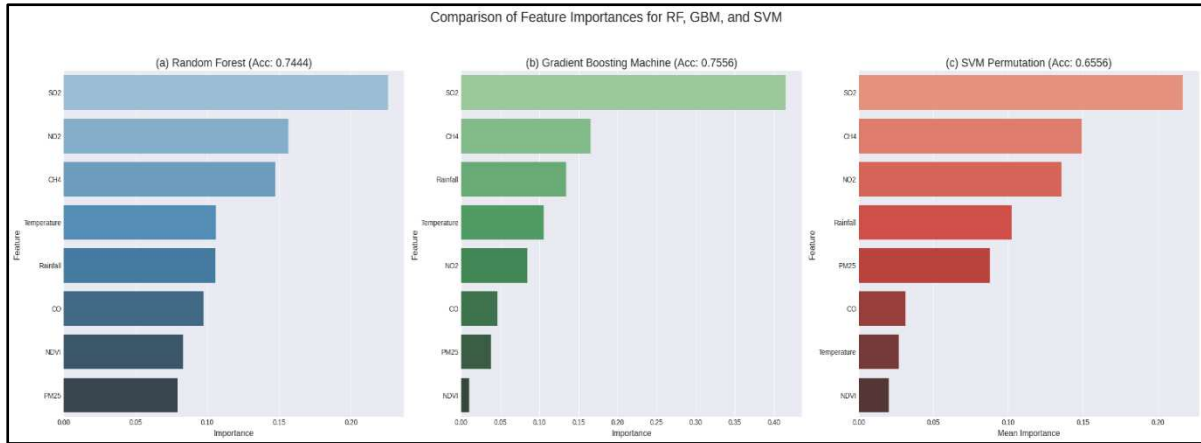


Figure 5. Importance of pollution assessment variable factors based on (a) RF, (b) GBM, and (c) SVM models.

To delve further into understanding the agreement of the model performance in mapping pollution, a prediction overlap between models was performed (Figure 6). The Prediction Overlap analysis, visualised in the heatmap and summarised by the agreement percentages, suggests how consistently the three machine learning models (RF, GBM, and SVM) classify the pollution samples. The predictive models showed varying levels of accuracy, with RF and GBM, the results suggest an agreement of approximately 83.3% of the predictions. This indicates a strong consistency between these two tree-based ensemble methods. The RF and SVM show an agreement of approximately 78.9% of the predictions. This is a satisfactory level of agreement between a tree-based model and a kernel-based model. GBM and SVM show an agreement on approximately 71.1% of the predictions. This is the lowest pairwise agreement, suggesting that the GBM and SVM models have slightly more divergent decision boundaries compared to the other pairs. For the three-Way Agreement, the 66.7% three-way agreement means that all three models (RF, GBM, and SVM) made the same prediction for two-thirds of the test samples. A higher three-way agreement suggests more confidence in the combined predictions, as multiple diverse models are arriving at the same conclusion.

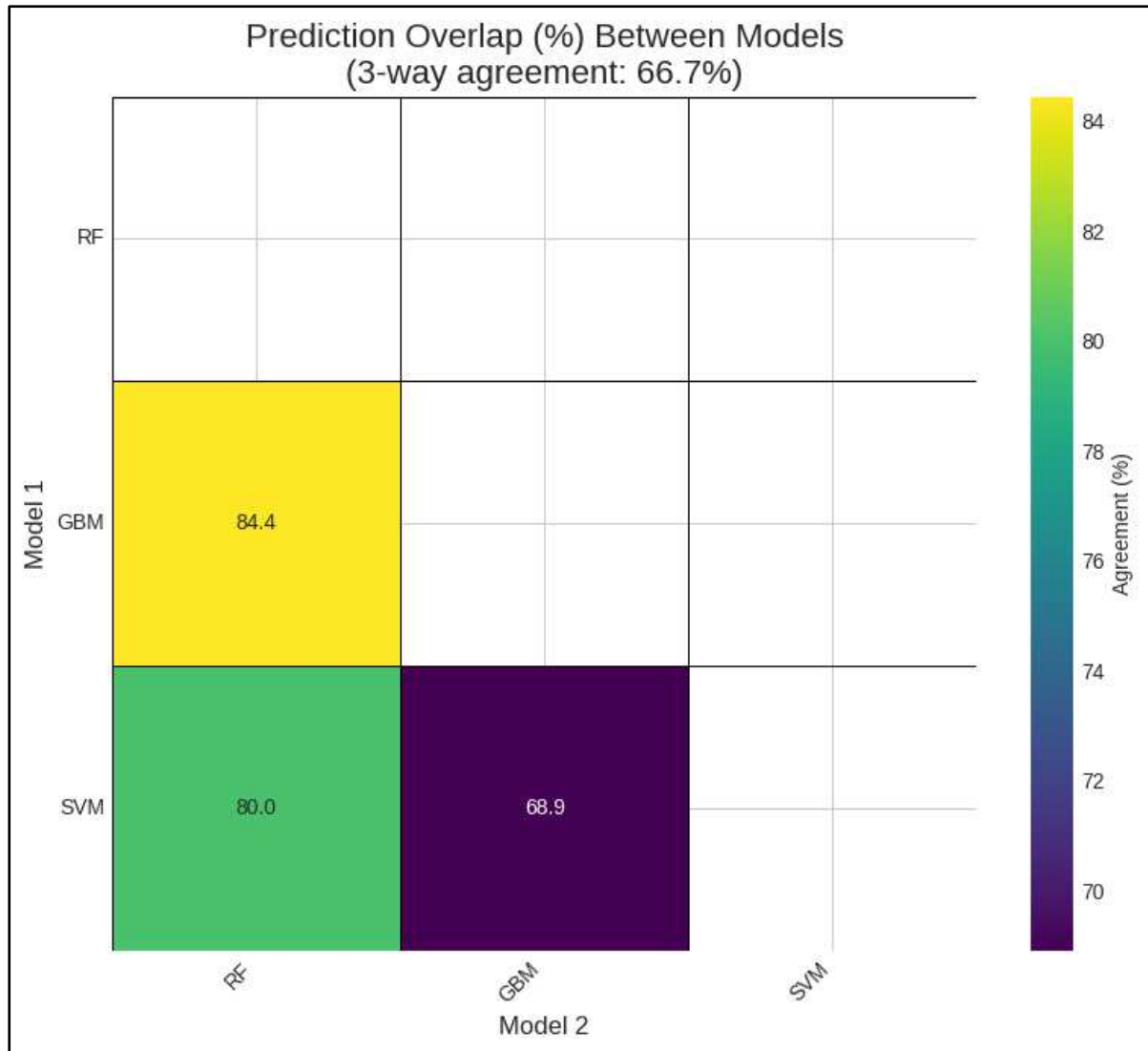


Figure 6. Agreement on the model performance consistency of the RF, GBM, and SVM.

To gain a deeper understanding of the spatial characteristics of the pollution classes as interpreted by each machine learning model, KDE plots were generated (Figure 7). This visualisation, presented as a figure with three subplots (a) Random Forest, (b) Gradient Boosting Machine, and (c) SVM, helps us understand where each model predicts different pollution classes across the geographical area. Each 3D KDE map shows X and Y axes, which are represented by longitude and latitude, indicating the geographical location of the samples. The Z-axis represents the density of predicted classes. Across all three models, the spatial surfaces exhibit highly similar patterns, indicating strong agreement among the classifiers regarding the geographic distribution of the predicted phenomenon. The density values range approximately from 0.0 to 0.65, as indicated by the colour bars and vertical elevation. Several prominent high-density zones are visible, for instance, a major peak occurs around longitude 31.7 to 31.9 and latitude -24.6 to -24.3. Secondary elevated regions appear near longitude 31.4 to 31.6 and latitude -24.8 to -24.4. Lower-density areas dominate the outer margins, especially toward the northern and western extents of the study region.

The Random Forest model produces a relatively smooth but detailed surface with multiple localised peaks. The model captures spatial heterogeneity effectively, showing nuanced

variations in density across the landscape. The GBM surface closely resembles the RF output but appears slightly smoother and more generalised. The high-density regions remain concentrated within the same spatial extents. However, transitions between peaks and surrounding areas are more gradual. This indicates that GBM successfully captures the dominant spatial structures while reducing some localised variability. The model may therefore provide stronger generalisation and reduced overfitting compared to RF. The SVM model also demonstrates strong agreement with RF and GBM. The spatial peaks occur in nearly identical geographic locations, although the surface appears slightly more balanced and regularised.

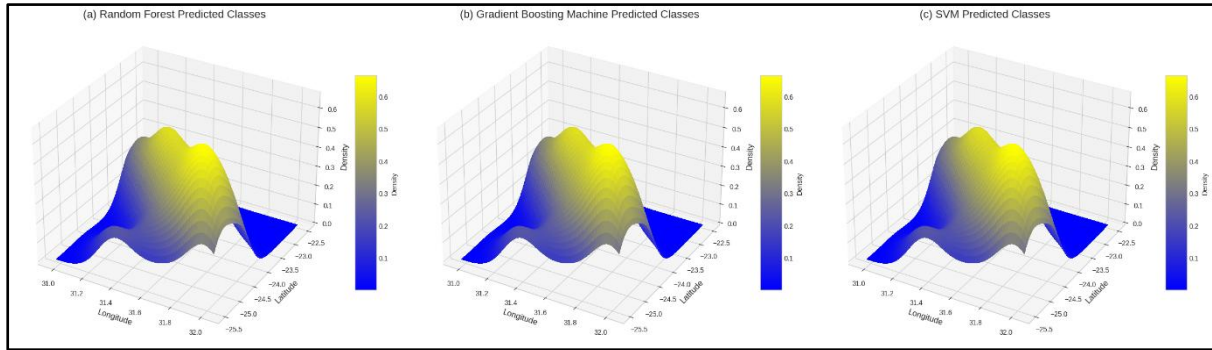


Figure 7. Spatial density analysis of the predicted classes for each of the three models: Random Forest, Gradient Boosting Machine, and SVM

3.4 Pollution vulnerability mapping (PVM)

Five classes (very low, low, moderate, high, and very high) were included in the pollution classification prediction models (Figure 8). For the random forest classification results (Figure 8a), the very low class covers 14.4% of the overall area, the low class covers 16.7% of the area, the moderate class covers 28.9%, the high class covers 17.8%, and the very high class covers 22.2% of the study area. For the SVM, the very low class covers 16.7% of the overall area, the low class covers 10.0%, the moderate class covers 28.9%, the high class covers 25.6%, and the very high class covers 18.9% of the study area. For the GBM model in Figure 8c, the very low class covers 16.7%, the low class covers 14.4%, the moderate class covers 27.8%, the high class covers 17.8%, and the very high class covers 23.3% of the total study area. These percentages suggest the proportion of data points (or estimated area) that each model classifies into a particular pollution level. For instance, both Random Forest and SVM predict that Class 2 (Moderate pollution) covers the largest percentage of the study area, followed by Class 4 (Very High pollution) for RF and GBM, and Class 3 (High pollution) for SVM. This offers a quantitative view of how each model delineates the spatial extent of different pollution categories.

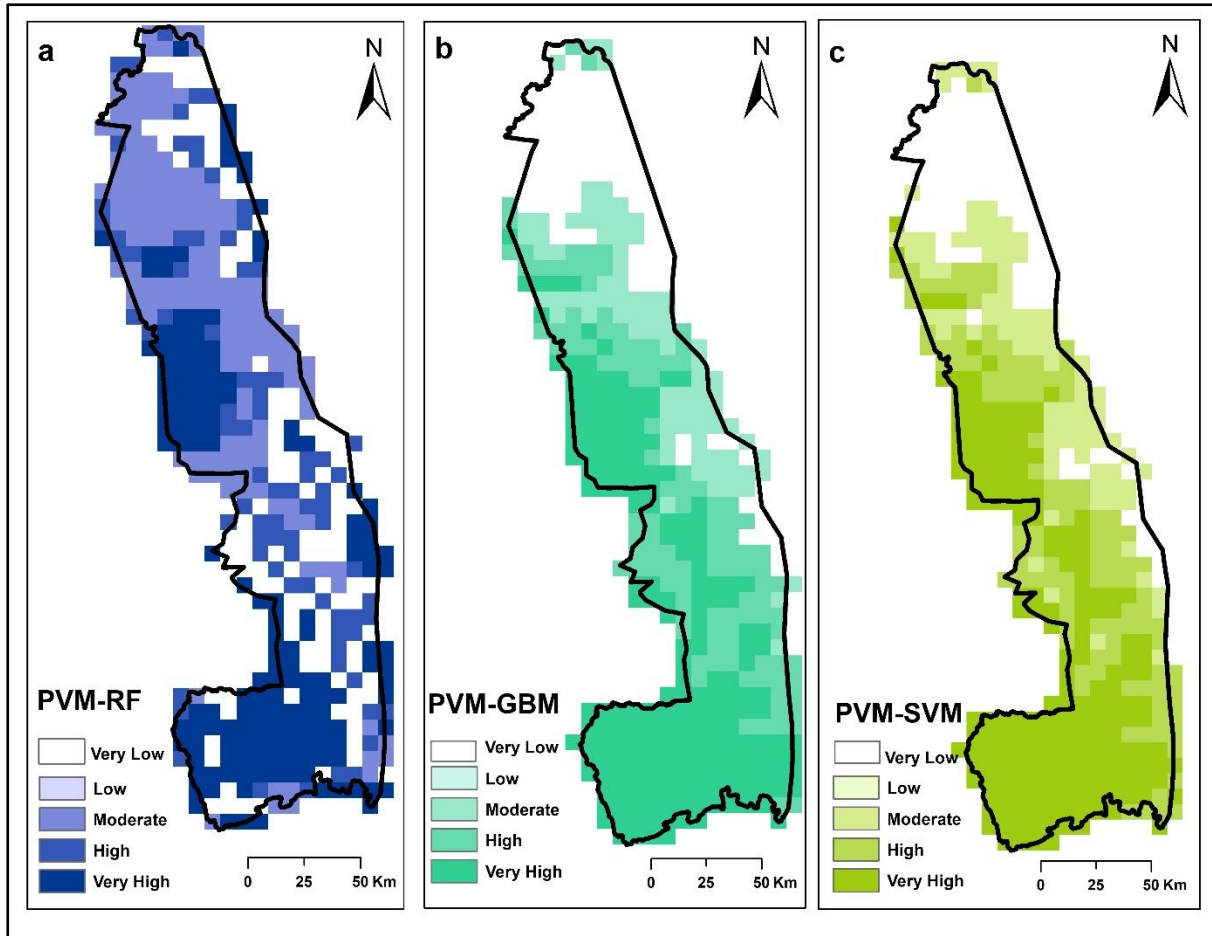


Figure 8. Pollution vulnerability maps of KNP using (a) RF, (b) SVM, and (c) GBM models

Receiver operating characteristic (ROC) curves were created to test the validity of the three model predictions (Figure 9). The ROC analysis suggests strong discriminatory performance across all three classifiers. The significant area under the curve (AUC) values and the prominent curvature toward the upper-left corner of the ROC space show that the ROC analysis exhibits strong discriminatory performance across all three classifiers. The RF model (Figure 9a) achieved an accuracy of 0.74 and the highest AUC of 0.96, suggesting satisfactory classification among classes and a high true positive rate across pseudo-positive thresholds. The GBM classifier yielded the highest overall classification accuracy of 0.76 but had a slightly lower AUC of 0.92 relative to the RF classifier. The SVM yielded the lowest classification accuracy of 0.66. This suggests that although SVM threshold-dependent accuracy is relatively lower, its overall ranking ability and classification capability remain acceptable.

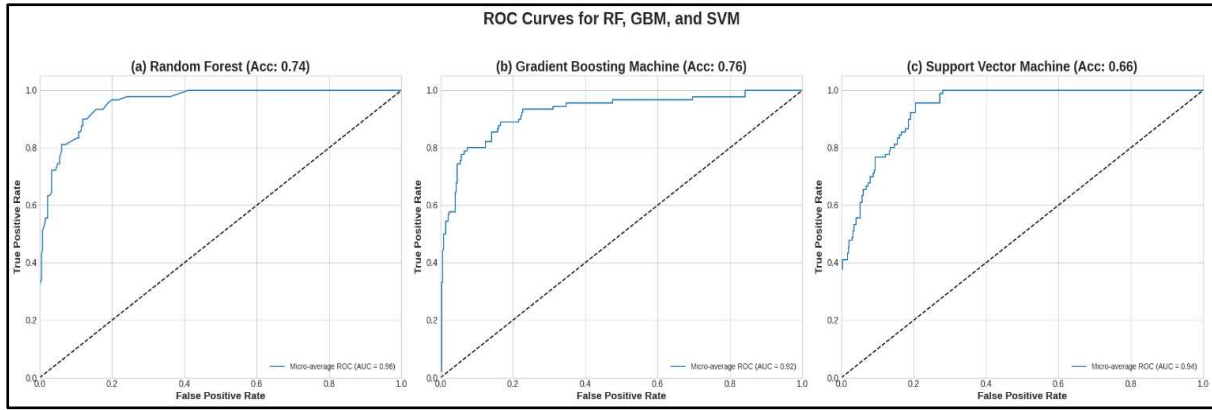


Figure 9. Validation of the pollution susceptibility of (a) RF, (b) GBM, and (c) SVM models

3.5 Model Validation Using Literature Point as Reference Inventory

To further ensure the validity of the model, an inventory was created to reference the known areas prone to pollution. The inventory comprised 12 river-associated reference locations, represented by the white circular points in Figure 10. The locations were distributed across four major river systems associated with the KNP landscape. The Crocodile River was represented by four reference locations (Karino, Montrose, Dalton [Kaap River tributary], and Malelane Gate). The Sabie River was represented by four locations (Sabie Town, Perry's Bridge/Hazyview, Phabeni Gate, and Lower Sabie). Two reference locations were associated with the Olifants River, namely Phalaborwa Barrage and Mamba Weir, while the Luvuvhu River was represented by Mutale River and Mhinga. Spatially, the reference points formed a north–south distribution around the KNP boundary and along the major river corridors. The southern portion of the inventory included the Crocodile and Sabie River locations, while the central and northern portions included the Olifants and Luvuvhu River-associated locations. Each reference location was subsequently used to extract the corresponding SVM pollution/degradation class from the classified raster, providing a spatial inventory for comparison with the modelled five-class pollution/degradation pattern. The results show a consistent outcome as predicted by the ground data. This suggests that the model performance has provided a good prediction.

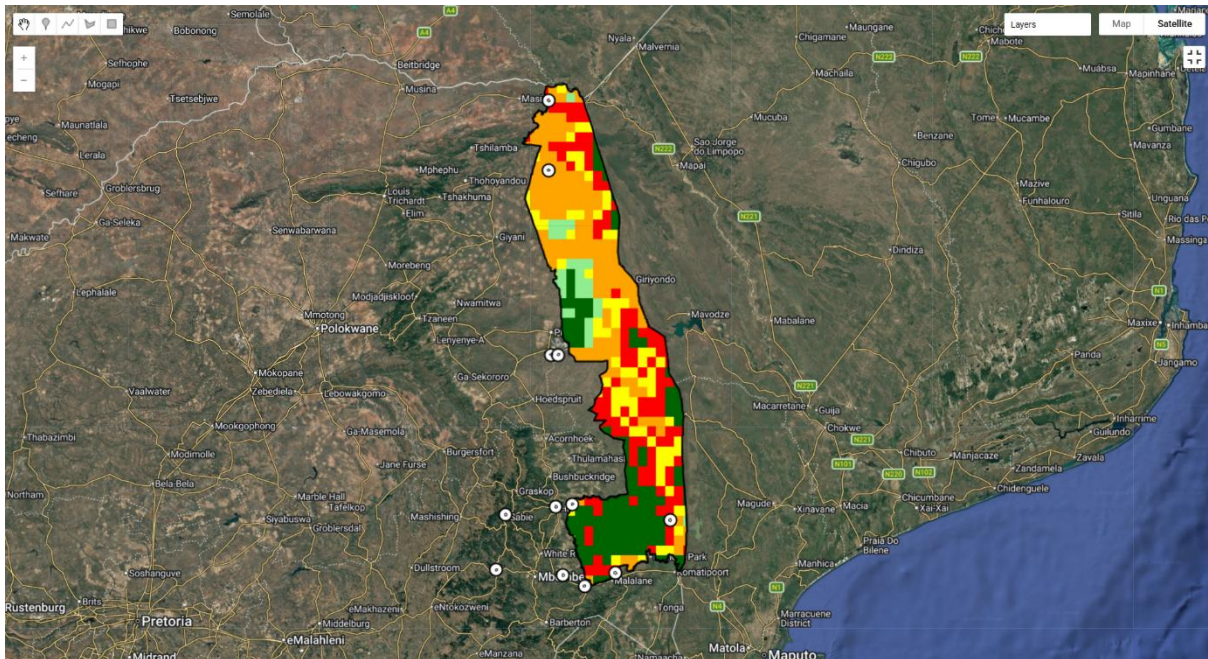


Figure 10. Spatial distribution of the reference locations (white dots) used for spatial comparison of the SVM-derived

4. Discussion

The presented study shows the effectiveness of integrating remote sensing, GIS, and ML techniques for pollution vulnerability assessment within a protected savanna ecosystem. The findings support the study objectives by successfully integrating atmospheric, climatic, and vegetation variables into a unified modelling framework capable of predicting and spatially mapping pollution vulnerability across KNP. The results further confirm the study hypothesis that ensemble-based machine learning approaches, particularly RF and GBM, outperform kernel-based approaches such as SVM in modelling complex environmental interactions associated with atmospheric pollution (Islam, 2025; Merabet et al., 2025). The spatial distributions presented in Figures 3 and 4 demonstrate substantial environmental heterogeneity across the park, reflecting the complex interactions between atmospheric pollutants, vegetation dynamics, and climatic conditions within protected savanna ecosystems. Elevated concentrations of NO₂ and SO₂ observed in several regions of the park may be associated with atmospheric transport processes, biomass burning, nearby anthropogenic activities, and regional emission sources, which have previously been identified as major contributors to atmospheric pollution in southern African environments (Aram et al., 2024; Morapedi & Obagbuwa, 2023). Variations in NDVI, rainfall, and temperature further highlight the ecological sensitivity of KNP, where climatic conditions strongly influence vegetation condition, fire susceptibility, and atmospheric composition (Van Wilgen et al., 2022). The successful integration of these multi-source datasets demonstrates the effectiveness of remote sensing and cloud-based geospatial platforms for environmental monitoring in data-scarce conservation regions.

The correlation analysis shown in Figure 4 provides further insights into the interactions among the investigated variables. The strong positive correlation between NO₂ and SO₂ ($r = 0.95$) suggests that both pollutants may come from related combustion-based processes, consistent with previous atmospheric studies conducted in southern Africa and other developing regions (Ketu, 2022; Liu et al., 2021; Vallejo et al., 2025). The strong negative relationship between rainfall and temperature ($r = -0.74$) aligns with established climatic dynamics in semi-arid savanna systems, where increased temperatures are frequently associated with drought conditions and reduced moisture availability. NDVI demonstrated positive correlations with NO₂ and SO₂, which may reflect the complex ecological and atmospheric interactions within KNP, where regions with denser vegetation can simultaneously experience elevated pollutant accumulation due to biomass burning, atmospheric deposition, and localised transport processes. This finding contributes to the significance of integrating vegetation dynamics within pollution vulnerability assessments rather than treating vegetation solely as an indicator of ecosystem health.

Moreover, the study aimed to identify and rank the primary drivers of pollution vulnerability. The hypothesis predicted that SO₂ and NO₂ would dominate, with NDVI and meteorological variables playing secondary roles. The feature importance results (Figure 5) across all three models consistently confirm the hypothesis that SO₂ is the most important predictor in the three models (RF, GBM, and SVM), followed closely by NO₂ and CH₄. This is systematically plausible, as SO₂ and NO₂ are key combustion by-products from industrial and vehicular sources, and their long-range transport into KNP has been documented by previous studies (Aram et al., 2024; Kamaludeen et al., 2025). The correlation matrix (Figure 4) further supports this, showing a very strong positive correlation between SO₂ and NO₂ ($r = 0.95$). NDVI, despite showing strong correlations with NO₂ ($r = 0.67$) and SO₂ ($r = 0.60$), ranked relatively low in feature importance, especially in SVM (Figure 5c). This suggests that while vegetation density covaries with pollutant levels, possibly due to shared spatial gradients (e.g., higher anthropogenic activity in greener, wetter areas), it is not a direct causal driver of pollution vulnerability. Temperature and rainfall showed moderate importance (Figure 5a, 5b), consistent with their well-known roles in pollutant dispersion and deposition (Yadav et al., 2026; Ravindiran et al., 2023). CH₄ exhibited a strong negative correlation with NO₂ ($r = -0.61$) and SO₂ ($r = -0.50$), which may reflect

different source signatures (biogenic vs. combustion). The primary drivers are, in order, $\text{SO}_2 > \text{NO}_2 > \text{CH}_4 > \text{temperature} > \text{rainfall}$, with NDVI and PM_{2.5} playing lesser roles.

We further involved implementing and optimised RF, SVM, and GBM to capture non-linear pollution environment relationships. Our hypothesis stated that integrating multi-source satellite data with collaborative ML would yield accurate vulnerability maps. The ROC-AUC values (Figure 9) strongly support this, as RF achieved an AUC of 0.96, GBM 0.92, and SVM 0.90, all indicating excellent discriminatory power. These values compare favourably with urban studies in (Lafortezza et al., 2026) and internationally, despite the greater environmental heterogeneity of a protected wilderness area. Regarding our second hypothesis (GBM has the highest accuracy, RF is the most interpretable), the results provide partial support. GBM yielded the highest overall classification accuracy (0.76), consistent with its sequential error-correcting mechanism (Mgomezulu et al., 2025; Uthappa et al., 2025). However, RF achieved the highest AUC (0.96), suggesting superior ranking ability despite slightly lower threshold-dependent accuracy (0.74). For interpretability, RF's feature importance ranking (Figure 5a) clearly identified SO_2 as the dominant driver, followed by NO_2 and CH_4 , confirming RF's well-known advantage in providing transparent variable importance via Gini impurity or MDI (Hoshyarzadeh et al., 2026). GBM also produced interpretable rankings (Figure 5b), while SVM's permutation-based importance (Figure 5c) was less intuitive. Thus, hypothesis 2 is accepted with the implication that RF excels in interpretability, while GBM leads in raw accuracy. The stronger performance of RF and GBM may also be attributed to their ability to handle nonlinear relationships and variable interactions more effectively within heterogeneous environmental landscapes such as KNP. RF, through bootstrap aggregation and random feature selection, provides strong generalisation capabilities, while GBM improves prediction accuracy through sequential error correction. This explains the relatively high agreement observed between RF and GBM predictions (83.3%) compared to the lower agreement between GBM and SVM (71.1%) shown in Figure 6. The strong three-way agreement (66.7%) further demonstrates the reliability and stability of the modelling framework. These findings indicate that despite methodological differences, the models consistently identified similar spatial pollution patterns across the park.

We conducted a comparative evaluation using ROC curves. The ROC analysis (Figure 9) visibly shows that all three models are highly effective, but with notable differences. RF's AUC of 0.96 indicates near-perfect class separability, surpassing values reported in similar AQI forecasting studies (Kamaludeen et al., 2025; Liu et al., 2021). GBM's accuracy of 0.76 is the highest among the three, aligning with findings that boosting frameworks (XGBoost, LightGBM) often outperform bagging methods in predictive tasks involving temporal dynamics (Songsom et al., 2025; Xie & Tummala, 2025). SVM, while still acceptable (accuracy 0.66, AUC 0.90), was the weakest performer, likely due to its sensitivity to parameter tuning and the high dimensionality of the feature space (Morapedi & Obagbuwa, 2023; Vallejo et al., 2025). The prediction overlap analysis (Figure 6) reinforces these rankings, where RF and GBM agreed on 83.3% of predictions, whereas GBM and SVM agreed on only 71.1%, suggesting that tree-based ensembles share more similar decision boundaries than kernel-based methods. RF provides the best overall discrimination (AUC), GBM the highest accuracy, and SVM a viable but less optimal alternative.

Lastly, we sought to produce actionable, high-resolution spatial maps identifying pollution hotspots. Figure 8 presents the vulnerability maps for RF (8a), SVM (8b), and GBM (8c), each reclassified into five categories (very low to very high). For RF, very high vulnerability covers 22.2% of KNP, for GBM, 23.3%, and for SVM, 18.9%. Moderate vulnerability is consistently the largest class across all models (28%). Spatial density analysis (Figure 7) reveals that all three models converge on the same geographic hotspots, particularly around longitude 31.7 to 31.9 and latitude -24.6 to -24.3 (the south-central region of KNP, near the Sabie and Crocodile River catchments). Secondary hotspots appear around 31.4 to 31.6 and -24.8 to -24.4. These areas are known to experience higher traffic from park entry gates and adjacent human settlements (Christie, 2024; Smit, 2013), supporting the interpretation that anthropogenic activity gradients drive the observed pollution patterns. The resulting maps (Figure 8) reveal distinct spatial concentrations of moderate-to-high pollution vulnerability across several sectors

of the park. The similarity of hotspot distributions across RF, GBM, and SVM suggests strong spatial consistency in the identified vulnerable zones. The KDE spatial density analyses presented in Figure 7 further confirm this consistency by showing concentrated high-density prediction zones around the central and eastern portions of the study area. These findings may reflect the combined influence of atmospheric transport, fire occurrence, vegetation structure, climatic variability, and anthropogenic pressures surrounding the park boundaries. Previous studies have shown that protected areas located near industrial corridors or transboundary pollution pathways are particularly susceptible to atmospheric contamination despite their conservation status (Aram et al., 2024; Morapedi & Obagbuwa, 2023).

The three-way agreement of 66.7% (Figure 6) indicates that all models concur with the classification of two-thirds of the test samples, providing high confidence in the combined vulnerability mapping. For park management, the RF-based map (Figure 8a) offers the highest spatial detail and interpretability, while the GBM map (Figure 8c) provides the best overall accuracy. The GIS-ready maps locate specific zones requiring priority intervention, such as targeted air quality monitoring or fire management adjustments in high-vulnerability areas. Despite the strong performance demonstrated in this work, there are gaps worth noting. The pollution vulnerability modelling framework was developed using satellite-derived environmental and atmospheric datasets at varying spatial resolutions (Nwaogu et al., 2024). Although these datasets provide extensive spatial coverage and are highly valuable for regional-scale environmental assessment, differences in sensor resolution and data resampling may introduce minor spatial inconsistencies within localised hotspot identification. Future studies should incorporate higher-resolution environmental datasets and additional ecological variables, such as wildfire occurrence, land-use dynamics, wind patterns, and topographic influences, to further improve the spatial precision of pollution vulnerability mapping. The inclusion of field-based validation measurements and seasonal monitoring could also enhance model calibration and strengthen ecological interpretation.

The findings have implications beyond the identification of spatial pollution patterns because they provide a spatial evidence base for sustainable protected-area management. The strongest alignment is with SDG 15 (Life on Land), as the pollution vulnerability maps can assist conservation managers in identifying areas requiring further environmental monitoring and management attention (Huan & Zhu, 2023; Mohanty et al., 2024). Such spatial information can support the protection of terrestrial ecosystems and biodiversity by helping managers incorporate pollution-related pressures into conservation planning. The study also contributes to SDG 13 (Climate Action) by integrating rainfall and temperature with atmospheric and vegetation indicators, thereby providing information that can support climate-sensitive environmental assessment and planning (Gaddam et al., 2026; Hamer et al., 2025). The relationship with SDG 3 (Good Health and Well-being) is indirect. Although the study does not assess human exposure, morbidity, or mortality, improved identification and management of pollution-vulnerable environments can contribute to broader efforts to reduce environmental pollution. Therefore, the principal contribution of the study to the SDGs is through spatial environmental monitoring and sustainable ecosystem management rather than direct measurement of health or climate outcomes.

5. Conclusion

This study developed an integrated GIS and machine learning framework for mapping pollution vulnerability in South Africa's Kruger National Park using remotely sensed, atmospheric, and climatic variables. The results demonstrate the potential of machine learning algorithms, particularly Random Forest, Gradient Boosting Machine, and Support Vector Machine, to identify and spatially represent areas exhibiting different levels of pollution vulnerability. The modelling results further indicate the importance of atmospheric pollutants, particularly SO₂ and NO₂, within the assessment framework.

The resulting spatial information provides a basis for targeted environmental monitoring and can support conservation managers in identifying areas that may require additional investigation or management intervention. The integration of GIS, remote sensing, and machine learning therefore provides a practical spatial decision-support approach for sustainable protected-area management. From a sustainable-development perspective, the study contributes most directly to SDG 15 (Life on Land) by supporting the protection and sustainable management of terrestrial ecosystems. It also supports SDG 13 (Climate Action) by incorporating climatic variables into environmental vulnerability assessment and contributes indirectly to SDG 3 (Good Health and Well-being) through improved environmental pollution monitoring. Future research should incorporate longer temporal datasets, field-based measurements, and additional environmental and socio-ecological variables to improve the validation and interpretation of pollution vulnerability patterns.

Declarations

Ethical approval

Not applicable. This study did not involve human participants, human subjects, identifiable personal data, or identifiable images.

Consent to participate

Not applicable. The study did not involve human participants.

Consent to publish

Not applicable. The manuscript does not contain identifiable information or identifiable images of individual participants.

Data availability

The datasets used in this study are publicly available through Google Earth Engine and the respective data providers identified in Table 1. The processed datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request, subject to the terms and conditions of the original data providers. The study did not generate personal or confidential data.

Competing interests

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.

Reference

- Al-Raei, M. (2024). Artificial intelligence for climate resilience: Advancing sustainable goals in SDGs 11 and 13 and its relationship to pandemics. *Discover Sustainability*, 5(1), 513. <https://doi.org/10.1007/s43621-024-00775-5>
- AlThuwaynee, O. F., Kim, S.-W., Najemaden, M. A., Aydda, A., Balogun, A.-L., Fayyadh, M. M., & Park, H.-J. (2021). Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms. *Environmental Science and Pollution Research*, 28(32), 43544–43566. <https://doi.org/10.1007/s11356-021-13255-4>
- Aram, S. A., Nketiah, E. A., Saalidong, B. M., Wang, H., Afitiri, A.-R., Akoto, A. B., & Lartey, P. O. (2024). Machine learning-based prediction of air quality index and air quality grade: A comparative analysis. *International Journal of Environmental Science and Technology*, 21(2), 1345–1360. <https://doi.org/10.1007/s13762-023-05016-2>
- Christie, N. C. (2024). *Assessment of the Possible Relationships and Influence of Climate Variability and Water Flow on Water Quality in Major Rivers of the Kruger National Park—ProQuest*. <https://www.proquest.com/openview/714dad371f3e1d64ce244c83093fb844/1?pq-origsite=gscholar&cbl=2026366&diss=y>
- Coetsee, C., Jacobs, S., & Govender, N. (2012). An Overview of Nitrogen Cycling in a Semiarid Savanna: Some Implications for Management and Conservation in a Large African Park. *Environmental Management*, 49(2), 387–402. <https://doi.org/10.1007/s00267-011-9779-0>
- Fazil, A. W., Hakimi, M., Akbari, R., Quchi, M. M., & Khaliqyar, K. Q. (2023). Comparative Analysis of Machine Learning Models for Data Classification: An In-Depth Exploration. *Journal of Computer Science and Technology Studies*, 5(4), 160–168. <https://doi.org/10.32996/jcsts.2023.5.4.16>
- Gaddam, V. K., Dutta, S., Ranjan, R., Kulkarni, A. V., & Karuppannan, S. (2026). Remote sensing based assessment of climate change impacts and ecosystem health in the Baspa River Basin, Western Himalaya, India: Implications for SDG 13. *Environmental Sciences Europe*, 38(1), 142. <https://doi.org/10.1186/s12302-026-01425-5>

- Goyal, K., & Goyal, S. (2024). Predicting PM2.5 Air Quality Using Random Forest Regression Enhanced with Polynomial Features. *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, 1–6. <https://doi.org/10.1109/IS61756.2024.10705219>
- Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Journal of Environmental and Public Health*, 2023, 1–26. <https://doi.org/10.1155/2023/4916267>
- Hamer, E. A., Oyedotun, T. D. T., & Nedd, G. A. (2025). Evaluating coastal agroecological dynamics using Landsat-derived vegetation and environmental indices embedded in Decision Support System and Monitoring Tools: Insights from Guyana towards achieving SDGs. *Discover Sustainability*, 6(1), 379. <https://doi.org/10.1007/s43621-025-01203-y>
- Hoshyarzadeh, R., Hafshejani, L. D., Tishehzan, P., & Parsaie, A. (2026). A hybrid framework of feature selection and interpretability for dissolved oxygen prediction in drinking water treatment plants. *Scientific Reports*, 16(1), 6912. <https://doi.org/10.1038/s41598-026-37276-6>
- Hossain, M. K. (2026). An Artificial Intelligence Integrated Geospatial Decision-Support Framework for Scenario-Based Climate-Adaptive Land-Use Optimization and Environmental Risk Reduction. *American Journal of Data Science and Analytics*, 7(06), 425–483. <https://doi.org/10.63125/nyc7sh45>
- Huan, Y., & Zhu, X. (2023). Interactions among sustainable development goal 15 (life on land) and other sustainable development goals: Knowledge for identifying global conservation actions. *Sustainable Development*, 31(1), 321–333. <https://doi.org/10.1002/sd.2394>
- Islam, M. M. (2025). MACHINE LEARNING-BASED PREDICTIVE MODELING FOR ASSESSING BRIDGE LOAD CAPACITY USING REAL-TIME SENSOR DATA. *Journal of Sustainable Development and Policy*, 4(03), 01–37. <https://doi.org/10.63125/v5y21788>
- Kamaludeen, A. M., Yusoff, I. M., & Yusuf, S. (2025). A Review: Identifying and Analyzing Air Pollution Hotspots Using Machine Learning and Remote Sensing Techniques. *International Journal of Environmental Sciences*, 11(22).

- Ketu, S. (2022). Spatial Air Quality Index and Air Pollutant Concentration prediction using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF): A case study in India. *Natural Hazards*, 114(2), 2109–2138. <https://doi.org/10.1007/s11069-022-05463-z>
- Lafortezza, R., Giordano, F., Capolongo, D., Refice, A., Lovergine, F. P., Elia, M., D’Alò, D., Amoroso, N., Nutricato, R., Zingaro, M., Parisi, A., Caporusso, G., Rizzo, F., Costantino, A., Ursi, A., Sacco, P., Virelli, M., & Tapete, D. (2026). Leveraging AI and multi-source satellite imagery for multi-hazard monitoring and susceptibility mapping in urban areas. *Scientific Reports*, 16(1), 21268. <https://doi.org/10.1038/s41598-026-52139-w>
- Liu, M., Chen, H., Wei, D., Wu, Y., & Li, C. (2021). Nonlinear relationship between urban form and street-level PM2.5 and CO based on mobile measurements and gradient boosting decision tree models. *Building and Environment*, 205, 108265. <https://doi.org/10.1016/j.buildenv.2021.108265>
- Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2024). *Applying Grid Search, Random Search, Bayesian Optimization, Genetic Algorithm, and Particle Swarm Optimization to fine-tune the hyperparameters of the ensemble of ML models enhances its predictive accuracy for mud loss*. In Review. <https://doi.org/10.21203/rs.3.rs-5187887/v1>
- Martins, F. P., Closs, J. G., Waked, D., Saldiva, P. H. N., & Veras, M. M. (2024). Positive Impacts of Air Pollution Reduction on SDG 3 Targets in Urban Environment. In N. Rezaei (Ed.), *Integrated Science for Sustainable Development Goal 3: Universal Good Health and Well-Being* (pp. 269–292). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64292-0_11
- McNeely, J. A. (1994). Protected areas for the 21st century: Working to provide benefits to society. *Biodiversity & Conservation*, 3(5), 390–405. <https://doi.org/10.1007/BF00057797>
- Mcneely, J. A. (2020). Today’s protected areas: Supporting a more sustainable future for humanity. *Integrative Zoology*, 15(6), 603–616. <https://doi.org/10.1111/1749-4877.12451>
- Merabet, K., Di Nunno, F., Granata, F., Kim, S., Adnan, R. M., Heddami, S., Kisi, O., & Zounemat-Kermani, M. (2025). Predicting water quality variables using gradient boosting machine:

- Global versus local explainability using SHapley Additive Explanations (SHAP). *Earth Science Informatics*, 18(3), 298. <https://doi.org/10.1007/s12145-025-01796-y>
- Mgomezulu, W. R., Thangata, P., Mkandawire, B., & Amoah, N. (2025). Advancing predictive analytics in child malnutrition: Machine, ensemble and deep learning models with balanced class distribution for early detection of stunting and wasting. *Human Nutrition & Metabolism*, 42, 200340. <https://doi.org/10.1016/j.hnm.2025.200340>
- Mohanty, S., Pandey, P. C., Pandey, M., Srivastava, P. K., & Dwivedi, C. S. (2024). Wetlands contribution and linkage to support SDGs, its indicators and targets- A critical review. *Sustainable Development*, 32(5), 5348–5392. <https://doi.org/10.1002/sd.2948>
- Morapedi, T. D., & Obagbuwa, I. C. (2023). Air pollution particulate matter (PM_{2.5}) prediction in South African cities using machine learning techniques. *Frontiers in Artificial Intelligence*, 6, 1230087. <https://doi.org/10.3389/frai.2023.1230087>
- Nadkarni, S. B., Vijay, G. S., & Kamath, R. C. (2023). Comparative Study of Random Forest and Gradient Boosting Algorithms to Predict Airfoil Self-Noise. *RAiSE-2023*, 24. <https://doi.org/10.3390/engproc2023059024>
- Ncube, M. M., & Ngulube, P. (2024). Enhancing environmental decision-making: A systematic review of data analytics applications in monitoring and management. *Discover Sustainability*, 5(1), 290. <https://doi.org/10.1007/s43621-024-00510-0>
- Nwaogu, C., Ajeyomi, S. A., Diagi, B. E., Nwaiwu, E. R., Anyalewechi, C. J., Alabi, B., Nwachukwu, J. I., Ejiogu, C. C., Iwuji, M. C., Mohammad, K. D., Fajembola, T., Cherubin, M. R., Ogbuagu, D. H., Lazarus, C., & Oduaro, I. J. (2024). The Role of Agriculture and Environmental Drivers on Soil Carbon in Brazil: A Remote Sensing, GIS, MLA, and Geostatistical Approach. *2024 IEEE International Conference on Agrosystem Engineering, Technology & Applications (AGRETA)*, 21–26. <https://doi.org/10.1109/AGRETA61912.2024.10949013>
- Ogwu, M. C., & Izah, S. C. (2024). Air Pollution and the Sustainable Development Goals. In M. C. Ogwu & S. C. Izah (Eds.), *Sustainable Strategies for Air Pollution Mitigation: Development*,

- Economics, and Technologies* (pp. 201–233). Springer Nature Switzerland.
https://doi.org/10.1007/698_2024_1118
- Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*, 236(7), 464. <https://doi.org/10.1007/s11270-025-08122-8>
- Partha Protim Roy, Md. Shahriar Abdullah, & Iqtiair Md. Siddique. (2024). Machine learning empowered geographic information systems: Advancing Spatial analysis and decision making. *World Journal of Advanced Research and Reviews*, 22(1), 1387–1397.
<https://doi.org/10.30574/wjarr.2024.22.1.1200>
- Pokharel, P. R. (2026). Advancing Sustainable Development Goal 15 (Life on Land): The Role of Corporate Environmental Initiatives. *Sustainable Development*, 34(S1), 769–782.
<https://doi.org/10.1002/sd.70221>
- Ramdani, F., & Furqon, M. T. (2022). The simplicity of XGBoost algorithm versus the complexity of Random Forest, Support Vector Machine, and Neural Networks algorithms in urban forest classification. *F1000Research*, 11, 1069. <https://doi.org/10.12688/f1000research.124604.1>
- Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, 338, 139518.
<https://doi.org/10.1016/j.chemosphere.2023.139518>
- Riddell, E. S., Govender, D., Botha, J., Sithole, H., Petersen, R. M., & Shikwambana, P. (2019). Pollution impacts on the aquatic ecosystems of the Kruger National Park, South Africa. *Scientific African*, 6, e00195. <https://doi.org/10.1016/j.sciaf.2019.e00195>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. <https://doi.org/10.58496/BJML/2024/007>
- Sharma, G., Khurana, S., Saina, N., Shivansh, & Gupta, G. (2024). Comparative Analysis of Machine Learning Techniques in Air Quality Index (AQI) prediction in smart cities. *International Journal of System Assurance Engineering and Management*, 15(7), 3060–3075.
<https://doi.org/10.1007/s13198-024-02315-w>

- Smit, I. (2013). Systems approach towards surface water distribution in Kruger National Park, South Africa. *Pachyderm*, 53, 91–98. <https://doi.org/10.69649/pachyderm.v53i.328>
- Songsom, V., Jaruk, P., & Suteerasak, T. (2025). Examining the spatiotemporal dynamics of urban heat island and its impact on air pollution in Thailand. *Environmental Challenges*, 19, 101120. <https://doi.org/10.1016/j.envc.2025.101120>
- Sudjianto, A., & Zhang, A. (2024). *Model Validation Practice in Banking: A Structured Approach for Predictive Models* (arXiv:2410.13877). arXiv. <https://doi.org/10.48550/arXiv.2410.13877>
- Tırink, S. (2025). Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM. *PLOS One*, 20(10), e0334252. <https://doi.org/10.1371/journal.pone.0334252>
- Uthappa, A. R., Das, B., Chavan, S. B., Raizada, A., Desai, S., Paramesha, V., Kumar, P., Jha, P. K., & Prasad, P. V. V. (2025). Comparison of machine learning models for mapping Arecanut based agroforestry system in Goa by enhancing precision and efficiency. *Scientific Reports*, 15(1), 44280. <https://doi.org/10.1038/s41598-025-27845-6>
- Vallejo, F., Yáñez, D., Viñán-Guerrero, P., Díaz-Robles, L. A., Oyaneder, M., Reinoso, N., Billartello, L., Espinoza-Pérez, A., Espinoza-Pérez, L., & Pino-Cortés, E. (2025). Enhancing air quality predictions in Chile: Integrating ARIMA and Artificial Neural Network models for Quintero and Coyhaique cities. *PLOS ONE*, 20(1), e0314278. <https://doi.org/10.1371/journal.pone.0314278>
- van Wilgen, B. W., Strydom, T., Simms, C., & Smit, I. P. J. (2022). Research, monitoring, and reflection as a guide to the management of complex ecosystems: The case of fire in the Kruger National Park, South Africa. *Conservation Science and Practice*, 4(4), e12658. <https://doi.org/10.1111/csp2.12658>
- Van Wilgen, B. W., Strydom, T., Simms, C., & Smit, I. P. J. (2022). Research, monitoring, and reflection as a guide to the management of complex ecosystems: The case of fire in the Kruger National Park, South Africa. *Conservation Science and Practice*, 4(4), e12658. <https://doi.org/10.1111/csp2.12658>

- Venter et al., (2008). *The Evolution of Conservation Management Philosophy: Science, Environmental Change and Social Adjustments in Kruger National Park | Ecosystems*.
<https://link.springer.com/article/10.1007/s10021-007-9116-x>
- Wahba, M. (2024). *Forecasting of flash flood susceptibility mapping using random forest regression model and geographic information systems*.
- Xie, Y., & Tummala, S. P. (2025). Machine Learning for Sensor Analytics: A Comprehensive Review and Benchmark of Boosting Algorithms in Healthcare, Environmental, and Energy Applications. *Sensors (Basel, Switzerland)*, 25(23), 7294. <https://doi.org/10.3390/s25237294>
- Yadav, N., Singh, S. K., & Sharma, D. (2026). Comparative Analysis of Random Forest, Support Vector Regression, Gradient Boosting and Phyllo-ViT Model for the Prediction of Air Pollutants. *2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, 732–735.
<https://doi.org/10.1109/IC3ECSBHI67834.2026.11468953>