Vol 6 No 2 (2024) E-ISSN: 2682-860X

International Journal on Robotics, Automation and Sciences

Forecasting PM2.5 Concentrations in Chiang Mai using Machine Learning Models

Manlika Ratchagit*

***Abstract* – Particulate matter 2.5 poses a significant threat to human life. Over the past decade, there has been a significant increase in the number of articles dedicated to studying and forecasting PM2.5 concentrations. Thailand, particularly Chiang Mai, has elevated levels of dangerous PM2.5 throughout the hot season. The primary objective of this study is to evaluate the efficacy of three widely used machine learning models, namely artificial neural network (ANN), long short-term memory network (LSTM), and convolutional neural network (CNN), in predicting the levels of PM2.5 particles in Chiang Mai. The raw data are obtained from the Pollution Control Department, Ministry of Natural Resources and Environment Thailand between January 2014 and June 2023, a total of 3,468 observations. We split the data into three sets namely, training, validation, and test sets. The criterion to evaluate three machine learning techniques is the median absolute error. The experimental results confirm that all three machine learning models provide similar movements of PM2.5 dust pollution. Moreover, the artificial neural network technique provides better results than the others regarding error measurement.**

Keywords— PM 2.5 Concentrations, Machine Learning, Artificial Neural Network, Long Short-Term Memory Network, Convolutional Neural Network.

I. INTRODUCTION

One of the most important issues facing society today is air pollution. Air pollution commonly encompasses several forms such as particulate matter, ground-level ozone, automobile pollutants, and other pollutants that have long-term impacts on both human health and the ecosystem. The investigation of particulate matter, particularly particulate matter 2.5 (PM2.5), was conducted ten years ago. The characteristics of PM2.5 particles are their small diameter, light weight, high reactivity, and capacity to float and stay suspended in the environment for long periods of time. The presence of suspended particles can lead to a reduction in air visibility, hence causing haze conditions. Additionally, these particles can influence the radiation balance and the Earth's biological cycle [1]. The human risk significantly affects individuals' physical and mental well-being. Two primary issues that PM2.5 can lead to are respiratory disorders (such as asthma, bronchitis, and chronic obstructive pulmonary disease, or COPD) and cardiovascular disorders (including heart attacks, strokes, arrhythmias, and heart disease). PM2.5 can have long-term negative impacts on health, including death, a greater risk of lung function, and reduced lung function and development [2]. The presence of PM2.5 in the environment has been found to have detrimental effects on various aspects, including the degradation of materials and structures, the deposition of acid, and the elevation of ozone levels [3]. Chiang Mai, Thailand, is ready for a PM2.5 disaster area because, during the first week of April 2024, the average concentration of fine particulate matter smaller than 2.5 microns exceeded 150 micrograms per cubic

*Corresponding author. Email: manlika@mju.ac.th ORCID: 0000-0001-8600-5387
Manika Ratchagit*, Assistant Professor, Program in Statistics and Information Management, Faculty of Science, Maejo University, Chiang Mai 50290, Thailand.

International Journal on Robotics, Automation and Sciences (2024) 6, 2:37-41
https://doi.org/10.33093/ijoras.2024.6.2.6
Manuscript received: 14 June 2024 | Revised: 2 Aug 2024 | Accepted: 28 Aug 2024 | Published: 30 Sep 2024

Published by MMU PRESS. URL: http://journals.mmupress.com/ijoras

meter (µg/m3), which is extremely dangerous [4]. Chiang Mai is the capital city of the northern part of Thailand. High PM2.5 levels in the northern region have caused numerous residents to seek medical assistance for respiratory ailments such as asthma and inflammation. According to a report by Maharaj Nakorn Chiang Mai Hospital on March 19, 2024, 30,339 people sought medical attention for respiratory issues from January 1 to March 15, 2024. This figure represents an increase compared to the corresponding period in the previous year, during which 12,671 individuals sought treatment [5].

This study utilized three widely used machine learning techniques to predict the occurrence of PM2.5 particles in Chiang Mai. The outcomes of this research will assist healthcare professionals and the general public with a better understanding of air quality, boosting awareness about the toxicity of PM2.5 in Chiang Mai, Thailand.

II. LITERATURE REVIEW

Wongrin et al. [6] examined statistics and deep learning neural networks for predicting daily average PM2.5 concentrations in northern Thailand. The study employs three statistical models, including Holt-Winters exponential smoothing (ETS), autoregressive integrated moving average (ARIMA), and dynamic linear model (DLM). Two widely studied deep learning neural networks are the recurrent neural network (RNN) and the long-short term memory (LSTM). The root mean square method (RMSE) is used to determine the suitable method. According to the authors, the ARIMA technique outperforms deep learning neural networks in most stations. Amnuaylojaroen [7] used multivariate linear regression models to analyze hourly PM2.5 values in northern Thailand. The provinces of Chiang Mai, Lampang, and Nan are utilized in this study. To compare the performance of two multivariate linear regression models, the root mean square error (RMSE) is utilized. In this work, the author suggested multivariate linear regression models for humid and rainy seasons incorporating meteorological parameters and several gaseous pollutants, including SO2, NO2, CO, and O3. Thongrod et al. [8] utilized three techniques for PM2.5 and PM10 forecasting in Chiang Mai between 2012 and 2021: support vector machine (SVM), artificial neural network (ANN), and multiple linear regression (MLR). The numerical findings show that the support vector machine outperforms the others by providing the lowest root mean square error (RMSE), mean absolute error (MAE), and mean absolute percent errors (MAPE). Srikamdee and Onpans [9] employed linear regression, neural networks, and genetic programming to forecast the daily air quality index in northern Thailand. They stated that two linear equations developed from linear regression and genetic programming are suitable for prediction since they produce an average accuracy of more than 70%. The summary of papers that have been reviewed for this section is presented in Table 1.

TABLE 1. The literature summary.

Reference	Data	Method	Error Measures
Wongrin et.al. [6]	Daily average PM2.5	ETS, ARIMA, DLM, RNN, LSTM	RMSE
Amnuaylojaroen [7]	hourly PM2.5	multivariate linear regression models	RMSE
Thongrod et. Al [8]	hourly PM2.5	SVM, ANN, MLR	RMSE, MAE, MAPE
Srikamdee and Onpans [9]	Daily AQIs	LR, NN, GP	RMSE

TABLE 2. The classification of PM2.5 values into five distinct warning classifications [7].

AQI	Meaning	Color
0 – 50	Good	Green
51-100	Moderate	Yellow
101-150	Unhealthy for sensitive group	Orange
151-200	Unhealthy	Red
201-300	Very unhealthy	Purple
301-500	Hazardous	Maroon

III. RESEARCH METHODOLOGY

In this article, three well-known ML strategies are discussed, which are supervised techniques that allow us to collect data or generate output based on prior knowledge.

A. Artificial Neural Network (ANN) [10]

ANN represents the initial model for nonparametric nonlinear time series. The network is commonly called a feed-forward neural network due to its exclusive processing of inputs in a forward direction. An artificial neural network (ANN) comprises a network of computer units known as neurons. In artificial neural networks, the neurons are depicted as nodes. Weights are used to connect the nodes. ANN is now widely used in research due to its flexibility and ability to represent nonlinear phenomena. ANN is beneficial when the data exhibit non-stationarity and possess an uncertain statistical distribution. The architecture of ANN is presented in Figure 1.

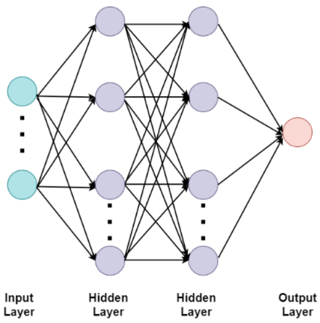


FIGURE 1. Architecture of ANN [11].

B. Long Short-Term Memory Network (LSTM)

The main aims of Long Short-Term Memory (LSTM) models are to effectively capture extended dependencies and determine the most suitable temporal delay (lag) in time series scenarios. There are two distinct subcategory states in the LSTM model: a short-term state, which has a resemblance to the RNN, and a long-term state. The data is gathered to capture the long-term dependencies between the current and previous hidden states across time. The long-term state moves from left to right through a forget gate. An additional set of memories is appended through the addition operation, while others are deleted. The structure of the LSTM is depicted in Figure 2.

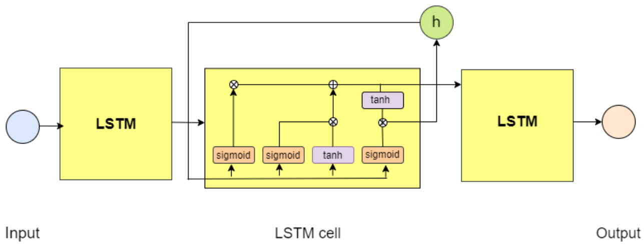


FIGURE 2. Architecture of LSTM [11].

C. Convolutional Neural Network (CNN)

Convolutional neural networks (CNN) belong to the category of deep learning techniques. There exist three distinct dimensions of convolutional neural networks (CNNs). These dimensions are commonly employed to analyze time series, image, and 3D image data, respectively. CNN has an advantage over its predecessors by automatically identifying significant features without human supervision [12]. The architecture of CNN is revealed in Figure 3.

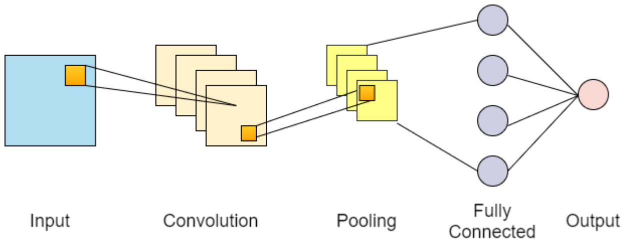


FIGURE 3. Fully CNN architecture [11].

The raw data used in this work are daily average PM2.5 pollution in Chiang Mai from the Pollution Control Department, Ministry of Natural Resources and Environment, Thailand [13]. A total of 3468 observations were collected between January 2014 and June 2023. We split data into three parts: training, validation, and testing. The training set (70%) contains 2428 observations from January 2014 to August 2020, which are used to fit (train) a prediction model. Next, 520 observations from September 2021 to January 2022 (15%) are used to validate network performance. The remaining 520 data points are used as a test set (15%). Before applying the data to the ML model, it is essential to scale it to the range [0, 1] using min-max normalization. Normalization of data generally accelerates learning and convergence. All experimental procedures are conducted within the Google Colaboratory (Google Colab) platform. To illustrate the efficacy of the predictive model, the error measurement is implemented. In this work, the median absolute error (MdAE) is presented because this error measurement falls in scale-dependent measures which is suitable when comparing numerous models for the same data set [14]. The median absolute error is shown in equation (1).

$$\text{MdAE} = \text{Median}(\left|Y_t - F_t\right|), \qquad (1)$$

here Y_t is an actual PM2.5 pollution at time t

F_t is predicted PM2.5 pollution at time t

IV. RESULTS AND DISCUSSIONS

To analyze the pattern of our dataset, it is necessary to generate a plot of the time series. Figure 4 illustrates the PM 2.5 particle pattern.

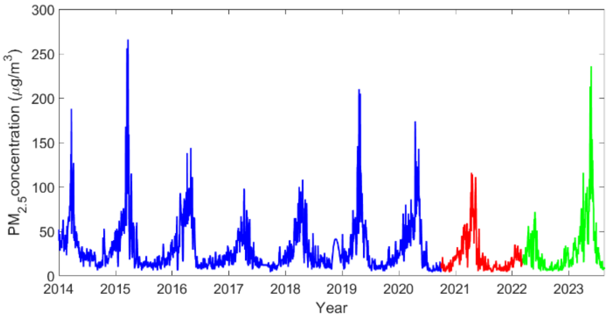


FIGURE 4. Time series plot of PM2.5 concentrations from 2014 to 2023.

From Figure 4, the training, validation, and test sets are represented by the blue, red, and green lines, respectively. The data presented in Figure 4 clearly demonstrates that the PM2.5 pollution levels exhibit seasonal variations during the hot season. Table 3 contains descriptive statistics for the observed daily average PM2.5 values.

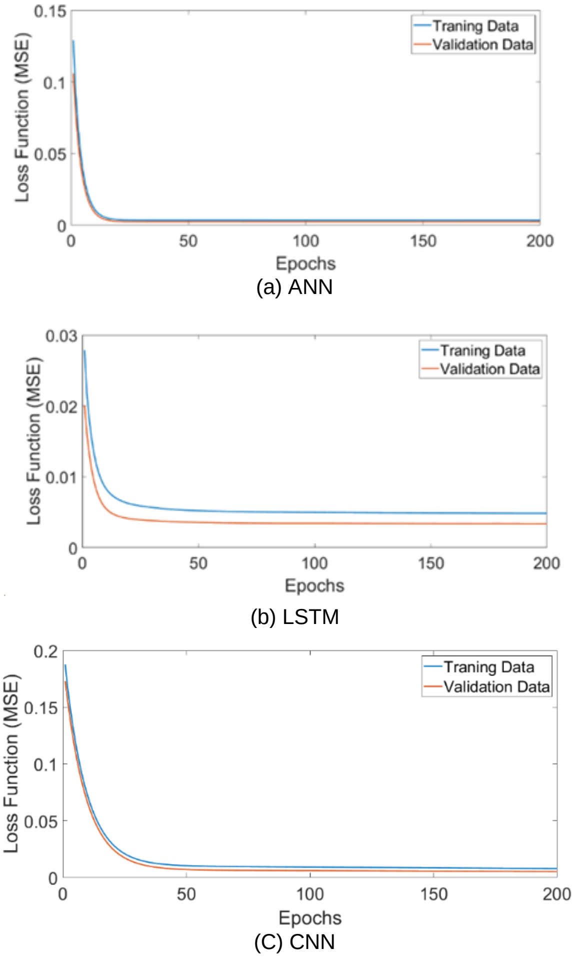
TABLE 3. Descriptive statistics of the original daily average PM 2.5 dust pollution.

N	3,468
Missing	151
Min	4
Max	266
Median	20
Mean	29.3463
Std	27.3492

According to Table 3, the data used in this study is 3,468, with 151 missing points. Spline interpolation is used to fill in any missing values. The greatest daily average PM 2.5 dust pollution was 266 micrograms per cubic meter on March 16, 2015, during Thailand's summer, and it reduced to 4 micrograms per cubic meter on September 7, 2019. The average and standard deviation for this set are 29.3463 and 27.3492 μg/m³, respectively. In all trials, we set the epoch to 200 and the batch size to 32. The loss function is the Mean Squared Error (MSE), whereas the activation function is the ReLU. Next, the optimal hyperparameters for all three Machine Learning techniques are determined using the Sherpa algorithm [15].

TABLE 4. The optimal hyperparameters of ANN, LSTM, and CNN models.

Model	Parameters	Values
ANN	Hidden unit1	94
	Hidden unit2	213
	Learning rate	0.039859
LSTM	LSTM(layer1)	121
	Learning rate	0.002608
CNN	Conv1D	10
	Kernel size	2
	Hidden unit1	74
	Learning rate	0.042326



(a) ANN

(b) LSTM

(C) CNN

FIGURE 5. The loss function between the training and validation sets during training (a) ANN, (b) LSTM, and (c) CNN.

We then examine the loss function's convergence plot. Figure 5 compares the loss function between the training and validation sets for all three machine learning models.

As illustrated in Figure 5, the training and validation losses settled and decreased near each other. This indicates that our model architecture can be applied to predict future PM2.5 concentrations. We then compute and compare the error measurements of the ANN, LSTM, and CNN models across the training, validation, and test sets. The MdAE comparisons are detailed in Table 5. The MdAE obtained from the ANN, LSTM, and CNN methods is presented in Table 5.

TABLE 5. MdAE of three ML models.

Model	Training set	Validation set	Test set
ANN	3.2300	2.3602	3.2957
LSTM	7.0118	7.3218	8.4065
CNN	3.5083	2.5608	3.5523

Table 5 shows that the ANN technique produces considerably lower error measurements than other models. The comparison between the predicted PM2.5 pollution obtained from each ML method and its actual observations on the test set is illustrated in Figure 6.

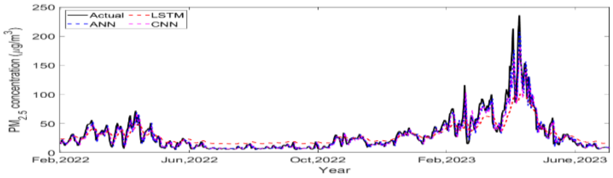


FIGURE 6. The comparison of actual data together with each ML procedure on the test set

Figure 6 presents the observed data pattern and compares the forecast PM2.5 for all three ML techniques between February 2022 and June 2023. The black solid lines represent the actual PM2.5 levels. The PM2.5 predictions obtained from the ANN, LSTM, and CNN are represented by the blue, red, and pink dashed lines, respectively. It is clear that the predicted PM2.5 from all ML methods follows a similar trend to the observed PM2.5. However, the ANN outperforms all other models in terms of error measurement.

V. CONCLUSION

PM2.5 concentrations are a serious concern globally, particularly in Chiang Mai, Thailand, which will be the world's most polluted city on April 9, 2024 [16]. Residents are suffering from respiratory issues. Numerous researchers have sought to investigate the impact of PM2.5 concentrations. Forecasting is one of the most frequent subjects for PM2.5 concentrations. This study uses well-known ML methods such as the ANN, LSTM, and CNN models to predict daily average PM2.5 concentrations in Chiang Mai, Thailand. The primary sources of information are 3,468 observations obtained from the Pollution Control Department of the Ministry of Natural Resources and Environment between January 2014 and June 2023. We divided the data into training, validation, and test sets. The criterion for evaluating the efficacy of our three ML methods is the median absolute error. The results indicate that the ANN technique is superior to the alternatives in terms of error measurement. This can serve as a warning to both locals and the government to prioritize air quality improvement. For future work, more individual ML models and hybrid techniques should be considered to improve the performance of the forecasting method.

ACKNOWLEDGMENT

We thank the anonymous reviewers for the careful review of our manuscript.

FUNDING STATEMENT

There is no financial support from any agencies for this research.

AUTHOR CONTRIBUTIONS

Manlika Ratchagit: Writing – Original Draft Preparation;

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

REFERENCES

[1] N. Jia, Y. Li, R. Chen and H. Yang, “A review of global pm2.5 exposure research trends from 1992 to 2022,” *Sustainability*, vol. 15, pp.105099, 2023. DOI: https://doi.org/10.3390/su151310509

[2] P. Vathesatogkit, "The health risks of pm 2.5," *Bumrungrad hospital health-blog,* URL: https://www.bumrungrad.com/en/health-blog/january-2024/the-health-risks-of-pm-2-5 (Accessed 1 Aug, 2024)

[3] QAir, “PM2.5”, *IQAIR*, URL: https://www.iqair.com/world-most-polluted-cities (Accessed 1 Aug, 2024)

[4] T. Panumate, “Chiang Mai prepares for pm2.5 disaster areas,” *bangkokpost news,* URL: https://www.bangkokpost.com/thailand/general/2541546/chiang-mai-prepares-for-pm2-5-disaster-areas (Accessed 1 Aug, 2024)

[5] Bangkok Post, “Lung cancer, PM2.5 deaths surge in the North, ”*bangkokpost news,* URL:https://www.bangkokpost.com/thailand/general/2542879/lung-cancer-pm2-5-deaths-surge-in-the-north (Accessed 1 Aug, 2024)

[6] W. Wongrin, K. Chaisee and K. Suphawan, “Comparison of statistical and deep learning methods for forecasting pm 2.5 concentrationss in northern Thailand,” *Polish Journal of Environmental Studies,* vol. 32, pp.1419-1431, 2023. DOI: https://doi.org/10.15244/pjoes/157072

[7] T. Amnuaylojaroen, “Prediction of PM2. 5 in an urban area of northern Thailand using multivariate linear regression model,” *Advances in Meteorology,* vol. 2022, pp.1-9, 2022. DOI: https://doi.org/10.1155/2022/3190484

[8] T. Thongrod, A. Lim, T. Ingviya and B.A.Owusu, “Prediction of pm2. 5 and pm10 in Chiang Mai province: a comparison of machine learning models, ” *2022 37th International Technical Conference on Circuits/Systems, Computers and Communications*, pp. 1-4, 2022. DOI: https://doi.org/10.1109/ITC-CSCC55581.2022.9894884

[9] S. Srikamdee and J. Onpans, “Forecasting daily air quality in northern Thailand using machine learning techniques, ” *2019 4th International Conference on Information Technology*, pp. 259-263, 2019. DOI: https://doi.org/10.1109/INCIT.2019.8912072

[10] M. Ratchagit, “Statistical analysis of delay in time series,” Ph.D. dissertation, EECMS., Curtin University, Perth, Australia, 2023. URL: https://espace.curtin.edu.au/handle/20.500.11937/91414 (Accessed 1 Aug. 2024).

[11] M. Ratchagit and H. Xu “A two-delay combination model for stock price prediction,” *Mathematics*, vol. 10, pp. 3447, 2022. DOI: https://doi.org/10.3390/math10193447

[12] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, 2021. DOI: https://doi.org/10.1186/s40537-021-00444-8

[13] Pollution Control Department, “Daily Average PM2.5”, *Air4Thai*. URL: http://air4thai.pcd.go.th/webV3/#/Home (Accessed 1 Aug, 2024)

[14] M.Theodosiou, “Forecasting monthly and quarterly time series using STL decomposition,” *International Journal of Forecasting*, vol.27, no.4, pp.1178-1195, 2011. DOI: https://doi.org/10.1016/j.ijforecast.2010.11.002

[15] L. Hertel, J. Collado, P. Sadowski, J. Ott and P. Baldi, “Sherpa: Robust hyperparameter optimization for machine learning,” *SoftwareX*, vol.12, pp. 100591, 2020. DOI: https://doi.org/10.1016/j.softx.2020.100591

[16] Bangkok Post, “Chiang Mai again world's most polluted city,” *bangkokpost news.* URL: https://www.bangkokpost.com/thailand/general/2543360/chiang-mai-again-worlds-most-polluted-city (Accessed 1 Aug, 2024)