Final Project: Gopal Bhusal

Comparing ML and DL Models for PM2.5 Concentration

1.Introduction

Air pollution, particularly fine particulate matter (PM2.5), is a major global health and environmental concern, with the Kathmandu Valley and other areas experiencing severe problems. Due to its unusual geographic location, fast urbanization, and dense population, the valley has consistently high PM2.5 levels, which exacerbate respiratory conditions and degrade the ecosystem (Shakya et al., 2017). Implementing successful mitigation solutions requires precise PM2.5 concentration modeling and forecast.

This study investigates the **research question**: *Do machine learning (ML) and deep learning (DL) models differ in performance on PM2.5 data in Kathmandu Valley under varying parameter configurations?*

The comparative analysis focuses on understanding how these methods, under different parameter settings, handle the complexities of PM2.5 prediction in this region.

Previous research has shown that ML approaches like Random Forest and Support Vector Machines are useful in air quality modeling because of their capacity to efficiently process structured datasets (Chen et al., 2020). Similarly, deep learning models, notably convolutional and recurrent neural networks, have showed potential in detecting nonlinear correlations and temporal patterns in PM2.5 data (Zhang et al., 2021). However, their comparative effectiveness in a localized and demanding setting such as the Kathmandu Valley has received little attention.

By assessing the performance of various models, this study hopes to add to the larger discussion about the usefulness of ML and DL for environmental monitoring. Furthermore, it seeks to advise policymakers and stakeholders by finding the best predictive frameworks for tackling air quality challenges in the Kathmandu Valley.

Location and Data

This study focuses on Kathmandu Valley, Nepal, a region well-known for its distinctive geography and persistent air quality issues. To better understand the area's environmental dynamics, a study area map was developed using ArcGIS Pro, highlighting key features and boundaries of the region.

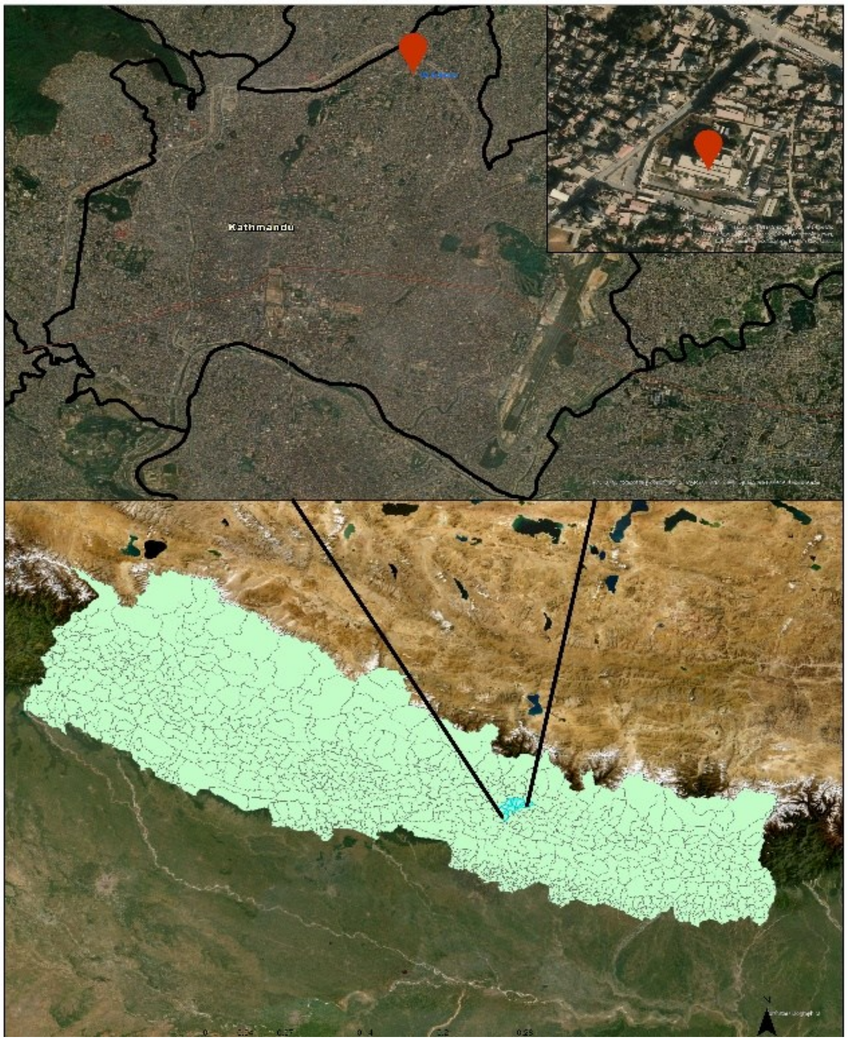


Fig1. Study Area Map

This map shows the study area, with the PM2.5 data collected at the U.S. Embassy in Kathmandu, Nepal. The site, located at about 1,400 meters above sea level, lies in the Kathmandu Valley, surrounded by hills. The area's geography and weather patterns significantly affect air quality, making it an important location for monitoring PM2.5 and understanding its environmental impacts.

2.Methodology

This chapter explains the data, and methods used to create the model in this research

Data Sources:

1. **PM2.5 Concentration Data**

   i. Source: Open Data Nepal

   ii. Type: Time-series data

   iii. Variable: Daily PM2.5 concentrations (μg/m^3)

   iv. Volume: 22,832 hourly records spanning January 2021 to December 2023

   Relevance: Key variable for analyzing air pollution levels.

2. **Climate Data**
   Climate variables were sourced from the ERA5 dataset, covering the same time period. Key variables include:

   i. Daily surface temperature (t2m) (i.e. temperature at 2m from the earth's surface) measured in Kelvin.

   ii. Horizontal wind speed (u100) and vertical wind speed (v100), both measured in m/s The data, provided as grid-based values, were interpolated to align with the study area.

3. **Shapefile Data**
   The administrative boundaries of Kathmandu Valley were acquired from Open Data Nepal in shapefile format, offering a geographic framework for spatial analysis.

Methods:

The study's approach involved four key stages, from data preprocessing to model evaluation, to investigate the relationship between PM2.5 concentrations and meteorological factors

A. Data Preprocessing

To prepare the datasets for machine learning, several preprocessing steps were carried out. First, missing PM2.5 values were imputed using time-series interpolation to maintain consistency in the dataset. Outliers in both PM2.5 and meteorological variables were removed to ensure data accuracy. The PM2.5 and meteorological data (temperature, wind components) were then synchronized on a daily time scale to match the temporal resolution of both datasets. A logarithmic transformation was applied to the temperature variable given the skewed distribution of temperature data. Finally, all variables were normalized to ensure uniformity and consistent input for machine learning models

B. Machine learning and deep learning workflow:

Using Linear Regression, Random Forest, Support Vector Machines (SVMs), and 1D Convolutional Neural Networks (1D-CNNs), the association between PM2.5 concentrations and meteorological variables was studied. While Random Forest caught non-linear interactions, Linear Regression served as a baseline for fundamental linkages. Because SVMs could handle high-dimensional data, they were used, while 1D-CNNs made use of the data's temporal patterns. With PM2.5 as the goal variable, the predictors were log-transformed temperature (t2m), u100, and v100. The data was divided into 25% testing and 75% training, and grid search was used to optimize the hyperparameter tuning.

C. Model Evaluation

Models were assessed using two key metrics:

R^2: Indicates how well the model explains the variance in PM2.5 concentrations.
Mean Squared Error (MSE): Measures prediction accuracy by calculating the average squared difference between predicted and observed values.

The flow chart diagram to show all of the above work flow description is shown below:

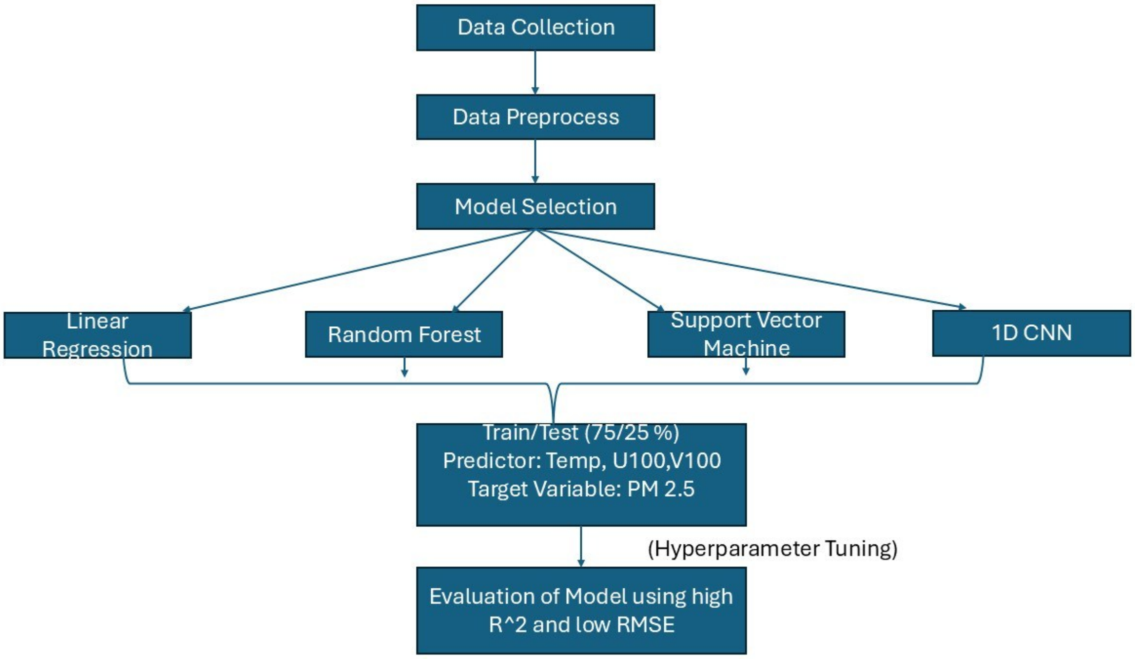


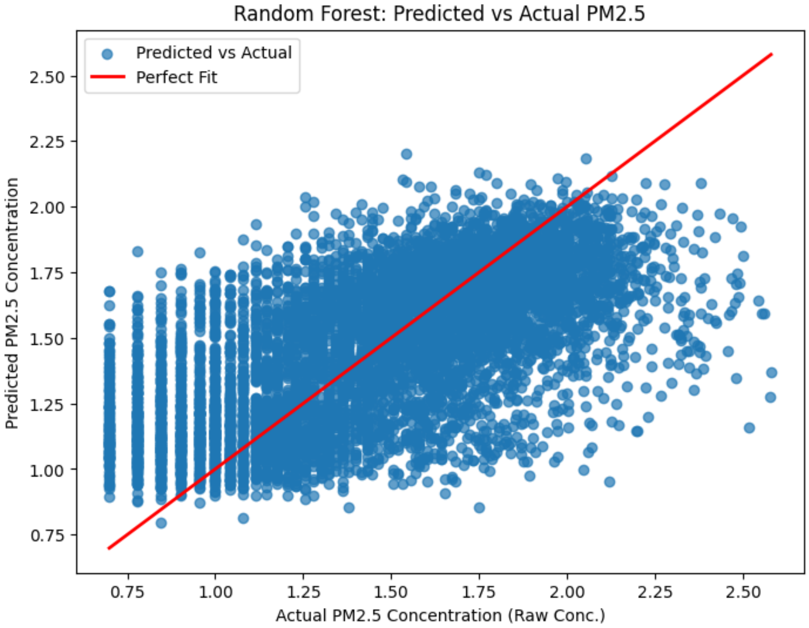
Fig 2: Overview of the Methodology

3.Results

The findings show how to use the techniques described in the Methods section to forecast PM2.5 concentrations based on meteorological factors. Four models—Random Forest, Linear Regression, Support Vector Machine (SVM), and Convolutional Neural Network (CNN)—are analyzed, with hyperparameter adjustment and performance evaluation included. MSE, R2, and RMSE metrics are used to assess each model's capacity to represent the link between the predictors and PM2.5.

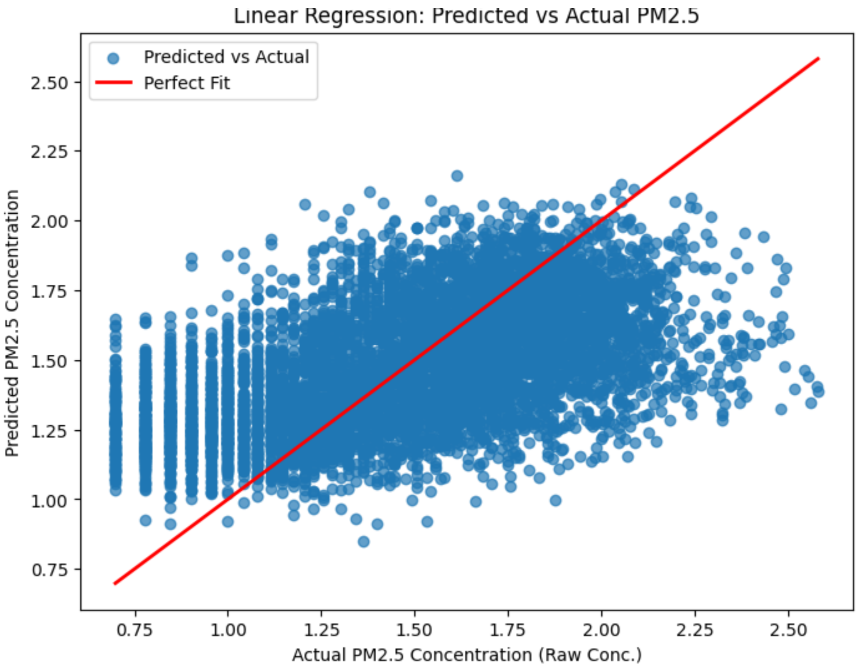
A. Random Forest Model

The Random Forest model demonstrated moderate predictive capability with an MSE of **0.092** and an R^2 of **0.395**. This model effectively captured non-linear interactions among the features. Its performance indicates that it can model relationships between PM2.5 and meteorological data but leaves room for improvement compared to other methods.



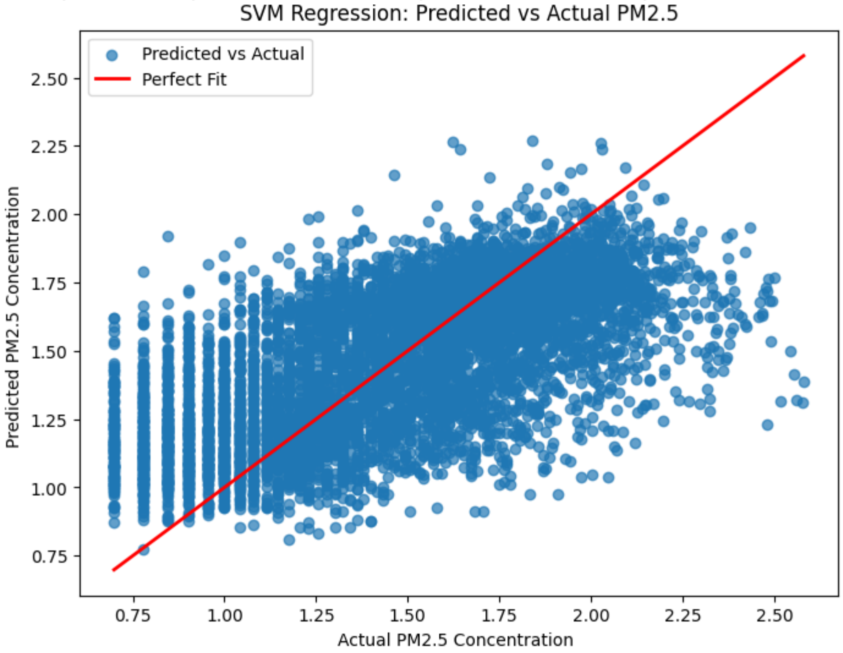
B. Linear Regression Model

As the simplest baseline model, Linear Regression achieved an MSE of **0.104** and an R^2 of **0.316**. While it provided an initial understanding of the variable relationships, its limitations in handling complex data interactions are apparent, as reflected in the relatively lower R^2 value.



C. Support Vector Machine (SVM)

The SVM model, utilizing a radial basis function kernel (Kernel = 'rbf', C = 1, epsilon = 0.1), outperformed the traditional models with an MSE of **0.086** and an R^2 of **0.432**. These results highlight its ability to model intricate patterns in the data effectively, making it a strong candidate among machine learning approaches.



D. Convolutional Neural Network (CNN)

The 1D CNN model was evaluated across three configurations of learning rate and batch size. This model is chosen because it is particularly effective in handling sequential and time-series data. PM2.5 levels and meteorological variables are often recorded at regular intervals, and a 1D CNN is capable of detecting temporal patterns and dependencies over time. The architecture included:

Input Layer: Daily PM2.5 and meteorological variables (temperature and wind components).
Two Convolutional Layers: 32 and 64 filters, kernel size (3,3), with ReLU activation.
Two Pooling Layers: MaxPooling with pool size (2,2).
Fully Connected Layer: Dense layer with 128 units and ReLU activation.
Output Layer: A dense layer with a single unit for regression output.

The results shown in the table focus on a batch size of 16, as this configuration provided the most consistent and interpretable results across varying learning rates. While other batch sizes (32, 64) were also tested, the batch size of 16 demonstrated comparable or better performance, making it the primary focus of this analysis

Batch Size	Learning Rate	Train R2	Test R2	Validation Loss
16	0.001	0.458	0.444	0.553
16	0.01	0.443	0.430	0.565
16	0.1	0.424	0.409	0.685

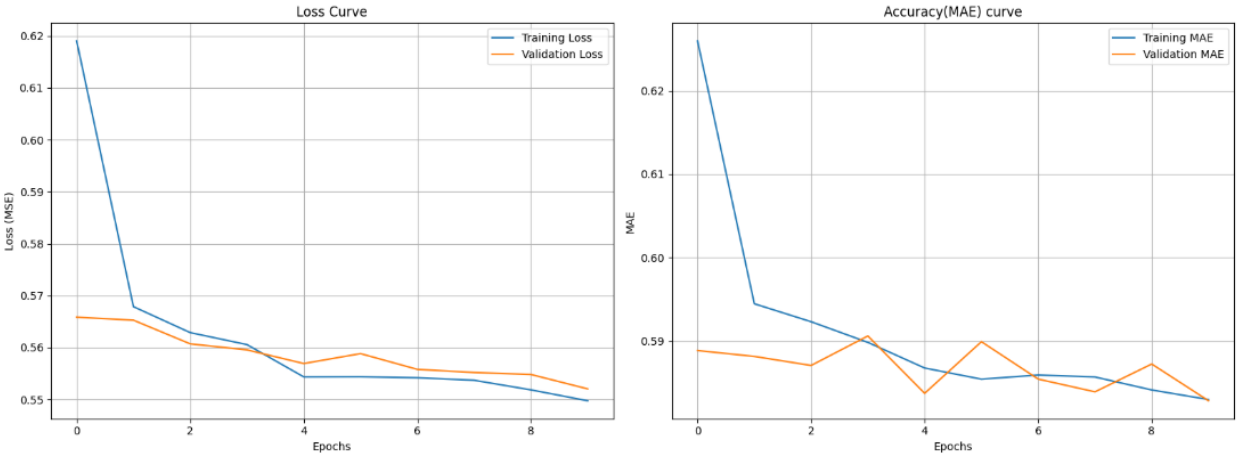
The CNN with a learning rate of 0.001 demonstrated the best performance, achieving a **Train R^2 of 0.458** and a **Test R^2 of 0.444**, with a validation loss of **0.5536**.

Analysis of Results:

The CNN model provided the most accurate predictions, benefiting from its ability to capture complex patterns in the data through its deep learning architecture. SVM also performed well, making it a strong alternative for smaller datasets. The results indicate that advanced methods, particularly CNN, are better suited for modeling PM2.5 concentrations in this context.

Visual Analysis

- **Loss Curve**: Training and validation loss curves show a steady decline over epochs, indicating effective training for CNN.
- **MAE Curve**: The Mean Absolute Error (MAE) plots for training and validation data demonstrate improved accuracy over time.



4. Discussion

The results of this study demonstrate that both machine learning and deep learning methods are capable of modeling PM2.5 concentrations in Kathmandu Valley, but their performance varies significantly. The Support Vector Machine (SVM) model outperformed other traditional ML methods, while the Convolutional Neural Network (CNN) showed the highest accuracy overall, particularly with a batch size of 16 and a learning rate of 0.001. These findings suggest that advanced methods like CNNs are better suited for capturing complex, non-linear relationships in air quality data.

Similar studies have highlighted the efficacy of machine learning for air quality modeling. For instance, Chen et al. (2020) demonstrated the superiority of SVM over Random Forest in predicting PM2.5 levels due to its ability to handle high-dimensional data. Moreover, Zhang et al. (2021) showed that CNNs excel in extracting patterns from temporal and spatial datasets, consistent with the results here.

Despite these successes, the CNN's dependence on hyperparameter optimization poses challenges, and its higher computational cost compared to simpler models like Linear Regression must be acknowledged. Furthermore, the relatively low R^2 values across all models indicate room for improvement, possibly by incorporating additional predictors such as humidity or traffic data.

5.Conclusion

The result above shows that both machine learning and deep learning models are effective at predicting PM2.5 concentrations in the Kathmandu valley. Among the models examined, the 1D CNN performed the best, capturing the intricate temporal patterns in the data. The study emphasizes the need of adopting advanced approaches such as CNN for air quality prediction. While the models showed potential, they can be improved by integrating more variables like humidity and traffic data, as well as expanding data gathering to numerous sites for greater generalizability.

Limitations

i. The CNN model required significant computational resources due to hyperparameter optimization.
ii. The models showed relatively low R^2 values, suggesting that important variables may be missing.
iii. The data was collected from only one location, which limits the generalizability of the results.
iv. Variables like humidity and traffic data were not included, which could improve the model's accuracy.

Future Work

Future research could explore integrating more meteorological variables, such as humidity and precipitation, to enhance model accuracy. Additionally, testing other deep learning architectures like LSTMs may improve temporal predictions. Expanding the study to include data from multiple observation sites within Kathmandu Valley would also allow for spatial generalization of the models.

This study addresses the research question by confirming the differences in performance between ML and DL models in predicting PM2.5 concentrations, with CNN proving to be the most effective. However, further refinements are needed for broader applicability.

References

Chen, Z., Huang, Y., & Zhang, Y. (2020). Machine learning approaches for PM2.5 prediction: A review. *Environmental Modelling & Software, 124*, 104600. https://doi.org/10.1016/j.envsoft.2020.104600

Shakya, K. M., Rupakheti, M., & Aryal, R. (2017). Air pollution trends in Kathmandu Valley: A review of monitoring results. *Journal of Environmental Science*.

Zhang, J., Liu, X., & Wang, T. (2021). Deep learning in air quality prediction: Advances and challenges. *Atmospheric Environment, 244*, 117956. https://doi.org/10.1016/j.atmosenv.2020.117956

ERA5 Reanalysis Data. (n.d.). Retrieved from https://cds.climate.copernicus.eu

Open Data Nepal. (n.d.). Air quality data in Kathmandu. Retrieved from https://opendatanepal.com

https://opendatanepal.com/dataset/new-political-and-administrative-boundaries-shapefile-of-nepal