Journal of Hazardous Materials 474 (2024) 134666

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Research Article

What is the spatiotemporal pattern of benzene concentration spread over susceptible area surrounding the Hartman Park community, Houston, Texas?

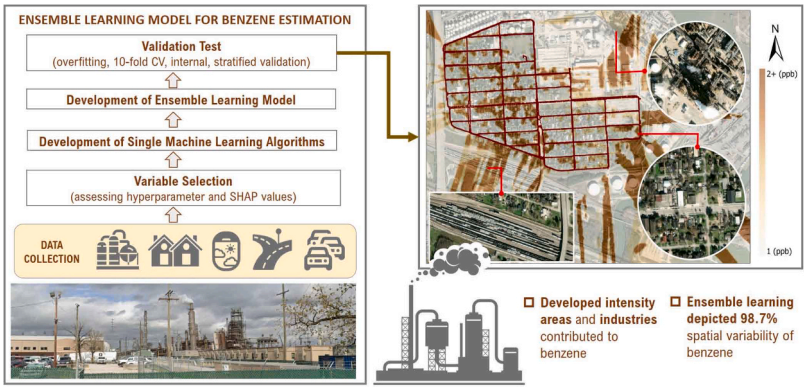
Aji Kusumaning Asri [a], Galen D. Newman [b], Zhihan Tao [b], Rui Zhu [b], Hsiu-Ling Chen [c], Shih-Chun Candice Lung [d,e,f], Chih-Da Wu [a,g,h,i,*]

[a] Department of Geomatics, College of Engineering, National Cheng Kung University, Tainan 701, Taiwan, ROC
[b] Department of Landscape Architecture and Urban Planning, School of Architecture Texas A&M University, 3137 TAMU, College Station, TX 77843, USA
[c] Department of Food Safety Hygiene and Risk Management, National Cheng Kung University, Tainan 701, Taiwan, ROC
[d] Research Center for Environmental Changes, Academia Sinica, Taipei, Taiwan, ROC
[e] Department of Atmospheric Sciences, National Taiwan University, Taipei, Taiwan, ROC
[f] Institute of Environmental Health, School of Public Health, National Taiwan University, Taipei, Taiwan, ROC
[g] National Institute of Environmental Health Sciences, National Health Research Institutes, Miaoli, Taiwan, ROC
[h] Innovation and Development Center of Sustainable Agriculture, National Chung Hsing University, Taichung City 402, Taiwan, ROC
[i] Research Center for Precision Environmental Medicine, Kaohsiung Medical University, Kaohsiung 804, Taiwan, ROC

HIGHLIGHTS

- Estimating spatiotemporal patterns of benzene in environmentally disadvantaged area.
- Utilizing ensemble learning model to accurately estimate benzene concentration.
- Ensemble model depicted 98.7% benzene spatial variability in Hartman Park community.
- Intensively developed areas and industry-related factors contribute to benzene.
- High benzene levels were spread over industrial area to residential area nearby.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Benzene concentration
Ensemble learning model
Hazardous pollution zone
Vulnerable community
Machine-learning algorithms

ABSTRACT

The Hartman Park community in Houston, Texas-USA, is in a highly polluted area which poses significant risks to its predominantly Hispanic and lower-income residents. Surrounded by dense clustering of industrial facilities compounds health and safety hazards, exacerbating environmental and social inequalities. Such conditions emphasize the urgent need for environmental measures that focus on investigating ambient air quality. This study estimated benzene, one of the most reported pollutants in Hartman Park, using machine learning-based approaches. Benzene data was collected in residential areas in the neighborhood and analyzed using a

* Correspondence to: Department of Geomatics, National Cheng Kung University, No. 1 University Road, Tainan 701, Taiwan, ROC.
E-mail addresses: akusumaning@gmail.com (A.K. Asri), chidawu@mail.ncku.edu.tw (C.-D. Wu).

https://doi.org/10.1016/j.jhazmat.2024.134666
Received 24 March 2024; Received in revised form 7 May 2024; Accepted 19 May 2024
Available online 21 May 2024

combination of five machine-learning algorithms (i.e., XGBR, GBR, LGBMR, CBR, RFR) through a newly developed ensemble learning model. Evaluations on model robustness, overfitting tests, 10-fold cross-validation, internal and stratified validation were performed. We found that the ensemble model depicted about 98.7% spatial variability of benzene (Adj. R^2 =0.987). Through rigorous validations, stability of model performance was confirmed. Several predictors that contribute to benzene were identified, including temperature, developed intensity areas, leaking petroleum storage tank, and traffic-related factors. Analyzing spatial patterns, we found high benzene spread over areas near industrial zones as well as in residential areas. Overall, our study area was exposed to high benzene levels and requires extra attention from relevant authorities.

1. Introduction

Air pollution is a global issue, but an obtrusive problem in Houston, Texas, USA, ranking as the sixth-worst city for air quality [1]. The proximity of industrial facilities to residential neighborhoods and communities has created environmental issues in Houston [2]. As a result, air pollution issues often disproportionately affect socio-economically disadvantaged groups, who are more likely to reside near industrial areas and busy roadways [3]. The Hartman Park community in Houston is surrounded by medium-high intensity development and industries. Texas Environmental Justice Advocacy Services [4] reported that people living in this area are surrounded by industrial facilities that have contributed to benzene emissions and other toxic chemicals. In September 2017, a chemical storage facility belonging to Kinder Morgan also experienced a 10-day accidental release, involving 350 pounds of benzene, ethylbenzene, hexane, toluene, xylene, and an additional 9571 pounds of other volatile organic compounds into the community [5]. Further, the air and ground in this area are contaminated with pollutants, and health issues such as asthma and cancer are frequently reported among the residents [6]. During Hurricane Harvey, for instance, a plume of benzene was identified in Manchester, TX, a neighborhood located close to Hartman Park [5]. According to the city's Health Department, the benzene plume was caused by a leaking tank at the nearby Valero refinery. The Texas Commission on Environmental Quality (TCEQ) also reported that although the benzene concentrations in Manchester dropped after this incident; the overall air quality problem in this area remains ongoing.

Numerous studies have extensively investigated air pollution exposure assessment across countries, with a particular focus on benzene [7-9]. Benzene, a toxic chemical prevalent in industrial areas like the Hartman Park community, raises significant health concerns due to its presence in the air from nearby industrial activities. As a carcinogen, benzene poses various health risks including anemia, harm the bone marrow, increasing the chance for infection, excessive bleeding, and immune system damage [10]. Furthermore, systematic reviews have consistently linked benzene exposure to increased risks of lung cancer, blood malignancies, lymphoid malignancies, and cardiovascular diseases [11-13]. The dense concentration of benzene in industrial pollution areas threatens the community residents' health and safety, contributing to environmental and social disparities [4]. This situation highlights the urgent need for environmental measures and justice to address the air quality issues in vulnerable areas such as Hartman Park community area, which is surrounded by emission sources.

Digital technologies have been widely applied to portray critical environmental conditions, including the use of machine learning algorithms to map air pollution concentrations [14,15]. The primary advantage of machine learning in air pollution estimation is not only being able to deal with big data but also the ability to construct high accuracy estimation models [16,17]. Machine learning algorithms are also able to deal complex linear and non-linear interactions between predictor variables related to air pollution [18].

In the development of estimation techniques, previous studies have integrated several machine learning algorithms, known as ensemble learning models [19-23]. This method is effective since each integrated algorithm offers advantages in handling data and has a cumulative effect that leads to a more accurate model performance than would a single use of machine-learning algorithm [24]. In its implementation, a study by Huang et al. was able to depict around 85% spatial variability of nitrogen dioxide in China by using an ensemble model [25]. A prior study in Taiwan also used the ensemble model to estimate several air pollution sources such as ozone, nitrogen dioxide, and carbon monoxide [21].

While the ensemble method has been shown to perform better than individual models, some limitations exist. *First*, the ability of their developed models was limited to predicting temporal variation. In this case, a scaling approach with fine resolution was not feasible in wider areas due to the constraint of computing power [25]. *Second*, machine learning-based approaches find it difficult to accurately select the most potential predictors before training each algorithm [16]. In fact, selecting irrelevant predictors can lead to overfitting issues, which then reduce computational efficiency and provide a limited interpretation of air pollution estimations. Accordingly, it is important to determine potential predictors before training every single machine-learning algorithm [26].

This study addressed the spatiotemporal distribution of benzene concentrations around the Hartman Park community by developing several machine-learning algorithms and an ensemble learning model that overcome limitations identified in prior research [27,28]. In this case, they have employed stepwise linear regression to identify key predictor variables for subsequent non-linear machine learning-based algorithms in model estimation. However, our study will improve the predictor selection method by adopting a machine learning-based approach, utilizing SHapley Additive exPlanations (SHAP) value. Unlike previous studies that only incorporated predictions from certain algorithms to construct the ensemble model, our study incorporates predictions from all developed single algorithms. This inclusive approach allows all algorithms to contribute to enhancing performance of the ensemble learning model. Furthermore, several validation tests were also performed for evaluation such as overfitting tests, 10-fold cross-validation, internal data validation, and stratified validation. To our knowledge, this approach has limited applied to estimate air pollution exposure in the Hartman Park community. Thus, by emphasizing air pollution estimation techniques with high resolution and accuracy, this study also generated a more reliable model to identify susceptible areas of benzene exposure that warrant more attention.

2. Materials and methods

2.1. Study area

As noted, and illustrated in Fig. 1, this study was conducted in the Hartman Park and Manchester community, in Houston, Texas. Geographically, the north and east of the study site is surrounded by industrial areas (Valero Houston Refinery) and lies close to the Houston Ship Channel; to the south is the railroad, and the east is directly adjacent to the East Loop Highway-Interstate 610. Located at an altitude of approximately 10 m above sea level, the average temperature in the study area during measurements was approximately 69°F (20 °C), with a relative humidity of 58%. In this case, the location where benzene measurements were carried out covered the entire residential area with an extent of around 66 ha. As described in the background, the Hartman

Park community was selected as the focus of this investigation considering the significant impact of air pollution from the surrounding industrial areas that have contributed to intense toxic emissions, including benzene. This condition poses multiple health risks and has an excessive effect on its predominantly Hispanic and lower income population [4].

2.2. Benzene data collection

Raw data of benzene were obtained from direct measurement using a mobile laboratory equipped with a global positioning system (GPS) sensor for determining coordinates in the Hartman Park and Manchester community area (Fig. 1). This on-board measurement was conducted for five days between 08/11/22 to 10/11/22 (p1) and 16/11/22 to 17/11/22 (p1), resulting in the collection of approximately 53,528 samples of benzene data. The main purpose of these 5-day measurements was to get representative data of morning to evening emission. A Proton Transfer Reaction – Mass Spectrometry (PTR-MS) tool was utilized to collect the benzene traces every second. Typically, this measurement was done in a bidirectional format. In this case, a bidirectional format allows the car to be driven in either direction, forwards or backwards. Additionally, weather conditions during the measurement periods were sunny and temperatures averaged 78°F during the first measurement period (p1), and decreased to approximately 56°F during the second period (p2).

2.3. Predictor Variables

2.3.1. Meteorological and topographical factors

As listed in Table S1, several variables were set as predictors for estimating benzene concentrations.

Meteorological factors obtained from direct measurements along with benzene data were considered as predictors. Those data include temperature, relative humidity, atmospheric pressure, and wind data.

By using a Magellan weather system mounted on the moving car, the meteorological data were measured. Since this system was installed together with GPS and PTR-MS, the meteorological data will be matched with time on GPS and benzene concentrations. To synchronize meteorological data, 125 points without matched time information were removed. Minimizing the possibility of missing data, we turn on the meteorological device before the PTR-MS which was used for benzene measurements. Considering topographical factors, we assessed elevation and slope using the digital elevation model (DEM). The data with 30-m spatial resolution were obtained from the United States Geological Survey and resampled to 5-m spatial resolution. The resampling process was performed considering all variables were generated at 5-m spatial resolution. In this case, the "nearest-neighbor resampling" function provided by ArcGIS Pro was applied. This resampling method is commonly employed in image processing that involves the nearest value of the original value [29]. By bypassing interpolation between data, this method has the advantage of maintaining the original values.

2.3.2. Land use/land cover datasets

Prior studies have reported that land use and land cover data were determinants of air pollution compounds such as benzene [30,31]. In terms of land cover, we used data obtained from Houston-Galveston Area Council [32]. In this case, 14 types of land cover were assessed, such as developed-open space, developed-low intensity, developed-medium intensity, developed-high intensity, deciduous forest, evergreen forest, mixed forest, shrub, woody wetland, open water, emergent herbaceous wetlands, crops, herbaceous, and hay. Meanwhile, for land use, we utilized the railroad data provided by the City of Houston Geographic Information System (COHGIS) as predictors [33]. In addition, we also counted green land cover data from satellite-based vegetation index provided by the National Aeronautics and Space Administration (NASA). In this case, greenness was represented by the normalized difference vegetation index (NDVI) with a temporal resolution of 16 days.

2.3.3. Traffic-related factors

Traffic elements are indicated to be one of the contributors to benzene [34]. In this study, we collected airport data from COHGIS in 2020

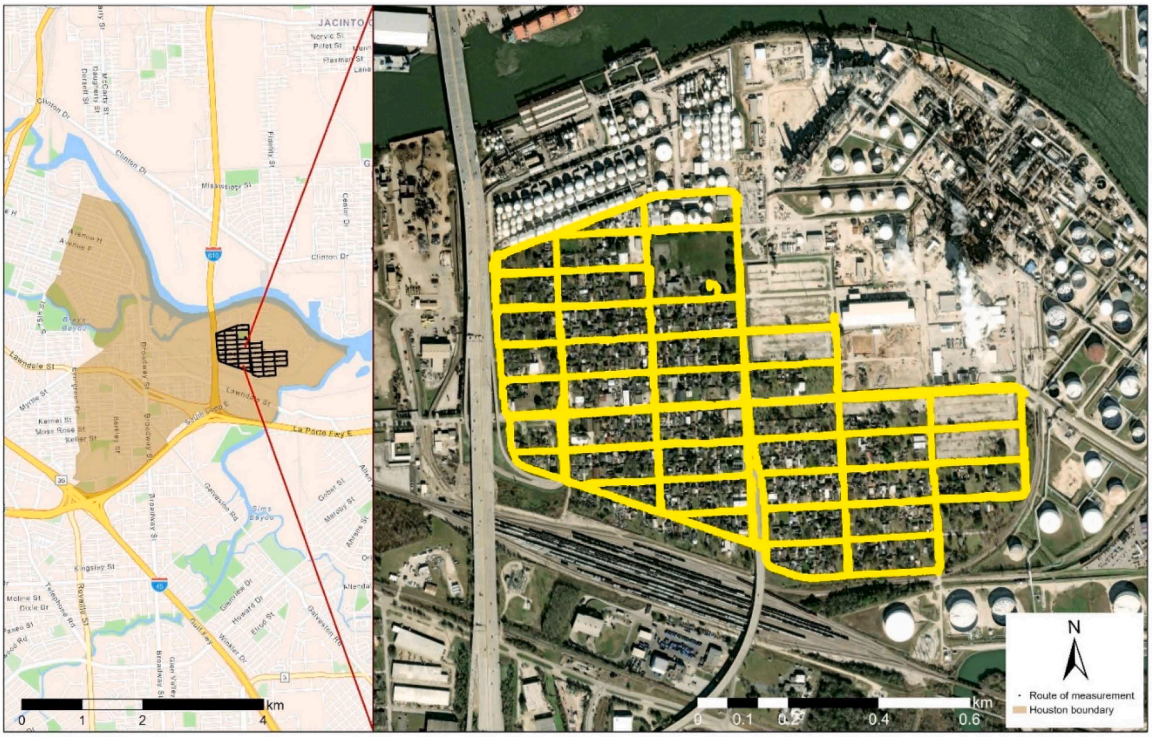


Fig. 1. The site of benzene measurement in the Hartman Park and Manchester community area in Houston, Texas, United States. The yellow line indicates the route of on-board measurement for benzene couple with meteorological data collection.

[35] because it can be a source of chemicals due to the use of diesel fuel in aircraft engines and reference ground support equipment [36]. Further, the presence of roads can be a significant contributor to benzene emissions [37,38]. Considering the condition of the road network around the study site, eight main types of roads were assessed, including freeway, major arterial, minor arterial, access, frontage, abandon, high occupancy vehicle, and ramp. The data were collected from The City of Houston Geographic Information System (COHGIS) in 2018 [39].

2.3.4. Industrial data

Since benzene is primarily produced from petroleum and used in industrial processes [40], several industrial sources were considered. Here, we assessed industrial waste storage data which includes petroleum storage tanks (PST), leaking petroleum storage tanks (LPST), and industrial and hazardous waste corrective action (IHCWA) as predictors. In addition, knowing that some abandoned or unused industrial properties may contribute as the emitters, brownfields, landfills, and wastewater outfalls data were also assessed [41]. Those data were provided by the Texas Commission on Environmental Quality in 2023 and were publicly accessible.

2.4. Development of machine learning-based models

We fully developed an ensemble learning model by integrating several machine learning algorithms to estimate benzene concentrations. As presented in Fig. 2, several steps were conducted including: (1) database preparation; (2) calculating hyperparameter for every single machine learning algorithm; (3) estimating SHAP value to determine the most potential predictors contributed to benzene emission; (4) developing five single machine learning algorithm to estimate benzene concentrations; (5) validation test of single machine learning models; (6) developing ensemble learning model; (7) validation test of all developed models and ranking analysis; (8) mapping benzene concentrations based on the best estimates model, which was ensemble learning model. The following section gives further detailed explanations of each step.

2.4.1. Data preparation and variable selection

In the initial step, after benzene coupled with meteorological data were measured, several predictor variables were then prepared. In this case, geospatial-based variables including land use, land cover, topography, traffic-related factors (road network, trail, transportation), and industrial waste storage data surrounding the measurement site were estimated at buffer ranges of 250, 500, 750, 1000, 1250, 1500, 1750, 2000 m. We used point density, line density, and focal statistics function to estimate the density of point-based, polyline-based, and polygon-raster-based predictor variables, respectively. Additionally, the Euclidean distance method measured distances between variables, such as the distance between land cover and the measurement site. This method calculates the distance from each cell in the raster to the closest source. Using the geoprocessing tool in ArcGIS Pro, this function estimated the distance of the central point of emission sources from the measurement site [42]. In developing the main database, we finally generated around 200 predictor variables to estimate benzene concentrations as listed in Table S1. In our model development, benzene was designated as the dependent variable (Y), while the other variables were treated as independent variables (X).

Since this study fully used a machine-learning based approach, we then randomly split the main database (N = 53,528) into 70% training data (N = 37,469), 20% testing data (N = 12,847), and remained 10% internal data validation (N = 3212). After data splitting, we calculated hyperparameter for every single machine learning algorithm to control model structure and performance (Table S2). Hyperparameters are the best combination of values for machine learning parameters [43]. Since they are set before the learning process begins and control the learning process, allowing users to tweak model performance for optimal results. We have five single algorithms developed in this study including gradient boosting regressor (GBR), extreme gradient boosting regressor (XGBR), light gradient boosting machine regressor (LGBMR), categorical boosting regressor (CBR), and random forest regressor (RFR). Differing from prior studies that employed stepwise linear regression [27,44,28], this study used the order of SHAP values to initially determine the most potential predictors. Involving a game-theoretic concept that measures each feature's contribution to the outcome [45], SHAP value has the advantage of demonstrating the effect proportion of each predictor attributed to benzene concentrations. In this case, game theory refers to a mathematical framework used to analyze interactions between rational decision-makers, e.g., players and agents [46]. Considering that this study developed machine-learning-based algorithms, utilizing SHAP values that can be used to interpret any machine-learning models is a more reasonable compared to linear regression-based method. The main criterion for variable selection was based on SHAP values, with variables contributing to changes in R^2 by more than 0.001

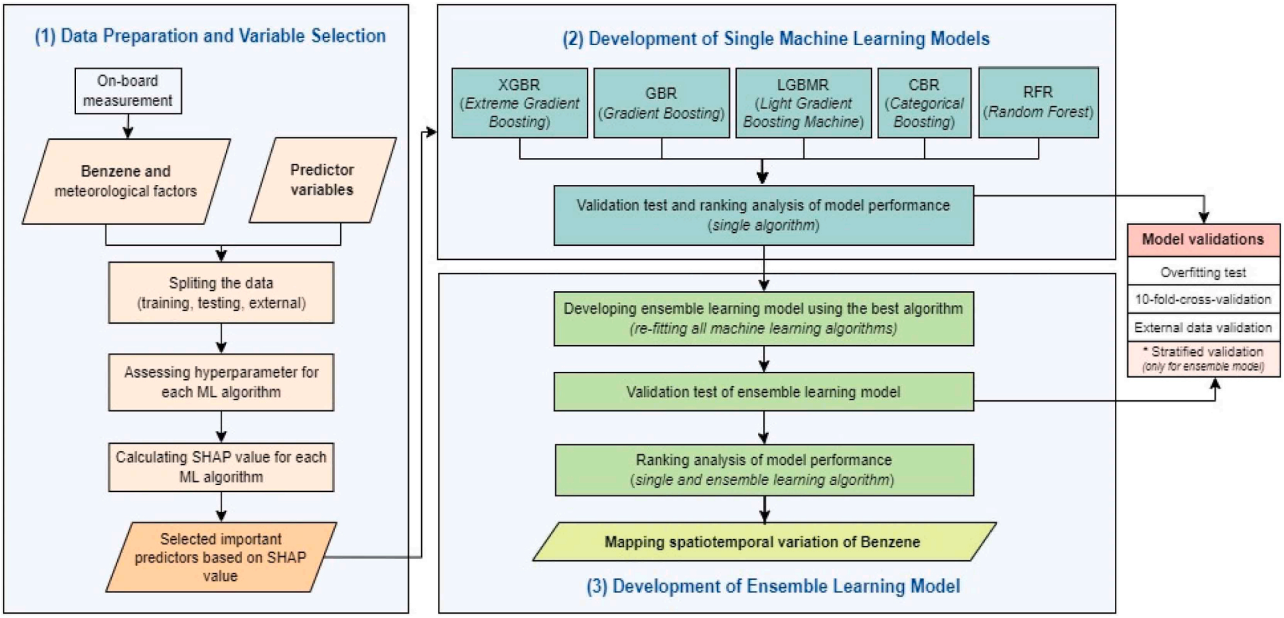


Fig. 2. The Framework of benzene estimation in the Hartman Park Community area, Houston by using machine learning-based approach.

being retained for algorithm development. Since five different machine learning algorithms were employed, SHAP values were computed individually for each algorithm. This method allows the SHAP analysis to identify the most relevant variables for assessment in each algorithm, recognizing that variable importance can vary across different models.

2.4.2. Development of single machine-learning and ensemble learning models

As noted, after potential predictors were determined, five single machine learning algorithms including GBR, XGBR, LGBMR, CBR, and RFR were developed. In this case, each algorithm assesses different predictor variables based on SHAP value analysis. The five machine-based algorithms were selected by considering their respective advantages for performing the estimation process. *First*, by incorporating several decision trees and applying gradient descending functions, GBR can optimize a smoothed estimation of the compatibility index which enhances accuracy [47,48]. *Second*, as a branch of the boosting algorithm, XGBR can deal with missing data, improve accuracy, work well on small datasets, can process data more quickly and flexibly compared to original gradient boosting [49]. *Third*, one type of gradient-boosting algorithm that applies a leaf-wise tree growth approach, LGBMR can effectively reduce memory usage, prevent overfitting issues, and provides better model accuracy [50,51]. *Fourth*, adopting the advantages of general gradient boosting and applying an ordered boosting approach to train data, CBR has the advantage of handling gradient bias and is more appropriate for dealing with categorical variables than other algorithms [52,53]. *Fifth*, used for both classification and decision tree regressions [54], RFR can deal with overfitting, is flexible to perform regression, can handle large variables, and offers a reliable alternative to traditional parametric-semiparametric statistical analysis.

This study also utilized automated machine learning, or autoML, to improve efficiency and reduce manual interference in data processing, modeling, and hyperparameter optimization. AutoML libraries and frameworks are freely available in repositories for public use. In this part, all five machine-learning algorithms were integrated with this automation framework using the "auto_ml" developed by Parry [55]. AutoML was chosen because this framework supports analysis of tree-based algorithms including gradient boosting-based algorithms and regression-based models. Further, suitable for both analysis and real-time estimates, autoML allows big data processing, robust data scaling, feature selection, and determination of the best model for generating accurate predictions [56]. We used Python 3.7 which was integrated with the Jupyter Notebook platform to run the automation process in the machine learning analyses. In addition, we also utilized several statistical-based software for model analysis, including Ms. Excel and R 3.6.3.

Apart from developing a single algorithm, this study developed an ensemble learning model by refitting all five single machine learning algorithms. The main algorithm in generating the ensemble model was the result of ranking analysis of model performance indicators. In this case, the ranking analysis was established by comparing the performances of training, testing, and 10-fold cross-validation (10-fold CV) of single machine learning algorithm. Several indicators that were compared included the coefficient of the determinant (R^2 and Adj. R^2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and overfitting values. Of the five algorithms compared, the performance of the algorithm selected as the best model was re-examined to develop an ensemble learning model. After all of the single algorithms and ensemble learning algorithms were developed, we re-ranked them to ensure the best model for estimating benzene concentrations. In this case, by summing all ranking values, we determine the algorithm with the lowest total value as the best model for generating benzene estimate maps. In addition, we also generated graphs of SHAP values to depict the predictor variables that were most influential for benzene emission in each developed model.

2.5. Validation tests

Several validation tests were carried out to evaluate the model performance, including 10-fold CV, overfitting test, internal data validation, and stratified test. In the 10-fold CV, the data was randomly divided into 10 subsets. Nine of these subsets were used to develop the training, while the remaining subset was used for testing. This process was reiterated until all subsets were examined, ensuring each fold had a chance to be the held-out test set. The models were deemed robust if there was not a significant change in the performance of their indicators. For the overfitting test, we checked the value of Adj. R^2 and calculated the difference between the training and testing model, as well as the difference between the training and 10-fold CV. Further, internal data validation was performed by examining 10% of data which was not examined for training and testing. This validation was intended to ensure whether the algorithms could work well even though it uses different datasets. Stratified validation by time variations and emission sources (LPST, IHCWA, freeway, ramp, all road, and areas with high intensity development) were further carried out. For the stratified test by time variation, we developed a model for each measurement day for the five days' worth of measurements. Meanwhile, for stratified tests based on emission sources, the models were developed based on the distance between the emitter location and the measurement site. Here, near, and far were distinguished based on the median value of the distance between the emission source and the measurement site. After model development and validation tests were done, a spatiotemporal mapping of benzene concentration was generated using the best-selected model, which was the ensemble learning model. ArcGIS Pro was then employed for mapping and spatial analysis.

3. Results

3.1. Statistical trend of benzene concentrations

Fig. 3 depicts the trend of benzene concentrations surrounding the Hartman Park community over the study periods. Aggregated per minute, measurement on the first day showed an average benzene concentration of 0.352 ppb. The average benzene on the second day measurement was 0.421 ppb. Then, measurement on the third day showed an average concentration of 0.686 ppb. Measurement on the fourth day showed an average benzene concentration of 0.410 ppb. Also, measurement on the fifth day showed an average concentration of 0.433 ppb. Visualizing measured concentrations, we also incorporated the plotting of on-board benzene measurement in Fig. S1**.** Overall, we found that on the third day of measurement (morning time), the benzene concentration was higher compared to other days. Descriptive statistics for each variable examined in this study were also performed in Table S3**.**

3.2. The performance of all developed models

Of the six algorithms, the ensemble learning model demonstrated the best performance for estimating benzene concentrations compared to the other five machine learning models. As shown in Table 1, the ensemble model yielded the highest coefficient of determination (Adj. R^2) of 0.987 and the lowest RMSE of 0.056 ppb. This indicated that the ensemble model can explain about 98.7% benzene variability. Not significantly different, modeling using the XGBR algorithm also resulted in a high Adj.R^2 of 0.986 with an RMSE of around 0.058 ppb. Further, we noted that other single algorithms perform well with Adj. R^2 values were 0.968, 0.947, 0.907, and 0.757 ppb for CBR, LGBMR, RFR, and GBR, respectively. Meanwhile, the RMSE were 0.087, 0.113, 0.156, and 0.273 ppb. For assessing testing data, 10-fold CV, and internal verification, model robustness was evaluated. By using testing data, the ensemble learning model consistently showed the best performance (Adj. R^2 = 0.928, RMSE 0.145 ppb), followed by XGBR, LGBMR, CBR, RFR, and

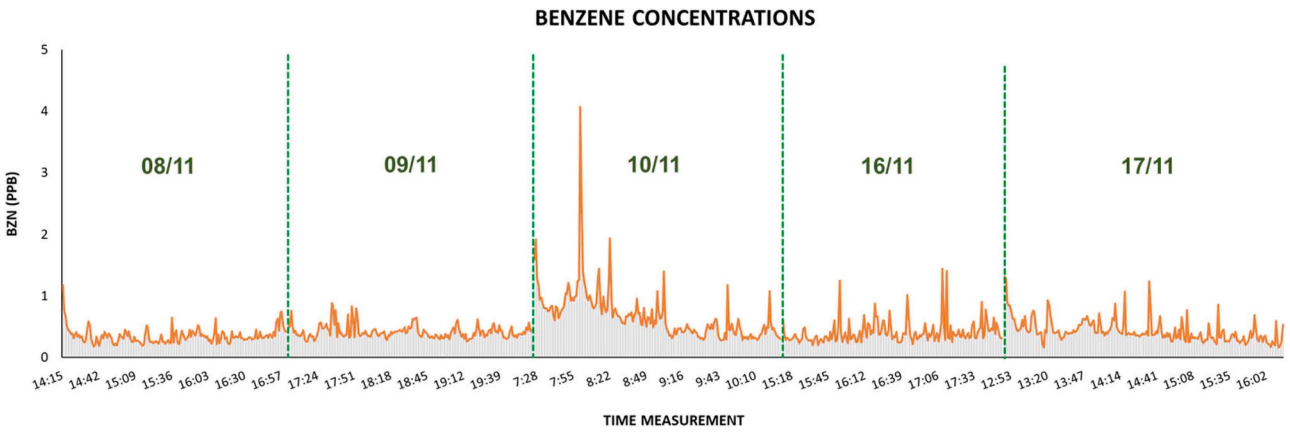


Fig. 3. Temporal trends of average benzene concentrations were derived from direct measurements in the Hartman Park community area. The numbers were aggregated every minute over five-day-period.

Table 1
Model performance of all developed algorithms.

Algorithm	Number of Predictors	Process	R^2	Adj. R^2	RMSE	MSE	MAE	Overfitting
XGBR	6	Training	0.986	**0.986**	0.058	0.003	0.004	0.062 *
		Testing	0.923	**0.923**	0.149	0.022	0.011	
		10-fold CV	*0.863*	*0.863*	*0.157*	*0.025*	*0.011*	0.108#
		Internal	*0.878*	*0.878*	*0.169*	*0.029*	*0.012*	
GBR	8	Training	0.757	**0.757**	0.273	0.074	0.093	0.007 *
		Testing	0.749	**0.749**	0.310	0.096	0.091	
		10-fold CV	*0.884*	*0.884*	*0.165*	*0.027*	*0.031*	0.127#
		Internal	*0.638*	*0.638*	*0.286*	*0.082*	*0.095*	
LGBMR	7	Training	0.947	**0.947**	0.113	0.013	0.026	0.043 *
		Testing	0.905	**0.905**	0.168	0.028	0.029	
		10-fold CV	*0.836*	*0.836*	*0.196*	*0.038*	*0.029*	0.111#
		Internal	*0.846*	*0.845*	*0.173*	*0.030*	*0.030*	
CBR	7	Training	0.968	**0.968**	0.087	0.008	0.032	0.114 *
		Testing	0.853	**0.853**	0.201	0.040	0.034	
		10-fold CV	*0.910*	*0.910*	*0.145*	*0.021*	*0.034*	0.058#
		Internal	*0.942*	*0.942*	*0.101*	*0.010*	*0.033*	
RFR	8	Training	0.907	**0.907**	0.156	0.024	0.011	0.061 *
		Testing	0.846	**0.846**	0.224	0.050	0.015	
		10-fold CV	*0.793*	*0.793*	*0.231*	*0.053*	*0.016*	0.114#
		Internal	*0.759*	*0.759*	*0.215*	*0.046*	*0.015*	
Ensemble Learning	6	Training	0.987	**0.987**	0.056	0.003	0.004	0.059 *
		Testing	0.928	**0.928**	0.145	0.021	0.011	
		10-fold CV	*0.981*	*0.981*	*0.067*	*0.005*	*0.004*	0.006#
		Internal	*0.856*	*0.855*	*0.161*	*0.026*	*0.011*	

* Overfitting 1, the difference between the training and testing model
# Overfitting 2, the difference between the training and 10-fold CV model

GBR (Adj. R^2 were 0.923, 0.905, 0.853, 0.846, and 0.749, respectively). Further, the 10-fold CV test showed the best performance of the ensemble model compared to other single models. However, we also acknowledge that in validation using internal data, the best performance was shown by the model built using the CBR algorithm. In addition, the overfitting values of the ensemble model were low compared to others (less than 0.06) indicating the high accuracy of the estimation model. When examining all indicators of model performance in ranking analysis (Table S4) (i.e., R^2, Adj. R^2, RMSE, MSE, MAE, and overfitting values), we still found that the ensemble learning model was the best algorithm to estimate benzene with the lowest total ranking value. Through this process, we identified several predictor variables that have the most influence on contributing to benzene emissions including temperature, developed low-medium-high intensity areas, and LPST. The contributions from the road network such as major arterial, frontage, ramp, access, and freeway were also recognized. In details, this information was presented as SHAP value's plot in the supplementary file, Fig. S2.

3.3. Stratified Validations

The results of the stratified tests were shown in Fig. 4. Analyzed using data from each measurement day (Fig. 4**a**), we found that the ensemble model showed better performance than others. This was demonstrated by the high R^2 (0.830 to 0.988) and low error (MAE) on each day. From the presented graph, the GBR algorithm performed poorly compared to other single machine learning algorithms. The best model performance was found on the third day of measurement, and this differed from other algorithms that did not show significant changes in model performance between days. Dividing the data based on the distance to the emission sources (Fig. 4**b**), the ensemble model demonstrated stable results with high performance (R^2 = 0.898 to 0.981). This validation test indicates the robustness of the ensemble model in handling data to estimate benzene even though the datasets controlled for various characteristics of emission sources. From several validation tests constructed, we finally selected an ensemble learning model to generate spatiotemporal maps of benzene concentrations.

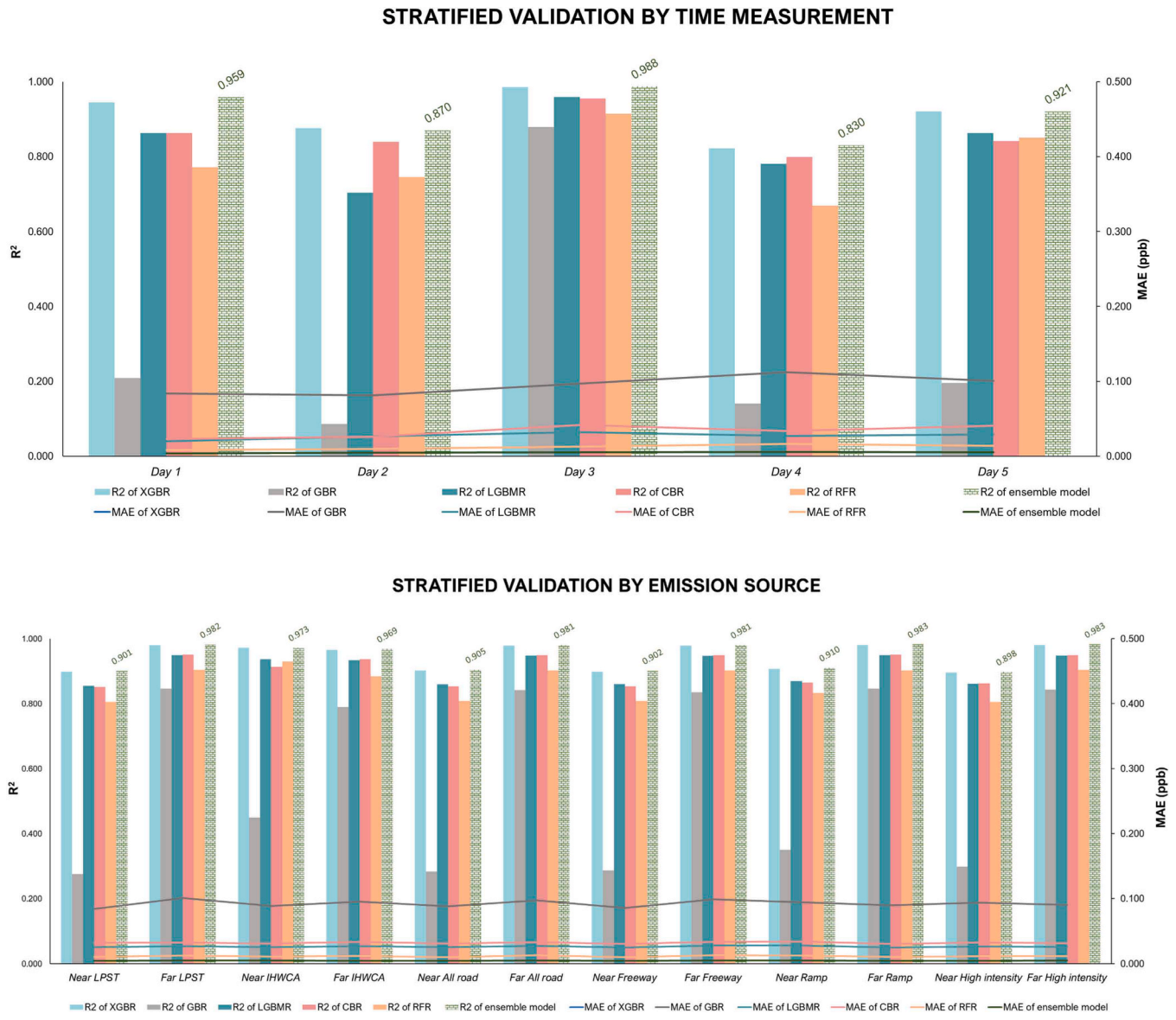


Fig. 4. (a) Validation tests in each measurement day (day 1 to day 5). Illustrated the values of R^2 and MAE for all developed algorithms, i.e., XGBR, GBR, LGBMR, CBR, RFR, and ensemble learning.Fig. 4 **(b)** Validation tests by the distance of emission sources. Illustrated the values of R^2 and MAE for all developed algorithms, i.e., XGBR, GBR, LGBMR, CBR, RFR, and ensemble learning.

3.4. Spatiotemporal analysis of benzene concentrations

Using the estimation results from the ensemble learning model, Fig. 5 depicted the spatial distribution of benzene concentrations at the study site. Based on five-day measurements, we identified high benzene levels (dark brown areas), $>$ 1.0 ppb, at several locations. To the north and east, high concentrations were observed in high-intensity development zones dominated by industrial areas. In the south, we found high exposure to benzene in the railroad area which is a medium-high intensity development zone. Meanwhile, the western part of the freeway was also identified as contributing to high benzene emissions. We also noticed a high exposure to benzene in residential areas surrounded by road networks where direct measurements were carried out. Mapping the spatial distribution of benzene on each day, the estimation results showed that high concentrations were always concentrated in the north and east which areas of the community, where the primary industrial areas such as the Valero Refinery are located. From the in-depth analysis, Fig. 6 showed the spatial distribution of benzene with high concentrations exceeding the threshold, 1 to $>$ 2 ppb. The results showed a similar pattern where high concentrations can be seen in several locations mentioned previously. Conducting scale adjustment as shown in Fig. S3, we found that nearly the entire study area was exposed to benzene with concentrations $>$ 0.5 ppb.

4. Discussion

Preliminary analysis of direct measurements in the Hartman Park community revealed elevated benzene concentrations during morning hours, while measurements over other timeframes showed no significant changes. We assumed that the high morning concentration of benzene is closely related to the operational hours of industrial and traffic activities particularly those adjacent to congested freeway [57,58,31]. To present the spatial patterns of benzene concentrations in this susceptible area, we developed multiple machine learning-based algorithms and proposed an ensemble learning model. As the main result, the selected ensemble learning model yielded efficient and accurate estimates of benzene concentrations, demonstrating high explanatory power with an adj-R^2 of 0.987. The superior performance of the ensemble learning

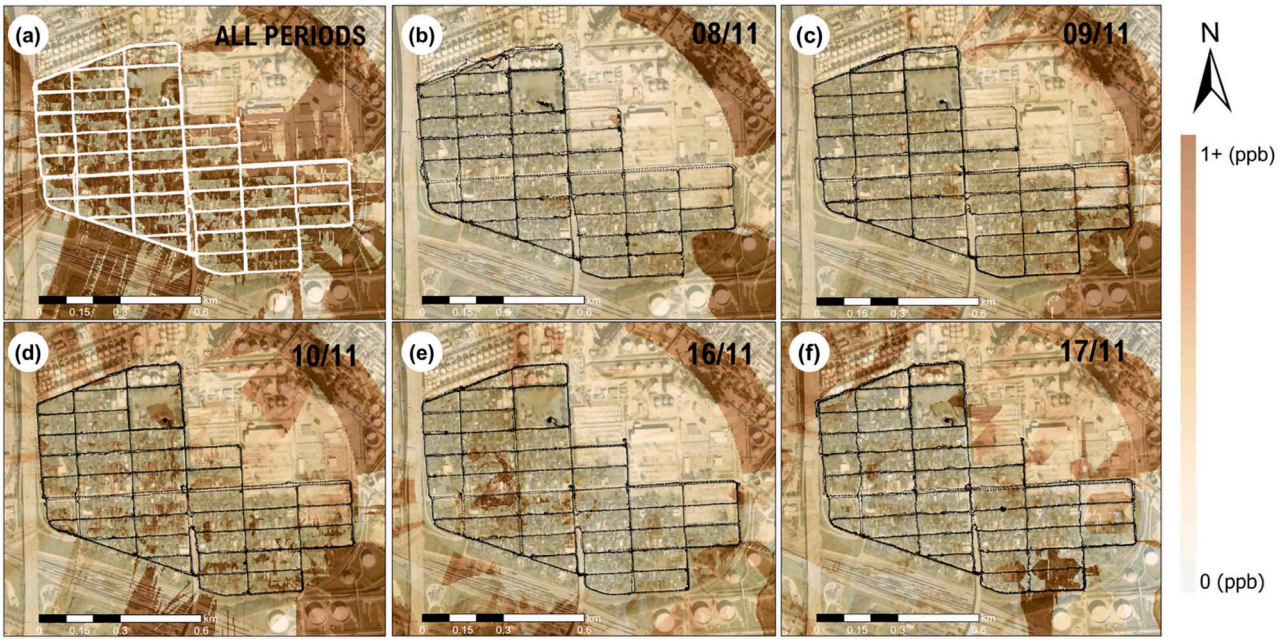


Fig. 5. Spatiotemporal distribution of benzene concentrations (0 to 1 + ppb) in the Hartman Park community area. Generated using estimated values from an ensemble learning model, it incorporated data from all measurement times (a) and daily measurements (b-f).

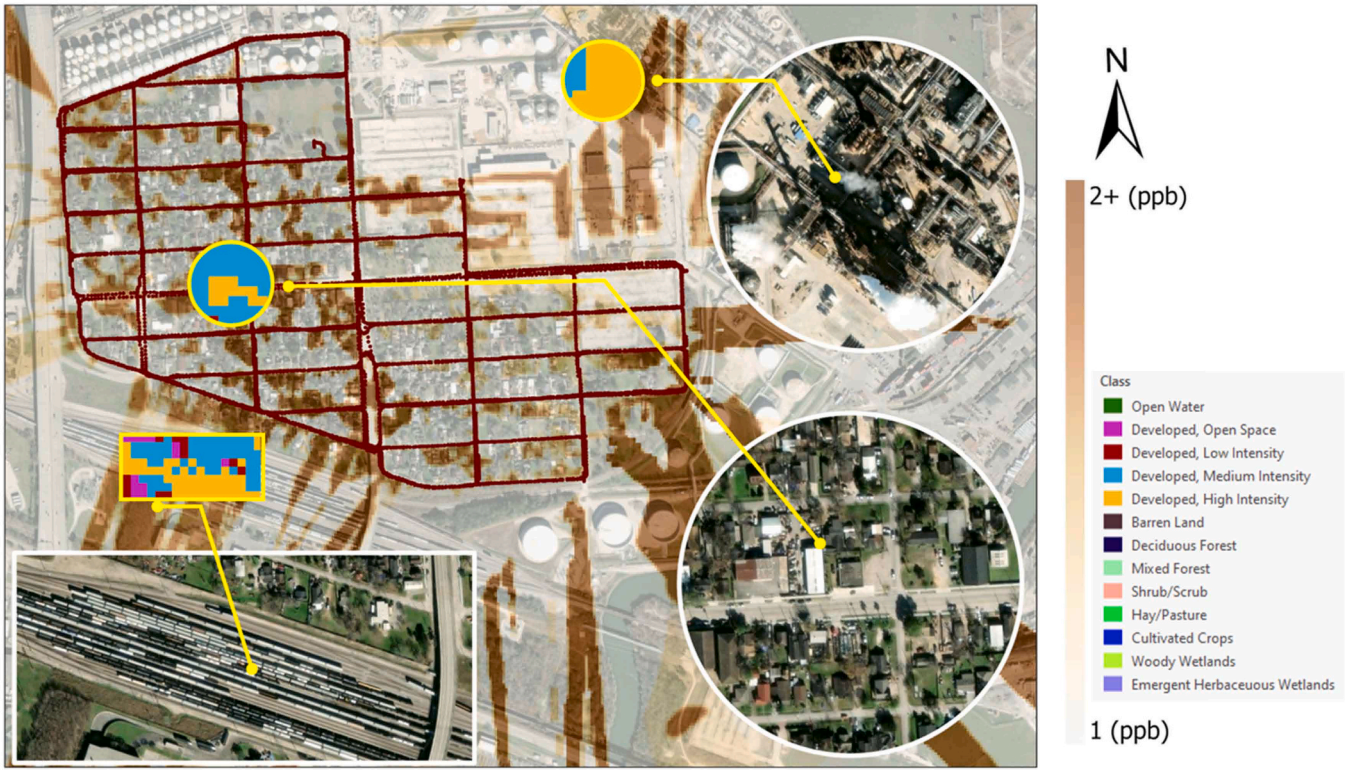


Fig. 6. Spatial pattern of benzene concentrations (1 to 2 + ppb) in the Hartman Park community area. Generated using estimated values from an ensemble learning model and it incorporated data from all measurement times.

model in estimating benzene concentrations may be attributed to its ability to minimize errors, and deal with overfitting issues [59-61]. This was demonstrated by lower RMSE, MSE, and MAE compared to other single algorithms, with overfitting values less than 0.06. Evaluation by testing multiple subsets of data through stratified validation analysis, the ensemble learning model exhibited better and more stable performance compared to single algorithms. Further, the ensemble model combines the strengths of multiple algorithms also resulting in higher performance within a relatively short processing time (Table S5). Moreover, ensemble methods exhibit greater robustness to data noise compared to single algorithms, allowing them to maintain good performance even when outliers were retained in the estimation. Although ensemble learning is not a novel approach, our study design incorporated several improvements that surpass prior studies [62,25,63].

Specifically, our study involved fine-tuning hyperparameters, estimating SHAP values for variable selection and determining important predictors, and creating an ensemble model that refits all developed machine learning algorithms, rather than specific ones. Finally, these enhancements enabled us to identify the most influential predictors and generate high-resolution estimates of pollutant concentrations.

We also found that GBR exhibited poorer performance compared to other algorithms. However, despite this, GBR still demonstrated reasonable effectiveness by capturing over 75% spatial variability of benzene. We opined that the lower performance of GBR is due to its sensitivity to noisy data and its lesser ability to handle it compared to other algorithms. Moreover, GBR has fewer hyperparameters to tune and limited regularization compared to other boosting tree-based algorithms, which may simplify the modeling process but also provide less flexibility in optimizing model performance [48,64]. Besides the main model, the weakness of GBR in estimation became apparent from the results of stratified validations. We inferred that the model's underperformance was due to GBR's limitations in evaluating data with a small sample size, potentially resulting in overfitting. As known, in the stratified test, the main database was divided into several subsets, e.g., five subsets for stratified validation based on measurement time.

Interpreting the benzene estimations generated from ensemble model, we observed spatiotemporal emission patterns. Throughout the study periods, we noted that nearly every day, the study site experienced benzene exposure exceeding 0.5 ppb. Upon adjusting the color scale on the estimation map, we observed that high concentrations of benzene (>1 ppb) were consistently dispersed in the industrial area, located to the north and east of residential zones. This pattern remained consistent across each day. This result was in accordance with the Collins [65]'s report which stated that this refinery had an average benzene of 3.3 micrograms ($\pm$ 1 ppb) and east refinery emitted benzene at levels exceeding the environmental protection administration limit. The study area that is also close to the Houston Ship Channel to the north, may increase air toxic emissions due to the involvement of petrochemical compounds [66]. Furthermore, our estimation maps depicted high benzene levels in the railroad area. In this case, exposure to diesel exhaust from railroad activities may contribute to benzene emissions, as diesel fumes contain many types of chemicals, including benzene [67,68]. Examining daily exposures, we identified a spatial pattern of elevated benzene on the third and fourth days of measurement, with concentrations exceeding 1 ppb dispersed throughout residential area. We assumed that the dense exposure to benzene from the surrounding industrial zone, freeways, railroads, and road activities nearby has contributed to the heightened benzene levels in this area. In addition, our generated map also noticed the high concentrations in areas close to freeway. This makes sense in that benzene is a component of traffic-related air pollution [69]. Performing additional analysis, Fig. S4 revealed that the benzene concentrations in the area close to the industrial zone and the main chimney of air pollutant release was around 4 ppb. Referring to the federal threshold set by the Texas Commission on Environmental Quality (TCEQ), the average allowable benzene emission is 9 $\mu g/m^3$ or equal to 3.44 ppb. Thus, we underlined that the benzene emission in this area exceeded the permissible standard.

Implementing concepts to measure the variable contributions, our ensemble model can determine the most influential predictors of predicted pollutant and outperformed many previous studies that apply ensemble learning but were unable to identify potential predictors [63]. Several variables were identified as strongly contributing to benzene concentrations in Hartman, including the temperature, developed low-medium-high intensity areas, and LPST. The involvement of temperature in relation to benzene concentrations in the air can be explained through its influence on traffic, manufacturing activities, and atmospheric conditions such as cold start emissions, emission rate variability, and vapor pressure [70-72]. In relation to developed low intensity areas as a potential predictor, we identified that this sector is dominated by residential areas where population activities may contribute to benzene emission. Furthermore, this study also highlights that developed, medium and high intensity areas are major contributors to benzene emissions in Hartman. This cannot be denied considering that this area is dominated by large-scale industrial areas to the north and east, while, to the south, is the railroad and, to the west, is the freeway. The United States Environmental Protection Agency [73] confirmed that industrial activities are the largest contributor to benzene exposure as emitted by petroleum refineries, through process vents, storage tanks, equipment leaks, transfer operations, and wastewater collection and treatment [74].

This study also found the involvement of LPST as a benzene emitter. This makes sense considering that this carcinogen can be released from various sources at petroleum storages, especially equipment that have a history of leaks [75,76]. In fact, TCEQ reported that storage tank leaks have occurred in Houston at more than 250 locations from 1985 until the end of 2022 [77]. Apart from industry-related factors, the presence of road network activities from major arterial, frontage, ramp, access, and freeway also contributed to benzene exposure in Hartman. Moreover, from a geographical aspect to the west of the study site is the East Loop Highway-Interstate 610, the worst road segment with the most congested traffic, which has a high possibility of emitting air pollution [78,79].

To our understanding, estimating, and modeling the spatial distribution of benzene concentrations in vulnerable areas near the Valero Houston Refinery, Hartman Park community, is very limited. As our strength, this study provided interesting findings by identifying a distribution pattern of benzene with high concentrations within an industrial zone which then also spread to surrounding residential areas. Generated benzene maps reveal critical points of benzene exposure that called for further environmental measures. Technically, by using machine learning-based algorithms, this study resulted in estimates of benzene concentrations with high explanatory power and accuracy, verified by several validation tests. When applying this concept to measure the portion of variable contributions to model performance, this study can also identify predictors that may have a strong possibility as benzene sources. We also acknowledge that the model selection substantially impacts the estimated outcomes, with implications for environmental decision-making and future research directions. Accurate estimation of benzene concentrations is crucial for identifying environmental risks and guiding mitigation efforts. The ensemble learning model, by effectively capturing spatial patterns of benzene distribution, can pinpoint areas requiring immediate attention most effectively. Furthermore, identifying influential factors aids in prioritizing environmental interventions to address challenges in vulnerable community areas near benzene emission sources. This highlights the importance of selecting models that offer both accuracy and interpretability to inform strategies for effective environmental management. Finally, this study design can also be widely adopted and implemented in other regions with numerous improvements, adjusted to the study area characteristics.

As for improvement, we also noted several study shortcomings. *First*, even considering time variations (i.e., morning-afternoon-evening), benzene measurements were only conducted over a short period of five days and cannot capture seasonal variations. We suggest that for future studies benzene can be estimated using long-term data and measured from fixed stations which are installed around residential areas or areas with potential as emission sources. *Second*, this study was unable to account for the operating time of industrial facilities that may contribute to benzene emission variations. Further studies should consider this factor to strengthen the study analysis.

5. Conclusions

During initial measurement, our study pointed out that the benzene concentrations were identified to be higher in the morning and tended to be similar at other time ranges. By developing several machine learning-

based algorithms, the ensemble learning model revealed the best performance in depicting the spatiotemporal variability of benzene concentrations with an explanatory power of around 98% (RMSE 0.056 ppb). Validation tests also showed the robustness of the selected ensemble model with no significant changes in performance even after several adjustments. In addition to high-accuracy estimates, our model was able to identify several predictors that may have major contributions to benzene emissions such as temperature, low-medium-and-high intensity developed areas, LPST, and traffic-related factors. Interpreting the spatial pattern of exposure distribution, we found that high benzene concentrations (>1ppb) scattered to the north and west near the industrial zone, and further towards residential areas. Lastly, this study emphasized the critical conditions in the Hartman Park community which are exposed to high benzene concentrations and called attention to relevant authorities for further environmental measures and justice.

CRediT authorship contribution statement

Chih-Da Wu: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Shih-Chun Candice Lung:** Writing – review & editing, Funding acquisition. **Aji Kusumaning Asri:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Rui Zhu:** Writing – review & editing, Data curation. **Hsiu-Ling Chen:** Writing – review & editing, Conceptualization. **Galen D. Newman:** Writing – review & editing, Investigation, Funding acquisition, Data curation, Conceptualization. **Zhihan Tao:** Writing – review & editing, Data curation.

Environmental Implications

This study underscored the critical environmental challenges faced by the Hartman Park community in Houston, Texas-USA, particularly concerning air quality issue. The concentration of benzene, a hazardous pollutant, was estimated using advanced machine learning techniques, revealing alarming levels in residential areas adjacent to industrial zones. The identification of contributing factors, including developed intensity areas, industries, and traffic-related emissions, highlighted the complex nature of air pollution in the study area. The findings emphasized the urgent need for targeted interventions and policy measures to protect public health and mitigate the environmental impact on vulnerable communities.

Declaration of Competing Interest

The authors declare no conflict of interest. The funders had no role in the study design; data collection, analyses, interpretation; writing of the manuscript; nor in the decision to publish the results.

Data Availability

Data will be made available on request.

Acknowledgements

This study was granted by the National Science and Technology Council, Taiwan (MOST 110-2628-M-006-001-MY3; NSTC 112-2121-M-006-015-; NSTC 112-2123-M-001-008-; NSTC 112-2121-M-006-004-), and the “Innovation and Development Center of Sustainable Agriculture” from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. Additional funding support was provided by the Research Center for Precision Environmental Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and by Kaohsiung Medical University Research Center Grant (KMU-TC113A01). We also acknowledge support from the National Institute of Environmental Health Sciences Superfund Grant #P42ES027704–01. Data availability was supported by the City of Houston Geographic Information System (COHGIS), Houston-Galveston Area Council (H-GAC), the National Aeronautics and Space Administration (NASA), and the United States Geological Survey (USGS) for providing satellite-derived data.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2024.134666.

References

[1] Sessions, K. (2023). Report: Houston ranks sixth on list of U.S. cities for worst air pollution. Available at: www.chron.com/news/houston-texas/article/houston-air-pollution-17830025.php, last access: 01 February 2024.
[2] Lam, Y., Sivasubramanian, R., Guerrero, M., & Parras, J. (2021). Toxic air pollution in the Houston ship channel: disparities show urgent need for environmental justice. www.nrdc.org/sites/default/files/air-pollution-houston-ship-channel-ib.pdf.
[3] Sexton, K., Linder, S., Delclos, G., Stock, T., Abramson, S., Bondy, M., & Ward, J.. (2005). A closer look at air pollution in Houston: Identifying priority health risks. The Institute for Health Policy, University of Texas School of Public Health, Health Science Center at Houston.
[4] Texas Environmental Justice Advocacy Services. (2016). Air Toxics and Health in the Houston Community of Manchester. Available at: www.ucsusa.org/sites/default/files/attach/2016/06/ucs-manchester-air-toxics-and-health-factsheet-2016.pdf, last access: 01 February 2024.
[5] Hersher and Schaped. (2017). Air Pollution from Industry Plagues Houston in Harvey’s Wake. National Public Radio, United States. Available at: www.npr.org/sections/health-shots/2017/09/14/550472740/air-pollution-from-industry-plagues-houston-in-harveys-wake, last access: 01 February 2024.
[6] Lindwall, C. (2023). Community Science is changing how people can fight pollution. The Natural Resources Defense Council. Available at: www.nrdc.org/stories/community-science-changing-how-people-can-fight-pollution, last access: 01 February 2024.
[7] Doris, M., Daley, C., Zalzal, J., Chesnaux, R., Minet, L., Kang, M., Hatzopoulou, M., 2024. Modelling spatial & temporal variability of air pollution in an area of unconventional natural gas operations. Environ Pollut 348, 123773. https://doi.org/10.1016/j.envpol.2024.123773.
[8] Hsu, C.Y., Xie, H.X., Wong, P.Y., Chen, Y.C., Chen, P.C., Wu, C.D., 2022. A mixed spatial prediction model in estimating spatiotemporal variations in benzene concentrations in Taiwan. Chemosphere 301, 134758. https://doi.org/10.1016/j.chemosphere.2022.134758.
[9] Jephcote, C., Mah, A., 2019. Regional inequalities in benzene exposures across the European petrochemical industry: a bayesian multilevel modelling approach. Environ Int 132, 104812. https://doi.org/10.1016/j.envint.2019.05.006.
[10] Agency for Toxic Substances and Disease Registry, United States. (2007). Toxicological Profile for Benzene - Health Effect. Available at: www.ncbi.nlm.nih.gov/books/NBK591289/, last access: 01 February 2024.
[11] Chiavarini, M., Rosignoli, P., Sorbara, B., Giacchetta, I., Fabiani, R., 2024. Benzene exposure and lung cancer risk: a systematic review and meta-analysis of human studies. Int J Environ Res Public Health 21 (2), 205. https://doi.org/10.3390/ijerph21020205.
[12] He, Y., Qiu, H., Wang, W., Lin, Y., Ho, K.F., 2024. Exposure to BTEX is associated with cardiovascular disease, dyslipidemia and leukocytosis in national US population. Sci Total Environ, 170639. https://doi.org/10.1016/j.scitotenv.2024.170639.
[13] Zahed, M.A., Salehi, S., Khoei, M.A., Esmaeili, P., Mohajeri, L., 2024. Risk assessment of benzene, toluene, ethyl benzene, and xylene (BTEX) in the atmospheric air around the world: a review. Toxicol Vitr, 105825. https://doi.org/10.1016/j.tiv.2024.105825.
[14] Peng, Z., Zhang, B., Wang, D., Niu, X., Sun, J., Xu, H., Shen, Z., 2023. Application of machine learning in atmospheric pollution research: a state-of-art review. Sci Total Environ, 168588. https://doi.org/10.1016/j.scitotenv.2023.168588.
[15] Tang, D., Zhan, Y., Yang, F., 2024. A review of machine learning for modeling air quality: overlooked but important issues. Atmos Res, 107261. https://doi.org/10.1016/j.atmosres.2024.107261.
[16] Sarker, I.H., 2021. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2 (3), 1–21. https://doi.org/10.1007/S42979-021-00592-X/FIGURES/11.
[17] Zhou, L., Pan, S., Wang, J., Vasilakos, A. v, 2017. Machine learning on big data: opportunities and challenges. Neurocomputing 237, 350–361. https://doi.org/10.1016/J.NEUCOM.2017.01.026.
[18] Ren, X., Mi, Z., Georgopoulos, P.G., 2020. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air

pollution: Modeling ozone concentrations across the contiguous United States. Environ Int 142, 105827. https://doi.org/10.1016/J.ENVINT.2020.105827.
[19] Arowosegbe, O.O., Röösli, M., Künzli, N., Saucy, A., Adebayo-Ojo, T.C., Schwartz, J., Kebalepile, M., Jeebhay, M.F., Dalvie, M.A., de Hoogh, K., 2022. Ensemble averaging using remote sensing data to model spatiotemporal PM_{10} concentrations in sparsely monitored South Africa. Environ Pollut (Barking, Essex: 1987), 310. https://doi.org/10.1016/J.ENVPOL.2022.119883.
[20] Huang, S.K., Chen, S.-Y., Chou, K.-L., Hsu, W.C., Lai, K.-H., Chueh, T.-H., Kuo, L., Lu, W., 2022. Optimizing the $PM_{2.5}$ tradeoffs: the case of Taiwan. Aerosol Air Qual Res 22 (10), 210315. https://doi.org/10.4209/AAQR.210315.
[21] Lai, W.I., Chen, Y.Y., Sun, J.H., 2022. Ensemble machine learning model for accurate air pollution detection using commercial gas. Sens Sens 22 (12), 4393. https://doi.org/10.3390/S22124393.
[22] Li, Z., Yim, S.H.L., Ho, K.F., 2020. High temporal resolution prediction of street-level $PM_{2.5}$ and NO_x concentrations using machine learning approach. J Clean Prod 268, 121975. https://doi.org/10.1016/j.jclepro.2020.121975.
[23] Liu, J., Chen, W., 2022. First satellite-based regional hourly NO_2 estimations using a space-time ensemble learning model: A case study for Beijing-Tianjin-Hebei Region, China. The Science of the Total Environment 820. https://doi.org/10.1016/J.SCITOTENV.2022.153289.
[24] Pintelas, P., Livieris, I.E., 2020. Special Issue on Ensemble Learning and Applications. Algorithms 13 (6), 140. https://doi.org/10.3390/A13060140.
[25] Huang, C., Sun, K., Hu, J., Xue, T., Xu, H., Wang, M., 2022. Estimating 2013–2019 NO2 exposure with high spatiotemporal resolution in China using an ensemble model. Environ Pollut 292, 118285. https://doi.org/10.1016/j.envpol.2021.118285.
[26] Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A.Z., Li, B., Zhang, H., 2021. Machine learning: new ideas and tools in environmental science and engineering. Environ Sci Technol. https://doi.org/10.1021/acs.est.1c01339.
[27] Babaan, J., Hsu, F.T., Wong, P.Y., Chen, P.C., Guo, Y.L., Lung, S.C.C., Chen, Y.C., Wu, C.D., 2023. A Geo-AI-based ensemble mixed spatial prediction model with fine spatial-temporal resolution for estimating daytime/night time/daily average ozone concentrations variations in Taiwan. J Hazard Mater 446, 130749. https://doi.org/10.1016/J.JHAZMAT.2023.130749.
[28] Wong, P.Y., Su, H.J., Lung, S.C.C., Wu, C.D., 2023. An ensemble mixed spatial model in estimating long-term and diurnal variations of $PM_{2.5}$ in Taiwan. Sci Total Environ 866, 161336. https://doi.org/10.1016/J.SCITOTENV.2022.161336.
[29] Brandsma, T., Können, G.P., 2006. Application of nearest-neighbor resampling for homogenizing temperature records on a daily to sub-daily level. Int J Clim: A J R Meteorol Soc 26 (1), 75–89.
[30] Barros, N., Tulve, N.S., Bailey, K., Heggem, D.T., 2019. Outdoor air emissions, land use, and land cover around schools on tribal lands. Int J Environ Res Public Health 16 (1). https://doi.org/10.3390/IJERPH16010036.
[31] Kheirbek, I., Johnson, S., Ross, Z., Pezeshki, G., Ito, K., Eisl, H., Matte, T., 2012. Spatial variability in levels of benzene, formaldehyde, and total benzene, toluene, ethylbenzene and xylenes in New York City: a land-use regression study. Environ Health: A Glob Access Sci Source 11 (1), 1–12. https://doi.org/10.1186/1476-069X-11-51/FIGURES/4.
[32] Houston-Galveston Area Council (2020), Houston-Galveston Area Land Cover Data. Retrieved from https://h-gac.sharefile.com/d-s50e70145aa9c464fad5e3a73d5587710.
[33] The City of Houston Geographic Information System (2021), Land Use Dataset, Retrieved from https://cohgis-mycity.opendata.arcgis.com/datasets/MyCity::coh-land-use/about.
[34] Olin, M., Kuuluvainen, H., Aurela, M., Kalliokoski, J., Kuittinen, N., Isotalo, M., Timonen, H.J., Niemi, J. v, Rönkkö, T., Dal Maso, M., 2020. Traffic-originated nanocluster emission exceeds H2SO4-driven photochemical new particle formation in an urban area. Atmos Chem Phys 20 (1), 1–13. https://doi.org/10.5194/ACP-20-1-2020.
[35] The City of Houston Geographic Information System (2020), COH AIRPORTS, Retrieved from https://cohgis-mycity.opendata.arcgis.com/datasets/860985f5c46f449b858056a9c5bbf9f8_19/explore?location=29.816953%2C-95.439800%2C10.14.
[36] Bendtsen, K.M., Bengtsen, E., Saber, A.T., Vogel, U., 2021. A review of health effects associated with exposure to jet engine emissions in and around airports. Environ Health 20 (1), 1–21. https://doi.org/10.1186/S12940-020-00690-Y.
[37] Eltarkawe, M., Thomas, G., Miller, S.L., 2022. Modeling county-level benzene emissions using transportation analysis zones in the Denver metro area. Atmos Environ: X 15, 100180. https://doi.org/10.1016/J.AEAOA.2022.100180.
[38] Whaley, C.H., Galarneau, E., Makar, P.A., Moran, M.D., Zhang, J., 2020. How much does traffic contribute to benzene and polycyclic aromatic hydrocarbon air pollution? Results from a high-resolution North American air quality model centred on Toronto. Can Atmos Chem Phys 20 (5), 2911–2925. https://doi.org/10.5194/ACP-20-2911-2020.
[39] The City of Houston Geographic Information System (2018), Houston, Texas Roads, Retrieved from https://cohegis.houstontx.gov/cohgispub/rest/services/EGIS/GeoCitizen_wm/MapServer/5.
[40] Rosebrook, D.D., Worm, G.H., 1993. Industrial sources of benzene exposure? Environ Health Perspect 101 (1), 13–16. Available at: ehp.niehs.nih.gov/doi/pdf/10.1289/ehp.9310113, last access: 01 February 2024.
[41] Megharaj, M., Naidu, R., 2017. Soil and brownfield bioremediation. Microb Biotechnol 10 (5), 1244. https://doi.org/10.1111/1751-7915.12840.
[42] ESRI. Understanding Euclidean distance analysis. ArcGIS Desktop: Release 10.7.1. Redlands, CA: Environmental Systems Research Institute. Retrieved from https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/understanding-euclidean-distance-analysis.htm#:~:text=The%20Euclidean%20distance%20output%20raster%20contains%20the%20measured%20distance%20from,cell% 20center%20to%20cell%20center. (Accessed in June 2023).
[43] Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415, 295–316. https://doi.org/10.1016/J.NEUCOM.2020.07.061.
[44] Hsu, C.Y., Lin, T.W., Babaan, J.B., Asri, A.K., Wong, P.Y., Chi, K.H., Ngo, T.H., Yang, Y.H., Pan, W.C., Wu, C.D., 2023. Estimating the daily average concentration variations of PCDD/Fs in Taiwan using a novel Geo-AI based ensemble mixed spatial model. J Hazard Mater 458, 131859. https://doi.org/10.1016/J.JHAZMAT.2023.131859.
[45] Molnar, C., 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub, United States.
[46] Bhuiyan, B.A., 2018. An overview of game theory and some applications. Philos Prog 111–128. https://doi.org/10.3329/PP.V59I1-2.36683.
[47] Chen, Y., Jia, Z., Mercola, D., Xie, X., 2013. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. Comput Math Methods Med 2013. https://doi.org/10.1155/2013/873595.
[48] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann Stat 29 (5), 1189–1232. https://doi.org/10.1214/AOS/1013203451.
[49] Rusdah, D.A., Murfi, H., 2020. XGBoost in handling missing values for life insurance risk prediction. SN Appl Sci 2 (8), 1–10. https://doi.org/10.1007/S42452-020-3128-Y/TABLES/5.
[50] Wang, Y., Chen, J., Chen, X., Zeng, X., Kong, Y., Sun, S., Guo, Y., Liu, Y., 2021. Short-term load forecasting for industrial customers based on TCN-LightGBM. IEEE Trans Power Syst 36 (3), 1984–1997. https://doi.org/10.1109/TPWRS.2020.3028133.
[51] Yang, S., Zhang, H., Yang, S., Zhang, H., 2018. Comparison of several data mining methods in credit card default prediction. Intell Inf Manag 10 (5), 115–122. https://doi.org/10.4236/IIM.2018.105010.
[52] Hancock, J.T., Khoshgoftaar, T.M., 2020. CatBoost for big data: an interdisciplinary review. J Big Data 7 (1), 1–45. https://doi.org/10.1186/S40537-020-00369-8/FIGURES/9.
[53] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2017. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 6638–6648. https://doi.org/10.48550/arxiv.1706.09516.
[54] Abedin, M.Z., Moon, M.H., Hassan, M.K., Hajek, P., 2021. Deep learning-based exchange rate prediction during the COVID-19 pandemic. Ann Oper Res 1, 52. https://doi.org/10.1007/s10479-021-04420-6.
[55] Parry, P. (2019). Automated machine learning for production and analytics: auto_ml.PyPI. MIT License. Retrieved from https://pypi.org/project/auto_ml/#files, last access: 01 February 2024.
[56] Elshawi, R., Sakr, S., 2020. Automated machine learning: techniques and frameworks. Lect Notes Bus Inf Process 390, 40–69. https://doi.org/10.1007/978-3-030-61627-4_3/FIGURES/6.
[57] Al Madhoun, W.A., Ramli, N.A., Yahaya, A.S., Yusuf, N.F.M., Ghazali, N.A., Sansuddin, N., 2011. Levels of benzene concentrations emitted from motor vehicles in various sites in Nibong Tebal. Malays Air Qual Atmosphere Health 4 (2), 103–109. https://doi.org/10.1007/S11869-010-0083-6/METRICS.
[58] Garg, A., Gupta, N.C., Tyagi, S.K., 2019. Levels of benzene, toluene, ethylbenzene, and xylene near a traffic-congested area of East Delhi. Environ Claims J 31 (1), 5–15. https://doi.org/10.1080/10406026.2018.1525025.
[59] Asri, A.K., Lee, H.Y., Chen, Y.L., Wong, P.Y., Hsu, C.Y., Chen, P.C., Lung, S.C.C., Chen, Y.C., Wu, C.D., 2024. A machine learning-based ensemble model for estimating diurnal variations of nitrogen oxide concentrations in Taiwan. Sci Total Environ 916, 170209. https://doi.org/10.1016/J.SCITOTENV.2024.170209.
[60] Marsland, S. (2014). Machine learning: An algorithmic perspective. Machine Learning: An Algorithmic Perspective, Second Edition, 1–452. doi.org/10.1201/B17476/MACHINE-LEARNING-STEPHEN-MARSLAND.
[61] Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecol Model 411, 108815. https://doi.org/10.1016/J.ECOLMODEL.2019.108815.
[62] Guo, C., Liu, G., Chen, C.H., 2020. Air pollution concentration forecast method based on the deep ensemble neural network. Wirel Commun Mob Comput 2020. https://doi.org/10.1155/2020/8854649.
[63] Requia, W.J., Di, Q., Silvern, R., Kelly, J.T., Koutrakis, P., Mickley, L.J., Sulprizio, M.P., Amini, H., Shi, L., Schwartz, J., 2020. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. Environ Sci Technol 54 (18), 11037. https://doi.org/10.1021/ACS.EST.0C01791.
[64] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min 785–794. https://doi.org/10.1145/2939672.2939785.
[65] Collins, C. (2020). Six Texas oil refineries are among the nation's worst benzene polluters, data shows - The Texas Observer. Available at: ⟨www.texasobserver.org/benzene-oil-refineries-texas-coast/⟩, last access: 01 February 2024.
[66] Olaguer, E.P. (2020). Overview of the benzene and other toxics exposure (BEE-TEX) Field Study. doi.org/10.1177/EHLS15654.
[67] Madl, A.K., Paustenbach, D.J., 2002. Airborne concentrations of benzene due to diesel locomotive exhaust in a roundhouse. J Toxicol Environ Health Part A 65 (23), 1945–1964. https://doi.org/10.1080/00984100290071487.
[68] Pronk, A., Coble, J., Stewart, P.A., 2009. Occupational exposure to diesel engine exhaust: a literature review. J Expo Sci Environ Epidemiol 2009 19 (5), 443–457. https://doi.org/10.1038/jes.2009.21.

[69] Han, X., Naeher, L.P., 2006. A review of traffic-related air pollution exposure assessment studies in the developing world. Environ Int 32 (1), 106–120. https://doi.org/10.1016/J.ENVINT.2005.05.020.
[70] Rich, A.L., Orimoloye, H.T., 2016. Elevated atmospheric levels of benzene and benzene-related compounds from unconventional shale extraction and processing: human health concern for residential communities. Environ Health Insights 10. https://doi.org/10.4137/EHI.S33314/ASSET/IMAGES/LARGE/10.4137_EHI.S33314-FIG3.JPEG.
[71] Wine, O., Vargas, A.O., Campbell, S.M., Hosseini, V., Koch, C.R., Shahbakhti, M., 2022. Cold climate impact on air-pollution-related health outcomes: a scoping review. Int J Environ Res Public Health 19 (3), 1473. https://doi.org/10.3390/IJERPH19031473/S1.
[72] Wongaree, M., Choo-In, S., 2019. Monitoring of benzene in an ambient air on the roadside at Udon Thani of Thailand. IOP Conf Ser: Earth Environ Sci 281 (1), 012007. https://doi.org/10.1088/1755-1315/281/1/012007.
[73] United States Environmental Protection Agency. Environmental Fact Sheet Final Standards Promulgated for Petroleum Refining Waste. (1998). retrieved from: www.epa.gov/osw (Acessed in January 2024).
[74] Lucas, M.R., 2002. Petroleum Refinery Source Characterization and Emission Model for Residual Risk Assessment. US Environmental Protection Agency: Office of Air Quality Planning and Standards, Research Triangle Park, NC, USA.
[75] Wallace, L.A., 1989. Major sources of benzene exposure. Environ Health Perspect 82, 165–169. https://doi.org/10.1289/EHP.8982165.
[76] Zhou, M., Xu, C., Xu, X., Li, X., 2023. Accident consequence assessment of benzene leakage from storage tank in a chemical park in Bengbu City, China. Process Saf Prog 42 (3), 440–447. https://doi.org/10.1002/PRS.12463.
[77] Texas Commission on Environmental Quality. (2023). TCEQ Data and Records Home. Available at: www.tceq.texas.gov/agency/data/lookup-data/download-data.html, last access: October 2023.
[78] Begley, D. (2018). If you thought traffic on these Houston highways are the worst, the state of Texas agrees. The Houston Chronicle. Available at: www.chron.com/news/houston-texas/transportation/article/Two-Houston-freeway-segments-top-list-of-13327848.php. last access: 01 February 2024.
[79] Briggs, J. (2023). Before the Highway: Houston, Texas. The American Association of Retired Persons. Available at: www.aarp.org/livable-communities/getting-around/info-2023/before-the-highway-houston-texas-rose-childress.html, last access: 01 February 2024.