

Enhancing Air Quality Forecasting by Predicting Volatile Organic Compounds Using Machine Learning

Shreyas R. Rajalingari

Monta Vista High School, 21804 McClellan Rd, Cupertino, CA, 95014, USA; go4shreyas@gmail.com

ABSTRACT: Although current research on air quality reveals that certain VOCs (like formaldehyde), which have large mass increases (>33%) during wildfire events, may be relevant for air quality prediction, they are not consistently and accurately reported or included in AQI. These issues have also severely limited the prediction of VOCs into the future. This paper attempts to test the accuracy of various machine learning models for predicting formaldehyde given historical AQI, month, and wildfire count data for counties in California from 2015 to 2025. Based on experimental results, the correlation coefficient was greater than 0.3 in a majority of cases and ranged to around 0.8, suggesting at least a moderate, positive correlation between AQI and formaldehyde. In addition, the Gradient Boosting Regressor (GBR) model performed the best (without overfitting) for predicting formaldehyde based on AQI, month, and wildfire count. The GBR model achieved an RMSE of 1.36 after hyperparameter tuning. This research suggests that there is a correlation between AQI and formaldehyde that should be explored further through higher-quality formaldehyde sensors and more consistent reporting of formaldehyde. Incorporating meteorological data and deep learning models may improve the accuracy of formaldehyde prediction as well.

KEYWORDS: Earth and Environmental Science, Atmospheric Science, Air Quality, Volatile Organic Compounds, Formaldehyde.

■ Introduction

The Air Quality Index (AQI) is the most widely adopted system for reporting air quality across the United States. The Environmental Protection Agency currently calculates AQI by selecting the highest pollutant index among 5 regulated criteria pollutants: particulate matter (less than 2.5 and 10 microns), ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide. The pollutant index is defined as follows:

$$\text{Pollutant Index} = \frac{AQI_{Hi} - AQI_{Low}}{Conc_{Hi} - Conc_{Low}} \times (Conc_i - Conc_{Low}) + AQI_{Low}$$

In the equation above, $Conc_i$ refers to the measured concentration of the pollutant, $Conc_{Hi}$ and $Conc_{Low}$ refer to the pollutant-specific breakpoints immediately greater than and less than the measured concentration, respectively, and AQI_{Hi} and AQI_{Low} refer to the corresponding AQI value for the Hi and Low breakpoints, respectively. The pollutant-specific breakpoints come from the EPA National Ambient Air Quality Standards (NAAQS).¹

Despite its widespread usage, the current AQI system has several limitations. For example, the criteria pollutants used for AQI do not include larger particles that are produced during wildfires, including ash. In addition, these pollutants do not include volatile organic compounds (VOCs), which are harmful toxins often produced by wildfires. These omissions may be especially important in states with historically high wildfire counts, such as California. One study by Jaffe *et al.* notes that in addition to particulate matter, wildfire smoke can contain VOCs, including benzene, toluene, xylene, acrolein, and formaldehyde,² which are associated with certain health risks like cancer and neurological defects. Compared to the 5 criteria pollutants measured in AQI, VOCs are often unreliably

reported due to limited monitoring frequency and poor-quality sensors, making it difficult to obtain an accurate model for air quality that integrates them.

Another study by Rice *et al.* shows that formaldehyde increased by around 34% on average during wildfires, the second largest mean mass increase during wildfire events among all VOCs (second to acrolein at 43%).³ These results show that certain VOCs may be meaningful predictors of general air quality. Since formaldehyde was also more consistently tracked by the EPA than acrolein (although still to an irregular extent), this paper will focus on formaldehyde as a case study.

Existing ML-based air quality forecasting studies are typically limited to AQI or the 5 criteria pollutants that are found in AQI. For example, a study conducted by Fan *et al.* forecasted ozone and PM_{2.5} in the Pacific Northwest,⁴ while a study by Raffuse *et al.* forecasted PM_{2.5} in wildfire conditions.⁵ However, neither of these studies considers volatile organic compounds. This highlights a clear gap in research focusing on the forecasting of VOCs like formaldehyde, which can have equally consequential health effects.

Current research on air quality reveals that while certain VOCs are relevant for air quality prediction and have dangerous health effects, they aren't accurately reported or included in AQI. This paper attempts to examine the relationship between AQI and formaldehyde and assess whether AQI actually predicts formaldehyde concentration in order to measure the efficacy of extrapolating AQI to the concentrations of harmful VOCs. In addition, while substantial research exists attempting to predict AQI into the future, the prediction of formaldehyde has been severely limited due to the inadequate monitoring of formaldehyde and the lower quality of formaldehyde sensors.

This paper attempts to close this gap by testing the accuracy of various machine learning models for predicting formaldehyde given historical AQI, month, and wildfire count information for counties in California from 2015 to 2025. Data from California was specifically chosen to more accurately determine the relevance of wildfires in formaldehyde prediction, as California is known to be a wildfire hotspot. This paper first attempts to establish a correlation between AQI and formaldehyde and then trains various simple machine learning models on AQI, month, and fire count (the number of wildfires in the county of the measurement on that particular date) and attempts to accurately predict formaldehyde. It was hypothesized that there is a positive correlation between AQI and formaldehyde, with AQI, month, and fire count all being important factors for predicting formaldehyde.

■ Methods

Data:

The AQI and VOC datasets⁶ are publicly available and are found on the EPA website, specifically under Pre-generated Data Files on the EPA Air Quality System (AQS) page. The specific files used are Hourly VOC data and Daily AQI data from 2015 through 2025. The temporal range 2015 through 2025 was chosen to obtain the most recent and accurate VOC data. Entries for VOC data are typically averages taken over 1, 3, 8, or 24 hours of a single day, while all entries for AQI data are averages taken over the entire day. The wildfire data⁷ is also publicly available and found on the CALFire website under Incidents/Incident Data.

Materials:

Google Colaboratory, a hosted Jupyter Notebook service, was used as the development environment and is where all of the code for this paper was written. All of the code for this paper was written in Python. The data downloaded from the sources listed in the Data subsection above was stored in a folder on Google Drive, and Google Colaboratory allowed this drive to be mounted directly to the notebook itself. The following Python libraries were used during code development: pandas, sklearn, pathlib, typing, matplotlib, glob, and os. The pandas, pathlib, typing, glob, and os libraries were used to compile the data and build the pandas DataFrames, while sklearn was used to build the models themselves.

Data Preprocessing:

1. AQI Data Preprocessing: First, the CSV file containing the AQI dataset was converted to a pandas DataFrame. Then it was filtered by state name such that only California AQI data from 2015 to 2025 remained.

2. VOC Data Preprocessing: First, the CSV file containing the VOC dataset was converted to a pandas DataFrame. Then the DataFrame was filtered by state name and parameter name such that only formaldehyde data from California from 2015 to 2025 remained. In cases where the formaldehyde data contained multiple entries for the same date and location, the entries were deduplicated by keeping rows in the following order of precedence: highest sample duration (time over

which the sample was taken), then highest arithmetic mean (concentration of formaldehyde for that entry). Higher sample durations were prioritized because they provide estimates that are more accurate for an entire day. The highest arithmetic mean was prioritized because the maximum concentration was also the most dangerous concentration for that location on that date.

3. Merging the Datasets: First, the AQI and VOC datasets were merged using the columns for county name and date to keep only entries where both AQI and VOC data were captured. Then the CALFire data was used to add a column called “fire counts,” which counted the number of wildfires burning in that entry’s location for that particular day.

4. AQI and VOC data were compared to establish any correlation between them. This step is described in more detail in the Data Analysis subsection.

5. Removing Outliers and Compiling the Final Dataset: First, data where the VOC observation percent—the percentage of samples taken out of the total scheduled to be taken—was less than 50% were removed to remove data that was primarily incomplete. This step also removed all but four of the outlier entries for the formaldehyde data because most of the original outliers had observation percentages that were less than 50%. Since the number of remaining outliers was extremely small compared to the number of observations, they had a negligible impact on model performance. Finally, the dataset was filtered to 6 columns, which comprised the final dataset that was used to train the machine learning models: state name, county name, date, AQI, formaldehyde, and fire count. Training the machine learning models is described in more detail in the Models subsection.

Data Analysis:

1. Normalizing the dataset: First, all of the columns in a copy of the final dataset were z-score normalized using StandardScaler from the sklearn library.

2. Establishing a Relationship Between AQI and Formaldehyde: To identify a relationship between AQI and formaldehyde, a dual-axis line plot was created for every entry in the dataset with AQI and normalized formaldehyde plotted on the y-axis (with an equal scale for both) and the date plotted on the x-axis. The correlation coefficient between normalized AQI and normalized formaldehyde was then calculated to quantify the correlation between AQI and formaldehyde.

Models:

1. Splitting Feature and Target Datasets: The feature DataFrame, containing the AQI, month (came from date), and fire count columns, was separated from the target DataFrame containing the formaldehyde column.

2. Normalizing the Feature Dataset: All of the features in the feature DataFrame were z-score normalized using a StandardScaler.

3. Splitting the Data into Testing and Training Datasets: Both the feature and target DataFrames were separated into training and testing data, using the train_test_split function

from sklearn, such that 75% of the data was used for training and 25% was used for testing.

4. Building and Testing the Baseline Models: First, a Linear Regressor was built using the open-source model found in the sklearn library. The linear regression model attempts to find the line of best fit for data that is linearly correlated. The Linear Regressor was trained on the training set and tested using the testing dataset. The following regression line was found after testing:

$$\text{formaldehyde} = 3.0504 + 0.9221 \times \text{AQI} + 0.1709 \times \text{Fire Count} + 0.2814 \times \text{Month}$$

A dummy regressor from the sklearn library was also built in order to provide a baseline to measure how much each model improved over a useless model. The dummy regressor always predicts the average of the target variable.

5. Building and Testing the Models: First, the following models were built from open-source models found in the sklearn library: Random Forest Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, and SVR. The Random Forest Regressor trains many individual decision trees and then averages the results. The Gradient Boosting Regressor sequences many decision trees, with each one learning from the errors of the previous model. The XGBoost and LightGBM models are variants of the Gradient Boosting Regressor. The SVR model maps data into high-dimensional space and attempts to find the best-fitting hyperplane to minimize loss. Next, each of the models was trained on the training set and tested using the testing dataset.

6. Tuning the Models: After training and testing, the hyperparameters of each model (where applicable) were tuned using GridSearchCV from the sklearn library. GridSearchCV searches through all combinations of a specified hyperparameter dictionary to find the combination with the best average R-squared score (percentage of variance in the target variable captured by the input features). Finally, the root mean squared error (RMSE) metric was used to assess the performance of each model, and the best-performing model was identified by selecting the one that improved the most over the baseline after hyperparameter tuning. Percentage improvement over the baseline was used as the ranking metric to highlight the relative usefulness of each machine learning model over a naive approach, even though raw RMSE after tuning would yield the same ranking.

■ Results and Discussion

Results:

First, dual-axis line plots were created showing normalized AQI and formaldehyde vs date for each of the counties in the time period from 2015 to 2025 in order to establish a relationship between AQI and formaldehyde. Each line plot also plotted semi-translucent orange bands in the background, which corresponded to the periods when each fire for that county was active. After each of the plots was plotted, the correlation coefficient between AQI and formaldehyde was calculated for each. Three examples of these plots are shown below.

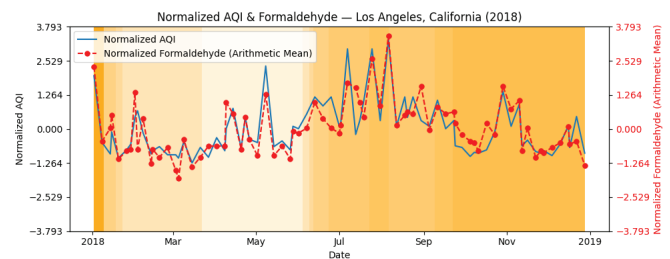


Figure 1: Dual-axis line plot of normalized AQI and formaldehyde vs date for Los Angeles County in 2018, before removing outliers, with semi-translucent orange bands plotted in the background to represent the periods during which each fire for 2018 in Los Angeles County was active. The correlation coefficient between normalized AQI and normalized Formaldehyde is approximately 0.8012.

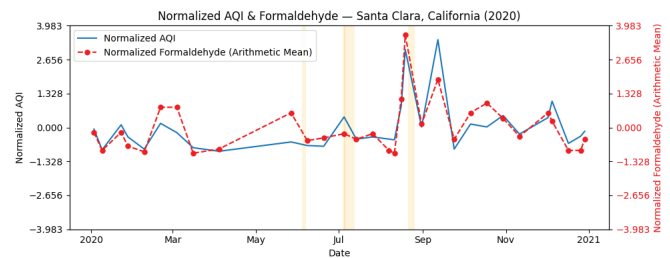


Figure 2: Dual-axis line plot of normalized AQI and formaldehyde vs date for Santa Clara County in 2020, before removing outliers, with semi-translucent orange bands plotted in the background to represent the periods during which each fire for 2020 in Santa Clara County was active. The correlation coefficient between normalized AQI and normalized Formaldehyde is approximately 0.8459.

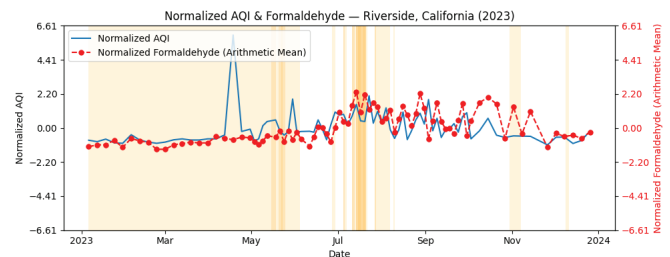


Figure 3: Dual-axis line plot of normalized AQI and formaldehyde vs date for Riverside County in 2023, before removing outliers, with semi-translucent orange bands plotted in the background to represent the periods during which each fire for 2023 in Riverside County was active. The correlation coefficient between normalized AQI and normalized Formaldehyde is approximately 0.3450.

After a relationship between AQI and formaldehyde was established, the Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, and SVR models were built and tested. Each of the models was trained on the training set and tested on the testing dataset. After this, each of the models, with the exception of the Linear Regression, was hyperparameter-tuned using GridSearchCV. After this, a dummy regressor was built in to provide a baseline, and the improvement of the hyperparameter-tuned models over the dummy regressor was measured as a percentage. The results are summarized below.

Table 1: RMSE on the testing dataset before and after hyperparameter tuning, the percentage of RMSE improvement over the Dummy Regressor, and the best hyperparameters found after tuning using GridSearchCV for each model. The Gradient Boosting Regression and XGBoost models had the highest percent RMSE improvement over the dummy regressor.

Model	RMSE Prior to Hyperparameter Tuning	RMSE After Hyperparameter Tuning	Percent RMSE Improved Over Dummy Regressor	Best Hyperparameters Found using GridSearchCV
Dummy Regressor	1.98	N/A	N/A	N/A
Linear Regression	1.88	N/A	5.05	N/A
Random Forest Regression	1.57	1.39	29.8	max_depth: 10, max_features: "sqrt", min_samples_split: 10, n_estimators: 100
Gradient Boosting Regression	1.38	1.36	31.31	learning_rate: 0.1, max_depth: 3, n_estimators: 100
XGBoost	1.56	1.36	31.31	colsample_bytree: 0.7, learning_rate: 0.05, max_depth: 3, n_estimators: 200, subsample: 0.9
LightGBM	1.40	1.37	30.81	learning_rate: 0.5, max_depth: 3, n_estimators: 100, num_leaves: 20, reg_alpha: 0.5, reg_lambda: 0.1
SVR	1.40	1.38	30.30	C: 1, epsilon: 0.2, kernel: "rbf"

The percentage increase in RMSE from training to testing for each of the models was also calculated in order to establish which models overfit and underfit the data. These results are summarized below.

Table 2: RMSE on the training and testing dataset after hyperparameter tuning using GridSearchCV (when applicable) and percentage increase in RMSE for each model. The XGBoost model had a higher percentage increase in RMSE from training to testing than the Gradient Boosting Regression model.

Model	RMSE on the Training Dataset	RMSE on the Testing Dataset	Percentage Increase in RMSE From Training to Testing
Dummy Regressor	1.90	1.98	4.04
Linear Regression	1.59	1.88	15.43
Random Forest Regression	1.17	1.39	15.83
Gradient Boosting Regression	1.27	1.36	6.62
XGBoost	0.97	1.36	28.68
LightGBM	1.20	1.37	12.41
SVR	1.36	1.38	1.45

Table 3: Mean Absolute Error (MAE) and R^2 on the testing dataset after hyperparameter tuning using GridSearchCV for each model.

Model	Mean Absolute Error (MAE) on the Testing Dataset After Hyperparameter Tuning	R^2 on the Testing Dataset After Hyperparameter Tuning
Random Forest Regression	1.070	0.367
Gradient Boosting Regression	0.922	0.510
XGBoost	1.031	0.375
LightGBM	0.954	0.495
SVR	0.917	0.498

Tree-based ensemble models like Gradient Boosting Regressor are highly interpretable, allowing for feature importance to be assessed. Below is the feature importance for each of the features used to train the tuned Gradient Boosting Regressor model.

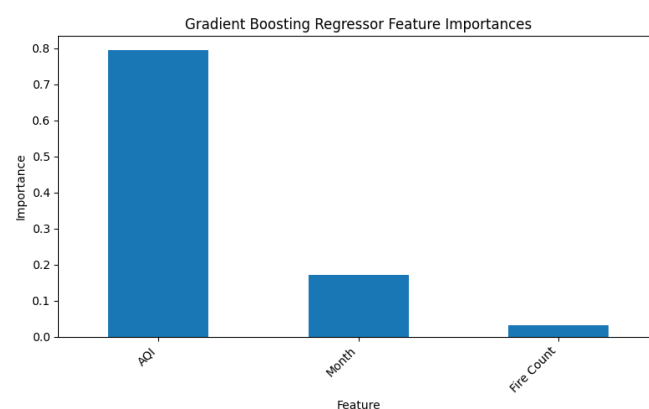


Figure 4: Bar chart showing the importance of each feature for the Gradient Boosting Regressor model after hyperparameter tuning. AQI was the most important, the month was the second most important, and fire count was the least important.

Finally, the Gradient Boosting Model predictions were plotted on a scatterplot showing the true values for formaldehyde for each normalized AQI value in the testing dataset. The results are shown below.

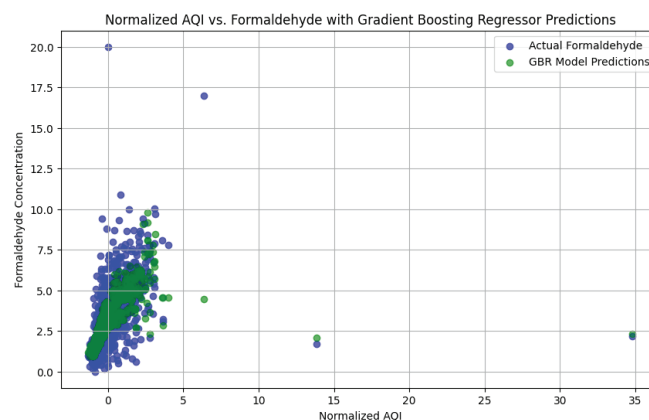


Figure 5: Scatter plot showing formaldehyde vs normalized AQI for each entry in the testing dataset and the Gradient Boosting Regressor model's formaldehyde predictions in green for each entry in the testing dataset. The Gradient Boosting Regressor model's predictions were close to the actual formaldehyde values.

Discussion:

A majority of the calculated correlation coefficients between AQI and formaldehyde were greater than 0.3, indicating that there was at least a moderate positive correlation between AQI and formaldehyde. In some cases, for counties where a greater amount of formaldehyde data existed, the correlation coefficient was greater, suggesting a stronger correlation. For example, for Los Angeles County in 2018 (Figure 1), a county that had a high level of formaldehyde that year had a correlation coefficient of approximately 0.8012, which indicated a strong positive correlation. However, this was not always the case, as seen in Riverside County in 2023 (Figure 3), which had a correlation coefficient of approximately 0.3450 despite high levels of formaldehyde reporting. Outliers in the AQI dataset, especially the large spike in mid- to late April, may have contributed to this weaker relationship. This conclusion is supported by the fact that qualitatively there was a clear correlation between AQI and formaldehyde. In addition, it can

also be determined qualitatively that AQI and formaldehyde generally had greater values when there were many fires than when there were no fires.

Based on the results in Table 1, it can be established that the Gradient Boosting Regression and XGBoost models performed best on the testing dataset since those models had the greatest improvement in RMSE (31.31%) over the dummy regressor, which was the baseline model. Gradient Boosting Regression and XGBoost are models that perform well for complex, non-linear relationships between the features and the target, which is exactly what was found during the data analysis from Figures 1-3. However, after hyperparameter tuning, each of the models, except for linear regression, performed similarly overall, suggesting that if careful hyperparameter tuning is performed, any non-linear model could be a viable one.

Based on the results in Table 2, it is likely that XGBoost overfitted the training dataset, since its percentage increase in RMSE was relatively high, which suggests overfitting. The Gradient Boosting Regressor model, however, did not seem to overfit the data because its percentage increase in RMSE was relatively low. This suggests that the Gradient Boosting Regressor model was able to generalize better, making it the best overall model to predict formaldehyde based on the AQI, month, and fire count. This conclusion is supported by Table 3, which shows that the Gradient Boosting Regressor model had the second-lowest MAE and the highest R^2 value.

From the feature importances in Figure 4, it can be seen that AQI was the most important feature at around 0.79 importance, and fire count was the least important at around 0.04. This shows that AQI was an important feature in predicting formaldehyde, which furthers the idea that AQI is a good predictor of formaldehyde overall. The fire count was the least important, which was surprising given that formaldehyde tends to increase during fires. The dominance of AQI suggests that it tracks short-term VOC fluctuations well, while fire activity contributes only indirectly due to its primarily long-term effects.

Based on Figure 5, it can be seen that the Gradient Boosting Regression model's formaldehyde predictions are relatively close to the actual values in the testing data and that the predictions seem to capture the overall trend between AQI and formaldehyde well.

Limitations:

While it can preliminarily be established that there is at least a moderate positive correlation between AQI and formaldehyde, more consistent reporting and better formaldehyde data quality are necessary to determine a more precise correlation. Also, AQI is likely an incomplete predictor because the same AQI value can have different compositions of the 5 criteria pollutants, which this experiment did not account for.

Additionally, it is important to note that there are confounding factors like meteorological conditions and quality differences between AQI sensors and VOC monitors that should be examined further to establish a more definitive correlation or relationship. Such factors, or noise in the data, may explain the four outliers in Figure 5, where formalde-

hyde spiked without a corresponding increase in AQI or AQI spiked without a corresponding increase in formaldehyde.

When training each of the models, the amount of data was only around 5000 rows, and there may still have been noise in the data (particularly in the AQI data), both of which could lead to overfitting. If the amount of high-quality data were increased, it is likely that the issue of overfitting seen in most of the models would subside, but further analysis with more data needs to be conducted to test that conjecture.

Furthermore, the initial data split did not preserve temporal ordering, which may have caused data leakage during training because the model could learn from future data to predict past observations. Any future repetitions of this experiment should preserve temporal ordering to alleviate this risk. However, when tested, preserving temporal ordering yielded negligible differences in results.

Next Steps:

One step that can be taken to potentially improve the models is to incorporate meteorological data as features. This could include wind speed, carbon monoxide, humidity, and temperature, which Khan *et al.* found to be the most important features for predicting AQI.⁸ This suggests that these features could be important for predicting formaldehyde as well, due to the established relationship between AQI and formaldehyde. In addition, training deep learning models or neural networks should be tested as the next step to evaluate whether those models can achieve better performance than the simpler models that were used for this paper.

Conclusion

This paper examined the relationship between AQI, month, wildfires, and formaldehyde by first establishing a relationship between AQI and formaldehyde and then training 6 regression models to see how accurately AQI, month, and fire count can predict formaldehyde using historical data from California. From the experiments conducted, it can be concluded that there is at least a moderate, positive correlation between AQI and formaldehyde. In addition, the Gradient Boosting Regressor was the best of the models tested for predicting formaldehyde based on AQI, month, and fire count. When tuned properly, however, all of the non-linear models produced similar results, suggesting that, with proper tuning, any of the non-linear models tested can be viable for predicting formaldehyde. Surprisingly, fire count was not as important a feature as the other two for predicting formaldehyde. With more high-quality data, a more precise relationship between AQI and formaldehyde can be calculated. Finally, incorporating meteorological data or deep learning models should be tested as a next step as a potential way to improve model performance.

Acknowledgments

I would like to acknowledge Shiang Chin, PhD. for his invaluable guidance throughout the project.

■ References

1. *AQI Breakpoints | Air Quality System | US EPA*. aqs.epa.gov. https://aqs.epa.gov/aqswb/documents/codetables/aqi_breakpoints.html.
2. Jaffe, D. A.; O'Neill, S. M.; Larkin, N. K.; Holder, A. L.; Peterson, D. L.; Halofsky, J. E.; Rappold, A. G. Wildfire and Prescribed Burning Impacts on Air Quality in the United States. *Journal of the Air & Waste Management Association* **2020**, 70 (6). <https://doi.org/10.1080/10962247.2020.1749731>.
3. R. Byron Rice; Boaggio, K.; Olson, N. E.; Foley, K. M.; Weaver, C. P.; Sacks, J. D.; McDow, S. R.; Holder, A. L.; LeDuc, S. D. Wildfires Increase Concentrations of Hazardous Air Pollutants in Downwind Communities. *Environmental Science & Technology* **2023**. <https://doi.org/10.1021/acs.est.3c04153>.
4. Fan, K.; Dhammapala, R.; Harrington, K.; Lamb, B.; Lee, Y. Machine learning-based ozone and PM_{2.5} forecasting: Application to multiple AQS sites in the Pacific Northwest. *Frontiers* **2023**. <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2023.1124148/full>.
5. Raffuse, S.; O'Neill, S.; Schmidt, R. *A model for rapid PM_{2.5} exposure estimates in wildfire conditions using routinely available data: Rapid-fire v0.1.3*. Geoscientific Model Development 2024. <https://doi.org/10.5194/gmd-17-381-2024>.
6. *AirData website File Download page*. aqs.epa.gov. https://aqs.epa.gov/aqswb/airdata/download_files.html#Annual.
7. CAL FIRE. *2025 Fire Season Outlook | Welcome to CAL FIRE*. Ca.gov. <https://www.fire.ca.gov/incidents>.
8. Khan, R. R.; Dwivedi, S.; Yadav, S.; Sharf, R.; Maurya, S.; Pandey, U. P.; Kumari, S. Integrating AI with IoT Sensors for Real-Time Air Quality Monitoring and Prediction. *International Journal of Environmental Sciences* **2025**, 2274–2287. <https://doi.org/10.64252/pc0vbk56>.

■ Author

Shreyas Rajalingari is a student at Monta Vista High School in Cupertino, CA. His intended major in college is computer science, AI, or data science. He loves to play football and baseball, hike, and bike. He also likes to travel around the world and meet new people.