Satelite PM2.5 Retrieval Using Machine Learning Models-Maryland Case Studies

Fikewa Akindolire*, Gracelyn Arunachalam†, Elton Mawire‡, Sogidechukwu Unegbu§, Elainia Ross-Jones¶, Temidayo Fapohunda*, and Daniel Okon*, and Xiaowen Li‖
*Morgan State University, Baltimore, MD 1700, USA
†University of Maryland, College Park, MD 20742, USA
‡Alabama Agricultural and Mechanical University, Normal, AL 35762, USA
§Community College of Baltimore County, MD 21237, USA
¶Trinity Washington University, Washington, DC, 20017, USA
‖University of Chicago, Chicago, IL 60637, USA
fbv

I. Abstract

Particulate Matter 2.5 (PM2.5) is one of the deadliest air pollutants, posing a significant threat to human respiratory health. In this research, we collected Aerosol Optical Depth (AOD) data from satellite observations and merged it with PM2.5 and meteorological data from ground stations located in Howard and Padonia. This integrated dataset was used to train and evaluate several machine learning models, including Random Forest, Gradient Boosting, and a Long Short-Term Memory (LSTM) neural network. The Gradient Boosting model achieved the highest performance, reaching an R^2 score of 0.83, indicating strong predictive accuracy. These results highlight the value of combining satellite and ground-based data with advanced learning algorithms to enhance PM2.5 forecasting in urban environments and support public health decision-making.

***Index Terms*—aerosols, python, PM2.5, Random Forest, Gradient Boost, Neural Networks, satellite**

II. Introduction

Even though air pollution may seem like it is visible to the human eye as smoke, pollution exists in the form of invisible solid or liquid partcles in the air known as aerosols. Among other aerosols in the atmosphere, PM2.5 (Particulate Matter with a radius of 2.5 microns or less) is considered one of the deadliest pollutants. Small particles like these can accumulate in the trachea and the lungs and cause asthma and other lung/respiratory diseases. These pollutants are released into the atmosphere through natural and anthropogenic sources. Natural sources include tornadoes, dust storms, etc., while anthropogenic sources include construction activities, mining sites, etc. Once released into the air these particles linger in the air and end up in sinks, which are usually clouds [1] or on the ground through rain droplets. Despite their lethal properties, these aerosols help in the process of cloud and rain formation and also regulate the temperature in the atmosphere by deflecting UV rays from the Sun [2].

A. Problem Statement

Regulating such pollutants in the atmosphere is vital; however, monitoring it has posed difficulties. Ground LiDAR stations can accurately measure the concentration of PM2.5 within a limited radius of their location. However, the installation and maintenance of these stations are expensive, and thus there is a limited number of ground stations in the US. This limits the ability to predict the amount of PM2.5 in some regions in the United States. The only other option available to us is to use NASA satellites to measure the amount of contamination in the atmosphere. However, this method is less accurate because of clouds that could interfere with the satellite measurements [3]. Our project aims to use the availability of satellite data and the accuracy of ground stations to predict the amount of PM2.5 given satellite AOD (Aerosol Optical Depth) data.

B. Motivation

With advancements in AI and ML, there is hope in being able to predict the PM2.5 concentrations in a specific place given AOD data from satellites. ML models are experts at finding patterns and relationships between two variables. Incorporating ML in this project will make predictions accurate and reliable.

This paper describes the use of AI in the field of climate science to predict the amount of pollutants in urban atmospheres specifically Howard and Cockeysville (Padonia) given satellite data. This paper documents the background, methodology, results and analysis and conclusion of the project and will focus on data from two specific stations, the Howard and Padonia Stations.

LaTeX. Please observe the conference page limits.

III. Background

Particulate Matter 2.5 (PM2.5) is one of the most dangerous classes of air pollutants due to its ability to linger invisibly in the atmosphere for extended periods—often hundreds of days—unlike larger particles (e.g., =5 micrometers) that settle more quickly [4]. PM2.5 poses a serious threat to respiratory

health and the environment. Urban areas, particularly low-income and historically marginalized communities, face elevated risks due to sustained exposure [5].

A study by some authors [6] highlights this environmental injustice by revealing that communities in cities like Baltimore continue to be exposed to pollutants from coal mine emissions. The study also underscores the contribution of mobile sources, such as trucks, where frequent stop-and-go traffic—especially at traffic lights—can result in higher localized PM2.5 levels, even in areas farther from the pollution source [6].

Understanding the transport and transformation of aerosols is critical to comprehending their health impact. A recent study by several researchers [7] challenges long-held beliefs about natural removal mechanisms, revealing that clouds can act not only as sinks (removal agents) but also as sources of aerosols under certain conditions. If natural phenomena like precipitation are less effective at aerosol removal than previously assumed, then real-time modeling and prediction of pollutants—especially PM2.5—become even more imperative.

One of the significant challenges in PM2.5 prediction is the lack of widespread, high-resolution ground monitoring stations, which provide the most accurate data [8]. On the other hand, satellite-derived Aerosol Optical Depth (AOD) offers broader spatial coverage but less accuracy. AOD measures the degree to which aerosols prevent the transmission of light and can be used as a proxy to estimate PM2.5 levels.

To bridge this gap, many studies have developed hybrid or data-driven models that combine ground PM2.5 measurements with AOD and meteorological data to train machine learning models. These range from ensemble methods to spatiotemporally weighted deep learning models. For instance, [9] some authors demonstrated that ensemble models like Random Forest and Gradient Boosting perform well in urban PM2.5 prediction, achieving an R^2 value of 0.81. [10] Others further highlight that deep learning models such as LSTM can yield even higher prediction performance.

In this study, we build on these insights by applying Random Forest, Gradient Boosting, and LSTM models to merged satellite (AOD) and ground-based PM2.5 datasets from Howard and Padonia. Our aim is to explore and improve the predictive capability of these models for PM2.5 concentrations in urban settings.

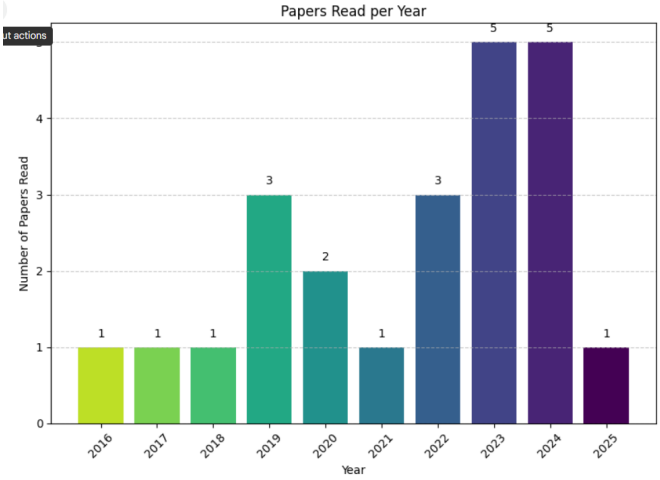

Fig. 1. Literature covered

IV. Methodology

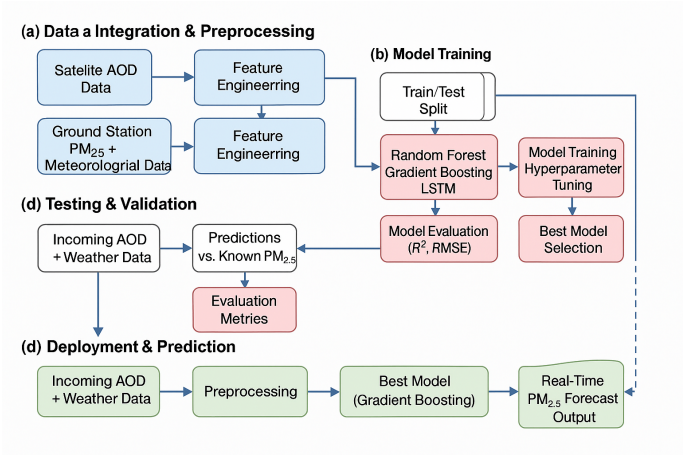

Fig. 2. Flow Chart of the Machine Learning Pipeline.

A. Data Collection and Preprocessing

1) Satellite Data
We primarily focused on analyzing Particulate Matter 2.5 (PM2.5) concentrations in Baltimore from 2019 to 2022 at the Padonia and Howard Lidar Station to maximize the utility of our data for our machine learning models and the extraction of critical information regarding PM2.5 concentrations. Using satellite data to capture Aerosol Optical Depth (AOD) at 550 nm, we were able to predict PM2.5 in the two specified locations accurately. The satellite data was acquired from both NASA Giovanni and Dark Target. The NASA Dark Target Retrieval algorithm is able to measure AOD from 342.5 nm to 865 nm; however, we chose 550 nm because of its ability to measure observable light in the electromagnetic spectrum. Although the retrieval of AOD can

be sourced from ground-level lidar stations, it has its restrictions, such as limited coverage, being expensive, and the accessibility of ground-based facilities in cities or densely populated areas. For that reason, we decided that satellite-derived AOD would be the most productive option because of the strong correlation with PM2.5.

2) Ground-Based and Meteorological Data
We used ground-station PM2.5 data from both Padonia and Howard, serving as a comparison of the actual vs the predicted data from our machine-learning models. This data spans the period from 2019 to 2022, located in the Baltimore Metropolitan area. We integrated temperature, wind speed, relative humidity, and meteorological data acquired from the Maryland Department of the Environment (MDE), because they influence the dispersion and accumulation of pollutants.

3) Satellite Data Before selecting our models, we cleaned and preprocessed our data to be understood and effectively used by our machine learning models. Data cleaning included removing "bad" or unusable data by dropping NaN (missing values), placing conditions on values out of range (e.g., negative temperature), and removing duplicate records. After cleaning, we preprocessed the data, preparing the data for modeling. This included converting objects into datetime format, converting strings into integers and floats, merging data sets, renaming column names for consistency across all data frames, and interpolating data. These steps are essential in ensuring the model's predictive accuracy.

B. Machine Learning Models

1) Model Selection

An assortment of machine learning models was chosen to predict PM2.5 concentrations from AOD and meteorological/ground station data.The models are as follows:

Random Forest (RF) - According to [11], "RF is an ensemble learning method used for classification and regression. Developed by Breiman (2001), the method combines Breiman's bagging sampling approach (1996a), and the random selection of features, introduced independently by Ho (1995; 1998) and Amit and Geman (1997), to construct a collection of decision trees with controlled variation."

Gradient Boosting (GBR) - According to [12], "Gradient Boosting serves as a baseline model, building weak learners sequentially to minimize the overall loss."

Long Short-Term Memory (LSTM) - According to [2],, "Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), are renowned for their efficacy in capturing intricate 4 temporal dependencies within sequential data."

FeedForward Neural Network (FNN) - According to [12], "FNN is a parametric model composed of layers of interconnected nodes, where the primary learnable parameters are weights and biases. Training involves the forward pass, where inputs are transformed through hidden layers using activation functions (typically ReLU), and the output layer produces classification probabilities."

C. Model Training

The model is trained using the features (x), the input variables (e.g., temperature, wind speed, and relative humidity), and the target (y), which is the output variable (e.g., PM2.5 concentrations). We then split the data, using 80% to train the model and 20% to evaluate the learning performance, (Fig. 3.) Using the defined X (X_train) and Y (Y_train) values, you can now input the line of code .fit(). This is how the model learns from our training data. In this, the model is building small decision trees with each tree trying to correct the errors from the previous one, constantly fine-tuning itself to get a better prediction of PM2.5, (Fig. 4.) After the model is trained, we can now give it new data (AOD and Meteorological data), and it will predict the PM2.5 concentrations.

```
# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Fig. 3. Split Data.

```
# Fit the model
model.fit(X_train, y_train)
```

Fig. 4. Model Fitting.

D. Model Evaluation (RSME, MSE, MAE, R^2)

To evaluate the performance of our models, we used evaluation metrics. The metrics include: Mean Squared Error (MSE), which measures how far off our predictions are on average; Root Mean Square Error (RMSE), which is the square root of MSE, written in the same unit as PM2.5 (e.g., plus or minus 3.2 micrograms per cubic meter); Mean Absolute Error

(MAE), calculates the average absolute difference between the predicted and actual values; R$\hat{2}$ Score (R-squared), explains the variation in the data. [htbp]

$$\mathbf{R^2:}\quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\mathbf{MSE:}\quad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\mathbf{RMSE:}\quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\mathbf{MAE:}\quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Evaluation Metrics Equation.

V. Results and Analysis

The results are presented in both tabular and graphical formats (Scatterplot and Bar charts). These results compare the performance of various AI models in both the Howard County dataset and the Padonia station dataset.

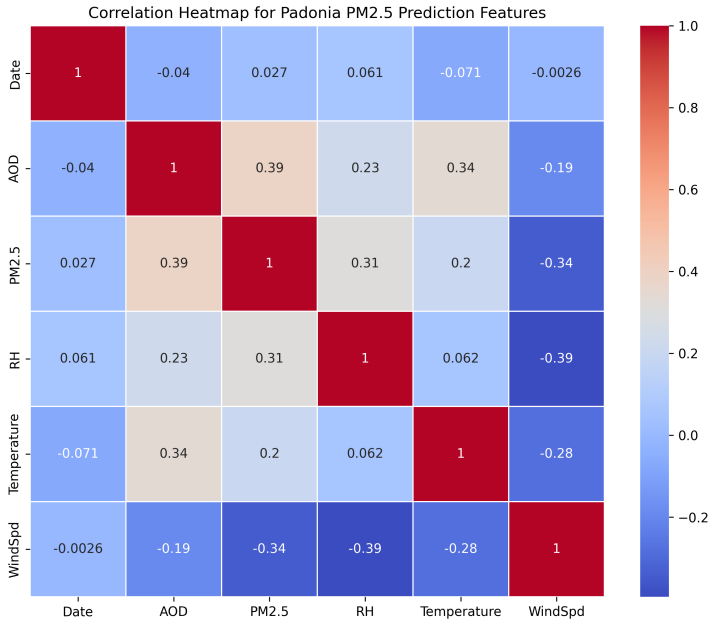

Fig. 5. Correlation Heatmaps (Padonia Station).

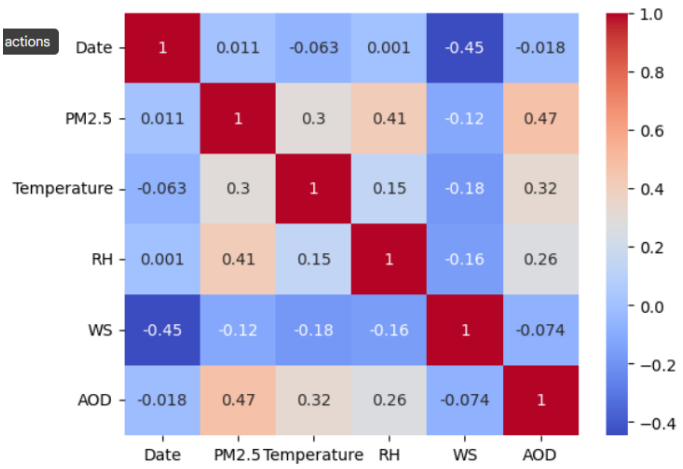

Fig. 6. Correlation Heatmaps (Howard County).

The image in Figure 5 and 6 is a correlation matrix. This tells us about the relationships of the features at a glance. In a heatmap:

1) A positive number show that the variables are directly proportional
2) A negative number shows that the variables are inversely proportional
3) A zero means no relation and a one means a perfect relation.

With a look at the heatmap, we can tell that each feature has some relationship with one another. The results of the model training agrees with this.

TABLE I
AI Accuracy comparison (Howard County Station)

AI Model	**Accuracy parameters (4 sig figs)**			
	r2 score	***MAE***	***MSE***	***RMSE***
Random forest	0.7012	1.7382	5.7293	2.3936
Gradient Boosting	0.8430	2.8293	1.6821	1.2163
LSTM	0.7694	0.0314	0.0022	0.0468
FNN	0.8033	1.7229	1.1163	2.9683

TABLE II
AI Accuracy comparison (Padonia Station)

AI Model	**Accuracy parameters**			
	r2 score	***MAE***	***MSE***	***RMSE***
Random forest	0.73	1.8	7.1	2.67
Gradient Boosting	0.81	1.5	5.0	2.25
LSTM	0.76	1.6	5.1	2.26
FNN	0.81	1.1	12.6	1.7

Tables 1 and 2 show us the AI accuracy comparison for Howard and Padonia stations, respectively. With this we can see that the models got good R2 scores throughout and the error parameters were on the lower end.

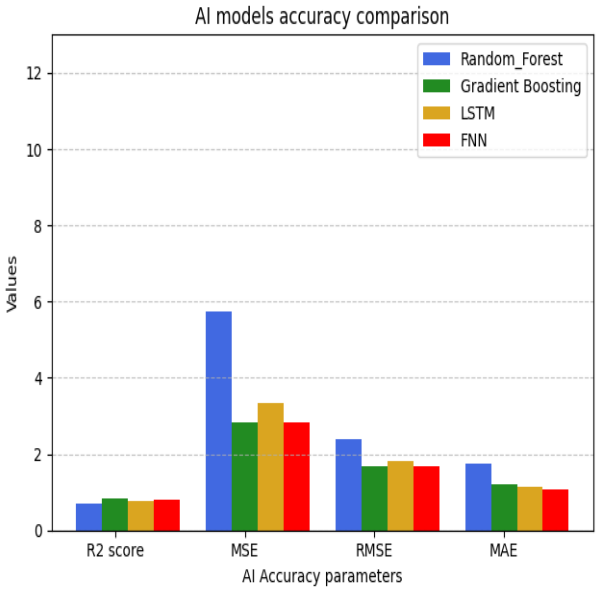

Fig. 7. AI models comparison(Howard County).

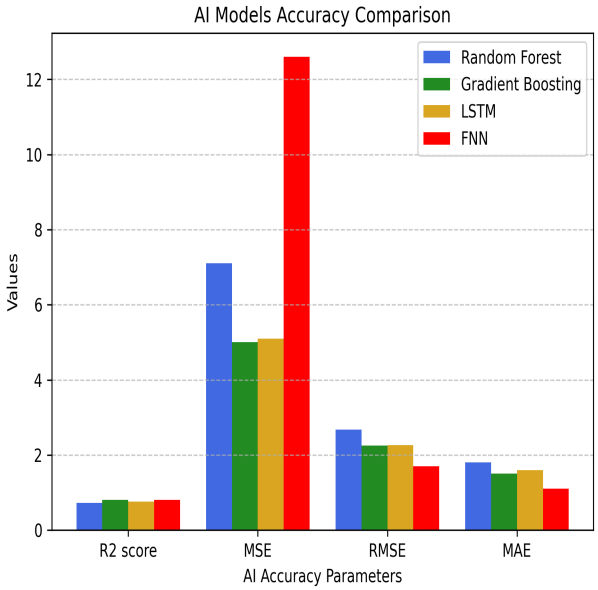

Fig. 8. AI models comparison(Padonia Station).

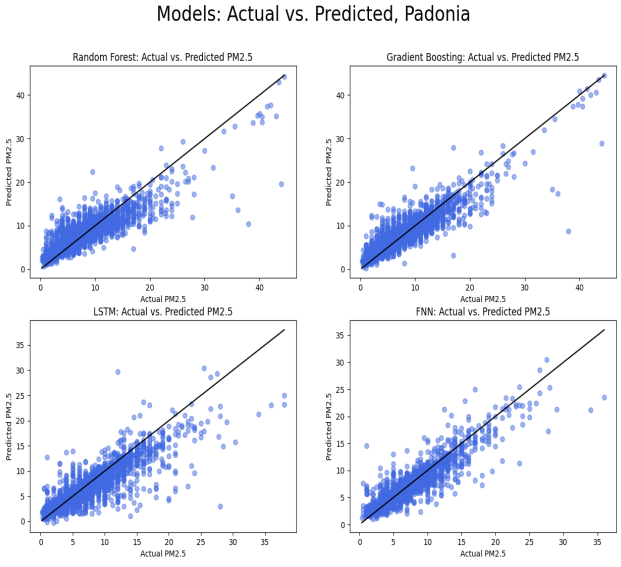

Fig. 9. Models Actual Vs Predicted PM2.5 (Padonia).

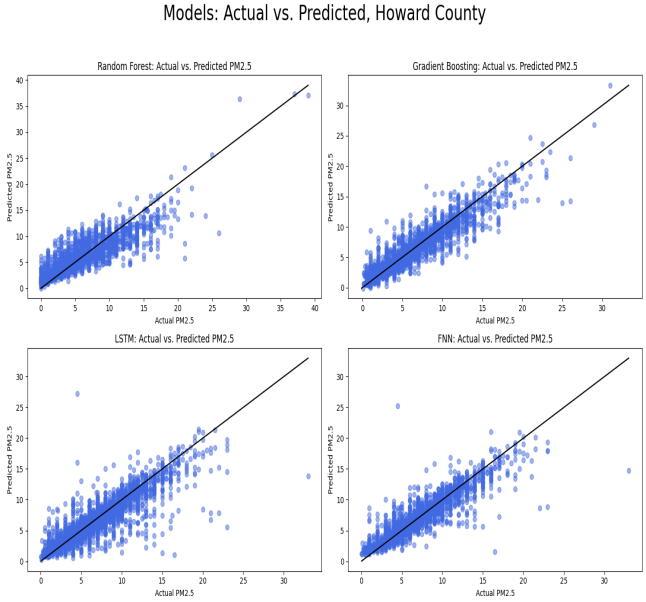

Fig. 10. Models Actual Vs Predicted PM2.5(Howard County).

Graphs 7 and 8 show the values in the table as a bar chart to give a visual comparison of the accuracy of each model. At a glance, we can see that the R2 score was close to one and the other parameters show that our error is low although not zero which means we still have room for improvements

Figure 9 shows the scatter plot graph of the prediction of the trained model against the actual value of PM2.5 for the Padonia station. Figure 10 shows the scatter plot graph of the prediction of the trained model against the actual value of PM2.5 for the Howard County station.

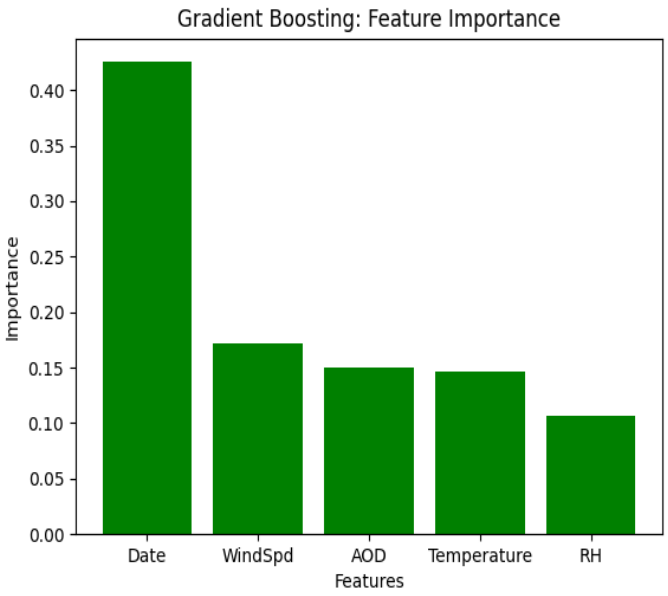

Fig. 11. Gradient Boosting feature importance (Padonia Station)

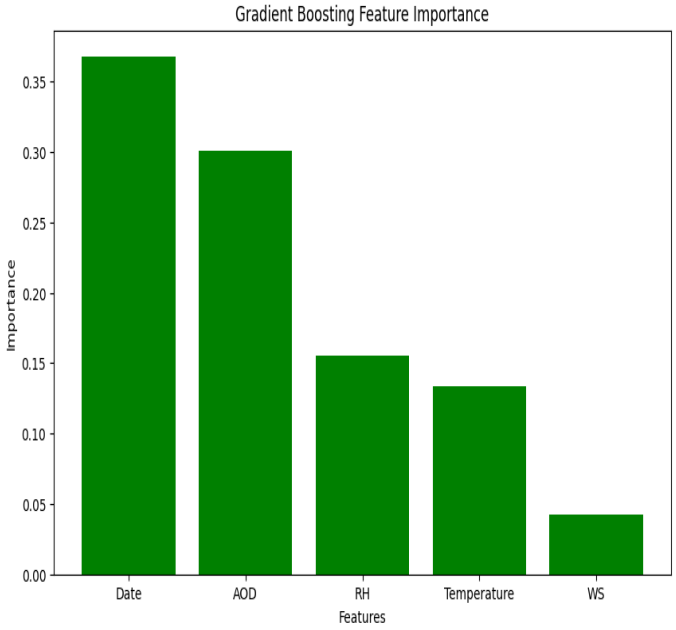

Fig. 12. Gradient Boosting feature importance (Howard County Station)

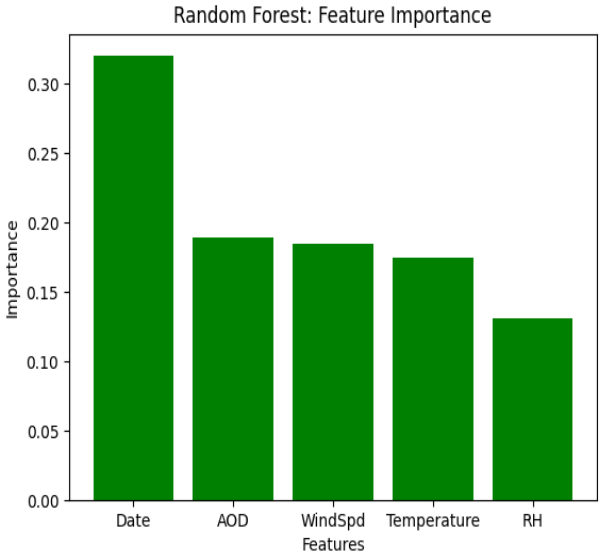

Fig. 13. RF feature importance Padonia station.

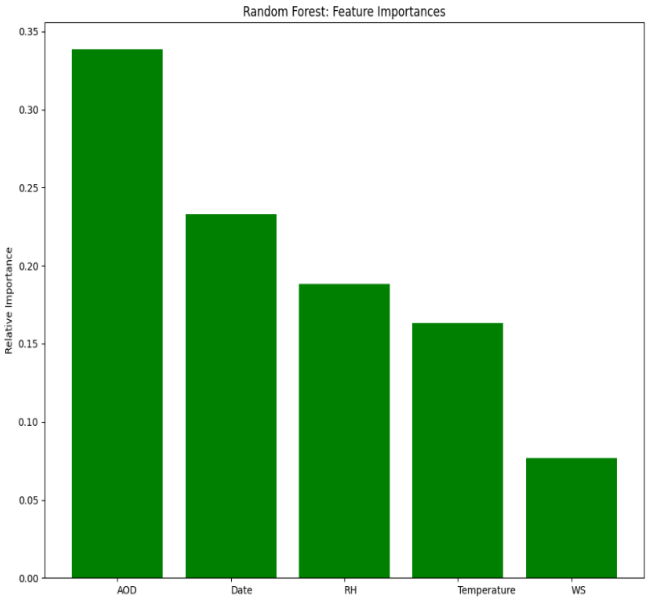

Fig. 14. RF feature importance Howard County.

Figures 11-14 show the feature importance for each model used through this we can see the features and how much of an impact it had on the AI models prediction.

The concept of feature importance is deep and we only scratched the surface which is why we were unable to talk about the feature importance for the neural network models that we used (LSTM and FNN) .

VI. Conclusion

As seen in previous sections, the implementation of AI in the field of PM2.5 monitoring was successful to some extent. From both stations, the Gradient Boosting Model served to yield the best predicted versus actual overlap. The random forest model, on the other hand, was slightly less accurate.

The LSTM model, on the other hand, performed worse than expected. Based on the theory for Deep Learning models, they must learn the most from the input data set and make more accurate predictions. However, in the case of this project, due to the inconsistencies in the datasets, the LSTM model struggled to make accurate predictions on non-sequential data. FNN was adopted as a deep learning model to work with inconsistent data, and also yielded good results. Despite not being exemplified in this paper, the data was also used to train Linear Regression and Standard Vector Regression models. The results obtained with these models were much lower than those reflected in the paper.

A. Challenges and Solutions

Even though the results seem good and pleasing, there were many challenges encountered in the process of data collection and visualization.

1) Inconsistent AOD Data from Satellites:
   Satellites measure AOD from the amount of sunlight that reflected and deflected by aerosols in the atmosphere. Therefore, they do not measure AOD during night periods. Due to this fact, the dataset of AOD had less than half of the theoretical points (about 75000 data points). In the field of research and data collection it is also possible that there are outliers in the dataset, which had to be dropped when preprocessing our dataset. All these factors put together reduced the dataset to a total of about 13000 points for the Howard Station and 12000 points for the Padonia Station.

   Solution: More accurate data was obtained from NASA's Dark Target Website.

2) Over-fitted Data Interpolation:
   In order to reverse the effect of losing data points due to various factors, the data was interpolated to train the different models. However, with more than half of the data being fake, the model created an over-fitted model with extremely high R2 scores in the range of 0.92-0.95.

   Solution: Instead of interpolating data, raw data was used to train models.

3) Overcomplicated Models:
   Apart from the models discussed in this paper, the GTW (Geographically and Temporally Weighted) tree model appeared to yield pleasing results, as stated by [13]. As stated in the name, this model yields results based on location i.e. latitude and longitude data. Since this project only focused on two locations this model seemed to be slightly confusing to understand its implementation with the datasets from Howard and Padonia.

   Solution: Models that depended on variables other than the meterological data were dropped.

4) Long Runtimes for ML Models:
   When attempting to test the models that were developed with more trees (for Random Forest and Gradient Boosting) and epochs (for LSTM), they ran for about 30 minutes minimum. This hindered the ability to try out combinations of hyperparameters to increase the R2.

   Solution: Models were ran simultaneously on different Google Colab notebooks. The graphs to generate and models were split up between teammates.

B. Future Work

1) Most of the challenges described above were caused by the lack of consistent data. Therefore, the next step in the field of climate science would be to figure out a way to measure and collect AOD data during the night season.
2) This project only focused on two specific locations, thus expanding our geographical coverage by exploring the PM2.5 predictions for various other locations. These locations would also include rural areas, and multiple other urban locations. This will open up ways to understand the impact of spacial variation on PM2.5.
3) Instead of using a date as a variable, it may be easier to include the Julian date of that day. The Julian calendar consists of a continuous count of numbers from 4000 BC. The continuity of the dates on the Julian calendar make calculations and visualization easier.
4) Exploring other deep learning models would be a great way to expand research on this project. Discovering new models such as the GTW tree, would enlarge the scope of this project by introducing new interpretations and significant features of the data obtained from the two stations.

REFERENCES

[1] Y.-H. Ryu, S.-K. Min, and C. Knote, "Contrasting roles of clouds as a sink and source of aerosols: A quantitative assessment using wrf-chem over east asia," *Atmospheric Environment*, vol. 277, p. 119073, 2022.

[2] B. Mmame, P. Sunitha, and K. Samatha, "Identification of sources and sinks of atmospheric aerosols and their impact on east african rainfall," *Acta Geophysica*, vol. 71, no. 3, pp. 1335–1346, 2023.

[3] T. Holloway, D. Miller, S. Anenberg, M. Diao, B. Duncan, A. M. Fiore, D. K. Henze, J. Hess, P. L. Kinney, Y. Liu, *et al.*, "Satellite monitoring for air quality and health," *Annual review of biomedical data science*, vol. 4, no. 1, pp. 417–447, 2021.

[4] T. Haszpra, "Intricate features in the lifetime and deposition of atmospheric aerosol particles," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 7, 2019.

[5] L. Liang, J. Daniels, C. Bailey, L. Hu, R. Phillips, and J. South, "Integrating low-cost sensor monitoring, satellite mapping, and geospatial artificial intelligence for intra-urban air pollution predictions," *Environmental Pollution*, vol. 331, p. 121832, 2023.

[6] R. R. Dickerson, P. Stratton, X. Ren, P. Kelley, C. D. Heaney, L. Deanes, M. Aubourg, K. Spicer, J. Dreessen, R. Auvil, *et al.*, "Mobile laboratory measurements of air pollutants in baltimore, md elucidate issues of environmental justice," *Journal of the Air & Waste Management Association*, vol. 74, no. 11, pp. 753–770, 2024.

[7] T. Khadir, I. Riipinen, S. Talvinen, D. Heslin-Rees, C. Pohlker, L. Rizzo, L. A. Machado, M. A. Franco, L. A. Kremper, P. Artaxo, *et al.*, "Sink, source or something in-between? net effects of precipitation on aerosol particle populations," *Geophysical Research Letters*, vol. 50, no. 19, p. e2023GL104325, 2023.

[8] P. Nath, B. Roy, P. Saha, A. I. Middya, and S. Roy, "Hybrid learning model for spatio-temporal forecasting of pm 2.5 using aerosol optical depth," *Neural Computing and Applications*, vol. 34, no. 23, pp. 21367–21386, 2022.

[9] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "Pm2. 5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data," *Atmosphere*, vol. 10, no. 7, p. 373, 2019.

[10] L. Zhao, Z. Li, and L. Qu, "A novel machine learning-based artificial intelligence method for predicting the air pollution index pm2. 5," *Journal of Cleaner Production*, vol. 468, p. 143042, 2024.

[11] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering: An Open Access Journal*, vol. 2, no. 1, pp. 602–609, 2014.

[12] D. Tran and A. W. Tham, "Accuracy comparison between feedforward neural network, support vector machine and boosting ensembles for financial risk evaluation," *Journal of Risk and Financial Management*, vol. 18, no. 4, p. 215, 2025.

[13] T. Li, Y. Wang, and J. Wu, "Deriving pm2. 5 from satellite observations with spatiotemporally weighted tree-based algorithms: enhancing modeling accuracy and interpretability," *npj Climate and Atmospheric Science*, vol. 7, no. 1, p. 138, 2024.

VII. Meet The researchers

Fig. 15. Elton Mawire

Elton Mawire is a senior Electrical Engineering Major at Alabama A&M with a concentration in computer engineering. Elton was born and raised in Zimbabwe, where he developed a strong affinity for physics, mathematics, and problem-solving. These interests have led him to research spaces spanning from chips to aerosols. He continues to propel forward, taking each step at a time with an aspiration to utilize his skills for the betterment of humanity.

Fig. 16. Fikewa Akindolire

Fikewa is a business management student at Morgan State University, aspiring to become a Product Manager in AI and Machine Learning. She combines her entrepreneurial spirit—shown through her hair braiding business—with a passion for innovation and a background in community engagement. A former lacrosse player, she values discipline, teamwork, and creativity. Her journey began with a curiosity about how technology shapes daily life, which evolved into a goal of building smart, inclusive products. Always seeking new opportunities, Fikewa is focused on growing the skills needed to lead in tech.

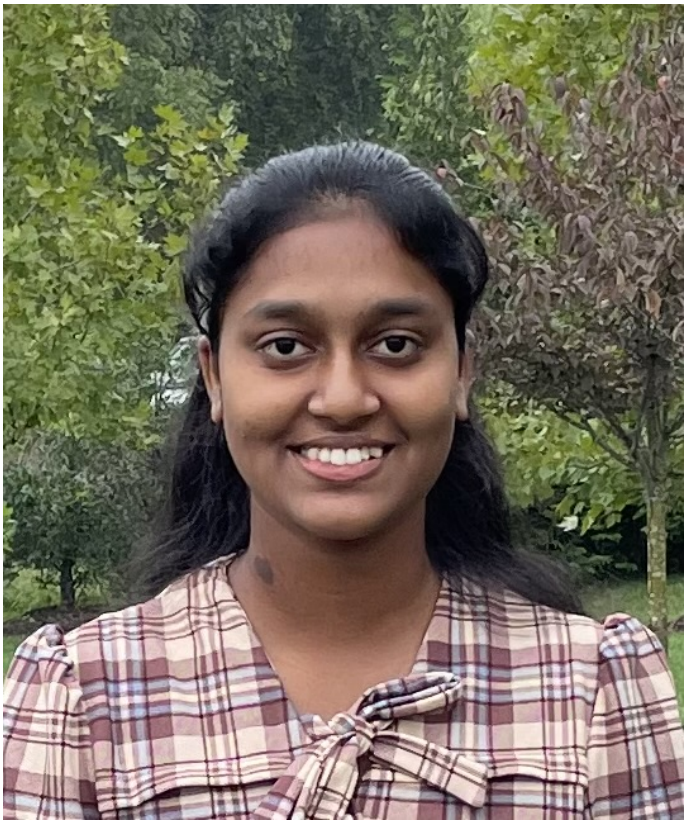

Fig. 17. Gracelyn Ruth Arunachalam

Gracelyn Ruth Aruanchalam is a rising Junior at the University of Maryland, College Park majoring in Computer Engineering. She enjoys learning about the mechanics and real-life applications of AI and ML models and wishes to specialize in this field in the upcoming years. Through this project she gained a general idea of the implementation of AI in real world issues, specifically in Climate Science.

Fig. 18. Unegbu M Sogidechukwu

Unegbu Sogidechukwu is a student currently studying Computer Science at Community College of Baltimore County in Baltimore, Maryland. He is interested in becoming a machine learning specialist in the future. From May 2025 to July 2025, he was a Research Assistant with the CEAMLS 2025 Summer Institute. When he is not working on research, and self improvement. He enjoys reading, playing games, and learning new things