Communication

Quick Estimation Model for the Concentration of Indoor Airborne Culturable Bacteria: An Application of Machine Learning

Zhijian Liu [1,†], Hao Li [2,3,*,†] and Guoqing Cao [4,*]

[1]   Department of Power Engineering, School of Energy, Power and Mechanical Engineering,
     North China Electric Power University, Baoding 071003, China; zhijianliu@ncepu.edu.cn
[2]   Department of Chemistry, The University of Texas at Austin, 105 E. 24th Street, Stop A5300,
     Austin, TX 78712, USA
[3]   Institute for Computational and Engineering Sciences, The University of Texas at Austin, 105 E. 24th Street,
     Stop A5300, Austin, TX 78712, USA
[4]   Institute of Building Environment and Energy, China Academy of Building Research, Beijing 100013, China
*   Correspondence: lihao@utexas.edu (H.L.); cgq2000@126.com (G.C.)
†   These authors contributed equally.

Abstract: Indoor airborne culturable bacteria are sometimes harmful to human health. Therefore, a quick estimation of their concentration is particularly necessary. However, measuring the indoor microorganism concentration (e.g., bacteria) usually requires a large amount of time, economic cost, and manpower. In this paper, we aim to provide a quick solution: using knowledge-based machine learning to provide quick estimation of the concentration of indoor airborne culturable bacteria only with the inputs of several measurable indoor environmental indicators, including: indoor particulate matter ($PM_{2.5}$ and PM_{10}), temperature, relative humidity, and CO_2 concentration. Our results show that a general regression neural network (GRNN) model can sufficiently provide a quick and decent estimation based on the model training and testing using an experimental database with 249 data groups.

1. Introduction

Many indoor microorganisms are potential threats to human health [1–6]. High concentrations of airborne bacteria or fungi may lead to some common diseases, such as respiratory symptoms, allergies, and asthma [7–9]. Studies have shown that outdoor particulate matter (PM) air pollution could cause formidable public health risk due to inhalable microorganisms that exist in $PM_{2.5}$ and PM_{10} pollutants [10]. Our recent studies further found that relatively high concentrations of indoor $PM_{2.5}$ and PM_{10} pollutants are qualitatively related to high concentrations of indoor inhalable bacteria and fungi [8,11], which seriously jeopardizes human health.

To reduce the potential threats of indoor inhalable microorganisms, a common method is the use of germicides [12] or antifungal agents [13]. However, before deciding whether to use these products, people usually struggle with a simpler question: what are the approximate concentrations of bacteria and fungi in my house? Experimentally, the measurements of bacterial or fungal concentrations are complicated, including sampling and culture, which require high economic cost and a long timescale. To estimate the indoor fungal concentration, we previously found a linear relationship between PM and indoor fungal concentration [8], which is applicable for indoor fungi estimation. However, to our

Int. J. Environ. Res. Public Health **2017**, *14*, 857

2 of 9

best knowledge, very few reports have shown an effective estimation method for the concentration of indoor airborne culturable bacteria.

To estimate the concentration of indoor airborne culturable bacteria, a good first step is to make full use of the indoor environmental indicators. With the assumption that some important indoor indicators (e.g., PM, temperature, relative humidity and CO_2 concentration) correlate (or partially correlate) with the growth of bacteria, they can be used as the independent variables for the prediction of bacterial concentration. Setting the bacterial concentration as the dependent variable, we can further use predictive machine learning methods trained from the experimental database. A good machine learning model can effectively "learn" from the existing database and perform precise and robust predictions. With the algorithm developments made during the past few decades, there are currently a large number of machine learning methods that can be performed for numerical regression, such as artificial neural network (ANN) [14] and support vector machine (SVM) [15]. Artificial neural network (ANN) is a proven powerful machine learning tool for numerical prediction [16,17], classification [18], and pattern recognition [19]. During recent years, it has been widely used in chemical [20,21], biological [22], medical [23], environmental [24], and engineering [17,25,26] applications.

In this communication, we aim to propose a quick estimation method of the concentration of indoor airborne culturable bacteria using ANN models. Our results show that with the simple inputs of indoor $PM_{2.5}$ and PM_{10}, temperature, relative humidity, and CO_2 concentration, the model trained from our experimental database with 249 data groups can effectively predict the concentration of indoor airborne culturable bacteria with relatively low root mean square errors (RMS errors).

2. Modeling Methods

The structure of a typical ANN consists of three different types of layers, including the input, hidden, and output layers. Each layer consists of a certain number of neurons. Each neuron in the input layer represents an independent variable, while the neurons in the output layer are the dependent variables. Each neuron interconnects with all the neurons in the adjacent layer(s). The training of an ANN is essentially the searching of optimal weights between all the pairs of neurons. Currently, there are a large number of ANN algorithms, such as back-propagation neural network (BPNN) [27–29], general regression neural network (GRNN) [17,22], and extreme learning machine (ELM) [30–32]. Though they have different weight calculation strategies, the basic principles and structures are similar. The introduction of a general schematic ANN structure can be seen in Reference [33]. In this study, we mainly used a GRNN as the ANN model for training. Compared to conventional ANN methods, GRNN is a state-of-the-art algorithm that has the advantages of fast training and fixed structure [17]. People do not need to compare varying numbers of neurons in the hidden layer during the training and testing, which saves time and computational cost. Details of the principle of a typical GRNN can be found in the work conducted by Specht [34].

To develop a model for estimating the concentration of indoor airborne culturable bacteria, we chose a series of independent variables that can be easily measured inside a building, including: (i) indoor $PM_{2.5}$, (ii) indoor PM_{10}, (iii) temperature, (iv) relative humidity, and (v) CO_2 concentration. All the data were measured in various buildings in Baoding, one of the cities that recently suffered from the most serious $PM_{2.5}$ pollutions in China. Descriptive statistics of the 249 measured data groups are shown in Table 1, which shows that the data ranges of all our measured variables are wide enough for a machine learning model training. Details on the measurements have been mentioned in our previous articles [8,11,35]. With the assumption that all these factors are correlated (or partially correlated) to the concentration of indoor airborne bacteria, these independent variables were considered as the inputs of the GRNN for model training. The measured bacterial concentration was then assigned as the output of the GRNN. Statistical analysis of the potential relationships between the independent variables and the dependent variable are shown in Section 3.1. For each training and testing process, the dataset was divided into two different subsets, the training and testing sets. Training sets were used for the data training, from which the GRNN model can "learn" via a "black-box" fitting process;

testing sets were used for the testing of the precision of the trained GRNN models. In our study, we used different percentages of training and testing sets for model developments and compared the precision and robustness of each model. The training and testing of each percentage were repeated 200 times. RMS errors were calculated from the testing results for comparison. Details on the training and testing results are shown in Section 3.2. For comparison, we also performed training and testing processes on a multilayer feedforward neural network (MLFN) [36] with varying numbers of hidden nodes (2–15) in a single hidden layer. The number of trials for each MLFN training was set as 100,000. All the modeling works were performed on a laptop (ThinkPad series, Lenovo, Morrisville, NC, USA).

Table 1. Descriptive statistics of variables for the 249 measured data groups.

Item	Indoor $PM_{2.5}$ ($\mu g/m^3$)	Indoor PM_{10} ($\mu g/m^3$)	Temperature (°C)	Relative Humidity (%)	CO_2 Concentration (ppm)	Bacterial Concentration (CFU/m^3)
Maximum	398	431	28	66	1789	3522
Minimum	18	36	12	18	491	122
Range	380	395	16	48	1298	3400
Average	157	198	20	31	881	877.3
Standard Deviation	76	91	3	10	204	596.6

Note: PM: particulate matter; CFU: colony forming units.

3. Results and Discussion

3.1. Statistical Analysis

Figure 1 shows the potential relationships between the measured bacterial concentration and the measured independent variables. It can be seen that the indoor $PM_{2.5}$ and PM_{10}, temperature, and relative humidity correlate moderately with the concentration of indoor airborne culturable bacteria. However, in terms of CO_2 concentration, the correlation is insignificant. Interestingly, the concentrations of indoor $PM_{2.5}$ and PM_{10} have less correlation with the bacterial concentration than with the indoor fungal concentration reported in a previous study [8], which indicates that the bacterial concentration is less dependent than the fungal concentration on these environmental indicators. This can also probably explain why there are always some data points that significantly deviate from the regression lines in Figure 1. Also, the correlation between relative humidity and the concentration of indoor airborne culturable bacteria is consistent with the previous studies [37]. All these results show that we cannot use a simple linear regression model with a single independent variable to predict the bacterial concentration, because the growth of bacterium usually depends on more than one environmental factor. Therefore, to more precisely predict the bacterial concentration, a non-linear fitting technique is necessary.
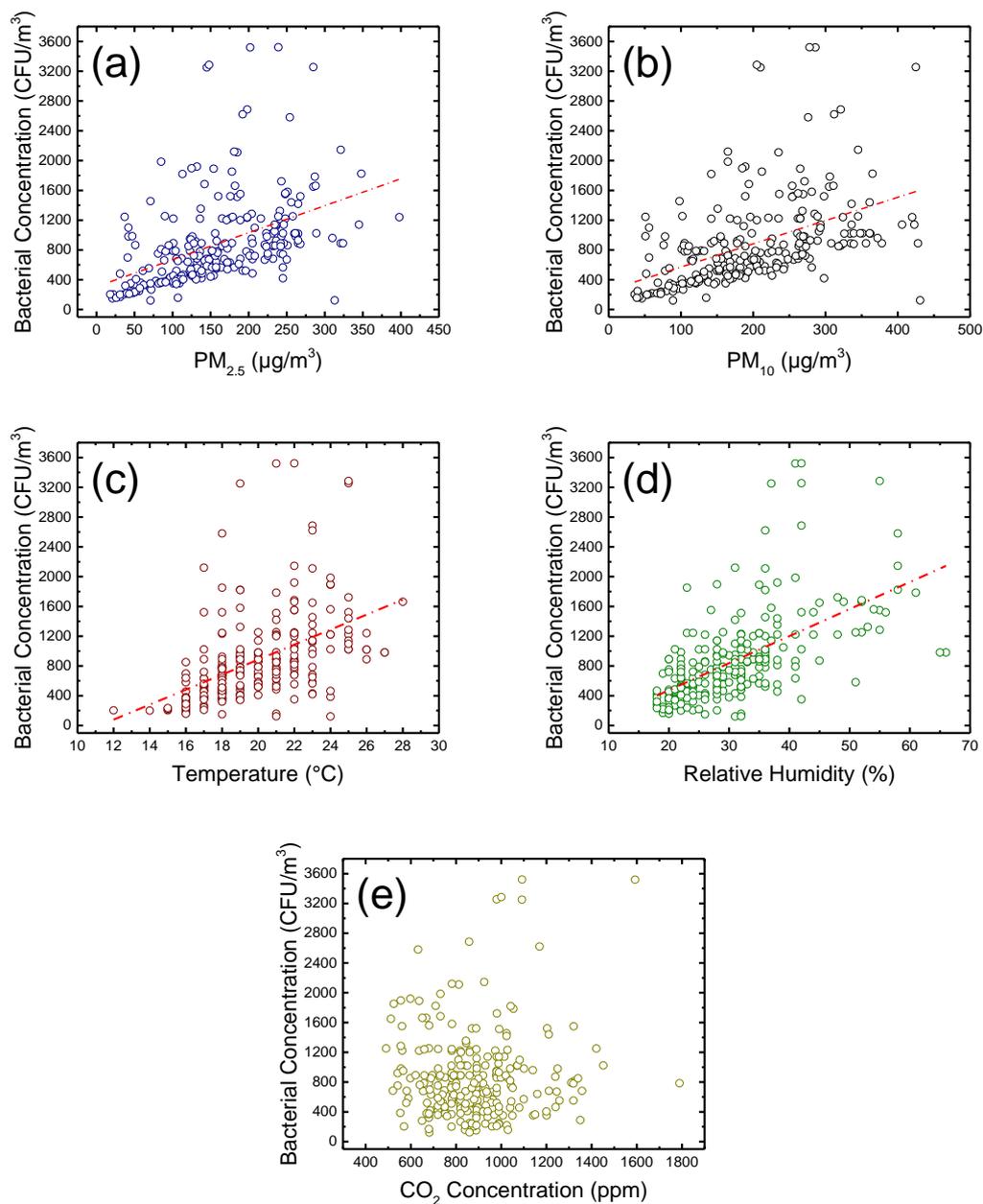
Figure 1. Relationships between the concentration of indoor airborne culturable bacteria vs. (**a**) indoor $PM_{2.5}$, (**b**) indoor PM_{10}, (**c**) temperature, (**d**) relative humidity, and (**e**) CO_2 concentration.

3.2. Machine Learning-Based Prediction

During the development of the GRNN, the training sets of 95%, 85%, 75%, 65%, and 55%, were respectively used. The average RMS errors (acquired from 200 repeated training and testing processes) are shown in Figure 2. It can be clearly seen that with a higher percentage of training set, the average RMS error in the testing process becomes lower. This is consistent with the principle of a non-linear fitting process: training with a larger dataset will lead to a lower risk of over-fitting. Due to the high economic cost for the bacterial concentration measurement in our study, we so far could only acquire 249 data groups. Though this database is large enough for a promising model training, for practical applications, it is always recommended that using a larger experimental database can lead to better predictive performance.
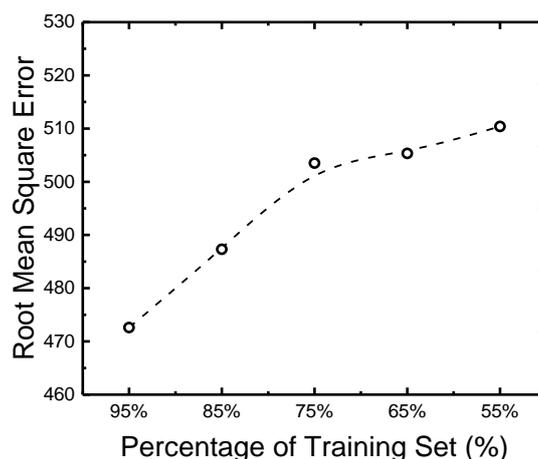
Figure 2. Average root mean square errors (RMS errors) of the testing processes with the training sets percentage varying from 55% to 95%. The training with each percentage was repeated 200 times.

Meanwhile, the performance of the GRNN was also compared with the multiple linear regression and MLFN models with varying numbers of hidden nodes (Table 2). The RMS errors (calculated from the testing set) shown in Table 2 were extracted from 15 repeated model training processes. It can be seen that the average RMS error of the GRNN is significantly lower than those of other models we developed. Though its minimum RMS error is higher than some other models, the range between the maximum and minimum RMS error of the GRNN is the smallest among all the models we developed. These results show that the GRNN not only has generally low average RMS error, but also has very good stability. For this reason, we conclude that with the training using our experimental database (Table 1), the GRNN outperforms the MLFN and multiple linear regression models.

Table 2. RMS errors of different predictive models. Multilayer feedforward neural networks (MLFNs) with different numbers of hidden neurons are represented as MLFN-*x*, where *x* represents the number of hidden neurons. The average RMS error of each model was acquired from 15 repeated training and testing processes.

Model	Average RMS Error (Testing)	Maximum RMS Error (Testing)	Minimum RMS Error (Testing)	Range
Linear Regression	488.51	742.40	183.63	558.77
GRNN	412.69	630.87	243.59	387.28
MLFN-*2*	501.51	842.10	191.61	650.49
MLFN-*3*	567.56	936.49	261.92	674.57
MLFN-*4*	551.14	762.22	238.96	523.26
MLFN-*5*	537.75	905.98	221.88	684.09
MLFN-*6*	618.37	1168.41	169.72	998.69
MLFN-*7*	613.02	1232.63	137.23	1095.40
MLFN-*8*	554.64	852.59	239.91	612.68
MLFN-*9*	617.65	873.16	198.35	674.81
MLFN-*10*	646.42	924.84	255.50	669.34
MLFN-*11*	614.87	1015.63	218.95	796.68
MLFN-*12*	644.37	1313.63	223.55	1090.08
MLFN-*13*	675.69	1033.43	279.78	753.65
MLFN-*14*	746.24	1304.57	321.40	983.17
MLFN-*15*	716.42	1761.50	250.64	1510.86

To show the training and testing results of the GRNN more clearly, we picked three typical modeling results for discussions, as shown in Figure 3 (a and b: training and testing sets of 95% and 5%, respectively; c and d: training and testing sets of 85% and 15%, respectively; e and f: training and testing sets of 75% and 25%, respectively). It can be seen that with relatively high percentages of training set (e.g., 95% in Figure 3a and 85% in Figure 3c), the training results look quite decent (except for several points that deviate from the diagonals), leading to good testing results, as shown in Figure 3b,d,

Int. J. Environ. Res. Public Health **2017**, *14*, 857

6 of 9

respectively. A plausible explanation for the discrete points is that the data groups with relatively high bacterial concentration are not enough to ensure a good fitting in a high concentration area. With our experimental conditions, it was found that the average bacterial concentration was around 877.3 colony forming units per m^3 (CFU/m^3) (Table 1), which was much lower than the maximum we have observed (3522 CFU/m^3, Table 1). We expect that, with a larger database in practical applications, this problem can be addressed properly. In terms of the training set with a relatively low percentage (e.g., a training set of 75%, as shown in Figure 3e), the training fails to precisely fit the data groups with a relatively large bacterial concentration, leading to weak prediction performance in the testing set (Figure 3f). We also found that using the percentages of the training set lower than 75% resulted in even worse predictive performance in their corresponding testing sets. These results, together with Figure 2, show that with a relatively large training set, the developed GRNN method is a promising ANN model for predicting the concentration of indoor airborne culturable bacteria with a few inputs of measurable indoor environmental indicators.
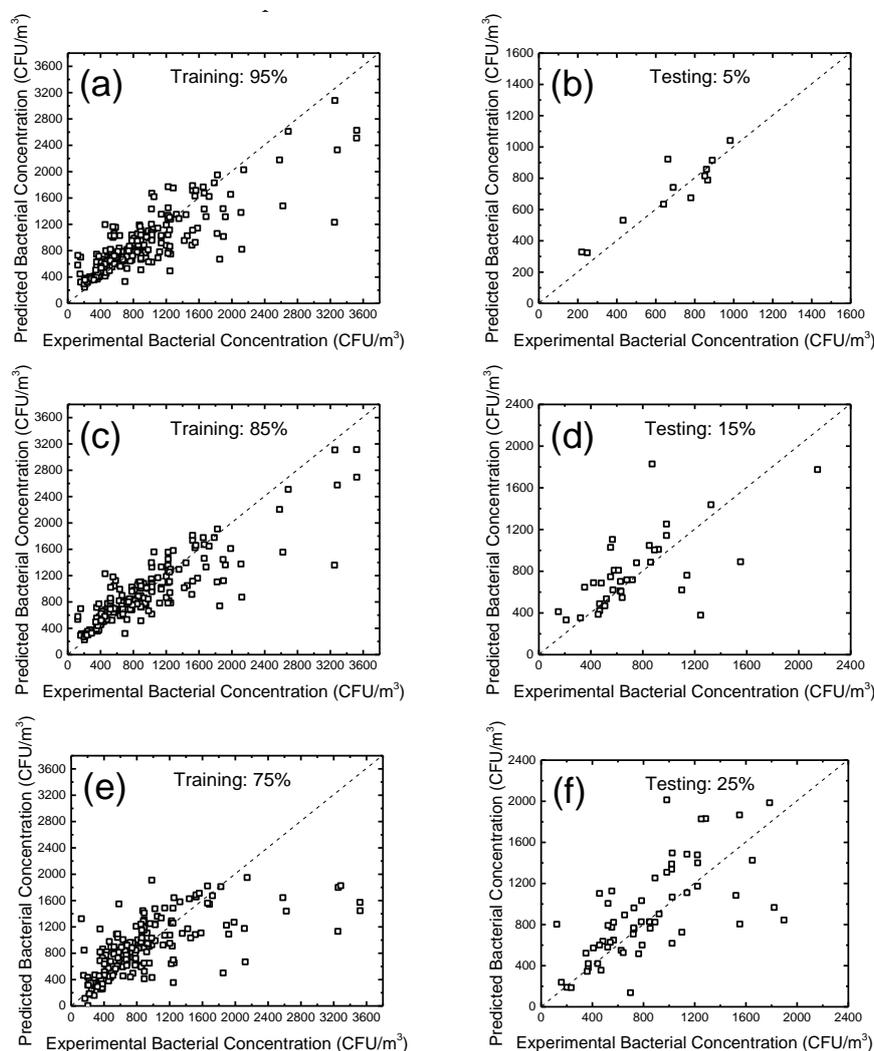

Figure 3. Typical comparative results of predicted bacterial concentration vs. experimental bacterial concentration in various percentages of training and testing sets. (**a**,**b**) Training and testing results with the training and testing percentages of 95% and 5%, respectively. (**c**,**d**) Training and testing results with the training and testing percentages of 85% and 15%, respectively. (**e**,**f**) Training and testing results with the training and testing percentages of 75% and 25%, respectively. Diagonals represent the function of $y = x$.

A good potential application of such a model is the real-time measurement of the concentration of indoor airborne culturable bacteria. Because all the independent variables used in this model can be acquired in a short period, the real-time bacterial concentration can also be evaluated once the values of the independent variables are inputted into the GRNN model. The flow chart of the proposed measurement method is shown in Figure 4. We expect that this quick evaluation method can dramatically shorten the measurement of bacterial concentration from days to seconds. In future study, we will continue to develop a user-friendly software that can be applicable for real measurements.
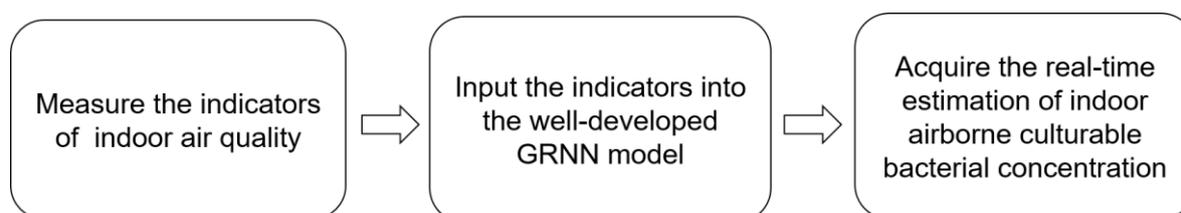

Figure 4. A proposed framework for the real-time estimation of the concentration of indoor airborne culturable bacteria.

4. Conclusions

In this communication, we have shown that a machine learning-based method can perform fast estimation of the concentration of indoor airborne culturable bacteria. We found that with the inputs of some indoor environmental indicators that can be easily measured (indoor $PM_{2.5}$ and PM_{10}, indoor temperature, relative humidity, and CO_2 concentration), a well-trained GRNN model can help to quickly acquire the estimated concentration. This novel estimation can dramatically reduce the measurement time from days to seconds, saving much time, economic cost, and manpower. We expect that this estimation method can also be applied to the quick measurement of indoor fungal and virus concentrations.

Author Contributions: Z.L. provided experimental data. H.L. did all the modeling and statistical studies and wrote the manuscript. G.C. participated in the discussions and revised the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Prussin, A.J.; Garcia, E.B.; Marr, L.C. Total concentrations of virus and bacteria in indoor and outdoor air. *Environ. Sci. Technol. Lett.* **2015**, *2*, 84–88. [CrossRef] [PubMed]
2. Mentese, S.; Mirici, N.A.; Otkun, M.T.; Bakar, C.; Palaz, E.; Tasdibi, D.; Cevizci, S.; Cotuker, O. Association between respiratory health and indoor air pollution exposure in Canakkale, Turkey. *Build. Environ.* **2015**, *93*, 72–83. [CrossRef]
3. Bragoszewska, E.; Mainka, A.; Pastuszka, J.S. Bacterial and fungal aerosols in rural nursery schools in Southern Poland. *Atmosphere (Basel)* **2016**, *7*. [CrossRef]
4. Hospodsky, D.; Qian, J.; Nazaroff, W.W.; Yamamoto, N.; Bibby, K.; Rismani-Yazdi, H.; Peccia, J. Human occupancy as a source of indoor airborne bacteria. *PLoS ONE* **2012**, *7*. [CrossRef] [PubMed]
5. Spendlove, J.C.; Fannin, K.F. Source, significance, and control of indoor microbial aerosols: Human health aspects. *Public Health Rep.* **1983**, *98*, 229–244. [PubMed]
6. Husman, T. Health effects of indoor-air microorganisms. *Scand. J. Work Environ. Health* **1996**, *22*, 5–13. [CrossRef] [PubMed]
7. Daisey, J.M.; Angell, W.J.; Apte, M.G. Indoor air quality, ventilation and health symptoms in schools: An analysis of existing information. *Indoor Air* **2003**, *13*, 53–64. [CrossRef] [PubMed]

8. Liu, Z.; Li, A.; Hu, Z.; Sun, H. Study on the potential relationships between indoor culturable fungi, particle load and children respiratory health in Xi'an, China. *Build. Environ.* **2014**. [CrossRef]

9. Lanthier-Veilleux, M.; Baron, G.; Généreux, M. Respiratory diseases in university students associated with exposure to residential dampness or mold. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1154. [CrossRef] [PubMed]

10. Cao, C.; Jiang, W.; Wang, B.; Fang, J.; Lang, J.; Tian, G.; Jiang, J.; Zhu, T.F. Inhalable microorganisms in Beijing's $PM_{2.5}$ and PM_{10} pollutants during a severe smog event. *Environ. Sci. Technol.* **2014**, *48*, 1499–1507. [CrossRef] [PubMed]

11. Liu, Z.; Zhu, Z.; Zhu, Y.; Xu, W.; Li, H. Investigation of dust loading and culturable microorganisms of HVAC systems in 24 office buildings in Beijing. *Energy Build.* **2015**, *103*, 166–174. [CrossRef]

12. Pham, T.D.; Lee, B.K. Feasibility of silver doped TiO_2 glass fiber photocatalyst under visible irradiation as an indoor air germicide. *Int. J. Environ. Res. Public Health* **2014**, *11*, 3271–3288. [CrossRef] [PubMed]

13. Rogawansamy, S.; Gaskin, S.; Taylor, M.; Pisaniello, D. An evaluation of antifungal agents for the treatment of fungal contamination in indoor air environments. *Int. J. Environ. Res. Public Health* **2015**, *12*, 6319–6332. [CrossRef] [PubMed]

14. Zhang, G.; Eddy Patuwo, B.; Y. Hu, M. Forecasting with artificial neural networks. *Int. J. Forecast.* **1998**, *14*, 35–62. [CrossRef]

15. Suykens, J.A.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]

16. Kalogirou, S.A.; Panteliou, S.; Dentsoras, A. Artificial neural networks used for the performance prediction of a thermosiphon solar water heater. *Renew. Energy* **1999**, *18*, 87–99. [CrossRef]

17. Liu, Z.; Li, H.; Zhang, X.; Jin, G.; Cheng, K. Novel method for measuring the heat collection rate and heat loss coefficient of water-in-glass evacuated tube solar water heaters based on artificial neural networks and support vector machine. *Energies* **2015**, *8*, 8814–8834. [CrossRef]

18. Hile, R.; Cova, T.J.; Aubrecht, C.; Kainz, W. Exploratory testing of an artificial neural network classification for enhancement of the social vulnerability index. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1774–1790. [CrossRef]

19. Zanardi, M.M.; Sarotti, A.M. GIAO C-H COSY Simulations merged with artificial neural networks pattern recognition analysis. Pushing the structural validation a step forward. *J. Org. Chem.* **2015**, *80*, 9371–9378. [CrossRef] [PubMed]

20. Perea, J.D.; Langner, S.; Salvador, M.; Kontos, J.; Jarvas, G.; Winkler, F.; Machui, F.; Görling, A.; Dallos, A.; Ameri, T.; et al. Combined computational approach based on density functional theory and artificial neural networks for predicting the solubility parameters of fullerenes. *J. Phys. Chem. B* **2016**, *120*, 4431–4438. [CrossRef] [PubMed]

21. Li, H.; Tang, X.; Wang, R.; Lin, F.; Liu, Z.; Cheng, K. Comparative study on theoretical and machine learning methods for acquiring compressed liquid densities of 1,1,1,2,3,3,3-heptafluoropropane (r227ea) via song and mason equation, support vector machine, and artificial neural networks. *Appl. Sci.* **2016**, *6*, 25. [CrossRef]

22. Chen, F.; Li, H.; Xu, Z.; Hou, S.; Yang, D. User-friendly optimization approach of fed-batch fermentation conditions for the production of iturin A using artificial neural networks and support vector machine. *Electron. J. Biotechnol.* **2015**, *18*, 273–280. [CrossRef]

23. Khan, J.; Wei, J.S.; Ringnér, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **2001**, *7*, 673–679. [CrossRef] [PubMed]

24. Hsu, K.; Gupta, H.V.; Sorooshian, S. Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **1995**, *31*, 2517–2530. [CrossRef]

25. Liu, Z.; Li, H.; Liu, K.; Yu, H.; Cheng, K. Design of high-performance water-in-glass evacuated tube solar water heaters by a high-throughput screening based on machine learning: A combined modeling and experimental study. *Sol. Energy* **2017**, *142*, 61–67. [CrossRef]

26. Sanaye, S.; Asgari, H. Thermal modeling of gas engine driven air to water heat pump systems in heating mode using genetic algorithm and Artificial Neural Network methods. *Int. J. Refrig.* **2013**, *36*, 2262–2277. [CrossRef]

27. Wang, Z.; Wang, F.; Su, S. Solar irradiance short-term prediction model based on BP neural network. *Energy Procedia* **2011**, *12*, 488–494. [CrossRef]

Int. J. Environ. Res. Public Health **2017**, *14*, 857

9 of 9

28. Wang, J.Z.; Wang, J.J.; Zhang, Z.G.; Guo, S.P. Forecasting stock indices with back propagation neural network. *Expert Syst. Appl.* **2011**, *38*, 14346–14355. [CrossRef]

29. Peng, H.; Ling, X. Optimal design approach for the plate-fin heat exchangers using neural networks cooperated with genetic algorithms. *Appl. Therm. Eng.* **2008**, *28*, 642–650. [CrossRef]

30. Huang, G.-B.; Zhu, Q.; Siew, C.; Ã, G.H.; Zhu, Q.; Siew, C.; Huang, G.-B.; Zhu, Q.; Siew, C. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]

31. Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2012**, *42*, 513–529. [CrossRef] [PubMed]

32. Liu, Z.; Li, H.; Tang, X.; Zhang, X.; Lin, F.; Cheng, K. Extreme learning machine: A new alternative for measuring heat collection rate and heat loss coefficient of water-in-glass evacuated tube solar water heaters. *SpringerPlus* **2016**. [CrossRef] [PubMed]

33. Kalogirou, S. Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy* **2000**, *25*, 479–491. [CrossRef]

34. Specht, D.F. A general regression neural network. *Neural Netw. IEEE Trans.* **1991**, *2*, 568–576. [CrossRef] [PubMed]

35. Li, A.; Liu, Z.; Zhu, X.; Liu, Y.; Wang, Q. The effect of air-conditioning parameters and deposition dust on microbial growth in supply air ducts. *Energy Build.* **2010**, *42*, 449–454. [CrossRef]

36. Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to multi-layer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 43–62. [CrossRef]

37. Arundel, A.V.; Sterling, E.M.; Biggin, J.H.; Sterling, T.D. Indirect health effects of relative humidity in indoor environments. *Environ. Health Perspect.* **1986**, *65*, 351–361. [CrossRef] [PubMed]