Predicting Air Quality in Californian Counties

Abraham Yeung [1] Yashas Mattur [1]

[1]Department of Computer Science, Stanford University

Stanford | ENGINEERING
Computer Science

Predicting

Air quality forecasting is crucial for public health, especially in wildfire-prone California, where timely AQI predictions can help mitigate risks. We developed a hybrid model combining Graph Neural Networks (GNNs) and Long Short-Term Memory (LSTM) networks to predict daily AQI for 52 Californian counties using 10 years of historical data. Our model integrates spatial relationships through GNN embeddings and captures temporal dependencies using LSTMs. Compared to traditional machine learning baselines, our approach improved prediction accuracy, achieving lower RMSE. These results highlight the effectiveness of combining deep learning with spatial and temporal modeling for air quality forecasting.

Data

Our dataset comes from the U.S. Environmental Protection Agency's (EPA) Outdoor Air Quality API, covering daily air quality data from 52 Californian counties between 2015 and 2024. Each row represents a single county's daily air quality index (AQI), with columns including the date, county name, AQI value, and geographic coordinates (mean latitude/longitude of monitoring stations). The dataset is labeled with ground truth AQI values, and we applied the Interquartile Range (IQR) method to remove extreme outliers.

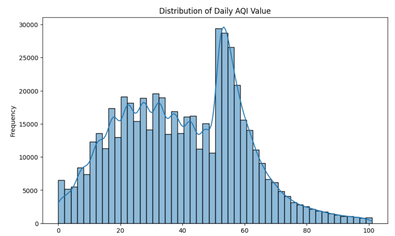


Figure 1. AQI Dataset Distribution after Outlier

Features

Our dataset includes both raw and derived features. Raw input features include AQI values, county name, date, and mean latitude/longitude of monitoring stations. We derived additional features, such as a k-nearest neighbor (k-NN) graph to model spatial relationships and Fourier-transformed components to capture seasonal patterns. These features are well-suited for our task because they enable our model to leverage both spatial dependencies (via GNN embeddings) and temporal trends (via LSTM), improving predictive accuracy.

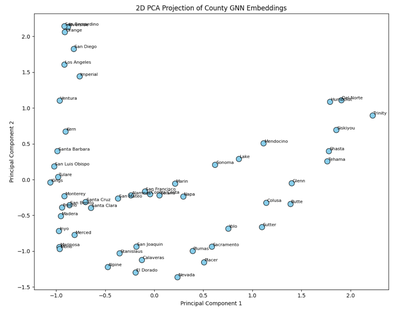


Figure 2. PCA of Geographical Embeddings based on Geographical Data

Models

Baseline and Classical ML Methods:
The baseline model predicts AQI using the last observed value with learned biases. Classical ML methods include XGBoost, CatBoost, and Random Forests, leveraging gradient boosting and ensemble learning.

Long Short-Term Memory Network:
LSTM captures long-range dependencies in sequential data using gating mechanisms and incorporates county embeddings for spatial context.

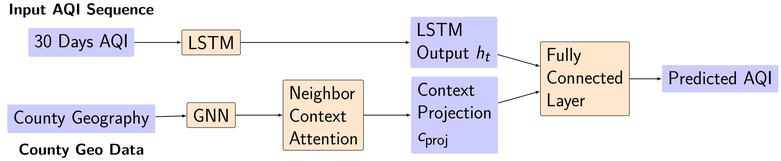
Our Novel Models:

1. *GNN-LSTM Model:*
   - GNN generates county embeddings $z_c \in \mathbb{R}^d$.
   - County-neighbor attention mechanism:

$$\bar{z}_c = \sum_{n \in N(c)} \frac{\exp(z_c^\top z_n)}{\sum_{j \in N(c)} \exp(z_c^\top z_j)} z_n \tag{1}$$

   - Final AQI forecast:

$$\hat{y}_{t+1} = W_{\text{fc}}[h_t \oplus c_{\text{proj}}] + b_{\text{fc}} \tag{2}$$

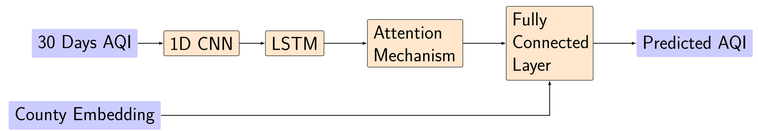


2. *CNN-LSTM with Attention:*
   - Extracts local features using 1D convolution.
   - Computes attention-weighted context vector:

$$c = \sum_{t=1}^{L} \frac{\exp(w_a^\top h_t)}{\sum_{s=1}^{L} \exp(w_a^\top h_s)} h_t \tag{3}$$

   - Final projection:

$$\hat{y}_{t+1} = W_{\text{fc}}(h_L \oplus z_{\text{proj}}) + b_{\text{fc}} \tag{4}$$



Results

The table below summarizes the training and test Root Mean Squared Error (RMSE) for each model we built. The training set contains 400,000 samples, while the test set contains 50,000 samples, ensuring a robust evaluation of model performance.

Model	Training Error (RMSE)	Test Error (RMSE)
Persistent Baseline	17.50	17.77
XGBoost	12.85	15.32
CatBoost	12.74	15.27
Random Forest	12.91	15.41
County-Conditioned LSTM	13.02	14.16
GNN-LSTM	12.89	14.12
GNN-LSTM + County-Neighbor Attention	12.78	14.10
CNN-LSTM + Hidden State Attention	12.70	14.08

Table 1. Training and Test Errors for Different Models

Discussion

Deep learning models outperformed baselines. Our best models: **LSTM with CNN Embeddings and Hidden State Attention** (RMSE 14.08) and **LSTM with GNN Embeddings and County-Neighbor Attention** (RMSE 14.10). GNN embeddings and county-neighbor attention improved spatial modeling, while CNN embeddings enhanced temporal pattern recognition. The combination of sequential modeling, spatial embeddings, and attention mechanisms allowed our models to capture both temporal and regional dependencies effectively.

Nevertheless, models struggled with extreme pollution events due to data imbalance and loss function bias. High-AQI volatility areas (e.g., Los Angeles) had larger errors, likely due to urban sprawl and localized pollution, which our dataset did not fully capture. Adding urban-specific features like traffic density and industrial emissions could improve predictions in these regions. Figure 3 shows RMSE variations across counties, highlighting areas for improvement.

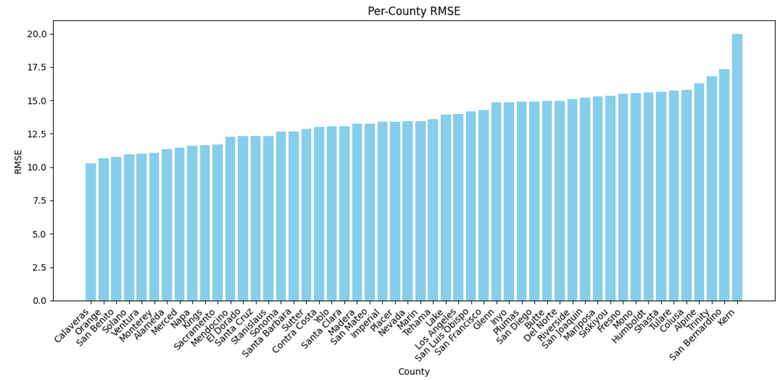


Figure 3. Models' RMSE Across All California Counties

Future Steps

If given another six months, we would first incorporate meteorological and satellite data to improve spatial and temporal feature representation. Exploring self-supervised learning techniques like masked autoencoders could help leverage unlabeled data for better generalization. Additionally, optimizing hyperparameters with a more extensive Genetic Algorithm search would likely enhance model performance further.

References

[1] W. Leong, R. Kelani, and Z. Ahmad, "Prediction of air pollution index (api) using support vector machine (svm)," *Journal of Environmental Chemical Engineering*, 2019.

[2] W. Gao, T. Xiao, L. Zou, H. Li, and S. Gu, "Analysis and prediction of atmospheric environmental quality based on the autoregressive integrated moving average model (arima model) in hunan province, china," *Sustainability*, 2024.

[3] G. Drewil and R. Al-Bahadili, "Air pollution prediction using lstm deep learning and metaheuristics algorithms," *Measurement: Sensors*, vol. 24, 2022.

[4] L. Chen, J. Xu, B. Wu, Y. Qian, Z. Du, Y. Li, and Y. Zhang, "Group-aware graph neural network for nationwide city air quality forecasting," 2021.

[5] D. Iskandaryan, J. Ramos, and S. Oliver, "Graph neural network for air quality prediction: A case study in madrid," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[6] D. Lorenzo, V. Champaney, C. Ghnatios, and E. Cueto, "Physics-informed and graph neural networks for enhanced inverse analysis," *Engineering Computations*, 2024.

[7] J. Zuo, W. Li, M. Baldo, and H. Hacid, "Opportunistic air quality monitoring and forecasting with expandable graph neural networks," *arXiv*, 2023.

ykmattur@stanford.edu

ayeung16@stanford.edu