Received September 1, 2019, accepted September 9, 2019, date of publication September 11, 2019, date of current version September 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940910

Environmental and Human Data-Driven Model Based on Machine Learning for Prediction of Human Comfort

FUBING MAO[1,2,3], XIN ZHOU[1,2], AND YING SONG[1,2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798
[3]School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China

Corresponding author: Xin Zhou (enoche.chow@gmail.com)

This work was supported in part by the Delta-NTU Corporate Laboratory for Cyber-Physical Systems with funding support from Delta Electronics, Inc., and in part by the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.

ABSTRACT Occupants' comfort level has a strong correlation with health problems. Providing a comfortable environment for the occupants will bring the benefits of improved health. To achieve this goal, it is necessary to have a reliable human comfort model for predicting the occupants' comfort level and subsequently controlling the involved comfort condition. However, the comfort perception of occupants is subjective. There is a lack of objective indices for measuring comfort level. Furthermore, human comfort is affected by various environmental factors. Such situations make it difficult to set up a model for measuring human comfort. To address the challenges, we use Blood Pulse Wave (BPW) as an objective comfort index and adopt a data-driven approach to predict human comfort level based on data including both environmental factors and human factors. We propose a framework for collecting the data followed by investigating the relationship between the factors with the purpose of building a scalable comfort model. In consideration of the nonlinear relationship present in the dataset, we opt for support vector regression with radial basis function (SVR-RBF) algorithm to establish the comfort model. To validate the predication performance of this method, we have applied the other six popular machine learning models on the same dataset. In order to choose an optimal model, we apply the holdout method and k-folder cross-validation method together with the grid search. The comparison results show that the SVR-RBF has the best performance for comfort prediction according to the mean squared error, mean absolute error and R-squared score.

INDEX TERMS Machine learning, human comfort prediction, support vector regression.

I. INTRODUCTION

Human comfort plays a key role in individuals' health and wellbeing and also has great impact on their work efficiency. Nowadays 90% of people spend most of their time in buildings which are generally built for individuals' working and living [1], [2]. Comfortable indoor conditions would considerably improve the occupants' health, well-being and work performance. Therefore, there has been an increasing demand for the improvement of indoor comfort. To achieve this objective, a reliable human comfort model is required to be established to delineate the relationship between external factors and human comfort. The validated model can be used to predict the occupants' comfort level so as to adjust the external factors to suit their comfort needs.

On the other hand, development of human comfort model would contribute to the improved energy efficiency of buildings. Currently buildings all around the world consume a significant amount of energy which accounts for about one third of the total energy consumption [2]. The majority of the energy consumption in buildings is resulted from Heating Ventilation and Air Conditioning (HVAC) systems [2]. To activate HVAC systems according to occupants' comfort needs instead of based on fixed criteria would definitely help reduce energy consumption.

Comfort is usually characterized as a lack of hardship or a sense of psychology or physical ease [3]–[7]. It involves some dimensions of the physical environment and depends

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Cheng.

on occupants' physical and psychological factors. It can be assessed in terms of thermal comfort [3], [8], respiratory comfort [4], [9], visual comfort [10], [11] and acoustic comfort [12], [13]. Comfort is a subjective term. Thus, it is difficult to quantify it. In order to represent or predict the comfort level of individuals, firstly a common indicator is required. Previous studies have proposed some thermal comfort indicators. For instance, the Predicted Mean Vote (PMV), the Predicted Percentage of Dissatisfied (PPD), the percentage of local dissatisfaction (PD) and the extended adaptive version aPMV are adopted as indices [14]–[17]. However, these indicators involving occupants' votes are subjective and cannot accurately reflect human comfort or thermal comfort.

A reliable human comfort model with high accuracy to predict or improve human comfort is also demanded and necessary. Basically, previous studies mainly carry out research on thermal comfort models. These models are generally classified into three categories: Regression [18]–[24], Classification [25]–[27] and other models [14]–[17], [28]–[30]. Regression models primarily use the PMV index as the subjective comfort index to predict continuous comfort levels. Classification models only predict discrete comfort states. Other models such as PMV model and PDD model have low prediction accuracy about 41.68-65.5% as reported in [31] and are difficult to handle diverse factors and scenarios.

To address the above-mentioned problems, in this paper we use blood pulse wave (BPW) [32] as a proxy for comfort. Blood Pulse Wave is an objective composite index which qualitatively measures how the blood pulsate over time. In order to continuously monitor individuals' comfort levels and take into account the nonlinear relationship investigated based on our dataset, we adopt an environmental and human data-driven model namely support vector regression with radial basis function (SVR-RBF) to predict human comfort [33], [34]. SVR-RBF is selected based on our consideration that this model is able to learn complex patterns and can deal with nonlinear relationship and mass data well. In addition, this model is scalable, which has the ability to incorporate multiple factors. The model is set up based on input environmental data collected from physical sensor nodes and output human data extracted from medical-level wearable sensors. This model has been compared with other popular machine learning models according to the criteria of Mean Squared Error (MSE) [35], Mean Absolute Error (MAE) [36] and Adjusted R-squared score [37].

The main contributions of our paper are as follows.

- We present a detailed framework to obtain the environmental and human data.
- We use Blood Pulse Wave (BPW) as an objective comfort index.
- Prior to building the comfort model, we visualize the data and explore the relationship between the environmental (input) factors and the human (output) factor from one dimension to multiple dimensions so as to guide the selection of suitable model.
- Based on the analysis of the data, we adopt the Support Vector Regression with radial basis function kernel for human comfort prediction.
- To further validate the performance of the SVR-RBF model, we compare its analysis result with those of other six popular regression models.

The remainder of the paper is organized as follows. Section II gives a brief introduction to previous studies related to human comfort. Section III presents a framework for data collection and the adopted prediction model. Section IV introduces the methodology and model optimization. Experiment results are described in Section V. We discuss our findings and current limitations in Section VI. Section VII concludes the paper.

II. RELATED WORK

A number of studies related to human comfort have been conducted and previous research mainly focus on modelling thermal comfort [28]. As we mentioned above, our work focuses on the regression model since we aim to continuously monitor the comfort levels of individuals. Thus, we primarily review the regression models [18]–[24] and other continuous prediction models [14]–[17], [28]–[30].

Regression models: A two-stage empirical PMV regression model was presented and it considered the architectural parameters and control variables [18]. A regression approach was proposed in [19] to analyse human thermal comfort. An artificial neural network (ANN) was built in [20] to analyse the relationship between the estimated PMV index and the input parameters. The study presented in [21] adopted an autoregressive neural network to establish thermal comfort models for controlling heating settings. A recursive least square estimation approach was applied to learn the thermal comfort profile [22]. The study presented in [23] developed a thermal comfort model with a kernel based method which was used to learn occupants' thermal comfort profile. Gaussian Process (GP) regression was proposed in [24] to extract subjects' thermal preferences. However, these models mainly concentrate on thermal comfort and involve human thermal perception votes which is subjective in nature.

Other continuous prediction models: The PMV [28], PPD [28], [29], and extended PMV models which were in the form of equations were proposed to evaluate thermal comfort focusing on specific factors [14]–[17]. A fuzzy rule-based model was presented to set up a predictive model for thermal comfort [30]. However, the PMV model and its extended models have low prediction accuracy [14], [15] and are difficult to handle multiple factors. It is tough to tune the controller parameters of the fuzzy rule-based models [30].

III. DATA COLLECTION AND THE ADOPTED PREDICTION MODEL

A. IoT BASED FRAMEWORK FOR DATA COLLECTION

In this section, we briefly describe the data information of our dataset and introduce our adopted machine learning based prediction model. We also need to protect user privacy and

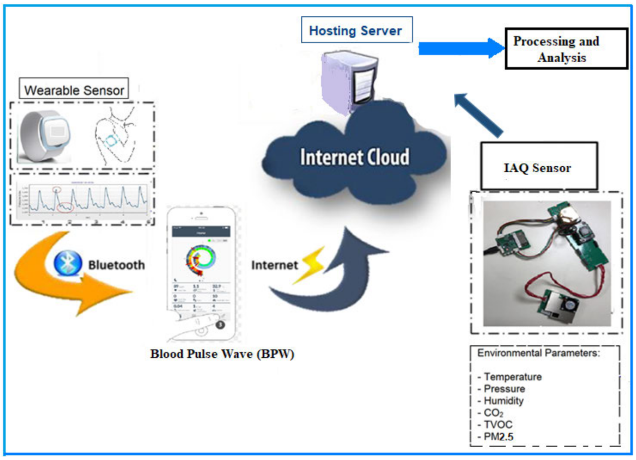


FIGURE 1. IoT sensor based framework for data collection.

keep academic ethics of our work. Thus, we briefly present how the data is processed.

We study integrating a sensor network into an office workplace, which can also compile the collected data and compute the comfort level in the given context. We propose an Internet of thing (IoT) sensor based framework shown in Fig. 1 to collect data. The design introduces the equipment including portable air conditioner, heater, humidifier and air purifier to control the environmental parameters. This is for widening the range of the value of individual environmental parameters which serve as the inputs to the comfort model. Based on the sensor data, we aim at designing an intelligent human comfort model to automatically estimate indoor parameters for occupants' comfort. For developing the human comfort model, the wearable sensor data monitoring human vital signs are collected in correlation with the environmental sensor data. In our work, the wearable sensors collect human data per second and the indoor air quality (IAQ) sensors acquire environmental data every minute. Both the environmental and human data have the time stamps for each data record. To achieve data matching between the environmental data and human vital sign data, a load sensor is adopted to detect the presence of a human subject at his/her workstation and the matched data must be acquired at the same time stamp. We finally obtain the matched data including environmental data and human data per minute.

We describe the detailed steps for data collection in Fig. 1. Firstly, IAQ sensors are used to measure the temperature, air pressure, humidity, CO_2, TVOC and PM2.5 of the environment. The sensor data is transmitted wirelessly via a Wi-Fi or Bluetooth Low Energy (BLE) to a server. These sensors are placed on human volunteer's workstation to measure the exact environment parameters surrounding him/her. Then, a wearable device is used for the continuous monitoring of the certain vital sign of individual human volunteer. It is an armband and worn on the upper arm. It sends captured data via Bluetooth wireless connection to a smartphone from which it is then transmitted to cloud storage. Currently the device measures the following vital sign: Blood Pulse Wave (BPW).

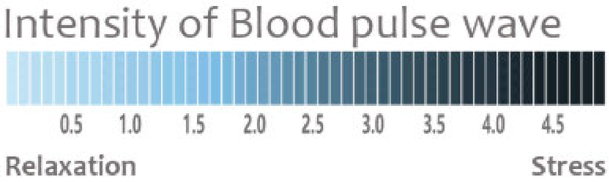


FIGURE 2. Stress indication-"blood pulse wave (BPW)" reflects the stress (happiness) to some extent. Colour shows low blood pulse wave (light grey/blue) indicates a state of relaxation (low stress level).

Finally, we process and analyze the data from the servers. Specifically, we explore the distribution of each variable and relationship between environmental factors and human factors. We also standardize and filter data for the model. In our experiments, we have six volunteers and conduct the experiments for 33 days.

B. ADOPTED MACHINE LEARNING BASED PREDICTION MODEL

Based on the data from environmental sensors and wearable sensors, we find the relationship between them. More specifically, we take the six environment parameters including temperature, pressure, humidity, carbon dioxide (CO_2), total volatile organic compounds (TVOC) and particulate matter (PM2.5) [38] as input variables and take the blood pulse wave (BPW) [32] as the output/target variable. Previous studies indicate that there is no common/popular and deterministic indicator to reflect whether people are comfortable or not. However, the Biovotion company has tested and done experiments on BPW and shown that it can reflect stress (comfort) of people to some extent [32]. Furthermore, the quality of the BPW has been proven by an increasing number of papers published independently from different renown resources such as leading universities, pharma companies, government agencies and so on [32], [39]. Thus, in our work we take the BPW as the objective comfort index to reflect people's comfort and the range of the BPW is $[0, 5.1]$. The relationship of the stress and the value of BPW is shown in Fig. 2 [32].

Since the BPW has continuous values and the relationship between input factors and output factors is nonlinear obtained in Section IV, our adopted environmental and human data-driven model is based on support vector regression with radial basis function kernel (SVR-RBF) [40] for human comfort prediction. The approach belongs to regression and supervised machine learning algorithms and can predict continuous values. The overview of the approach is shown in Fig. 3. We take the environmental factors as the input data, and the BPW as the output data. We take all the inputs as important and train and test some popular regression models. We select the regressor with the minimum mean squared error (MSE) [35] and the minimum mean absolute error (MAE) [36] and the largest R-squared score [37] to predict comfort level.

C. SUPPORT VECTOR REGRESSION

The support vector regression (SVR) technique is derived from the support vector machine (SVM) which was invented

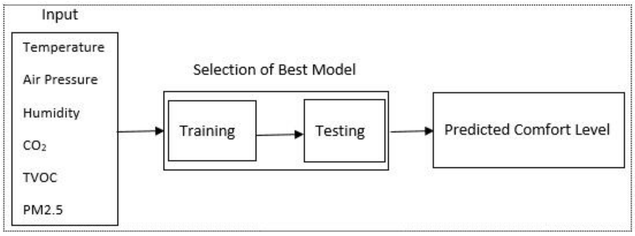


FIGURE 3. Overview of our comfort prediction approach.

by Vapnik and his co-workers in 1995 [41] and based on statistical learning theory and Vapnik-Cervonenkis (VC) theory [42]. The fundamental idea of SVM is to map the input data into a high-dimensional feature spaces using nonlinear mapping, and then a linear problem is acquired in the feature space. We briefly introduce an overview of the SVR and the more detailed descriptions of the theory can be found in [42], [43].

Given a set of training data with N data points $(x_1, y_1), \ldots, (x_N, y_N)$, where N is the size of the training data and each $x_i \in R^n$ represents the input sample in the input space and has a corresponding target value $y_i \in R(1 \leq i \leq N)$. SVR aims to determine a function $f(x)$ to fit the data accurately such that the function has at most ϵ $(\epsilon \geq 0)$ deviations from the actually acquired data for all the input data. The basic SVR function adopts the following form.

$$f(x) =< w, \phi(x) > +b \tag{1}$$

where $w \in R^n$ is the coefficient (weight) vector, $b \in R$ is an offset scalar, <> is the dot product, ϕ represents a nonlinear transformation from R^n to high-dimensional space, $\{\phi(x_i)\}_{i=1}^{N}$ denotes the high-dimensional feature spaces. We focus on finding the values of w and b to minimize the regression risk function $R_{regression}(f)$ in (2).

$$R_{regression}(f) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\Gamma_\epsilon(f(x_i), y_i) \tag{2}$$

The first item $\frac{1}{2}\|w\|^2$ is the regularized term which is served as a flatness measurement of $f(x)$. C is a user defined constant and can be used to control the weights to minimize the error. The definition of the ϵ-insensitive loss function is given as follows.

$$\Gamma_\epsilon(f(x_i), y_i) = \begin{cases} 0 & |y_i - f(x_i)| \leq \epsilon \\ |y_i - f(x_i)| - \epsilon & otherwise \end{cases} \tag{3}$$

ϵ used to fit the training data is a user defined insensitive bound with non-negative value. As we mentioned above, SVR aims to minimize the function $R_{regression}(f)$ in (2) with the constraint (3). By introducing the slack variables ξ_i and ξ_i^*, the problem can be written as

$$minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{4}$$

$$subject\ to \begin{cases} y_i - < w, \phi(x) > -b \leq \epsilon + \xi_i & i = 1, \ldots, N \\ < w, \phi(x) > +b - y_i \leq \epsilon + \xi_i^* & i = 1, \ldots, N \\ \xi_i, \xi_i^* \geq 0 & i = 1, \ldots, N \end{cases} \tag{5}$$

The constrained optimization problem can be solved by adopting the lagrange multiplier techniques. The problem can be transformed as the following problem.

$$\begin{aligned} &L(w, \xi, \xi^*, \alpha, \alpha^*, \lambda, \lambda^*) \\ &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ &\quad - \sum_{i=1}^{N}\alpha_i(\epsilon + \xi_i - y_i + < w, \phi(x) > +b) \\ &\quad - \sum_{i=1}^{N}\alpha_i^*(\epsilon + \xi_i^* + y_i - < w, \phi(x) > -b) \\ &\quad - \sum_{i=1}^{N}(\lambda_i\xi_i + \lambda_i^*\xi_i^*) \end{aligned} \tag{6}$$

Here, L is the Lagrangian and $\alpha, \alpha^*, \lambda, \lambda^*$ are Lagrange multipliers. For finding the optimal solution, the partial derivatives of L with respect to primary variables w, b, ξ, ξ^* are 0 and they are given as below.

$$\begin{aligned} \frac{\partial L}{\partial w} &= 0 \Rightarrow w = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i) \\ \frac{\partial L}{\partial b} &= 0 \Rightarrow \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ \frac{\partial L}{\partial \xi_i} &= 0 \Rightarrow \lambda_i = C - \alpha_i \\ \frac{\partial L}{\partial \xi_i^*} &= 0 \Rightarrow \lambda_i^* = C - \alpha_i^* \end{aligned} \tag{7}$$

Thus, we obtain the following equations.

$$w = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i) \tag{8}$$

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i).\phi(x) + b \tag{9}$$

The extreme value problem can be transformed into the following problem.

$$f(x) = \sum_{i=1}^{N}\beta_i < \phi(x_i), \phi(x) > +b \tag{10}$$

β_i is the coefficient corresponding to each (x_i, y_i). In SVR, we can obtain the same solution by using all the training data or only the support vectors.

A kernel function can be applied to estimate the inner product in the feature space which allows the inner product to be implemented in high-dimensional feature space by adopting low-dimensional space input data while the transformation function is unknown. We can rewrite the above equation (10) by using the kernel function $K(x_i, x)$ (x_i and x are two samples). Here, the kernel function must meet the Mercer's condition [44].

$$f(x) = \sum_{i=1}^{N} \beta_i K(x_i, x) + b \tag{11}$$

In our work, we use the RBF kernel shown in (12) to model the environmental factors as well as the human factor. σ is a free parameter and the x and x' are any two samples.

$$K(x, x') = exp(-\frac{\|x - x'\|^2}{2\sigma^2}) \tag{12}$$

When the data including the environmental data (Temperature, Air Pressure, Humidity, CO_2, TVOC, PM2.5) and human data (BPW) is obtained, the environmental data is regarded as the input and the human data is considered to be the output. We take the obtained environmental data and human data as the training data and introduce them into the (11) to acquire the values of w and b. Finally, we obtain the model. If we get a new environmental data, we use the obtained model to predict comfort value.

D. ACADEMIC ETHICS CONSIDERATION

When we collect user data, we need to take the academic ethics seriously. We have signed the ethic agreement form. We use a series of steps to protect users' privacy involved in our dataset and keep the ethics. First, all collected raw data are stored in the protected cloud server. Second, we clean and combine the data depending on the user id, device id and the time. Finally, we delete all the personal information related to users in the dataset and keep the data anonymous. Thus, we get the dataset which only covers statistical information of the users and environment information. We cannot trace the actual users at all.

IV. METHODOLOGY AND MODEL OPTIMIZATION

A. INPUT AND OUTPUT FACTORS

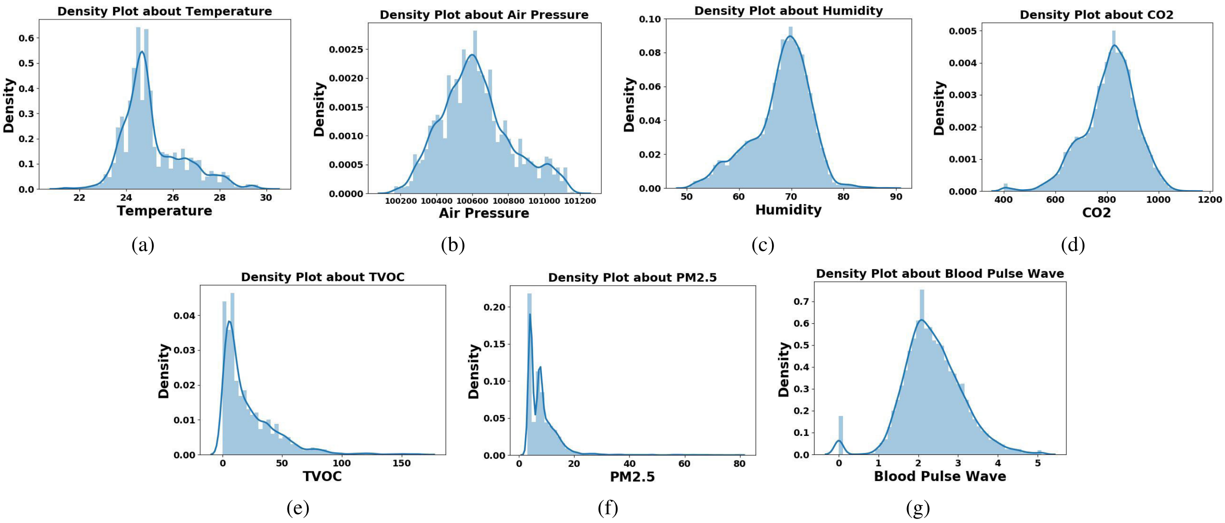
We first introduce the input factors (parameters): Temperature, Pressure, Humidity, CO_2, TVOC and PM2.5 [38] and the output factor Blood Pulse Wave (BPW) [32]. The reasons for choosing these six environmental factors to measure the comfort are given as follows.

- The factors are considered in the indoor environmental quality (IEQ) which has significant impact on occupant comfort, health, and productivity [45], [46]. Furthermore, they are the commonly used parameters that can be measured by the commercially available IEQ sensors.
- These factors can be measured easily from an indoor environment and are representative. For example, Kansas State University developed an empirical equation which expressed the Predicted Mean Vote (PMV) index [18], [47] for measuring the thermal comfort. The equations are only related to temperature and partial vapor pressure. It has been adopted by ASHRAE [48] which is a professional association which makes thermal comfort standards and guidelines.
- Our model is designed for real-time control systems. However, the traditional Fanger's model [18], [47] is not suitable for real-time control purpose due to its complex nature and the difficulty in acquiring certain input parameters. For example, it is impractical to get human subjects wearing a sensor all the time for measuring their activity and clothing level. Besides, previous studies have reported that the traditional Fanger's model has low prediction accuracy about 41.68-65.5% which is far from good (the best accuracy of the model is 100%) [31].

To obtain a quick impression about the factors, we plot the distribution of each factor in Fig. 4. In our work, we conduct all the experiments in the indoor office room.

- **Temperature**: It is a quantity to express hot and cold. It can make occupants feel comfortable and have a restful sleep. In our work, it represents the indoor temperature. We plot the distribution of this quantity in Fig. 4(a), it shows a normal distribution with a mean value at 25.1.
- **Air (barometric) Pressure**: It is the pressure within the atmosphere of Earth. It can contribute to understand why arthritic pains happen. It represents the indoor air pressure. Fig. 4(b) shows its distribution and it has a normal distribution and its mean value is 100622.
- **Humidity**: It is the amount of water vapour present in air. It can serve to minimize moisture. Here, it represents the indoor humidity. We also plot its distribution in Fig. 4(c). It is under a normal distribution and its mean value is 68.
- **Carbon Dioxide (CO_2)**: It is an atmospheric gas in the atmosphere. It can affect people's mind. It represents the indoor carbon dioxide. Fig. 4(d) shows it also has a normal distribution and its mean value is 811.
- **Total Volatile Organic Compounds (TVOC)**: VOC are gases emitted by different types of solids and liquids. It can be useful to know which chemical and products to keep out. It has a beta distribution shown in Fig. 4(e).
- **Particulate Matter (PM2.5)**: It is microscopic or liquid matter found in the atmosphere of Earth with a diameter less than 2.5 micrometres. It helps prevent dust build-up before infections occur. It also has a beta distribution shown in Fig. 4(f).
- **Blood Pulse Wave (BPW)**: When the heart contracts, blood is ejected generating a pulse wave that travels through the circulatory system. BPW is a measure of the wave and describe the rhythmicity and shape of the wave. It is a powerful proxy indicator) for stress (happiness). It basically has a normal distribution with a mean value at 2.37 shown in Fig. 4(g). It has a slight fluctuate around zero. The reason is that BPW is a human factor and it can be changed abruptly by many factors such as environmental factors, psychological factors and so on.

We also explore the relationships between each two factors and their pairwise relationships are shown in Fig. 5. Figures in



(a) (b) (c) (d)

(e) (f) (g)

FIGURE 4. **Univariate distribution: seven factors (X axis) and density (Y axis).**

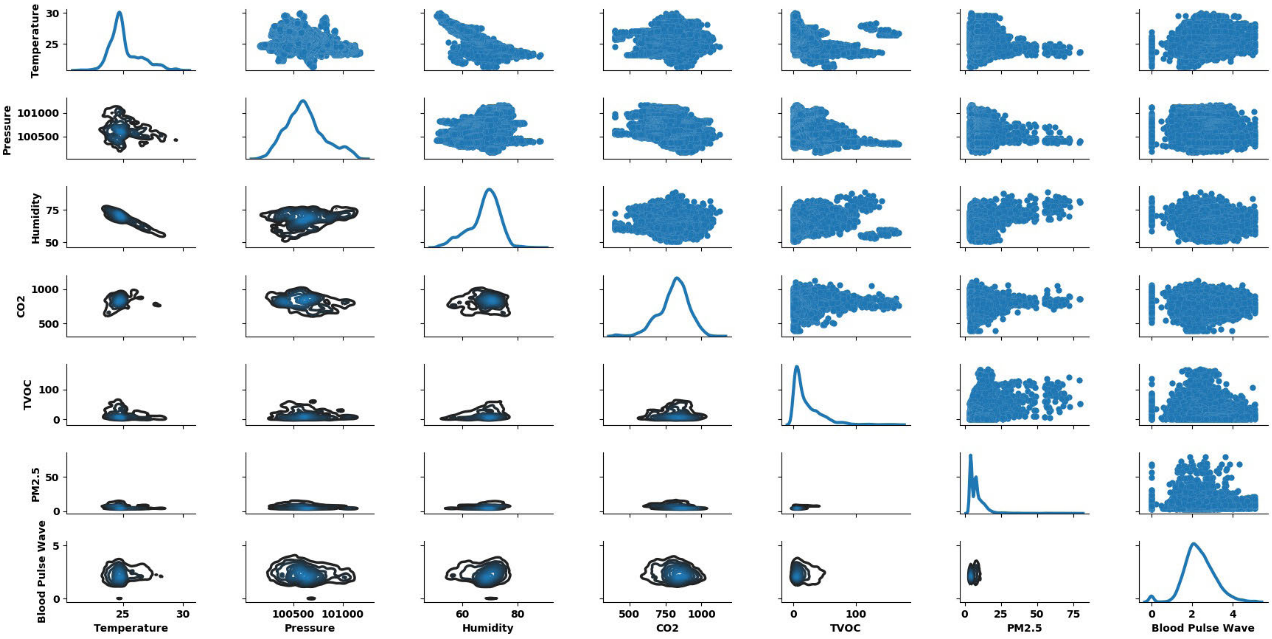
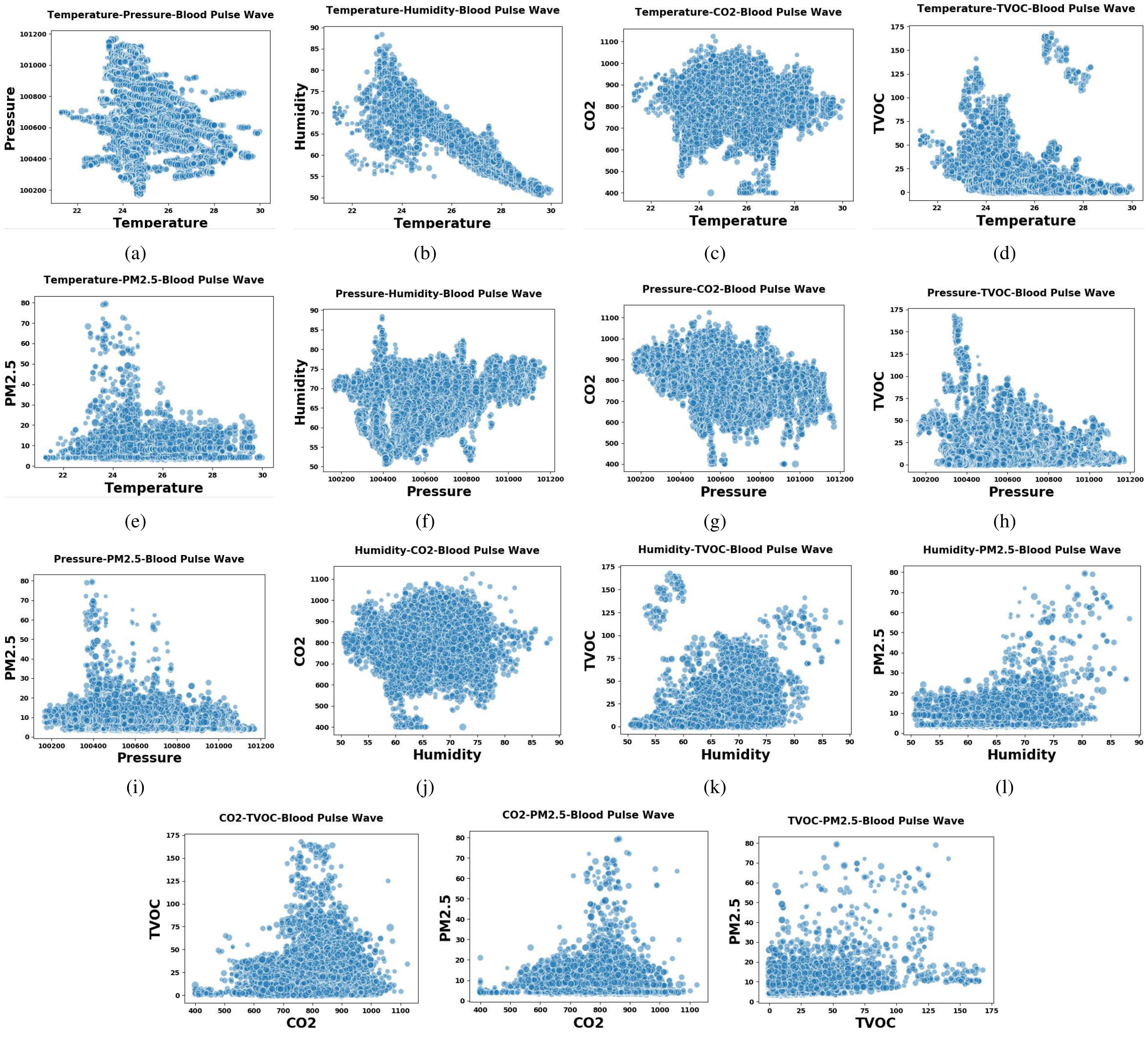


FIGURE 5. **The scatterplot to visually assess the nature of association between two factors.**

each row represent the relationship between the variable of this row and other variables including itself. Figures on the diagonal show the marginal distribution of each variable. Specifically, we apply kernel density estimation (KDE), a nonparametric technique for density estimation, to estimate multivariate densities on the lower triangle and univariate density on the diagonal. For the upper triangle, we draw a scatter plot. The upper and lower triangles are mirrored along the diagonal. The figures on the lower triangle show the contour plot of each bivariate density and the contour lines define regions of probability density from high (inner circle) to low (outer circle). Color is the probability density at each point and the regions with darker color has higher densities than other regions. Fig. 5 shows that PM2.5 is close to have a linear relationship with the other environmental (input) factors and the humidity is close to have a linear relationship with the temperature. For other factors, they have a nonlinear relationship. For the diagonal, the figures have the same properties as we described in the previous paragraph. With respect to the upper triangle, the coordinates of each point are defined by the two variables and each point is denoted by a filled circle. The plot show the distribution of the values of the points. The denser the region, the more points it has. For example, the figure in the 4th row (counted from top to



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o)

FIGURE 6. Three-dimensional graph: two environment factors with the human (output) factor.

bottom) representing CO_2 and 6th column (counted from left to right) denoting PM2.5 reflects the relationship between CO_2 and PM2.5. More points are located in the region (CO_2: 500-1000, PM2.5: 3-28) and the two variables are more likely have the nonlinear relationship.

We then explore the relationships between every two environment (input) factors and the output factor (BPW) and their relationship are shown in Fig. 6. The point in the figures represents the value of the BPW. The larger the value, the big the size of the point. The figures from Fig. 6(a) to Fig. 6(o) shows that every two environmental factors cannot clearly distinguish the BPW values which also showed they have the nonlinear relationship.

Finally, we explore the relationship between all the input environmental factors and the output factor (BPW), and the relationship is shown in Fig. 7. We first use the t-SNE [49], a tool for visualizing high-dimensional data, to fit the six environmental factors and the human factor into a two-dimensional embedded space. T-SNE runs at perplexity 50 which is a recommended parameter and the plot is made with 5000 iterations which is generally enough for convergence. Next, we study the relationship between the two-dimensional factors and the output factor (BPW). Since each original item of the data containing seven factors (before transforming) corresponds to an unique value of BPW, we can use the color to mark each item denoting the new point in the two-dimensional embedded space and the larger the BPW value, the darker blue the point. If all the points in the embedded space are well separated, we may understand their relationship better. Fig. 7 shows that the different values of BPW are irregularly distributed in the two-dimensional embedded space, which means that the

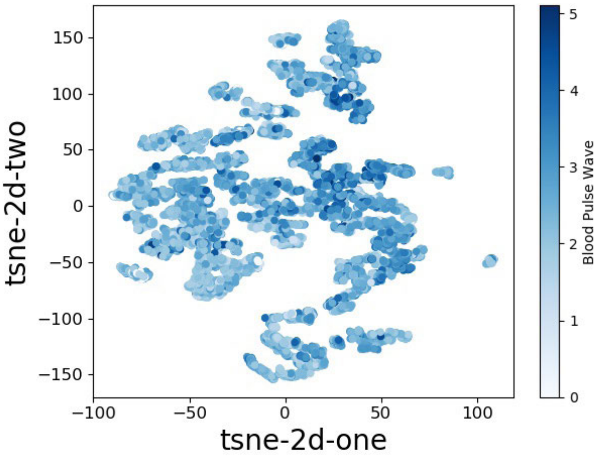


FIGURE 7. T-SNE results: Two-dimensional factors with the output BPW.

environmental factors and the output factor seem to be a nonlinear relationship.

B. DIFFERENT MACHINE LEARNING MODELS

We tested different models on the same dataset. The different models [33], [34] are briefly introduced as follows.

(1) **Linear Regression (LinearR)**: It models relationships between independent variables (input) and a dependent variable (output) using linear predictor functions. One case occurs when the two (or more) of the input variables are very strongly correlated. In such cases, it may make a large variance in the final parameter estimates.

(2) **Ridge Regression with Built-In Cross-validation (RidgeRCV)**: It performs generalized cross-validation. It is biased and accepts little bias to reduce the mean squared error and variance, and make the prediction more accurate, which has more stable solutions.

(3) **Bayesian Ridge Regression (BayesianRR)**: It estimates a probabilistic model of the regression problem.

(4) **Linear Support Vector Regression (LinearSVR)**: It sets a margin of tolerance (epsilon) and minimizes error, individualizing the hyperplane which maximizes the margin.

(5) **Multi-layer Perceptron Regressor (MultiLPR)**: It is a type of artificial neural network and each node is a neuron using a nonlinear activation function. It can distinguish data which is non-linear separated.

(6) **SVR with Kernel Function (RBF) (SVR-RBF)**: It is the support vector regression with radial basis function kernel which maps a lower dimensional data to a higher dimensional data.

(7) **Kernel Ridge Regression (RBF) (KernelRidge-RBF)**: It combines ridge regression with radial basis function kernel.

A comparison on these models is given as follows [50].

- **Linear Regression** assumes that the dependent (output) variable and its predictors (input variable) have a linear relationship. It offers a straightforwardly interpretable model. However, it is inaccuracy to predict nonlinear relationship [50].
- **Ridge Regression** is a technique adopted when the data suffers from multicollinearity (independent variables are highly correlated). It does not differentiate the importance of different predictors and adds just a bias to make the estimates reliable. It also performs well when the number of predictors larger than the number of obtained data. However, ridge regression does not perform feature selection and meet the requirements to reduce the input features.
- **Bayesian Ridge Regression** is a linear regression approach. When the regression model has errors with a normal distribution, and if a particular form of prior distribution is assumed, the results of the posterior probability distributions of the models' parameters are available. It offers a natural and principled way to combine prior information with data under a decision theoretical framework, and abide by the likelihood principle. However, it does not tell us how to select the prior and what is the best way to choose the prior. It also has a high computational cost.
- **Linear Support Vector Regression** is a regression model which mainly solves the linear problems.
- **Multi-layer Perceptron Regressor** is a powerful approach for nonlinear regression. It usually has high accuracy and does not need the prior knowledge. However, it has the black-box nature that it is difficult to interpret the results, which makes a limitation for us to deep into the reasons for performance degradation and impact of different scenarios. It is also very insentive to outliers [50].
- **SVR with Kernel Function (RBF)** is a regression approach which applies the RBF in the support vector regression. It has strong theoretical guarantees and sparse solutions. It also can make use of different kernels. Compared with other models, it is more generalized and deals with nonlinear relationship well.
- **Kernel Ridge Regression (RBF)** is a kernel-based regularized form of regression which combines ridge regression with the kernel trick. It utilizes the kernel trick to operation data in high-dimensional feature space without computing the coordinates of the data in that space. It only needs to computing the inner products between all pairs of data in the original feature space. The method has strong theoretical guarantees and can use different kernels. However, the solution is not sparse and has long traing time for large matrices.

C. MODEL OPTIMIZATION

Before modelling the relationship between environmental factors and the human factor. We perform a data pre-processing step. To standardize the range of the independent input variables, we do feature standardization [51]. Kalman filtering is an algorithm which uses a series of measurements observed over time and noise to generate

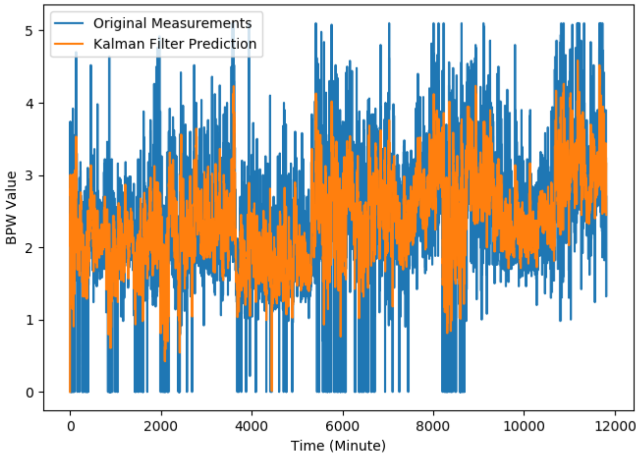


FIGURE 8. The values of the BPW between the original measurements and the kalman filtering results.

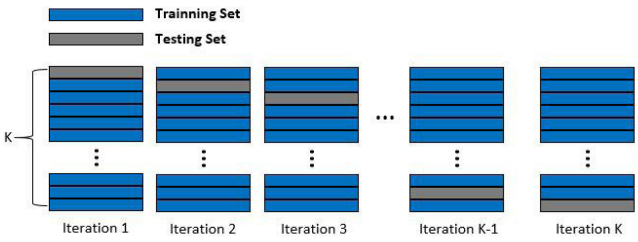


FIGURE 9. The diagram of k-folder cross-validation.

estimates of unknown variables for each timeframe [52]. It tends to have a more accurate measurement. To reduce the influence of noise, we also apply the kalman filtering [52] to the BPW (human factor) [32]. Fig. 8 shows that the values of BPW between the original measurement and the filtered results. After filtering, we can get a more reliable result.

In order to find the best optimized parameters for SVR, we combine the grid search with the k-folder cross-validation method [53], and apply them on the model. The grid search is adopted to choose the best combination parameters for the SVR. First, we specify the range of parameters for a model. For each group of parameters, we employ k-folder cross-validation approach on the data to evaluate the model. The cross-validation technique averages measures of the score (fitness) of the prediction and find a more accurate prediction model [53]. For the diagram of k-fold cross-validation method shown in Fig. 9, the original data is randomly partitioned in to k equal subsamples. K-1 subsamples are the training data, and the remaining one is the testing data. It repeats k times and each of the k subsamples is used as the testing data once [53]. The k-folder cross-validation method can be used to reduce underfitting and overfitting.

The framework of parameter selection and model optimization is briefly presented in Alg. 1. Lines 1 through 3 introduce the initialization of the the variables and parameters. From line 4 to line 17, the procedure employs grid search and k-folder cross-validation technique to find the optimized parameters so as to obtain the best model. Specifically, lines 5 through 6 initialize the variable and divide the data. Lines 7 through 11 indicate the k-folder cross-validation step. Lines 12 through 16 evaluate and compare the models. Line 18 returns the best model and indicates the procedure finishes.

Algorithm 1 The Framework of Parameters Selection and Model Optimization

```
1:  Given a dataset $D$
2:  Specify the range of the parameters of a model and let the
    $P$ denotes the parameter set
3:  tmpscore $= 0$ //initial the value. record the score of the
    model based on the data
4:  for each group of parameters $\in P$ do
5:      tempscorea $= 0$ //“mean absolute error” acts as criteria. the smaller the value, the better the model
        //apply the k-folder cross-validation
6:      Divide the dataset $D$ into $k$ subsamples. $k - 1$ subsamples act as training data, the left one acts as testing data.
7:      for $i = 1$ to $k$ do
            //each part of $k$ subsamples will act as testing data
            once in the k-folder cross-validation
8:          Calculate the score and its value is denoted by temponetimescore
9:          tempscorea $=$ tempscorea $+$ temponetimescore
10:         Use currentModel to record both the current model
            and the corresponding parameters
11:     end for
12:     tmpscorea $=$ tempscorea / $k$ //get the mean value (k
        times)
13:     if tempscorea $<$ tmpscore then
            //if find a better model
14:         tempscore $=$ tempscorea
15:         model$_{best}$ $=$ currentModel
16:     end if
17: end for
18: Return the best model model$_{best}$ ///it will be used to
    predict
```

We also adopt the holdout method [53] to find the optimal model. We randomly divide the data set into the training set and the testing data in the holdout method. We train a model based on the training data only and use the model to predict results based on the testing data.

V. EXPERIMENT RESULTS

We conduct the experiments and obtain our experimental data in indoor office using indoor air quality (IAQ) sensors [38] and wearable sensors [32] described in previous sections. We adopt the support vector regression with radial basis function (SVR-RBF) model (**SVR-RBF**) in our work and use the mean squared error (MSE) [35], mean absolute error (MAE) [36] and R-squared score (R2 Score) [37] of the

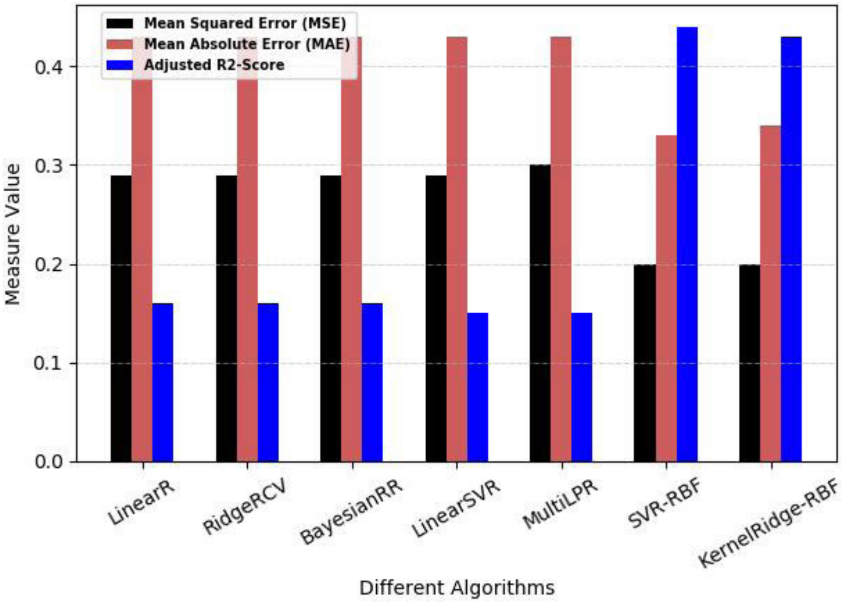


FIGURE 10. The different measure values under seven different models.

testing set as the criteria to evaluate the efficiency of the model. We also compare the other six widely adopted regression models including Linear Regression (**LinearR**), Ridge Regression with Built-In Cross-validation (**RidgeRCV**), Bayesian Ridge Regression (**BayesianRR**), Linear Support Vector Regression (**LinearSVR**), Multi-layer Perceptron Regressor (**MultiLPR**) and Kernel Ridge Regression (RBF) (**KernelRidge-RBF**) on the same data set with the same criteria. We employ the holdout method [53] and K-folder cross-validation method [53] to measure each model. In our experiment, we use 70% data as the training data and 30% data as the testing data, widely used in previous studies, in the holdout method and run 10 times with different random seeds for each experiment and present the average result for the holdout method. We adopt the 10-folder cross-validation method since we choose a trade-off between the computation cost and the accuracy. The parameters used in SVR-RBF are that ϵ and C are 0.1, 200 respectively. The total number of data is 11812.

A. THE RESULT OF THE HOLDOUT METHOD

We obtain the average result of the holdout method for seven different machine learning models and measure each model based on three criterions: Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared score (R2 Score). The average results are shown in Fig. 10 and the black bars show the MSE values, the red bars show the MAE values and the blue bars show the Adjusted R2-Score values under seven different models.

1) MEAN SQUARED ERROR (MSE)

MSE is the average squared difference between the original values and the estimated values. It is non-negative. The smaller the value, the better the model [35], [54]. MSE is a measure of the quality of the estimator.

From the black bars in Fig. 10, we observe that SVR with RBF kernel function (**SVR-RBF**) and Kernel Ridge Regression with RBF kernel function (**KernelRidge-RBF**) have better results than other models since they have the smaller MSE.

2) MEAN ABSOLUTE ERROR (MAE)

MAE is the average of absolute difference between the original values and the estimated values [36]. It is non-negative. The smaller the value, the better the model. From the red bars in Fig. 10, we observe that SVR with RBF kernel function (**SVR-RBF**) has the best result than other models since it has the smallest MAE.

3) ADJUSTED R2 SCORE

R2 score is the coefficient of determination and is the proportion of the variance of the output variable which can be predictable from the input variables [37]. Adjusted R2 score penalizes the statistic when extra variables are considered in the model. The maximum of the R2 score is one. The larger the R2 score, the better the model. From the blue bars in Fig. 10, we observe that SVR with RBF kernel function (**SVR-RBF**) has the better result than other models since it has the largest adjusted R2 score.

In summary, based on these three criterions we obtain Support Vector Regression with Radial Basis Function Kernel (**SVR-RBF**) is the best model for the prediction of comfort level. The reason is that compared with other models, the SVR-RBF model is more generalized and deals with nonlinear relationship better.

TABLE 1. Measure values under different models.

Model	MSE(Train)	MAE (Train)	Adjusted R2-score(Train)	MSE (Test)	MAE (Test)
Linear Regression	0.29	0.42	0.16	0.37	0.49
Ridge Regression Cross-validation	0.29	0.42	0.16	0.37	0.49
Bayesian Ridge Regression	0.29	0.42	0.16	0.37	0.49
Linear Support Vector Regression	0.29	0.42	0.15	0.39	0.49
Multi-layer Perceptron regressor	0.29	0.42	0.15	0.38	0.5
SVR Kernel with RBF	**0.18**	**0.31**	**0.48**	**0.36**	**0.47**
Kernel Ridge Regression with RBF	**0.18**	0.32	0.47	0.4	0.49

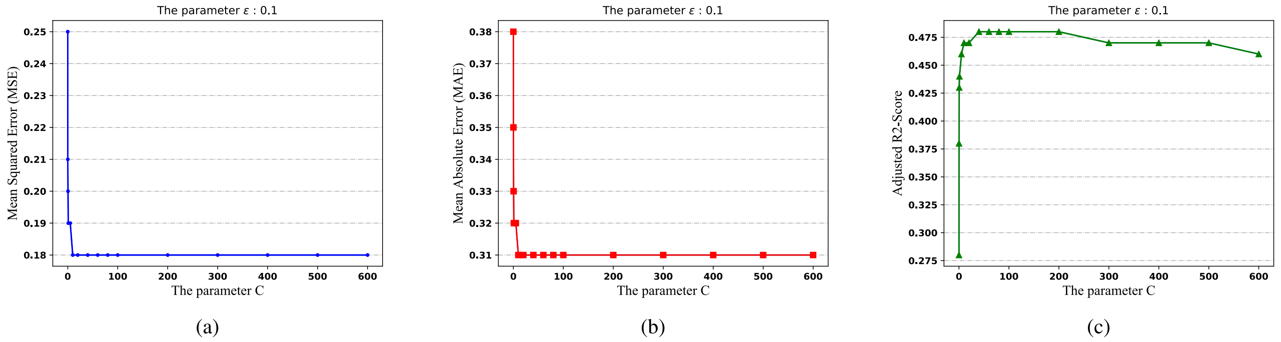


FIGURE 11. Performance analysis of the impact of parameter C.

B. THE RESULT OF THE 10-FOLDER CROSS-VALIDATION METHOD

We also use the 10-folder cross-validation method for these models and the results containing the MSE, MAE and R2 score are shown in the Table 1. From Table 1, we know that the MSE, MAE and adjusted R2 score on testing data under the support vector regression with radial basis function kernel (SVR-RBF) model are 0.36, 0.47 and 0.48 respectively. The results show that the SVR-RBF is also the best model for the prediction of comfort level since it has the smallest MSE, the smallest MAE and the largest R2 score compared with other models.

C. PARAMETER ANALYSIS OF METHOD

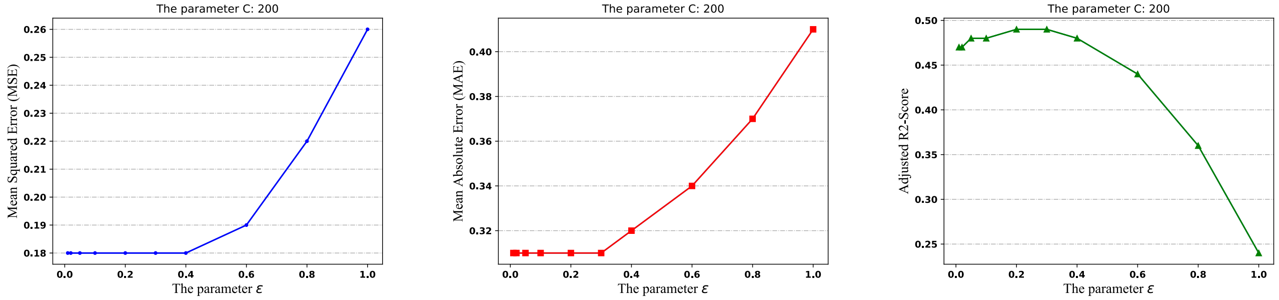
We explore the impact of parameter C and parameter ϵ on the performance.

Fig. 11 shows the variation of the MSE, MAE and R2 score with different values of parameter C (within the range [0.01,600]) under the fixed ϵ ($\epsilon = 0.1$). The curves in Fig. 11(a) (circular marks, blue color), Fig. 11(b) (rectangular marks, red color) and Fig. 11(c) (triangular marks, green color) show the variations of the MSE, MAE, R2 score with the different values of parameter C, respectively. Fig. 11(a) indicates that when the parameter C varies from 0.01 to 5, the MSE decreases quickly. After that, the MSE does not change. The reason is that when the C is small, it has little impact on minimizing the error. With an increasing C, it reaches a good level which can minimize the error. Fig. 11(b) shows that when the parameter C varies from 0.01 to 5, the MAE decreases quickly. After that, the MAE does not change. The reason is the same as that for MSE. Fig. 11(c) indicates that when the parameter C varies from 0.01 to 20, the R2 score increases quickly. When C varies from 40 to 200, the R2 score keeps at a steadily state. After that, the R2 score slowly decreases. The reason is that the large value of Parameter C impacts the fitting of the model. These figures show that the best range for choosing parameter C should be [40, 200].

Fig. 12 shows the variation of the MSE, MAE and R2 score with different values of parameter ϵ (within the range [0.01,1]) under the fixed parameter C (C = 200). The curves in Fig. 12(a) (circular marks, blue color), Fig. 12(b) (rectangular marks, red color) and Fig. 12(c) (triangular marks, green color) show the variations of the MSE, MAE, R2 score with the different values of parameter ϵ, respectively. Fig. 12(a) indicates that when the parameter ϵ varies from 0.01 to 0.4, the MSE keeps a small value. After that, the MSE increases quickly. The reason is that when the ϵ is small, the model has a high accuracy. With an increasing ϵ, it becomes inaccurate. Fig. 12(b) shows that when the parameter ϵ varies from 0.01 to 0.3, the MAE has a small value. After that, the MAE increases rapidly. The reason is the same as that for MSE. Fig. 12(c) indicates that when the parameter ϵ varies from 0.01 to 0.3, the R2 score increases. After that, the R2 score decreases quickly. The reason is that the large value of Parameter ϵ makes the model inaccuracy. Based on the criteria, the best range for choosing parameter ϵ should be [0.01, 0.2].

VI. DISCUSSION

We present a framework to obtain environmental and human data, and adopt a data-driven machine learning model namely support vector regression with radial basis function (SVR-RBF) to predict the human comfort. We have explored the relationships between environmental factors and the



(a) (b) (c)

FIGURE 12. **Performance analysis of the impact of parameter ϵ.**

human factor. From the plotted figures, we know that there is a nonlinear relationship between the environmental factors and the human factor. We also prove that the best model that fits these data is SVR-RBF which is a nonlinear model. This agrees with the conclusion of our preceding exploration on the characteristics of the data. However, the results of our model do not reach the best theoretical value according to the MSE, MAE and the adjusted R2 score. The reason could be that we only considered part of the factors related to human comfort which are not comprehensive. Thus, we need to investigate more human factors and improve the model. Previous studies do not have a common good indicator which can represent human comfort well. We adopt the BPW as an objective comfort index from the verified sensor to reflect the human comfort to some extent. Although it is an objective index, a single parameter of BPW alone may not be able to fully represent the overall human comfort. Thus, a better composite reliable indicator is needed.

The dataset used in our work and the corresponding results are limited. Due to the lack of volunteers and resources (funding, office room etc.) to conduct the experiments, in this study we develop a general model by combining BPW of all the volunteers. However, individuals are different from each other in terms of their comfort range, ages, health conditions etc which may lead to biased results. The more the acquired data, the more information we can obtain. In order to gain the more generalized and reliable model, we need to involve more volunteers with widened range of physiological and psychological factors in the experiments. To predict comfort status more accurately for an individual occupant, we may consider developing personalized comfort model for each individual.

As we mentioned above, people spend most of their time in the indoor buildings. Building a robust and highly accurate comfort model for individuals plays an important role for improving human comfort and saving energy. The comfort model can be used to monitor the comfort status of people in real time. Based on the predictions from the comfort model, the environmental condition for the individuals can be well controlled. For instance, if it is raining outside, the indoor temperature of a building equipped with air conditioner decreases that makes the people feel cold or discomfort. Through comfort model, the temperature can be automatically adjusted to a high value that saves the energy. In addition, improvement of human comfort not only brings positive effects on human wellbeing, but also promotes the work efficiency. Specifically, promoting health at work contributes to employees's engagement and productivity as well as leads to significantly savings in operating cost for employers.

VII. CONCLUSION

In this paper, we proposed a framework for data collection and adopted an environmental and human data-driven model namely support vector regression with radial basis function (SVR-RBF) to detect comfort level of occupants in the indoor buildings. We took environmental factors from indoor air quality sensors as input and the human factor (blood pulse wave: BPW) from wearable sensors as the output to derive a machine learning model. The model predicted the comfort level (value) based on the environment input, which helped monitor the comfort status of the occupants with the aim to prevent health problems in advance. We have explored the relationship between environmental factors and the human factor which shows they have nonlinear relationship. We also studied several popular regression models based on the same dataset and evaluated them by the hand-out method and the 10-folder cross-validation method. The experimental results showed that the SVR-RBF achieved the best prediction results according to the values of the mean squared error (MSE), mean absolute error (MAE) and the R-squared score. The results also showed that the factors have nonlinear relationship which matches our previous investigation. The employed approach provided a potential solution to improve the health of occupants which was significantly important to our life. In future work, we aim to include in our model more human factors and other external factors such as different regions. This would help build smart buildings which has certain requirements on energy efficiency.

REFERENCES

[1] M. Frontczak and P. Wargocki, "Literature survey on how different factors influence human comfort in indoor environments," *Building Environ.*, vol. 46, no. 4, pp. 922–937, 2011.

[2] H. Gao, C. Koch, and Y. Wu, "Building information modelling based building energy modelling: A review," *Appl. Energy*, vol. 238, pp. 320–343, Mar. 2019.

[3] P. K. Rosier, "Comfort theory and practice: A vision for holistic health care and research," *Clin. Nurse Spec.*, vol. 19, no. 1, p. 49, Jan./Feb. 2005.

[4] L. Stalder, "Sensing human comfort: An inclusive implementation of indoor environmental data collection," M.S. thesis, Univ. Fribourg, HUMAN-IST Inst., Fribourg, Switzerland, Oct. 2015.

[5] G. Kyung, M. A. Nussbaum, and K. Babski-Reeves, "Driver sitting comfort and discomfort (Part I): Use of subjective ratings in discriminating car seats and correspondence among ratings," *Int. J. Ind. Ergonom.*, vol. 38, nos. 5–6, pp. 516–525, 2008.

[6] S. Pinto, L. Fumincelli, A. Mazzo, S. Caldeira, and J. C. Martins, "Comfort, well-being and quality of life: Discussion of the differences and similarities among the concepts," *Porto Biomed. J.*, vol. 2, no. 1, pp. 6–12, 2017.

[7] L. Fortney and M. Taylor, "Meditation in medical practice: A review of the evidence and practice," *Primary Care, Clinics Office Pract.*, vol. 37, no. 1, pp. 81–90, 2010.

[8] M. S. Mustapa, S. A. Zaki, H. B. Rijal, A. Hagishima, and M. S. M. Ali, "Thermal comfort and occupant adaptive behaviour in japanese University buildings with free running and cooling mode offices during summer," *Building Environ.*, vol. 105, pp. 332–342, Aug. 2016.

[9] J. Guttmann, H. Bernhard, G. Mols, A. Benzing, P. Hofmann, K. Geiger, C. Haberthür, D. Zappe, and B. Fabry, "Respiratory comfort of automatic tube compensation and inspiratory pressure support in conscious humans," *Intensive Care Med.*, vol. 23, no. 11, pp. 1119–1124, Nov. 1997.

[10] T. Hwang and J. T. Kim, "Effects of indoor lighting on occupants' visual comfort and eye health in a green building," *Indoor Built Environ.*, vol. 20, pp. 75–90, Feb. 2011.

[11] G. Yun, K. C. Yoon, and K. S. Kim, "The influence of shading control strategies on the visual comfort and energy demand of office buildings," *Energy Buildings*, vol. 84, pp. 70–85, Dec. 2014.

[12] *The Noise Rating*, Standard ISO/R 1996:1971, Mar. 2016, p. 47.

[13] A. Gramez and F. Boubenider, "Acoustic comfort evaluation for a conference room: A case study," *Appl. Acoust.*, vol. 118, pp. 39–49, Mar. 2017.

[14] N. Djongyang, R. Tchinda, and D. Njomo, "Thermal comfort: A review paper," *Renew. Sustain. Energy Rev.*, vol. 14, no. 9, pp. 2626–2640, 2010.

[15] J. A. O. García, "A review of general and local thermal comfort models for controlling indoor ambiences," in *Air Quality*, A. Kumar, Ed. London, U.K.: IntechOpen, 2010, pp. 309–326.

[16] S. Carlucci and L. Pagliano, "A review of indices for the long-term evaluation of the general thermal comfort conditions in buildings," *Energy Buildings*, vol. 53, pp. 194–205, Oct. 2012.

[17] R. Holopainen, P. Tuomaala, P. Hernandez, T. Häkkinen, K. Piira, and J. Piippo, "Comfort assessment in the context of sustainable buildings: Comparison of simplified and detailed human thermal sensation methods," *Building Environ.*, vol. 71, pp. 60–70, Jan. 2014.

[18] S. Wu and J.-Q. Sun, "Two-stage regression model of thermal comfort in office buildings," *Building Environ.*, vol. 57, pp. 88–96, Nov. 2012.

[19] L. Barrios and W. Kleiminger, "The Comfstat—Automatically sensing thermal comfort for smart thermostats," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2017, pp. 257–266.

[20] M. Abdallah, C. Clevenger, T. Vu, and A. Nguyen, "Sensing occupant comfort using wearable technologies," in *Proc. Construct. Res. Congr.*, May 2016, pp. 940–950.

[21] K. Katić, R. Li, J. Verhaart, and W. Zeiler, "Neural network based predictive control of personalized heating systems," *Energy Buildings*, vol. 174, pp. 199–213, Sep. 2018.

[22] Y. Zhao, Q. Zhao, F. Wang, Y. Jiang, and F. Zhang, "On-line adaptive personalized dynamic thermal comfort (PDTC) model using recursive least square estimation," in *Proc. Int. Conf. Intell. Building Manage.*, 2011, pp. 275–279.

[23] C. Manna, N. Wilson, and K. N. Brown, "Personalized thermal comfort forecasting for smart buildings via locally weighted regression with adaptive bandwidth," in *Proc. 2nd Int. Conf. Smart Grids Green IT Syst. (SMARTGREENS)*, Jan. 2013, pp. 32–40.

[24] D. Fay, L. O'Toole, and K. N. Brown. "Gaussian process models for ubiquitous user comfort preference sampling; global priors, active sampling and outlier rejection," *Pervasive Mobile Comput.*, vol. 39, pp. 135–158, Aug. 2017.

[25] F. Salamone, L. Belussi, C. Currò, L. Danza, M. Ghellere, G. Guazzi, B. Lenzi, V. Megale, and I. Meroni, "Integrated method for personal thermal comfort assessment and optimization through users' feedback, IoT and machine learning: A case study," *Sensors*, vol. 18, no. 5, p. 1602, May 2018.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA, USA: Wadsworth, 1984.

[27] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.

[28] D. Enescu, "A review of thermal comfort models and indicators for indoor environments," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 1353–1379, Nov. 2017.

[29] F. R. d'Ambrosio Alfano, B. W. Olesen, B. I. Palella, and G. Riccio, "Thermal comfort: Design and assessment for energy saving," *Energy Buildings*, vol. 81, pp. 326–336, Oct. 2014.

[30] F. Jazizadeh, A. Ghahramani, B. Becerik-Gerber, T. Kichkaylo, and M. Orosz, "Human-building interaction framework for personalized thermal comfort-driven systems in office buildings," *J. Comput. Civil Eng.*, vol. 28, no. 1, pp. 2–16, Feb. 2014.

[31] T. Chaudhuri, Y. C. Soh, H. Li, and L. Xie, "Machine learning based prediction of thermal comfort in buildings of equatorial singapore," in *Proc. IEEE Int. Conf. Smart Grid Smart Cities (ICSGSC)*, Jul. 2017, pp. 72–77.

[32] Biovotion Company. (2019). *The Biovotion 'Spiral' Explained*. [Online]. Available: https://biovotion.zendesk.com/hc/en-us/articles/213542089-The-Biovotion-Spiral-explained

[33] Q. Guo, Z. Li, B. An, P. Hui, J. Huang, L. Zhang, and M. Zhao, "Securing the deep fraud detector in large-scale e-commerce platform via adversarial machine learning approach," in *Proc. ACM World Wide Web Conf. (WWW)*, 2019, pp. 616–626.

[34] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, Dec. 2013.

[35] S. Singh and P. Dayan, "Analytical mean squared error curves for temporal difference learning," *Mach. Learn.*, vol. 32, no. 1, pp. 5–40, Jul. 1998.

[36] B. Zeng and Y. Neuvo, "Optimal parallel stack filtering under the mean absolute error criterion," *IEEE Trans. Image Process.*, vol. 3, no. 3, pp. 324–327, May 1994.

[37] T. Chen and U. M. Braga-Neto, "Maximum-likelihood estimation of the discrete coefficient of determination in stochastic Boolean systems," *IEEE Trans. Signal Process.*, vol. 61, no. 15, pp. 3880–3894, Aug. 2013.

[38] uHoo Company. (2019). *Frequently Asked Questions*. [Online]. Available: https://uhooair.com/faq_category/air-quality-indicators/

[39] Biovotion Company. (2019). *Newsletter*. [Online]. Available: https://www.biovotion.com/blog/

[40] W. Kim, M. S. Stanković, K. H. Johansson, and H. J. Kim, "A distributed support vector machine learning over wireless sensor networks," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2599–2611, Nov. 2015.

[41] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[42] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[43] F. Javed, G. S. H. Chan, A. V. Savkin, P. M. Middleton, P. Malouf, E. Steel, J. Mackie, and N. H. Lovell, "RBF kernel based support vector regression to estimate the blood volume and heart rate responses during hemodialysis," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, Sep. 2009, pp. 4352–4355.

[44] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer, 2000.

[45] J. Kim, Y. Zhou, S. Schiavon, P. Raftery, and G. Brager, "Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning," *Building Environ.*, vol. 129, pp. 96–106, Feb. 2018.

[46] J. G. Allen, P. MacNaughton, J. G. C. Laurent, S. S. Flanigan, E. S. Eitland, and J. D. Spengler, "Green buildings and health," *Current Environ. Health Rep.*, vol. 2, no. 3, pp. 250–258, Sep. 2015.

[47] P. O. Fanger, *Thermal Comfort. Analysis and Applications in Environmental Engineering*. New York, NY, USA: McGraw-Hill, 1972.

[48] *Thermal Environmental Conditions for Human Occupancy*, Ashrae 55-2013 Standard 55-2013, ASIN, ASHRAE, 2013, p. 54.

[49] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

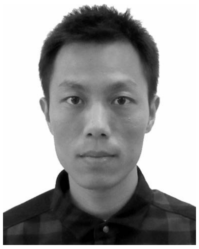
[50] E. Ould-Ahmed-Vall, J. Woodlee, C. Yount, and K. A. Doshi, "On the comparison of regression algorithms for computer architecture performance analysis of software applications," in *Proc. Workshop SMART Co-Located HIPEAC*, Ghent, Belgium, Jan. 2007.

[51] S. H. Myers and B. M. Huhman, "Enabling scientific collaboration and discovery through the use of data standardization," *IEEE Trans. Plasma Sci.*, vol. 43, no. 5, pp. 1190–1193, May 2015.

[52] C. Pantaleon and A. Souto, "Comments on 'An aperiodic phenomenon of the extended Kalman filter in filtering noisy chaotic signals,"' *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 383–384, Jan. 2005.

[53] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, Mar. 2010.

[54] X. Zhou, Y. Murakami, T. Ishida, X. Liu, and G. Huang, "ARM: Toward adaptive and robust model for reputation aggregation," *IEEE Trans. Autom. Sci. Eng.*, to be published.



FUBING MAO received the M.S. degree from the School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, and the Ph.D. degree from the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He studied in the Department of Computer Science and Technology, Tsinghua University, for two years, as a Visiting Student during his master's degree. He is currently a Postdoctoral Fellow with the Delta-NTU Corporate Laboratory for Cyber-Physical Systems, Nanyang Technological University, Singapore. His research interests include machine learning, complex networks, FPGA, the layout optimization of IC designs, and optimization algorithms.

XIN ZHOU received the M.S. degree in computer science from Shanghai Jiao Tong University, in 2012, and the Ph.D. degree in informatics from Kyoto University, in 2016. He is currently a Research Fellow with Nanyang Technological University. His current research interests include the IoT, service computing, mobile computing, recommender systems, trust, and reputation management in e-commerce systems.

YING SONG received the M.S. degree in bioengineering from the National University of Singapore, in 2007. From 2007 to 2018, she was a Lecturer with the School of Engineering, Republic Polytechnic, Singapore. Since May 2018, she has been a Research Associate with the Delta-NTU Corporate Laboratory for Cyber-Physical Systems, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Her research interests include the applications of the Internet of Things (IoT), smart living, and biomedical signal processing.