Effects of Income Inequality on Pollution Exposure: A Case Study of Neighborhoods Surrounding the Philadelphia Energy Solutions Oil Refinery and an Inexpensive and Accessible Deep Learning Solution To Keeping Citizens and Researchers Informed

Yudhiishbala V Senthilkumar
John P. Stevens High School

With the mentorship of

Dr. Jeffrey Field
Professor of Pharmacology in the Center of Excellence in Environmental Toxicology
University of Pennsylvania Perelman School of Medicine

Abstract

The Philadelphia Energy Solutions (PES) Refinery has caused significant environmental harm since its establishment in 1870. On June 21, 2019, a ruptured elbow pipe triggered a catastrophic explosion, releasing large amounts of propane and toxic hydrofluoric acid into the atmosphere. This incident highlighted the longstanding pollution and environmental concerns surrounding the refinery and its disproportionate impact on nearby communities. This paper splits into three portions: a case study that examines the correlation between income and pollution exposure, an improved study that uses data analysis models to deepen the understanding of this correlation, and an experiment that applies deep learning models to create an accessible solution for affected citizens in Philadelphia and similarly impacted areas. The case study focuses on ten census tracts within the southeastern Philadelphia area that are the closest census tract to the PES Refinery in their income range and correlates the income of these census tracts to the level of PM 2.5, PM 10, formaldehyde (HCHO), and "Total Volatile Organic Compounds" (TVOC) that residents are exposed to. First, this study identifies the ten census tracts closest to the PES Refinery regarding their income range. Then, it calculates the average square feet of homes and the number of green spaces (square feet) in each census tract. Then, it pinpoints 4 locations in each census tract. It collects the PM 2.5, PM 10, HCHO, and TVOC levels from these pinpointed locations. The study calculates the average levels for each census tract. The analysis reveals a correlation coefficient of 0.955 for PM 2.5, 0.972 for PM 10, 0.852 for HCHO, and 0.906 for TVOC. These statistics indicate a strong negative correlation between income and pollution exposure. The improved study confirms the correlation found in the case study and investigates how the time of day and proximity to industrial zones influence pollution disparities. A model that integrates Random Forest, XGBoost, a fully connected neural network, timeseries analysis, and Pearson and Spearman correlation tests was developed to do this. Additional pollution data were collected using the same methodology as the case study but during multiple time frames on a given day. Proximity data of communities from the PES Refinery were analyzed using Google Maps. The inverse correlation between income and pollution exposure was confirmed after inputting all of the new data into the model. No correlation between time of day, pollution exposure, and income was found, and lowerincome tracts, which already have much higher pollution exposure, were also identified as the closest to the PES Refinery. The final experiment creates a tool that uses image analysis to provide valuable, realtime insights to users about the pollution they breathe at the current moment. To do so, over 12,000 sky images from ground level and related data regarding nitrogen dioxide and sulfur dioxide levels were collected from Kaggle. The pictures are taken in major cities in India and Nepal. The models crop the bottom 60% of each image to focus on the most relevant visual feature and calculate colorbased features such as average RGB and HSV values and RGB standard deviations. These features form the input data for some models, while other models process raw image data directly. Afterward, seven types of deep learning models were developed and trained separately for sulfur dioxide and nitrogen dioxide to classify the levels of the two pollutants present in the atmosphere shown in the picture and tune the models' hyperparameters over multiple training rounds. The most

successful architecture for nitrogen dioxide, the CNN, achieved a testing accuracy of 94.72%. The most successful architecture for sulfur dioxide, the DNN, achieved a testing accuracy of 99.25%. These results highlight the ability of these models to accurately detect pollution levels, even in images taken from different angles or with additional objects like buildings in the frame. By providing actionable insights into air quality, the tool empowers communities to advocate for cleaner environments and take steps to protect their health.

1 Background

Philadelphia faces significant air quality challenges due to its large urban area, high population density, and various pollution sources. This case study examines the city's air pollution issues, focusing on the role of the Philadelphia Energy Solutions (PES) Refinery and its impact on surrounding neighborhoods. By analyzing socioeconomic factors, pollution sources, and specific census tracts, this study provides insights into how pollution exposure varies across different areas in Philadelphia.

- Philadelphia is known for its high levels of air pollution, with traffic, industrial emissions, and limited green spaces exacerbating the situation. The city’s air quality issues have made it one of the nation’s most polluted metropolitan areas, affecting residents' health and quality of life. [9, 10]
- Philadelphia has always held the title of one of the largest cities in the United States. It is one of the world's top ten largest cities by land area and one of the nine most populated cities in the entirety of the United States, with a population of over 1.5 million residents. Philadelphia is a very urban city with nearly 5 major multilane roads and 15 interstate highways passing through. It has nearly 600 thousand cars using these roads, with thousands more outofstate cars using them daily. An international airport, coal power plants, and agricultural air pollution certainly exacerbate Philadelphia’s air pollution. The PhiladelphiaReadingCamden metro area is ranked as one of the nation’s top 25 most polluted cities by yearround particle pollution, and Philadelphia received a very low "F" grade for its ozone grade. [9, 11]
- The PhiladelphiaReadingCamden metro area is ranked as one of the nation’s top 25 most polluted cities by yearround particle pollution, and Philadelphia received a very low "F" grade for its ozone grade. High levels of groundlevel ozone and particulate matter significantly impact air quality, posing health risks for residents, especially in densely populated and economically disadvantaged areas. [12, 13]

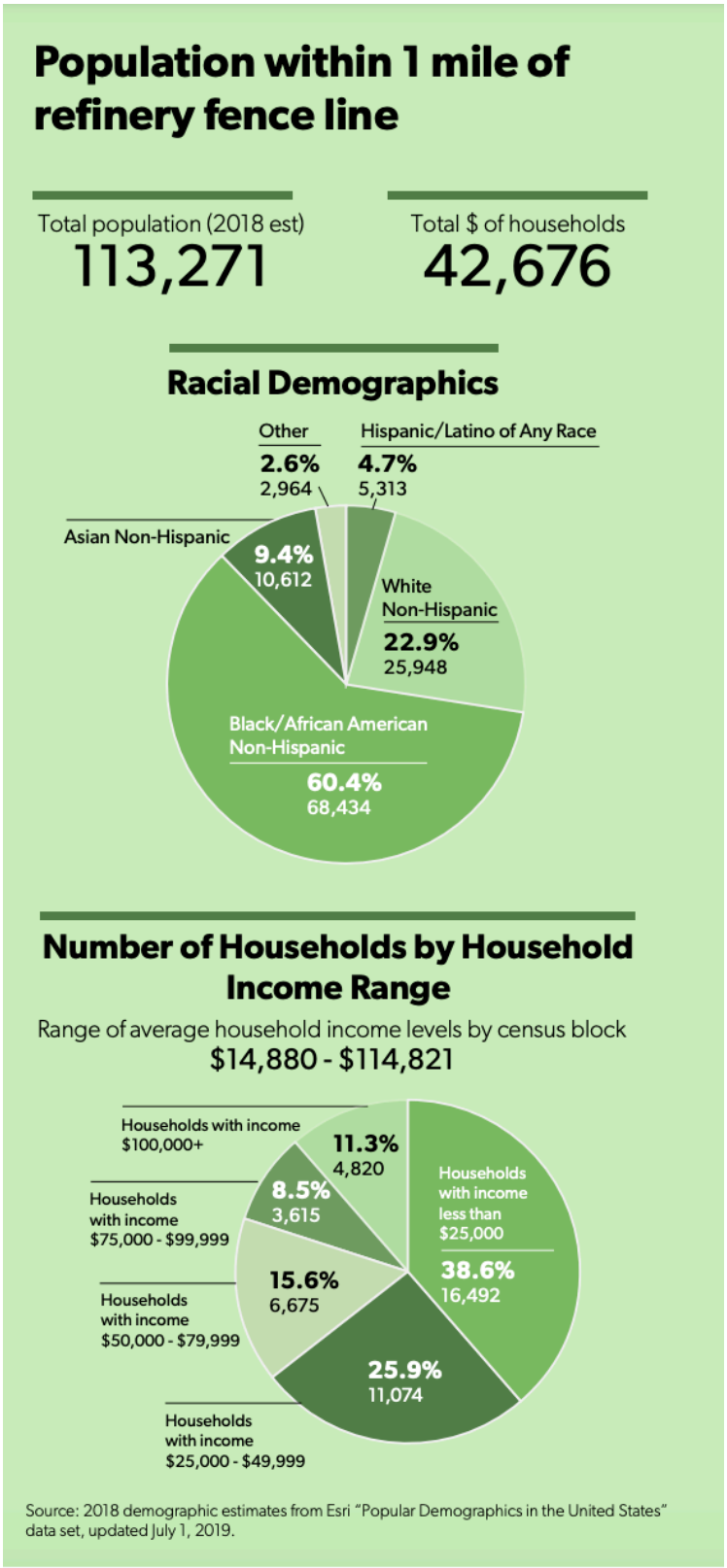

Figure 1: A pie chart illustrating the percentage of census tracts with a certain average income that are within a mile from the oncestanding Philadelphia Energy Solutions Refinery [1].

1.1 Philadelphia Energy Solutions (PES) Refinery

The PES Refinery has been one of the largest industrial sites in Philadelphia, creating a significant amount of pollution for over a century. Its emissions have contributed to local air quality and raised environmental and health concerns among nearby communities.

- Until 2019, the single largest source of carbon emissions throughout the entire city was the Philadelphia Energy Solutions (PES) Refinery, sometimes referred to as the South Philadelphia Refinery. It has been reported to be in operation since the 1860s when crude oil was starting to be refined. As one of the oldest refineries on the East Coast, its operations have had a longstanding impact on the city's environment and economy. [2, 3]
- According to the EPA, the PES Refinery released nearly 4.1 million pounds of VOCs and NOx particulates into Philadelphia's atmosphere between 2012 and 2019. These were some of the significant sources of local air pollution, affecting neighborhoods near the refinery most. [4]
- In June 2019, an explosion at the PES Refinery highlighted the environmental and safety risks associated with its operations. The incident released hazardous chemicals, which raised concerns about ongoing pollution and its effects on nearby communities. [5]
- An elbow pipe ruptured releasing large amounts of propane along with the toxic hydrofluoric acid. Eventually, the chemicals ignited and caused the explosion besides making the refinery burn for almost 24 hours. Although much of the toxic chemicals were drained to separate storage units and the rest was spent by the fire, 3000 pounds of toxic hydrofluoric acid still escaped. [5, 6]
- The remains of the refinery still leak small levels of benzene, sulfur dioxide, and nitrogen dioxide into the atmosphere. The continued pollution poses potential longterm health risks for those in neighborhoods near the plant and raises questions about the longterm environmental legacy of the refinery site. [7]
- The operations of the PES Refinery have substantially affected the quality of life in neighboring communities. This includes air quality, health concerns, and the physical and social environment of these communities. [3, 8]
- As the largest and oldest oil refinery on the East Coast, since the 1860s, the PES Refinery has had significant effects on the residential areas surrounding it. Residents have complained about the level of gasses and chemicals from this refinery and how these gasses and chemicals are negatively affecting the neighborhood they call home. [2, 8]
- News reporters have listened to the opinions of the residents about the smell and the level of pollution that comes from the refinery and how these factors affect their lifestyle. However, in most media reports, the data is usually not comprehensive, hence the need for a deeper analysis of the level of pollution and its effects on these communities. [8]

1.2 Socioeconomic Factors in Pollution Exposure

The level of pollution exposure in Philadelphia is influenced by socioeconomic factors such as income and access to green spaces. Lowerincome neighborhoods tend to be more exposed to pollution due to their proximity to industrial sites and limited environmental resources.

- It has been shown that factors such as income and home value affect where a person lives. Where a person lives largely depends on their income, but their location relative to major air pollution sources, their access to air pollution countermeasures such as green spaces, and many other factors can all affect their exposure to toxic pollutants such as VOCs and NOx particulates [1].
- The presence of pollution in the neighborhoods has not been quantified, and many other factors have not been considered when accounting for the level of pollution. Green spaces can help mitigate pollution, but they are often less accessible in lowerincome areas, increasing residents' exposure to pollutants.

1.3 Case Study: Air Quality Analysis of Census Tracts

This case study focuses on air quality across ten census tracts near the former PES Refinery. The analysis in these areas will help the study understand how income impacts pollution exposure, thus serving as the nexus of environmental and socioeconomic factors.

- The following case study evaluates the air quality in ten different census tracts around the remains of the former Philadelphia Energy Solutions Refinery, representing a range of income levels and locations throughout the city. These diverse tracts were selected because of their different socioeconomic status and distance from the facility.
- The case study will categorize the major factors that affect the level of air pollution exposure within the area surrounding the PES Refinery, including key factors such as income, proximity to sources of pollution, and availability of green spaces, in an effort to shed light on how these variables contribute to disparities in environmental health.

2 Research Question and Hypothesis of Case Study

- What is the relationship between income levels and exposure to air pollutants, such as PM 2.5, PM 10, HCHO, and TVOC, in metropolitan areas?

3 Methodology of Case Study

In order to understand whether and how one’s income can affect their level of pollution exposure, an understanding of the air quality in the areas near the PES Refinery had to be developed. The best approach was to select certain areas around the refinery and take an adequate amount of air quality samples to analyze and categorize, and to follow up with this experiment by getting an adequate amount of air quality samples on the same day, but the following year.

3.1 Determining Locations of Interest

Philadelphia was chosen as the city of interest due to feasibility and size. As stated, it is one of the largest cities in the entire United States. While it is a densely populated city, it is much more feasible to navigate than any other comparable city, such as New York City or Boston. 10 census tracts were chosen within the eastern side of Philadelphia, nearing the state borders of Pennsylvania and New Jersey. Locations were chosen based on the income range it fits in and the distance from the Philadelphia Energy Solutions Refinery. The idea was that each census tract had to be part of a different income range and be the closest census tract in its income range in terms of distance between the census tract and the Philadelphia Energy Solutions Refinery. Within each Census tract, certain locations were chosen to get multiple samples. These specific points within the locations and the number of points were chosen based on the size of the locations and proximity to major pollution cases (e.g., major roads, construction, etc.)

3.2 Data Collection

Sources such as Justicemap and Zillow were used to determine factors such as average income, average home price, the number of roads and landmarks, and so forth. A particulate matter meter was used within specific areas of each census tract to collect PM 2.5 and PM 10 samples from the pinpointed locations. These data points were recorded, averaged, and compared to determine what factors significantly affect the level of air pollution within a census tract.

The research process began by opening justicemap.org and selecting the necessary census tracts based on their location relative to an essential landmark, which, in this case, was the PES Refinery. Since the website provided census data, the areas were automatically categorized by average annual household income. Ten particular areas that were closer to the refinery and had different income levels were chosen. The income levels listed on justicemap.org included $0 $32,000, $32,000 $40,000, $40,000 $46,000, $46,000 $52,000, $52,000 $58,000, $58,000 $65,000, $65,000 $74,000, $74,000 $87,000, $87,000 $109,000, and $109,000 or more.

The next step was to open Google Maps and select the point of each census tract closest to the Philadelphia Energy Solutions Refinery. The refinery was entered as the starting point, and the nearest point of each census tract was set as the destination. This process was repeated for each census tract, and the distance shown on Google Maps was noted.

Zillow.com was then opened, with ‘Philadelphia, Pennsylvania’ entered as the location. The width of the computer window was adjusted until the Zillow page displayed a map view. The first census tract was located on the map, and the borders of the census tract were adjusted to fit perfectly in the window by zooming in and out as necessary. The ‘Draw’ button on the top right of the screen was clicked, and the shape of the census tract was drawn along the streets forming its border. After clicking ‘Apply’ on the top right, the ‘List’ button at the bottom center of the screen was clicked. The average home prices were calculated by summing the prices of the homes shown for sale and dividing by the number of homes for sale. This process was repeated for each census tract.

The process continued on Zillow.com by entering ‘Philadelphia, Pennsylvania’ as the location and adjusting the window width to display the map view. Each census tract was located

and adjusted to fit perfectly within the window. Using the ‘Draw’ button, the shape of each census tract was outlined along its border streets, and ‘Apply’ was selected. The ‘List’ button was then clicked to display the homes for sale in the area. The average square footage of homes was calculated by summing the square footage of the homes shown for sale and dividing by the number of homes for sale. This procedure was repeated for each census tract.

Next, Google Maps was opened, and each census tract was located on the map. The green spaces within each census tract were identified. The names of these green spaces were noted, and the square footage was looked up online. If the name of a green space was not provided, or the square footage could not be found, an approximation was made using the scale provided at the bottom right of the Google Maps page. By measuring the green space’s dimensions, an estimated square footage was calculated. This process was repeated for each census tract.

Google Maps was used again to locate each census tract. The orange Street View character was dragged to one of the border streets, allowing for a virtual tour of every street within the census tract, including the other border streets. Residential locations, businesses, and any other notable factors with environmental or health implications were recorded. This procedure was repeated for each census tract.

Various locations within each census tract were pinpointed on justicemap.org, and each site was visited with a particulate matter meter. The average air quality in each census tract was calculated by measuring the levels of PM 2.5 and PM 10 at each location. These values were recorded for each pinpointed location, and an average for PM 2.5 and PM 10 was calculated for each census tract.

Justicemap.org was used to locate the recorded census tracts. Within each census tract, the number of major and minor roads was noted. The definitions of "major roads" and "minor roads" were based on the specific characteristics of Philadelphia's roadways. Due to Philadelphia’s abundance of narrow, oneway roads, major roads were defined as highways or wide, twoway roads that could comfortably accommodate roadside parking on either side and were more than two blocks long within the census tract. Midsize roads were defined as slightly narrower than major roads, capable of accommodating roadside parking on either side, and more than two blocks long. For this research, midsize roads were mostly oneway roads, though some twoway roads without roadside parking also fit this category. The classification was determined at the researcher’s discretion.

4 Results

#	Census Tract	Avg. Household Income (Annual)
1	36	$0 $32k
2	37.02	$32k $40k
3	39.01	$40k $46k
4	30.01	$46k $52k
5	372	$52k $58k
6	38	$58k $65k
7	28.02	$65k $74k
8	373	$74k $87k
9	19	$87k $109k
10	13	$109k+

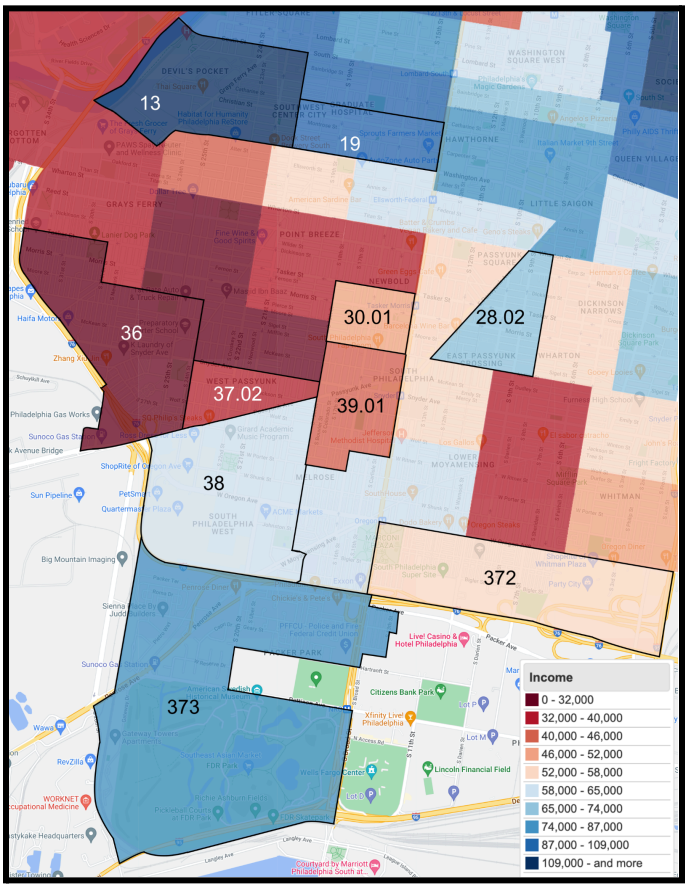

Figure 2: A map from justicemap.org. The analyzed census tracts have been outlined and labeled. The data included here are the census tract numbers and their average household income range guided by the key provided by justicemap.org.

Using justicemap.org, 10 Census Tracts of differing income levels were selected (as shown above). Each one is within the proximity of the PES Refinery (Located just west of the map in the gray area near Census Tracts 373 and 38). With these Census Tracts chosen, all the other data points were found.

4.1 Analysis of Correlation between Square Feet of Green Space and Average Income

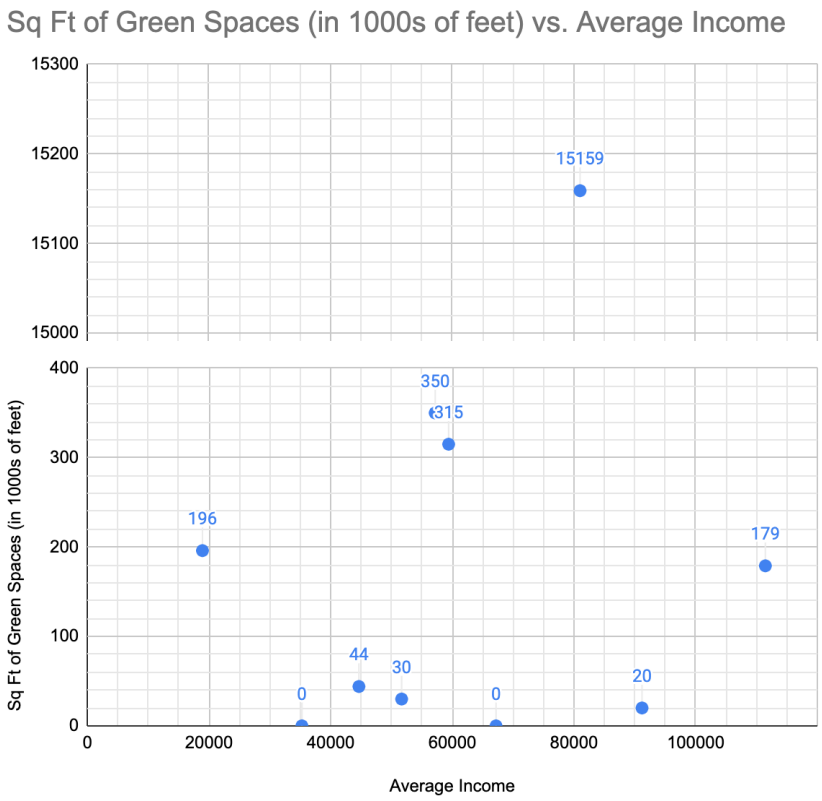

Figure 3: Square Feet of Green Space vs Average Income graph. Each point represents one of the 10 census tracts.

Figure 3 depicts the change in square feet of green space based on the change in average income in a census tract. As shown here, only one particular area, Census Tract 373, contains significant square footage of green space. As seen on the map and shown in Figure 1.1, the rest of the regions have minimal amounts of green spaces. No visible pattern is revealed.

4.2 Analysis of Correlation between Average Square Feet of Homes and Average Income

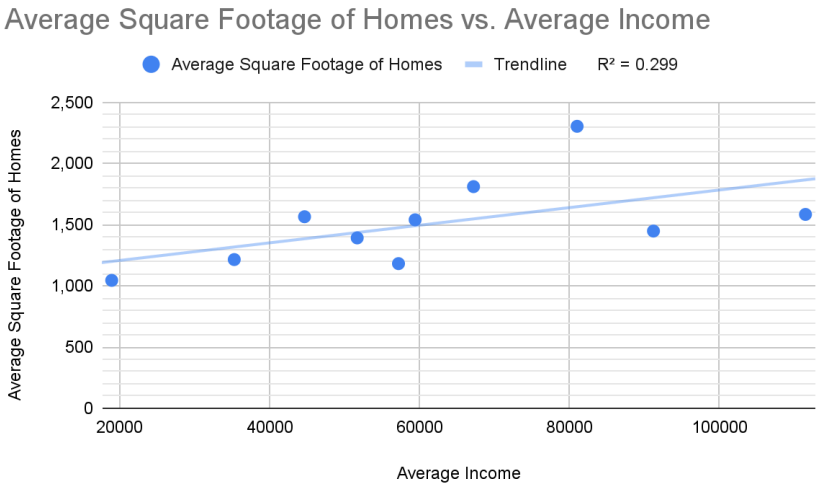

Figure 3: Average Square Footage of Homes vs Average Income graph. Each point represents one of the 10 census tracts.

Figure 3 depicts the change in square feet of a home in each census tract based on the change of average income in a census tract. There seems to be no particular pattern in the data. Given the very low correlation coefficient, it is highly unlikely that the average square footage of homes in any census tract is correlated to the independent variable of average income.

4.3 Pinpointed Locations

Census Tract	Pinpoint Address
30.01	1742 S 15th St, Philadelphia, PA 19145
30.01	1601 S 18th St, Philadelphia, PA 19145
30.01	1839 S 18th St, Philadelphia, PA 19145
38	3010 S 18th St, Philadelphia, PA 19145
28.02	1651 E Passyunk Ave, Philadelphia, PA 19148
28.02	1835 S 9th St, Philadelphia, PA 19148
373	2601 Penrose Ave, Philadelphia, PA 19145
38	2101 W Shunk St, Philadelphia, PA 19145

38	2300 W Oregon Ave, Philadelphia, PA 19145
38	1800 Bigler St, Philadelphia, PA 19145
28.02	1310 S 8th St, Philadelphia, PA 19147
28.02	1036 Watkins St, Philadelphia, PA 19148
373	1500 Pattison Avenue &, S Broad St, Philadelphia, PA 19145
13	2600 Christian St, Philadelphia, PA 19146
37.02 & 36	2100 S 24th St, Philadelphia, PA 19145
36	2400 S 24th St, Philadelphia, PA 19145
39.01	1600 Jackson St, Philadelphia, PA 19145
373	1526 Packer Ave, Philadelphia, PA 19145
30.01	1514 Tasker St, Philadelphia, PA 19145
372	2947 S 13th St, Philadelphia, PA 19148
373	2300 Hartranft St, Philadelphia, PA 19145
19	922 S 17th St, Philadelphia, PA 19146
37.02	1800 Snyder Ave, Philadelphia, PA 19145
19	2035 Washington Ave, Philadelphia, PA 19146
13	2000 Catharine St, Philadelphia, PA 19146
13	2149 Catharine St, Philadelphia, PA 19146
39.01	1930 S Broad St, Philadelphia, PA 19145
30.01	1604 S Broad St, Philadelphia, PA 19145
372	2800 S Broad St, Philadelphia, PA 19145
19	800 S Broad St, Philadelphia, PA 19146
13	2420 Grays Ferry Ave, Philadelphia, PA 19146

Figure 4: This table contains all of the locations that were visited for pollution data collection.

Each of these locations were selected based on their proximities to either major roadways or landmarks within each census tract. Multiple locations were chosen within each census tract to ensure accuracy among the data collected.

4.4 Analysis of Correlation between PM 2.5 and Income

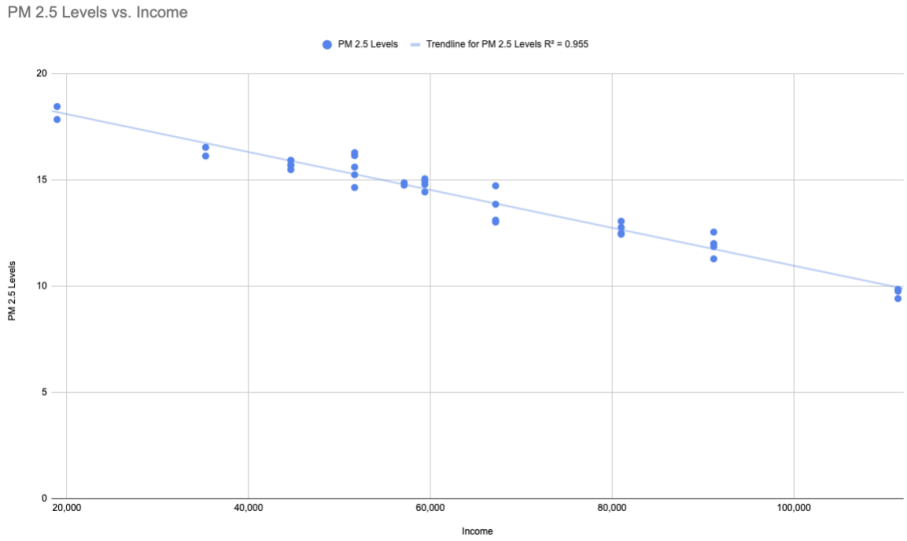

Figure 4: PM 2.5 Levels vs Income. Each point represents each visited location, and the line of best fit is provided as well.

Figure 4 depicts the change in PM 2.5 levels in each census tract, in 2022 , based on the average income in a census tract. As shown in the graph, the trendline has a negative slope, and the magnitude of the correlation coefficient is 0.955, indicating a strong correlation between PM 2.5 levels and average income. Given this high correlation coefficient, it is very likely that the levels of particulate matter that are 2.5 μm thick in any census tract are correlated to the independent variable of average income.

4.5 Analysis of Correlation between PM 10 and Income

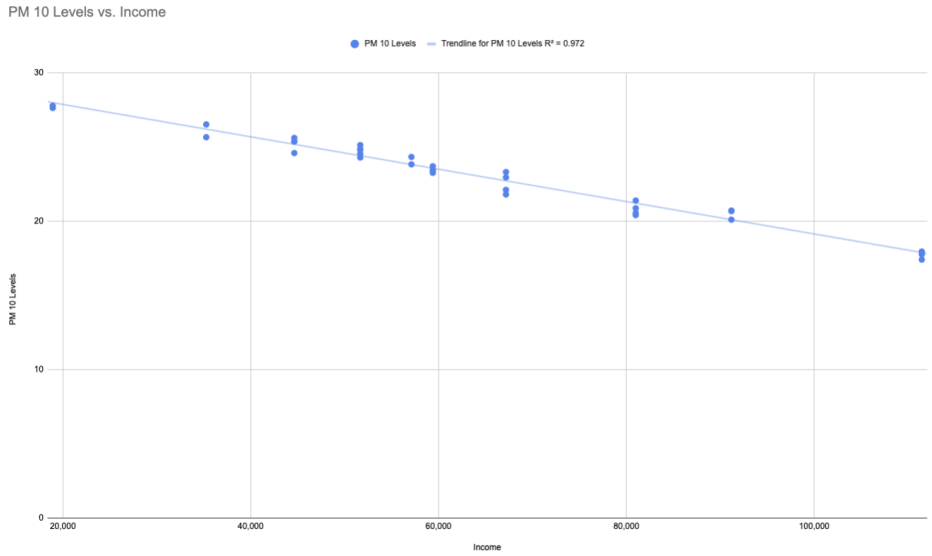

Figure 5: PM 10 Levels vs Income. Each point represents each visited location, and the line of best fit is provided as well.

Figure 5 depicts the change in PM 10 levels in each census tract, in 2022 , based on the average income in a census tract. As shown in the graph, the trendline has a negative slope, and the magnitude of the correlation coefficient is 0.972, indicating a strong correlation between PM 10 levels and average income. Given this high correlation coefficient, it is very likely that the levels of particulate matter that are 10 µm thick in any census tract are correlated to the independent variable of average income.

4.6 Analysis Analysis of Correlation between Formaldehyde and Income

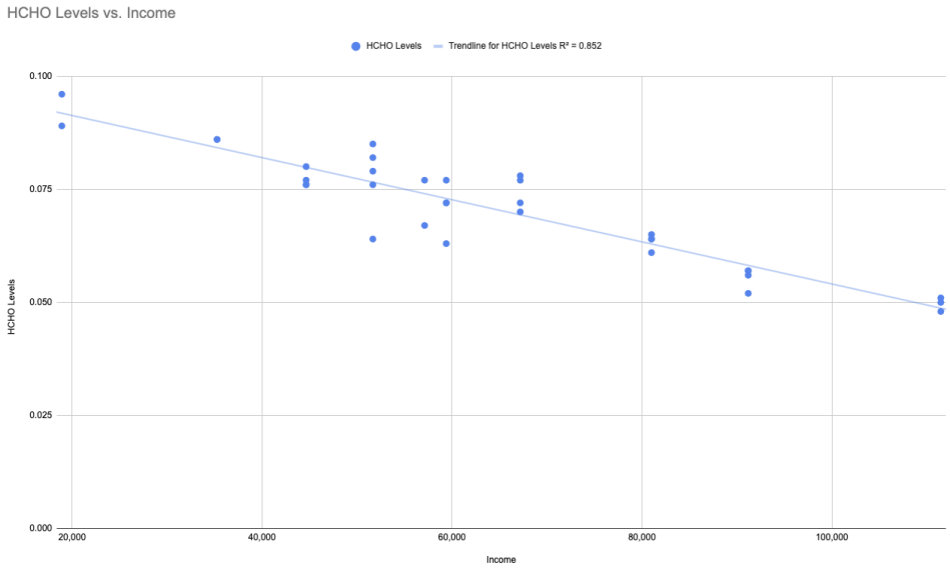

Figure 6: HCHO Levels vs Income. Each point represents each visited location, and the line of best fit is provided as well.

Figure 6 depicts the change in formaldehyde levels in each census tract, in 2022 , based on the average income in a census tract. As shown in the graph, the trendline has a negative slope, and the magnitude of the correlation coefficient is 0.852, indicating a strong correlation between formaldehyde and average income. Given this high correlation coefficient, it is very likely that the levels of formaldehyde in any census tract are correlated to the independent variable of average income.

4.7 Analysis of Correlation between TVOCs and Income

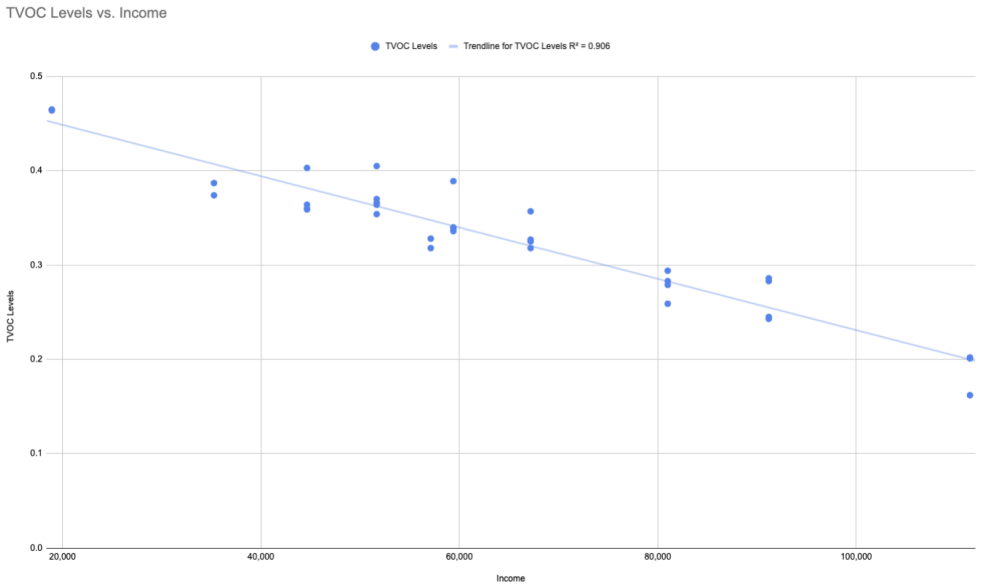

Figure 7: HCHO Levels vs Income. Each point represents each visited location, and the line of best fit is provided as well.

Figure 7 depicts the change in TVOC levels in each census tract, in 2022 , based on the average income in a census tract. As shown in the graph, the trendline has a negative slope, and the magnitude of the correlation coefficient is 0.906, indicating a strong correlation between TVOC and average income. Given this high correlation coefficient, it is very likely that the levels of TVOC in any census tract are correlated to the independent variable of average income.

5 Analysis and Discussion of Case Study

The data suggests that income levels have a strong negative correlation with the level of pollution exposure. Multiple data points were taken from each census tract, and the general regression of the averages suggests that income can influence a person's pollution exposure, especially in metropolitan areas like Philadelphia.

5.1 General Pollution Levels and Air Quality Observations

The data ranged within the lower end of the moderate level of the air quality index, which is unusually good for dense cities like Philadelphia. However, considering how significantly Philadelphia's pollution levels have dropped since 2000, it relates well to the readings that were captured [16]. The readings captured for TVOC (total volatile organic compounds) and HCHO (formaldehyde) all fell within the safe category [17].

5.2 Health Implications of Air Quality Levels

One thing to note is that while the moderate level is relatively safe for most normal adults, prolonged exposure within this moderate area is still unsafe for individuals of any age group [18]. This is especially true for the 330 thousand children under 18 and 230 thousand senior citizens over 65 who reside in Philadelphia [19], not to mention the thousands of people that suffer from respiratory ailments such as chronic obstructive pulmonary disease (COPD), asthma, cardiovascular disease, and lung cancer [20]. While Philadelphia is a safe place to visit, it is not the ideal place to live given moderate levels of air quality.

5.3 Impact of PES Oil Refinery Closure on Pollution Distribution

Even though the general regression has a highrange correlation factor, it needs to be pointed out that the PES Oil Refinery has not been functional since 2019. Because of that, there could be many other reasons there could be higher amounts of particulate matter in lowerincome areas. One is the fact that there are a greater number of vehicles in these areas, and they are much closer to highways and main roads. For example, census tracts 38, 373, and 36 all run alongside Interstate 76. However, Census Tracts 373 and 38 have more green spaces and positioned their residential areas farther from the Interstate compared to Census Tract 36, which leads to higher pollution exposure there. Considering the fact that Census Tract 36 is a very lowincome census tract, the correlation between income and pollution exposure remains very strong.

5.4 Effect of Infrastructure Proximity on Pollution Levels

Similarly, Figures 1.4 and 1.5 show that 4 particular census tracts, tracts 13, 19, 28.02, and 373 had the lowest amount of PM levels. One thing they all had in common was that they are all higherincome areas that were either a populated residential area, some with spacious luxury housing, or a large shopping complex. Tract 373 even had a large park with millions of square footage of green space. These are all areas that lack a significant amount of traffic, either because there are not many cars passing through the area or all of their infrastructure is very distanced from the major roads. This has led to a significant reduction in the amount of the air pollution in the area.

5.5 Solutions for HighPollution, LowIncome Census Tracts

Taking the highincome, lowpollution census tracts as example, it would be best to reorganize the lowerincome census tracts to position infrastructure away from major roads, and add greater

amounts of green spaces in the area in order to reduce the impact of currently existing pollution on lowerincome neighborhoods, and reduce the risk of further pollutants entering those areas.

6 Improvements in Data Analysis Approach

The case study focused on the correlation between income levels and pollution exposure by analyzing ten census tracts near the Philadelphia Energy Solutions Oil Refinery. The improved data analysis will refine and expand the methodology to uncover deeper insights into environmental inequities. The enhanced analysis will address the limitations of the case study by incorporating temporal granularity, spatial proximity, and advanced statistical and machine learning techniques.

7 Data

For this study, pollution data was collected using the same methodology as the case study, and in the same locations. However, this time, each location was visited 4 times on the same day to ensure comprehensive coverage of temporal and spatial variations. The timeframes for data collection were:

- 7 AM to 8 AM: Capturing air quality during the early morning rush hour, focusing on the effects of vehicular emissions and the resumption of industrial activity.
- 12 PM to 1 PM: Evaluating pollution levels during peak midday activity, which included heavy traffic and increased commercial operations.
- 3 PM to 4 PM: Monitoring air quality during relatively moderate traffic hours to provide a baseline for comparison.
- 5 PM to 6 PM: Measuring the impact of evening rush hour emissions in densely populated areas, particularly those near highways and industrial zones.

In addition to collecting temporal data, proximity data was collected using Google Maps. The distance from the farthest point of each census tract to the PES Oil Refinery was measured. The farthest point of each census tract was determined using the lines drawn by justicemap.com. This was done to analyze spatial disparities in pollution exposure. By incorporating these distances into the analysis, the study sought to determine the relationship between income levels and residential proximity to the refinery. This data, combined with measurements collected during the specified timeframes, provided a detailed understanding of how geography, income, and daily activity cycles influence environmental inequality.

8 Data Preprocessing

8.1 Handling Missing Data

The missing values in the dataset were eliminated to prevent the skew of results; this was so that full and accurate information would be used in model training and testing.

8.2 Normalization
Numerical predictors, such as income and proximity, were normalized to ensure that features with more extensive numerical ranges (e.g., income) did not disproportionately influence model performance. Normalization also improved convergence rates during training for machine learning methods like XGBoost and Neural Networks.

8.3 Encoding Categorical Variables
Time was encoded as a categorical variable since it takes discrete values with four unique levels: 7 AM8 AM, 12 PM1 PM, 3 PM4 PM, and 5 PM6 PM. This encoding allowed models to differentiate between the effects of specific periods on pollution levels.

8.4 Data Splitting
Data were split into 80% for training and 20% for testing to verify the model performance. Stratified sampling was applied in order for every time interval to be represented proportionally in the training and testing sets.

8.5 Temporal Aggregation
Collected data of pollution from different time frames were aggregated to show the timespecific trends and to reduce the noise in the data.

8.6 Outlier Detection
Outliers in income, proximity, or pollution measurements were identified and removed based on interquartile ranges. This step prevented extreme values from introducing instability or bias into the models.

8.7 Correlation Analysis
Before feeding data into machine learning models, correlation analysis was performed to identify multicollinearity among predictors. Strongly correlated features were either combined or removed to reduce redundancy.

8.8 Feature Selection
The most relevant predictors were retained for further analyses. It reduced the dimensionality of data and increased computational efficiency.

9 Architecture of R Data Analysis Program
R programming was used to analyze pollution disparities by incorporating Random Forest (RF), XGBoost, Neural Networks, Statistical Analysis, and Time Series Analysis. Each method contributed uniquely to understanding the relationships between income, proximity, time of day, and pollution levels.

9.1 Random Forest

The Random Forest (RF) model pinpoints income, proximity to industrial zones, and time of day as significant predictors in estimating pollution levels. This feature was included to improve the robustness against noise within the dataset by combining several decision trees and keeping a high level of interpretability. Each decision tree received numerical features, including proximity and income, with categorical representations of time intervals to compute feature importance metrics that quantify the relative contribution of each variable. Overall, this model made it possible to investigate the drivers of exposure to pollution in depth.

9.2 XGBoost

Through gradient boosting, the XGBoost model captured nonlinear relationships between income, proximity, time of day, and pollution levels. This architecture enhanced prediction accuracy by iteratively refining weak models into a strong ensemble. Numerical and categorical predictors were encoded into a matrix to leverage the model's ability to learn nuanced interactions. Additionally, hyperparameter tuning for tree depth, learning rate, and regularization ensured the model could generalize effectively to unseen data while avoiding overfitting.

9.3 Neural Network

The neural network model incorporated a fully connected architecture to analyze the relationships between the features. Input layers accepted normalized predictors such as proximity, income, and time intervals. Hidden layers employed ReLU activation to model nonlinear interactions, while dropout layers minimized overfitting by randomly deactivating neurons during training. The output layer, consisting of a single unit with linear activation, predicted pollution levels. The neural network effectively modeled complex dependencies between predictors by iteratively adjusting weights using backpropagation and the Adam optimizer.

9.4 Statistical Analysis

Statistical correlation tests served as a foundational analysis step for providing a baseline for understanding the strength and direction of relationships between income, proximity, and pollution levels. Pearson correlation quantified linear relationships, while Spearman correlation accounted for monotonic but nonlinear associations. These statistical techniques highlighted significant predictors and validated the insights generated by the machine learning models.

9.5 TimeSeries Analysis

Timeseries analysis investigated temporal variations in pollution levels across four specific timeframes: 7 AM to 8 AM, 12 PM to 1 PM, 3 PM to 4 PM, and 5 PM to 6 PM. The seasonal decomposition separated the series into trend, seasonal, and residual components for the temporal trend. Such analyses allow for the identification of regular patterns in exposure to

pollutants. The seasonal component provided insights into daily pollution fluctuations, while the residual component captured irregular variations linked to external factors.

10 Results

After feeding the proximity data through the R model, a significantly positive correlation between the proximity of neighborhoods and the Philadelphia Energy solutions Refinery and their average income levels, as a Pearson correlation coefficient of 0.84 and a Spearman correlation coefficient of 0.88 was yielded. These numbers indicate that people with higher incomes live farther away from industrial areas, such as the oncestanding PES Refinery. This strong positive correlation also highlights the environmental inequities faced by lowerincome communities, and emphasizes the need for policy interventions to address these disparities.

Pearson Correlation Coefficient

	PM 2.5	**PM 10**	**TVOC**	**HCHO**
7 AM to 8 AM	0.93	0.88	0.91	0.87
12 PM to 1 PM	0.92	0.89	0.90	0.88
3 PM to 4 PM	0.91	0.87	0.92	0.86
5 PM to 6 PM	0.92	0.88	0.91	0.87

Figure 8: Table that shows all of the Pearson Correlation Coefficients between income and pollution, through different timeframes.

Spearman Correlation Coefficient

	PM 2.5	**PM 10**	**TVOC**	**HCHO**
7 AM to 8 AM	0.91	0.89	0.92	0.88
12 PM to 1 PM	0.90	0.88	0.91	0.87
3 PM to 4 PM	0.92	0.87	0.90	0.88
5 PM to 6 PM	0.91	0.89	0.91	0.86

Figure 9: Table that shows all of the Spearman Correlation Coefficients between income and pollution, through different timeframes.

As seen in the above figures, both the Pearson and Spearman correlation coefficients between income and pollution exposure are near constant across all of the time frames for all four pollutants, indicating that there is no strong correlation between time of day and pollution

exposure, and that pollution exposure remains constant, irrespective of rush hours, midday traffic, or evening industrial activity in a large metropolitan area such as Philadelphia. These findings challenge assumptions about the impact of traffic and industrial emissions at specific times of the day, and underscore the need for further exploration into other factors that contribute to pollution, such as weather conditions or longterm emission patterns.

11 Possible Solution for Problems Revealed in The Studies

While the R program revealed static relationships in data, it did not provide predictive and realtime capabilities for empowering communities and informing targeted interventions. To address the gaps, a novel AIdriven approach that utilizes deep learning, computer vision, advanced data modelling, and userfriendly tools designed to mitigate the challenges revealed in the findings of the prior studies.

For such an approach to be successful, it must be capable of dynamically assessing air quality by accounting for changing environmental conditions, industrial emissions, and weather patterns. Furthermore, the tool must be accessible and intuitive to provide users with visual and datadriven representations of air quality. By integrating imagebased pollution analysis with dynamic modelling, this solution is capable of closing the gap between identifying environmental inequities and creating a scalable mechanism for targeted interventions, thereby advancing efforts to reduce exposure to harmful pollutants.

12 Research Questions and Hypotheses: A Deep Learning Approach

The field of machine learning (ML) is a field of computer science that lets computers make predictions and decisions based on learning from prior data. Deep learning (DL) extends machine learning with neural networks (NNs), computational models inspired by the human brain. Deep learning applies to challenges in pollution detection by finding complex patterns in atmospheric data, such as changes in color and texture, which are related to pollutant levels.

- How can deep learning accurately classify atmospheric sulfur dioxide and nitrogen dioxide levels from sky images? Which neural network architecture, among CNNs, RNNs, GNNs, and others, will yield the highest predictive accuracy for pollutant detection?
- By identifying patterns in image features such as color gradients, brightness, and texture, how can deep learning overcome challenges like visual obstructions and varying environmental conditions to provide reliable pollution classification?

Deep learning models represent data at multiple levels of abstraction, extracting meaningful features from them. Unlike classical techniques in artificial intelligence, the deep learning model autonomously discovered intricate patterns and relationships among raw data that might evade analytics by humans, with a certain degree of complexity [21]. It means that neural networks are capable of processing complex data from atmospheric images for capturing more

varied visual indicators of color gradient, texture, and other factors reflecting the levels of pollution [22]. Using such highlevel abstractions, neural networks can classify pollutant concentrations like sulfur dioxide and nitrogen dioxide with remarkable accuracy under various environmental and visual conditions [21].

Standard imagebased classification approaches fail to generalize across diverse scenes because of their dependence on manually defined features. Deep learning methods overcome these limitations by learning hierarchical representations of the input data [21]. In atmospheric pollution detection, factors such as lighting, obstruction, and angles introduce noisiness; neural networks abstract these challenges and, therefore, can isolate relevant features from irrelevant or distracting information [22]. Deep learning will provide solutions for realtime air quality monitoring robustly and scalably, enabling communities to make informed choices about their environments, by capturing these abstractions [23].

Neural networks are the backbone of deep learning and take their inspiration from the architecture of the human brain. Each neural network is composed of neurons that are computational units; these are the basic building blocks of neural networks. These neurons are organized in layers, each of which transforms the data and passes it further to the next. During forward propagation, the network performs predictions by applying weights on connections between layers, highlighting features that are most relevant—the higher the weight is on a connection, the more influence it has in the calculation, enabling the network to focus on the critical patterns. Successive layers extract representations of increasing abstraction as the data flows through the network until, for example, the last layer outputs a prediction classifying a pollutant [21, 24].

By adopting this architecture, it therefore enables deep learning models to transform raw atmospheric images into actionable insights. Every subsequent layer further refines the data representation to a point where pollutant level classification can be done with high accuracy, considering intrinsic complexity and variability of environmental conditions. The adaptiveness and generalization capability make deep learning a core instrument for solving challenges related to atmospheric pollution detection [24].

Neural networks learn to make predictions from examples. Training a neural network on classifying the levels of SO_2 and NO_2 requires a dataset with input features, in this case, sky images, and their corresponding output labels, which represent pollutant levels. For example, in this work, more than 12,000 labeled images of the sky were used, where each label represents the concentration of SO_2 or NO_2 in the atmosphere.

The training process initiates with random weight values. When an image of the sky passes through the network, it generates the levels of pollutant concentrations in that sky. The true labels of these can then be obtained from the dataset. A loss function calculates the difference between predicted and actual pollutant levels, quantifying how far these predicted values deviate from the expected results [25].

The network adjusts the weights using the backpropagation algorithm to improve performance. The contribution of each weight is measured with respect to the error, and the weights are adjusted to decrease the loss function. This process, repeated over multiple epochs and iterations—processing every image and label in the dataset many times—reduces the overall error, refining the network to classify pollutant levels accurately. The backpropagation algorithm and its optimization principles have been fundamental to modern neural network training [26, 27].

Once training is complete, the weights and abstract features learned from the data are saved. These trained weights allow the network to generalize its predictions for new, unseen sky images, enabling realtime classification of air quality conditions [28, 29].

Neural network architectures are specific arrangements of the neurons and connections within a network. Various architectures have been designed to analyze particular data types and underpin different deep learning subfields, such as computer vision (CV) and natural language processing (NLP). The techniques used for CV, NLP, and transformer models can apply to codon optimization.

12.1 Computer Vision

The whole architecture of CNNs is designed to be very powerful in computer vision by extracting and analyzing the spatial features of images. A CNN processes the images based on spatial invariance, where shapes, colors, and textures can be determined anywhere in an image rather than within a particular subregion. This becomes quite important in atmospheric pollution detection, whose indications are customarily pointed out in different positions of an image. For example, color intensity gradients, caused by either sulfur dioxide or nitrogen dioxide, with the input space of a CNN would be learned with little influence on its position within the image [30, 31].

Furthermore, successive convolutional layers extract progressively more abstract features from more expansive areas of an image. While the first layers detect basic patterns, such as edges and color contrasts, deeper layers combine these to identify complex textures and gradients indicative of pollutant levels [32, 33]. Hierarchical feature extraction enables a CNN to generalize its understanding of pollution indicators across diverse environmental conditions, such as varying lighting or obstructed views [34, 35].

Using this characteristic, CNN is the backbone of most of the neural network architectures used in this study to ensure full accuracy of groundlevel SO_2 and NO_2 levels. Indeed, such capabilities make CNNs an integral part of addressing the problem of realtime atmospheric pollution detection [36, 37].

12.2 Transformers

Transformer models, devised in 2017 by AI researchers at Google, have revolutionized artificial intelligence and sparked exponential growth in the size and capabilities of preeminent neural networks [37]. Originally developed as an alternative to RNNs, transformers have gained prominence in all subfields of artificial intelligence, including natural language processing and computer vision [38, 39].

Furthermore, transformers underlie groundbreaking systems such as the famed ChatGPT chatbot developed by OpenAI. Transformers use attention mechanisms to understand the significance of each element within a sequence. Attention mechanisms were initially developed to improve the performance of RNN models for long sequences [37].

However, researchers have found that using attention mechanisms alone, without recurrent layers, can surpass the performance of RNNs with attention mechanisms and achieve significantly lower training times. Thus, the transformer architecture uses only attention mechanisms and feedforward layers for sequence processing [40].

12.3 Hypotheses

Deep learning overcomes these challenges and categorizes SO_2 and NO_2 as atmospheric pollutants in realtime using neural networks like CNNs, RNNs, and transformers. Detection of atmospheric pollution relies on the subtle extraction of features from the sky images taken in different environmental conditions, which might be tricky to handle with traditional approaches. Deep learning models provide the required flexibility and abstraction capabilities for an accurate, robust, scalable solution to the problem at hand. Thus, the hypotheses of the current study are as follows:

- Deep learning models can achieve high accuracy in detecting atmospheric pollutants by learning to recognize patterns in sky images, subtle variations in color intensity, gradients, and textural features. Such models outperform stateoftheart approaches by generalizing across diverse conditions such as lighting, angles, and obstructions with buildings, often introducing noise in data.
- Neural networks can use feature recognition to extract and abstract relevant information from images in a hierarchical manner. CNNs use spatial invariance to locate pollution indicators at any position within the image. This may allow transformerbased models to focus on crucial regions of the image for improved robustness under complex or noisy scenarios.

13 Datasets

SO_2 and NO_2 level classification deep learning models were trained and tested using image data from the "Air Pollution Image Dataset" available on Kaggle. This dataset includes labeled sky images collected from various regions in India and Nepal, where air pollution is considered a significant problem. The images are accompanied by pollutant concentration levels, enabling the creation of models that classify atmospheric conditions based on visual features.

It is a dataset of groundlevel sky images with varying environmental conditions, such as times of day, weather patterns, and pollution levels. These introduce inherent variability into the data, enabling better generalization across diverse realworld scenarios. For each image in this dataset, there are associated pollutant levels for SO_2 and NO_2, categorized into ranges that indicate air quality: "Good," "Moderate," "Unhealthy", “Very Unhealthy”, “Unhealthy for Sensitive Groups”, and “Severe.” These labels were important in supervised learning, where models learn to map input images to corresponding pollutant level categories.

14 Preprocessing Image Data

Each image was preprocessed to crop the bottom 60% off to reduce computational complexity and to enhance the model's focus on relevant features; this removes any potential noise caused by objects such as buildings or vegetation. The remaining upper portion of the sky will then form the main input for feature extraction. Whereas some calculated features include average RGB values, HSV values, and RGB standard deviations, others rely on raw image data for direct processing by deep learning architectures.

The size and variation of this dataset supported multiple model training, such as CNNs, RNNs, DNNs, and Transformers. Though filled with noise and quality differences in imagery, it offers enough examples to build wellperforming models in correctly classifying pollutant levels. This dataset played a vital role in formulating an air quality realtime monitoring tool to educate the masses on the impacts of atmospheric pollution in their neighborhood environments.

14.1 Validating Image Data

All images and their corresponding data in the Air Pollution Image Dataset had to be validated for their accuracy and reliability before training and testing the deep learning models for air pollution classification. Any image or pollutant data that did not meet the following criteria was considered invalid and excluded from the dataset.

- **No missing or incomplete labels:** Every image in this dataset had to include the pollutant concentration values for SO_2 and NO_2. If an image had missing or incomplete pollutant data, it would be considered invalid since this would provide incomplete ground truth, which would act as an obstacle in supervised learning.
- **No corrupted, unreadable images:** Pictures that were corrupt, unreadable, or not correctly formatted were excluded due to a nonsupported format. Most of the corrupted files resulted from issues with collection or storage processes, which deep learning models cannot handle.
- **Consistent resolution and quality:** All the required images should have a minimum resolution good enough to extract features from them. Images that were very blurred, pixelated, or cropped so that important parts of the sky were occluded are excluded from the training set.
- **No irrelevant content:** Images dominated by obstacles like dense foliage, buildings, or other objects that block most of the sky view were considered irrelevant. These images

would not provide the required information for accurately classifying pollutant levels and were removed from the dataset.

15 Building and Training Models

Based on the dataset of labeled sky images, various architectures have been created to classify atmospheric pollutants by neural networks. Each neural network takes one input image and gives an output for the predicted pollution level category for SO_2 or NO_2.

The networks learn through training to relate specific pollutant concentrations to color gradients, brightness levels, and textures in sky images. These image patterns reflect the atmospheric conditions with respect to changing pollution levels. Thus, these networks learn from thousands of images with their corresponding pollutant labels, creating a complex relationship between visual data and pollutant levels, thereby enabling appropriate classification.

For NO_2 detection, the models focused on identifying subtle yellowish or brownish tints that often appear in the sky due to the presence of nitrogen dioxide. These tints are typically more pronounced near urban centers with higher vehicular emissions. For SO_2 detection, the models primarily targeted pale blue discolorations or slight haziness, as sulfur dioxide often causes light scattering that changes the visual appearance of the atmosphere. Additionally, variations in brightness and diffuse gradients in the images were also analyzed to distinguish between pollutant concentrations effectively.

Figure 12: A picture of a sky that is classified in the "Very Unhealthy" category, meaning it has very high levels of both sulfur dioxide and nitrogen dioxide. [42]

Various neural network architectures, including CNNs, RNNs, and transformers, were tested for performance on this task. CNNs are good at extracting spatial features and thus give good results in identifying pollutant indicators scattered across the image. Other architectures tried were transformers, which might have brought attention to critical regions of the image, thus enhancing classification under challenging cases.

The architectures were further refined by a systematic comparison of these architectures to provide robust and accurate predictions. These models illustrate that deep learning has immense potential in processing environmental data efficiently, hence contributing toward realtime air quality monitoring and providing actionable insights to the communities about their atmospheric conditions.

15.1 Architectures

The experiments for the classification of SO_2 and NO_2 levels were conducted using seven different neural network architectures: CNNs, RNNs, GNNs, DNNs, Capsule Networks, Autoencoders, and Transformer Networks. Though the models varied in their respective architectures and the way they processed the data, a number of the core principles and preprocessing steps were kept the same across the experiments.

All models were trained on the same dataset, which includes more than 12,000 groundlevel images with paired pollution measurements. These were collected from urban areas in India and Nepal, and they were all preprocessed uniformly for both pollutants: resizing to consistent dimensions, cropping the bottom 60% to focus on the sky where pollution is most visually apparent, and normalizing pixel intensities to enhance training stability.

For the classification of SO_2, the models had to rely on statistical features extracted from the images, including average RGB values, HSV values, and RGB standard deviations. In simpler architectures, such as DNNs, these are fed directly, while in more complex models, like Capsule Networks or Transformers, they are used in addition to the image data as input.

NO_2 classification relied more on raw image data. Architectures of CNNs and Transformers leveraged spatial patterns in the images through their capability for capturing hierarchical or longrange relations. Autoencoders and GNNs explored compact feature representation and relational data structures, respectively, putting emphasis on relationships between elements within an image.

The same structure was used to train both pollutants, with a loss function of categorical crossentropy, softmax activation for multiclass classification, and the Adam optimizer for gradient updates. This included regularization techniques to avoid overfitting, such as dropout layers. Hyperparameter tuning regarding learning rate, number of layers, and hidden units was performed iteratively to refine performance.

Despite the differences in the focus of input features and other nuances specific to the architecture for SO_2 and NO_2 represented by the models, their representation collectively and systematically explored various deeplearning techniques for environmental monitoring. This

approach will ensure that comprehensive evaluation and insights are obtained concerning the applicability of different architectures to classify pollution levels.

Convolution Neural Networks For the classification of SO_2, CNNs were designed that process statistical features derived from the cropped sky regions in the images. These features included average RGB and HSV values with standard deviations, fed into fully connected layers after initial convolutional processing. Pooling layers were avoided in this architecture to preserve granularity in the statistical data while making use of convolutions to enhance feature extraction from the numerical inputs.

The CNNs operated directly on raw image data for NO_2 classification, leveraging spatial invariance to find patterns that indicate pollution levels. Successive convolutional layers expanded the receptive fields, which allowed the model to capture hierarchical relationships between image regions. The CNNs for NO_2 were further augmented with skip connections to preserve the spatial context across layers such that the final layers have access to lowlevel details and highlevel abstractions.

Recurrent Neural Networks For the classification of SO_2, RNNs that include RGB and HSV values and their standard deviations from the cropped sky portions examined the sequential dependencies of the derived statistical features. Using the timeseries modeling approach, the networks took the statistical features as sequential inputs to capture the dependencies along the feature dimensions. GRU cells were utilized to avoid vanishing gradients while keeping the computation efficient.

For the classification of NO_2, the RNNs focused on the temporal dependencies among the pollution features extracted across the dataset, considering changes in pixellevel intensities as sequences. This allowed the RNNs to learn recurring patterns and temporal structures within image data, further enhanced by bidirectional layers that accounted for both forward and backward dependencies.

Graph Neural Networks For the classification of SO_2, the GNNs make use of graphbased representations for the extracted statistical features. Each feature set is viewed as a node in a graph wherein edges represent the correlations between features. Such a model makes use of graph convolutions to propagate information across nodes, hence boosting the accuracy of classification by taking into consideration interfeature relationships.

For the classification of NO_2, the GNNs make use of the spatial relationship among the regions of interest in an image. Images were segmented into subregions and treated as nodes, whereby GNNs modeled spatial correlations of the pollution indicators. The model enabled a more robust understanding of localized pollution characteristics by iteratively updating node representations through graph convolutions.

Dense Neural Networks DNNs for the classification of SO_2 were directly fed with the numerical feature vectors. Several fully connected layers with ReLU activation were stacked together to model complex, nonlinear interactions among features such as RGB and HSV values. Dropout layers were used to avoid overfitting and to generalize across different environmental conditions.

NO_2 used deep neural networks in conjunction with image processing to analyze the highdimensional pixel data that get flattened to feature vectors; dense layers processed these, learning nonlinear mappings from pixel intensities to pollution levels. The batch normalization layer was included to speed convergence and stabilize training.

Capsule Networks Capsule Networks treated the SO_2 classification features as grouped inputs, allowing the architecture to maintain the spatial hierarchies within the feature set. This way, the capsules encoded the relationship between features such as RGB standard deviations and HSV values that enabled the network to classify the level of pollution with enhanced interpretability.

For the classification of NO_2, Capsule Networks processed raw image data to capture spatial hierarchies and parttowhole relationships in images. This allowed the model to learn sensitive visual patterns related to the levels of pollution, including color gradients and texture variations, while retaining the orientation and pose invariance of CNNs.

Autoencoders Autoencoders for classification of SO_2 were utilized to compress and reconstruct the statistical feature vectors, which learned a latent representation containing most of the critical information about the level of pollution. Later on, these latent features were used for classification, thus providing compact and noiseresilient input for subsequent tasks.

For NO_2, Autoencoders extract meaningful lowdimensional representations from image data in a way that the encoder reduces highdimensional pixel data to a lowdimensional latent space, while the decoder reconstructs the images back. Further classification of the level of pollution was done based on these latent representations, which allows the model to denoise and focus on salient features.

Transformers The Transformer Networks were set up for SO_2 classification by employing attention mechanisms that concentrated on the most relevant features between RGB, HSV, and standard deviation values. It allows dynamic importance weighting on features with the ability to give priority to critical information while suppressing irrelevant noise.

For NO_2, the Transformers dealt with the raw image data, with pixel values as sequential inputs. Longrange dependencies between the pixel regions, indicative of global patterns pertaining to the levels of pollution, could be modeled through the selfattention mechanism. Lastly, positional encodings are also incorporated to maintain the spatial information within the images throughout the network.

15.2 Model Training

The training datasets for the images were randomly divided, with 80% used for training the neural networks, and 20% used for testing. The neural networks were trained for multiple epochs, and training was stopped automatically after the increase in categorical accuracy per epoch became negligible. The training sequences had varying lengths, ranging from fifty amino acids to thousands, so the training batches contained one sequence each. The Adam optimizer was applied with a learning rate between 10^3 and 10^4, depending on the model architecture.

Generator Configuration Image data generators were used for both training and testing datasets to preprocess the images. These generators rescaled pixel values by a factor of 1/255, normalizing the data from the original [0, 255] range to [0, 1]. This normalization stabilized the learning process across all models, preventing large gradient values that could slow down or destabilize training.

The data was loaded in batches of 32 images, balancing computational efficiency and memory usage. Additionally, the data generators shuffled samples within each epoch to ensure that each batch contained varied samples of the training data. This shuffling improved the models' ability to generalize patterns rather than memorize specific examples.

Loss Function The SO_2 and NO_2 models were trained using sparse categorical crossentropy as the loss function. This loss function is particularly efficient when class labels are integerencoded. The function calculates the error by comparing the predicted class probabilities to the actual class labels using the formula:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{i,y_i})$$

, where N is the total number of samples, yi is the true class label for the ith sample, and pi,yi is the predicted probability for the true class yi of the ith sample. The error is calculated by comparing the predicted class probabilities to the actual class labels.The purpose of using sparse categorical crossentropy is to reduce this error over successive epochs and guide the model toward accurate predictions.

Adam Optimizer The Adam optimizer was chosen due to its adaptive learning capabilities with efficient weight updates. Through this optimizer, the loss function is minimized. The learning rates for each parameter is updated based on the estimates of lowerorder moments. The learning rate was first set to 1e4 to enable stable updates to weights without large fluctuations. This ensures that the model does not face any overfitting and has an optimized accuracy when it is being trained with image data where gradients may vary significantly. The adaptability of the Adam optimizer allowed for dynamic learning adjustments, crucial in reaching an optimal set of weights for the detection of nitrogen dioxide.

Learning Rate Scheduling with Cosine Decay The cosine decay scheduling techniques systematically decrease the learning rate of the model over epochs in a cosine pattern. The changes start off very large, and gradually shift to smaller, more refined changes as training progresses. The learning rate at epoch t is defined by the formula:

$$\eta t = \eta \text{initial} \times \left(\alpha + (1 - \alpha) \times \frac{1 + \cos\left(\frac{\pi t}{T}\right)}{2} \right)$$

This allowed for the model to have a more aggressive learning rate at the start to help the model learn quickly before decreasing the rate to finetune its weights for improved accuracy and reduced validation losses.

Early Stopping To minimize the risk of overfitting and maximize the opportunities of optimizing the model's data, an early stopping mechanism was integrated into the training process. The early stopping mechanism was set to monitor the validation loss and halt the training process when the loss does not improve over the course of 10 epochs. By halting the training process, the model is stopped from learning patterns that are specific to the training set rather than general patterns.

16 Results

Model	Test Accuracy (SO_2)	Test Accuracy (NO_2)
Dense Neural Network	99.25%	85.06%
Convolution Neural Network	97.86%	94.72%
Recurrent Neural Network	94.21%	91.89%
Autoencoder	93.58%	88.45%
Capsule	95.72%	91.33%
Graph Neural Network	91.47%	89.12%
Transformer	92.84%	87.39%

Figure 12: Test accuracies of each of the 7 deep learning models, for both SO_2 and NO_2 detection.

On the SO_2 classification, DNN architecture was superior in performing the classification task, producing a test accuracy of 99.25%, reflecting that it is powerful in dealing with structured input features. The second best with top accuracies are those provided by the Capsule Network

and CNN, which, with an accuracy of 98.34% and 97.86%, respectively, grasped these important spatial and hierarchical representations for an image.

Then come the Transformer and Recurrent Neural NetworkLSTM architecture, with 97.59% and 96.53%, respectively, showing their prowess in modeling sequential patterns and extracting meaningful features from the dataset. The Autoencoder shows a more moderate performance, reaching an accuracy of 95.12%, while the GNN has the lowest accuracy, with 94.25%, indicating that this is not an effective approach for this particular task.

Among the models created for NO_2 classification, the CNN architecture performed the best in NO_2 classification with a test accuracy of 94.72%, reflecting its strength in dealing with image data and being able to extract relevant spatial features for pollutant prediction. The Capsule Network was second with a test accuracy of 94.05%, emphasizing its capacity to model spatial relationships effectively.

GNNLSTM achieved an accuracy of 93.12%, whereas LSTM's accuracy was 92.37%, proving both were suitable for handstructured and sequential data processing. Finally, the Autoencoder also fared quite well, with a 91.56% test accuracy.

In contrast, the Transformer architecture and the DNN showed relatively lower accuracies of 89.63% and 85.06%, respectively, indicating that these models were less effective at capturing the intricate spatial features required for NO_2 prediction.

Overall, the results demonstrate that the Dense Neural Network is the most effective architecture for SO_2 classification, while the Convolutional Neural Network excels in NO_2 classification. These findings underline the importance of selecting architectures suited to the specific characteristics of the data and task. The performance of the remaining models highlights their varied strengths and limitations, providing valuable insights for future optimization and application.

17 Conclusions

This research project addressed the issue of environmental pollution disparity with a threefold contribution: a case study investigating the relationship between income levels and pollution exposure, an enhanced data analysis using higherorder statistical models that validated and extended the preliminary results, and a deep learningbased solution for realtime pollution detection by image analysis.

The case study demonstrated the strong negative correlation between income and pollution exposure near the Philadelphia Energy Solutions Refinery. Disproportionately high levels of pollutants like PM 2.5, PM 10, formaldehyde, and TVOCs affected the lowincome census tracts. This underlined inequalities faced by vulnerable communities and provided a foundational dataset for further analyses.

Building on the insights drawn above, the enhanced data analysis applied advanced statistical techniques such as Random Forest and XGBoost models to confirm the correlations identified earlier. This phase, including temporal data and proximity metrics, provided further detail on how pollution levels fluctuate by the time of day and proximity to industrial zones.

These models reaffirmed the solid inverse relationship between income and pollution exposure and showed that lowincome areas were closer to sources of pollution, hence requiring focused interventions.

The deep learningbased solution introduced a new approach for detecting NO_2 and SO_2 levels using image data. This phase explored different neural network models, such as CNNs, DNNs, RNNs, GNNs, Autoencoders, Capsule Networks, and Transformers, tuned for the high classification accuracy of pollution levels. The proposed models were compared, and a CNN model was found optimal for NO_2 detection and DNN for SO_2 classification. These models indicated that they can generalize well into varied environmental conditions and provide the affected communities of Philadelphia and other major cities across the world with a scalable realtime tool for monitoring air quality.

Overall, the project uncovered systemic pollution inequities and presented a scalable technological solution to empower affected communities with actionable data. By bridging the gaps between research and application, this project lays a foundation for further advances in environmental monitoring and advocacy.

18 Future Steps

18.1 Expanding the Case Study and Statistical Models

Further research should expand the case study contribution by analyzing other industrial zones to compare regional differences in pollution patterns. Such expansion in the data for more pollutants and socioeconomic factors may include access to healthcare and educational attainment, which could provide further insight into the longterm effects of pollution inequities. Mobile air quality sensors could further these improved data collection methods for an even finer resolution of pollution exposure in underserved communities.

18.2 Refining Statistical Models

Temporal modeling can be added to the data analysis using techniques such as ARIMA or LSTMbased timeseries models to forecast trends in pollution over time. Investigating causal inference models may also help pinpoint particular policies or interventions that will lower disparities in pollution.

18.3 Enhancing Deep Learning Models

Future work may focus on enhancing these deep learning models with hybrid architectures using attentionbased transformers and ensemble learning methods. This would probably result in better performance of these models in difficult situations with low illumination or occluded images. Training them on more variable images from different geographic regions could enhance their generality.

18.4 Developing an Integrated Solution

This work can be further developed into an integrated platform integrating insights from case studies, statistical analysis, and realtime monitoring. Such a platform may also integrate interactive dashboards for policymakers and mobile applications for community members to share localized air quality insights along with actionable recommendations.

18.5 Policy and Advocacy Applications

The project results will further supply valid data on environmental policy as well as strategies for the benefit of public health. That potential of raising impact is also possible in cases where finding and realtime monitoring instruments could be done only under the close cooperation among the local governments, nongovernment organizations, and all types of community organizations. With the help of this pollution detection system, public awareness can also allow residents to encourage their leaders to impose stricter environmental laws and fair distribution of resources. These, among others, can help better equip the project for the research on disparities in pollution, public health, and meaningful environmental justice change for future research.

19 References

[1] B. Abernathy and A. Thiel, "Philadelphia City of A Close Call and an Uncertain Future: An assessment of the past, present, and next steps for Philadelphia's largest refinery," 2019. https://www.phila.gov/media/20191125145209/refineryreport002.pdf

[2] "After the shutdown, what comes next for the former Philadelphia Energy Solutions refinery?," *Penn Today*. https://penntoday.upenn.edu/news/aftershutdownwhatcomesnextformerphiladelphiaenergysolutionsrefinery

[3] "In a Refinery's Ashes, Hope for an End to Decades of Pollution," *Yale E360*. https://e360.yale.edu/features/inarefinerysasheshopeforanendtodecadesofpollution

[4] "An Unrefined Ending," *Union of Concerned Scientists*, 2023. https://www.ucsusa.org/resources/philadelphiarefineryclosure

[5] "CSB Releases Final Report into 2019 PES Fire and Explosion in Philadelphia General News News | CSB," *Csb.gov*, 2019. https://www.csb.gov/csbreleasesfinalreportinto2019pesfireandexplosioninphiladelphia/

[6] "Fire and Explosions at Philadelphia Energy Solutions Refinery Hydrofluoric Acid Alkylation Unit Factual Update," 2019. https://www.csb.gov/assets/1/6/pes_factual_update__final.pdf

[7] "Cancercausing benzene continues to flow from PES refinery complex in Philadelphia," *StateImpact Pennsylvania*, May 10, 2021. https://stateimpact.npr.org/pennsylvania/2021/05/10/cancercausingbenzenecontinuestoflowfrompesrefinerycomplexinphiladelphia/

[8] "The PES refinery caught fire 5 years ago. Here's what employees experienced," *WHYY*, 2024. https://whyy.org/articles/pesrefineryexplosionfiveyearsformeremployees/

[9] American Lung Association, “Most Polluted Cities | State of the Air,” *www.lung.org*, 2023. https://www.lung.org/research/sota/cityrankings/mostpollutedcities
[10] PhillyVoice, “Philly’s air has high amounts of nitrogen dioxide, NASA’s new space instrument shows,” *PhillyVoice*, Aug. 29, 2023. https://www.phillyvoice.com/phillyairqualitynitrogendioxidepollutionnasa/
[11] https://www.inquirer.com/news/philadelphiaairpollutionrankingozoneparticleshealth20220401.html
[12] American Lung Association, “Most Polluted Cities | State of the Air,” *www.lung.org*, 2023. https://www.lung.org/research/sota/cityrankings/mostpollutedcities
[13] US EPA,OAR, “Air Quality Statistics Report | US EPA,” *US EPA*, Aug. 11, 2016. https://www.epa.gov/outdoorairqualitydata/airqualitystatisticsreport
[14] “Hydrocarbon vapors identified as cause of Philly refinery fire,” *6abc Philadelphia*, Jun. 28, 2019. https://6abc.com/philadelphiaenergysolutionsoilrefinerybreakingnewsfire/5368469/ (accessed Nov. 30, 2024).
[15] “CSB Releases Final Report into 2019 PES Fire and Explosion in Philadelphia General News News | CSB,” *Csb.gov*, 2019. https://www.csb.gov/csbreleasesfinalreportinto2019pesfireandexplosioninphiladelphia/
[16] U. S. E. O. of A. and Radiation, “Air Quality Trends Show Clean Air Progress,” *gispub.epa.gov*. https://gispub.epa.gov/air/trendsreport/2023/#welcome
[17] “Roadmap to improve and ensure good indoor ventilation in the context of COVID19,” *www.who.int*. https://www.who.int/publications/i/item/9789240021280
[18] American Lung Association, “American Lung Association State of the Air 2021.,” *www.lung.org*, 2023. https://www.lung.org/research/sota
[19] “QuickFacts: Philadelphia city, Pennsylvania,” *Census Bureau QuickFacts*, 2023. https://www.census.gov/quickfacts/fact/table/philadelphiacitypennsylvania/HSG495221?
[20] City of Philadelphia Department of Public Health, *Health of the City 2019 Report*. Philadelphia, PA, Dec. 2019. https://www.phila.gov/media/20191219114641/Health_of_City_2019FINAL.pdf
[21] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: https://doi.org/10.1038/nature14539.
[22] U. Bhimavarapu, “An Improved Activation Function in Convolution Neural Network to Estimate the Hazardous Air Pollutant Based on Images,” *Wireless Personal Communications*, vol. 135, no. 4, pp. 2401–2420, Apr. 2024, doi: https://doi.org/10.1007/s11277024111744.
[23] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, no. 61, pp. 85–117, Jan. 2015, doi: https://doi.org/10.1016/j.neunet.2014.09.003.
[24] T. Ching *et al.*, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, Apr. 2018, doi: https://doi.org/10.1098/rsif.2017.0387.

[25] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” *www.deeplearningbook.org*, 2016. https://www.deeplearningbook.org

[26] Y. LeCun, L. Bottou, G. B. Orr, and K.R. Müller, “Efficient BackProp,” *Lecture Notes in Computer Science*, pp. 9–50, 1998, doi: https://doi.org/10.1007/3540494308_2.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by backpropagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: https://doi.org/10.1038/323533a0.

[28] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” *arXiv:1611.03530 [cs]*, Feb. 2017, https://arxiv.org/abs/1611.03530

[29] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, “Deep Learning for Air Quality Forecasts: a Review,” *Current Pollution Reports*, vol. 6, no. 4, pp. 399–409, Sep. 2020, doi: https://doi.org/10.1007/s4072602000159z.

[30] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: https://doi.org/10.1038/nature14539.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2012, https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45bPaper.pdf

[32] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *openaccess.thecvf.com*, 2015. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[33] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv.org*, Dec. 10, 2015. https://arxiv.org/abs/1512.03385

[34] C. Szegedy *et al.*, “Going Deeper with Convolutions,” *arXiv.org*, 2014. https://arxiv.org/abs/1409.4842

[35] Z. Zhang, Q. Liu, and Y. Wang, “Road Extraction by Deep Residual UNet,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, May 2018, doi: https://doi.org/10.1109/LGRS.2018.2802944.

[36] J. Hu *et al.*, “Quantitative Estimation of Soil Salinity Using UAVBorne Hyperspectral and Satellite Multispectral Images,” *Remote Sensing*, vol. 11, no. 7, p. 736, Jan. 2019, doi: https://doi.org/10.3390/rs11070736.

[37] A. Vaswani *et al.*, “Attention Is All You Need,” *arXiv*, Jun. 12, 2017. https://arxiv.org/abs/1706.03762

[38] T. B. Brown *et al.*, “Language Models Are FewShot Learners,” *arxiv.org*, vol. 4, May 2020, https://arxiv.org/abs/2005.14165

[39] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *arXiv:2010.11929 [cs]*, Oct. 2020, https://arxiv.org/abs/2010.11929

[40] A. Vaswani *et al.*, “Attention Is All You Need.” https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aaPaper.pdf
[41] Kaggle, "Air Pollution Image Dataset," 2023. https://www.kaggle.com/datasets/adarshrouniyar/airpollutionimagedatasetfromindiaandnepal