Engineered Science
DOI: https://dx.doi.org/10.30919/es1375

Deep-Learning-Based Prediction System for Ultrafine Particulate Matter ($PM_{0.1}$) Concentration Using Meteorological Factors

Apaporn Tipsavak,[1] Thanathip Limna,[2, 8] Racha Dejchanchaiwong,[3, 8] Perapong Tekasakul,[4, 8] Kirttayoth Yeranee,[5] Bukhoree Sahoh[6, 9] and Mallika Kliangkhlao[7, 9, 10,*]

Abstract

Ultrafine particulate matter ($PM_{0.1}$) is a global and significant environmental issue because it can deeply translocate the human body, causing adverse health effects and leading to a high mortality rate. This study investigates the relationship between meteorological factors and $PM_{0.1}$ concentrations, providing insights into the formation and distribution of ultrafine particles. However, accurate measurement of the $PM_{0.1}$ concentration information is challenging due to sophisticated processes and expensive instruments that make it difficult to access. This study addresses these concerns with a new deep-learning regression model for $PM_{0.1}$ concentration prediction based on meteorological factors. The model is designed and developed to explore the optimal model structures (hidden layers and neurons) to achieve standard laboratory-based $PM_{0.1}$ measurement. The model structures are verified by root mean squared error (RMSE) and coefficient of determination (R^2) based on predictive performance to prove the laboratory-based standard's accomplishment. The results demonstrate that the proposed model can estimate $PM_{0.1}$ concentration with high performance, R^2 = 92.52%, and RMSE = 0.26 μg/m^3, which is precise and reliable for an intelligent-driven $PM_{0.1}$ concentration prediction system to support preventive health decision-making. This approach contributes to a more comprehensive understanding of atmospheric composition by enabling widespread monitoring of $PM_{0.1}$, a critical but often unmeasured component of air pollution.

Keywords: Artificial intelligence; Atmospheric nanoparticles; Machine learning; Dust pollutant; Air pollution.
Received: 03 November 2024; Revised: 23 December 2024; Accepted: 09 January 2025.
Article type: Research article.

1. Introduction

Particulate matter (PM) is one of the most harmful air pollutants to human health, such as $PM_{2.5}$ (its aerodynamic diameter is smaller than 2.5 μm).[1] In 2019, the World Health Organization (WHO) notified that PM caused 4.2 million premature deaths worldwide, of which 89% occurred in low- and middle-income countries (*e.g.*, Southeast Asia).[2] Ultrafine particulate matter ($PM_{0.1}$), whose aerodynamic diameter is smaller than 100 nm, is an important portion of $PM_{2.5}$, causing most unhealthy and toxic symptoms because it can efficiently translocate through the blood systems to critical body organs.[3]

Mao *et al.*[4] investigated the correlation between $PM_{0.1}$ and cardiovascular disease, indicating that it could damage the cardiovascular system. In this way, the urgent need to monitor and detect the concentration of $PM_{0.1}$ becomes apparent, as it may help people avoid risky environments and prevent severe health conditions. Phairuang *et al.*[5] measured $PM_{0.1}$ using an ambient nano-sampler and highlighted that its characteristics varied by different meteorological factors, producing different concentrations. However, accurate measurement of $PM_{0.1}$ needs to include the device's intensive maintenance and repair and the requirement of advanced technologies with the expensive equipment cost.[6] An intelligent-driven system for $PM_{0.1}$ concentration prediction based on available meteorological factors (*e.g.*, temperature, humidity, wind

[1] *College of Graduate Studies, Walailak University, Tha Sala, Nakhon Si Thammarat, 80160, Thailand*
[2] *Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, 90110, Thailand*
[3] *Department of Chemical Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, 90110, Thailand*
[4] *Department of Mechanical and Mechatronics Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, 90110, Thailand*
[5] *School of Mechanical Engineering, Shanghai Jiao Tong University, Minhang, Shanghai, 200240, China*

speed and direction, and precipitation) is critical in dealing with this challenge.

Machine learning (ML) regression plays a significant role in PM concentration prediction, estimating continuous PM values based on multiple meteorological factors. Fernández *et al*.[7] proposed IoT-based ML for $PM_{2.5}$ and PM_{10} concentration prediction and discussed that their approach could achieve high performance (R^2 for $PM_{2.5}$ and PM_{10} = 0.74 and 0.91, respectively). Hong proposed ML regression-based models for predicting the PM_{10} concertation and found that random forest (RF) could forecast accurately (R^2 = 0.98).[8] Kalantari *et al*.[9] proposed ML regression-based models for predicting the PM_{10} concertation's missing values and found that eXtreme Gradient Boosting (XGBoost) could solve the problem effectively (R^2 = 0.78). Zhou *et al*.[10] reviewed the role of ML regression in the concertation prediction of $PM_{2.5}$. They discussed that deep learning (DL) regression has massive potential for PM_{10} concertation prediction in future applications. Bhatti *et al*.[11] proposed a novel DL regression based on a graph-based convolutional network. Their proposed model could predict PM concentrations with superior performance (R^2 for $PM_{2.5}$ and PM_{10} = 0.75 and 0.76, respectively).

Although recent studies have intensively investigated PM concentration prediction based on ML regression, they focused on $PM_{2.5}$ and PM_{10} and have not yet proposed a model for $PM_{0.1}$ concentration prediction. In contrast, Schraufnagel studied the impact of $PM_{0.1}$ and discussed that the more ultra-small the size of PM, the more dangerous the effect on human health.[12] This makes $PM_{0.1}$ concentration prediction demanding and critical to support environmental policymakers in understanding air pollution problems and producing policy interventions for better air pollution control. Moreover, the capability of environmental management depends on the accurate prediction of $PM_{0.1}$ concentration so that decision-makers can precisely identify specific geolocations, design strategies, and provide the right action on time. In this way, the DL regression model must be designed and developed to fulfill the challenge, and the main contributions of this research are:

- To introduce the $PM_{0.1}$ concentration prediction framework based on ready-to-use meteorological factors that can address the problems of expensive equipment and installation costs.
- To propose a DL regression model for $PM_{0.1}$ concentration prediction based on meteorological factors, including $PM_{2.5}$, temperature, humidity, wind speed and direction, and precipitation.
- To prove that the DL regression model can predict $PM_{0.1}$ concentration outstandingly given available meteorological factors compared to basic algorithms such as artificial neural network (ANN) regression.

$PM_{0.1}$ and DL regression for PM concentration prediction are investigated in Section 2. The overview system of $PM_{0.1}$ concentration prediction is illustrated in Section 3. Section 4 concentrates on PM0.1 and meteorological factor acquisition, while Section 5 focuses on the DL regression-based model for $PM_{0.1}$ concentration prediction. Section 6 evaluates the effectiveness of the DL regression-based model using quantitative experiments. Conclusions and future directions appear in Section 7.

2. Related works and background knowledge

This section considers potential technologies and their applications in personal thermal comfort. They are relevant to physiological IoT and ML techniques to deal with complex meteorological factors.

2.1 Ultrafine particulate matter ($PM_{0.1}$)

$PM_{0.1}$ refers to a nanoparticle of about 0.1 micrometers (micron) in diameter or smaller, which can quickly and deeply translocate all over the human body systems. $PM_{0.1}$ causes tissue damage and is more harmful to human health because the toxicity of PM depends on size. Wang *et al*.[13] examined $PM_{0.1}$ concentrations concerning traffic-related air pollution. They employed the NanoScan based on ultrafine size measurement to monitor and gather the $PM_{0.1}$ and revealed that the near-road site increased about 30% of total $PM_{0.1}$ concentrations. Pham *et al*.[14] investigated the characteristics of $PM_{0.1}$ concentrations in burning agricultural crop residue. They used the nanoparticle sampler to monitor $PM_{0.1}$ concentrations and highlighted that concentrations significantly increase during the open burning of rice straw. Moreover, Mahasakpan *et al*.[15] explored the effect of $PM_{0.1}$ concentrations on health based on the nanoparticle sampler and found that it triggered a potential carcinogenic risk approximately 1.5 times. Bergmann *et al*.[16] investigated a correlation between $PM_{0.1}$ concentrations and mortalities, triggering respiratory and cardiovascular disease. In this manner, accurate measurement of $PM_{0.1}$ based on real-time sensing systems may help people be aware of its concentration and enable the environmental government authorities to recognize and manage pollution efficiently.

However, the $PM_{0.1}$ measurement process is sophisticated; for example, Kurotsuchi *et al*.[17] introduced advanced technology for a nanoparticle sampler that can measure $PM_{0.1}$ concentrations accurately. It requires skillfully laboratory-based experiments for professional experts to control quality

[6] *School of Informatics, Walailak University, Tha Sala, Nakhon Si Thammarat, 80160, Thailand*
[7] *School of Engineering and Technology, Walailak University, Nakhon Si Thammarat, 80160, Thailand*
[8] *Air Pollution and Health Effect Research Center, Prince of Songkla University, Hat Yai, Songkhla, 90110, Thailand*
[9] *Informatics Innovation Center of Excellence, Walailak University, Tha Sala, Nakhon Si Thammarat, 80160, Thailand*
[10] *Research Center for Intelligent Technology and Integration, Walailak University, Nakhon Si Thammarat, 80160, Thailand*
**Email:* mallika.kl@mail.wu.ac.th (M. Kliangkhlao)

extensively with expensive instrumentation, and it is almost impossible for ordinary people and government decision-makers to access information on $PM_{0.1}$ concentration, which is especially true in underdeveloped and developing countries where the technological infrastructures are unprepared. Therefore, an intelligent-driven system for $PM_{0.1}$ prediction plays an essential role in simplifying the measurement process of $PM_{0.1}$ concentrations. It produces results that support decision-makers and helps them be aware of the health risks of PM's impact at the right time.

2.2 DL regression for $PM_{0.1}$ concentration prediction

An intelligent-driven system imitates professional experts, encoding complex processes based on automation technologies such as ML. It may employ the available meteorological factors from a dynamic environment and compute and produce the results based on human-like specialists. This may help address the limitations of $PM_{0.1}$ concentration measurement because of the expensive instrumentation required by professional experts.

Tripathi *et al.*[18] proposed a novel approach based on ML regression to measure PM concentration in the mining industry. They summarized that ML regression-based random forest could achieve the best predictive performance (*RMSE* of PM concentration = 1.49 μg/m^3). Fernández *et al.*[19] proposed an ML regression model for PM concentration prediction and discussed that their approach could achieve high performance (R^2 of PM = 0.83). Their studies focused on applying the Internet of Things (IoT)-based PM sensors, the tools for real-time data acquisition. These studies aim to predict $PM_{1.0}$, $PM_{2.5}$, and PM_{10} based on off-the-shelf sensor technologies but did not investigate the techniques for $PM_{0.1}$ concentration measurement. This is because the materials for measuring $PM_{0.1}$ concentration are scarce and are still being researched. This causes $PM_{0.1}$ to be unconsidered and is a vital gap that needs to be filled. Mishra and Gupta1 expressed that DL regression could significantly improve the superior predictive performance of PM concentration prediction.[20] DL regression is a subset of ML, one of the most potent elements of the intelligent-driven system. Therefore, it may benefit $PM_{0.1}$ concentration prediction to deal with its unknown patterns in a dynamic environment.

DL regression extends the traditional concept of ANN regression to model the complex environmental patterns of PM pollution and produce continuous outcomes such as the $PM_{0.1}$ concentration. It may consist of multiple hidden layers depending on the complexity of the task (nonlinear systems) that wishes to be modeled.[21] DL regression encodes the patterns based on three main aspects, including (1) structure design, (2) neuron function, and (3) automatic learning improvement.

Structure design refers to the pre-defined process to determine the model's depth based on several hidden layers ($layer_l$) and neurons ($x_i^{layer_l}$). Each hidden layer connects through the neuron that can be computed as follows:

$$h_j^{layer_l} = \sum (w_{ij}^{layer_l} \times x_i^{layer_{l-1}}) + b_i^{layer_l},\ w_{ij}^{layer_l} \in \mathbb{R}^{layer_l \times layer_{l-1}} \quad (1)$$

where $h_j^{layer_l}$ is the raw output computed by a particular neuron based on weight ($w_{ij}^{layer_l}$), and bias ($b_i^{layer_l}$), and prior neuron value ($x_i^{layer_{l-1}}$), the weight encodes a coefficiency between the previous and present neurons between two hidden layers. The bias encodes the positive or negative direction of the coefficiency between the two, which lets DL regression recognize the complex patterns based on nonlinear relationships.

In this way, the value of $h_j^{layer_l}$ is a critical ingredient and must be verified whether its values should be attached to the next neuron or not, which is the role of the neuron function. The neuron function decides whether the two neurons between hidden layers should be connected (transferring if useful) or ignored (squashing if useless). It is called the activation function (α), assigning the neuron's role in taking action in a particular prediction process,[22] which can be computed as follows:

$$x_j^{layer_l} = \alpha^{layer_l}(h_j^{layer_l}) \quad (2)$$

The $x_j^{layer_l}$ represents a neuron role based on α^{layer_l} threshold that is authorized to be active or inactive. It lets DL regression handle the complex pattern based on dynamic connections between two hidden layers. As seen, the accurate action of neuron function depends on the summation of weights and biases, and automatic learning improvement is essential in adjusting and discovering these suitable values. Automatic learning improvement for DL regression investigates the loss between actual outcomes ($y_k^{layer_{output}}$) made by the laboratory-based measurement and predicted outcomes ($\hat{y}_k^{layer_{output}}$) made by algorithms given input features such as available ambient factors.[23] It employs mean squared error (MSE) to measure the loss (L_{MSE}) and can be computed as follows:

$$l(w_{ij}^{layer_l}, b_i^{layer_l}) = \text{ArgMin}\ \frac{1}{n} \sum_{k=1}^{n} L_{MSE}(\hat{y}_k^{layer_{output}}, y_k^{layer_{output}}) \quad (3)$$

where l is a tradeoff function, the loss in the model is compared, and weights and biases are adjusted to fit with all training data n. This process is automatically recursive until the loss is minimized (ArgMin function), meaning that actual and predicted outcomes are almost identical. In other words, the model improves itself until it achieves the level of the laboratory-based measurement.

The three components, structure design, neuron function, and automatic learning improvement, help DL regression address the challenge of complex patterns in nonlinear systems. It is highly potential for applying DL regression to $PM_{0.1}$ concentration predictions.

3. Overview system of $PM_{0.1}$ concentration prediction

The proposed real-time system employs meteorological

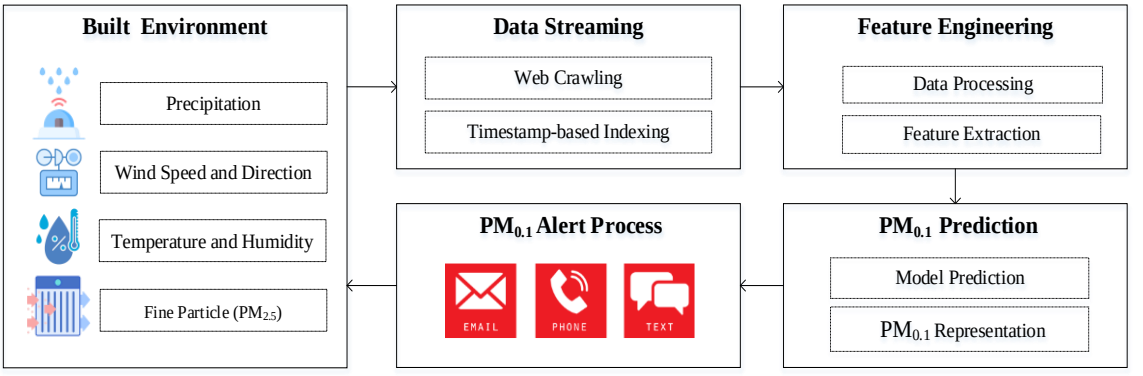


Fig. 1: The overview system of $PM_{0.1}$ concentration prediction based on fundamental ambient factors.

factors to predict $PM_{0.1}$ concentration in a particular environment. It consists of five sub-processes, including (1) Built environment, (2) Data streaming, (3) Feature engineering, (4) $PM_{0.1}$ prediction, and (5) $PM_{0.1}$ alert system.

According to Fig. 1, the built environment represents urban settings that serve human living, working, and recreational spaces. It directly interacts with people and can harm human health if it is polluted. Our proposed system measures the change of meteorological factors (*i.e.*, fine particles, temperature, humidity, wind speed and direction, and precipitation) for $PM_{0.1}$ concentration prediction. Data streaming is the first component that connects the built environment and the prediction process. It gathers the raw meteorological factor signals from the weather station provided by the official government based on credible web sources. Web crawling technology is an automatic search engine that continuously retrieves fresh weather data. Different sources may produce different ambient factors based on various time scales (*e.g.*, per hour, day, or week), and timestamp-based indexing allocates and schedules weather data and maps them into a single time-index format that can be used to compute $PM_{0.1}$ concentration. However, weather data is collected by various sensors that may include missing values, outliers, and inconsistent units (*e.g.*, Fahrenheit or Celsius). In this way, feature engineering handles those problems using data preprocessing and feature extraction to highlight each ambient factor to help support $PM_{0.1}$ concentration prediction. The $PM_{0.1}$ prediction process employs preprocessed-based meteorological factors to estimate the ultrafine particle's concentration and represent it in a human-readable format. Finally, the $PM_{0.1}$ alert process directly feedbacks the outcomes to the built environment to support people in realizing the current $PM_{0.1}$ conditions and making better decisions.

Each sub-system systematically computes based on a sequential process (see directed arrows between sub-systems in Fig. 1) and runs cycling in real time to serve the needs of $PM_{0.1}$ concentration prediction. The following section examines the sub-prediction system's step-by-step detail-based design and methodology.

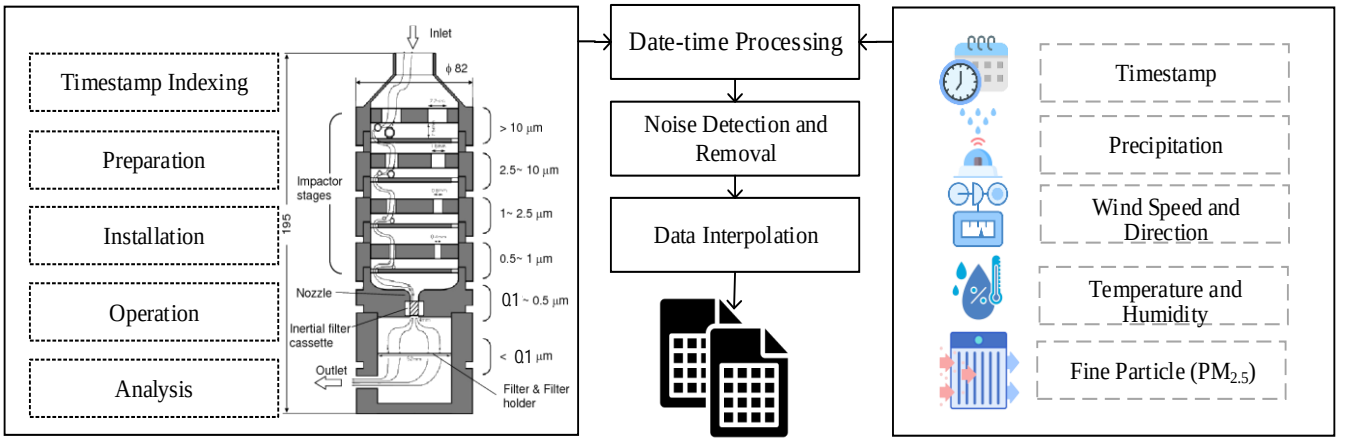
4. $PM_{0.1}$ and meteorological factor acquisition

The proposed system predicts $PM_{0.1}$ concentration based on the quality of the ambient air samplings. The prediction system's effectiveness depends on the dataset's quality for the ML process (*e.g.*, training and testing sets). In this way, the $PM_{0.1}$ dataset generation plays a critical role in the system, and its efficient dataset production process is shown in Fig. 2.

There are three main sub-processes in Fig. 2: (a) ambient $PM_{0.1}$ acquisition, (b) $PM_{0.1}$ data engineering, and (c) meteorological factor acquisition. Ambient $PM_{0.1}$ acquisition is a $PM_{0.1}$ concentration measurement process in which aerosol experts must control and operate the empirical study. It produces $PM_{0.1}$ concentration, an actual outcome as the target variable that is a critical task that the ML model wants to predict. Meteorological acquisition is a weather-related feature extraction process discovered from local meteorological data. Urban weather stations publicly provide it under the support of the meteorological department. The features are measurable variables demonstrated on real-time web applications to support ordinary people in freely accessing information concerning ambient conditions. These variables can be input features, and $PM_{0.1}$ concentration can be a predicted label. However, both processes produce the outputs from different sources, and combining them into the same format must be performed. Data engineering is a critical core of the ML model development process, encoding the relationship between labels and features. It gathers, cleans, transforms, stores, and delivers datasets (input features and labels) to build and enhance the DL regression-based model. Our proposal systematically integrates the meteorological factor and laboratory-based $PM_{0.1}$ concentrations and produces reliable datasets. In the following section, each sub-process is discussed in more detail.

4.1 Ambient $PM_{0.1}$ acquisition

This section aims to gather the $PM_{0.1}$ concentration according to the standards of professional experts as a ground truth or label for the ML dataset generation. In this way, a nanoparticle sampler is employed as a tool for $PM_{0.1}$ concentration acquisition, as contributed by Furuuchi *et al.*[24] It is designed, controlled, measured, and analyzed based on experiments with systematic procedures, and its acquisition process is shown in Fig. 2a.



(a) Laboratory-based $PM_{0.1}$ Acquisition (b) $PM_{0.1}$ Data Engineering (c) Meteorological Factor Acquisition

Fig. 2: The efficient process of $PM_{0.1}$ dataset generation consists of three sub-processes, including (a) laboratory-based $PM_{0.1}$ acquisition, (b) $PM_{0.1}$ data engineering, and (c) meteorological factor acquisition.

Fig. 2 shows the five steps of $PM_{0.1}$ concentration acquisition using the nanoparticle sampler. Preparation of the sampling includes filter treatment to remove impurities by pre-heating all filters at 350 °C for one hour, conditioning in a desiccator at 25 ± 2 °C and 50 ± 5% relative humidity (RH) for 72 h,[5,15] weighing in a 5-digit analytical balance, filter storage in an aluminum foil wrap and plastic bag before sampling, and sampler cleanup by an ultrasonic bath. The nanosampler consists of (1) four impactor stages, (2) an inertial filter cassette, and (3) a backup filter. Each impactor stage consists of a filter acting as an impactor below the corresponding acceleration nozzle to collect particles where the cut-off aerodynamic diameters are 10, 2.5, 1.0, and 0.5 microns, respectively. Inertial filters capture particles with an aerodynamic diameter >0.1 micron, and backup filters collect all the particles smaller than 0.1 micron or $PM_{0.1}$. The filters are placed in the sampler before setting it up in a sampling house stationed roughly 10 m above the ground.[14] The airflow rate is set to 40 L min^{-1} using a vacuum pump and a calibrated rotameter. The sampling period is usually 72-120 hours, depending on ambient particle concentration. A timer records actual sampling time to avoid errors in possible power outages. After sampling, the filters are folded in half, treated in the same desiccator under identical conditions as the pre-treatment, stored in the same aluminum wraps and plastic bags, weighted, and refrigerated in a freezer before chemical analysis. Finally, the recorded $PM_{0.1}$ concentration is tagged by timestamp to identify digital signature events.

Ambient $PM_{0.1}$ measurement of the abovementioned procedure makes each sample reliable and can be a label or actual outcome of meteorological factors to construct a complete dataset. The following section will examine the primary meteorological factor collection that can be gathered from available sources.

4.2 Meteorological factor acquisition

Smart cities offer real-time meteorological data based on a web application that comprises meteorological factors. It can be accessed anywhere at any time and represented at different levels (*e.g.*, global scale, regional scale, local scale).[25] Yu *et al.*[26] investigated universally available data sources of ground-based monitoring stations worldwide. They found that such relevant ambient factors were produced by various temporal resolutions such as hourly, 3-hourly, daily, monthly, and yearly. Our system might utilize these existing meteorological factors as input features.[27] The essential meteorological factors include $PM_{2.5}$, temperature, wind speed, wind direction, humidity, precipitation, radiation, atmospheric pressure, and planetary boundary layer height. Chen *et al.*[28] investigated the role of meteorological factors affecting PM. They found that (1) radiation, (2) planetary boundary layer height, and (3) atmospheric pressure are secondary factors that can be derived from temperature, humidity, and wind. To this extent, we can cut these indirect factors out of our input system and finally consist of six critical features: (1) temperature, (2) humidity, (3) wind speed, (4) wind direction, (5) precipitation, and (6) $PM_{2.5}$, vital components of the $PM_{0.1}$ prediction process. This highlights that our model's design and development process is built based on comprehensive data and generates reliable outcomes.

The web crawler can be applied to collect relevant meteorological factors for our input system. It is an essential technology that retrieves meteorological factors automatically. It employs a fixed schema based on metadata to continuously collect updated ambient information from real-time web applications at a specified period (*e.g.*, every hour).

4.3 Feature engineering

Feature engineering is the first step of DL regression-based model development, transforming the raw data into meaningful features corresponding to actual outcomes. It may concern data quality control, including outliers, missing values, duplications, and standardizations (*e.g.*, date-time format from different sources) that affect the model's predictive

performance.

Date-time processing is a core data generation process when various sources may produce relevant factors that are precursors of input features. Moreover, it is an index of the relationship between the input feature and the actual labels. In other words, an inconsistent date-time format may cause the system not to produce an accurate dataset. Furthermore, date-time may provide hidden features such as rush hour, weekend, and season that impact the $PM_{0.1}$ prediction process.

Outliers are a common problem when sensor devices produce electrical noise that is undesirable data points caused by electronic circuit issues. For instance, if an ambient temperature is detected as 185 °C by a defect sensor, it can be interpreted that such malfunction output may be affected by an inaccurate measurement. The dataset must not include it since that condition never occurs in real life. In this way, noise detection and removal are critical processes in controlling data quality.

Missing values are another common issue when sensor devices cannot recode desirable data points of certain meteorological factors caused by inconsistent government policy, improper maintenance, or failure due to human error. For example, the meteorological department may recode and store the temperature every hour and the $PM_{2.5}$ every 2 hours, causing $PM_{2.5}$ to become unknown values at the temperature time index scale. The data interpolation process estimates the missing values between two recoded values. In other words, data interpolation can estimate and fill the unknown $PM_{2.5}$ values and represent them in the same temperature time index scale.

All laboratory-based $PM_{0.1}$ and primary meteorological factor values are set in the standard time index scale, named the complete dataset, based on the machine-readable format. Pandas, an open-source Python data analysis library, can be employed to resolve the issues of the three processing components. DL regression-based model can learn the task from the dataset and improve its performance automatically.

5. DL regression-based model for $PM_{0.1}$ concentration prediction

The outstanding role of the DL regression-based model is to discover the hidden features from complex and dynamic input, such as meteorological factors aligned with labels like $PM_{0.1}$. Moreover, it supports advanced computing-based hardware like GPUs and TPUs that help learn more profoundly and faster than traditional models. This section elaborates on the critical components of the DL regression model design and development.

5.1 Area study-based data description

The air sampling was set up at the open rooftop of Sirindhorn Applied Research Engineering Building, Faculty of Engineering, Prince of Songkla University (PSU; 7°00'21.8" N, 100°30'08.6" E), Hat Yai, Songkhla, Thailand. Hat Yai is one the major cities in southern Thailand, with a population of approximately 500,000 in 2022, covering roughly 850 square kilometers. This location can be categorized by three seasons: the dry season between January and April, the pre-monsoon between May and August, and the monsoon between September and December. The wind speed averages and directions during the dry and monsoon season are 13–22 km/h east-bound and 9–11 km/h southwest-bound, respectively. In other words, the season changes (*e.g.*, meteorological factors) may characterize the behavior of the $PM_{0.1}$ concentration differently.[14]

We continuously collected the meteorological factors covering all three seasons from the local weather station under the service of the Thai Meteorological Department between January 1, 2016, and December 31, 2023. The station was close to the sampling location (around 1.5 kilometers), which can infer that factors could represent the ambient around our area study. The factors were recorded every three hours, including (1) temperature, (2) relative humidity, (3) wind speed, (4) wind direction, (5) precipitation, and (6) $PM_{2.5}$ concentration. A statistical description of the dataset is shown in Table 1 as follows:

Table 1: A statistical description of the dataset.

No.	Feature	Statistical description
1	Temperature (°C)	20.90 – 36.40 (μ: 27.77 ± 2.84)
2	Relative humidity (%)	40.00 – 100.0 (μ: 80.82 ± 13.5)
3	Wind speed (m/s)	0.000 – 10.00 (μ: 1.661 ± 2.58)
4	Wind direction (degree)	0.000 – 310.0 (μ: 42.93 ± 77.4)
5	Precipitation (millimeter)	0.000 – 48.72 (μ: 5.092 ± 8.91)
6	$PM_{2.5}$ concentration ($\mu g/m^3$)	5.000 – 37.06 (μ: 17.28 ± 5.59)
7	$PM_{0.1}$ concentration ($\mu g/m^3$)	0.000 – 6.104 (μ: 1.556 ± 1.13)

In the same period, we collected $PM_{0.1}$ concentration using the nanosampler to label the relevant meteorological factors. The procedure for acquiring $PM_{0.1}$ followed the steps outlined in section 4.1. Finally, data engineering was applied to standardize and transform relevant factors into input features and labels (complete dataset). All meteorological factors and $PM_{0.1}$ concentration were interpolated based on three-hour-based transactions following the frequency of the Thai Meteorological Department service (https://data.tmd.go.th/dataset/index.php.).[29] Finally, the dataset was stored for 23,892 transactions. However, our datasets are limited by area study with the complex data acquisition process, so we applied statistical technique-based 10-fold-cross-validation that randomly and fairly shuffles the dataset into training and testing datasets. They can be used to design and develop a DL regression-based model for $PM_{0.1}$ concentration prediction.

5.2 DL regression-based model design and development

The design of the DL regression-based model is essential for automatic knowledge discovery, such as predicting $PM_{0.1}$ concentration given complex meteorological factors. It is a process of pre-defined hyperparameters that includes (1) a model structure, (2) learning improvement, and (3) regularization. These three processes help the model determine unknown values of the model's parameters: weight (w) and bias (b). In other words, the suitable setup of the hyperparameter can automatically estimate and calibrate fitted w and b.

Pre-defined hyperparameters are technical designs that explore and combine optimal points between (1) a model structure, (2) learning improvement, and (3) regularization based on large search spaces to fit w and b. It is based on the accomplishment of the best predictive performance. The hyperparameters of each process are shown in Table 2.

Table 2: Hyperparameter configuration for DL regression-based model design.

Configuration	Hyperparameter	Search Space
Model Structure	Hidden layer	[1, 2, 3, …10]
	Neuron	[8, 16, 32, …1024]
	Activation function	ReLU, Leaky ReLU, Parametric ReLU
Learning improvement	Epoch	[100, 200, 300, …1500]
	Batch size	[4, 8, 16, …1024]
	Optimizer	Stochastic gradient descent (SGD), Adam, RMSprop
	Learning rate	[0.001, 0.050, 0.100]
	Loss function	Mean squared error (MSE) loss, Mean absolute error (MAE) loss
Regularization	Dropout	[0.1, 0.2,0.3, …1.0]
	Batch normalization	Batch size-based mean and standard
	Weight decay	[0.001, 0.050, 0.100]

Table 2 shows the hyperparameter configuration for DL regression-based model design based on includes (1) a model structure, (2) learning improvement, and (3) regularization. The model structure consists of a set of hidden layers and neurons where the purpose of the hidden layer is to encode the nonlinear function (unobservable patterns) between input and output and the neuron to encode the dynamic flow between layers.

Learning improvement automatically estimates the w and b of each neuron between hidden layers, given the training dataset. It employs epoch, batch size, optimizer, learning rate, and loss function to adjust and fit w and b given complex meteorological factors as input features and $PM_{0.1}$ concentration as desirable outcomes. Epoch represents the number of times DL repeatedly learns from the complete dataset. Batch size is the number of subsamples in each epoch that DL employs to update and adjust the w and b. The Optimizer tunes the w and b based on batch size and controls the speeds of updates using learning rates. It aims to minimize errors by comparing input features and desirable outcomes using the loss function.

Regularization applies to our case study with a high-quality measurement process, but the data was explicitly gathered in a single-sided setup. It causes overfitting problems because the model can encode the patterns in the area study but might not recognize the rest. Batch normalization rescales the input features based on the standard ranges. Dropout introduces additional patterns by randomly removing neural units that may not occur in the dataset. Weight decay generalizes the significant adjustment of w and b (exploding gradients), which is the critical cause of overcoming overfitting. It helps the learning improvement process compute w and b smoothly and generates a stable connection between layers.

Implementing hyperparameter configuration for DL regression-based model design utilized PyTorch framework and ecosystem, an open-source-based Python project for deep learning development. The development process was run based on GPU-accelerated applications with the CUDA model (NVIDIA GeForce RTX 4070 Ti). The technical concept of grid search was applied to discover optimal hyperparameters of the DL regression model for $PM_{0.1}$ concentration prediction. It explores the best hyperparameter combination, scoring the optimal discovery of the best model structure, which can minimize loss during learning in each batch. Our study based on the training set found that the pre-configured hyperparameters of the epoch, batch size, dropout, learning rate, and weight decay were 500, 16, 0.2, 0.001, and 0.001, respectively. It was based on MSE Loss, ReLU activation, and Adam optimizer, which were suitable for the DL regression modeling for the $PM_{0.1}$ concentration prediction system.

However, the recent loss measured the errors within the algorithms given training data to benefit the computer program to diminish errors (*e.g.*, misestimation of w and b between layers); it did not reflect the model performances based on a meaningful hypothesis to benefit human understanding. For instance, can the DL regression model accurately predict $PM_{0.1}$ concentration? The question uses real-world assumptions to compare the DL model's ability with human-like intelligence. In this way, the following section will validate the model performance to highlight how the proposed model achieves laboratory-based $PM_{0.1}$ measurement standards as human-like intelligence to support decision-making.

5.3 Experiment

The effectiveness of the prediction system for $PM_{0.1}$ concentration depends on the performance estimation of the DL regression model, which means it can generate the outcome of achieving laboratory-based $PM_{0.1}$ measurement

standards given meteorological factors.

The dataset comprises 23,892 transactions (as mentioned in section 5.1), presenting a limited sample size. To evaluate the generalization performance of the DL regression model and mitigate the risk of overfitting, we employed 10-fold cross-validation on the training portion. In this approach, the training dataset is randomly partitioned into ten equal-sized subsets. In each fold, nine subsets are used for training the model, while the remaining subset serves as a validation set. This iterative process ensures that every data point in the training portion is used for training and validation once during the 10-fold cross-validation procedure, and the model is evaluated on unseen validation data in each fold, providing a robust assessment of the model's generalization ability. The optimal model structure is determined by averaging the results across all ten folds, yielding a more reliable estimate of its true performance, and evaluating the test set with the selected model.

We wish to verify the optimal structure of the DL regression model that can explain the variance in the $PM_{0.1}$ concentration based on meteorological factors. Moreover, we evaluate the predictive ability, comparing the average prediction error between the DL regression model's ability (predicted outcome) and laboratory-based $PM_{0.1}$ measurement standard (actual outcome).

5.4 Evaluation metrics

This study employs two evaluation-based statistical metrics of a fundamental regression model: R-squared (R^2) and root mean squared error (*RMSE*). These metrics quantify how well the DL regression models predict the $PM_{0.1}$ concentration against actual outcomes. R^2 is a coefficient of determination metric that measures the goodness-of-fit of our proposed models, the proportion of the $PM_{0.1}$ concentration variance explained by the meteorological factors. It can be calculated as follows:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \quad (4)$$

where $\Sigma(y_i - \hat{y}_i)^2$ computes the sum of squares of errors where y_i represents the actual outcome of the $PM_{0.1}$ concentration, and $\hat{y}_i$ represents the predicted outcome of the $PM_{0.1}$ concentration. $\Sigma(y_i - \bar{y}_i)^2$ computes the sum of squares that $\bar{y}_i$ represents the simple mean of the $PM_{0.1}$ concentration (without considering the influence of the input features or meteorological factors). R^2 may take a value between 0 and 1, and 1 means 100% that the DL regression model based on meteorological factors can effectively encode all the variations in the $PM_{0.1}$ concentration.

RMSE measures the predictive errors around the line of models' best fit using the Euclidean distance between the actual measurement of $PM_{0.1}$ concentration and predicted outcomes. It can be calculated as follows:

$$RMSE = \sqrt{\sum \frac{(y_i - \hat{y}_i)^2}{n}} \quad (5)$$

where n is the number of transactions in testing data, the *RMSE* signifies how well the DL regression model can accurately estimate the $PM_{0.1}$ concentration. The *RMSE* value expresses the distance of the model estimation from the laboratory-based $PM_{0.1}$ measurement standard. If the *RMSE* value is close to 0, the DL regression model can correctly predict $PM_{0.1}$ concentration as perfectly as the laboratory-based $PM_{0.1}$ concentration measurement.

6. Results and discussion

This experiment investigates the model effectiveness of the prediction system for $PM_{0.1}$ concentration. The ML models were evaluated by *RMSE* and R^2, implemented by the sci-kit-learn, a simple and efficient Python-based library for ML validation. It explores the best-fit ML regression models, focusing on distance-based algorithms to encode the hidden pattern between the X- and Y-axises. The traditional algorithms, including linear regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), and ANN, are distance-based computing techniques that are applied to be a baseline of our proposed DL regression model. The models were implemented based on the sci-kit-learn library with default hyperparameters, which were set up as follows: KNN (K = *5*, weights = *uniform*, and metric = *Euclidean distance*), SVM (regularization parameter (*C*) = 1.0 and kernel = radial basis function (*RBF*), and ANN (hidden layer's neurons = *1024*). We repeated the experiment ten times to ensure the results did not occur by chance (*e.g.*, by initializing intercept and coefficient). This way, the best results of each traditional distance-based algorithm's learning performance were compared based on the ten-round *RMSE* and R^2 mean with standard deviation (SD), as shown in Table 3.

Table 3: The comparative effectiveness of four traditional ML regression models using *RMSE* and R^2.

ML model	R^2 mean	R^2 SD	*RMSE* mean	*RMSE* SD
LR	0.4436	0.0077	0.7098	0.0101
KNN	0.5991	0.0078	0.6025	0.0114
SVM	0.5791	0.0129	0.6172	0.0093
ANN	0.7718	0.0204	0.4540	0.0194

Table 3 compares four traditional ML regression models, and overall performances are quite low. LR produced the lowest R^2, achieving 44.36% of $PM_{0.1}$ concertation variations, and the highest *RMSE*, producing an error of 0.7098 ± 0.0101 $\mu g/m^3$. This means that LR could deal with the uncertainty of complex meteorological factors poorly, while KNN and SVM could control uncertainty moderately. This is because these traditional models do not support regularization and limit the learning process's ability to achieve accurate performance. This suggests that they may not fit to encode uncertain patterns of $PM_{0.1}$ concentration in dynamic environments.

In contrast, ANN performed better than the rest because it had an optional function within hidden layers and neurons to model nonlinear patterns of $PM_{0.1}$ concentration. The ANN's

results are acceptable, achieving 77.18% of $PM_{0.1}$ concertation variations and the lowest *RMSE*, approximately 0.4540 ± 0.0194 µg/m^3. However, improving the baseline models' performance to be close to the laboratory-based measurement is challenging. In this way, we highlight this impressive gap by proposing DL regression models that employ pre-configuration of automatic learning improvement and regularization hyperparameters (see Table 2). The comparative results of seven DL regression models are shown in Table 4.

Table 4: The comparative effectiveness of seven DL regression models using *RMSE* and R^2.

Layer	Neuron	R^2 mean	R^2 SD	*RMSE* mean	*RMSE* SD
1-hidden layer	1024	0.8571	0.0077	0.3566	0.0071
2-hidden layer	1024-256	0.8869	0.0019	0.3173	0.0026
3-hidden layer	512-1024-512	0.9098	0.0030	0.2833	0.0035
4-hidden layer	512-1024-256-128	0.9229	0.0028	0.2620	0.0032
5-hidden layer	256-512-1024-256-128	0.9252	0.0027	0.2581	0.0042
6-hidden layer	128-256-512-1024-512-256	0.9158	0.0039	0.2738	0.0053
7-hidden layer	128-256-512-512-256-128-128	0.9151	0.0016	0.2749	0.0026
Average		0.9046	0.0034	0.2954	0.0040

Table 4 determines seven DL regression models based on their hidden layers and neurons. Average R^2 approved which models effectively fit meteorological factors aligned with $PM_{0.1}$ concertation, and average *RMSE* approved the best predictive performance of the models. Overall, the models could fit $PM_{0.1}$ concertation perfectly, which is 90.46 ± 0.34%, and could predict excellent performance with a slight error of 0.2954 ± 0.0040 µg/m^3. This highlights our research contribution that the DL regression model could estimate $PM_{0.1}$ concentration using available meteorological factors, including $PM_{2.5}$, temperature, humidity, wind speed and direction, and precipitation. However, delving into the specific aspects of the models, we can see competent differences in details.

The 1- and 2-hidden layer-based models could fit variations of $PM_{0.1}$ concertation, which achieved 85.71% and 88.69%, respectively. The 1-hidden layer-based model produced the highest error, 0.36 µg/m^3, but outperformed the standard ANN because it employed the hyperparameters-based configuration (see Table 1). If the 1-hidden layer-based model predicts $PM_{0.1}$ concentration to be 4.10 µg/m^3, it might be far from the laboratory-based measurement standard, around 3.74 and 4.46 µg/m^3. This is because of the underfitting problems; the 1-hidden layer cannot encode the multi-dimensions of meteorological input features, which dynamically change in nature and complex patterns, and some features may depend on one another.

On the other hand, the 4- and 5-hidden-layer-based models perform the best-fit models, which reached 92.29% and 92.25%, respectively. They could effectively estimate $PM_{0.1}$ concertation with slight errors of 0.2620 ± 0.0032 and 0.2851 ± 0.0042 µg/m^3 compared to the laboratory-based $PM_{0.1}$ measurement. The additional layers help encode the complexity of meteorological factors. The results highlight that the DL regression models can improve the predictive performance higher than the Artificial Neural Network (see 1-layer-based model results) by approximately 6%.

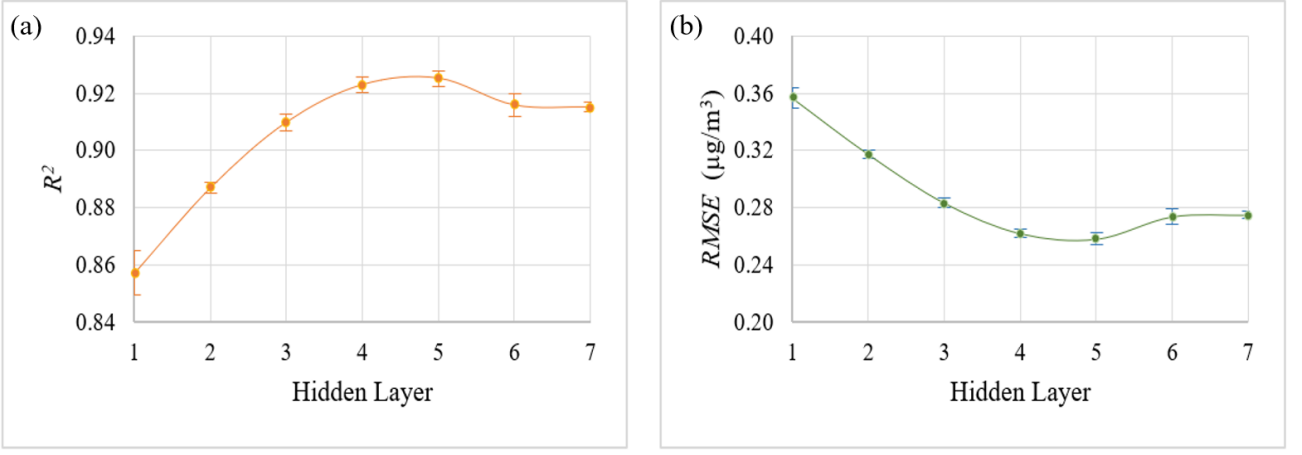


Fig. 3: The trends and directions of seven DL regression models are based on different hidden layers, using *RMSE* for predictive ability and R^2 for learning ability. (a) The learning ability of DL regression models encodes the variance in the $PM_{0.1}$ concentration based on the primary meteorological factors. (b) The predictive ability is based on error comparisons between the DL regression model's ability and human-like intelligence.

In the 6- and 7-hidden-layer-based models, however, R^2 and *RMSE* were worse than the 4- and 5-hidden-layers while having more layers and neurons. Moreover, the R^2 and *RMSE* SDs were widespread and inconsistent in the prediction when the unseen data were fed. This is because it suffered from overfitting that learned strictly fitting with training data and encoded too many patterns based on intensive layers and neurons. This suggests that hidden layers over five may be wasteful for $PM_{0.1}$ concertation modeling.

Fig. 3 demonstrates the trends and directions of the R^2-based learning ability and *RMSE*-based predictive ability of seven DL regression models under different hidden layers. Figs. 3(a) and (b) expressed that the number of hidden layers lower and higher than 4 and 5 could be underfitted and overfited, respectively, and adding more layers and neurons into the model structures cannot help improve the predictive and learning abilities. The graphs also highlight that the performances of the 4- and 5-hidden-layer-based models are almost identical, but the 4-hidden-layer-based model may be more suitable. This is especially true if computational costs are concerned. Since more layers and neurons are added, more computational costs are required. This suggests that the 4-hidden-layer-based model is good enough for an intelligent-driven $PM_{0.1}$ concentration prediction based on meteorological factors, which helps produce accurate results based on laboratory-based $PM_{0.1}$ measurement standards. However, some meteorological features still face challenges in discovering new input features that may improve their model accuracies. For instance, recent input features did not consider the sequences of events based on long-term and short-term dependencies, which are critical features in estimating $PM_{0.1}$ concentration patterns.

This suggests that the 4-hidden-layer-based model is good enough for an intelligent-driven $PM_{0.1}$ concentration prediction based on meteorological factors, which helps produce accurate results based on laboratory-based $PM_{0.1}$ measurement standards.

Fig. 4 illustrates the optimal DL regression model's performance, including training and validation loss curves obtained through 10-fold cross-validation and comparing outcomes between predicted $PM_{0.1}$ concentrations and those measured in the laboratory.

In summary, this research demonstrates the potential of a DL regression model to accurately predict $PM_{0.1}$ concentrations in urban environments, offering a valuable system for air quality monitoring and management. The system employs available meteorological data from open services to provide near real-time predictions of $PM_{0.1}$ concentrations at specific city locations, yielding results comparable to laboratory-based $PM_{0.1}$ measurements. It is based on DL regression-based technologies that provide an alternative way to overcome sophisticated processes and expensive instruments of laboratory-based $PM_{0.1}$ measurement. The vital contribution is that our proposed model can improve the predictive performance by approximately 15% compared to traditional ANN, with particularly strong performance in urban cities with warm and high-humid climates.

The model proved particularly effective in handling our city's complex meteorological patterns, successfully incorporating multiple factors, including $PM_{2.5}$, temperature, humidity, wind patterns, and precipitation. The seven-year dataset effectively captures periodic and seasonal variations, including dry seasons, monsoons, and notable climate events (*e.g.*, high pollution episodes in the harvesting season affected by agriculture hotspots), especially in its analysis of $PM_{0.1}$ dynamics.

This suggests that it can support environmental authorities in accessing $PM_{0.1}$ concentration information to issue timely air quality advisories and plan adaptation strategies for vulnerable populations, inform urban planning decisions, and guide the development and implementation of effective air quality regulations and pollution control strategies. Moreover, this benefits regions with limited access to expensive laboratory-based $PM_{0.1}$ monitoring equipment, enabling more cost-effective and accessible air quality assessments.

Future research can focus on evaluating and refining the model's performance in diverse urban settings and incorporate

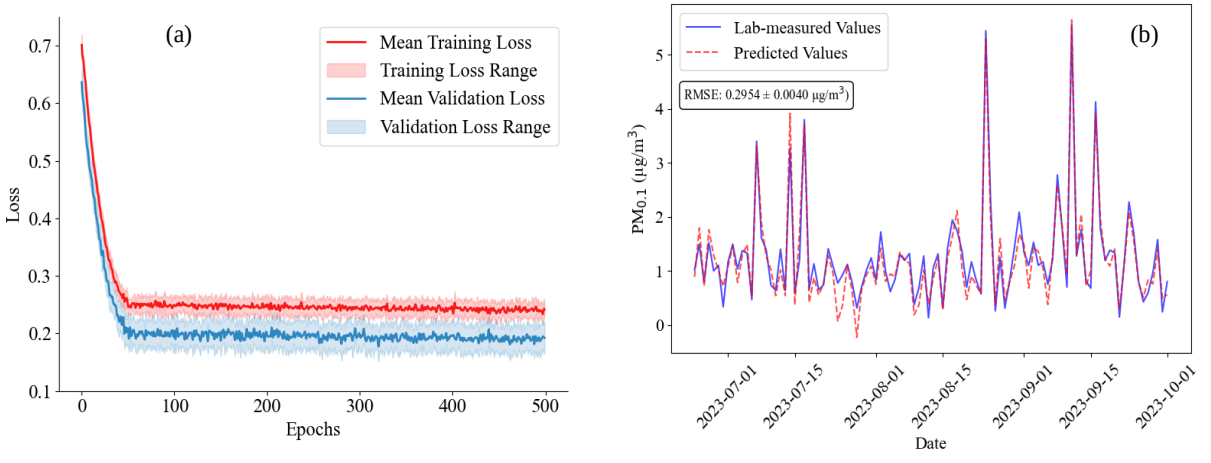


Fig. 4: The performance of the optimal DL model, (a) Training and validation loss curves of the 10-fold cross-validation, (b) Comparison of outcomes between predicted and laboratory-measured $PM_{0.1}$ concentrations.

additional environmental factors like land use and traffic data. Moreover, some meteorological features still face challenges in discovering new input features that may improve their model accuracy. For instance, recent input features did not consider the sequences of events based on long-term and short-term dependencies, which are critical features in estimating $PM_{0.1}$ concentration patterns.

7. Conclusion

This research has proposed a DL regression model for the $PM_{0.1}$ concentration prediction system to estimate accurate results based on laboratory-based measurement standards. This approach advances our understanding of atmospheric composition by providing a novel method to assess ultrafine particle concentrations. It is especially needed when the government cannot provide the $PM_{0.1}$ measurement infrastructure (because of sophisticated processes and expensive instruments), and environmental authorities cannot access $PM_{0.1}$ information. This limitation in data availability hinders comprehensive analysis of atmospheric components and their impacts. Lack of information causes decision-makers and policymakers to be challenged to manage and control the situations. In contrast, our proposed model enables environmental authorities to monitor and detect the concentration of $PM_{0.1}$, avoid risky environments, and prevent severe health conditions.

The models employ existing meteorological factors (*i.e.*, $PM_{2.5}$, temperature, humidity, wind speed and direction, and precipitation) and fit with $PM_{0.1}$ collected by the nanoparticle sampler, an accurate device for $PM_{0.1}$ measurement. The design and development of the DL regression-based model focuses on investigating the best-fit model structures based on different hidden layers and neurons (w and b). The suitable hyperparameters are explored and utilized to pre-configure for model structure discovery to encode complex $PM_{0.1}$ concentration patterns with precise prediction and produce reliable outcomes.

Our experiments based on R^2 and *RMSE* prove that the proposed models are reasonably acceptable, employ simple meteorological factors, and predict accurate $PM_{0.1}$ concentration in laboratory-based measurement standards. Moreover, the results demonstrate that the model structure based on 4-hidden layers is the best fit, which is optimal for the $PM_{0.1}$ prediction system (R^2 = 92.29% and *RMSE* = 0.26 $\mu g/m^3$). This suggests that the proposed model can add critical significance to the $PM_{0.1}$ concentration prediction system to generate accurate and reliable results. It ensures that even countries with poorly prepared technologies, environmental authorities, decision-makers, and policymakers can still employ alternative information to understand the situation and manage $PM_{0.1}$ pollution confidently and efficiently.

Our proposed model focuses on a single site, a big-picture area study. However, the characteristics of $PM_{0.1}$ concertation may vary based on the geographical location, contributing to spatial variations in atmospheric composition. For instance, the near-roadway area and industrial region may produce $PM_{0.1}$ concentration differently, reflecting diverse anthropogenic sources and their impacts on local air quality. Moreover, cultural activities such as burning incense in Asia, part of everyday life, may produce intensive $PM_{0.1}$ concentration, highlighting the need to understand region-specific contributions to atmospheric particulate matter. In the future, we will expand the models for $PM_{0.1}$ concertation prediction to new environments, focusing on various urban areas with specific problems to help support particular occupants (*e.g.*, older people and people with respiratory conditions). This expansion will enhance our understanding of $PM_{0.1}$ distribution and transformation processes across atmospheric conditions. We also plan to consider advanced model mechanisms to identify cause-and-effect relationships between meteorological factors that may semantically influence $PM_{0.1}$ concertation. This may benefit ordinary occupants by helping them understand why such $PM_{0.1}$ concertation is generated and helping government authorities design the policy for long-term sustainability and occupant wellbeing.

Acknowledgment

This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation (Grant No. B40G660042) and the Research and Innovation Institute of Excellence, Walailak University (Grant No. WU68212).

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable.

References

[1] V. K. Yadav, S. Bijekar, A. Gacem, A. M. Alkahtani, K. K. Yadav, M. A. Alreshidi, P. Kumar, T. Ghosh, R. K. Verma, S. Mishra, A. Patel, N. Choudhary, The impact of fine particulate matters (PM_{10}, $PM_{2.5}$) from incense smokes on the various organ systems: a review of an invisible killer, *Particle & Particle Systems Characterization*, 2024, **41**, 2300157, doi: 10.1002/ppsc.202300157.

[2] I. Gryech, C. Asaad, M. Ghogho, A. Kobbane, Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: A systematic literature review, *Engineering Applications of Artificial Intelligence*, 2024, **137**, 109182, doi: 10.1016/j.engappai.2024.109182.

[3] S. F. I. Abdillah, Y.-F. Wang, Ambient ultrafine particle ($PM_{0.1}$): sources, characteristics, measurements and exposure implications on human health, *Environmental Research*, 2023, **218**, 115061, doi: 10.1016/j.envres.2022.115061.

[4] Z. Mao, Y. Wu, L. Kong, L. Zhou, X. Zhang, A. Geng, J. Cai,

H. Yang, P. Huang, Changes in cargoes of platelet derived extracellular vesicles heterogeneous subpopulations induced by PM0.1: Undisclosed cardiovascular injury communication mechanism, *Environmental Pollution*, 2024, **348**, 123845, doi: 10.1016/j.envpol.2024.123845.

[5] W. Phairuang, M. Inerb, M. Furuuchi, M. Hata, S. Tekasakul, P. Tekasakul, Size-fractionated carbonaceous aerosols down to PM0.1 in southern Thailand: Local and long-range transport effects, *Environmental Pollution*, 2020, **260**, 114031, doi: 10.1016/j.envpol.2020.114031.

[6] S. D. Lowther, K. C. Jones, X. Wang, J. Duncan Whyatt, O. Wild, D. Booker, Particulate matter measurement indoors: a review of metrics, sensors, needs, and applications, *Environmental Science & Technology*, 2019, **53**, 11644-11656, doi: 10.1021/acs.est.9b03425.

[7] E. I. Fernández, A. J. Jara Valera, J. T. Fernández Breis, Embedded machine learning of IoT streams to promote early detection of unsafe environments, *Internet of Things*, 2024, **25**, 101128, doi: 10.1016/j.iot.2024.101128.

[8] W. Y. Hong, Meteorological variability and predictive forecasting of atmospheric particulate pollution, *Scientific Reports*, 2024, **14**, 14, doi: 10.1038/s41598-023-41906-8.

[9] E. Kalantari, H. Gholami, H. Malakooti, M. Eftekhari, P. Saneei, D. Esfandiarpour, V. Moosavi, A. R. Nafarzadegan, Evaluating traditional versus ensemble machine learning methods for predicting missing data of daily PM_{10} concentration, *Atmospheric Pollution Research*, 2024, **15**, 102063, doi: 10.1016/j.apr.2024.102063.

[10] S. Zhou, W. Wang, L. Zhu, Q. Qiao, Y. Kang, Deep-learning architecture for $PM_{2.5}$ concentration prediction: A review, *Environmental Science and Ecotechnology*, 2024, **21**, 100400, doi: 10.1016/j.ese.2024.100400.

[11] M. A. Bhatti, Z. Song, U. A. Bhatti, S. M. S, AIoT-driven multi-source sensor emission monitoring and forecasting using multi-source sensor integration with reduced noise series decomposition, *Journal of Cloud Computing*, 2024, **13**, 65, doi: 10.1186/s13677-024-00598-9.

[12] D. E. Schraufnagel, The health effects of ultrafine particles, *Experimental & Molecular Medicine*, 2020, **52**, 311-317, doi: 10.1038/s12276-020-0403-3.

[13] C. Wang, J. Xiang, E. Austin, T. Larson, E. Seto, Quantifying the contributions of road and air traffic to ambient ultrafine particles in two urban communities, *Environmental Pollution*, 2024, **348**, 123892, doi: 10.1016/j.envpol.2024.123892.

[14] C. T. Pham, T. D. Nghiem, H. T. Le, H. D. Chu, T. Tran Viet, K. Sekiguchi, N. Tang, K. Hayakawa, A. Toriba, Size distribution of airborne particle-bound polycyclic aromatic hydrocarbons during rice straw open burning in Hanoi, Vietnam, *Atmospheric Pollution Research*, 2024, **15**, 102115, doi: 10.1016/j.apr.2024.102115.

[15] N. Mahasakpan, P. Chaisongkaew, M. Inerb, N. Nim, W. Phairuang, S. Tekasakul, M. Furuuchi, M. Hata, T. Kaosol, P. Tekasakul, R. Dejchanchaiwong, Fine and ultrafine particle- and gas-polycyclic aromatic hydrocarbons affecting southern Thailand air quality during transboundary haze and potential health effects, *Journal of Environmental Sciences*, 2023, **124**, 253-267, doi: 10.1016/j.jes.2021.11.005.

[16] M. L. Bergmann, Z. J. Andersen, A. Massling, P. A. Kindler, S. Loft, H. Amini, T. Cole-Hunter, Y. Guo, M. Maric, C. Nordstrøm, M. Taghavi, S. Tuffier, R. So, J. Zhang, Y.-H. Lim, Short-term exposure to ultrafine particles and mortality and hospital admissions due to respiratory and cardiovascular diseases in Copenhagen, Denmark, *Environmental Pollution*, 2023, **336**, 122396, doi: 10.1016/j.envpol.2023.122396.

[17] Y. Kurotsuchi, K. Sekiguchi, Y. Hayakawa, Divisive refinement of metal fiber at the PM0.1 classification stage for PM0.5-0.1 sampling with nanosampler, *Aerosol and Air Quality Research*, 2023, **23**, 220439, doi: 10.4209/aaqr.220439.

[18] A. K. Tripathi, M. Aruna, S. Parida, D. Nandan, P. V. Elumalai, E. Prakash, J. S. C. Isaac JoshuaRamesh Lalvani, K. S. Rao, Integrated smart dust monitoring and prediction system for surface mine sites using IoT and machine learning techniques, *Scientific Reports*, 2024, **14**, 7587, doi: 10.1038/s41598-024-58021-x.

[19] E. Illueca Fernández, I. C. Martínez, J. T. F. Breis, A. J. J. Valera, Design and evaluation of a dryer system for IoT hyperlocal particulate matter monitoring device, *IEEE Sensors Journal*, 2024, **24**, 11152-11165, doi: 10.1109/JSEN.2024.3364537.

[20] A. Mishra, Y. Gupta, Comparative analysis of Air Quality Index prediction using deep learning algorithms, *Spatial Information Research*, 2024, **32**, 63-72, doi: 10.1007/s41324-023-00541-1.

[21] Y. Bengio, Deep learning of representations for unsupervised and transfer learning, *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop*, 2011, **27**, 17-37, doi: 10.5555/3045796.3045800.

[22] A. Apicella, F. Donnarumma, F. Isgrò, R. Prevete, A survey on modern trainable activation functions, *Neural Networks*, 2021, **138**, 14-32, doi: 10.1016/j.neunet.2021.01.026.

[23] Q. Wang, Y. Ma, K. Zhao, Y. Tian, A comprehensive survey of loss functions in machine learning, *Annals of Data Science*, 2022, **9**, 187-212, doi: 10.1007/s40745-020-00253-5.

[24] M. Furuuchi, K. Eryu, M. Nagura, M. Hata, T. Kato, N. Tajima, K. Sekiguchi, K. Ehara, T. Seto, Y. Otani, Development and performance evaluation of air sampler with inertial filter for nanoparticle sampling, *Aerosol and Air Quality Research*, 2010, **10**, 185-192, doi: 10.4209/aaqr.2009.11.0070.

[25] A. Kaginalkar, S. Kumar, P. Gargava, D. Niyogi, Review of urban computing in air quality management as smart city service: An integrated IoT, AI, and cloud technology perspective, *Urban Climate*, 2021, **39**, 100972, doi: 10.1016/j.uclim.2021.100972.

[26] W. Yu, J. Song, S. Li, Y. Guo, Is model-estimated $PM_{2.5}$ exposure equivalent to station-observed in mortality risk assessment? A literature review and meta-analysis, *Environmental Pollution*, 2024, **348**, 123852, doi: 10.1016/j.envpol.2024.123852.

[27] F. Zeng, C. Pang, H. Tang, Sensors on Internet of Things systems for the sustainable development of smart cities: a systematic literature review, *Sensors*, 2024, **24**, 2074, doi:

10.3390/s24072074.
[28] Z. Chen, D. Chen, C. Zhao, M. P. Kwan, J. Cai, Y. Zhuang, B. Zhao, X. Wang, B. Chen, J. Yang, R. Li, B. He, B. Gao, K. Wang, B. Xu, Influence of meteorological conditions on $PM_{2.5}$ concentrations across China: a review of methodology and mechanism, *Environment International*, 2020, **139**, 105558, doi: 10.1016/j.envint.2020.105558.
[29] N. Phumkokrux, P. Trivej, Investigation of temperature, precipitation, evapotranspiration, and new thornthwaite climate classification in Thailand, *Atmosphere*, 2024, **15**, 379, doi: 10.3390/atmos15030379.