JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 12, pp. 267-275, December 2022
https://doi.org/10.9708/jksci.2022.27.12.267

Air Pollution Risk Prediction System Utilizing Deep Learning Focused on Cardiovascular Disease

Jisu Lee*, Yoo-Jin Moon*

*Student, Division of Global Business & Technology, Hankuk University of Foreign Studies, Gyeonggi, Korea
*Professor, Division of Global Business & Technology, Hankuk University of Foreign Studies, Gyeonggi, Korea

[Abstract]

This paper proposed a Deep Neural Network Model system utilizing Keras for predicting air pollution risk of the cardiovascular disease through the effect of each component of air on the harmful virus using past air information, with analyzing 18,000 data sets of the Seoul Open Data Plaza. By experiments, the model performed tasks with higher accuracy when using methods of sigmoid, binary_crossentropy, adam, and accuracy through 3 hidden layers with each 8 nodes, resulting in 88.92% accuracy. It is meaningful in that any respiratory disease can utilize the risk prediction system if there are data on the effects of each component of air pollution and fine dust on oil-borne diseases. It can be further developed to provide useful information to companies that produce masks and air purification products.

▸ **Key words**: Deep neural network, Keras, Cardiovascular disease, Air pollution, Risk prediction

[요 약]

이 논문은 대기오염의 심장병에 대한 위험도를 예측하기 위하여 Keras를 활용한 Deep Neural Network Model 시스템을 제안하였다. 연구 데이터로 서울열린데이터광장의 서울시 기간별 시간평균 대기환경 데이터 18,000개의 데이터 셋을 분석하여, 심장병 질병에 미치는 영향에 대한 정보를 얻을 수 있었다. 이 모델은 각각 8개의 노드를 가진 3개의 은닉층, Sigmoid, Binary_crossentropy, Adam과 Accuracy를 사용했을 때 88.92%의 높은 정확도를 얻을 수 있었다. 이 시스템은 각 지역별 대기오염에 따른 심장병 질병 위험도를 예측하여 유용한 질병 예방의 지표로 활용 가능하다고 사료되고, 대기오염과 미세먼지의 각 성분이 유해질환에 미치는 영향에 대한 데이터만 존재한다면 어떠한 호흡기 질환이든 위험도 예측 결과를 알 수 있다는 것에 의미가 있다. 이 시스템을 더욱 발전시킨다면, 마스크 및 공기정화제품 생산기업에게 유용한 정보를 제공하여 기업의 기술개발에 도움이 될 수 있다고 사료된다.

▸ **주제어**: 딥 뉴럴네트워크, 케라스, 심장병, 공기 오염, 위험도 예측

• First Author: Jisu Lee, Corresponding Author: Yoo-Jin Moon
*Jisu Lee (gksm1215@gmail.com), Division of Global Business & Technology, Hankuk University of Foreign Studies
*Yoo-Jin Moon (yjmoon@hufs.ac.kr), Division of Global Business & Technology, Hankuk University of Foreign Studies

• Received: 2022. 09. 14, Revised: 2022. 12. 09, Accepted: 2022. 12. 09.

I. Introduction

Air pollution satellite maps around the world showed that Korea and China had the highest pollution level[1], and Korea had the highest level of pollution compared to its national territory. In addition, it is said that if Korea would not properly establish fine dust measures in the near future, it will be the country with the highest early mortality rate and economic damage from air pollution among member countries of the Organization for Economic Cooperation and Development (OECD) by 2060. In addition, air pollution-related diseases are increasing in the world.

However, people's awareness of the air pollution has been insufficient, and the purpose of wearing a mask has been limited to the COVID-19[2]. So, it is expected that wearing a mask would be neglected and interest in air pollution would decrease if the With Corona situation and the virus are expected to end[3, 4].

Based on this situation, the research aimed to predict air pollution risk of the cardiovascular disease on the effect of each component of air on the harmful virus using air information in the Seoul districts, thereby providing awareness of the risk of air pollution[5-7]. It focused on the cardiovascular disease utilizing deep learning.

The project is meaningful in that any respiratory disease can utilize the risk prediction system if there are data on the effects of each component of air pollution and fine dust on oil-borne diseases.

Even though there were many types/numbers of data required to predict information according to the existing standby information, the complexity of calculation and the scope of information provision were wide[8, 9]. However, this research would be utilized by establishing a system that predicted the risk of harmful diseases in more detail at lower cost through past standby data and deep learning[10-12]. In addition, figures that the existing fine dust prediction system did not solve could be comprehensively predicted.

II. Previous Studies and Preliminaries

1. Previous studies

Table 1. Fine Dust Prediction Accuracy in Each Model

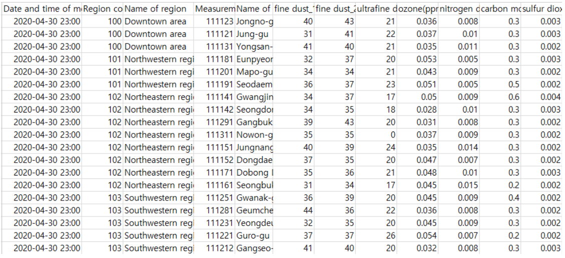
Model	Accuracy				
	Good	Normal	Bad	Very Bad	Total Acc.
Model-2	0.7077	0.8693	0.2222	0.0	0.7533
Model-4	0.8000	0.7778	0.5000	0.0	0.7632
Model-6	0.7308	0.7974	0.3333	0.333	0.7368
Model-8	0.7407	0.8431	0.3333	0.0	0.7437
RNN-1	0.0846	0.9804	0.1111	0.0	0.5362
LSTM-3	0.6154	0.7516	0.1666	0.0	0.6513

Recently, various studies on fine dust have been conducted[13-17]. The previous study[13] found that carbon monoxide, nitrogen dioxide, sulfur dioxide, ozone, and fine dust were statistically significant in evaluating the risk of urban air pollution on cardiovascular disease. And, the project[14] in Table 1 conducted by Hanyang University predicted the level of fine dust in Korea and designed the CRNN (combination of RNN and CNN, Model-2) model using domestic and foreign fine dust, wind direction, and wind speed data. As shown in Table 1, the proposed model achieved about 76% accuracy in Model-2 for predicting the level of fine dust [14] by differentiating parameters for the functions in the CRNN model. The prvious studies[15-17] were researches focused on cardiovascular disease rather than on air pollution, published by the journals related to cardiology.

The Hanyang University project[14] is different from this research in that it performed the risk through fine dust prediction and its level by analyzing various factors such as fine dust, wind direction, and wind speed data home and abroad, while this research conducted data analysis from each component of air pollutants to determine the effect of each component on the specified disease and the risk level of disease incidence.

Though there were various artificial intelligence systems related to fine dust prediction, they did not deal with the effect of air pollution on the disease, that is, the actual risk[18, 19]. Thus, it is expected that this artificial intelligence system will be able to provide useful information of the effect on the specified disease from each component of air pollutants[20-22].

2. Big data sources



Date and time of m	Region co	Name of region	Measurem	Name of	fine dust_1	fine dust_2	ultrafine d	ozone(ppr	nitrogen d	carbon mc	sulfur diox
2020-04-30 23:00	100	Downtown area	111123	Jongno-g	40	43	21	0.036	0.008	0.3	0.003
2020-04-30 23:00	100	Downtown area	111121	Jung-gu	31	41	22	0.037	0.01	0.3	0.003
2020-04-30 23:00	100	Downtown area	111131	Yongsan-	41	40	21	0.035	0.011	0.3	0.002
2020-04-30 23:00	101	Northwestern regi	111181	Eunpyeor	32	37	20	0.053	0.005	0.3	0.003
2020-04-30 23:00	101	Northwestern regi	111201	Mapo-gu	34	34	21	0.043	0.009	0.3	0.002
2020-04-30 23:00	101	Northwestern regi	111191	Seodaem	36	37	23	0.051	0.005	0.5	0.002
2020-04-30 23:00	102	Northeastern regi	111141	Gwangjin	34	37	17	0.05	0.009	0.6	0.004
2020-04-30 23:00	102	Northeastern regi	111142	Seongdoı	34	35	18	0.028	0.01	0.3	0.003
2020-04-30 23:00	102	Northeastern regi	111291	Gangbuk	39	43	20	0.031	0.008	0.3	0.002
2020-04-30 23:00	102	Northeastern regi	111311	Nowon-g	35	35	0	0.037	0.009	0.3	0.002
2020-04-30 23:00	102	Northeastern regi	111151	Jungnang	40	39	24	0.035	0.014	0.3	0.002
2020-04-30 23:00	102	Northeastern regi	111152	Dongdae	37	35	20	0.047	0.007	0.3	0.002
2020-04-30 23:00	102	Northeastern regi	111171	Dobong I	35	36	21	0.048	0.01	0.3	0.003
2020-04-30 23:00	102	Northeastern regi	111161	Seongbul	31	34	17	0.045	0.015	0.2	0.002
2020-04-30 23:00	103	Southwestern regi	111251	Gwanak-(	36	39	20	0.045	0.009	0.4	0.002
2020-04-30 23:00	103	Southwestern regi	111281	Geumche	44	36	22	0.036	0.008	0.3	0.002
2020-04-30 23:00	103	Southwestern regi	111231	Yeongdeı	32	35	20	0.045	0.009	0.3	0.002
2020-04-30 23:00	103	Southwestern regi	111221	Guro-gu	37	37	26	0.054	0.007	0.2	0.002
2020-04-30 23:00	103	Southwestern regi	111212	Gangseo-	41	40	20	0.032	0.008	0.3	0.003

Fig. 1. Time Average Air Environment Information by Period in Seoul (18,000 data sets)

The data illustrated in Fig. 1 are information on the average air environment by Seoul Metropolitan Government provided by Seoul Open Data Plaza[23]. The data consist of measurement date and time, region code, region name, measurement station code, measurement station name, fine dust for 1 hour, fine dust for 24 hours, ultrafine dust, ozone, nitrogen dioxide, carbon monoxide, and sulfur dioxide concentrations. The research processed the raw 18,000 data sets for fitting the deep learning system.

III. Performance Comparison of the System Architecture Utilizing Deep Learning

For construction of the deep neural network Chapter III explains the key elements of the system consistently - selection of the independent variables, data normalization and encoding, selection of the dependent variables, performance comparison by the number of hidden layers and nodes, selection of optimization methods.

1. Selection of the independent variables

The previous study "Analyzing Yellow Dust Effects in Air Pollution Risk Evaluation - Effects on Death Rate by Cause" published by the Korea Environmental Health Association, found that carbon monoxide, nitrogen dioxide, sulfur dioxide, ozone, and fine dust were statistically significant in evaluating the risk of urban air pollution[13]. They were selected as independent variables in the system to be proposed.

2. Data normalization and encoding

In the case of existing data, it was difficult to accurately calculate the impact due to the big differences in data values of each element.

Table 2. Converted Data Set after Data Normalization

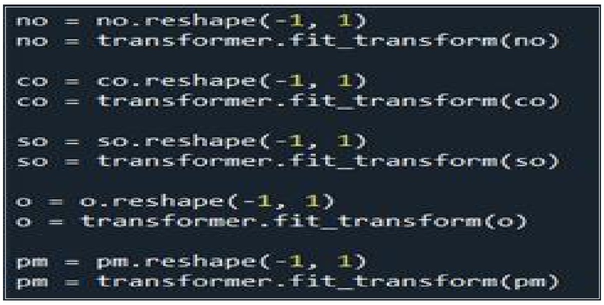
fine dust	ozone	nitrogen dioxide	carbon monoxide	sulfur dioxide
40	0.036	0.008	0.3	0.003
31	0.037	0.01	0.3	0.003
41	0.035	0.011	0.3	0.002
32	0.053	0.005	0.3	0.003
34	0.043	0.009	0.3	0.002
36	0.051	0.005	0.5	0.002
34	0.05	0.009	0.6	0.004
34	0.028	0.01	0.3	0.003

⇩

PMnew	Onew	NOnew	COnew	SOnew
0.145985	0.371134	0.098765	0.272727	0.176471
0.113139	0.381443	0.123457	0.272727	0.176471
0.149635	0.360825	0.135802	0.272727	0.176471
0.116788	0.546392	0.061728	0.272727	0.176471
0.124088	0.443299	0.111111	0.272727	0.176471
0.131387	0.525773	0.061728	0.454545	0.176471
0.124088	0.515464	0.111111	0.545455	0.235294
0.124088	0.28866	0.123457	0.272727	0.176471

In order to extract accurate result values, data of each existing component were normalized as shown in Table 2 with the Python codes in Fig. 2. In Table 2, the column values of fine dust, ozone, nitrogen dioxide, carbon monoxide, and sulfur dioxide gas were converted into those of PMnew, Onew, NOnew, COnew, SOnew.

The data were converted into a number between -1 and 1, by normalizing to the corresponding value. Gaussian Normalization was used as a method of normalizing x' = (x - means) / standard deviation instead of the input x.

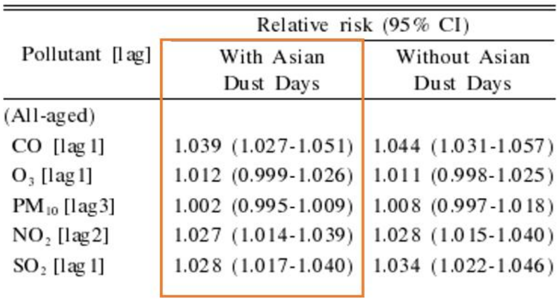


```
no = no.reshape(-1, 1)
no = transformer.fit_transform(no)

co = co.reshape(-1, 1)
co = transformer.fit_transform(co)

so = so.reshape(-1, 1)
so = transformer.fit_transform(so)

o = o.reshape(-1, 1)
o = transformer.fit_transform(o)

pm = pm.reshape(-1, 1)
pm = transformer.fit_transform(pm)
```

Fig. 2. Codes for Element Data Normalization

3. Selection of the dependent variables

3.1 Standard weights on air pollution risk of cardiovascular disease



Pollutant [lag]	Relative risk (95% CI)	
	With Asian Dust Days	Without Asian Dust Days
(All-aged)		
CO [lag 1]	1.039 (1.027-1.051)	1.044 (1.031-1.057)
O_3 [lag1]	1.012 (0.999-1.026)	1.011 (0.998-1.025)
PM_{10} [lag3]	1.002 (0.995-1.009)	1.008 (0.997-1.018)
NO_2 [lag2]	1.027 (1.014-1.039)	1.028 (1.015-1.040)
SO_2 [lag 1]	1.028 (1.017-1.040)	1.034 (1.022-1.046)

Fig. 3. Analysis for the Effects of Each Component of Air Pollution on Cardiovascular Disease

To produce risk data for air pollution in order to calculate dependent variables, the previous study[13] published by the Korea Environmental Health Association was referenced. The data illustrated in Fig. 3 [13] were obtained through Poisson regression analysis for the effects of each component of air pollution on cardiovascular disease, which were produced by the experts of the air pollution and are believed to be trustworthy at this moment.

According to the Fig. 3, it was found that CO (carbon monoxide), NO2 (nitrogen dioxide), SO2 (sulfur dioxide gas), O3 (ozone), and PM (particular matter, fine dust) showed a statistically significant positive association with diseases caused by air pollution. Referring to the data in Fig. 3, new virtual data in Table 3 were produced by calculating the influence of each component from the existing data set. The formula of Risk was set as follows.

*Risk: CO*1.039 + NO2*1.027 + SO2*1.028 + O3*1.012+ PM*1.002*

Table 3. Risk by Calculating the Influence Value of Each Component after Normalization

PMnew	Onew	NOnew	COnew	SOnew	Risk	result 20	result 30	result 4
0.145985	0.371134	0.098765	0.272727	0.176471	1.088073	2	2	1
0.113139	0.381443	0.123457	0.272727	0.176471	1.090951	2	2	1
0.149635	0.360825	0.135802	0.272727	0.176471	1.058863	2	2	1
0.116788	0.546392	0.061728	0.272727	0.176471	1.198141	1	1	1
0.124088	0.443299	0.111111	0.272727	0.176471	1.091337	2	2	1
0.131387	0.525773	0.061728	0.454545	0.176471	1.320341	1	1	1
0.124088	0.515464	0.111111	0.545455	0.235294	1.568706	1	0	1
0.124088	0.28866	0.123457	0.272727	0.176471	1.008025	2	2	1
0.142336	0.319588	0.098765	0.272727	0.117647	0.97178	2	2	2
0.127737	0.381443	0.111111	0.272727	0.117647	1.032429	2	2	1
0.145985	0.360825	0.17284	0.272727	0.117647	1.093243	2	2	1

3.2 Production of result value elements according to Risk.

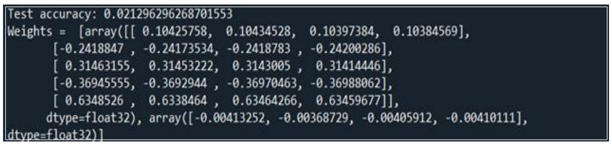
In Table 3 an element 'result20' classified the existing data set into 3 risk categories - hazardous data of the top 20% as 0, normal data as 1, and safe data of the bottom 20% as 2, respectively. An element 'result30' also produced data classified based on 30% instead of 20% in result20. Since the risk value of an element 'result4' had a value from 0 to about 2.38, it was divided into one of three cases: 2 or more (about top 12%), between 2 and 1, and 1 or less (about bottom 14%), respectively, and expressed as 0, 1, and 2. That is, the value of the result4 decided whether the air was prone to the cardiovascular disease.

Among these three elements of 'result20', 'result30' and 'result4,' the element 'result4' showed the highest accuracy from the logical viewpoint. Based on the formula of Risk, CO, NO2, SO2, O3 and PM before normalization were selected as independent variables, and the element 'result4' was selected as a dependent variable.

4. Performance comparison by the number of hidden layers and nodes

By varying the number of hidden layers and nodes, performance comparisons of the deep neural network were conducted when Activation was set to softmax, Loss Function binary_crossentropy, Optimizer adam, and Accuracy binary_accuracy. For the performance comparison, the number of hidden layer gradually increased by 1 and the number of nodes by 4. The representative comparison results were as follows.

1) When 1 hidden layer with 4 nodes

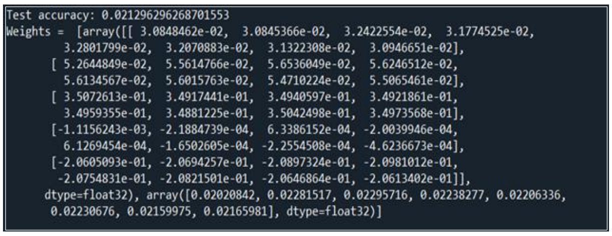


```
Test accuracy: 0.021296296268701553
Weights =  [array([[ 0.10425758,  0.10434528,  0.10397384,  0.10384569],
       [-0.2418847 , -0.24173534, -0.2418783 , -0.24200286],
       [ 0.31463155,  0.31453222,  0.3143005 ,  0.31414446],
       [-0.36945555, -0.3692944 , -0.36970463, -0.36988062],
       [ 0.6348526 ,  0.6338464 ,  0.63464266,  0.63459677]],
      dtype=float32), array([-0.00413252, -0.00368729, -0.00405912, -0.00410111],
dtype=float32)]
```

Fig. 4. When 1 hidden layer with 4 nodes, the accuracy was 2.12%.

When there was one hidden layer with four nodes, the accuracy was 2.12% as shown in Fig. 4, and each weight of CO, NO2, SO2, O3 and PM was [0.10425758, 0.10434528, 0.10397384, 0.10384569], [-0.2418847, -0.24173534, -0.2418783, -0.24200286], [0.31463155, 0.31453222, 0.3143005, 0.31414446], [-0.36945555, -0.3692944, -0.36970463, -0.36988062], [0.648526, 0.6338464, 0.63464266, 0.63459677].

2) When 1 hidden layer with 8 nodes



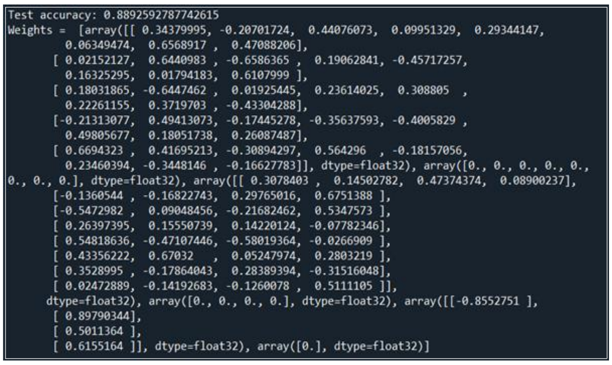
```
Test accuracy: 0.021296296268701553
Weights =  [array([[ 3.0848462e-02,  3.0845366e-02,  3.2422554e-02,  3.1774525e-02,
         3.2801799e-02,  3.2070883e-02,  3.1322308e-02,  3.0946651e-02],
       [ 5.2644849e-02,  5.5614766e-02,  5.6536049e-02,  5.6246512e-02,
         5.6134567e-02,  5.6015763e-02,  5.4710224e-02,  5.5065461e-02],
       [ 3.5072613e-01,  3.4917441e-01,  3.4940597e-01,  3.4921861e-01,
         3.4959355e-01,  3.4881225e-01,  3.5042498e-01,  3.4973568e-01],
       [-1.1156243e-03, -2.1884739e-04,  6.3386152e-04, -2.0039946e-04,
         6.1269454e-04, -1.6502605e-04, -2.2554508e-04, -4.6236673e-04],
       [-2.0605093e-01, -2.0694257e-01, -2.0897324e-01, -2.0981012e-01,
        -2.0754831e-01, -2.0821501e-01, -2.0646864e-01, -2.0613402e-01]],
      dtype=float32), array([0.02020842, 0.02281517, 0.02295716, 0.02238277, 0.02206336,
       0.02230676, 0.02159975, 0.02165981], dtype=float32)]
```

Fig. 5. When 1 hidden layer with 8 nodes, the accuracy was 2.12%.

When there was one hidden layer with eight nodes, the accuracy was 2.12%, as shown in Fig. 5.

3) When 3 hidden layers with each 8 nodes

When there were three hidden layers with each eight or four nodes, the accuracy was 88.92%, as shown in Fig. 6.



```
Test accuracy: 0.8892592787742615
Weights =  [array([[ 0.34379995, -0.20701724,  0.44076073,  0.09951329,  0.29344147,
         0.06349474,  0.6568917 ,  0.47088206],
       [ 0.02152127,  0.6440983 , -0.6586365 ,  0.19062841, -0.45717257,
         0.16325295,  0.01794183,  0.6107999 ],
       [ 0.18031865, -0.6447462 ,  0.01925445,  0.23614025,  0.308805  ,
         0.22261155,  0.3719703 , -0.43304288],
       [-0.21313077,  0.49413073, -0.17445278, -0.35637593, -0.4005829 ,
         0.49805677,  0.18051738,  0.26087487],
       [ 0.6694323 ,  0.41695213, -0.30894297,  0.564296  , -0.18157056,
         0.23460394, -0.3448146 , -0.16627783]], dtype=float32), array([0., 0., 0., 0., 0.,
0., 0., 0.], dtype=float32), array([[ 0.3078403 ,  0.14502782,  0.47374374,  0.08900237],
       [-0.1360544 , -0.16822743,  0.29765016,  0.6751388 ],
       [-0.5472982 ,  0.09048456, -0.21682462,  0.5347573 ],
       [ 0.26397395,  0.15550739,  0.14220124, -0.07782346],
       [ 0.54818636, -0.47107446, -0.58019364, -0.0266909 ],
       [ 0.43356222,  0.67032   ,  0.05247974,  0.2803219 ],
       [ 0.3528995 , -0.17864043,  0.28389394, -0.31516048],
       [ 0.02472889, -0.14192683, -0.1260078 ,  0.5111105 ]],
      dtype=float32), array([0., 0., 0., 0.], dtype=float32), array([[-0.8552751 ],
       [ 0.89790344],
       [ 0.5011364 ],
       [ 0.6155164 ]], dtype=float32), array([0.], dtype=float32)]
```

Fig. 6. When 3 hidden layers with each 8 nodes, the accuracy was 88.92%.

From the performance comparison for the number of hidden layers and nodes, when there was one hidden layer it could be seen that the accuracy decreased when the node increased. When the number of the hidden layers increased, the accuracy increased compared to one hidden layer. Based on the experiments, the system selected the model through 3 hidden layers with each 8 nodes.

5. Selection of optimization methods

To suggest the appropriate model of the neural network for the system, the research conducted performance tests for each optimization method and selected the high performance methods in the parameters of Activation Function, Loss Function, Optimizer and Accuracy.

5.1 Activation Function comparison

To compare the methods for Activation Function, the remaining function parameters were set as follows. Loss Function was set to binary_crossentropy, Optimizer adam, and Accuracy binary_accuracy.

When the sigmoid was applied to the Activation Function parameter, the accuracy was 88.92%, and the weights of CO, NO2, SO2, O3 and PM were [136.7423], [32.51897], [138.43831], [198.22136], and [12.213113].

When the softmax was applied to the Activation Function parameter, the softmax accuracy was 88.92%.

When the tanh was applied to the Activation Function parameter, the tanh accuracy was 88.92%.

When the relu was applied to the Activation Function parameter, the relu accuracy was 88.92%.

In the case of Activation Function, the accuracy was all high, but the results of the weight values were different, so sigmoid, tanh and relu which did not produce negative values were confirmed to be suitable.

5.2. Loss Function comparison

To compare the methods for Loss Function, the remaining function parameters were set as follows. Activation Function was set to sigmoid, Optimizer adam, and Accuracy accuracy.

When the MAE(Mean Absolute Error) was applied to the Loss Function parameter, the MAE accuracy was 88.92%, and the weights of CO, NO2, SO2, O3 and PM were [4.9927692], [4.5129194], [5.716214], [5.2808733], and [5.468247].

When the binary_crossentropy was applied to the Loss Function parameter, the binary_crossentropy accuracy was 88.92%.

When the MSE(Mean Square Error) was applied to the Loss Function parameter, the MSE accuracy was 88.92%.

Since the accuracies were all high, it can be seen that the weight values were similar to the actual reflected values and stable in the MAE, and binary_crossentropy and MSE.

5.3. Optimizer comparison

To compare the methods for Optimizer, the remaining function parameters were set as follows. Activation Function was set to sigmoid, Loss Function MAE, and Accuracy binary_accuracy.

When the adam was applied to the Optimizer parameter, the adam accuracy was 88.92%, and the weights of CO, NO2, SO2, O3 and PM were [4.467369], [5.42335], [5.405407], [5.6855927], and [4.6777765].

When the sgd was applied to the Optimizer parameter, the sgd accuracy was 88.92%.

It could be seen that both adam and sgd had high accuracy, but the adam weight values had accuracy more similar to those of the existing theory than sgd.

5.4. Accuracy comparison

To compare the methods for Accuracy, the remaining function parameters were set as follows. Activation Function was set to sigmoid, Loss Function binary_crossentropy, and Optimizer adam.

When the accuracy was applied to the Accuracy parameter, the method of accuracy had 88.92%, and the weights of CO, NO2, SO2, O3 and PM were [1.2493447], [0.30460027], [0.45930934], [0.9662148], and [-0.28202996].

When the MSE was applied to the Accuracy parameter, the MSE accuracy was 137.69%.

When the binary_accuracy was applied to the Accuracy parameter, accuracy of the binary_accuracy was 88.92%.

It could be seen that methods of accuracy and binary_accuracy were both stable in accuracy and weight values.

Table 4. Optimization Methods Working in the Deep Learning Model Suggested

Functions for Optimization	Recommended Methods
Activation Function	sigmoid, tanh, relu
Loss Function	MAE, binary_crossentropy, MSE
Optimizer	adam
Accuracy	binary_accuracy, accuracy

Overall, optimization methods with the high performance in the deep learning model suggested could be recommended in Table 4.

IV. Results of the Proposed Deep Learning Model

When comparing the methods for Activation Function, Loss Function, Optimizer, Accuracy, number of layers, and number of nodes respectively in the section III, the success rates of the model were often similar. So the research further compared the risk values reflected when creating weight data.

1. Standard risk data of weight

Table 5. Standard Risk Values of Weight

Ranking	Variable name.	weight
1st	CO	1.039
2nd	SO2	1.028
3rd	NO2	1.027
4th	O3	1.012
5th	PM	1.002

Table 5 illustrates standard risk values of weight reflected from the previous study [13], where the risk variable CO has the weight of 1.039, SO2 1.028, NO2 1.027, O3 1.012 and PM 1.002.

2. The Proposed Deep Learning Model-1

Table 6. Weight Values Most Similar to Standard Risk Values

success rate	0.8892	
weight	CO	5.2149105
	NO2	5.559767
	SO2	5.9619613
	O3	4.9028573
	PM	4.6480546

Ranking	Variable name.
1st	SO2
2nd	CO
3rd	NO2
4th	O3
5th	PM

Table 6 shows the weight values most similar to standard risk values with success rate 88.92% in the proposed deep learning model-1. The proposed model-1 obtained most similarly to weights of the standard risk data when Activation Function was sigmoid, Loss Function MAE, Optimizer adam, and Accuracy binary_accuracy, through 3 hidden layers with each 8 nodes. Except for the differences in ranking 1st and 2nd, the others were similar to the standard reflection ratio.

3. The Proposed Deep Learning Model-2

Table 7. Weight Values Similar to Standard Risk Values

success rate	0.881	
weight	CO	3.459279
	NO2	2.4554453
	SO2	2.564329
	O3	2.0333827
	PM	2.0411696

Ranking	Variable name.
1st	CO
2nd	SO2
3rd	NO2
4th	PM
5th	O3

Table 7 shows the weight values the 2nd most similar to standard risk values with success rate 88.1% in the proposed deep learning model-2. The proposed model-2 obtained 2nd most similarly to weights of the standard risk data when Activation Function was tanh, Loss Function MAE, Optimizer adam, and Accuracy binary_accuracy, through 3 hidden layers with each 8 nodes. Except for the differences in ranking 4th and 5th, the others were similar to the standard reflection ratio.

V. Conclusions

The research suggested the deep learning model system for predicting air pollution risk of the cardiovascular disease on the effect of each component of air on the harmful virus using air information, thereby providing awareness of the risk of air pollution. The suggested deep learning system showed 88.92% accuracy as a result of the prediction, and can be easily used for predicting air quality focused on the cardiovascular disease.

Even though there were many types/numbers of data required to predict information according to the existing standby information, the complexity of calculation and the scope of information provision were wide. However, this research would be utilized by establishing a system that predicted the risk of harmful diseases in more detail at lower cost through past standby data and deep learning. In addition, figures that the existing fine dust prediction system did not solve could be comprehensively predicted.

The Proposed Deep Learning Model-1 and Model-2 were described in Chapter IV. Each model had favorites compared to the other model. It depends on the users which model is to be selected.

The system aimed to alert people to the serious harmful effects of air pollution and fine dust, and it could be further developed to provide useful information to companies that produce masks and air purification products.

The project is meaningful in that any respiratory disease can utilize the risk prediction system if there are data on the effects of each component of air pollution and fine dust on oil-borne diseases. In addition to the predictive result on cardiovascular diseases due to air pollution, this system can draw various substances such as mortality from air pollution and mortality from respiratory diseases of air pollution.

Limitation of the research is that in the degree of air pollution hazardous data, normal data and safe data should be decided by the expert and the air pollution research. And it is expected that various air pollution events will need to be supplemented by learning additional data that can be embraced, not just reflecting past air pollution data.

ACKNOWLEDGEMENT

This work was supported by Hankuk University of Foreign Studies Research Fund of 2022.

REFERENCES

[1] Jun Wu, Jian Wang, Stephen Nicholas, Elizabeth Maitland, and Qiuyan Fan, “Application of Big Data Technology for COVID-19 Prevention and Control in China: Lessons and Recommendations”, Journal of Medical Internet Research, Vol. 22, No. 10: e21980, Oct. 2020. *doi:10.2196/21980.*

[2] Jaeyun Choi, “Effects of Social Risk from Corona-19 (COVID-19) on Consumer Sentiment and HMR Purchasing Patterns”, Graduate School of Living and Environment, Yonsei University, 2020, Seoul.

[3] Victor Fabius, Sajal Kohli, Bjorn Timelin, and sofia Moulvad Varanen, “How COVID-19 is Changing Consumer Behavior— Now and Forever”, McKinsey & Company, July 2020.

[4] Duralia Oana, “The Impact of the Current Crisis Generated by the Pandemic on Consumer Behavior”, Studies in Business and Economics, Vol. 15, Issue 2, Oct. 2020. *https://doi.org/10.2478/sbe.2020-0027*

[5] Bruce Lehrman, “Big Data’s Role in the Post-COVID Era”, Data Agility, Vol. 16, Issue 11, Sept. 2020. *www.pipelinepub.com.*

[6] Wesley Chai, Mark Labbe, and Craig Stedman, “Big Data Analytics,” 2021. *https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics*

[7] Sangho Kim, “A Study on Relationship of BDBA (Big Data Business Analytics) System and Supply Chain Management,” Journal of Korea Research Association of International Commerce, Vol. 19, No. 2, pp. 89-107, 2019.

[8] Honqmei Li, Jinying Huang, and Shuwei Ji, “Bearing Fault Diagnosis with a Feature Fusion Method Based on An Ensemble Convolutional Neural Network and Deep Neural Network,” Sensors (Basel, Switzerland), Vol. 19, Issue 9, pp. 2034, 2019.

[9] N. Yuvaraj, R. Arshath Raja, N.V. Kousik, Prashant Johri, and Mario José Diván, “Chapter15 - Analysis on the Prediction of Central Line-Associated Bloodstream Infections (CLABSI) Using Deep Neural Network Classification,” *Computational Intelligence and Its Applications in Healthcare*, Academic Press, pp. 229-244, 2020. *https://doi.org/10.1016/B978-0-12-820604-1.00016-9*

[10] Katy Warr, *“Strengthening Deep Neural Networks: Making AI Less Susceptible to Adversarial Trickery,”* O'Reilly Media, 2019.

[11] Jojo Moolayil, *“Learn Keras for Deep Neural Networks: A Fast-Track Approach to Modern Deep Learning with Python,”* Apress, 2019.

[12] Jen-Tzung Chien, *“Source Separation and Machine Learning,”* Elsevier: Academic Press, 2019.

[13] Ji-young Son, Yoon-shin Kim, Yeon-jung Kim, Jong-tae Lee, and Yong-sung Cho, “Yellow Dust Effect Analysis in Evaluating the Risk of Urban Air Pollution-The Effect of Total Deaths and the Mortality Rate by the Cause in Seoul,” Journal of the Korean

Environmental Health Association, Vol.35, No. 4, pp. 249-258, 2009.

[14] Ki-hyuk Lee, Woo-sung Hwang, and Myung-ryul Choi, "1-DCRNN Model Design for Predicting Fine Dust Risk Stages," Digital Convergence Research, Vol.19, No.2, pp.215-550, 2021.

[15] Sadeer G. Al-Kindi, Robert D. Brook, Shyam Biswal, and Sanjay Rajagopalan, "Environmental determinants of cardiovascular disease: lessons learned from air pollution,", Nature Reviews Cardiology, Vol.17, pp. 656-672, 2020.

[16] Sanjay Rajagopalan, Sadeer G. Al-Kindi, and Robert D. Brook, "Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review," Journal of American College of Cardiology, 2018 Oct, Vol.72, No.17, pp. 2054–2070, 2018.

[17] Sindana D. Ilango, and Rachel M. Shaffer, "Air Pollution, Cardiovascular Disease, and Dementia," JAMA Neurology, Vol.77, No.12, 2020. *doi:10.1001/jamaneurol.2020.4309*

[18] Gilbert Lim, Wynne Hsu, Mong Li Lee, Daniel Shu Wei Ting, and Tien Yin Wong, "Chapter 21 - Technical and Clinical Challenges of A.I. in Retinal Image Analysis," *Computational Retinal Image Analysis::Tools, Applications and Perspectives*, Academic Press, pp. 445-466, 2019. *https://doi.org/10.1016/B978-0-08-102816-2.00022-8*

[19] Wonil Lee, Byungjai Kim, and HyunWook Park, "Quantification of Intravoxel Incoherent Motion with Optimized B-values Using Deep Neural Network," Magnetic Resonance in Medicine, Feb. 2021. *DOI: 10.1002/mrm.28708*

[20] Russell, Stuart, *"Artificial Intelligence: A Modern Approach,"* Pearson Education, 2017.

[21] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *"Deep Learning,"* MIT Express, 2016.

[22] Ingook Cheon, *"Artificial Intelligence: Machine Learning and Deep Learning by Python,"* Infinity Books, 2020.

[23] Time Average Air Envitonment Information by Period in Seoul, Seoul Open Data Plaza, *http://data.seoul.go.kr*

Authors

Jisu Lee is a student of Hankuk University of Foreign Studies. She is studying Computer and Business. Her interested research area is artificial intelligence, database systems and data analysis.

Yoo-Jin Moon is a professor at Hankuk University of Foreign Studies. She received her Ph.D. from Seoul National University. Her interested area is artificial intelligence, natural language processing, big data analytics and database system etc.