Machine Learning for Air Quality Prediction in Alabama: An Environmental Justice Case-study

Arnav Maskey
Bob Jones High School
Madison, AL 35758
arnav.maskey@gmail.com

Abstract

Environmental justice means the fair treatment of all people from environmental impacts, regardless of race, color, national origin, education level, or income. This paper investigates using machine learning (ML) to model environmental justice issue caused by air quality disparities in Jefferson County, Alabama, particularly around Birmingham. This paper contributes to (1) ML-ready datasets curated from environmental sensors for air quality analysis and socio-economic sources, and (2) utilizes novel ML methods to identify, predict, and analyze areas with significant environmental injustice. It demonstrates the usability of this approach for extending it to analyze similar issues globally, enabling rapid identification and offering predictive capabilities for timely mitigation. The code, data, and exploratory data analysis related to this study are publicly available at https://github.com/amaskey/ej-aqi.

1 Introduction

Environmental Justice (EJ) addresses the disproportionate impact of environmental hazards on marginalized communities (groups of people who face discrimination based on their characteristics, such as race, gender, religion, or socioeconomic status). The US Environmental Protection Agency defines it as the fair treatment and meaningful involvement of all people, regardless of race, color, national origin, or income, in environmental law development and enforcement [3]. The EJ Advisory Council identifies affected communities as those with significant populations of vulnerable groups at risk of adverse health or environmental outcomes [2].

Typically, environmental hazards leading to injustices include natural disasters, drought, air and water pollution, urban heat exposure, and proximity to toxic emissions. Addressing these requires an interdisciplinary approach, combining Earth science, socio-economics, health sciences, and more, to understand environmental conditions and at-risk communities. Advocates use heterogeneous datasets to address these issues. Improved data fusion and prediction techniques are needed to identify environmental injustice hotspots and measure impacts. Increasing awareness and demonstrating these impacts can encourage broader change.

Machine Learning (ML) has become a powerful tool for quickly analyzing complex datasets to identify and predict patterns. This paper explores using ML to address air quality issues in Jefferson County, Alabama and presents a ML-ready dataset along with benchmarking results. This region is known for its significant air quality problems and has historically faced environmental challenges that disproportionately affect low-income and minority communities. Living nearby, I have personally experienced its poor air quality.

This study's primary contributions are: (1) curation of comprehensive ML-ready datasets for air quality index (AQI) coupled with socio-economic information and (2) development and application of a Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) network to identify and predict EJ issues. The social impacts of this work include providing actionable insights into air quality

Preprint. Under review.

disparities using advanced ML techniques. It benefits marginalized communities with accurate early predictions and helps policymakers create timely action plans. Additionally, researchers benefit from ML-ready datasets for advancing EJ research. The rest of the paper is structured as follows: Section 2 covers related work, Section 3 introduces the method, Section 4 presents experiments and results, and Section 5 concludes with potential future work.

2 Related work

Although, the concept of EJ emerged in the 1980s, research on using ML for EJ is limited. Chakraborty and Maantay [4] illustrated the use of geographic information system (GIS) in mapping environmental hazards near vulnerable populations, a method that ML could enhance. Recent studies have applied various methods to predict environmental hazards. For example, Li et al. [9] introduced a deep learning fusion model for predicting PM2.5 concentrations, combining convolutional neural networks and deep bidirectional long short-term memory with an attention mechanism optimized by the sparrow search algorithm. Hogrefe et al. [7] analyzed regional air quality variations and their correlations with socio-economic indicators. They used simulations with atmospheric components and projections based on scenarios from the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emission Scenarios to predict air quality. In the context of Southeastern US, the focus area of my study, Gutierrez and LePrevost [5] highlighted the vulnerabilities of rural populations to climate change, stressing the importance of localized climate justice interventions to mitigate health risks from extreme heat and air pollution. Miranda et al. [10] highlighted that non-Hispanic Blacks are disproportionately present in areas with poor air quality, emphasizing the necessity for equitable air quality monitoring.

These past studies illustrate the importance of heterogeneous data integration and potential use of ML methods in addressing EJ issues. This research will focus on Birmingham, Alabama (Jefferson County), a region with a history of industrial activity affecting predominantly African American and low-income communities. It distinguishes from previous methods by compiling a comprehensive dataset and using ML method to identify and predict air quality disparities.

3 Method

Figure 1 shows the workflow for the proposed method, where two main contributions: data curation (left) and ML approach (right) are highlighted.

3.1 ML-ready data curation

In this study, air quality index (AQI) derived from the PM2.5 (particles less than 2.5 micrometers in diameter) [12] is considered. PM2.5 can severely irritate and impair lung function. Additional study has suggested that PM2.5 is directly correlated with daily mortality [11]. I have used AQI derived from environmental sensor data provided by DATAUSA[1] and population information derived from demographics data provided by USAFACTS[2]. The curated dataset includes 7 different locations with air quality sensors and population demographics in Jefferson County, AL (33.4914° N, 86.9824° W). A scaled ratio of AQI to racial population statistics was computed to account for correlations between AQI and socio-economic variables. A snapshot of the dataset is shown in Figure 2.

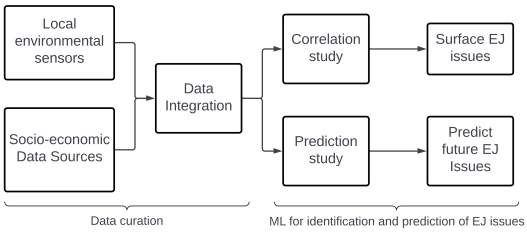

Figure 1: Workflow of overall data curation and ML approach for EJ.

3.1.1 Datasets

The study contributes two curated datasets. Dataset 1 consists of comprehensive daily records from the past 9 years for 7 locations in Jefferson County, AL. Each record includes the AQI and data

[1] https://datausa.io/profile/geo/jefferson-county-al

[2] https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population/state/alabama/county/jefferson-county/

on AQI variations relative to race and population. Dataset 2 aggregates daily data to annual level for the entire county over the same 9 years, excluding AQI. However, a new record is added to Dataset 2 incorporating population projection for 2025. Due to only annual updates of demographic information at the county level, annual data aggregation is necessary. Dataset 1 (about 15000 records) is used as an input for the CNN-LSTM model for training purpose. The trained model is used to predict AQI based on Dataset 2 and population projection. Figure 2 summarizes the AQI around 7 Zip codes in Jefferson County, Alabama. The Figure shows that 35207, a densely populated, predominantly African-American area near industrial sites, consistently experiences worse air quality than 35094. The time series reveal more days with unhealthy AQI in 35207 each year (red line indicates AQI=100).

3.2 CNN-LSTM model architecture

This study proposes a CNN-LSTM approach to predict AQI. Convolutional Neural Network (CNN) [8] layers extracts features from the input AQI data, which are provided as input to the Long Short-Term Memory (LSTM) [6]. CNNs are deep neural networks designed primarily for extracting features through spatial information; whereas LSTMs improve over Recurrent Neural Networks (RNNs) by incorporating a memory cell that retains information over time. LSTMs are able to identify long-term patterns by controlling three gates in the memory cell. The input gate controls what information goes into the memory cell, the forget gate can remove information from the cell, and the output gate controls what exits the memory cell. Since the input data includes variations over time, CNN-LSTM is an ideal method for curated AQI data. The CNN-LSTM architecture used in this study is shown in Figure 3.

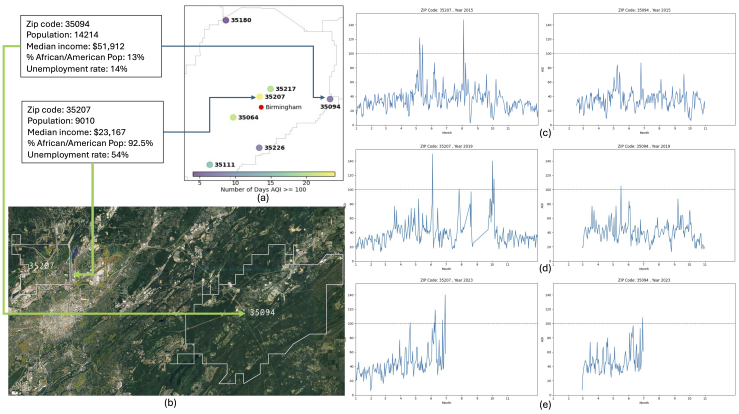

Figure 2: Summary view of air quality in Jefferson County, AL. (a): major air quality sensor locations and the days with AQI over 100 (unhealthy), (b): a satellite view of ZIP codes 35207 and 35094 with demographics information, (c)(d)(e): annual AQI timeseries (years 2015, 2019, and 2023), comparing 35207 (left) and 35094 (right).

4 Experiments and results

The experiment was conducted for 250 epochs with learning rate of 0.0001, mean squared error for loss function, and Adam optimizer. Training and validation losses stabilized after 150 epochs.

Overall AQI prediction: Daily AQIs for all locations in Jefferson County were used as input to the model. The correlation between predicted vs. true AQIs for all the test samples are shown in Figure 4. Table 1 shows the final metrics of the model after training.

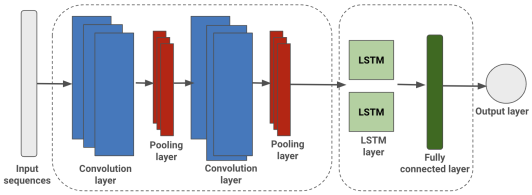

Figure 3: Proposed CNN-LSTM architecture

AQI prediction for communities grouped by race: The model also predicts AQIs with respect to marginalized communities as shown in Figure 5. As evident in Figure 5, the AQI follows a cyclic pattern throughout the times studied.

AQI with population projection: The model also predicts AQIs based on the projected population for 2025, which is derived from [1]. Based on the sequences learnt from the daily historical data the model estimates AQI values for 2025. The model learns the correlation between AQI and each population to

Table 1: Evaluation metrics.

Metrics	Result
Root Mean Square Error	8.499
Mean Absolute Error	5.15
R^2	0.6138
Overall Accuracy	89.64%

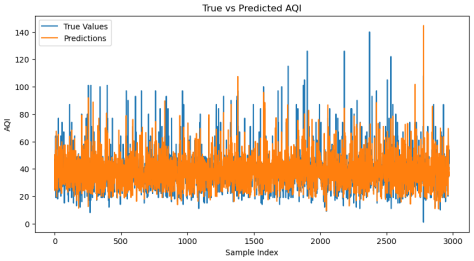

Figure 4: Predicted AQIs for the test samples overlaid on true AQIs

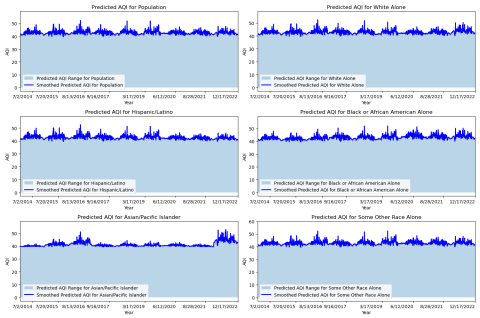

Figure 5: AQI prediction for different races

predict future AQI values for projected populations. This can surface problems that can arise in the future for each race within Jefferson County. The predicted mean AQI for Jefferson County for all residents for 2025 is 34.64.

5 Conclusion and future work

In this paper, I have successfully applied machine learning approach, especially CNN-LSTM, to environmental justice issues caused by poor air quality. I have also curated a comprehensive ML-dataset to study air quality coupled with demographics information. The dataset and the approach is reusable for other ML-based approaches. The results suggest that my approach can quickly identify potential environmental justice issues due to disparity in air quality and also predict future issues. In future, I intend to use inexpensive air quality sensors that I am currently building. These will be deployed in potential problematic areas. The data collected from these sensors will be correlated with satellite-based coarse resolution air quality measurements. The objective of the study will be to perform more granular temporal study (i.e., the hour of the day when AQI is particularly the worst) and expand the study to regions where air quality sensors are not available.

References

[1] Alabama Demographics 2013; Center for Business And Economic Research | The University of Alabama — cber.culverhouse.ua.edu. `https://cber.culverhouse.ua.edu/resources/alabama-demographics/`. [Accessed 06-01-2024].

[2] Environmental justice. `https://www.earthdata.nasa.gov/topics/human-dimensions/social-behavior/environmental-justice`. [Accessed 06-08-2024].

[3] Environmental Justice | US EPA — epa.gov. `https://www.epa.gov/environmentaljustice`. [Accessed 08-06-2024].

[4] J. Chakraborty and J. A. Maantay. Proximity analysis for exposure assessment in environmental health justice research. 2011.

[5] K. S. Gutierrez and C. E. LePrevost. Climate justice in rural southeastern united states: A review of climate change impacts and effects on human health. *International Journal of Environmental Research and Public Health*, 13(2), 2016.

[6] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[7] C. Hogrefe, B. Lynn, K. Civerolo, J.-Y. Ku, J. Rosenthal, C. Rosenzweig, R. Goldberg, S. Gaffin, K. Knowlton, and P. L. Kinney. Simulating changes in regional air pollution over the eastern united states due to changes in global and regional climate and emissions. *Journal of Geophysical Research: Atmospheres*, 109(D22), 2004.

[8] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[9] X. Li, N. Zou, and Z. Wang. Application of a deep learning fusion model in fine particulate matter concentration prediction. *Atmosphere*, 14(5), 2023.

[10] M. L. Miranda, S. E. Edwards, M. H. Keating, and C. J. Paul. Making the environmental justice grade: The relative burden of air pollution exposure in the united states. *International Journal of Environmental Research and Public Health*, 8(6):1755–1771, 2011.

[11] J. D. Schwartz, D. W. Dockery, and L. M. Neas. Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association*, 46 10:927–939, 1996.

[12] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian. The impact of pm2.5 on the human respiratory system. *Journal of Thoracic Disease*, 8(1), 2016.