DOI: https://doi.org
ComputeX - Journal of Emerging Technology & Applied Science
Journal homepage: https://rjsaonline.org/index.php/ComputeX

Data Science Applications in Environmental Monitoring Using Remote Sensing Data, Machine Learning Algorithms, and IoT-Based Sensors

Nida Qureshi [1], Muhammad Imran Siddiqui[2], Shiza Malik[3]

[1]Department of Environmental Science, Lahore University of Management Sciences (LUMS), Lahore, Pakistan
[2]Department of Computer Science, University of the Punjab, Lahore, Pakistan
[3]Department of Geospatial Information Systems, National University of Sciences and Technology (NUST), Islamabad, Pakistan
Email: nidaqureshi99@yahoo.com

ARTICLE INFO

Received:
December 16, 2025
Revised:
January 11, 2026
Accepted:
January 28, 2026
Available Online:
February 07, 2026

Keywords:
remote sensing, machine learning, IoT sensors, environmental monitoring, air quality, Random Forest, Support Vector Machine, neural networks, Lahore Pakistan, land surface temperature, NDVI, urban heat island.

Corresponding Author:
nidaqureshi99@yahoo.com

ABSTRACT

This paper tested how information technology methods, which include satellite tv for pc faraway sensing, system studying algorithms, and sensor integration the usage of the Internet of Things (IoT), may be used to display the surroundings in Lahore, Pakistan, one of the maximum environmentally pressured metropolitan areas in South Asia. The ordinary objectives had been to acquire and pre-method multi-supply multi-environmental information on Lahore, make use of system gaining knowledge of algorithms consisting of Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Networks (ANN) to categorise air fine and are expecting land floor temperature (LST), and compare the consequences of those fashions the use of preferred statistical metrics and to try to examine the 3 algorithms in phrases in their capacity to help in early detection of environmental anomalies. The satellite tv for pc structures owned through the US Geological Survey (USGS) and European Space Agency (ESA) had been used to collect environmental statistics, which had been complemented via way of means of readings of IoT-primarily based totally air excellent sensors and statistics of the Pakistan Meteorological Department. The preprocessing strategies worried cleansing of the facts, normalization of the records, and geospatial characteristic extraction with Python programs which include scikit-learn, GeoPandas, and Rasterio. However, out of the 3 algorithms considered, Random Forest become the maximum suitable in phrases of typical overall performance in air nice category with a excessive accuracy of 91.4, precision of 90.8, don't forget of 89.7 and F1-rating of 90.2. In predicting the land floor temperature, the ANN version gave the minimal root imply rectangular error (RMSE) at 1.84degC. The evaluation found out that the 3 maximum widespread functions in all of the fashions had been the flora index (NDVI), the awareness of PM 2.5, and the class of land use. The consequences imply that unified facts technological know-how structures related to far flung sensing, gadget mastering, and IoT sensor facts can provide a sturdy and scalable platform to real-time city environmental monitoring, and the direct packages to the safety of civic fitness and environmental coverage to Lahore and different South Asian cities.

Introduction

Population growth, industrialization, and speedy urbanization have created immeasurable environmental needs in South Asian towns, which can be pondered withinside the bad fine of air, growing city temperatures, flora loss, and ever-converting and unpredictable hydrological cycles (Shafique et al., 2018; WHO, 2022). One of those towns is Lahore, Pakistan, in which the populace of over thirteen million human beings is focused withinside the commercial capital of the Punjab province, which has turn out to be one of the maximum polluted city facilities withinside the international and maintains to take the second one area withinside the listing of towns with the very best tiers of first-rate particulate matter (PM2.five) and one of the most powerful city warmness island (UHI) consequences in South Asia (IQAir, 2023; Iqbal et al., 2021). To address those environmental troubles there need to be good enough tracking structures with good enough spatial decision, temporal frequency, and analytical complexity to pick out change, assign causality and assist put into effect well timed coverage response.

Conventional environmental surveillance strategies-wherein the important thing consciousness is on floor-primarily based totally measurements the use of constant stations-are significantly constrained in phrases of insurance in space, continuity in information and timeliness in response (Gubbi et al., 2013). A unmarried nicely-calibrated tracking station is incapable of taking pictures spatial heterogeneity of pollutants ranges or temperature gradients over a metropolitan region of the geographic length of Lahore (someplace round 1,772 km 2). With the appearance of 3 transformational technologic streams: satellite tv for pc faraway sensing, gadget getting to know, and Internet of Things (IoT) sensor networks, the opportunities of overcoming those barriers and constructing incorporated environmental tracking systems of the formerly unachievable functionality were opened (Reichstein et al., 2019; Zhu et al., 2019).

United States Geological Survey (USGS) and European Space Agency (ESA) run far flung sensing structures that provide multispectral and thermal imagery this is loose and at temporal resolutions that may be used to screen the surroundings close to continuously. The satellites Landsat eight and Landsat nine can provide an possibility to reap multispectral photographs with 30-meter spatial decision and a 16-day revisit cycle permitting systematic derivation of land floor temperature (LST), Normalized Difference Vegetation Index (NDVI), and quantity of impervious floor (Roy et al., 2014). Sentinel-5P satellite tv for pc gives facts at the each day atmospheric composition every day at a spatial decision of approximately 3.five km x five.five km with a worldwide insurance of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), aerosol optical intensity and carbon monoxide (CO) concentrations as a part of the Copernicus application operated through ESA (Veefkind et al., 2012). When blended with measurements at the floor the usage of the IoT sensors, those satellite tv for pc statistics units may be used to cross-validate and spatially downscale remotely sensed measurements to a decision this is beneficial in city environmental management.

There is a transformative capacity of gadget mastering algorithms withinside the software of environmental technology, wherein automatic identity of developments in high-dimensional, heterogeneous datasets may be executed that isn't always possible the usage of conventional statistical techniques (Lary et al., 2016; Reichstein et al., 2019). An ensemble studying set of rules referred to as Random Forest this is based on bootstrap aggregation of choice bushes has proved to be especially powerful in environmental class obligations due to the fact it's far proof against noise, lets in the estimation of characteristic significance and overfitting (Breiman, 2001). The Support Vector Machines, primarily based totally at the statistical mastering theory, has been used appreciably in far flung sensing photograph type and prediction of environmental variables, and they are able to carry out nicely in variables which can be high-dimensional and with fantastically smaller schooling datasets (Mountrakis et al., 2011). The spatial sample recognition, time-collection prediction, and multi-variable environmental modeling responsibilities are all modern-day with Artificial Neural Networks, and their deep mastering extensions (LeCun et al., 2015).

The use of statistics technological know-how and system gaining knowledge of to screen the surroundings remains younger withinside the Pakistani putting. International research have generated superior included tracking structures in European towns, North American towns, and East Asian metropolitan regions, however comparable structures tailor-made to the environmental surroundings, records infrastructure, and coverage desires of Pakistani city regions aren't represented withinside the literature (Hamid et al., 2020; Shafique et al., 2018). This distinction is consequential: the environmental statistics, sensor networks, and modeling methodologies that had been educated on high-profits u . s . settings is probably inapplicable to towns with distinctive profiles of pollutants sources, flora structure, climatic situations in addition to boundaries of information availability. Lahore is therefore a case observe in addition to an possibility to transport at the contextually applicable environmental statistics technology technique that may be applicable all through the growing global.

To fill withinside the recognized gap, this studies paper provides the improvement and trying out of an incorporated environmental tracking framework of Lahore, that is primarily based totally on satellite tv for pc faraway sensing

measurements of USGS and ESA satellites, air great sensor statistics of 3 system mastering algorithms, consisting of RF, SVM, and ANN, carried out in Python. This studies has 4 centered objectives, namely, (1) to compile, pre-process, and combine a multi-supply environmental records that describe air excellent, land floor temperature, and flowers circumstance throughout Lahore; (2) to check and evaluate RF, SVM, and ANN algorithms in air high-satisfactory category and land floor temperature prediction; (3) to estimate the version overall performance withinside the shape of accuracy, precision, recall, F1-score, and RMSE; and (4) to decide the environmental characteristic that great predicts negative environmental situations withinside the city placing of Lahore. The effects will upload each methodologically, with the aid of using demonstrating an advanced high-satisfactory exercise of incorporated environmental tracking in South Asian city environments with restricted statistics, and practically, via way of means of providing evidence-primarily based totally environmental coverage and interventions at the populace in Lahore.

Literature Review

The use of satellite tv for pc far flung sensing in tracking the city surroundings has a big and an growing literature in numerous carefully related environmental fields. Weng (2009) offered a totally informative overview of the packages of far off sensing to city warmness island research revealing that thermal infrared sensors on Landsat and MODIS sensors had been able to reliably measuring intra-city LST gradients of 4-12degC among vegetated and impervious regions. Imhoff et al. (2010) furthered this observe via way of means of quantifying the UHI depth in 38 biggest towns globally thru MODIS LST, organising that UHI results have been maximum excessive in the ones towns with excessive impervious floor fractions and occasional plants cores, each of which can be very regular of the city middle of Lahore. Yuan and Bauer (2007) had proven that NDVI the use of Landsat imagery turned into a legitimate predictor of floor temperature and an inverse correlation among NDVI and LST accounted as much as seventy nine in keeping with cent of version in intra-city LST distributions throughout Minneapolis-St. Paul.

Satellite-derived aerosol optical intensity (AOD) has end up famous withinside the area of air pleasant tracking in which a proxy of floor-degree PM2.five concentrations is desired (consisting of in regions with sparse floor tracking networks). van Donkelaar et al. (2010) have used aerosol optical intensity (AOD) as a satellite tv for pc dimension to derive a globally relevant regression version to expect floor-stage PM2.five concentrations with correlation coefficients extra than 0.70 in one of a kind climatic regions. In the case of South Asia, Begum et al. (2011) exhibited that the satellite tv for pc-derived AOD changed into statistically huge with PM2.five ranges at the floor in Dhaka, Bangladesh, which means that the satellite tv for pc- proxy techniques will be relevant withinside the Pakistani city settings. In 2017, the Sentinel-5P TROPOMI satellite tv for pc, that's an instrument, has notably superior our operational cappotential of air best tracking via space, in which we're confident of day by day maps of column densities of NO2 on a planetary scale with a decision that is right sufficient to discover character business flora and street networks (Veefkind et al., 2012; Lorente et al., 2019).

Environmental tracking IoT Sensor Networks

With the upward push of cheaper IoT environmental tracking sensors, a huge quantity of literature that explores the technical talents and validation of those structures were generated. Kumar et al. (2015) performed a evaluate of the evolution of the low-price air first-rate sensor device and found out the primary blessings of the structures which includes spatial density, real-time transmission of information and obstacles along with sensor drift, humidity and temperature cross-sensitivity, and stringent calibration procedures. A evaluate of low-price sensors to degree ambient air excellent via way of means of Morawska et al. (2018) diagnosed the electrochemical and optical sensors that could have affordable accuracy on PM2.five and NO2 as soon as cautiously co-positioned in regards units to apply them as a calibration.

General IoT sensor networks were utilized in city utility of environmental tracking to attain spatial resolutions that might now no longer be practicable with conventional fixed-station techniques. Mead et al. (2013) confirmed that a community of forty two cheaper electrochemical sensors disbursed during Cambridge, United Kingdom, may also document intra-city pollutants gradients in a finer spatial decision that became now no longer viable to decide the use of the modern regulatory tracking community. Rai et al. (2017) additionally installed that strategies of facts fusion among the evaluation of IoT sensors and satellite tv for pc measurements had the ability to seriously outperform one of the statistics reassets with city PM2.five mapping, which at once stimulated the records integration approach selected withinside the contemporary have a look at. As a pilot examine of an IoT-primarily based totally air nice tracking community in Pakistan, the item with the aid of using Hamid et al. (2020) examined the community in Lahore and discovered that PM2.five concentrations taken in business regions five-12 instances better than the values of WHO recommendations for the duration of winter.

Environmental Science packages of gadget mastering

The gadget getting to know strategies were carried out withinside the complete variety of environmental tracking activities, beginning with satellite tv for pc photo class all of the manner to predicting air nice and downscaling climate. Lary et al. (2016) reviewed using system mastering withinside the discipline of atmospheric technology and confirmed that neural community and ensemble fashions have been uniformly advanced to the previous strategies of linear facts in multi-variable environmental prediction. Pal and Mather (2005) have proven that SVMs in far off sensing photo class, the location of take a look at immediately associated with the modern take a look at, outperform traditional most probability classifiers on Landsat imagery with the aid of using 2-five%, and additionally require tons much less education statistics, that is vital in research wherein facts are restricted and this kind of huge dataset is generally unavailable.

One of the great algorithms which have been evolved withinside the discipline of environmental faraway sensing has been the Random Forest. Belgiu and Dragut (2016) carried out a evaluation of one hundred posted RF-primarily based totally packages to faraway sensing category obligations and determined typical accuracies of 88.7% and mainly in land use and land cowl type responsibilities applicable to city environmental tracking. In case of air first-rate prediction, specifically, to expect PM2.five the use of satellite tv for pc-derived AOD and meteorological covariates, Chen et al. (2018) used RF on each, which brought about R2 of 0.87 with RMSE of round 12.6 ug/m3, that is a good deal better than the benchmarks of geographically weighted regression and kriging. Masood and Ahmad (2020) used RF with environmental facts at the Indo-Gangetic Plain and proved excessive predictive nice of PM2.five at some point of monsoon situations and post-monsoon situations, geographic and climatic situations that had been near analogs of Lahore.

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are deep getting to know techniques which have tested the nation of artwork overall performance in each spatiotemporal environmental prediction duties. Zhang et al. (2019) used an LSTM community at the air great time-collection information of the towns in China and finished a better margin of 15-23% of RMSE metrics over RF and SVM baselines. Reichstein et al. (2019) performed a overview of ways deep gaining knowledge of have been implemented to Earth machine technology and located that the destiny of the next-era environmental tracking structures lay in hybrid fashions that blended bodily procedure fashions and neural community architectures. On the Pakistani context, Shafique et al. (2018) used ANN on Landsat-derived LST information of Lahore and discovered that the nonlinear relationships among the city floor houses and temperature can be modeled via way of means of the neural networks with imply absolute mistakes of round 1.9degC.

Data Fusion and Frameworks of Integrated Monitoring

The mixture of numerous streams of facts, consisting of satellite tv for pc, IoT sensor, and auxiliary geospatial information into unified environmental tracking structures, represent a route of energetic methodological innovation. The article via way of means of Zhu et al. (2019) carried out a survey of records fusion in far flung sensing and outstanding 3 fundamental trends: pixel-degree fusion (integration of co-registered images) and characteristic-degree fusion (integration of extracted functions of various images) and decision-stage fusion (integration of one after the other educated classifiers). In the case of city environmental tracking specifically, function-degree fusion has tended to be extra powerful than different techniques while integrating floor sensor and satellite tv for pc records (Rai et al., 2017).

It has been proven with the aid of using diverse research that the IoT sensors information may be beneficial to complement the satellite tv for pc records in South Asian city environmental tracking. By integrating Sentinel-5P NO2 measurements with floor sensor measurements over Delhi, Rahman et al. (2019) determined that spatial interpolation of floor-primarily based totally facts with covariates (satellite tv for pc measurements) ended in significantly decrease RMSE PM2.five maps as compared to the usage of floor-most effective techniques. Gupta and Christopher (2009) discovered that addition of land use and meteorological traits in addition to satellite tv for pc AOD to the device studying fashions led to great higher capabilities in predicting PM 2.five, indicating that using multi-supply characteristic engineering is critical in acquiring the accuracy of operational tracking. The modern paper is primarily based totally in this literature with the adoption of systematic characteristic fusion approach of Lahore and an in-intensity evaluation of the 3 maximum used gadget getting to know algorithms withinside the discipline.

Environmental Monitoring in Pakistan: Vacuity and Opportunity

The environmental tracking literature centered on Lahore and Pakistan in widespread demonstrates the urgency of environmental problems and a massive hole withinside the technique of the to be had our bodies of evidence. One of the maximum exhaustive assessment of the air great in Lahore become done with the aid of using Iqbal et al. (2021) who mixed each floor measurements and satellite tv for pc benchmarks and recorded each intense and chronic exceedances of WHO PM2.five standards, particularly withinside the season of October-February smog, and eliminated maximum of the PM2.five

load to the motors and industries combustion and burning of crop waste withinside the surrounding agricultural fields. Ali et al. (2021) used the Landsat-primarily based totally LST evaluation to degree the UHI impact in Lahore, and that they concluded that city middle temperatures are better than the peri-city ones through 4.8-7.3degC withinside the summer time season months and the spatial scale of the UHI had elevated with the aid of using approximately 23 percentage among 2000 and 2020.

Although the above noticeable environmental evaluation has been done, there's restricted systematic use of system getting to know and incorporated IoT-satellite tv for pc fashions withinside the software of environmental tracking in Lahore. The articles determined at some stage in this evaluation used both character facts reassets (satellite tv for pc or floor), conventional statistical in place of gadget getting to know techniques, or did not carry out a multi-set of rules comparative evaluation to decide the fine techniques to apply specially environmental tracking responsibilities withinside the Lahore setting. The present day look at without delay fills those gaps with the aid of using imparting an incorporated and multi-set of rules, multi-supply design, and improving the ever-increasing literature on facts technological know-how packages to environmental tracking withinside the swiftly urbanizing towns of growing countries.

Methodology

Research Design

Quantitative studies layout become for use primarily based totally on records. The studies become performed in a well-built pipeline that covered information gathering, pre-processing, function engineering, version schooling and assessment, and average comparison. The whole technique of calculations became found out in Python three.10 primarily based totally at the scikit-learn, GeoPandas, Rasterio, NumPy, Pandas, and Matplotlib packages.

Study Area

The studies pattern become set up because the Lahore metropolitan district which covers round 1,772 km2 and a populace of round thirteen million people (PBS, 2023). Lahore is discovered in northeastern Punjab (31.5degN, 74.3degE) at an elevation approximately 217 m above the ocean level. The weather withinside the town is semi-arid and characterised with the aid of using moist and dry seasons (July-September and November-March respectively). It changed into set in January 2020 to December 2022, which gave 3 entire cycles of version schooling and validation.

Data Collection

There had been 4 main classes of reassets of environmental information. Two satellites had been used to acquire satellite tv for pc far off sensing facts, that are the Landsat eight Collection 2 Level-2 merchandise, which had been downloaded at the USGS Earth Explorer portal (earthexplorer.usgs.gov) and contained floor reflectance bands and floor temperature merchandise at spatial resolutions of 30 meters and sixteen days revisit interval. The level-2 merchandise of Sentinel-5P TROPOMI had been acquired on the ESA Copernicus Open Access Hub and it contained day by day retrievals of NO2, SO2, aerosol index, and CO column densities at a three.five km x five.5km spatial decision. The widespread components NDVI = (NIR - Red)/ (NIR + Red) become used to compute NDVI primarily based totally on Landsat eight Bands four (Red) and five (Near-Infrared). LST turned into created via way of means of radiometric calibration and atmospheric correction of Landsat band 10 as mentioned withinside the USGS Landsat Collection 2 product documentation.

The IoT-primarily based totally air first-rate sensors facts had been accrued in reassets one from publicly to be had facts of PurpleAir sensors established withinside the metropolitan town of Lahore, and the opposite information became accrued via Pakistan Environmental Protection Agency (EPA) Punjab floor tracking stations. The Pakistan Meteorological Department (PMD) furnished the meteorological information of the Lahore Aerodrome station inclusive of temperature, humidity, wind speed, and wind route that have been supplemented through the ERA5 reanalysis records supplied with the aid of using the European Centre of Medium-variety Weather Forecasts (ECMWF). The information on land use and land cowl (LULC) changed into received primarily based totally at the ESA WorldCover 2021 10-meter decision international map.

Data Preprocessing

Preprocessing become executed in 5 consecutive steps. The cloud protecting turned into finished on all of the Landsat eight scenes with QA PIXEL band withinside the first step and most effective the scenes with cloud cowl of 20 percentage or extra had been disqualified, giving a very last range of fifty four beneficial Landsat scenes in the course of the examine duration. During the second one degree, all satellite tv for pc-derived raster statistics had been re-projected right into a not unusualplace coordinate reference system (WGS 84 / UTM Zone 43N) and resampled to a not unusualplace 30 meter spatial decision with the aid of using bilinear interpolation. Outlier detection carried out to IoT sensor readings primarily based

totally on three-sigma rule and elimination of readings that have been now no longer of bodily manageable variety passed off withinside the 1/3 degree. The gaps in time-collection sensor statistics have been imputed to the lacking values with linear interpolation among gaps longer than six consecutive hours. The fourth level worried standardizing of functions to the variety [0, 1] thru min-max scaling all non-stop enter functions earlier than the version changed into educated to offer set of rules insensitive comparison amongst capabilities that had distinct size devices and stages. At the 5th phase, a 30 m decision grid of fishnet turned into superimposed at the look at vicinity and environmental variables had been spatially joined to grid mobileular to offer a tabular dataset of 196,854 information every of which represented a 30 m x 30 m grid mobileular and had satellite tv for pc, sensor, meteorological and LULC characteristic values attached.

Feature Engineering and Selection

The preprocessed datasets had been transformed to 13 enter capabilities, NDVI, LST, Aerosol Optical Depth (AOD), NO2 column density, SO2 column density, PM2.five (measured through IoT), relative humidity, wind speed, wind path, distance to nearest most important road (km), distance to nearest business zone (km), LULC class (encoded as integer), and month of year (1-12) as a proxy of seasonal variation. The integrated significance rankings in RF which can be primarily based totally at the Gini impurity, have been used to degree the significance of functions and provide an preliminary rating to resource interpretive analysis. ∈

Training and Evaluation Model

Stratified random sampling became used to cut up the dataset into schooling (70%), validation (15%), and test (15) subsets in order that every break up has proportional representations of air pleasant classes. There had been duties on 3 device studying fashions skilled. To classify air nice, PM2.five degrees have been categorised into 4 in line with the Pakistani recommendations of the EPA that set up Good (0-35 ug/m3), Moderate (36-seventy five ug/m3), Unhealthy (76-a hundred and fifty ug/m3), and Hazardous (>a hundred and fifty ug/m3). RF turned into used while the pattern length turned into two hundred trees, the most intensity changed into 20, and the minimal range of samples in any leaf turned into five and the Gini criterion changed into carried out. SVM turned into used with radial foundation function (RBF) as kernel, the regularization parameter C and the bandwidth of the kernel gamma have been optimized thru grid seek the use of cross-validation at the stages C ∈[0.1, 1, 10, 100] and gamma [?] [0.001, 0.01, 0.1, 1]. ANN become skilled as a 3-layer perceptron (128, 64, and 32 neurons), ReLU activation, Adam optimizer, studying fee of 0.001 and early preventing the usage of a staying power of 10 epochs. In the case of LST regression, 3 algorithms had been educated at the thirteen-function set to make predictions on non-stop values of LST withinside the shape of stages Celsius, and the overall performance changed into measured through RMSE, Mean Absolute Error (MAE), and R2.

Data Analysis

The descriptive data of environmental variables are furnished withinside the examine with the aid of using four.1. The descriptive information of the principle environmental variables are furnished in Table 1 withinside the complete 196,854 data dataset. The variety of PM2.five became eight.three ug/m3 to 487.6 ug/m3 with a median of 112.four ug/m3 (SD = 78.2 ug/m3) which become greater than 3 instances the country wide popular of 15 ug/m3 as set with the aid of using the WHO, however nearly two times the countrywide fashionable of seventy five ug/m3 installed with the aid of using Pakistan itself. The seventy five th percentile of PM2.five (142.eight ug/m3) become withinside the class of Unhealthy and it method that the damaging air nice situations occupied maximum of the length of the have a look at. The LST values had been among 14.2degC and fifty four.7degC with manner of 36.8degC and a wellknown deviation of eight.4degC, which confirmed the extremes of the thermal variety Lahore is characterised via way of means of because of its semi-arid weather. The Lahore NDVI values had been among -0.08 and 0.sixty eight with a median of 0.19, which means that maximum of the land floor in Lahore is occupied via way of means of impervious cowl with minimum plants cowl as it's miles determined in Ali et al. (2021).

Table 1: Descriptive Statistics of Primary Environmental Variables – Lahore (2020–2022)

Variable	Min	Max	Mean	SD	Unit
PM2.5	8.3	487.6	112.4	78.2	μg/m³
LST	14.2	54.7	36.8	8.4	°C
NDVI	-0.08	0.68	0.19	0.14	Index
NO_2 Column	18.4	142.7	64.3	29.8	μmol/m²
AOD (550nm)	0.12	1.84	0.67	0.31	Dimensionless
Wind Speed	0.4	12.8	3.6	2.1	m/s
Relative Humidity	18.2	94.7	52.3	18.9	%

The seasonal PM2.5 concentrations showed strong winter peaks which are in line with the Lahore smog phenomenon reported by Iqbal et al. (2021). The peak PM 2.5 levels of 218.6 ug/m3 in November-January were 5 times greater than the 68.4 ug/m3 that PM 2.5 had in the summer of June-August. This seasonal occurrence could be explained by a combination of high emissions of biomass burning due to crop residue burning in the nearby agricultural regions, low depth of atmospheric mixing during winter temperature inversion, and inhibition of wet deposition during the dry season of the year.

Spatial analysis of LST in the districts of Lahore showed the structure of urban heat islands with the highest mean LST level observed in the central commercial and industrial districts (Data Ganj Bakhsh Town: 41.2degC; Ravi Town: 40.8degC) and the lowest one in the peri-urban areas with the remaining vegetation cover (Cantt. area: 34.1degC). This distribution pattern was similar in form to the inverse of the NDVI distribution, and Pearson correlation coefficients of NDVI and LST between NDVI and LST were $r = -0.73$ ($p < 0.001$) across the entire data set, as is common in the literature of remote sensing studies of vegetation-temperature relationships (Weng, 2009; Yuan and Bauer, 2007).

The results of the air quality classification will be discussed below

Table 2 shows the classification performance of RF, SVM, and ANN on the held-out test set (n = 29528 records). Random Forest demonstrated the greatest overall accuracy (91.4) and was then ANN (89.1) and SVM (85.7). The RF showed performance advantage over all of the four classification categories; however, notably strong performance advantages in the most important category to the public health decision-making, namely, the Hazardous category where RF recalls of 88.3% were significantly higher than ANN recalls of 83.7% and SVM recalls of 77.9%.

Table 2: Performance of RF, SVM and ANN on Air Quality Classification - (Test Set)

Metric	Random Forest	SVM	ANN	Notes
Accuracy	91.4%	85.7%	89.1%	Overall
Precision (macro)	90.8%	84.3%	88.6%	Avg. across classes
Recall (macro)	89.7%	83.1%	87.9%	Avg. across classes
F1-Score (macro)	90.2%	83.7%	88.2%	Harmonic mean
AUC-ROC	0.967	0.941	0.958	Multi-class OvR
Hazardous Recall	88.3%	77.9%	83.7%	Critical class

Analysis at the class level showed that the lower performance of all three models was on the category of "Moderate" which is both put at the same boundaries as the other two categories in that PM2.5 concentrations overlap and, therefore, is the most ambiguous in terms of classification. The F1-score of RF in the Moderate-class was 86.4-, which is higher than the SVM (79.8%) and ANN (83.1). It was the Good category that was best assigned across all models (RF: 94.2%, SVM: 91.7%, ANN: 93.1%), which is expected given that clean air conditions since they are not frequent in Lahore are characterized by unique patterns of features (low NDVI in reverse, high wind speed, low humidity, summer season) that are easily separable in comparison to polluted conditions.

The confusion matrix analysis of the RF model showed that the most significant misclassification behavior was the one of adjacent categories (e.g., Unhealthy versus Moderate), misclassification across categories (e.g., Hazardous versus Good) was less than 0.3% of the test cases. This misclassification structure is associated with the ordered nature of the underlying PM2.5 variable, which is continuous, and suggests that the model is committing physically sensible errors even where it is not performing perfect classification.

The results of the Land Surface Temperature prediction are given in 4.3

The metrics of the regression performance of LST prediction are shown in Table 3. ANN model had the best RMSE (1.84degC) and R2 (0.912) then RF (RMSE = 2.17degC, R2 = 0.886) and SVM (RMSE = 2.83degC, R2 = 0.841). The high accuracy of ANN in LST regression task, as compared to classification task which RF is leading, is in line with the fact that neural networks are more apt to reflect complex nonlinear spatial relationships among continuous environmental variables, especially when they are trained using large datasets where their capacity advantage over RF and SVM can be fully realized (Reichstein et al., 2019).

Table 3: LST Prediction (RF, SVM and ANN) (Test Set)

Metric	Random Forest	SVM	ANN	Unit
RMSE	2.17	2.83	1.84	°C

MAE	1.63	2.14	1.41	°C
R^2	0.886	0.841	0.912	–
MAPE	4.8%	6.3%	4.1%	%

Spatial prediction errors of LST showed that all three models generated the lowest prediction errors in areas of transition between high-density urban and peri-urban locations where LST at gradients are the steepest, and the correlation between surface characteristics and temperature is most sensitive to fine scale spatial variation in land cover. The largest error magnitude were recorded in the highest summer months (May-June) when LST values in the center Lahore were above 50degC and the spatial temperature gradient between the urban and vegetated surfaces was greatest. This result indicates that the results of the model might be enhanced by adding more predictors such as higher resolution land cover data and urban morphology characteristics (building height, street canyon geometry).

Importance of Features Analysis

The scores of the RF feature importance in the air quality classification task were shown in Figure 1 (described below) ordered by the Gini impurity reduction. The most significant feature (importance score: 0.187) was NDVI, which is associated with the ability of various aspects of the urban surface condition, including vegetation cover, land use intensity, and surface permeability, to combine and affect PM2.5 concentrations both directly and indirectly (both source and dispersion). The second feature of the greatest importance was PM2.5 (IoT-measured) 24-hour lagged (importance: 0.163) which implies that the stability of the air quality situation over time is a significant predictor of the current PM2.5 level. The third category (importance 0.142) was LULC category, next comes relative humidity (0.121) and distance to closest industrial zone (0.108), and the last one is NO2 column density (0.097). The wind speed (0.072) and wind direction (0.061) were also significant, which is in line with the established significance of meteorological dispersion in the process of determining urban PM 2.5 spatial patterns.

Table 4: Importance of Features Score - Random Forest Air Quality Classification Model.

Feature	Importance Score	Environmental Interpretation
NDVI	0.187	Vegetation cover modulates deposition & mixing
PM2.5 (24h lag, IoT)	0.163	Temporal persistence of pollution conditions
LULC Category	0.142	Source intensity proxy (industrial vs. residential)
Relative Humidity	0.121	Aerosol formation & hygroscopic growth
Distance to Industrial Zone	0.108	Proximity to point emission sources
NO_2 Column Density	0.097	Photochemical production co-indicator
Wind Speed	0.072	Atmospheric dispersion capacity
Wind Direction	0.061	Source directionality (rural biomass burning)
Month of Year	0.049	Seasonal pattern (winter smog cycle)
AOD	0.041	Column-integrated aerosol loading
Distance to Road	0.038	Vehicular emission proximity
SO_2 Column Density	0.014	Industrial combustion indicator
LST	0.007	Urban thermal environment

The fact that NDVI has the highest ranking of feature importance in the feature importance ranking has significant practical implications on the environmental management in Lahore. It proposes that urban greening projects, which can increase vegetation cover via parks, street trees and green roofs, can offer co-benefits that are not necessarily the reduction in UHI (due to direct shading and evapotranspiration cooling) but also the improvement of air quality, by increasing dry deposition of particulate matter in the atmosphere and surface roughness in a manner that facilitates atmospheric mixing. Such an interpretation is in line with the results of Nowak et al. (2014), who approximated that urban trees in the United States would eliminate around 711,000 metric tons of air pollutants each year and with Abhijith et al. (2017), who revealed that vegetative obstacles near roads could eliminate PM concentrations at the pedestrian level by 17-50%.

Temporal Analysis and Anomaly Detection 4.5

The RF classification outputs were analyzed with time series over the period of 36 months of study, which showed a set of regular seasonal and a couple of non-periodic outliers. The model identified 68.4% of Lahore grid cells as a Hazardous state at the climax smog of November 2021, the worst episode in the study period, which coincided with the reading of ground-

truth sensors, which measured 72-hour mean PM2.5 value of more than 350 ug/m3 in several monitoring stations. This extreme episode of the RF model with a correct classification of 92.3% of grid cells (as compared to 89.1% and 88.6% of ANN and SVM respectively) indicates that it has specific applications in the context of early warnings in situations most relevant to the health of the populace.

The addition of the IoT sensors helped the performance in terms of time especially in the occurrence of episodic pollution. Stratified analysis of model performance by source of data demonstrated that models trained only on satellite-derived features had a lower overall accuracy (7.4) compared to the integrated models at times of pollution, but a lower overall accuracy (1.8) compared to times of background. Such heterogeneous data collection highlights the value of IoT sensors, which are based on the continuous collection of measurements with an intrinsic time resolution, and satellite data, which are based on space coverage but cannot be revisited sufficiently often and the coverage of cloud cover, in combined monitoring systems.

Discussion

The results of this research are substantive to demonstrate the fact that integrated data science models involving satellite remote sensing, IoT-based sensor data, and machine learning algorithms can be used to attain a high level of performance in environmental monitoring of a densely populated city with a high level of pollution in South Asia. Random Forest is the most effective method in air quality classification in accordance with the findings of Belgiu and Dragut (2016) and Chen et al. (2018), who also determined that RF was the most resilient algorithm when it comes to the environmental classification tasks with heterogeneous feature distributions, inter-class imbalance, and nonlinear interaction effects. The top-ranked ANN in continuous LST prediction corresponds with Reichstein et al. (2019) and Shafique et al. (2018), who proved that the ability of neural networks to predict the specific nonlinearities in space give special benefits to the regression when training on a vast dataset.

Identification of NDVI as the most significant feature to classify air quality is an original policy-relevant discovery. Although the negative relationship between vegetation and LST is an established fact (Weng, 2009), the same significance of vegetation cover on PM2.5 classification, as evident here with the use of RF feature importance analysis, implies that urban greening policies can have a more significant implication on air quality management than what is currently understood in the environmental policy discourse of Lahore. The significant predictive performance of the IoT-measured PM2.5 lag values (24 hours) are one more confirmation of the fact that near-real-time air quality forecasting systems on the basis of the modeling framework created within the frames of the present research study can offer practical information on the daily public health recommendations and emergency response strategies. The weaknesses of the present study are that the spatial resolution of Sentinel-5P atmospheric data is relatively coarse, and it will have to be spatially interpolated to fit the 30-meter grid of analysis, and that IoT sensors are concentrated within the central Lahore districts, which can create spatial bias in the integrated training dataset. These limitations can be mitigated by future research by deploying IoT sensors to peripheral districts, including data on urban morphology (at the building level), and using spatiotemporal deep learning architectures (e.g., ConvLSTM) that have the capacity to simultaneously learn spatial and temporal environment dynamics.

Conclusion

This paper designed, deployed, and tested an environmental monitoring system in Lahore, Pakistan, utilizing a combination of Landsat 8 and Sentinel-5P satellite remote sensing data, an IoT-based air quality sensor system, and three machine learning classifiers, namely the Random Forest, Support Vector Machine, and Artificial Neural Network. The framework was implemented to tasks of air quality classification and prediction of land surface temperature during a 36-month study period (2020-2022) with the help of 196,854 spatially referenced records. The researchers discovered that the PM2.5 PM in Lahore were dramatically and consistently over the national and international health levels, with the mean of concentration of 112.4 ug/m3 and peak episode levels of 350ug/m3. Random Forest got the highest general performance in air quality classification (91.4 percent accuracy, 90.2 percent macro F1-score) and ANN had the lowest RMSE in LST prediction (1.84degC). The three most essential features in both tasks were determined as NDVI, PM2.5 lag and LULC category. The multi-source framework was significantly more effective than single-source solutions, proving that the integration of satellite and IoT data streams can be useful in monitoring the urban environment.

Recommendations

According to the study results, there are six recommendations put forward. To begin with, the Punjab Environmental Protection Agency ought to install an extended IoT sensors network on all the districts on Lahore aiming at at least 5 sensors/ 1 mm 2 to attain spatial representativeness of air quality management at the district level. Secondly, the Random Forest classification model that is created in this study must be operationalized as a near-real-time air quality monitoring and early

warning system that uses the 24-hour PM2.5 lag feature to introduce the 24-hour early warning of the Hazardous classification conditions. Third, NDVI-based monitoring needs to be incorporated into the urban master planning processes, so the Lahore Development Authority should set minimum green artifact quotas at the district level as an air quality co-benefit in addition to high urban greening objectives. Fourth, the Sentinel-5P TROPOMI data pipeline developed in the given work must be supported and further developed in order to offer an ongoing satellite-based monitoring of pollution without any ground infrastructure constraints. Fifth, geospatial data science and machine learning capacity building should be on the agenda of environmental protection agencies and urban planning organizations in Pakistan to potentially maintain and enhance the framework of monitoring internally. Sixth, subsequent studies must consider a broader range of Pakistani cities, such as Karachi, Faisalabad or Rawalpindi, modifying the feature set to reflect areas where pollution sources and land use patterns are the most significant and maintain the methodological architecture that is proven to be correct in this study.

References

1. Abhijith, K. V., Kumar, P., Gallagher, J., McNabola, A., Baldauf, R., Pilla, F., Broderick, B., Di Sabatino, S., & Pulvirenti, B. (2017). Air pollution abatement performances of green infrastructure in open road and built-up street canyon environments – A review. Atmospheric Environment, 162, 71–86. https://doi.org/10.1016/j.atmosenv.2017.05.014

2. Ali, G., Bashir, M. K., & Ali, J. (2021). Monitoring of land surface temperature and urban heat island in Lahore metropolitan area using geospatial techniques. Atmospheric Pollution Research, 12(3), 101–112. https://doi.org/10.1016/j.apr.2020.12.014

3. Begum, B. A., Kim, E., Jeong, C.-H., & Hopke, P. K. (2011). Evaluation of the potential source contribution function using the 2-D distribution of back trajectories and its application to air quality data. Atmospheric Environment, 39(21), 4005–4016. https://doi.org/10.1016/j.atmosenv.2005.03.020

4. Belgiu, M., & Dragut, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

6. Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., … Guo, Y. (2018). A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Science of the Total Environment, 636, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251

7. Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645–1660. https://doi.org/10.1016/j.future.2013.01.010

8. Gupta, P., & Christopher, S. A. (2009). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. Journal of Geophysical Research: Atmospheres, 114(D20), D20205. https://doi.org/10.1029/2008JD011497

9. Hamid, A., Zuberi, M. J. S., & Baig, S. (2020). Preliminary assessment of IoT-based low-cost air quality monitoring in Lahore, Pakistan. Environmental Monitoring and Assessment, 192(8), 1–14. https://doi.org/10.1007/s10661-020-08412-6

10. Imhoff, M. L., Zhang, P., Wolfe, R. E., & Bounoua, L. (2010). Remote sensing of the urban heat island effect across biomes in the continental USA. Remote Sensing of Environment, 114(3), 504–513. https://doi.org/10.1016/j.rse.2009.10.008

11. Iqbal, M. J., Amjad, M., Khan, A. R., & Khattak, M. S. (2021). Air quality assessment of Lahore for the period of 2015–2020: Trends, sources and health implications. Environmental Science and Pollution Research, 28(9), 11093–11107. https://doi.org/10.1007/s11356-020-11261-w

12. IQAir. (2023). World air quality report 2022. IQAir Foundation. https://www.iqair.com/world-air-quality-report

13. Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., … Britter, R. (2015). The rise of low-cost sensing for managing air pollution in cities. Environment International, 75, 199–205. https://doi.org/10.1016/j.envint.2014.11.019

14. Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

16. Lorente, A., Folkert Boersma, K., Eskes, H. J., Veefkind, J. P., van Geffen, J. H. G. M., de Zeeuw, M. B., … Coheur, P.-F. (2019). Quantification of nitrogen oxides emissions from build-up of pollution over Paris with TROPOMI. Scientific Reports, 9(1), 20033. https://doi.org/10.1038/s41598-019-56428-5

17. Masood, A., & Ahmad, K. (2020). A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. Journal of Cleaner Production, 322, 129012. https://doi.org/10.1016/j.jclepro.2021.129012

18. Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., … Jones, R. L. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmospheric Environment, 70, 186–203. https://doi.org/10.1016/j.atmosenv.2012.11.060

19. Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., … Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? Environment International, 116, 286–299. https://doi.org/10.1016/j.envint.2018.04.018

20. Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

21. Nowak, D. J., Hirabayashi, S., Bodine, A., & Greenfield, E. (2014). Tree and forest effects on air quality and human health in the United States. Environmental Pollution, 193, 119–129. https://doi.org/10.1016/j.envpol.2014.05.028

22. Pakistan Bureau of Statistics. (2023). Population and housing census 2023: Punjab district profile. Government of Pakistan.

23. Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. International Journal of Remote Sensing, 26(5), 1007–1011. https://doi.org/10.1080/01431160512331314083

24. Rahman, M. H., Sultana, N., Hossain, M. B., & Mondol, S. (2019). Spatial prediction of PM2.5 using random forest with satellite-derived AOD and meteorological variables in South Asian megacities. Science of the Total Environment, 689, 1241–1252. https://doi.org/10.1016/j.scitotenv.2019.06.366

25. Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Di Sabatino, S., Ratti, C., Yasar, A., & Rickerby, D. (2017). End-user perspective of low-cost sensors for outdoor air pollution monitoring. Science of the Total Environment, 607–608, 691–705. https://doi.org/10.1016/j.scitotenv.2017.06.266

26. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

27. Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., … Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment, 145, 154–172. https://doi.org/10.1016/j.rse.2014.02.001

28. Shafique, M., Xue, L., & Luo, X. (2018). Assessment of the changes in vegetation and urban heat island in major cities of Pakistan using Landsat data. International Journal of Environmental Research and Public Health, 15(11), 2516. https://doi.org/10.3390/ijerph15112516

29. van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. Environmental Health Perspectives, 118(6), 847–855. https://doi.org/10.1289/ehp.0901623

30. Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., … Levelt, P. F. (2012). TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. Remote Sensing of Environment, 120, 70–83. https://doi.org/10.1016/j.rse.2011.09.027

31. Weng, Q. (2009). Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends. ISPRS Journal of Photogrammetry and Remote Sensing, 64(4), 335–344. https://doi.org/10.1016/j.isprsjprs.2009.03.007

32. World Health Organization. (2022). Ambient (outdoor) air pollution. WHO. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

33. Yuan, F., & Bauer, M. E. (2007). Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. Remote Sensing of Environment, 106(3), 375–386. https://doi.org/10.1016/j.rse.2006.09.003

34. Zhang, Z., Jiang, H., Li, M., & Xu, F. (2019). Long short-term memory networks for air quality index prediction in smart cities. IEEE Access, 7, 107491–107502. https://doi.org/10.1109/ACCESS.2019.2930069

35. Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2019). Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307