Gas Diffusion Simulation Based on Ensemble Approach

K. M. Gwak and Young J. Rho

***Abstract*—The 4th [1] industrial revolution is promoting manufacturing industry to be vitalized again. The manufacturing industry requires many industrial materials. Among them, different gases are also used in many fields. While they are useful, industrial gases can be also hazardous at the same time. In order to control those bad features of gases, their dynamic characteristics are required to be understood. In this paper we tried to understand the characteristics by applying several machine learning methods such as MLP, DLP and LSTM. Two ensemble methods are applied to compensate the lack of raw data. Simulation outputs are compared each other to know which method is proper for this case.**

***Index Terms*—Data mining, neural networks, computer simulation, pattern recognition.**

I. Introduction

Recently, the 4th industrial revolution (the 4th IR in short) is a big wave to manufacturing industry. It forces the industry sector to be changed to a new paradigm of smart manufacturing. Various materials are not only involved in products but also applied to manufacturing processes. Hazardous gases are usually inevitable for the engineering processes. The gases must be under complete control because they can cause severe industrial accidents.

If gas dispersion can be simulated, industrial accidents by hazardous gases can be coped with. CFD (Computational Fluid Dynamics) [1] researches have been and being. Applying machine learning methods to the simulation of gas dispersion is a new approach to describe dynamic characteristics of gases if there are empirical data. However, it is not easy to find the factories that have enough data for machine learning.

In this research, ensemble method which is one of machine learning methods is applied to generate data for machine learning to simulate gas dispersion. The second part explains related researches and the third part describes the experiments, leaning models and algorithms that are used in this research. The fourth part explains the analysis of experimental data and the fifth part conclude.

II. Related Researches

Machine learning receives attention because of the 4th IR. It is inevitable for smart factories that generate data from their operations. Ensemble method is applied to these data to simulate gas dispersion.

Manuscript received October 29, 2019; revised May 20, 2020.

The authors are with the Department of Smart Factory and Computer Science, Korea Polytechnic University, Korea (e-mail: yrho@kpu.ac.kr, k010511@naver.com).

A. Multi-Layer Perceptron

Legacy perceptron has a XOR problem that arises from a legacy linear classifier. To solve this problem, linear discrimination of classification is derived by adding hidden layers between MLP perceptron layers [2]. This kind of MLPs is applied to bi-classification, multi-classification and value prediction. In this research, the applied MLP model has only one hidden layer to predict non-linear PPM (particles per million) values.

B. Deep Multi-Layer Perceptron

DLP is a kind of MLPs which has more than 2 hidden layers. It is possible that DLP models of more than 3 hidden layers show lower performance [3]. In this research, the DLP model of 2 hidden layers showed better performance than the MLP model of one hidden layer [4]. PPM values are predicted by the both models of MLP and DLP for ensemble.

C. Long-Short-Term Memory

LSTM is an upgraded form of RNN (Recurrent Neural Network), which is specialized in processing long sequential input. Each memory cell of hidden layers has an input gate, an output gate and a forget gate so that the memory cell brings its preceding memory and remove useless memory [5]. In this study, output values from the MLP and DLP are selected as input values for learning. Thus, its output becomes the final output.

D. Ensemble

Ensemble method is a model to combine multiple prediction models. Those component models can have a same algorithm or different ones. The models are trained with the same data set. The data set can be sampled in duplicate or not. In this study, voting and stacking approach as ensemble methods are applied [6].

E. Voting (Averaging)

Voting is to select one of predicted values from multiple models so that the final prediction value is determined [7]. In this study, MLP and DLP are used as component models, and their predicted values are combined into the final value.

F. Stacking

Stacking is a method to train with the data set from other machine learning algorithms [8]. The data sets from the MLP, DLP and the voting method are applied as input data set for the LSTM.

III. Experiments

The 3rd part describes about data collection, leaning models and methods, their outputs and combination method that are used in this study.

DOI: 10.7763/IJMO.2020.V10.765

A. Experimental Environment and Methods

CO^2 gas was used for the experiment to collect data. CO^2 is widely used in many industries and relatively safe and easy to handle. The level of its purity was 99.8%.

The place of experiment was an indoor room with enough windows for good ventilation. CO^2 gas in a safety tank was sprayed at the end of the table on the Fig. 1. Data collection point were set at the distances of 130, 160, 190 and 220cm from the spray point. The spraying strength was about 0.05~0.2 bar. Data were collected at the 4 locations. [9]

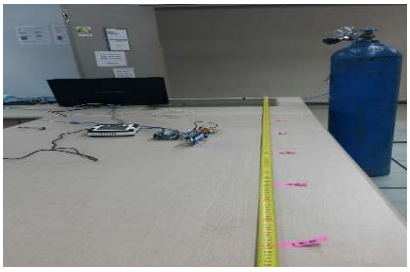

Fig. 1. Experimental environment.

The experiment has been executed in the ventilated indoor room as follows;

1) Start data collection,
2) Wait for 5seconds, and then start spraying gas,
3) Keep spraying gas for 15 seconds, and then stop it,
4) Keep collecting data until the monitoring sensors are stabilized,
5) Stop data collection and then ventilate the room for 5 minutes by opening the all the doors and windows of the experiment room.

Above process has been done 3 times to collect 222 data at each location of 130, 160, 190 and 220cm from the spraying point. The collected data are values between 0 and 5124. They are scaled into 0~1 because output data depend on activation function.

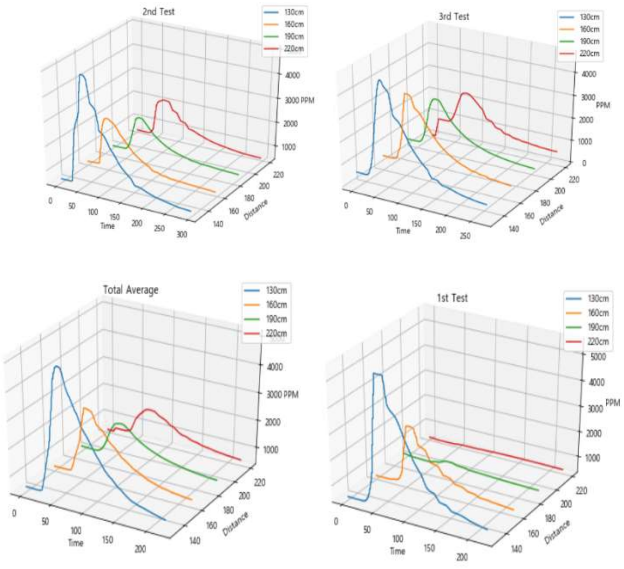

Fig. 2. Collected data from experiments.

Fig. 2 shows the data distributions of the three experiments and their averages for the last. Each axis of X, Y and Z represents for elapsed time, distance from the spraying point and monitored PPM values of CO_2 gas.

B. Model Definition and Input-output Data

1) Definition of MLP model

Fig. 3 shows the overall composition of the MLP layers that is used in this study. Data for elapsed time and distance values are used as input data. Four data sets of 221 data that were collected at the distances of 130, 160, 190, 220cm were applied for learning. The activation functions for learning are as follows;

- Input Layer -> Hidden Layer 1: Leaky ReLu
- Hidden Layer 1 -> Output Layer: Leaky ReLu
- Output Layer -> Output: Leaky ReLu

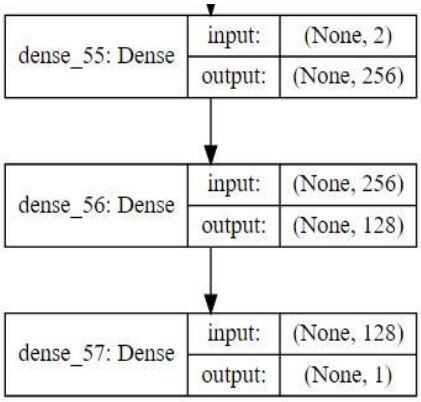

Fig. 3. Overall composition of the MLP model.

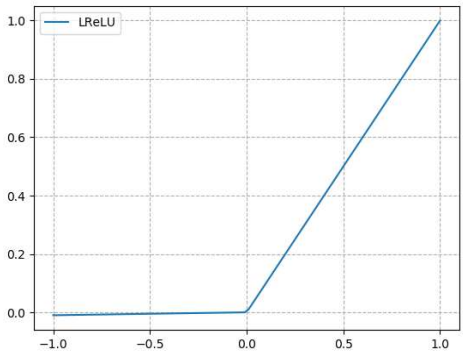

Fig. 4. Leaky ReLu function graph.

Fig. 4 is the graph for Leaky Relu Function. The definition of the function and its differential formula as follows; $F(x) = max\,(ax, x)$

$$f'(x) = \begin{cases} 1 & x > 0 \\ a & x \le 0 \end{cases}$$

Its output is values in PPM.

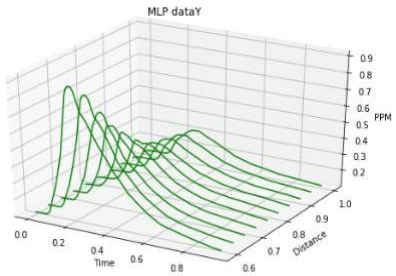

Fig. 5. 3D Graph of MLP output.

Fig. 5 visualizes the input data for the MLP model and the output data from the trained MLP in three dimensions. The distance depends on the divnum (to divide number) which is explained in the next section to explain the DLP model for this study. The distance values range from 130 to 220. To calculate real distance in centimeter (cm), the distance values on the graph have to be multiplied by 220.

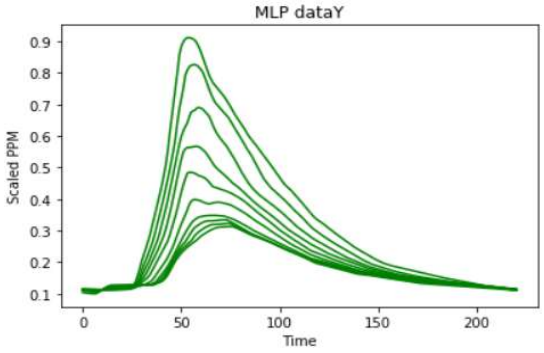

Fig. 6. 2D graph of MLP output.

Fig. 6 is a set of 2D graphs of Fig. 5 that shows elapsed time and ppm at each distance.

C. Definition of DLP Model

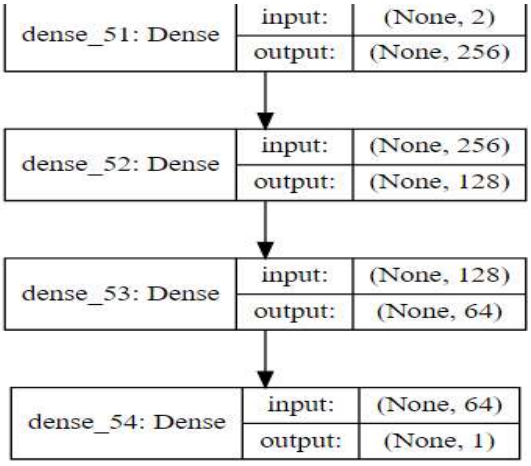

Fig. 7. Overall composition of the DLP model.

Fig. 7 shows the overall composition of the DLP model that is applied to this study. Its input data set is the same as that for the MLP model and the activation function is the same as well.

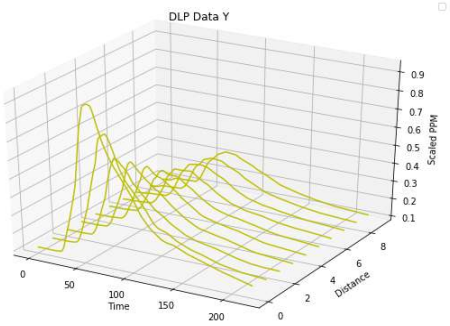

Fig. 8. 3D graph of DLP output.

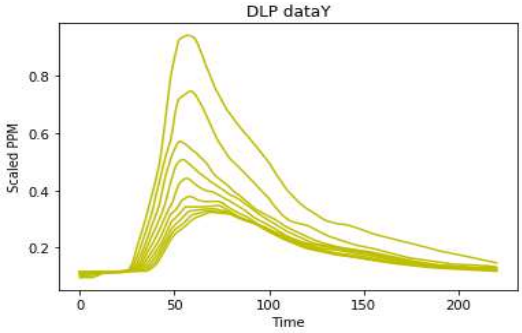

Fig. 9. 2D graph of DLP output.

Fig. 8 and 9 are the same graphs as Fig. 5 and 6, but are graphs using the DLP model instead of the MLP model.

D. Definition of LSTM Model

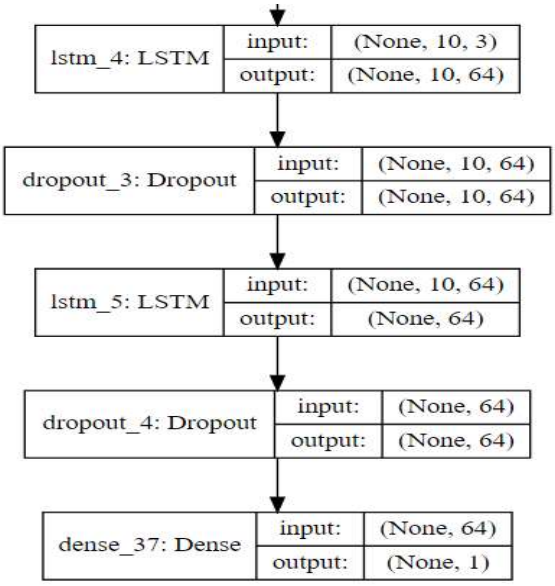

Fig. 10. Overall composition of LSTM model.

Fig. 10 is an overall composition of the LSTM model. Its input data set is in [10, 210, 10, 3]. The first index 23 is the values resulted from dividing the values of 130~220 by appropriate numbers as follows;

```
if (range = = 130~220 and divnum = = 10)
    while (index = =220, i = 0)
      index = 130 +( i * divnum)
      i++
```

If the size of divnum is too small, graphical visibility and explanatory power decrease. Otherwise, if too big, the amount of computation increases. In this study, the divnum was set to 10.

The divnum means the number of splits.

If divnum is 5,

Example: 130, 135, 140 ……215, 220cm

The second index 210 is the value from (data length – window_size). The first data does not have any preceding data so that the data after window_size are predicted.

The window_size is the data viewing size to a given data set.

If data length is 10, window_size is 5, for predict index 6 data it need previous 5(window_size) data, so you can predict index 6,7,8,9 data

The third index 10 stands for the window_size. This data set and the preceding set are used to predict the values after window_size.

The last index means the number of features such as PPM values, time and distance.

E. Technique to Combine Models

The MLP, DLP and LSTM models of the previous section are combined together with voting and stacking techniques.

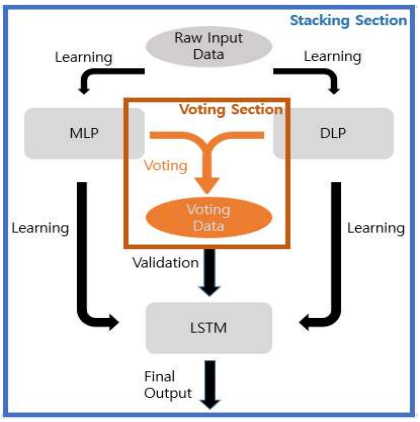

Fig. 11. Ensemble technique in overall.

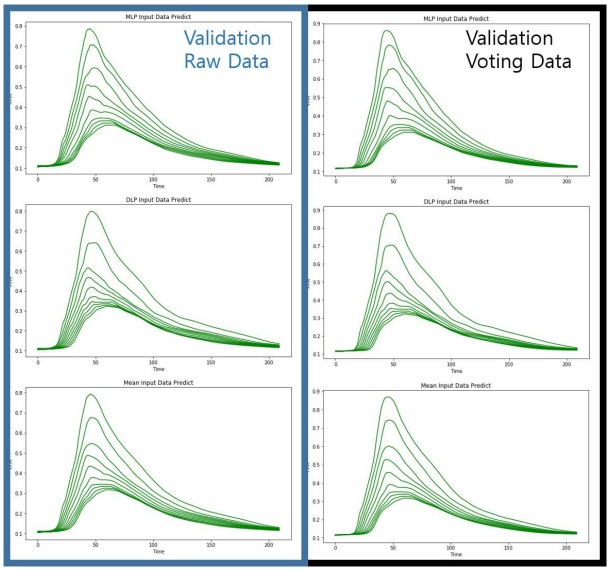

Fig. 12. 2D Graphs given different validation.

Fig. 11 shows overall features of the ensemble technique that is used in this study. The MLP and DLP are trained with the raw data that are collected at 130cm, 160cm, 190cm and 220cm, respectively. The trained MLP and DLP predict PPM values every 10cm from 130cm up to 220cm.

Voting with the MLP and DLP output data produces voting data as a result. Unlike conventional stacking, the data for validation is from voting output data instead of raw data. The advantage of this is that it is possible to obtain a variety of data without being taken into the raw data. Accuracy is verified in the section 4, Analysis and Results.

Analysis and Results.

Fig. 12 is a graph showing 2D LSTM output value when give MLP, DLP and Voting data to LSTM as input. The difference between the two cases is what to give the validation data. Left blue line graphs validation data is raw data, right black line validation data is voting data

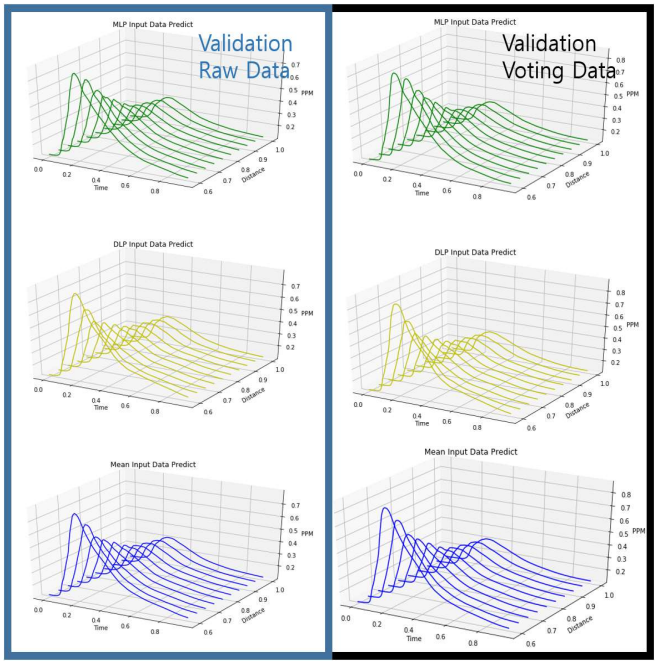

Fig. 13. 3D graphs given different validation.

Fig. 13 is a set of 3D graphs of Fig. 12. In Fig. 12 and 13, raw data and MSE loss are calculated by using data at distance 130, 160, 190 and 220cm.

```
DLP Data Score 130 :  0.006256255314143358
DLP Data Score 160 :  0.0018522914441605656
DLP Data Score 190 :  0.0001574350391O498706
DLP Data Score 220 :  3.992589433423826e-05
All DLP Data MSE Mean Score :  0.002076476922935787


MLP Data Score 130 :  0.00889183653524816
MLP Data Score 160 :  0.0004869785883260455
MLP Data Score 190 :  0.0001696899615752705S
MLP Score 220 :  2.25226340537185e-05
All MLP Data MSE Mean Score :  0.002392756929800938З


Voting Data Score 130 :  0.007574045924695759
Voting Data Score 160 :  0.0011696350162435B5
Voting Data Score 190 :  0.0001635625003401288
Mean Data Score 220 :  3.122426419397838e-05
All Voting Data MSE Mean Score :  0.002234616926368363
```

Fig. 14. Screen image of MSE score given validation voting data.

```
DLP Data Score 130 :  0.011557289797304911
DLP Data Score 160 :  0.002705323715305262
DLP Data Score 190 :  9.978870836900432e-05
DLP Data Score 220 :  2.3355724698577416e-05
All DLP Data MSE Mean Score :  0.003596439486419439


MLP Data Score 130 :  0.01424583147608051
MLP Data Score 160 :  0.001090171837222418
MLP Data Score 190 :  0.0001156318124875538б
MLP Score 220 :  4.3796535642551946e-05
All MLP Data MSE Mean Score :  0.0038738579153582585


Voting Data Score 130 :  0.012901560636692712
Voting Data Score 160 :  0.00189774777626384
Voting Data Score 190 :  0.0001077102604282790Э
Mean Data Score 220 :  3.357613017056468e-05
All Voting Data MSE Mean Score :  0.003735148700888848Б
```

Fig. 15. Screen image of MSE score given validation voting data.

Fig. 14 and 15 show the MLP, DLP and Voting output data as input data to the LSTM model used above and compares it with raw data. The difference between Fig. 14 and Fig. 15 is depends on what is used as validation data.

The data used in this study has a lower MSE score when using voting data than when using validation data as raw data. It means closer to raw data when voting data was used [10]. The ensemble technique also outputs different result values in different situations. Appropriate use is required for the situation.

IV. Conclusion

In this study, various data were generated by using machine learning models and two ensemble techniques using four data sets (130cm, 160cm, 190cm, 220cm), and the gas diffusion simulator was fabricated by combining them.

In addition to the models used in this study, more diverse data sets can be created by using other models. If a larger amount of data needs to be created, it can be completed in various model combinations as above, and the data can be modified by giving different validation values in the same combination.

This can be used to multiply data with a small number of raw data, and in a special situation such as gas diffusion simulation, it can be used as a simulator.

Further study is being planned to build a simulator by increasing the variables as mentioned above.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

KyungMin Gwak (K.M Gwak) performed ML research and collected data from experiments; Young J. Rho designed experiment and supervised entire research.

Acknowledgment

This research was partly supported by the research project 2019G00481 of Korea Research Foundation in Korea.

References

[1] Y. Yang, "Hazardous chemical dispersion analysis and surrogate model optimization using artificial neural network," University of Maryland at College Park, 2017.

[2] M. Clelland, L. James, D. E. Rumelhart, and G. E. Hinton, "Explorations in the microstructure of cognition: Foundations," *Parallel Distributed Processing*, vol. 1, 1986.

[3] Neural Network FAQ Document, AI FAQ Part3 Section 9, Oct 2019.

[4] K. M. Gwak and Y. J Rho, "Experimental machine learning study on CO_2 gas dispersion," in *Proc. 2019 9th IEEE Symposium on Computer Applications and Industrial Electronics,* 2019.

[5] A. Graves, "Supervised sequence labelling with recurrent neural networks," *Studies in Computational Intelligence*, 2008.

[6] Kaggle Ensembling Guide, Oct. 2019.

[7] Data Science School Insight, "Random forest," Oct. 2019.

[8] J. Brownlee, "How to develop a stacking ensemble for deep learning neural networks in python with keras," *Machine Learning Mastery*, Oct. 2019

[9] W. K. Lee, K. M. Gwak *et al*., "Development of meter to collect CO_2 dispersion data," *KSC,* pp. 1698-1700, 2018.

[10] Wikipedia, "Mean squared error," Oct. 2019.

Kyung-Min Gwak received his BS in computer engineering at Korea Polytechnic University in 2019. He is currently a master's course in the Department of Smart Manufacturing Engineering at Korea Polytechnic University.

Recently, he is a research on AI, data-mining technology.

Young J. Rho received his BS at Korea University, MS at FDU, Ph.D at UNSW. He has been a professor at the KPU since 2005. His interests are SW, HCI, IoT, ML.