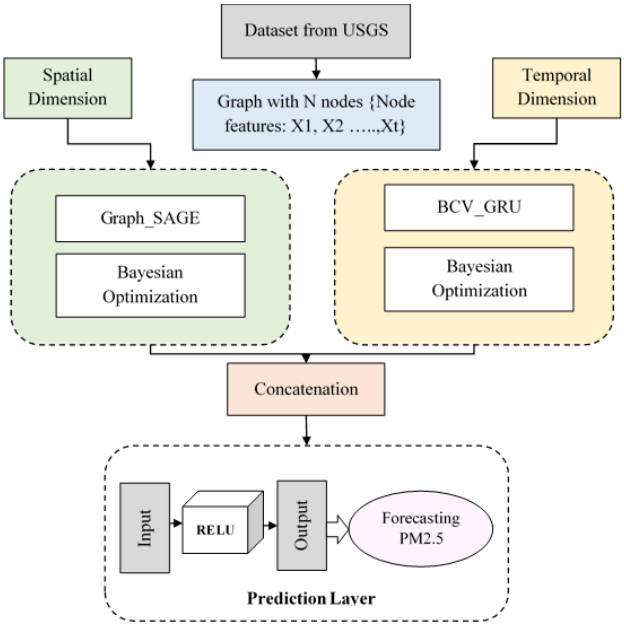
Global NEST Journal, Vol 26, No 2, 05598

Graph-based neural network model for predicting urban environmental air quality using spatio-temporal data optimization

Yogapriya J.[1]*, Deepa S.[2], Radha N.[3] and Madhumitha E.[4]
[1]Department of Computer Science and Engineering, Kongunadu College of Engineering and Technology, Trichy, Tamil Nadu, India
[2]Department of Physics, Government College of Engineering, Salem–636 011 Tamil Nadu, India
[3]Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering Chennai, Tamil Nadu, India
[4]Department of Artificial Intelligence and Data Science, Kongunadu College of Engineering and Technology, Trichy, Tamil Nadu, India
Received: 25/11/2023, Accepted: 01/01/2024, Available online: 13/01/2024
*to whom all correspondence should be addressed: e-mail: yogapriya.j@gmail.com
https://doi.org/10.30955/gnj.005598

Graphical abstract



Abstract

Environmental protection and the need for accurate pollutant forecasting have become increasingly important as worries about environmental issues and the harmful effects of pollution have grown. Predictive accuracy of air pollutants is generally unsatisfactory due to the fact that conventional methodologies prioritise time series analysis over the important spatial transmission dynamics among neighbouring locations. To address this inherent limitation, our proposed solution introduces an innovative Time Series Prediction Network, augmented by the auto-optimization capabilities of a Spatio-Temporal Graph-based Neural Network. This groundbreaking network comprises distinct spatial and temporal modules. The spatial module harnesses a Graph Sampling and Aggregation Network to extract essential spatial information from the data. Simultaneously, the temporal module integrates a Bayesian approach with a Complex Valued Graph Gated Recurrent Unit (BCV-GRU), seamlessly incorporating a graph network into the Gated Recurrent Unit (GRU) to capture temporal intricacies. Moreover, to manage the challenge of model inaccuracy stemming from inappropriate hyperparameters, Bayesian optimization was employed. The efficacy of our proposed method was validated using real PM2.5 data from the USGS website, showcasing a significant enhancement in prediction accuracy. This study puts forth a robust and effective approach for forecasting PM2.5 concentrations, bridging gaps in existing methodologies and contributing substantially to the evolution of environmental prediction models.

Keywords: Environmental protection, pollutant prediction, spatio-temporal data, graph-based neural network, bayesian optimization and PM2.5 forecasting

1. Introduction

Urban air pollution, an escalating issue rooted in urbanization and economic conditions, substantial risks towards the ecosystem, wellness of humans, including global warming (Zhao *et al.* 2020). The surge in human and industrial activities, coupled with ongoing fossil fuel usage, results in the emission of harmful air pollutions like NO_2s, O_3s, SO_2s, & COs. These pollutants substantially degrade air quality, impacting health and the environment.

Exposure to elevated NO_2s & SO_2s has a negative impact on the airways, particularly in young infants and older people. Short-term exposure to NO_2s can trigger asthma attacks, emergency room visits, and hospitalizations, while prolonged exposure can lead to respiratory infections (Zhuo *et al.* 2022) Additionally, NO_2 contributes to acid rain and SO_2 generates harmful tiny particles when interacting with other atmospheric elements, posing risks to human health and ecological balance (EPA, 2022a; EPA, 2022b). Elevated O_3 concentrations affect groups like children, the elderly, outdoor enthusiasts, and asthma patients. Research has connected repeated inhalation of O_3 to a number of health problems, such as chronic pulmonary inflammatory processes, difficulty inhaling

when outside, and irritation in the airways (Hu *et al.* 2023). Elevated CO levels deplete oxygen and have an impact on vital tissues such as the cerebral cortex and heart (EPAs, 2022s).

Notably, city air pollution exceeds WHO threshold limits (WHO, 2021), intensifying its impact on human health. Consequently, constant monitoring and predictive analysis of urban air quality, especially in emerging nations like Vietnam, are crucial. Emerging technology likes the IoTs & ICTs play pivotal roles. Finding a dependable solution to mitigate risks posed by poor air quality in urban areas is imperative. Forecasting the condition of the air is difficult, though. Although conventional techniques such as self-regressive techniques, average movement, and exponentially smoothed variables have been used (Chen *et al.* 2023), their efficacy is restricted because of the complex interplay among air pollution and weather variables. Thus, there's a growing need for more accurate, adaptable models that can continuously learn and update with new data

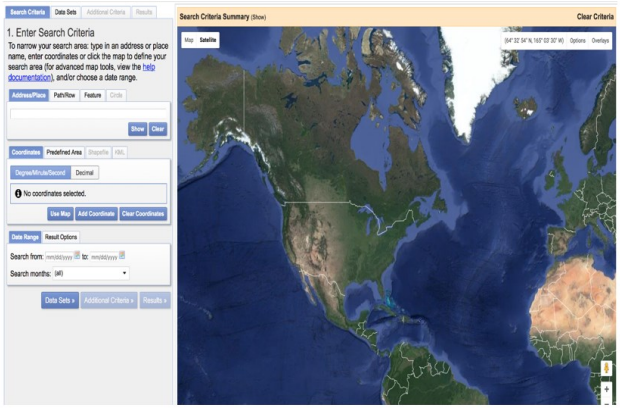


Figure 1. Monitoring stations observing air quality from USGS website (https:// earth explo rer. usgs. gov)

Several recent comprehensive reviews (Yin et al. 2023; Ban et al. 2023; Xiao et al. 2020; Zhang et a., 2021; Cheng et al. 2023; Luo et al. 2022; Zhu et al. 2022) highlight the utilization of various machine learning methods to check the pollutions. It constructed the individual methods; each pollutant requires substantial effort in deployment, maintenance, and monitoring.

While traditional time series forecasting methods like VARs, VARMAs, VAR-MAXs, SVARs utilized the multiple-series forecasted with a singularly techniques machine learning approaches have shown superiority over VAR-based methods (Li et al. 2023). Moreover, VAR-based approaches necessitate stationary period, it inadequate for non-stationary qualities of air data. This paper presents a global prediction technique that makes use of the N-BEATS in order to overcome these constraints. Comprising linked entirely interrelated levels with in both directions residual linkages, this framework seeks to forecast various air contaminants at the same time while accounting for connections underlying contaminants and variations in atmospheric scenarios. To understand temporal shifts and interrelationships across factors, geospatial and statistical correlation analyses were carried out. The key contributions of the proposed work is given below,

- Developing distinct spatial and temporal modules which incorporates a Graph Sampling and Aggregation Network for essential spatial information extraction.
- Temporal module integrates a Bayesian with Complex Valued Graph Gated Recurrent Unit (BCV-GRU) for capturing temporal intricacies.
- Employing Bayesian optimization to address the challenge of model inaccuracy arising from inappropriate hyperparameters enhances the robustness and reliability of the predictive model.
- Validating the proposed method using real PM2.5 data from the USGS website demonstrates a significant improvement in prediction accuracy compared to traditional methodologies.

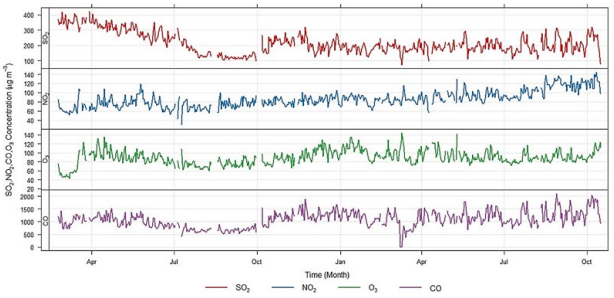


Figure 2. Air quality patterns on a daily basis (spanning February 2021 to August 2022)

2. Review of related works

2.1. Traditional model for predicting PM2.5

The conventional methods for predicting PM2.5 levels encompass statistical approaches and machine learning technologies. Statistical methods, due to their simplicity, focus primarily on understanding PM2.5 formation mechanisms and are extensively employed in air quality prediction. For example, (Shang and Luo, 2021) used ARIMAs with quantitative forecast to predictive daily and hourly levels of PM2.5 in Hong Kong. Using a period of twenty- and a generalized addition framework for forecasting, (Liu et al. 2023) investigated the relationship among PM2.5 and atmospheric conditions in Chengdu. However, these statistical methods have limitations due to the intricate nature of PM2.5 formation.

Machine learning techniques, leveraging historical data, address the nonlinearities in actual air pollution data, resulting in improved prediction accuracy. (Kamani et al. 2023) assessed hybrid models like ARIMAs, ANNs, SVMs, PCRs, DTs, & Catboosting, where Catboosting exhibited the most optimal evaluation. ARIMA-ANNs & DTs yielded satisfactory outcomes. (Palanisamy et al. 2023) created and assessed an SVMs approach to PM2.5 in crowded areas intricate topography, showcasing SVM's forecasting capabilities in areas that are comparable to tropical ones. used a MLPs to forecast and evaluate PM2.5 concentrations in eight major Chinese provincial cities,

emphasizing the need of lowering gaseous emissions in PM2.5 control.

Even the seemingly simple methodologies like machine learning and mathematical frameworks, however, are finding it difficult to meet the demands of large amounts of data due to the enormous rise in the amount of information. They tend to overfit and lack the sophisticated modeling capabilities needed for comprehensive analysis.

2.2. Top of form

2.2.1. Forecasting PM2.5 Using DL techniques

Deep learning has recently received a lot of attention for its skill in pollutants predictions because to its strong learning ability as well as capacity to represent nonlinear information. (Preethi and Asokan, 2020) used a hybrid model that combined two deep connections, CNN as well as LSTM. The CNNs extracted input data features, while the LSTM model accounted for pollutant dependencies over time. Their findings demonstrated enhanced prediction performance compared to traditional models. (Asokan and Preethi, 2021) utilized LSTM to forecast PM2.5 levels in Tehran, achieving an 80% explanatory capacity for PM2.5 variability. (Punarselvam et al. 2020) also suggested a CNN network method for analyzing previous sensor data and forecasting air pollution concentrations, which outperformed several sophisticated prediction mechanisms in this sector.

Despite the widespread application of these networks in air quality prediction, most fail to comprehensively capture data characteristics across both temporal and spatial dimensions. Several deep learning algorithms emphasize Euclidean space, however air quality observation points are typically not Geometrical.

2.3. Top of Form

2.3.1. Forecasting PM2.5 Utilizing GNNs Techniques

A technique known as GNNs uses a NNs to find, extracts, and interpret patterns and characteristics in graph-structured information. It may be used for a variety of graph teaching applications, such as differentiation, forecasting, grouping, and categorization. Typically, a graph having G = (V, E), where V = {V1, V2 . . . Vn} node, E = {E1, E2 . . . Em} edge set. GNNs excel in adapting to complex structures by delineating relationships between multiple concepts and delineating intricate nonlinear structures.

Compared to other neural network models, GNNs possess the capacity to model data characteristics from two facets: structure and function. Consequently, they offer enhanced versatility in modeling spatiotemporal data and extracting information.

In recent years, GNNs have found widespread application in air quality prediction due to these characteristics. (Kulurkar et al. 2023) presented a technique for predicting air pollution concentrations in Japan and China using MASTGNs. A PM2.5-GNN model that is skilled at detecting dependency over time was presented by (Punarselvam et al. 2021). integrated the PM2.5 distribution partial differential equation with the DPGN framework in order to combine interpretability and information extraction. A thorough forecasting technique centered on spatial dispersion among nearby locations was presented by (Preethi and Asokan, 2020; Tong et al. 2018). They fail to effectively combine GNNs' spatial information capture with recurrent neural networks' time series processing advantages.

Hence, traditional and deep learning-based prediction methods encounter challenges in adequately extracting data features across spatial and temporal dimensions. This article integrates spatiotemporal networking controlled recurrent units to present a unique time period forecasting system. This combination includes the gradient problem-solving ability of recurrent neural networks that are used in historical prediction, the geographical data extracting ability of GNNs, and the powerful hyperparameter optimization abilities of Probabilistic techniques.

3. Top of Form

3.1. Proposed BCV-GRU framework

BCV-GRU comprises two modules: BGraphSAGEs & BGraphGRUs collaboratively leveraging spatio-temporal relationships within the data. As shown in Figure 1, BGGRU starts with the dimension of space component and makes use of BGraphSAGE to acquire geographical characteristics form the supplied information, create nodes embedded data, and combine regional characteristics. The time dimension modules is covered in the next section. BGraphGRU is used to represent spatio-temporal relationships while maintaining spatial features. In the end, the model uses two completely linked regions to forecast predicted PM2.5 concentrations. The model uses the Bayesian approach to adjust hyper parameter settings for each component and the entire model in order to maximize effectiveness. Additionally, Figure 2 depicts the learning process, detailing model training and testing stages. During training, the model preprocesses training and validation datasets, initializes network parameters randomly, and generates the graph's adjacency matrix based on geographic information. The BGGRU network extracts spatio-temporal features and predicts future PM2.5 concentrations, adjusting network weights based on the calculated prediction error until reaching the specified number of epochs. During testing, the trained BGGRU processes test data to predict PM2.5 concentrations, evaluated using a designated evaluation function.

The model incorporates a skip connection between the spatial and temporal modules. In deep neural networks, deep layers often face challenges like overfitting and gradient issues, impeding parameter updates in shallow layers. The skip connection involves transmitting information from the current layer not only directly but also after nonlinear transformation, addressing overfitting and gradient issues to some extent. The model's bypass connection method is shown in Figures 3 as well as 4, where results from both the first BGraphSAGE &

BCV_GRU are received by the last entirely linked layer. This skip connection accelerates model stability and mitigates overfitting by incorporating information from different layers.

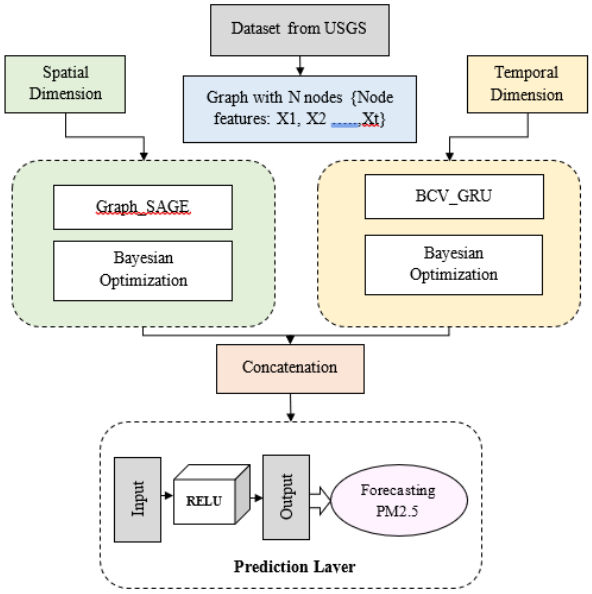


Figure 3. Flow of the proposed framework

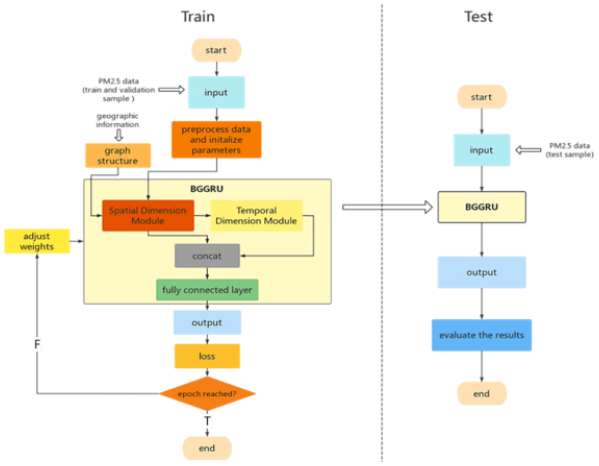


Figure 4. Flow chart for BGGRUs

3.2. The spatial aspect of BGGRU (BGraphSAGE) is its Spatial Dimension Module

Conventional graph embedding methods, relying on matrix decomposition and random walk, necessitate utilizing all node information iteratively to acquire vector representations. These methods inherently operate in a transductive manner. GraphSAGE introduces a technique involving sampling and aggregation. Initially, it uses node connections to sample their neighboring nodes, followed by the continuous fusion of adjacent node information using a multi-layer aggregation function. GraphSAGE's organizational structure is seen in Figure. 5. This paper's method uses GraphSAGE within an analogous manner to combine local node characteristics, model every node in the network, & retrieve their properties from the collected information.

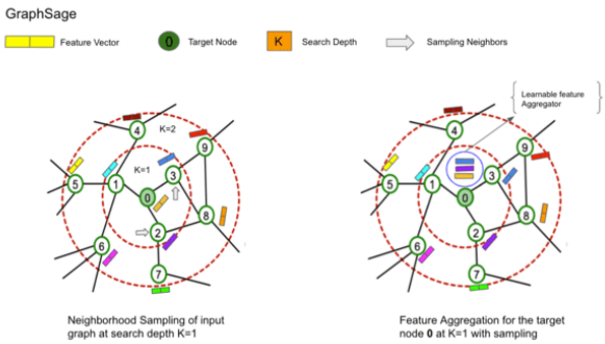


Figure 5. A graphical depiction showcasing the methodology of GraphSAGE sampling and aggregation

The graph, denoted as ς = (υ, ε), consists of a set of nodes (υ) and edges (ε). Each node collects surrounding embeddings of nodes in every repetition K by averaged such vectors. After that, the current embedded vectors and the aggregation vectors merge, and the output is computed using a linear layer of data with an activated sigmoid.

3.3. The temporal aspect of BGGRU, known as BGraphGRU, focuses on handling time-related dependencies within the BGGRU model

The GRU, a variant of LSTM, addresses gradient issues in long-term memory & back propagation. It is often used for extracting behavioral characteristics from historical information and frequently performs similarly to LSTM, however with less complicated implementation and computations. The suggested BGraphGRU preserves the same chain structure as GRU, but instead of using linear transformations to extract details of structure from snapshots at every step, it uses GraphSAGEs. With this change, BGraphGRU can now manage LSTMs dependencies & efficiently learning their temporal properties of incoming graphs, simulating spatio-temporal connections while maintaining a geographic framework.

The GRU introduces reset and update gates to alter how hidden states are calculated in recurrent neural networks. BGraphGRU follows a similar chain structure to GRU but uses GraphSAGE instead of the original data. The reset gate, akin to GRU, influences the hidden state, determining how much past information should be disregarded Here, at represents BGraphGRU's input at time t−1, Ht−1 represents the undetected state with moment t-1, and Wr & br represent the reset gate's weighted & bias vectors.

BGraphGRU's updating gates combines LSTM's memory and inputs gating. It influences both the current and previous hidden units, determining how much valuable information should be passed down. Its formula is:

$$r_t = \sigma\left(W_r A_t + GraphSAGE_r^k(h_{t-1}, A_{t-1}\right) + b_r \quad (1)$$

At time t, $A_t \in R^{N\times N}$ represents the input for GraphGRU, $h-1 \in ht-1 \in R^{N\times d}$ stands for the hidden state at time t - 1, $Wz \in R^{N\times d}$ and $bz \in R^{d}$ denote the weight and bias matrices respectively for the update gate.

$$Z_t = \sigma\left(W_z A_t + GraphSAGE_r^k(h_{t-1}, A_{t-1}\right) + b_z \quad (2)$$

following this, BCV-GRU proceeds to compute the calculation for subsequent hidden state .

$$\check{h}_t = tanh\left(W_{hA} A_t + W_{hh}(r_t * GraphSAGE_h^k\left(h_{t-1}\right)\right) + b_h) \quad (3)$$

Given the formula previously, the gate that resets controls the transition from the prior time step's concealed status to the current time step's prospective concealed state. The concealed state at the final time step may include all previous data in the time sequence up to that point. As such, past data that may not be relevant to the forecast may be eliminated via the reset gate. Finally, calculating the hidden state hht at time step t entails utilizing the current timed step's updates gateway zt and combining the prior time step's concealed state h1ht1 with the candidate concealed version heht at the current time step.

$$h_t = \left(1 - 2t\right) * GraphSQAE_h^k\left((h_{t-1}\right) + z_t * \check{h}_t \quad (4)$$

3.4. Optimization based on bayesian hyperparameter

Although the GNNs technique is a popular tool for predicting air quality, choosing a model's the hyperparameters is generally based on experience, much as when using different machine learning techniques. Nevertheless, this methodology not only necessitates significant time and resource investment but also bears no assurance about the achievement of ideal hyperparameters.

Therefore, the first crucial step in solving these problems is selecting an approach that works well for determining the ideal collection of parameters for the model. Bayesian optimization for black box functional situations has become the standard hyperparameter calculation method in recent years. The aim function of Bayesian optimization must respect local smoothing requirements such as Lipschitz or equal continuous in order for it to work globally. By using an acquisition functional for efficient discovery and usage, Bayesian optimization may approximate complicated function objectives with faster evaluation times. Thus, we made use of the Bayesian technique to maximize the model's parameters, guaranteeing the accuracy of the model used to predict.

Step 1 the hyperparameters losses the product where it is optimised to make a meaningful connection.

$$\rho^* = argminloss\left(p\right)p\varsigma p \quad (5)$$

Top of Form

The ideal set of hyperparameters identified by the Probabilistic a hyperparameter optimization approach is denoted by the symbol p∗ in this formula. The input to the the hyperparameter set is denoted by p, the complete set of the hyperparameters is represented by P, and the function of goal that has to be optimized is indicated by loss(·). Within our model, the hyperparameters slated for optimization encompassed the number of training epochs (*h*num_epoch), the overall model's learning rate (*learning_rate*), the count of BGraphSAGE layers (*num_layers* of BGraphSAGE), the edges per node (*edges_per_node*) in the graph, and the count of BGraphGRU layers (*num_layers* of BGraphGRU) the mean absolute error, outlined in the following formula:

$$loss\left(p_j\right) = \frac{1}{n}\sum_{j=1}^{n}\left|\check{y}_t\ \left(p_j - y_i\right)\right| \quad (5)$$

Within this equation, *pj* denotes the j-th hyper-parameter, *y* represents the actual range, and *y*^(*pj*) signifies their output utilizing the hype-rparameter *pj*. The subsequent phase of the Bayesian approach involves constructing a dataset*D*={(*x*1,*y*1),(*x*2,*y*2),…,(*xi*,*yi*)}, where *xi* denotes the i-th hyper-parameter , and *yi* represents the fault associated with the output derived from the particular set of hyper-parameters. The Bayesian technique then moves on by estimating the function's distributions and deriving a substitution model M using a restricted amount of observations data. This model adheres to a Gaussian distribution G characterized by a variance denoted ask and a mean denoted as μ. The posterior probabilities p(xi ,D) is derived from the datasetD.

$$\rho\left(loss\right) = G\left(loss; \mu, k\right) \quad (7)$$

$$\rho\left(loss|D\right) = G\left(loss; \mu_{loss|D,}\ k_{loss|D}\right) \quad (8)$$

The function responsible for establishing the guidelines for the subsequent observation point is termed the acquisition function *a*(*p*).

$$\rho^* = argmaxa\left(P, p\left(y|x\right)\right) \quad (9)$$

The primary aim of Bayesian hyperparameter optimization involves iteratively objective function *loss*(·) by gradually increasing observation points, guided by the acquisition function *a*(*p*). Finding the lowest threshold for the target variable loss (·) is the primary objective for this process. Thus, the previously indicated procedures are repeated until the optimum amount of repetitions is achieved and the ideal parameters are found **Investigational results**

It's critical to determine which characteristics or parameters have the most predictive power over the Radio Frequency once it has been trained. High relevance % factors are important predictors of the algorithm's result, having a big influence on the algorithm's results. Analyzing and quantifying variable importance aids in selecting pertinent features for the model.

For this study, different sets of variables, specifically 10 variables, were Variable relevance varied significantly among various air quality classes—good, mild, unhealthy delicate, and danger. The distance away from the road, LSTs and SAVIs were found to be the most important factors among the ten variable inputs in the present model. Band 4 was eliminated since its significance did not alter, suggesting that it had little bearing on the entire significance of the model.

4.1. Top of Form

4.2. Variables of Independent LSTs

Usually, when solar radiation and solar power cross paths with the earth, the earth warms. LSTs thereby measures

the thermal radiation that the land area emits. The LST provides an average annual depiction of land surface temperatures (refer to Figure. 6) measured in degrees Celsius via spectroradiometer imaging (LandSat 8). The research region has a prevailing sequence of mild to high surface-level temperatures, with LST ranging from 27.68 to 46.44 degrees Centigrade. LST may be calculated analytically with the use of the following formula (Eq. 10):

$$LST = \left(\frac{BT}{1 + \left(0.00115 * \frac{BT}{1.4388} \right) * Ln(\in)} \right) \quad (10)$$

4.2.1. Band of landsats

Images with nine different spectral bands were captured by the Landsat (8) OLIs and TIRSs. Organizations 1 though seven and nine have a field of view of thirty meters. Over a span of five years, five Landsat (8) bands (B2-B6) were obtained at no cost through the http://earthexplorer.usgs.gov. In the R these spectrum underwent correction & calibration procedures to prepare them for subsequent utilization in index estimation

4.2.2. SAVIs

In regions where vegetation cover is sparse, the Soil-Adjusted Vegetation Index (SAVI) serves to mitigate the impact of soil brightness on the NDVIs. Within these specific areas, SAVI values typically span between -0.77 and 0.69. SAVI's mathematical calculation is derived from the following equation (Eq. 11), where 'L' represents the soil brightness correction factor, typically assigned a value of 0.5 for various land cover types.

$$SAVI = \frac{(NIR - R)}{(NIR + R + L)} * (1 + L) \quad (11)$$

4.2.3. Top of Form

BU indices: By deducting the NDVI (Eq. 12) from an NDBs index (Eq. 11), the BUs indicator (Eq. 3) is derived. The NDBs list, also known as the normalization differential built-up directory, may be determined with the equations (Eq. 12) The Built-up index varies from -0.69 to 0.52 within the study area, indicating an average to high dispersal tendency over the region, as shown in Figure 6.

$$BU = NDVI - NIR \quad (12)$$

$$NDB = \frac{SWIR - NIR}{SWIR + NIR} \quad (13)$$

4.3. Proximity from the roads

The region's primary road network was taken from OpenStreetMap (www.openstreetmap.org). The resulting vector layers depicting the roadways have been processed in the ArcMap program to generate an abstract level that shows the separation from the road network (see Figure. 6). The Euclidean spacing function in ArcMap estimated distances between pollutants in the air detection sites and measurement locations (Figure. 6).

4.3.1. NDVIs

As per equation (Eq. 5), the NDVIs is computed, with its values consistently falling within the range of -1 to +1. While NDVI scores close to 0 are often linked to urban environments, a value of +1 indicates a significant probability of thick vegetation. Generally speaking, green vegetation is more reflective in the NIRs with green light frequencies than in other wavelength ranges. The research site's mean NDVIs score, which varies from -0.25 to +0.30, shows a lack of dense greenery in this region.

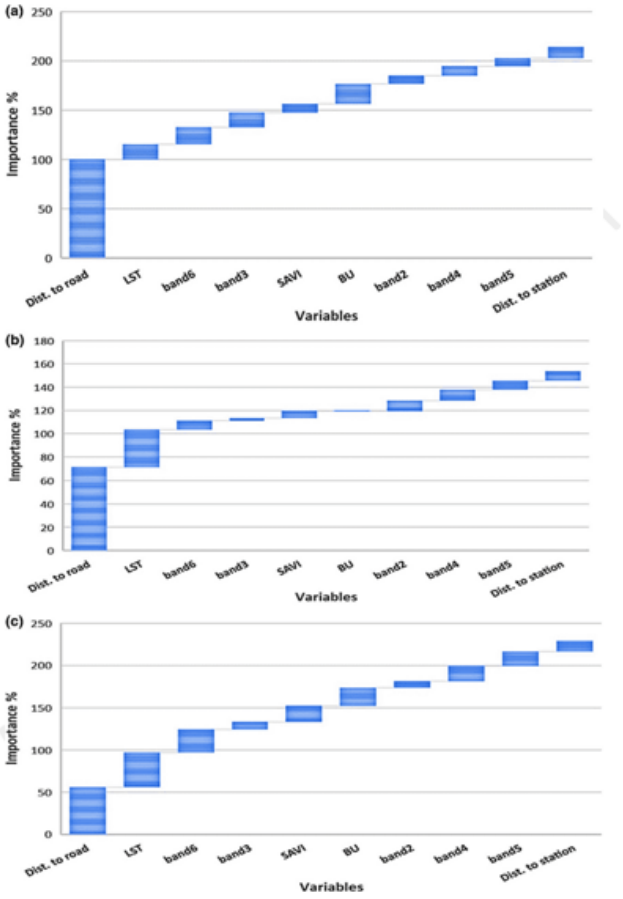


Figure 6. The significance of parameters for several categories: 'a' stands for excellent, 'b' for moderate, and 'c' for hazardous

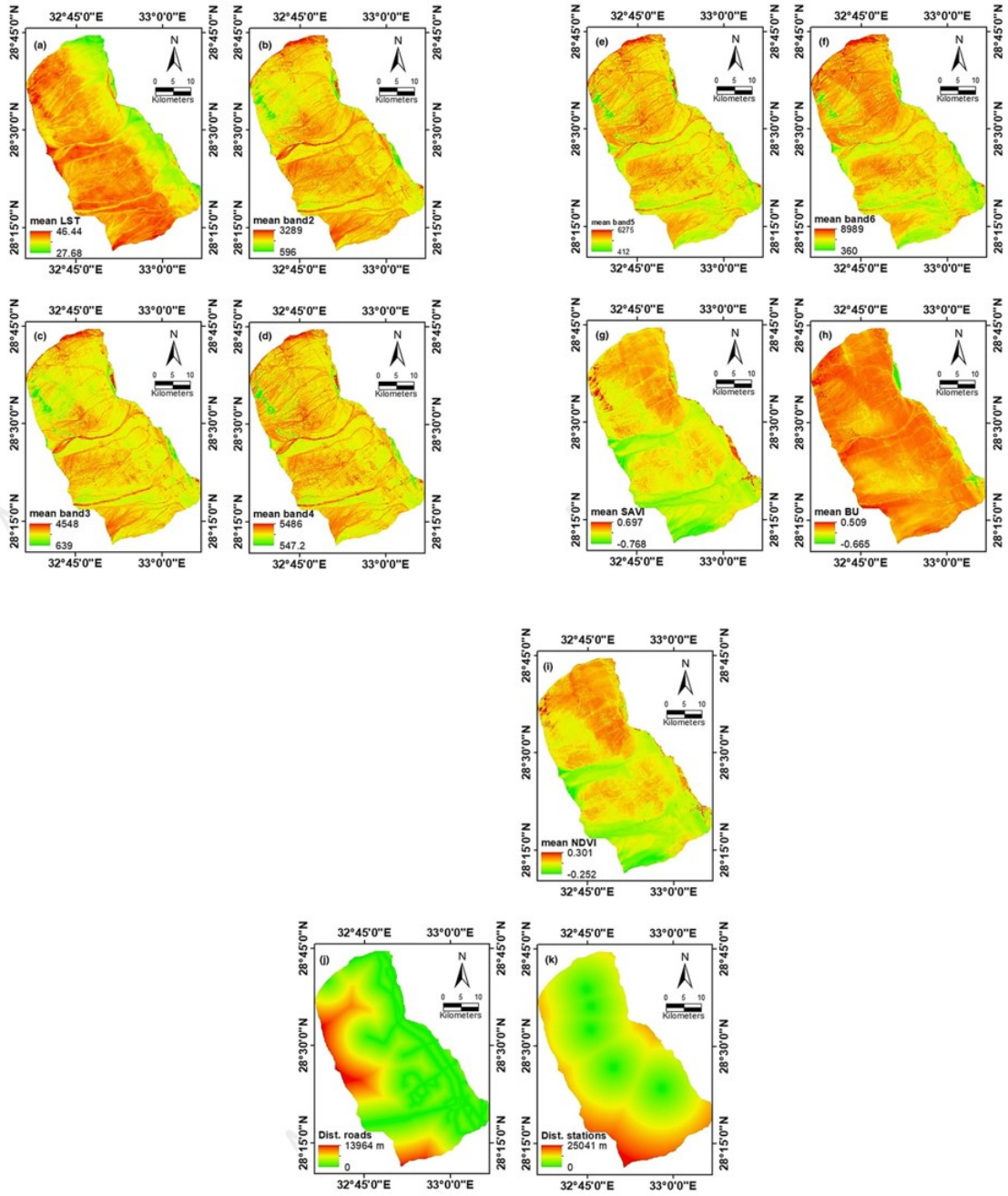


Figure 7. Mean LST (a), mean spectrum 2 (b), mean spectrum 3 (c), mean spectrum 4 (d), mean spectrum 5 (e), mean spectrum 6 (f), mean SAVI (g), average BU (h), average NDVI (i), proximity to highways (j), and proximity from stations (k) are the factors that affect the mapping of pollutants in the air

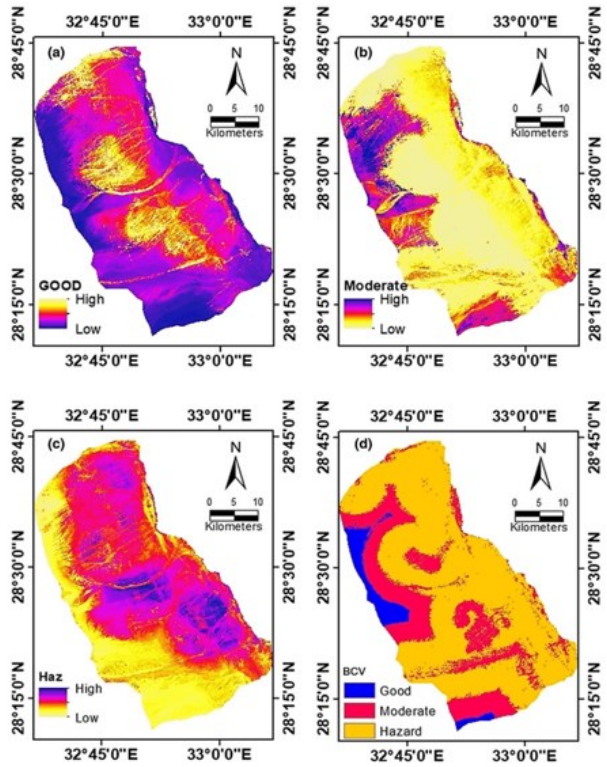


Figure 8. Four different classes of air quality have predictive models: the danger class (d), the excellent class (a), the intermediate class (b), and the random forest class (d)

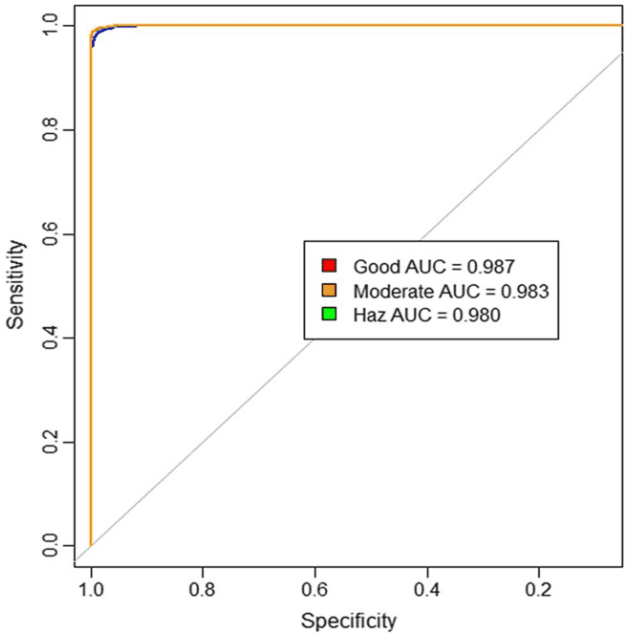


Figure 9. Validity structure for a database

4.4. Forecasting Air pollution

PM10 data spanning four years (2018–2021) were collected from 5 region situated in the North Ras Gharib region for analysis. The collected spatial dataset underwent preparation, stacking, and normalization to establish separate datasets for model training and testing. In order to guarantee uniform random distributions throughout every algorithm's procedure loops and avoid optimization breakdown, it was imperative to provide a seed value.

Among the stations, the East Arta station fell into the "good" air quality class, exhibiting average PM10 measurements within the clear and good air limits outlined by the EPA in 1997. Three sites, however—Hana feld, Hoshia feld, and Northwest Gharib feld—were placed in the "average" category as they PM10 analysis findings remained in line with the EPA's moderate air pollution criteria (1997). However, the Arta feld station was classified as "hazard" class, with its highest recorded PM10 value exceeding the EPA-recommended threshold of 424 µgm−3 (1997).

3 PM10 index classes—good, reasonable, with hazardous—and 10 variables were incorporated to create a BCVs framework (Figure. 7). Differential sensitivity patterns to all PM10 categories were identified across the Landsat 8 spectrum bands (2, 3, 4, 5, and 6) lacking considerable predictive ability. In the model, bands 2 and 5 showed very little reaction, but band 6 showed significant sensitivity in the majority of the predicted classes. Additionally, the thematic layer depicting distance from roads and LSTs demonstrated heightened sensitivity within the prediction model. Spatially, lower values were concentrated in the western parts across all classes (Figure. 8), suggesting a similarity in prediction model performance across different regions**Table 1.** Overview of the percentage distribution of prediction subclasses in the RG area.

Category	Area (km^2)	%
GOOD	86	6
MODERATE	437	28
HAZARD	1096	69

Table 2. Data on the accuracy & Kappa scores of the predictive techniques

Machine Learning Techniques	Details of ACC		SD with ACC
	ACC	Kappa	
RFs	99.30	97.56	0.0067

The land cover indexes, specifically SAVIs & BUs, exhibit diverse patterns within the predictive subcategory. Notably, SAVIs shows comparatively lower responsiveness across these subclasses among the various land cover indices. Meanwhile, the proximity of stations within petroleum activity areas demonstrates a pronounced correlation with the intensity of PM10.

The predictive model outcomes pertaining to PM10 classes highlight that a substantial portion of the area falls within the hazard class (Table 1). Notably, the PM10 values from the Arta station significantly impact the surrounding area according to the predictive model.

Although fine-tuning the modeling hyperparameters has marginally improved the results, enhancing the model's accuracy, the final predictive outcomes align with measurement expectations. Long-term uncertainty may, however, exacerbate the situation, making it necessary to precisely classify land cover, land use, industrial

development, and other distinct variables in order to make reliable projections.

4.5. ROC evaluation

The model's is most appropriate is usually measured by AUC ROC result reliability. Al does, however, issue a warning: a good showing on data that has been seen does not always imply a comparable showing on data that has not. Many researchers, such as have employed the ROC approach to assess air quality pollution model performance. Consequently, ROC is frequently considered the primary accuracy measure in much of the literature.

The evaluation of model performance utilizing ROC (Figure. 9) yielded an AUC of 99.30 over the four-year data measurement period. Despite this, while the model accuracy was high, the Kappa statistic showed reasonably good performance.

4.6. Monitoring process and prospective research areas

The absence of a comprehensive monitoring plan coupled with negligence toward environmental sustainability aggravates air pollution issues in the area, leading to extreme emissions from certain stations in this study. Achieving this necessitates ensuring the availability and reliability of air quality information across all levels without compromising pollution control efforts. There's a pressing need to enhance current air pollution monitoring capabilities by maintaining a comprehensive historical dataset.

Integrating forecasting methods with real-time data could significantly enhance risk management strategies. Moreover, deploying low-cost air quality sensors can substantially improve data collection quality, thereby enhancing data aggregation and management (Table 2).

Subsequent investigations in this field need to include a multitude of topics, such as the effects of air quality on the nearby ecology, detailed analyses of distinct contaminants, and a thorough discussion of any potential health hazards linked to air quality pollution. Beyond just disseminating recorded pollutant concentrations, communicating the health hazards caused by air quality contamination is critical.

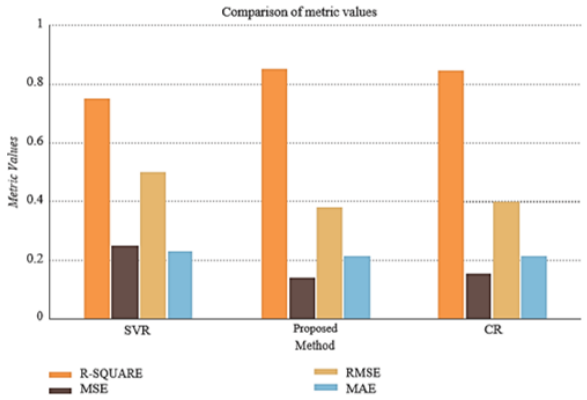


Figure 10. Comparative results of SVR, CR and proposed model

In Figure 10, a comprehensive comparison of key metrics—R-SQUARE, MSE, RMSE, and MAE—across support vector regression, proposed model, and CatBoost regression is presented, specifically for an imbalanced dataset where the optimization algorithm has not been utilized. This analysis provides valuable insights into the performance of these regression models under such challenging conditions. Remarkably, the proposed model emerges as the standout performer, boasting the highest R-SQUARE at an impressive 84 percent. This signifies the model's exceptional ability to explain the variance in the data, indicating a robust predictive capability. Notably, a high R-SQUARE is indicative of the model's effectiveness in capturing the underlying patterns and trends in the imbalanced dataset.

Equally noteworthy are the remarkably low error values associated with the proposed model. The lowest RMSE of 0.18 underscores the model's precision in predicting values, as it represents the square root of the average squared differences between predicted and observed values. Likewise, the MSE value of 0.28 and the MAE value of 0.21 further attest to the superior predictive accuracy of the proposed model. These findings collectively highlight the efficacy of the proposed regression model in handling imbalanced datasets without resorting to the optimization algorithm. The impressive R-SQUARE and minimal error metrics underscore its potential for practical applications where accurate and reliable predictions are paramount. This study not only contributes valuable insights into regression model performance but also emphasizes the significance of considering imbalanced datasets in real-world scenarios.

5. Conclusion

In this research, the scattering of PM10 in the North Ras Gharib region of Egypt and its environs was examined using the RF method. This method performs well when it comes to optimizing and displaying time-series information. The researchers wanted to look at the link between PM10 and Landsat 8 OLI spectral bands, as well as land use indices. A time series dataset that was gathered from five measurement sites in the research region and covered the years 2018 to 2021 was used. The predicted algorithm included ten factors, which included bands 2, 3, 4, 5, 6, LST, SAVI, BU, proximity to locations, and proximity from roadways.

The research region has a high vulnerability or hazard level in relation to PM10 pollution, according to results provided by the RF model. The prediction susceptibility was significantly impacted by variables like land surface temperature and proximity from highways. According to the distribution map, Arta field's high PM10 concentration considerably impairs air quality. Landsat's bands 4 and 5 had little effect on PM10 sdispersion. The primary distributors of PM10 were found to be isolated locations with high levels of oil extraction and activity.

Even with the small number of PM10 measurement sites, the model's accuracy was rather good. Future work might concentrate on increasing the number of data gathering locations around the region and adding other variables, such as wind velocity and acceleration, to improve the forecasting model.

References

Asokan R. and Preethi P. (2021). Deep learning with conceptual view in meta data for content categorization. In Deep Learning *Applications and Intelligent Decision Making in Engineering* (176–191). IGI Global.

Ban Y., Liu X., Yin Z., Li X., Yin L. and Zheng W. (2023). Effect of urbanization on aerosol optical depth over Beijing: Land use and surface temperature analysis. *Urban Climate*, **51**, 101655. doi: https://doi.org/10.1016/j.uclim.2023.101655

Chen J., Liu Z., Yin Z., Liu X., Li X., Yin L. and Zheng W. (2023). Predict the effect of meteorological factors on haze using BP neural network. *Urban Climate*, **51**, 101630. doi: https://doi.org/10.1016/j.uclim.2023.101630

Cheng Y., Lan S., Fan X., Tjahjadi T., Jin S., and Cao L. (2023). A dual-branch weakly supervised learning based network for accurate mapping of woody vegetation from remote sensing images, *International Journal of Applied Earth Observation and Geoinformation*, **124**, 103499. doi: https://doi.org/10.1016/j.jag.2023.103499

Hu F., Qiu L., Xiang Y., Wei S., Sun H., Hu H. and Zeng M. (2023). Spatial network and driving factors of low-carbon patent applications in China from a public health perspective. *Frontiers in Public Health*. doi: 10.3389/fpubh.2023.1121860

Kamani H., Baniasadi M., Abdipour H., Mohammadi L., Rayegannakhost S., Moein H. and Azari A. (2023). Health risk assessment of BTEX compounds (benzene, toluene, ethylbenzene and xylene) in different indoor air using Monte Carlo simulation in zahedan city, Iran. Heliyon, 9.

Kulkarni G.E., Muley A.A., Deshmukh N.K. and Bhalchandra P.U. (2018). Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra India. *Model. Earth Syst. Environm*. **4** (4), 1435–1444. https://doi. org/10.1007/s40808-018-0493-2.

Kulurkar P., kumar Dixit C., Bharathi V.C., Monikavishnuvarthini A., Dhakne A. and Preethi P. (2023). AI based elderly fall prediction system using wearable sensors: A smart home-care technology with IOT. Measurement: *Sensors*, **25**, 100614.

Li W., Wang C., Liu H., Wang W., Sun R., Li M., and Fu S. (2023). Fine root biomass and morphology in a temperate forest are influenced more by canopy water addition than by canopy nitrogen addition. *Frontiers in Ecology and Evolution*, **11**. doi: 10.3389/fevo.2023.1132248

Liu Z., Feng J. and Uden L. (2023). Technology opportunity analysis using hierarchical semantic networks and dual link prediction. *Technovation* **128**, 102872. https://doi.org/10.1016/j.technovation.2023.102872

Luo, J., Niu, F., Lin, Z., Liu, M., Yin, G. and Gao Z. (2022). Abrupt increase in thermokarst lakes on the central Tibetan Plateau over the last 50 years. CATENA, 217, 106497. doi: https://doi.org/10.1016/j.catena.2022.106497

Palanisamy P., Padmanabhan A., Ramasamy A. and Subramaniam S. (2023). Remote Patient Activity Monitoring System by Integrating IoT Sensors and Artificial Intelligence Techniques. *Sensors*, **23**(13), 5869.

Preethi P. and Asokan R. (2020). Neural network oriented roni prediction for embedding process with hex code encryption in dicom images. In Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India (18–19).

Preethi P. and Asokan R. (2020). Neural network oriented roni prediction for embedding process with hex code encryption in dicom images. In Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India (18–19).

Punarselvam E., Devi T.K., Britto A., Prakash N.B. and Suresh P. (2020). Segmentation Analysis Techniques and Identifying Stress Ratio of Human Lumbar Spine Using ANSYS, *Journal of Medical Imaging and Health Informatics*, **10**(10), 2308–2315.

Punarselvam E., Sikkandar M.Y., Bakouri M., Prakash N.B., Jayasankar T. and Sudhakar S. (2023). Retraction Note to: Different loading condition and angle measurement of human lumbar spine MRI image using ANSYS.

Punarselvam E., Sikkandar M.Y., Bakouri M., Prakash N.B., Jayasankar T. and Sudhakar S. (2021). Different loading condition and angle measurement of human lumbar spine MRI image using ANSYS. *Journal of Ambient Intelligence and Humanized Computing*, **12**, 4991–5004.

Shang M. and Luo J. (2021). The Tapio Decoupling Principle and Key Strategies for Changing Factors of Chinese Urban Carbon Footprint Based on Cloud Computing. *International Journal of Environmental Research and Public Health*, **18**(4), 2101. doi: 10.3390/ijerph18042101

Tong L., Alexis K.H. and Lau K.S. (2018). Time Series Forecasting of Air Quality Based on Regional Numerical Modeling in Hong Kong. *J. Geophys. Res. Atmos*, **123**, 4175–4196.

Xiao Y., Zuo X., Huang J., Konak A. and Xu Y. (2020). The continuous pollution routing problem. *Applied Mathematics and Computation*, **387**, 125072. doi: https://doi.org/10.1016/j.amc.2020.125072

Yin Z., Liu Z., Liu X., Zheng W. and Yin L. (2023). Urban heat islands and their effects on thermal comfort in the US: New York and New Jersey. *Ecological Indicators*, **154**, 110765. doi: https://doi.org/10.1016/j.ecolind.2023.110765

Zhang S., Bai X., Zhao C., Tan Q., Luo G., Wang J. and Xi H. (2021). Global CO2 Consumption by Silicate Rock Chemical Weathering: Its Past and Future. *Earth's Future*, **9**(5), e1938E-e2020E. doi: https://doi.org/10.1029/2020EF001938

Zhao M., Zhou Y., Li X., Cheng W., Zhou C., Ma T. and Huang K. (2020). Mapping urban dynamics (1992–2018) in Southeast Asia using consistent nighttime light data from DMSP and VIIRS. *Remote Sensing of Environment*, **248**, 111980. doi: https://doi.org/10.1016/j.rse.2020.111980

Zhu W., Chen J., Sun Q., Li Z., Tan W., and Wei Y. (2022). Reconstructing of High-Spatial-Resolution Three-Dimensional Electron Density by Ingesting SAR-Derived VTEC Into IRI Model. *IEEE Geoscience and Remote Sensing Letters*, **19**. doi: 10.1109/LGRS.2022.3178242

Zhuo Z., Du L., Lu X., Chen J. and Cao Z. (2022). Smoothed Lv Distribution Based Three-Dimensional Imaging for Spinning Space Debris. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1–13. doi: 10.1109/TGRS.2022.3174677