Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

Indoor Air Quality Prediction in Sick Building Using Machine and Deep Learning: Comparative Analysis

Hayder Qasim Flayyih[1,*], Jumana Waleed[1] and Amer M. Ibrahim[2]

[1] Department of Computer Science, College of Science, University of Diyala, Iraq
[2] Department of Civil Engineering, College of Engineering, University of Diyala, Baqubah, Diyala, 32001, Iraq

ARTICLE INFO

Article history:
Received April 23, 2024
Revised December 18, 2024
Accepted January 2, 2025
Available online March 1, 2025

Keywords:

Indoor air quality
Carbon Dioxide
Air pollution
Sick building syndrome
Machine and deep learning

ABSTRACT

Air pollution is a significant global concern that is continually increasing and threatening both the environment and human health. Air pollution is the principal factor leading to the deterioration of Indoor Air Quality (IAQ) in buildings. Carbon dioxide (CO2) significantly contributes to indoor pollution intensifying, primarily from human activities. The demand for effective IAQ systems has increased due to the necessity for sustainable building development. The artificial intelligence (AI) models presented in this work utilized Machine Learning (ML) and Deep Learning (DL) methodologies to train the available dataset. This dataset was collected by the indoor sensors in Shanghai from November 2016 to March 2017 to predict CO2 concentration and obtain pertinent information. The accuracy and the result of Ml and DL algorithms may differ depending on the datasets used and the algorithms' suitability for the specific data and application domain. Therefore, a significant benefit would be achieved by finding the best-fitted ML and DL models concerning the actual datasets and the application area. This necessity was fulfilled through an intensive review of the already existing DL and ML models. This analysis aims to implement the specified models and assess the efficiency of their prediction by computing several performance metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Median Absolute Error (Median AE), and Coefficient of determination (R2). Among implemented models, the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have performed better results in forecasting IAQ.

1. Introduction

Indoor Air Quality (IAQ) profoundly impacts human health, comfort, productivity, and general well-being. Poor IAQ can contribute to conditions, including Sick Building Syndrome (SBS) and Building-Related Illness (BRI), where occupants of buildings feel uncomfortable and have health problems due to their continued exposure to indoor pollutants [1]. Besides causing health and psychological problems, the occurrence of SBS indicates the regulatory frameworks relating to the quality of air, starting with the Air Quality Act of 1967, followed by its amendments in 1977 and 1990, which had been targeted to make amends in the air quality. While the primary concern of these regulations is related to outdoor air pollution, the same regulations assert the importance of maintaining a good IAQ in the environment [2].

Rapid urbanization and heavy infrastructure shortages, especially in developing countries, trigger rising pollution levels, notably from

* Corresponding author.
E-mail address: scicomphd222303@uodiyala.edu.iq
DOI: 10.24237/djes.2025.18112

This work is licensed under a Creative Commons Attribution 4.0 International License.

vehicle emissions. Transportation continues to be the leading source of urban air pollution, followed by industrial and agricultural activities. It is dangerous because such sources emit dangerous pollutants into the atmosphere that could affect human health, ranging from mild to severe, depending on the duration of exposure, concentration of pollutants, and health status of the people exposed. While outdoor air pollution is considered a widespread health threat, it is usually forgotten that indoor air quality can have profound implications for human health. Since more and more people nowadays work in 'closed' office spaces, recent research has placed greater emphasis on IAQ in workplaces, studying specific pollutants-emitting sources like photocopiers and printers. A prerequisite for good IAQ is the availability of purified outdoor air [3]. However, buildings located near sources of outdoor air pollution, such as freeways or markets, cannot consistently provide good indoor air quality owing to the infiltration of polluted outdoor air. Several surveys have reported that the detrimental effects of indoor air pollutants are equally harmful as those of outdoor, which alone can cause a wide range of health issues [4].

The three main indoor pollutant types identified by researchers include gases, particulate matter, and biological contaminants. Most indoor pollutants studied worldwide have included volatile organic compounds, aldehydes, ammonia, and particulate matter. Low ventilation has been linked with higher allergen concentrations and asthma, apart from the increased case of SBS. Buildings with too low ventilation consume more energy to maintain a comfortable indoor climate [5].

Naturally, ventilation usually cannot solve IAQ problems in heavily populated urban areas. It simply does not suffice to remove contaminated indoor air, especially when outdoor air is commonly or heavily polluted. Natural ventilation might be less adequate because of the timing of window operations for indoor activities. In addition to outdoor air quality, indoor Carbon Dioxide (CO_2) levels depend on human activity within the space [6]. One of the most effective ways to guarantee IAQ is by providing a constant supply of fresh air. However, this is usually difficult when the outdoor air is also of poor quality. The relationship between outdoor air quality and indoor health outcomes has been determined within the last several years. This is also reflected in the regulation of the Clean Air Act. Indeed, the Clean Air Act has evolved from the early beginnings of the Air Pollution Control Act of 1955 into a current thrust toward IAQ issues, particularly in heavily populated urban areas [7, 8]. Due to this, other regulations, such as the Clean Windows and Doors Act, were enacted in the 1960s and 1970s to better preserve indoor environments inside office buildings. Besides air pollution, other factors, such as noise pollution, also threaten the health of the occupants [9]. Several health disorders have been discovered that are directly linked to high levels of indoor contaminants; these include SBS and BRI, which can further result in chronic health disorders. The World Health Organization (WHO) estimated that indoor air pollutants accounted for 2.7% of the total global diseases in 2010 and estimated that as many as 3.8 million deaths were caused by poor IAQ in 2018. Besides, air pollutants like carbon compounds, nitrogen oxides, sulfur oxides, ultra-fine particles, and particulate matter can enter the buildings through ventilation facilities, thereby developing IAQ problems [10, 11].

Recent research targets Artificial Intelligence (AI) techniques to regulate and enhance IAQ. Application systems, with the incorporation of AI, fuzzy logic (FL), and genetic algorithms (GA), are starting to appear in literature as intelligent ventilation systems and show a promising reduction in CO_2 and other harmful pollutants [12]. Such approaches use variables like temperature and humidity to make decisions using decision trees that forecast CO_2 levels in smart homes. When incorporated into a building management system, these advanced models promise even better IAQ predictions and energy efficiency [13, 14]. Other studies compare methods to model CO_2 in offices and examine the effectiveness of Machine Learning (ML) and Deep Learning (DL) techniques in indoor pollutant-level predictions [15, 16]. The main contributions of this paper are as follows:

1. Implementing diverse ML and DL models to efficiently predict IAQ, including Decision Tree Regressor (DTR), Random Forest (RF), K-Nearest Neighbors(K-NN), Support Vector Regressor (SVR), and Gradient Boosting Regressor (GBR), Deep Adaptive Quantization Feature Fusion (DAQFF), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), Bidirectional Long Short-Term Memory (Bi-LSTM), Random Forest-Long Short-Term Memory (RF-LSTM), CNN-LSTM, and Deep Neural Network (DNN).
2. Checking the performance of implemented models in predicting indoor CO_2 concentration while finding out which methods are most effective for assuring the best condition related to IAQ inside buildings using a dataset with temperature, CO_2, relative humidity, Particulate Matter (PM10), PM2.5, and Volatile Organic Compounds (VOCs) variables.
3. Comparing the performance of implemented ML and DL models for indoor CO_2 prediction. While the GBR model outperforms existing ML algorithms with the highest benchmark for IAQ prediction, the LSTM model is strong enough to model temporal dependencies in time-series data.

To address these challenges, recent advancements in artificial intelligence techniques for IAQ prediction have gained prominence. The next section reviews the existing literature, highlighting the contributions and limitations of prior works in this domain and identifying the gaps this study aims to fill. The paper is organized as follows: Section 2 covers the related work, while Section 3 describes the Data Sources and Preparation. Section 4 explains the Methods utilized. Section 5 introduces the Accuracy Evaluation Metrics, while Section 6 presents the performance metrics used to evaluate the models. Section 7 discusses the results. Finally, Section 8 presents the conclusion.

2. Related works

Monitoring CO_2 levels in buildings has various applications: controlling Heating, Ventilation, and Air Conditioning (HVAC) systems, predicting occupancy, and even performing Computational Fluid Dynamics (CFD) analysis. Combining CO_2 monitoring with other building data may bring substantial value in optimizing building performance and energy use.

Li and Sun [17] presented an ML framework that could predict CO_2 emissions at a city level using open-access data. The feature selection methods used in this study included recursive feature elimination and Boruta to drive key variables for CO_2 emissions. Consequently, XGBoost produced the best predictions, with an R^2 higher than 0.98, recording lower errors than the rest of the models. By investigating the Sulfur Dioxide (SO_2) industrial emission as a socioeconomic predictor, it has been found that such a predictor can be used to estimate the CO_2 emissions in 182 Chinese cities.

Taheri and Razban [18], presented an ML model to forecast indoor CO_2 concentration for optimizing demand-controlled ventilation systems. The performance evaluation was accomplished for six algorithms involving the SVR, AdaBoost, RF, and Multilayer Perceptron (MLP) using data recorded in a classroom with variable occupation. Among all models, MLP gave the best results by providing highly accurate prediction values of CO_2 concentration. This control strategy reduced energy use by HVAC by 51.4% while maintaining compliance with the ASHRAE standards.

Marzouk and Atef [19], developed an IoT-based IAQ monitoring system for academic building environments. This system allowed for real-time measurement and forecasting of temperature, humidity, air pressure, CO_2, Carbon Monoxide (CO), and PM2.5 using sensor measures with AI models. It was suitable for the job, causing little interference with everyday activities, and hence capable of controlling and forecasting IAQ.

Zhu et al. [20], developed an LSTM-based, IoT-enabled CO_2 steady-state forecasting system to monitor IAQ. IoT sensors collect the

current CO_2 level in real time; hence, this developed system uses LSTM to predict future concentrations with as low as 5.5% error margins. This system focused on measuring CO_2 as a proxy for IAQ and infection risks, including COVID-19.

Dai et al. [21], proposed a hybrid model of RF, tree-structured Parzen estimator, and LSTM for indoor CO_2 concentration forecast in university classrooms is suggested, which includes RF selection of features, optimization by the estimator, and LSTM for time-series prediction. This model provided high accuracy with an R^2 of more than 98%, and the error has reduced to the least value with Mean Absolute Error (MAE) of 2.96 and Root Mean Squared Error (RMSE) of 5.54.

Kim et al. [22], developed a CO_2-driven ML model for the estimation of occupancy level inside buildings by considering IoT sensors, ventilation systems, and differential pressure gauges. This given research work has a very appropriate relevance to the correct estimation of occupancy in respect of IAQ management and infection transmission control. This model with RF and Artificial Neural Networks had an accuracy rate of 0.91 and 0.92 with CO_2 and ventilation inputs, respectively. Including differential pressure data decreased the accuracy slightly; hence, further studies should tune this integration for better predictions.

Kapoor et al. [2], conducted an exploratory investigation into the application of ML models for the prediction of CO_2 levels in office buildings. Real-time metrics such as occupancy, space per person, outside temperature, wind speed, humidity, and the IAQ were employed as input. In this study, the optimized Gaussian process regression model performed better among the compared models, providing an accuracy of 0.98 and RMSE of 4.20 ppm. These results anticipate that this model could serve well in applications related to ventilation design and IAQ testing of urban buildings.

Taheri and Razban [8] developed a ML based methodology to predict indoor CO_2 concentration and optimize DCV systems. SVR, AdaBoost, and MLP models were evaluated in this work using CO_2 and meteorological data measured in a classroom with variable occupancy. Among those models, the best performance for predicting CO_2 concentration was provided using the MLP. The developed control strategy achieved an average HVAC energy-use saving of 51.4%, and the calculated average IAQ was within the acceptable standard set by ASHRAE IAQ.

While previous studies demonstrate the potential of ML and DL models for IAQ prediction, a comprehensive comparison of their performance using diverse metrics remains underexplored. This study addresses these gaps by implementing and evaluating a wide range of ML and DL models, as described in the methodology section.

3. Artificial intelligence techniques

Computing systems that mimic human cognition, action, and behavior are known as AI technologies. AI is a subfield of computer science that aims to give computers the ability to think and reason like humans [23]. One branch of AI is ML, which allows computers to gradually improve their performance by analyzing data and applying what they've learned, all without human intervention or code. A subfield of ML is DL, which utilizes neural networks to detect subtle and complicated patterns within the data. ML is an aspect of AI's larger goal of automating tasks [24].

Several ML models improve the ability of a computing device to acquire knowledge and modify it without explicit instructions [25]. These algorithms use analytical and predictive modeling approaches to help computers distinguish different patterns. In ML, the presented information may be analyzed to determine the model's ability to extract and identify further hidden patterns and data correlations using data from the drop column. ML models are grouped into three types; supervised, unsupervised, and reinforcement learning. Each is distinct and created specifically to address the many issues connected with analyzing information and prediction [26]. DL is an advanced subset of AI that explores the complexities of ML. In contrast to traditional ML, it utilizes complex neural network topologies designed to emulate neurons

in the human brain. The networks are designed to analyze and learn from extensive datasets, making DL very effective for jobs that require precision and accuracy [27, 28]. DL models involve the development of deep neural networks that can effectively make accurate classifications and provide precise numerical forecasts. These models auto-discover features, distinguish complex patterns, and enhance performance over time with processed data to achieve more accuracy and reliability in the outcomes [29].

4. Methodology

In this work, various ML and DL models, were selected to predict IAQ. Each model was selected based on how it treats the dataset's complexities; thus, each may effectively predict CO_2 concentration. Figure 1 illustrates the design and functionality of ML and DL models' utilization in IAQ prediction.

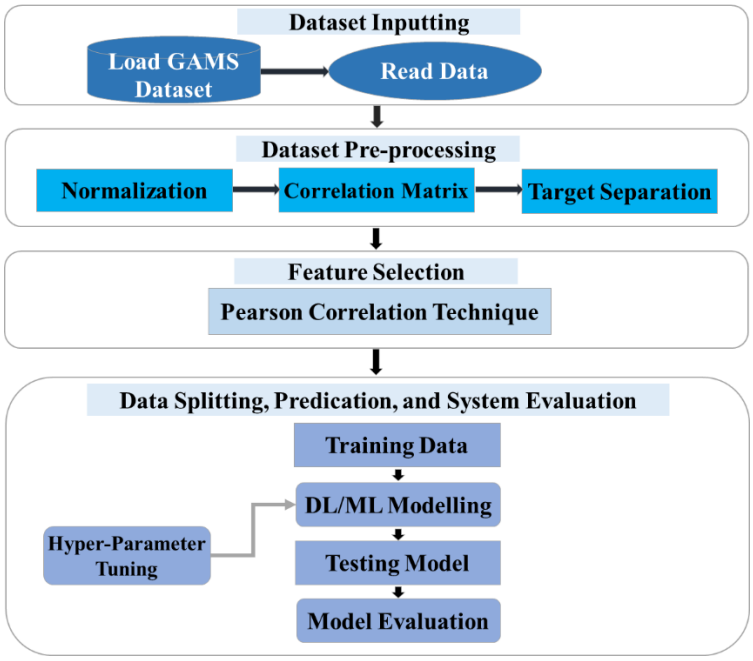


Figure 1. Diagram of IAQ prediction system

4.1 Data sources and pre-processing

The presented study was based on a detailed air quality dataset from GAMS sensors obtained from GAMS Environmental Monitoring [30]. This dataset provides high-resolution environmental data and pollutant levels, enabling the complete and in-depth analysis of air quality patterns and trends. A sample of the GAMS dataset is illustrated in Figure 2.

	Ts	CO2	Humidity	PM10	PM25	Temperature	VOC
0	2016-11-21 00:47:03	708.0	72.09	10.2	9.0	20.83	0.062
1	2016-11-21 00:48:03	694.0	70.95	10.9	10.1	21.01	0.062
2	2016-11-21 00:49:03	693.0	69.12	10.2	9.9	21.20	0.062
3	2016-11-21 00:50:03	692.0	68.83	9.6	9.6	21.37	0.062
4	2016-11-21 00:51:03	690.0	68.60	9.4	8.4	21.49	0.062

Figure 2. Sample of the GAMS dataset

The general information of this dataset (including the overall structure of the dataset, data collection frequency, and measured parameters) is depicted in Table 1. With this comprehensive dataset, we could delve into the complexities of air quality monitoring and its implications for the environment's health. The broader analysis of GAMS data may form an enlightening contribution to the capability of air quality monitoring technologies to protect public health.

Table 1: Detail of the indoor GAMS dataset

Terms	Details
Pollutants	CO2, PM10, PM2.5, VOC
Meteorology	Temperature @Humidity
Period	November 2016 – March 2017
No. of Samples	135099
Interval Reading	1.36 Minutes

The indoor dataset retrieved from the IAQ monitoring sensor GAMS was utilized. The documented factors are considered highly relevant for indoor environments. The dataset includes observations on indoor meteorological conditions such as CO_2, PM10, PM2.5, and VOCs recorded at intervals of several minutes, along with variables such as temperature and humidity. Spanning approximately five months, from November 2016 to March 2017, the dataset contains slightly more than 135,000 measurements. The dataset is divided into three subsets (training, validation, and testing) with a ratio of 60-20-20, as shown in Figure 3.

Additionally, before training the models, all data were normalized, with each pollutant scaled to a range between [0, 1], as provided in the following formula:

$$Norm(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where x denotes the count of pollutants, x_{min} denotes the smallest amount, and x_{max} denotes the largest amount in the dataset.

The imputation strategy was employed to deal with the missing values, which were replaced by the average of the available values in the relevant column. This imputation approach implies that missing data points occur at random and does not add major biases to the dataset, as provided in the following formula:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad (2)$$

Where $\bar{x}$ denotes the average of the missing values, x_i denotes the individual non-missing values, and N denotes the count of available data points. This approach was selected due to its simplicity, computational efficiency, and suitability for datasets with relatively sparse missing values.

Figure 4 presents all the statistical information (Standard Deviation, minimum and maximum values, median, and mean percentages) to provide a clear and helpful picture of the utilized dataset and to understand the fundamental characteristics of each numerical variable.

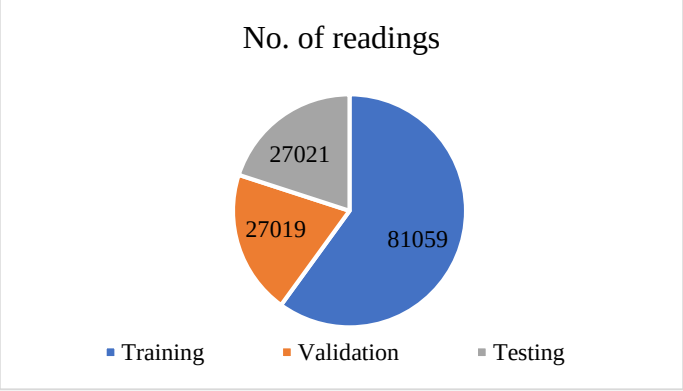


Figure 3. The number of readings in each subset after splitting

	CO2	Humidity	PM10	PM25	Temperature	VOC
Count	135099.000000	135099.000000	135099.000000	135099.000000	135099.000000	135099.000000
Mean	688.833011	37.879496	17.553535	15.801651	22.939613	0.121050
Std	385.845573	5.284216	12.603744	11.709474	2.051068	0.089947
Min	369.000000	21.970000	0.500000	0.500000	17.710000	0.062000
25%	429.000000	34.490000	8.400000	7.300000	21.410000	0.064000
50%	483.000000	37.640000	14.000000	12.300000	22.860000	0.076000
75%	852.000000	41.290000	23.400000	21.000000	24.650000	0.149000
Max	2626.000000	72.090000	142.600000	85.200000	27.960000	2.000000

Figure 4. Descriptive statistics for the indoor GAMS dataset

Figure 5 depicts a detailed description of the density distribution of several different indicators that are contained inside the GAMS dataset. These curves identify the outliers or anomalies in the datasets and give an idea of the behavior of each of the indicators. This view allows for a comparative examination of multiple GAMS columns, enabling us to identify widespread patterns or anomalies within the distributions. Investigations of this type are necessary because they allow the dependability and coherence of the recorded data to be reviewed quickly. It also helps map out the required information linked to the inquiries that may be proposed in the future.

4.2 Feature selection

The correlation matrix is a grid with one memory unit for each pair (i, j) of indicators. It determines how two data sets or two random variables are related. There are different correlation coefficients in correlation statistics, but the Pearson Correlation Coefficient is the most popular. This coefficient measures the linear correlation between two variables. Figure 6 shows the correlation matrix among GAMS features. There are decimal numbers in the Pearson coefficient range, from -1 to +1. A positive association exists between two variables when variable "A" raises simultaneously with variable "B". When variables A and B decrease simultaneously, this is a negative correlation. However, a correlation value of 0 means that A and B do not have a linear relationship. The two variables are strongly related if the correlation value is greater than 0. Correlation Matrix encompasses several characteristics:

- Symmetry: The correlation matrix is essentially symmetric, indicating that the correlation between variable X and variable Y is identical to that between variable Y and variable X.
- Main Diagonal Elements: The elements located along the principal diagonal of the correlation matrix are always equal to one. This occurs because the autocorrelation of any variable is always equal to one.
- Range of Values: All numerical values within the correlation matrix range from negative one to positive one (-1 to +1). A value of one indicates a perfect positive correlation, zero signifies no correlation, and a negative one denotes a perfect negative correlation.
- Non-Negative Eigenvalues: Every eigenvalue of a correlation matrix must be non-negative. This guarantees that the matrix is positive and semi-definite, a prerequisite for qualifying it as a legitimate correlation matrix.

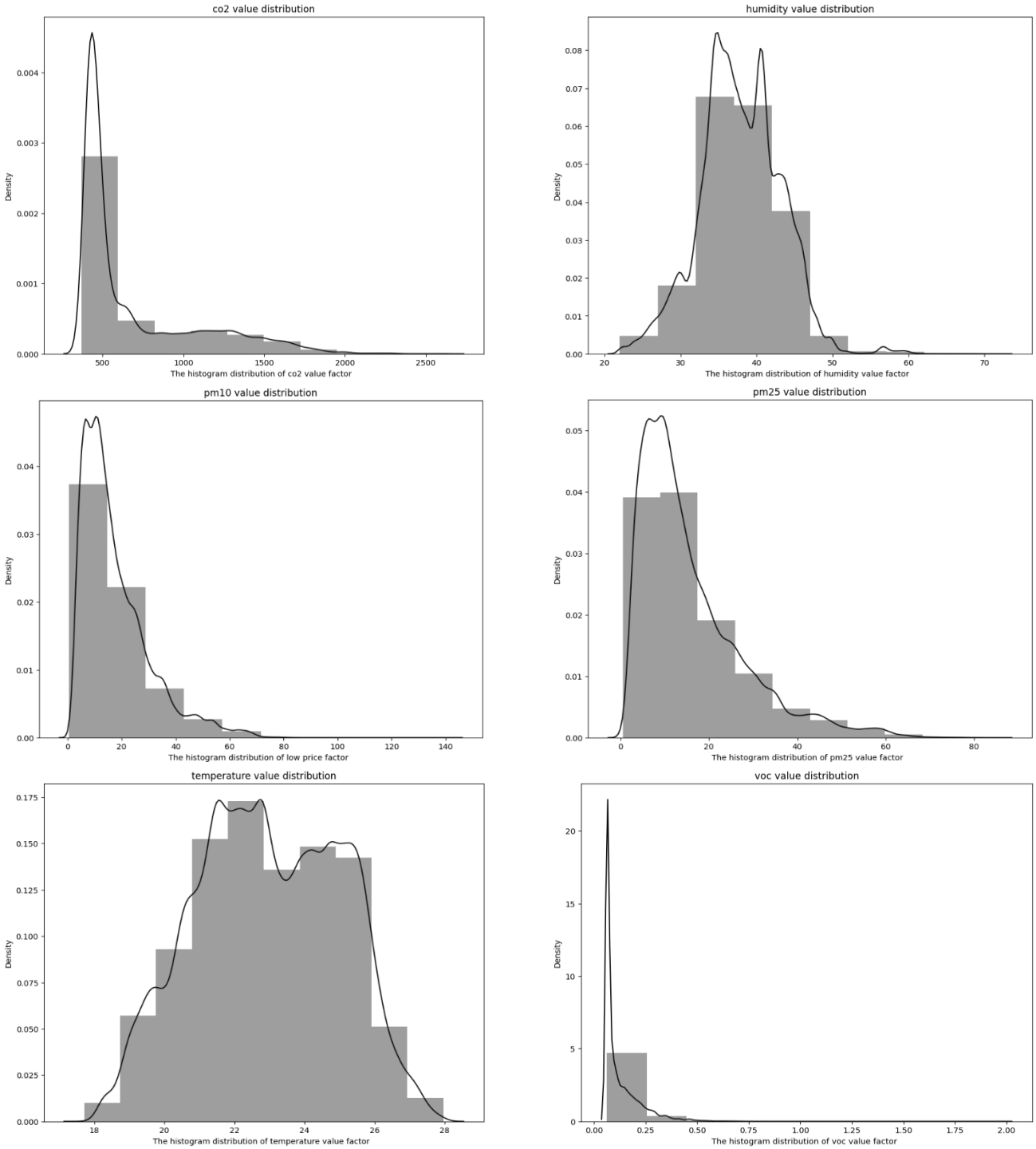


Figure 5. GAMS indicators' distribution

	CO2	Humidity	Pm10	Pm2.5	Temperature	VOC
CO2	1.000000	0.106859	-0.226913	-0.228081	0.713173	0.398333
Humidity	0.106859	1.000000	-0.022478	-0.011838	-0.046315	-0.099963
Pm10	-0.226913	-0.022478	1.000000	0.994504	-0.237087	0.020330
Pm2.5	-0.228081	-0.011838	0.994504	1.000000	-0.233693	0.011977
Temperature	0.713173	-0.046315	-0.237087	-0.233693	1.000000	0.214963
VOC	0.398333	-0.099963	0.020330	0.011977	0.214963	1.000000

Figure 6. Correlation matrix between the features of the GAMS dataset

4.3 Model selection and justification

In this work, some of the common types of ML models are used to predict indoor air CO_2 encompassing:

- DTR: It evaluates input features and generates interpretable rules to predict CO_2 levels in a closed environment.
- RF: It is an ensemble learning method designed to improve the accuracy of CO_2 estimates by aggregating many decision trees. To this end, it is capable of handling the complexity of the components that define the IAQ.
- K-NN: It classifies the objects depending on the proximity of those objects. It is often used in the classification of CO_2 or prediction of the indoor location due to the simplicity and efficacy of the algorithm.
- SVR: It efficiently classifies data by determining the level that produces the maximum possible distance between data points. This particular characteristic is very useful in the estimates of CO_2 generated from historical data, providing an accurate classification and forecast.
- GBR: It is an ensemble learning methodology that combines numerous decision trees to make the CO_2 forecasts more accurate over time.

Furthermore, several prominent DL methodologies are used to predict indoor air CO2 concentrations encompassing:

- DAQFF: It contains an integrated set of elements that work together to achieve specific goals.
- LSTM: It is a type of recurrent neural network characterized by its capability to use its internal state memory for superior processing sequences. The middle LSTM layer is a forgotten gate, and this forgotten portal is used to make decisions. By predicting the CO_2 data of the internal structure that should be saved, he forgets the data that should be saved. The middle layer will take the data from the input layer, while the output layer displays the result.
- CNN: It is a modern form of Ancient neural networks in which each layer consists of a neuron that corresponds to neurons in the next layer, and so on. The specialty of this model is a multilayer convolution. CNN has applied filters, and the filter size automatically recognizes its task.
- Bi-LSTM: It is considered more general for privacy when it has driving information because its output depends on all the above and the following. The structure of Bi-LSTM consists of a dual forward layer and a white layer, while the LSTM modules will consist of understanding future and past information.
- GRU: It is the simplest alternative to LSTM which consists of two gates: the "update gate," which consists of data entry gates and forget gates, and the "reset gate." The GRU architecture does not contain an additional memory cell to hold information, so it can only control the information contained within the unit.
- RF-LSTM: This hybrid model of RF-LSTM is presented in this work to predict CO_2 air quality. RF is employed to choose the data feature set, Subsequently, LSTM is employed, a highly effective model to forecast CO_2 levels and enhance the model's predictions.
- CNN-LSTM: Another hybrid model of the CNN-LSTM is also implemented in this work to forecast CO2 levels. Here, the CNN layer is adopted to extract features of CO_2 from the IAQ monitoring dataset generated from GAMS. After extraction, the captured features are transformed into a 1D- matrix to feed into the LSTM layer for conducting a time series feature analysis.
- DNN: It is similar in structure to ANN which creates more than one hidden layer between the input and output layers, so, it allows the developer to gain good, more sustained, high-quality accuracy results.

These models perform better with many dependencies on parameters and hyperparameters. Further model performance improvements are made based on the results from hyperparameter tuning. Table 2 and Table 3 elaborately overview the different parameters and hyperparameters used in any ML/DL model. These elements must be properly tuned and selected to enhance model accuracy and efficiency.

Table 2: The parameters and values of the various implemented ML models

Model	Parameter	Value(s)
DT	Max Depth	8
RF	Max Depth	8
	Number of Trees	4000
K-NN	Max Depth	8
	N-neighbors	10
SVR	Max Depth	8
	Cost factor(C)	25
	Cache_size	200
	Max_iter	-1
GBR	Max Depth	8
	Number of Trees	100
	Learning Rate	0.01

Table 3: The parameters and values of the various implemented DL models

Model	No. of Layers	No. of Units	Loss Function	Optimizer	LR.	No. of epoch
(DAQFF)	7	64, 128, 256, 32, 16, 8, 1	MAP	Adam	0.001	15
LSTM	6	64, 128, 256, 128, 64, 32	MSE	Adam	0.001	15
CNN	5	32, 64, 128	MAP	Adam	0.001	15
GRU	5	64, 128, 256, 128, 64	MSE	Adam	0.001	15
Bi- LSTM	7	64, 128, 256, 512, 256, 128, 64	MSE	Adam	0.001	15
RF-LSTM	7	64, 128, 256, 512, 256, 128, 64	MSE	Adam	0.001	15
RF-LSTM	7	64, 128, 256, 512, 256, 128, 64	MSE	Adam	0.001	15
DNN	6	64, 128, 256, 512, 256, 128	MAP	Adam	0.001	15

In summary, this study utilized a comprehensive dataset and a suite of ML and DL models to predict IAQ, with CO2 concentrations as the primary focus. The dataset underwent extensive preprocessing, including normalization and imputation of missing values, to ensure robust and accurate model performance. The selected models are tailored to leverage temporal dependencies, correlations, and non-linear relationships in the data. In the subsequent section, these models are evaluated, and their predictive performance is determined using MAE, RMSE, Median AE, and R^2 metrics.

5. Results and discussion

This section investigates the performance of the implemented ML and DL models in predicting CO2 levels for IAQ using the indoor dataset obtained from GAMS.

5.1 Evaluation metrics

In this work, four evaluation metrics are employed (RMSE, R^2, MAE, and Median AE) to compare the performance of different implemented models and find their relative effectiveness.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (5)$$

$$Median\ AE = median(|O_i - P_i|) \quad (6)$$

Where O_i denotes the observed value for the $i-th$ observation, P_i denotes the predicted value for the $i-th$ observation, $\bar{O}$ denotes the mean of the observed values, n denotes the total count of observations, the absolute errors $|O_i - P_i|$ are computed for each observation, and the median of these absolute errors are taken to find the Median AE.

5.2 Comparative analysis

Among the implemented ML models, the best performance was exhibited by GBR since, the boosting model becoming effective in dealing with complicated nonlinear relationships and large dispersions. GBR was able to catch a more granular pattern level between inputs and outputs, it also gave better predictions on IAQ prediction datasets that generally include many interdependent variables. The results were achieved by the ML models depicted in Table 4. The predicted result was compared with the actual outcomes, and the error performance metric was exploited to assess each model, as depicted in Figure 7.

Table 4: The results obtained using ML Models

Models	MAE	RMSE	Median AE	R^2
DTR	0.00217	0.00358	0.00113	0.099962
RF	0.17479	0.24796	0.10434	0.081029
K-NNs	0.00525	0.00950	0.00177	0.099734
SVR	0.00923	0.01185	0.00668	0.099587
GBR	0.00200	0.00353	0.00086	0.099963

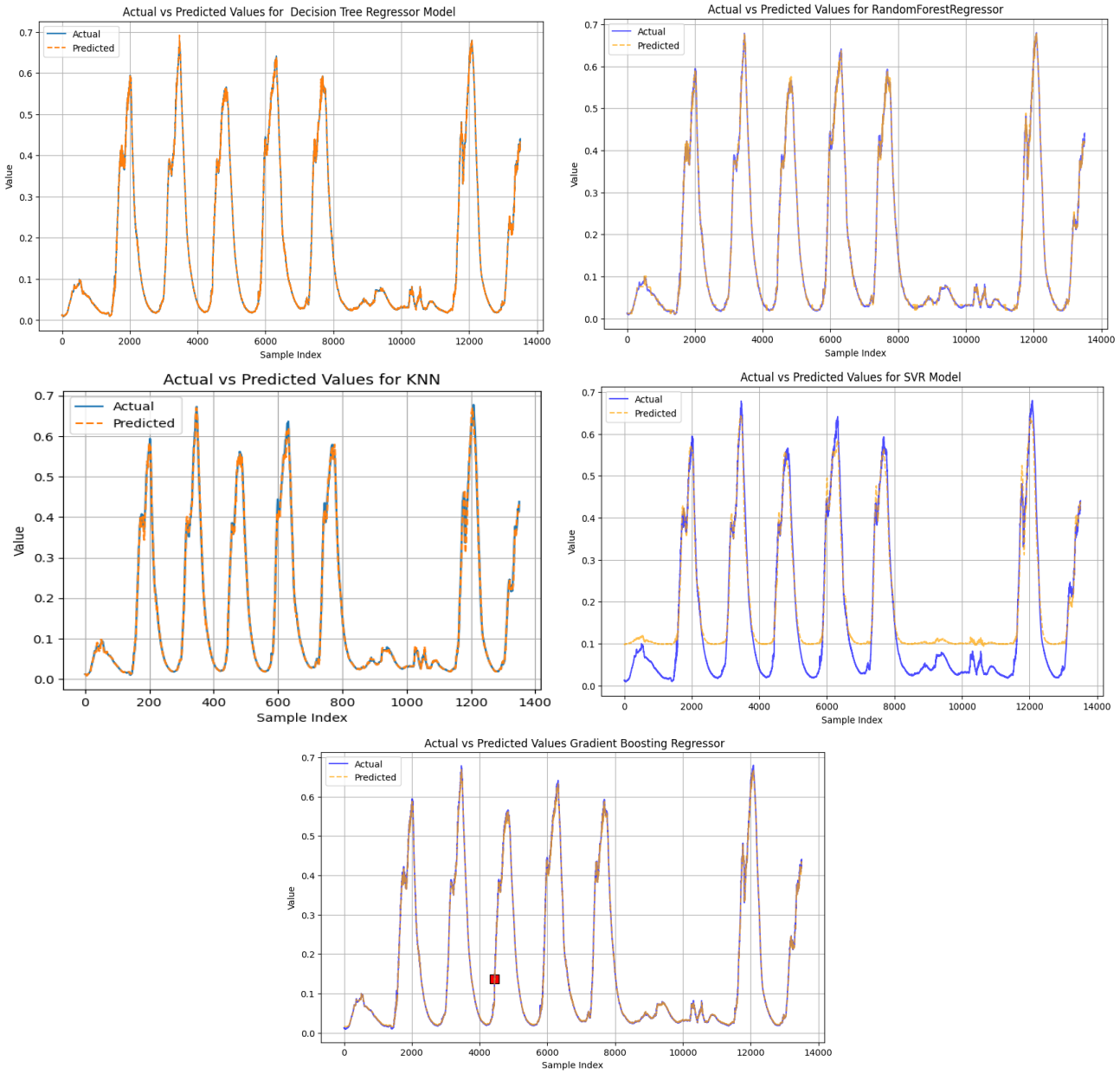
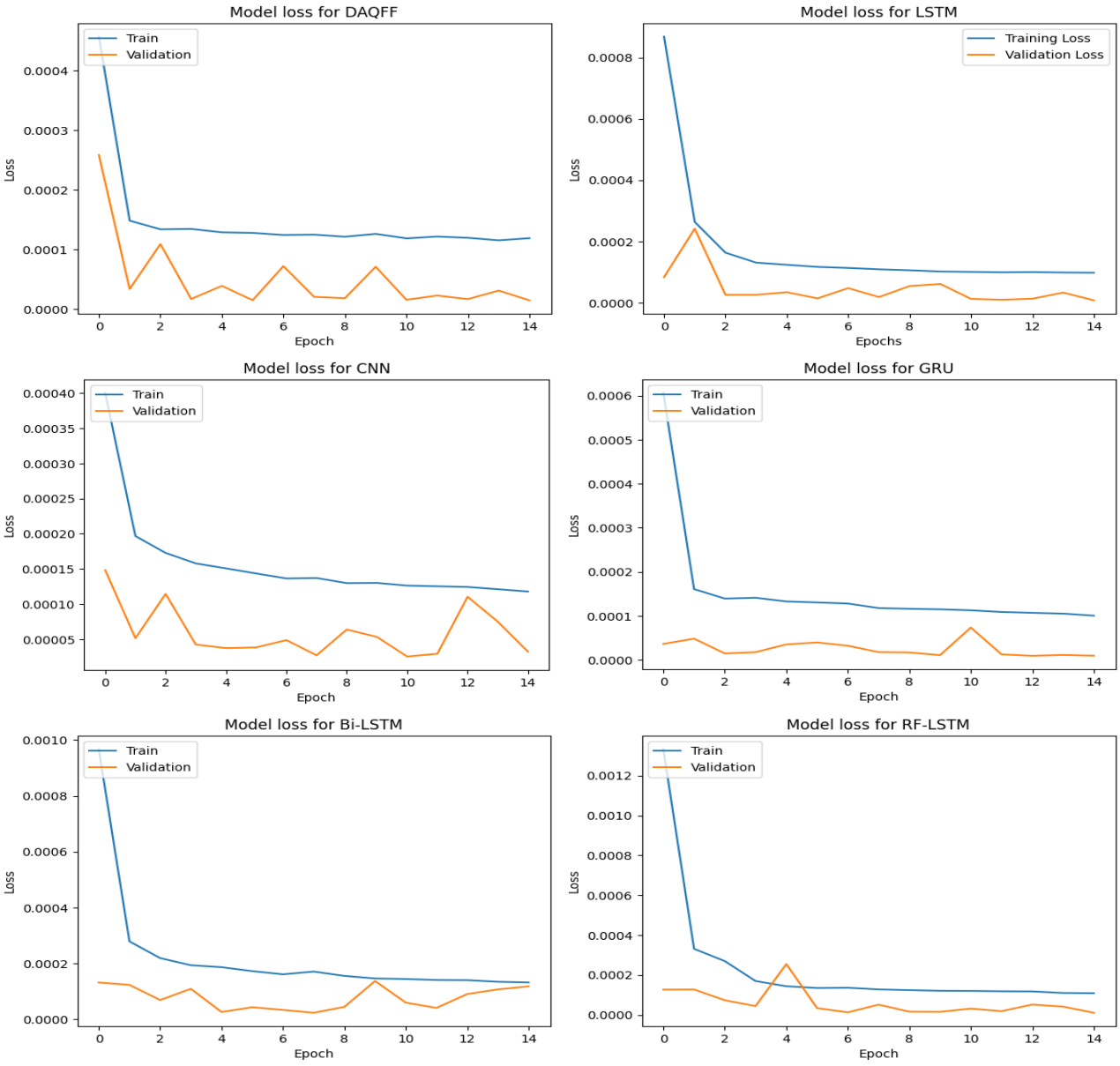


Figure 7. Actual versus predicted values for ML models

The LSTM model outperformed the other DL models due to its ability to retain information from previous periods without losing considerable details, making it an effective model for data prediction influenced by events from past times, such as temperature changes. The implemented DL models are tested on the same dataset, and the results are highlighted in Table 5 and Figure 8.

Table 5: The results obtained using DL Models

Models	MAE	RMSE	Median AE	R^2
DAQFF	0.002365	0.003715	0.001486	0.099956
LSTM	0.001661	0.002744	0.000961	0.099976
CNN	0.003189	0.005677	0.001286	0.099897
GRU	0.001829	0.003026	0.001033	0.099971
Bi-LSTM	0.010462	0.011502	0.009988	0.099576
RF-LSTM	0.002014	0.003067	0.001373	0.099970
CNN-LSTM	0.004360	0.006880	0.001797	0.099848
DNN	0.026897	0.034500	0.022202	0.096189



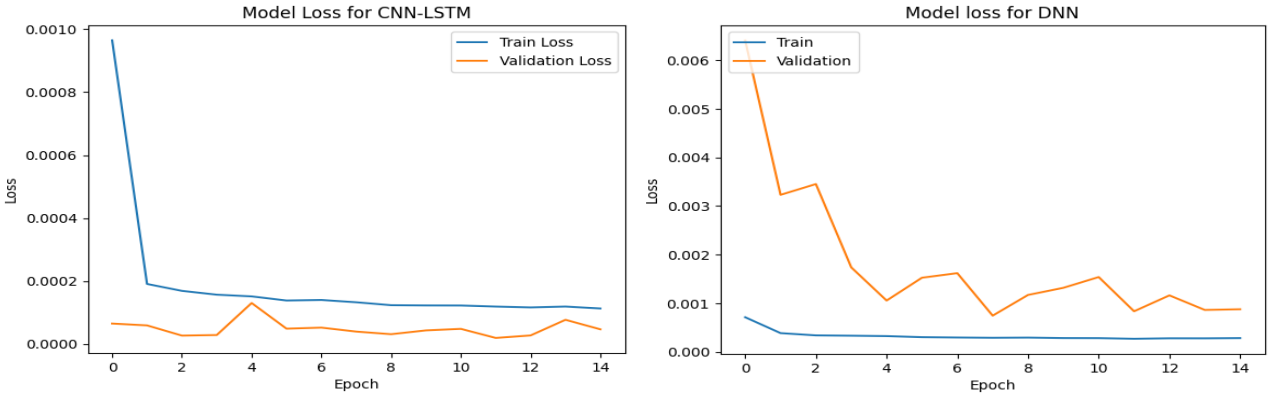


Figure 8. DL models' loss

Among the assessed models, the LSTM model exhibited more outstanding performance. It attained the minimal MAE of (0.001661), RMSE of (0.002744), and Median AE of (0.000961). The LSTM model demonstrates a substantial R^2 value of 0.999976, demonstrating that it is exceptionally proficient in detecting temporal relationships and intricate patterns in IAQ data. In contrast, the DNN model had the highest error rates with maximal MAE of (0.026897), RMSE of (0.034500), and R^2 of 0.096189, which is a pretty low result that would mean the DNN model did not express the IAQ data in its full complex details.

The results reveal that the LSTM and GRU models outperform others in predicting IAQ parameters, owing to their ability to capture temporal dependencies. These findings highlight the importance of model selection for time-series data, as discussed in the following section, which explores the implications, limitations, and potential applications of these results.

5.3 Implications of findings

Compared to other models, LSTM and GRU performed better in predicting the parameters of IAQ. Both models have inherent memory, which relates to their ability to handle dynamic data intrinsically in time series analysis, which is required to forecast the air quality inside buildings. The performance of these models confirms that they have the potential to be applied in operational real-time IAQ monitoring systems. This is because these models are becoming more accurate by now. However, the low performance of the DNN model can be considered as evidence that it is not the best suited for tasks involving IAQ prediction, especially in the case of time series.

This is particularly true in tasks where time series data must be handled. Therefore, careful consideration is needed when selecting models for specific data types and conditions. The results suggest that expanding the dataset's scope to include greater diversity in the types of sources could enhance model performance and generalization potential. A good example could be data from various residential and commercial buildings in different geographical areas. For further improvement, data could be collected during various seasons. Additionally, integrating external meteorological data on temperature and humidity could contribute to providing more accurate and detailed forecasts regarding indoor air quality, as this data offers insights into external weather conditions.

5.4 Limitations and considerations

While the findings of this study are encouraging, it is crucial to recognize the limitations that may impact the generalization and applicability of the results. A potential limitation is that the dataset refers to IAQ from environments that may differ in IAQ variations from other building types or regions. As a result, such models may be less successful when applied in contexts and situations outside those tested with this data. However, this limitation can be overcome by referring to other sources in further research to test the robustness of models and how well they perform under diverse conditions.

The most probable causes of this reduced generalizability are manifold. First, conditions from interior sites dominated the dataset on which training and testing were based. This naturally brings about bias because such models have been trained on data emanating from only one IAQ scenario. They can be less accurate if those models apply in diverse environments with different structures, ventilation systems, or occupation rates. Thus, it may be less applicable to IAQ predictions in various settings.

Another consideration is model selection bias. Whereas the GBR and LSTM models perhaps showed more outstanding performance in this study, their performance could be strictly optimal for the kinds of patterns represented in this dataset; poor performance might otherwise have been observed for patterns more complex or otherwise unexpected in other data, given the non-representative nature of not all relevant variables or higher-order interactions in the data at hand. Further, the variability of the dataset may not account for other extraneous factors that may be influencing the variation of IAQ, such as seasonal changes, long-term environmental changes, and differences in sources of pollutants. Therefore, future research should be directed at extending the dataset to run a wide range of data over different building types, geographical regions, and seasonal variations to make the model generalize better.

Overfitting is also risky, particularly for complex models like LSTM and CNN-LSTM, which try to learn minute patterns in time-series data. Though models might work fine during training, they usually fail to generalize on unseen data. If overfitting occurs, it drastically reduces the predictive power of the model when it is deployed. Such limitations require further studies with model refinement to make them more robust and generally applicable, thereby providing deeper insights into IAQ predictions.

In light of these findings, this study emphasizes the critical role of advanced AI techniques in enhancing IAQ prediction. The conclusion summarizes the key contributions of this work, along with suggestions for future research to further improve model generalizability and applicability.

6. Conclusions

In this work, various ML and DL approaches have been employed to predict indoor environmental CO_2 concentration. Various techniques adopted in this regard included the ML models comprising DTR, Random Forest, K-NN, SVR, and GBR, while DL models included DAQFF, LSTM, CNN, GRU, Bi-LSTM, RF-LSTM, and CNN-LSTM. The dataset obtained from GAMS includes variables such as CO_2, humidity, PM10, PM2.5, temperature, and VOC. The models obtained this way were compared afterward using a fivefold cross-validation, following the custom of using 80% of data for training and 20% for testing.

The outcome revealed GBR as the best method in predicting CO2 concentration, with low error metrics of MAE at 0.00200, RMSE at 0.00353, and Median AE at 0.00086, with an R^2 value close to 1. Similarly, the LSTM model did a great job, with an MAE of 0.001661, confirming its strength in time-series prediction. These results provide valuable insights into future studies on indoor CO_2 prediction given occupant health.

While the obtained predictions were promising, several directions for further research are open to remedy the identified limitations and extend the generalizability of the models. Furthermore, the system would be more robust when more diverse data, such as residential and commercial buildings across different geographic locations and seasons, are included in the dataset. In this respect, external weather data, including outdoor temperature and humidity, would further put into perspective those factors affecting indoor air quality.

The upcoming studies could also use more advanced models compared to what is already used in this study. For ML, attention could be given to the XGBoost and CatBoost models, which have proved to be considerably efficient in many applications and could enhance predictive accuracy. For DL, transformer models, or attention-based models, showing state-of-the-art performance in capturing intricate time-series relationships may be beneficial extensions. Lastly, more advanced

pre-processing techniques like sophisticated imputation of missing data and advanced smoothing of data could help significantly towards model performance improvement. Another direction of potential future work could be the application of ensemble learning techniques, including model stacking and model blending, where the strengths of different models are combined. Stacking traditional Ml models with DL approaches could explore further accuracy enhancement, again using the unique strengths of each method.

In conclusion, the results derived from the present study provide a strong foundation for anyone interested in predicting the indoor concentration of CO_2. Further research in these suggested areas can enhance the accuracy and applicability of such models for real-world use.

References

[1] L. Salvaraji, S. B. Shamsudin, R. Avoi, S. Saupin, L. K. Sai, S. B. Asan, H. R. B. Toha, and M. S. Jeffree, "Ecological study of sick building syndrome among healthcare workers at Johor primary care facilities", International Journal of Environmental Research and Public Health, vol. 19, no. 24, pp. 1-12, 2022. https://doi.org/10.3390/ijerph192417099.

[2] N. R. Kapoor, A. Kumar, A. Kumar, A. Kumar, M. A. Mohammed, K. Kumar, S. Kadry, and S. Lim, "Machine learning-based CO_2 prediction for office room: A pilot study", Wireless Communications and Mobile Computing, vol. 2022, pp. 1-16, 2022. https://doi.org/10.1155/2022/9404807.

[3] S. N. A. M. Noor and H. H. Ding, "Indoor environment quality (IEQ): Temperature and indoor air quality (IAQ) factors toward occupants satisfaction", IOP Conference Series: Materials Science and Engineering, vol. 864, no. 1, pp. 1-6, 2020. https://doi.org/10.1088/1757-899X/864/1/012012.

[4] İ. Arikan, Ö. F. Tekin, and O. Erbas, "Relationship between sick building syndrome and indoor air quality among hospital staff", La Medicina del Lavoro, vol. 109, no. 6, pp. 435-443, 2018. https://doi.org/10.23749/mdl.v110i6.7628.

[5] S. M. Yussuf, G. Dahir, A. M. Salad, M. Hayir T. M, S. A. Hassan, and A. Gele, "Sick building syndrome and its associated factors among adult people living in Hodan district, Mogadishu, Somalia", Frontiers in Built Environment, vol. 9, pp. 1-8, 2023. https://doi.org/10.3389/fbuil.2023.1218659.

[6] M. Mannan and S. G. Al-Ghamdi, "Indoor air quality in buildings: A comprehensive review on the factors influencing air pollution in residential and commercial structures", International Journal of Environmental Research and Public Health, vol. 18, no. 6, pp. 1-24, 2021. https://doi.org/10.3390/ijerph18063276.

[7] S. S. Korsavi, A. Montazami, and D. Mumovic, "Indoor air quality (IAQ) in naturally-ventilated primary schools in the UK: Occupant-related factors", Building and Environment, vol. 180, 2020. https://doi.org/10.1016/j.buildenv.2020.106992.

[8] W. Wei, O. Ramalho, and C. Mandin, "Indoor air quality requirements in green building certifications", Building and Environment, vol. 92, pp. 10-19, 2015. https://doi.org/10.1016/j.buildenv.2015.03.035.

[9] N. Mahanta and R. Talukdar, "Forecasting of domestic electricity consumption in Assam, India", International Journal of Energy Economics and Policy, vol. 13, no. 5, pp. 229-235, 2023. doi: 10.32479/ijeep.14509.

[10] V. Van Tran, D. Park, and Y. C. Lee, "Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality", International Journal of Environmental Research and Public Health, vol. 17, no. 8, 2020. doi: 10.3390/ijerph17082927.

[11] N. Fernandes and J. Gonçalves, "Multivariate and multi-output indoor air quality prediction using bidirectional LSTM," in 2023 11th International Symposium on Digital Forensics and Security (ISDFS), Chattanooga, TN, USA, 2023, pp. 1-6. doi: 10.1109/ISDFS58141.2023.10131695.

[12] S. Abirami and P. Chitra, "Regional air quality forecasting using spatiotemporal deep learning", Journal of Cleaner Production, vol. 283, 2021. doi: 10.1016/j.jclepro.2020.125341.

[13] H. Nan, "Apply RF-LSTM to predicting future share price," 2023 International Conference on Digital Economy and Management Science (CDEMS 2023), vol. 170, pp. 1-4, 2023. https://doi.org/10.1051/shsconf/202317002012.

[14] J. Lee, J. Woo, A. R. Kang, Y.-S. Jeong, W. Jung, M. Lee, and S. H. Kim, "Comparative analysis on machine learning and deep learning to predict post-induction hypotension", Sensors (Switzerland), vol. 20, no. 16, pp. 1-21, 2020. doi: 10.3390/s20163891.

[15] J. Lozano, J. I. Suárez, P. Arroyo, J. M. Ordiales, and F. Álvarez, "Wireless sensor network for indoor air quality monitoring", Chemical Engineering Transactions, vol. 30, pp. 54-59, 2012. doi: 10.3303/CET1230054.

[16] H. Q. Flayyih, J. Waleed, and A. M. Ibrahim, "IoT-based AI methods for indoor air quality monitoring

systems: A systematic review", International Journal of Computing and Digital Systems, vol. 16, no. 1, pp. 813-826, 2024. doi: 10.12785/ijcds/160159.

[17] Y. Li and Y. Sun, "Modeling and predicting city-level CO2 emissions using open access data and machine learning," Environmental Science and Pollution Research, vol. 28, no. 15, pp. 19260-19271, 2021. doi: 10.1007/s11356-020-12294-7.

[18] S. Taheri and A. Razban, "Learning-based CO2 concentration prediction: Application to indoor air quality control using demand-controlled ventilation", Building and Environment, vol. 205, 2021. doi: 10.1016/j.buildenv.2021.108164.

[19] M. Marzouk and M. Atef, "Assessment of Indoor Air Quality in Academic Buildings Using IoT and Deep Learning", Sustainability, vol. 14, no. 12, 2022, Art. no. 7015. doi: 10.3390/su14127015.

[20] Y. Zhu, S. A. Al-Ahmed, M. Z. Shakir, and J. I. Olszewska, "LSTM-Based IoT-Enabled CO2 Steady-State Forecasting for Indoor Air Quality Monitoring," Electronics, vol. 12, no. 1, pp. 1–12, 2023. doi: 10.3390/electronics12010107.

[21] Z. Dai, Y. Yuan, X. Zhu, and L. Zhao, "A Method for Predicting Indoor CO2 Concentration in University Classrooms: An RF-TPE-LSTM Approach", Applied Sciences, vol. 14, no. 6188, pp. 1-12, 2024. doi: 10.3390/app14146188.

[22] J. Kim, J. I. Bang, A. Choi, H. J. Moon, and M. Sung, "Estimation of Occupancy Using IoT Sensors and a Carbon Dioxide-Based Machine Learning Model with Ventilation System and Differential Pressure Data," Sensors, vol. 23, no. 585, pp. 1-18, 2023. doi: 10.3390/s23020585.

[23] A. E. Abdelkareem, "Performance Analysis of Deep Learning based Signal Constellation Identification Algorithms for Underwater Acoustic Communications", Diyala Journal of Engineering Sciences, vol. 17, no. 3, 2024. doi: 10.24237/djes.2024.17301.

[24] J. Graffelman and J. de Leeuw, "Improved Approximation and Visualization of the Correlation Matrix," The American Statistician, vol. 77, no. 4, pp. 432-442, 2023. doi: 10.1080/00031305.2023.2186952.

[25] A. J. Abdullah, T. M. Hasan, and J. Waleed, "An Expanded Vision of Breast Cancer Diagnosis Approaches Based on Machine Learning Techniques," in 2019 International Engineering Conference (IEC), Erbil, Iraq, pp. 177-181, 2019. doi: 10.1109/IEC47844.2019.8950530.

[26] K. Jain and S. Kaushal, "A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis," in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 483-487, 2018. doi: 10.1109/ICRITO.2018.8748793.

[27] R. J. Kolaib and J. Waleed, "Crime Activity Detection in Surveillance Videos Based on Developed Deep Learning Approach," Diyala Journal of Engineering Sciences, vol. 17, no. 3, pp. 98-114, 2024. doi: 10.24237/djes.2024.17307.

[28] A. Al-Saegh, A. Daood, and M. H. Ismail, “Dual Optimization of Deep CNN for Motor Imagery EEG Tasks Classification”, Diyala Journal of Engineering Sciences, vol. 17, no. 4, pp. 75-91, 2024. doi: 10.24237/djes.2024.17405.

[29] J. Waleed, S. Albawi, H. Q. Flayyih, and A. Alkhayyat, "An Effective and Accurate CNN Model for Detecting Tomato Leaves Diseases," in 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA), Najaf, Iraq, 2021, pp. 33–37. doi: 10.1109/IICETA51758.2021.9717816.

[30] J. Liu, "GAMS indoor air quality dataset," GitHub Repository, 2013. Available: https://github.com/twairball/gams-dataset.