JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(2), JUNE, 2025
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Deep Learning Models for Air Quality Prediction using Satellite Images

Shahida Abdullahi Umar, Badamasi Imam Ya'u, Fatima Umar Zambuk
Department of Mathematical Sciences, ATBU, Bauchi

ABSTRACT

According to World Health Organization, air pollution is considered to be one of the greatest environmental health threats. PM2.5, fine particles with a diameter that is generally 2.5 micrometers and smaller, is inhalable into the lungs and can induce adverse health effects. In order to mitigate the effects of PM2.5 on health outcomes, it is crucial to make accurate predictions of ambient concentrations. In the past, most studies developed traditional linear models from meteorological data or Environmental Protection Agency (EPA) ambient air quality monitors. However, the non-linear relationship between PM2.5 and other factors impacts the effectiveness of the models. Some other barriers are the sparseness of air quality monitors and the limited data sources, which also hinder researchers' ability to get accurate predictions. Recently, advanced technologies in satellite remote sensing have been widely used to estimate PM2.5 concentrations. The implementation of machine learning approaches has improved computational efficiency and accuracy. In this study, we used daily satellite imagery from January 2019 to October 2023 in 25 U.S. locations, and implemented an eXtreme Gradient Boosting (XGBoost) algorithm, a deep Convolutional Neural Network (CNN), and a CNN-XGBoost pipeline to make predictions based on the extracted features from each satellite image and atmospheric information along with meteorological conditions. To evaluate the performance of each model, we used daily EPA Federal Reference Method PM2.5 measurements as the validation data, and calculated the corresponding root mean squared error (RMSE) and the coefficient of determinant (R2). After combining each daily observation from 25 locations, the XGBoost approach demonstrated the highest performance with an RMSE of 3.98 μg m−3 and an R2 of 0.65. The CNN-XGBoost pipeline, tending to overestimate PM2.5 concentrations, had an RMSE of 5.87 μg m−3 and an R2 of 0.37. In conclusion, our study showed that XGBoost achieved reasonable PM2.5 prediction performance, indicating that the application of satellite remote sensing data and machine learning approaches has significant potential use in PM2.5 concentrations prediction.

ARTICLE INFO

Article History
Received: December, 2024
Received in revised form: February, 2025
Accepted: May, 2025
Published online: June, 2025

KEYWORDS

Deep learning, Satellite Images, Remote Sensing prediction, Air pollution, XGBoost, Machine Learning.

INTRODUCTION

Energy consumption and its consequences are inevitable in modern age human activities. Air pollution is a mixture of gaseous pollutants and Particulate Matter (PM), each of which has deleterious effects on human health. The anthropogenic sources of air pollution include emissions from industrial plants; automobiles; planes; burning of straw, coal, and kerosene; aerosol cans, etc. Various dangerous pollutants like CO, CO2, PM, NO2, SO2, O3, NH3, Pb, etc. are being released into our environment every day. Chemicals and particles constituting air pollution affect the health of humans, animals, and

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

even plants. Air pollution can cause a multitude of serious diseases in humans, from bronchitis to heart disease, from pneumonia to lung cancer, etc. In recent years, scientists applied satellite remote sensing technology to study climate change (Yang et al., 2013), surface air quality (Martin, 2008), and to estimate PM2.5 (Lin et al., 2015, Liu et al., 2005, Most studies used measurements of aerosol optical depth (AOD) or Moderate Resolution Imaging Spectroradiometer (MODIS) as predictors, and applied linear models to investigate the relationship between them and PM2.5. But the lack of accurate information on particle size distribution and composition leads to inaccurate predictions, suggesting a substantial amount of variability in PM2.5 concentrations cannot be explained by those models (Liu et al., 2005).

STATEMENT OF THE PROBLEM

Although machine learning algorithms have demonstrated their feasibility and accuracy in addressing nonlinear characteristics when using satellite images and data from other sources to make predictions, most of these algorithms have not been tested on weather data (Zheng et al., 2021). Due to complexity of the weather and variability of the seasons, meteorological measurements (temperature, dew point temperature, relatively humidity, surface pressure, precipitation and wind speed) and atmospheric data (AOD, cloud cover ratio, ground sampling distance of the image acquisition, water vapor concentration used, and ozone concentration used) have been identified as important factors in PM2.5 concentrations prediction (Stowell et al., 2020; Liu et al., 2019; Kleine et al., 2017).

Furthermore, to increase the accuracy of the prediction task, there is need for data from all locations were sensors are mounted to be considered (Zheng et al., 2021). This research work seeks to explore these angles in order to come up with a more accurate AQI prediction. This study intends to implement and evaluate the following two powerful methods to forecast daily PM2.5 concentrations of 25 locations in 17 states across the U.S.

Aim and Objectives of the Study

The ultimate goal is to improve air pollution forecasting using Machine Learning Algorithms by combining all observations from all locations recorded by the sensors. The specific objectives are to:

1. Use an XGBoost approach to model the association between PM2.5 and atmospheric data along with meteorological information.
2. Hybridize CNN-XGBoost pipeline, where a CNN will be used to process the extracted features that characterize the dynamic changes among the satellite imagery, and use XGBoost to model the association between PM2.5 and predictions from CNN atmospheric data along with meteorological information.
3. Evaluate the performance of the proposed models against the existing approach in (Zheng et al., 2021).

Scope of the Study

The study focuses on improving the work of Zheng et al., (2021) through two sets of experiments with air pollution dataset obtained from ULB Machine Learning Group, which is downloadable from https://www.kaggle.com. The first experiment will be using XGBoost, CNN and then secondly using CNN-XGBoost pipeline to further enhance the accuracy of the prediction task.

Conceptual Review

Previous research has worked on air pollution and adopted several methods/techniques in solving the problem. Gopalakrishnan (2021) combined Google's Street view data and ML to predict air quality at different places in Oakland city, California. He targeted the places where the data were unavailable. The author developed a web application to predict air quality for any location in the city neighborhoods. Sanjeev (2021) studied a dataset that included the concentration of pollutants and meteorological factors. The author analyzed and predicted the air quality and claimed that the Random Forest (RF)

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

classifier performed the best as it is less prone to over-fitting. The authors claimed to develop a novel method to model hourly atmospheric pollution. The authors reported that Adaptive Boosting (AdaBoost) and Stacking Ensemble are most suitable for air quality prediction but the forecasting performance varies over different geographical regions.

Air Pollution

According to Borghi (2021), pollutant sources can be divided into two groups: There are two types of sources: anthropogenic (man-made) and natural (natural). The natural sources of air pollution do not count much for the current increase in air pollution as compared to man-made sources. Anthropogenic sources, such as emissions from construction activities, industrial operations, fuel burning, and automobile pollutants, are the principal drivers of air pollution. Man-made pollution sources create sulphur, metal compounds, nitrogen, hydrogen, oxygen, and particulate matter, to name a few pollutants. Renewable energy sources like flat plate solar collectors, wind mills, etc. are also an important area to explore more for developing nations to achieve both economic growth and sustainable environment while reducing anthropogenic air pollution.

Modern approaches like ecological footprints, life cycle energy, etc. must are also nudging the society towards clean and sustainable environment. Natural pollution is caused by natural events that leak dangerous substances or have negative environmental repercussions like volcano eruptions, forest fires, etc. subfield of artificial intelligence called "machine learning" aims to develop techniques that let computers learn from data and get better at tasks without explicit programming. Machine learning algorithms provide predictions, classifications, or choices by processing input data, seeing patterns, and altering internal parameters (Jiang,2021, Lin et al.,2019).

Effects of air pollution

Air pollution affects the earth in various ways, from health to agriculture and economics. Each of them should be addressed separately.

Health effects of air pollution

The introduction of the chemicals mentioned above into our environment poses a threat to our health. Many illnesses are caused by toxins and certain people also die as a result of it. Air exposure affects a variety of respiratory issues, including difficulty breathing, wheezing, coughing, asthma, and heart complications. These factors have an effect on the human body and the body environment in general (Pi et al, 2019). The following is a list of the health consequences of air pollution, to be more specific:

1. Deaths
2. Cardiovascular disease
3. Lung disease
4. Lung cancer
5. Infants
6. Central nervous system

Several studies explored that due to extreme level of pollutants elderly and school children are the most affected age groups throughout the globe (Wong et al, 2004).

Gaps in the Literature

Based on the review, most of the approaches proposed in the literature employ the concept of traditional machine learning to predict air pollution. These approaches used less sophisticated machine learning algorithms which resulted in poor approximation and high variance estimation of model performance. The few researches that employed deep learning models require lots of hyper-parameter tuning and extensive feature selection to obtain optimal result.

It can also be noticed that the prediction models used didn't incorporate the use of metrological and environmental data. This has significantly impeded the model's performance. Developing models that are trained on data containing both metrological and environmental data will be a great research opportunity.

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

The study in Zheng et al., (2021) proposed ensemble deep convolutional neural networks and genetic algorithm to predict air quality. Promising predictive rate was recorded by the hybridised model. However, the research pays no attention to feature slicing in the feature selection of the model building process. This degraded the model performance (Rybarczyk & Zalakeviciute, 2021). Additionally, the model relies heavily on environmental data and treats all the features equally resulting in poor computational time and less accuracy (Gopalakrishnan, 2021)

METHODOLOGY

In this study, we will utilize XGBoost, CNN, and CNN-XGBoost hybridized model to predict daily PM2.5 concentrations in 25 locations across the U.S from April 2018 to September 2023. In designing our models, we will include daily measurements of PM2.5 from EPA monitors, satellite imagery and atmospheric information from Planet, and meteorological features from NASA POWER.

XGBoost algorithm will be used to model the association between PM2.5 and atmospheric data along with meteorological information, while the proposed CNN-XGBoost hybridized model, where a CNN is used to process the extracted features that characterize the dynamic changes among the satellite imagery, and an XGBoost is used to model the association between PM2.5 and predictions from CNN, atmospheric data along with meteorological information. CNN-XGBoost is a highly effective and versatile hybridized method that is capable of handling large-scale data and providing more accurate predictions.

Data Collection Technique

Satellite imagery data that we could obtain were limited by the Planet Scope Application Programming Interface (API). Among the locations with available satellite imagery, we will choose some of them for this study based on the 1) availability of daily data collected from EPA monitors, and 2) wide geographic distribution across the U.S. As a result, 25 locations across the U.S. will be selected because both PM2.5 measurements from EPA monitors and daily Planet Scope satellite images from surrounding area were abundant from 2017 to 2022.

Daily PM2.5 measurements from EPA Air Quality System (https://www.epa.gov/) across the U.S. from January 2017 to October 2022 will be downloaded from the repository. Those collected before 2017 will not be included since the Planet Scope satellite product was at the early stage of deployment and not working at full capacity.

The meteorological data will download from NASA POWER portal (https://power.larc.nasa.gov/data-access-viewer/). Meteorological data will be matched to image-PM2.5 pairs by corresponding location and date. The data collection that was employed by the researcher involved a group focus discussion. Three personnel were interview to gather accurate information about existing system and the requirements for the new system.

Architecture of the proposed model

In the proposed model, CPCB dataset is used alongside satellite images and XGBoost will be used to model the association between PM2.5 and predictions from CNN, atmospheric data along with meteorological information. CNN deep learning algorithm will be used to process the extracted features that characterize the dynamic changes among the satellite imagery. Finally, CNN-XGBoost pipeline, will be used train and build the model. This is depicted in Figure 2 below. Clearly, we can see in the existing model, since GA is used to extract the optimal parameters for the hyper parameter tuning, so handcrafted hyper parameter tuning is not needed and as such, the step for the task is not needed in the existing model.

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(2), JUNE, 2025
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

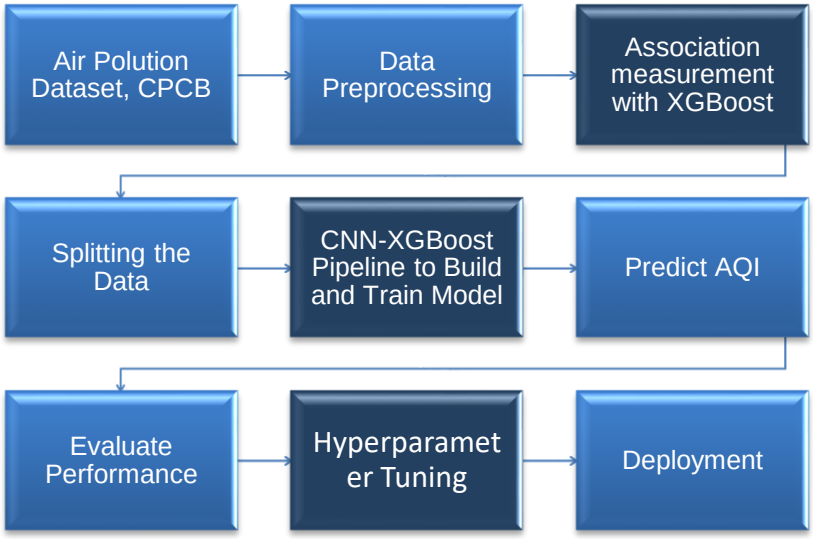

Figure 1: Framework of the proposed model

Performance Metrics

To access the performance of deep learning models, the commonly used performance indicators are; coefficient of correlation (R), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) (Fakharian et al., 2023; Rezazadeh et al., 2022). The pertinent expressions of R, MAPE, MAE, and MSE are shown in Equations 1 to 4.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\ (y_i - \bar{y})^2}} \quad \ldots (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{N} |x_i - y_i| \quad \ldots (2)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left|\frac{x_i - y_i}{x_i}\right| \times 100 \quad \ldots (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \quad \ldots (4)$$

Extreme Gradient Boosting for All Locations

XGboost is a gradient descent algorithm that minimizes the loss function by tuning parameters iteratively, where a loss function is used to measure how far an estimated value is from its true value. Common loss functions are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). XGBoost is a supervised learning algorithm that implements a boosting process to produce accurate predictions. Supervised learning refers to the task of inferring a predictive model from a set of labeled training examples. This predictive model can then be applied to new unseen examples or the testing examples. The data format of XGBoost input is usually a matrix with each row representing an observation and each column representing a feature.

In this analysis, we used vtreat package to convert the data frame into the required matrix format. We also incorporated 5-fold cross validations with 1,000 trees. The hyperparameters we tuned when implementing XGBoost were:

1. Maximum depth of a tree (max_depth): controls the complexity of the boosted ensemble. Default value is 6, and we set it to 6, 8, and 10.
2. Learning rate (eta): controls how quickly the algorithm proceeds down the gradient descent. If the learning rate is too small, the algorithm may take several iterations to find the minimum. If the learning rate is too large, it might jump across the minimum and end up

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(2), JUNE, 2025
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

further away than the starting point. Default value is 0.3, and we set it to 0.1, 0.2, and 0.3.

3. Minimum sum of instance weight needed in a child (min_child_weight): controls the minimum number of instances needed to be in each node, where a greater value leads to a more conservative algorithm. Default value is 1, and we set it to 2, 4, and 6.
4. Subsample ratio of the training instance (subsample): controls what proportion of the available training observations are used, where using less than 100% of observations means implementing stochastic gradient descent. This parameter is used to minimize overfitting and to avoid getting stuck in a local minimum or plateau of the loss function gradient. Default value is 1, and we set it to 0.6, 0.8, and 1.
5. Subsample ratio of columns (colsample_bytree): controls proportion of columns is used when constructing each tree. Default value is 1, and we set it to 0.6, 0.8, and 1.
6. To find the optimal parameters, we created a hyper parameters grid consisting of 243 different hyper parameter combinations (i.e. 3 possible values for each parameter→ run 35 = 243 models). The optimal tree booster we built had learning rate of 0.1, and maximum depth of each tree of 8. The minimum number of instances in each node was 6, with 80% of the training instance and 80% of columns of each tree. For XGBoost model, after reading in each image, we split it into 16 sections (shown in Figure 1) and extracted the mean and standard deviation of each band within each section. Thus, we obtained 128 summary statistics for each image (i.e. 16 sections ×4 bands × 2 summary statistics = 128).

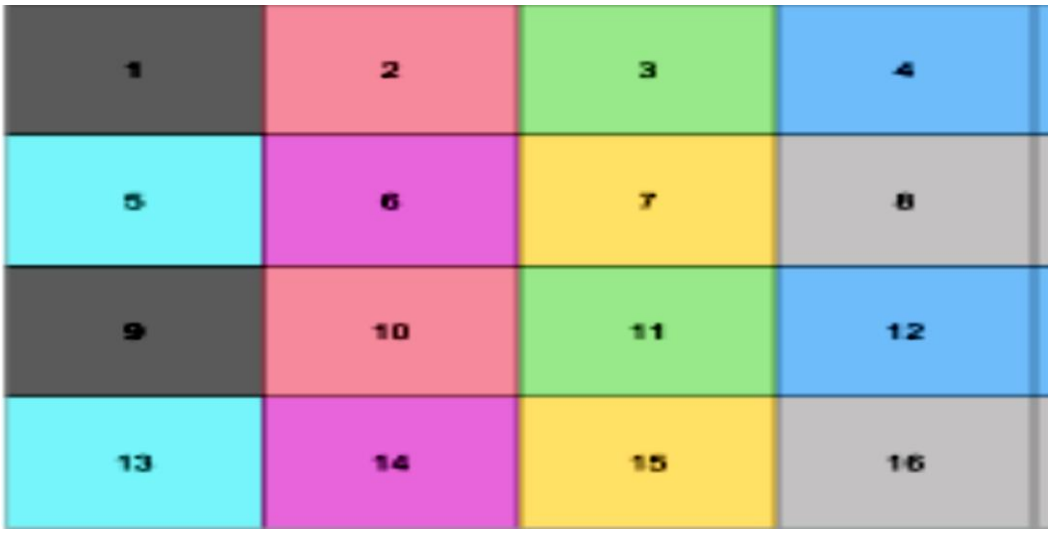

Figure 2: Illustration of Splitting an Image

PRESENTATION OF RESULTS

XGBoost

There were 31,565 available observations, where 18 observations with PM2.5 concentrations greater than 100 μg m−3. The 31,565 observations had an average PM2.5 of 9.08 μg m−3 (standard deviation (SD) = 7.30 μg m−3). The training set had an average PM2.5 of 9.12 μg m−3 (SD = 7.45 μg m−3). The testing set had a slightly lower average PM2.5 of 8.94 μg m−3

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

(SD = 6.68 µg m−3). The RMSE of testing set was 3.98 µg m−3 and R2 was 0.65.

Table 1 shows the RMSE of each month for the testing set. August had the highest average PM2.5 concentrations of 11.12 µg m−3 (SD = 9.68 µg m−3). It also had the second highest R2 of 0.76 and the highest RMSE of 4.73 µg m−3, following by December and November. Although it is not common to see high RMSE with high R2, August had a much higher standard deviation for PM2.5 concentrations than other months, indicating the observations were widely spread out from the average PM2.5 concentrations and thus leading to higher prediction errors. Also, since R2 is influenced by variance, as long as the RMSE is much smaller than the standard deviation of PM2.5 concentrations, we would have a high R2. June had the lowest RMSE of 2.95 µg m−3, with R2 of 0.43. Table 2 shows the RMSE of each year for the testing set. 2020 had the highest average PM2.5 concentrations of 9.51 µg m−3 (SD = 9.50 µg m−3). It had the highest R^2 of 0.75 and the highest RMSE of 4.75 µg m−3, following by 2019 and 2023. Similarly, although it seems contradictory to have high RMSE with high R2, 2022 had a much higher standard deviation than other years. 2021 had the lowest average PM2.5 concentrations of 7.92 µg m−3 (SD = 4.72 µg m−3). It had the lowest RMSE of 3.30 µg m−3 and the lowest R2 of 0.51. From these two tables, we concluded that when both the mean and the standard deviation of true PM2.5 concentrations were high, XGBoost model tended to have high RMSE but high R2 as well. Figure 1 is a scatter plot of the XGBoost predicted PM2.5 against EPA PM2.5, which shows a linear trend between predicted PM2.5 and EPA PM2.5. The fitted regression line (blue dashed line) overlapped with the 45◦ diagonal line (red solid line), suggesting that the XGBoost predictions were relatively consistent with the actual EPA measurements.

Table 1: XGBoost RMSE by month of testing set.

Month	#Obs	$mean(PM2.5)$ $[\mu g\ m^{-3}]$	$std(PM2.5)$ $[\mu g\ m^{-3}]$	$RMSE$ $[\mu g\ m^{-3}]$	R^2
January	397	8.53	4.99	3.87	0.41
February	360	9.04	6.78	4.24	0.61
March	492	7.05	5.03	3.83	0.42
April	539	7.43	4.39	3.24	0.45
May	537	7.51	4.26	3.09	0.47
June	640	7.81	3.92	2.95	0.43
July	699	9.90	5.93	4.10	0.52
August	651	11.12	9.68	4.73	0.76
September	618	9.62	9.26	4.23	0.79
October	482	8.84	7.32	4.33	0.65
November	448	9.47	6.49	4.35	0.55
December	449	10.45	6.65	4.50	0.54

Number of observations represents the number of observations in each month of testing set. mean(PM2.5) is the average PM2.5 of each month collected from EPA monitors during January 2019 to October 2023. std (PM2.5) is the standard deviation of PM2.5 of each month collected from EPA monitors during January 2019 to October 2023. RMSE is the root mean square error of each month. R2 is the R2 of each month.

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(2), JUNE, 2025
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Table 2: XGBoost RMSE by year of testing set.

Year	#Obs	$mean(PM2.5)$ $[\mu g\ m^{-3}]$	$std(PM2.5)$ $[\mu g\ m^{-3}]$	$RMSE$ $[\mu g\ m^{-3}]$	R^2
2019	818	8.87	5.84	4.03	0.52
2020	1427	9.22	5.70	3.78	0.56
2021	1493	7.92	4.72	3.30	0.51
2022	1450	9.51	9.50	4.75	0.75
2023	1124	9.28	6.05	3.93	0.58

Number of observations represents the number of observations in each year of testing set. Mean PM2.5) is the average PM2.5 of each year collected from EPA monitors during January 2019 to October 2023. std(PM2.5) is the standard deviation of PM2.5 of each year collected from EPA monitors during January 2019 to October 2023. RMSE is the root mean square error of each year. R2 is the R2 of each year.

Cloud Cover Issue

One of the major issues of the satellite-based methods is that they are dependent on availability of clear sky, and can provide clear images only when there are no clouds. Since PM2.5 concentrations cannot be estimated from satellite observations under cloudy conditions or bright surfaces such as snow or ice, we gathered all satellite imagery that had cloud coverage below 5%. To do this, we used one of the atmospheric information metrics, cloud, obtained from Planet. The 16,428 observations had an average PM2.5 of 9.02 µg m−3 (SD = 6.00 µg m−3). The training set had the same average PM2.5 as the overall data, and slightly higher standard deviation of 6.04µg m−3. The testing set had a slightly lower average PM2.5 of 9.00 µg m−3 (SD = 5.84 µg m−3). The RMSE of testing set was 3.67 µg m−3 and R2 was 0.60. Table 3 shows that eliminating all satellite imagery with cloud coverage above 5% was not the best approach since the resulting data set was not a good representation of the overall data.

For example, 80% of the observations from Yuma (CA) were selected, while only 19% of the observations from Broward (FL) were selected. Table 3 provides additional evidence. For example, 57% observations from June, July, September, and October were selected, while only 43% observations from January were selected.

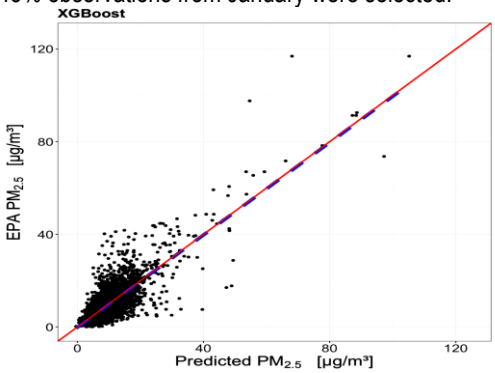

Figure 3: Scatter Plot of XGboost Predicted PM2.5 against EPA PM2.5.

The black points are the data points in the testing set. The red solid line represents the 45◦ diagonal line, while the blue dashed line represents the fitted regression line. Due to the complexity of terrains across the U.S., the atmospheric stability during winter time affects local PM2.5 concentrations variously (Karandana Gamalathge and Green, 2017). Table 2 shows that winter months (December, November, and January) had relatively high average PM2.5 values and variability. If the majority observations of winter months were removed due to high cloud coverage, the overall predictions would likely to be underestimated. Table 3 suggests that Broward (FL) had relatively low average PM2.5 values and variability. If the majority observations of Broward were removed, then the overall predictions would likely to be overestimated. To avoid the imbalance of data, we decided to use all observations for the following analysis.

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

Table 3: XGBoost Accounting for Cloud Cover Issue by Location.

Location	Total #Obs	#Obs with cloud < 0.05 (% of total obs)
Akron	1322	493 (37%)
Austin	1284	467 (37%)
Bronox	1247	595 (51%)
Broward	1372	265 (19%)
Cedar Rapids	1448	848 (59%)
Chicago	1188	531 (45%)
Essex	166	75 (45%)
Fort Collins	1230	822 (67%)
Los Angeles	1366	834 (61%)
Merced	1297	831 (64%)
Oakland	1505	768 (51%)
Oldtown	1425	587 (45%)
Orlando	1141	370 (32%)
Phoenix	1385	989 (71%)
Pittsburgh	1260	497 (39%)
Queens	1264	658 (52%)
Raleigh	1266	690 (55%)
Salt Lake	1545	834 (54%)
Seattle	1268	439 (35%)
St Louis	1433	781 (55%)
St Paul	1209	754 (62%)
Terre Haute	1316	655 (50%)
Timonium	948	416 (43%)
Tranqullity	1481	1087 (73%)
Yuma	1410	1133 (80%)

Number of observations represents the total number of observations in each location, collected from EPA monitors during January 2019 to October 2023. Number of observations with cloud < 0.05 represents the number of observations that have cloud coverage below 5% in each location, and its corresponding percentage of total number of observations.

Table 4: XGBoost Accounting for Cloud Cover Issue by Month

Month	Total #Obs	#Obs with cloud < 0.05 (% of total obs)
January	397	171 (43%)
February	360	164 (46%)
March	492	254 (52%)
April	539	288 (53%)
May	537	291 (54%)
June	640	366 (57%)

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(2), JUNE, 2025
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Month	Total #Obs	#Obs with cloud < 0.05 (% of total obs)
July	699	395 (57%)
August	651	310 (48%)
September	618	355 (57%)
October	482	273 (57%)
November	448	218 (49%)
December	449	200 (45%)

Number of observations represents the total number of observations in each month, collected from EPA monitors during January 2019 to October 2023. Number of observations with cloud < 0.05 represents the number of observations that have cloud coverage below 5% in each month, and its corresponding percentage of total number of observations.

Comparison of Two Models

After replacing NA's with 0's for those observations with missing bands information in certain sections of the satellite images, there were 33,562 observations. Table 5 shows the comparison of two models for each location. The RMSE of XGBoost was lower than that of CNN in every location, suggesting that XGBoost provided a more robust estimation. Merced (CA) and Tranquallity (CA) had relatively higher R2, but they also had relatively higher RMSE. Some locations had quite high CNN RMSE, so we did not include the CNN R2 in the table (i.e. those R2 can be negative and thus are not informative). Figure 2 shows 25 time series plots of CNN predicted PM2.5 against EPA PM2.5.

There were couple of negative predictions in almost every location, and some predictions were even below -10 μg m−3 in Austin (TX), Cedar Rapids (IA), Chicago (IL), and Essex (MD). In addition, the predictions of Fort Collins (CO), Oakland (CA), and Tranquillity (CA) were around 0. These unexpected behaviours suggested that the CNN model was not accurate and not able to capture the daily dynamic changes. Figure 4 shows 25 time series plots of XGBoost predicted PM2.5 against EPA PM2.5. In general, it performed better than CNN in every location and there was only one prediction below 0 (in Seattle, WA). It was able to capture most of the relatively high values, except the ones greater than 70 μg m−3 in Los Angeles (CA), Phoenix (AZ), and Tranquillity (CA).

Figure 4 shows 25 scatter plots of CNN predicted PM2.5 against EPA PM2.5, which indicates that most of the predictions overestimated the actual values. The majority of predictions were the same in Fort Collins (CO) and Tranquillity (CA), suggesting a different CNN architecture need to be used for depicting the changes. Figure 6 shows 25 scatter plots of XGBoost predicted PM2.5 against EPA PM2.5. While most of the predictions were consistent with the actual values, it was hard for XGboost to predict the relatively high values, such as the ones in Essex (MD), Los Angeles (CA), Oakland (CA), and Phoenix (AZ). All these tables and figures agree that XGBoost achieved better performances.

CNN-XGBoost Pipeline

The training set had an average PM2.5 of 9.08 μg m−3 (SD = 7.27 μg m−3). The testing set had a slightly higher average PM2.5 of 9.13 μg m−3 (SD = 7.42 μg m−3). The RMSE of testing set was 5.87 μg m−3 and R2 was 0.37. Figure 4-6 is a scatter plot of CNN-XGBoost predicted PM2.5 against EPA PM2.5, which shows a linear trend between predicted PM2.5 and EPA PM2.5. The fitted regression line (blue dashed line) was less steep than the 45◦ diagonal line (red solid line), suggesting that the CNN-XGBoost predictions overestimated the actual EPA measurements.

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

Table 5: Model Comparison for Each Location.

Location	#Obs	$mean\ (PM2.5)$	$std\ (PM2.5)$	$CNN\ RMSE$	$XGBoost\ RMSE$	$XGBoost\ R^2$
Akron	279	9.54	4.92	6.17	3.93	0.36
Austin	270	10.16	4.88	8.09	4.36	0.20
Bronox	249	8.22	4.39	6.58	3.88	0.22
Broward	291	7.04	4.16	10.68	3.59	0.26
Cedar Rapids	307	8.93	4.99	11.01	4.32	0.25
Chicago	252	8.50	4.69	6.81	4.14	0.22
Essex	33	8.65	7.44	11.69	6.62	0.21
Fort Collins	263	7.48	6.28	9.76	5.18	0.32
Los Angeles	291	13.16	11.43	12.37	10.99	0.08
Merced	276	13.94	14.14	10.25	9.52	0.55
Oakland	323	9.64	8.92	13.15	6.33	0.50
Oldtown	285	8.47	5.04	5.39	4.38	0.24
Orlando	245	6.92	3.18	5.63	3.11	0.04
Phoenix	293	8.54	8.52	7.97	7.94	0.13
Pittsburgh	265	14.22	9.18	9.78	7.70	0.30
Queens	265	7.14	4.48	5.10	4.27	0.09
Raleigh	269	8.82	3.73	4.84	3.68	0.03
Salt Lake	332	7.77	5.90	6.07	5.44	0.15
Seattle	267	6.29	6.30	6.31	6.12	0.06
St Louis	303	9.63	4.50	4.94	4.28	0.10
St Paul	262	7.50	4.54	8.37	4.18	0.15
Terre Haute	267	6.29	6.30	6.31	6.12	0.06
Timonium	200	8.89	4.82	6.68	4.38	0.17
Tranqullity	316	9.45	13.98	13.58	9.19	0.57
Yuma	298	8.86	3.98	5.26	3.91	0.03

Mean (PM2.5), std(PM2.5) and RSME were measured in [µg m^(-3)]. Number of observations represents the number of observations in testing set. mean (PM2.5) is the

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(2), JUNE, 2025
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

average PM2.5 concentrations in the testing set collected from EPA monitors during January 2019 to October 2023. std(PM2.5) is the standard deviation of PM2.5 concentrations in the testing set collected from EPA monitors during January 2019 to October 2023. CNN RMSE is the RMSE of the testing set from CNN model in each location. XGBoost RMSE is the RMSE of the testing set from XGBoost model in each location. XGBoost R2 is the R2 of the testing set from XGBoost model in each location. Note that the CNN R2's were not included in the table since those R2's were negative and thus are not informative.

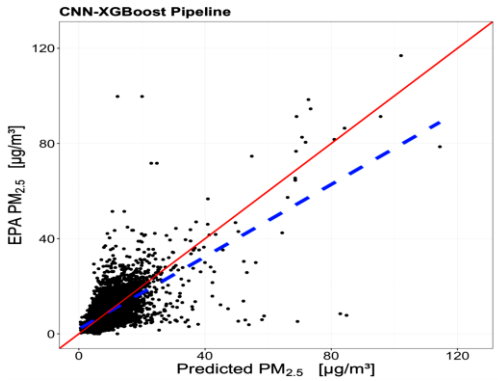

Figure 4: Scatter Plot of CNN-XGboost Predicted PM2.5 against EPA PM2.5.

The black points are the data points in the testing set. The red solid line represents the 45◦ diagonal line, while the blue dashed line represents the fitted regression line.

Comparison with Related Studies

Three machine learning techniques, including XGBoost, CNN, and CNN-XGBoost pipeline, were used to predict PM2.5 concentrations. The XGBoost technique demonstrated the highest performance and an acceptable time of training, with RMSE = 3.98 µg m−3 and R2 = 0.65. Within each location, the RMSE of XGBoost was lower than that of CNN, suggesting that XGBoost provided a more robust estimation. The CNN-XGBoost pipeline had an RMSE of 5.87 µg m−3 and an R2 of 0.37, and it tended to overestimate the actual PM2.5 measurements. Even though XGBoost outperformed the other models, it was not able to capture some of the relatively high values and variability, leading to lower R2. We compared our results with other studies

Table 6: Comparison of Model Performance with Other Studies.

Study	PM2.5 Mean (SD)	Model	RSME	R^2
Proposed Model	8.94 (6.68)	XGBoost	3.98	0.65
	9.13 (7.42)	CNN-XGBoost	5.87	0.37
Di, Kloog, et al., 2016		CNN	2.94	0.84
Wang et al., 2017		LME	24.5	0.86
Johnson, Bonczak, and Kontokosta, 2018	7.8	OLS	3.11	0.507
		RR	3.07	0.521
		GBRT	2.16	0.762
Zamani Joharestani et al., 2019	86.8 (33)	RF	14.47	0.78
		XGBoost	13.62	0.80
		ANN	14.56	0.77
Mukherjee et al., 2019		MLR		0.69
Datta et al., 2020	8.0 (6.0)	MLR	1.9	
T. Zheng et al., 2020	42.7 (42.9)	VGG16-RF	16.3	0.86

Abbreviations:

Linear Mixed Effect (LME); Random Forest (RF); Artificial Neural Network (ANN); Multiple Linear Regression (MLR); Ordinary Least Squares (OLS); Ridge Regression (RR); Gradient Boosting Regression Tree (GBRT); Visual

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

Geometry Group that supports 16 layers (VGG16).

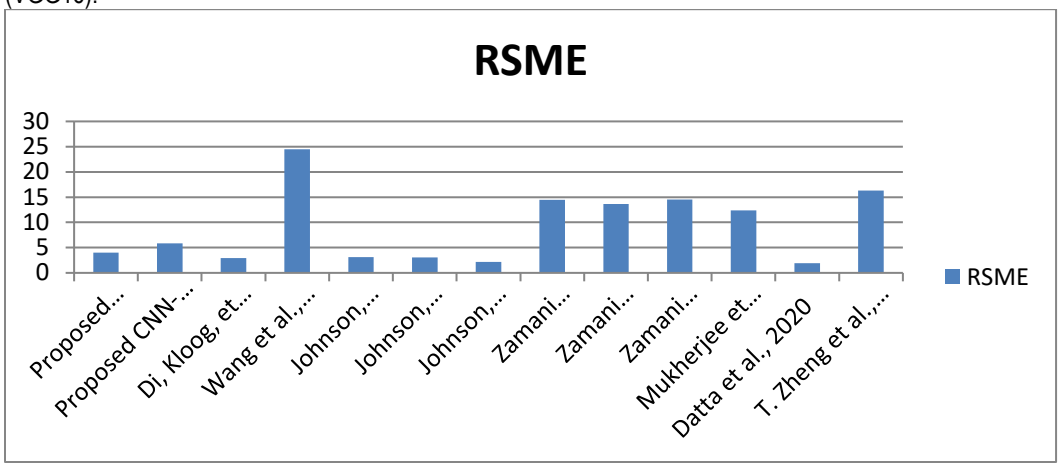

Figure 5: RSME Comparison of Model Performance with Other Studies.

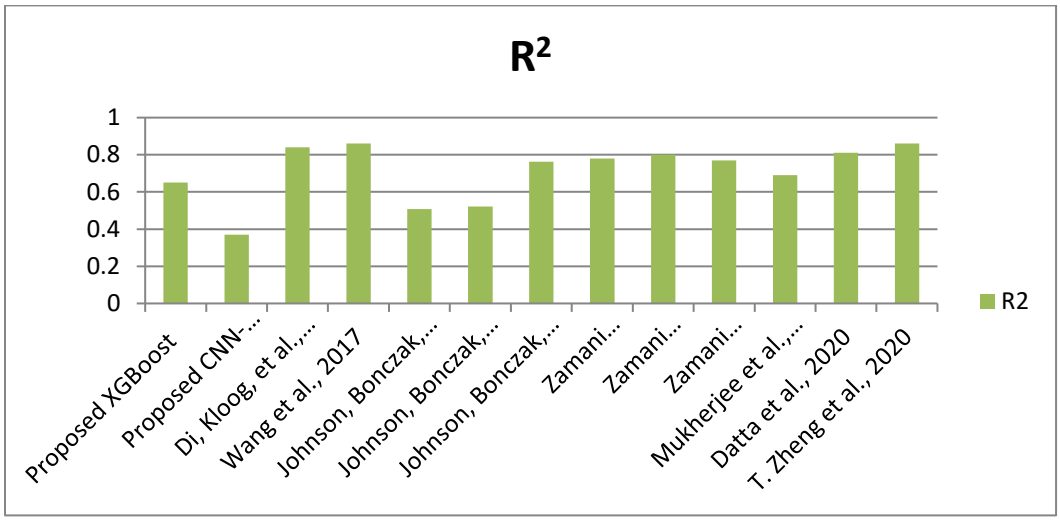

Figure 6: R^2 Comparison of Model Performance with Other Studies.

As seen in Table 5, the related works can be summarized to three main categories. In the first category, besides meteorological data, the satellite-derived data such as AOD products, were used to find a relationship between the AOD and PM2.5 measurements through an appropriate model. In the second category, low-cost sensors and meteorological data were used to improve the spatial and temporal resolution of $PM_{2.5}$ through some calibration methods. The third category, including our study, used machine learning approaches to make predictions of $PM_{2.5}$ concentrations based on satellite images and meteorological data. The most widely used method for $PM_{2.5}$ estimations from the satellite data is finding the relationship among the satellite-derived aerosol optical depth (AOD).

Wang et al., 2017 used AOD through an LME, and the RMSE was 24.5 μg m−3 and R^2 was 0.86. Even though the model performed well, their method was developed under clear-sky conditions and required filtering cloudy pixels. Zamani Joharestani et al., 2019 used AOD and conducted

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

RF, ANN, and XGBoost models to predict $PM_{2.5}$ of Tehran's urban area in Iran. They used interpolation to estimate and fill in the missing data. In addition, they applied standard normalizations to reduce the instability during the model training. XGBoost outperformed with RMSE of 13.62 μg m−3 and R^2 of 0.80, following by RF with RMSE of 14.47 μg m−3 and R^2 of 0.78. ANN had a slightly higher RMSE of 14.56 μg m−3 and a slightly lower R^2 of 0.77.

Since the baseline $PM_{2.5}$ concentrations in Tehran were much higher, the RMSE's were higher than our study's. But the high R^2's suggested that their models had better fits. Some additional features used in this analysis, such as latitude, longitude, altitude, day of week, and season, were identified to be important variables and may be utilized in our study as well. They set daily EPA measurements as reference, and included multiple input variables such as AOD data, surface reflectance, chemical transport model outputs, meteorological data, aerosol index data, land-use terms, regional and monthly dummy variables, and scaling factor. In addition, they filled in any missing values by a neural network or a linear interpolation.

A CNN model was used, where convolutional layers were implemented to account for the temporal and spatial autocorrelation. Their hybrid nationwide model achieved high performance with an average total R^2 of 0.84 (ranged from 0.74 to 0.88) and an average RMSE of 2.94 μg m−3 (ranged from 2.64 μg m−3 to 3.58 μg m−3). In terms of season and region, the model performed the best in summer and in the Eastern U.S.. Various types of variables helped the model achieve higher prediction accuracy, yet limited its application to regions or countries with fewer data available.

Zheng et al., 2020 employed a VGG16 algorithm for feature extraction from satellite images of Beijing from 2017 to 2019, and the extracted image features were given to a RF as input to estimate the $PM_{2.5}$ concentrations. The RMSE of the best model was 16.3μg m−3 and R^2 was 0.86. To validate the models, they used the same pipeline to make predictions in Shanghai, where the RMSE was 10.8 μg m−3 and R^2 was 0.87. Their approach achieved quite high prediction accuracy since they restrained themselves to only the images on approximately uncloudy days (*i.e.* cloud coverages below 5%). They also resized all their images to 224 × 224 pixels and subtracted the mean RGB values of the original ImageNet training set from each pixel, which made the model more flexible in being adopted to other locations.

Limitations and Future Work

Some limitations remain in this study. Although XGBoost achieved reasonable overall $PM_{2.5}$ prediction performance, the RMSE varied from 2.95 μg m−3 to 4.73 μg m−3 and R^2 varied from 0.40 to 0.76 in different months, indicating some seasonal patterns were not fully captured. Also, the CNN architecture used in this study was not optimal, since some of the predictions were negative.

Future studies will explore the possibility of adding other components, such as seasons and day of the week, as covariates. Studies have shown that relatively high $PM_{2.5}$ concentrations tend to appear on Fridays and Saturdays, while on Mondays and Sundays the $PM_{2.5}$ concentrations are generally lower than those on most other days of the week (Zhao et al., 2018). To eliminate the negative predictions from CNN architecture, we could take the logarithm of the outcomes. Also, we will find a more appropriate approach for dealing with the missing pixels in the satellite imagery, such as a Random Forest imputation or interpolation. In addition, we will investigate the application of a CNN architecture using ResNet with 50 layers (K. He et al., 2016) or a VGG16 architecture (T. Zheng et al., 2020).

CONCLUSION

In this study, we utilized XGBoost, CNN, and CNN-XGBoost pipeline to predict daily $PM_{2.5}$ concentrations in 25 locations across the U.S from January 2019 to October 2023. In designing our models, we included daily measurements of $PM_{2.5}$ from EPA monitors, satellite imagery and atmospheric information from Planet, and meteorological features from NASA POWER. Within each location, XGBoost provided more

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.

accurate estimates while CNN led to overestimation. Even though CNN resulted in lower prediction accuracy, it is flexible and is able to process the satellite imagery directly. It can detect the informative features and learn the dynamic changes without any human supervision. When we combined all observations from 25 locations, in comparison to CNN-XGBoost pipeline, XGBoost had a better performance of RMSE = 3.98 μg m−3 and R^2 = 0.65. Since XGBoost is designed with faster training speed and lower memory usage, it takes less time than CNN to produce outputs. It is a highly effective and versatile method that is capable of handling large-scale data and providing more accurate predictions.

REFERENCES

Borghi, Francesca et al. (2021). "Estimation of the Inhaled Dose of Pollutants in Different Micro-Environments: A Systematic Review of the Literature." In: Toxics 9.6, p. 140.

Gao, Meiling, Junji Cao, and Edmund Seto (2015). "A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2. 5 in Xi'an, China." In: Environmental pollution 199, pp. 56–65.

Guo, Bin et al. (2021). "Estimating PM2. 5 concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017." In: Science of The Total Environment 778, p. 146288.

Johnson, Nicholas E, Bartosz Bonczak, and Constantine E Kontokosta (2018). "Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment." In: Atmospheric environment 184, pp. 9–16.

Lin, Changqing et al. (2015). "Using satellite remote sensing data to estimate the highresolution distribution of ground-level PM2. 5." In: Remote Sensing of Environment 156, pp. 117–128.

Liu, Wei et al. (2019). "Meteorological pattern analysis assisted daily PM2. 5 grades prediction using SVM optimized by PSO algorithm." In: Atmospheric Pollution Research 10.5, pp. 1482–1491.

Stowell, Jennifer D et al. (2020). "Estimating PM2. 5 in Southern California using satellite data: factors that affect model performance." In: Environmental Research Letters 15.9, p. 094004.

Wang, Wei et al. (2017). "Deriving hourly PM2. 5 concentrations from himawari-8 aods over beijing–tianjin–hebei in China." In: Remote Sensing 9.8, p. 858.

Yang, Jun et al. (2013). "The role of satellite remote sensing in climate change studies." In: Nature climate change 3.10, pp. 875–883.

Zheng, Tongshu et al. (2020). "Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach." In: Atmospheric Environment 230, p. 117451.

Zhou, Xiaodan et al. (2021). "Excess of COVID-19 cases and deaths due to fine particulate matter exposure during the 2020 wildfires in the United States." In: Sci

Corresponding author: Shahida Abdullahi Umar
abdullahishahida10@gmail.com
Department of Mathematical Sciences, ATBU, Bauchi.