Prediction of PM2.5 concentration based on ARIMA, LSTM, and TCN models in Kigali, Rwanda

Yamlak Bogale, Choukouriyah Arinloye & Joselyne Muragijemariya

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction/Background

Air pollution, a critical global issue, has significant effects on health, the environment, and climate [1]. Among pollutants, PM2.5 (particulate matter with a diameter of 2.5 micrometers or smaller) is particularly concerning because it can penetrate deep into the respiratory system, causing conditions like asthma, cardiovascular diseases, and lung cancer. In Rwanda, and especially in its urban centres such as Kigali, the rapid pace of urbanization heightens air pollution risks, making it essential to monitor PM2.5 levels to protect public health [2].

This study leverages data from the Rwanda Environmental Management Authority (REMA) to forecast PM2.5 levels in Kigali using three machine learning models which analyze daily average PM2.5 levels, aiming to support air quality interventions and improve public health outcomes in Rwanda:

- AutoRegressive Integrated Moving Average (ARIMA)
- Long Short-Term Memory (LSTM)
- Temporal Convolutional Networks (TCN).

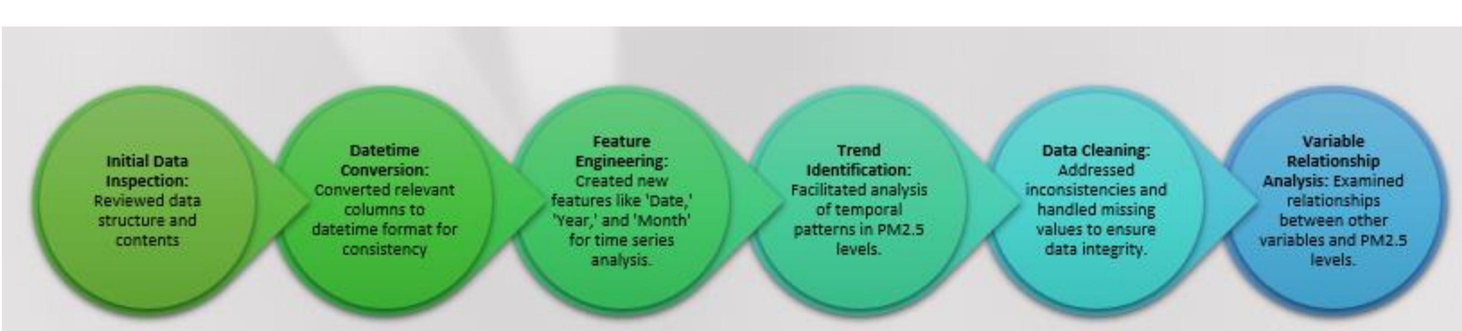
Data Collection & Pre-processing

Comprehensive data collection and pre-processing steps were applied to enhance model reliability.

The study uses air quality data from Kigali, Rwanda, collected from four monitoring stations: Gikondo-Mburabuturo, Jali, Mont Kigali, and Rusororo. The dataset covers the period from 2020 to 2024 and includes hourly PM2.5 concentration measurements obtained from the Rwanda Environment Management Authority (REMA). The key other variables used are presented in the table below:

Variables & Description	*Source*	*Health Impact*
SO_2, Sulfur Dioxide	Industry, power plants	Respiratory issues
CO, Carbon Monoxide	Vehicles, combustion	Reduces blood oxygen
PM10, Particles ≤10 micrometers	Construction, dust	Respiratory irritation
NO_2, Nitrogen Dioxide	Vehicles, industry	Aggravates respiratory conditions
O_3, Ground-level Ozone	NO_2 + VOCs in sunlight	Lung damage

Key Variables – Meaning, Source and Health impact



Data Pre-Processing chart flow

Methodology

Approaches for Time Series Forecasting and PM2.5 Prediction

Besides feature engineering, the study has tested three models and the specificities of each are outlined on the following image:

ARIMA Model

- Used for time series forecasting, with adjustments for stationarity.
- Applied differencing to achieve stationarity; then used AR and MA components to forecast PM2.5 levels.
- Configuration: p=2, d=1, and q=2, where p and q represent the orders of the AutoRegressive (AR) and Moving Average (MA)components, respectively, and d indicates the number of differences applied to achieve stationarity.
- Model fit on training data (80% of dataset), with forecast accuracy measured using Root Mean Squared Error (RMSE).

LSTM Model

- Implemented a bidirectional Long Short-Term Memory (LSTM) network for capturing sequential dependencies.
- Architecture: Three bidirectional LSTM layers with 150 units, followed by dense layers and dropout to prevent overfitting.
- Sequences of 30 past observations used to predict the next value; trained with Mean Absolute Error (MAE) loss and Adam optimizer.

TCN Model

- Temporal Convolutional Network (TCN) used to model long-range temporal dependencies with dilated causal convolutions.
- Configuration: 16 filters, kernel size of 3, with dilations set to capture dependencies.
- Sequences of 30 past observations created for forecasting; layer normalization and dropout rate of 0.02 applied.
- Optimized using SGD with learning rate of 0.001 and momentum of 0.8, evaluated with Mean Squared Error (MSE).

Methodology - Models' specificities

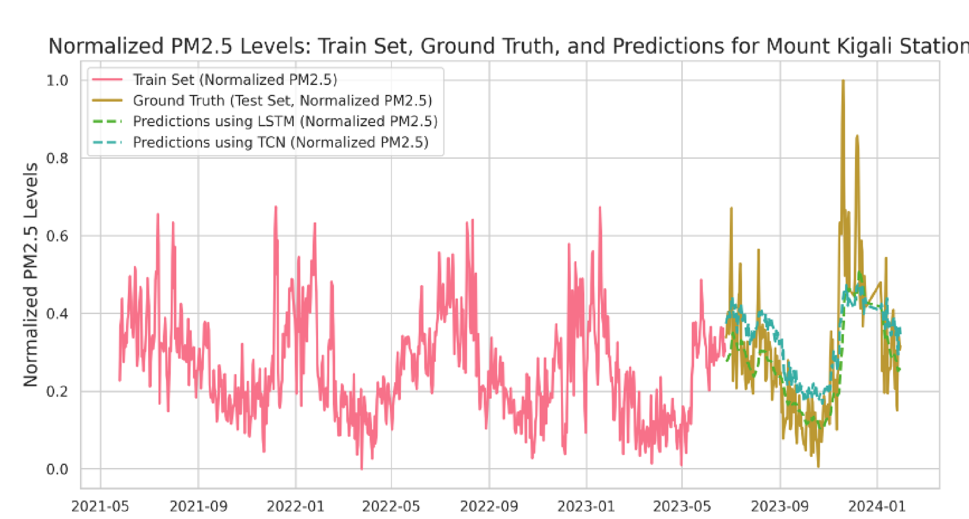
Results & Discussion

The performance of ARIMA, LSTM, and TCN models was evaluated across four locations, with metrics reported for both **RMSE** and **MAE**. As can be seen in the table below, **LSTM** and **TCN** models consistently outperform ARIMA across all locations, with LSTM generally achieving slightly better results. TCN also performs well but shows some variability, with performance close to that of LSTM in certain locations.

Region	Model	RMSE	MAE
Jali	ARIMA	0.558	0.492
	LSTM	0.273	0.212
	TCN	0.210	0.164
Mount Kigali	ARIMA	0.170	0.127
	LSTM	0.135	0.088
	TCN	0.155	0.103
Gikondo-Mburabuturo	ARIMA	0.188	0.123
	LSTM	0.161	0.098
	TCN	0.162	0.123
Rusororo	ARIMA	0.128	0.096
	LSTM	0.114	0.083
	TCN	0.133	0.111

Models Performance across regions in Kigali

The results highlight variability in model performance across sites, suggesting that effectiveness may be influenced by location-specific factors and dataset size. Notably, the **Jali** site had a shorter dataset period and a significant number of missing values, which may have affected the results.



PM2.5 Forecasting for Mount Kigali Station: Comparing the Two Best Performing Models

Implications & Conclusion

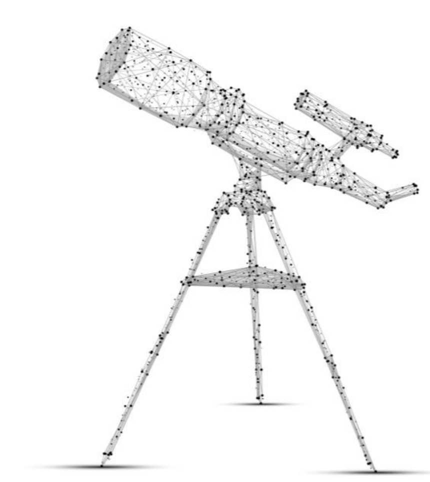
Advancing Air Quality management through Machine Learning insights

The success of machine learning models, particularly the LSTM and TCN, in forecasting PM2.5 levels highlights the potential of advanced algorithms in environmental monitoring. Given the accuracy of these models in capturing ***temporal dependencies***, similar methodologies could be applied to forecast other air pollutants, supporting a proactive approach in pollution management. By integrating ***ML***, these findings provide enhanced insights for climate policymakers and urban planners, enabling more accurate, data-driven strategies to improve air quality and mitigate environmental health risks in Kigali.

- This study demonstrates the effectiveness of Machine Learning approaches, with the LSTM model emerging as the most accurate for PM2.5 predictions.
- By extending this methodology, ***researchers and policymakers*** can develop more comprehensive air quality monitoring frameworks, contributing to healthier, sustainable urban environments in Kigali areas.

Future Work

Enhancing PM2.5 Prediction with Satellite Imagery and Deep Learning



Future work will address the limitations of our current research– namely to dive deeper into the anthropogenic factors, the spatio-temporal correlations and to reveal more multivariate dependencies. Moreover, future work could combine satellite imagery, like Sentinel 5P, with Deep Learning models [3,4,5] to enhance PM2.5 prediction accuracy by capturing environmental factors linked to air quality.

References

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: a review," *Frontiers in Public Health*, vol. 8, p. 14, 2020.

[2] "Urbanization in Rwanda: Building inclusive & sustainable cities," *UN Habitat*. [Online]. Available: https://unhabitat.org/rwanda.

[3] European Space Agency, "Introducing Sentinel-5P." [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-5P/Introducing_Sentinel-5P.

[4] arnavbansal1, "SatellitePollutionCNN/README.md at Master." [Online]. Available: https://github.com/arnavbansal1/SatellitePollutionCNN/blob/master/README.md.

[5] A. Rowley and O. Karakuş, "Predicting Air Quality via Multimodal AI and Satellite Imagery," *Remote Sensing of Environment*, vol. 293, p. 113609, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723001608. doi: 10.1016/j.rse.2023.113609.