Heliyon 9 (2023) e16114

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# A innovative wavelet transformation method optimization in the noise-canceling application within intelligent building occupancy detection monitoring

Jan Vanus, Jan Kubicek, Dominik Vilimek *, Marek Penhaker, Petr Bilik

*Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB - Technical University of Ostrava, 17. listopadu 15, Ostrava – Poruba, 708 00, Czech Republic*

ARTICLE INFO

*Keywords:*
Smart home
Prediction of room occupancy
Big data processing
Presence of person monitoring
Activities monitoring with indirect methods

ABSTRACT

The study deals with detection of the occupation of Intelligent Building (IB) using data obtained from indirect methods with Big Data Analysis within IoT. In the area of daily living activity monitoring, one of the most challenging tasks is occupancy prediction, giving us information about people's mobility in the building. This task can be done via monitoring of $CO_2$ as a reliable method, which has the ambition to predict the presence of the people in specific areas. In this paper, we propose a novel hybrid system, which is based on the Support Vector Machine (SVM) prediction of the $CO_2$ waveform with the use of sensors that measure indoor/outdoor temperature and relative humidity. For each such prediction, we also record the gold standard $CO_2$ signal to objectively compare and evaluate the quality of the proposed system. Unfortunately, this prediction is often linked with a presence of predicted signal activities in the form of glitches, often having an oscillating character, which inaccurately approximates the real $CO_2$ signals. Thus, the difference between the gold standard and the prediction results from SVM is increasing. Therefore, we employed as the second part of the proposed system a smoothing procedure based on Wavelet transformation, which has ambitions to reduce inaccuracies in predicted signal via smoothing and increase the accuracy of the whole prediction system. The whole system is completed with an optimization procedure based on the Artificial Bee Colony (ABC) algorithm, which finally classifies the wavelet's response to recommend the most suitable wavelet settings to be used for data smoothing.

## 1. Introduction

In order to optimize the management of operational and technical functions in intelligent buildings, Big data processing and IoT concepts are increasingly being applied to reduce the cost of building operation.

The presented work deals with the utilization and implementation of modern mathematical methods (Big Data processing, SVM, Wavelet transformation (WT) in noise canceling application) for processing of measured data as part of the occupancy monitoring in IB within IoT.

* Corresponding author.
*E-mail addresses:* jan.vanus@vsb.cz (J. Vanus), jan.kubicek@vsb.cz (J. Kubicek), dominik.vilimek@vsb.cz (D. Vilimek), marek.penhaker@vsb.cz (M. Penhaker), petr.bilik@vsb.cz (P. Bilik).

https://doi.org/10.1016/j.heliyon.2023.e16114
Received 5 April 2022; Received in revised form 30 April 2023; Accepted 5 May 2023

Available online 12 May 2023

Nowadays, it is more common to understand IoT as a data collection network obtained by sensors. Nevertheless, due to the large amount of information that these sensors generate in time, and the speed at which it occurs, the concept of Big Data (BD) emerges, being capable of generating, according to IBM and CSA [1], more than 2.5 quintillion bytes per day, to the point, that 90% of the world's data has been created during the last two years. In this investigation, we will focus on the application of IoT and Big Data on the detection of the presence or occupancy of people in Smart Home or intelligent buildings.

## 2. Related works

Occupancy monitoring or detection systems play a significant role in the possibility of energy-saving and consumption of this type of buildings [2,3]. According to a study published by Cisco, the Internet of things will connect around 50 billion devices worldwide in 2020, generating data traffic, which should be saved, analysed, and prepared for processing [4]. Therefore, we refer to the treatment and analysis of vast data repositories, so immensely large that it is impossible to treat them with the tools of conventional databases and analytics. The current data storage paradigm "cloud computing" and the various security concerns in IoT have led to the emergence of alternatives for data storage, organization, and processing, as well as technologies such as block chain that provide decentralized management of private data, as well as a higher level of security in IoT [5–7]. Currently, the most studies are focused on predicting the presence and number of people based on the recognition of the human silhouette [8], as well as with motion detection sensors or even through technologies such as Wi-Fi, Bluetooth [9,10], or images obtained from video cameras [11]. With this approach, an inconvenience arises, such as the privacy of people [12]. Since they would be continuously monitored, an alternative is proposed to detect if there is the presence of people or not in an IoT, such as the use of non-invasive methods using the current infrastructure of IoT or Infrastructure Mediated Sensing (IMS) [13]. These methods will also be useful for the care of elder people and the detection of possible falls or aging issues [14]. This study is focus on methods that predict occupancy in a non-intrusive manner for individuals. By leveraging the pre-existing infrastructure within a building, such as temperature sensors, CO2 sensors, HVAC units, surveillance cameras, and gas consumers, we can monitor occupancy without the need for user-installed software [15–20]. Various mathematical experiments have been conducted for occupancy recognition and people detection in Smart Homes, including Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), Random Forest (RF), Decision Trees (DT), rule induction, k-means clustering, and K-nearest neighbor algorithm (KNN), all of which have shown promising results [21].

The objective of this article is the design of a prediction method based on SVM, to determine the occupancy of the room, using non-invasive methods [22]. The article describes a proposal of a newly designed indirect method for detecting occupancy of monitored areas in intelligent buildings using the prediction of the $CO_2$ concentration trend from the operational measurement of non-electrical quantities (temperature indoor, relative humidity indoor) using SVM [23]. To enhance the accuracy of the newly proposed CO2 prediction method, an additional novel approach has been employed in this study. This approach optimally adjusts the parameters of the Wavelet Transformation (WT) mathematical method, which is utilized for suppressing additive noise in the predicted $CO_2$ waveform.

## 3. Material and methods

The practical implementation of the proposed Wavelet transformation method optimization in the noise-canceling application within Intelligent Building occupancy detection monitoring is divided into the following parts:

1. Part the SVM prediction of the $CO_2$ waveform
   (a) Data measuring, preprocessing and visualization.
   (b) Model design, model building, model evaluation, Test Developed, ShuffleSplit Configuration, Train_test_split Configuration, Model performance based on the SVM prediction of the $CO_2$ waveform with the use of sensors that measure indoor/outdoor temperature and relative humidity.
   (c) Implementation of the practical part experiments.
2. The new Wavelet transformation method optimization in the noise-canceling application
   (a) Motivation of Wavelet-based $CO_2$ signal smoothing.
   (b) Spatial wavelet response for $CO_2$ prediction.
   (c) A decomposition scheme for optimal wavelet selection.
   (d) Wavelet-based recommendation system.
   (e) A design of decomposition model for wavelet recommendation.
   (f) Definition of Fuzzy logic-based decomposition model.
   (g) Prediction model for estimation of vertex function.

### 3.1. Data preprocessing and data visualization

"The fundamental purpose of data preparation is to manipulate and transform raw data so that the information content enfolded in the data set can be exposed or made more easily accessible" [24]. One of the most important tasks in data analysis is data preprocessing [25]. It may be due to the impure nature of the data, which may result in the extraction of patterns or rules that are not very useful, and possibly as a result of:

- Uncompleted data
- Data noise
- Inconsistent data

Data preparation can generate a smaller data set than the original [26], which can improve the efficiency of the Data Mining process that includes:

- Relevant data selection: eliminating duplicate records, eliminating anomalies.
- Data Reduction: Selection of characteristics, sampling or selection of instances, discretization, and correlation coefficient.
- Recover incomplete information and outliers treatment.

Another crucial step in our methodology is the visualization of the data, which allows us to understand the initial distribution of our dataset [27]. Our analysis will be based on the following key aspects of data visualization:

- Data visualization is a way of displaying complex data graphically.
- The graphics can be precise to locate the implemented models as planned.
- It is possible allow greater ease to compare and interpret data, thus visualizing a large number of them quickly.
- It allows us to have a first global and fast image on how the data is distributed, as well as a time-saving [28].

In our case, it will allow us to choose what type of kernel to implement using SVM.

### *3.2. Model design*

Once the requirements analysis and preprocessing are finished, we will proceed to design each of the modules and software components of the system to be implemented during project development. In the same way, they must safely devise the algorithms and mechanisms with which they will solve the problem, and that will be implemented in the software components. The preprocessing of the data is critical for determining which method is appropriate for our problem and that the free lunch theorem does not apply [29,30].

### *3.3. Model building*

The phase will proceed to implement each of the component's software obtained during the design phase and with which it is intended to give solution to the problem posed. The Python programming language was selected for this study due to its inherent ease of use and the broad range of available libraries that facilitate the processing of data and the implementation of advanced mathematical and machine learning algorithms.

### *3.4. Model evaluation*

At this point, we will focus on the execution of the algorithm developed with the data obtained in order to obtain results that should be analyzed later. In this phase, it will also be advisable to compare the algorithm with others already implemented in order to know the performance of our algorithm. The phase analysis of results in which the operation of the developed system will be tested in order to find errors and improve the system. In turn, the results of the SVM algorithm to be implemented will be obtained, having used the dataset provided by the project tutor. This output will be analyzed, as well as conclusions will be given for them.

### *3.5. Test developed*

A set of tests have been carried out to obtain the generalization values of our model. For this, a division of the data set has been made into 80% of data for the train and 20% for tests. The division of the data has been divided in two ways, using a separation using split arrays or matrices into a random train and test subsets. Also, on the other hand, through ShuffleSplit, which will randomly sample the entire data set, in this way, we will be able to obtain the performance of our model by using the validation set. Various intervals will be used to conduct the generalization tests, including one day, three days, one week, two weeks, and one month. To ensure fairness in the tests, the first day and time of each dataset have been selected as the starting point. Testing could be conducted on random days, but this approach may not provide a fair comparison between the models. For each parameter configuration and method used in evaluating the model, the results will be presented in a table.

### *3.6. ShuffleSplit configuration*

The ShuffleSplit method, unlike other cross-validation strategies, does not guarantee that all folds will be distinct, though this is still highly probable for large datasets [31]. The parameters utilized for ShuffleSplit are as follows:

- n_split: The number of re-shuffling and splitting iterations; in this case, five splits have been used.

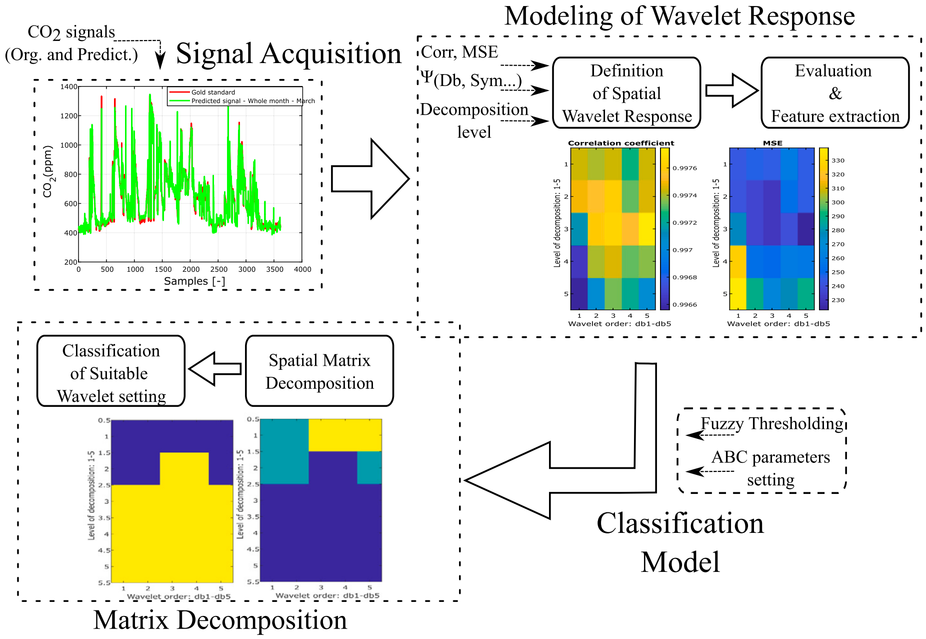


**Fig. 1.** Whole flow-chart of proposed system for spatial mapping of wavelet's response and consequent classification of suitable wavelet settings.

- test_size: This represents the percentage of the dataset that will be used in the test split; for this study, 20% (0.2) will be used.
- random_state: It is the seed used by the random number generator.

### 3.7. *Motivation of wavelet-based $CO_2$ signal smoothing*

Wavelet-based $CO_2$ signal smoothing is linked with parameters settings, which may significantly influence the smoothing results. When applying Wavelet transformation (Fig. 1), we investigate a degree of the correlation ($W(a,b)$) between the $CO_2$ signal ($s(x)$), as is denoted in equation (1) and respective wavelet ($\psi_{a,b}$) by the definition:

$$W(a,b) = \int_{-\infty}^{\infty} s(x)\psi_{a,b}(x)dx \tag{1}$$

The wavelet ($\psi_{a,b}$) can be described via using the equation (2):

$$\psi_{a,b} = \frac{1}{\sqrt{a}} \cdot \psi\left(\frac{x-b}{a}\right) \tag{2}$$

In this equation (2), the parameters $(a,b)$ represent scaling, respectively shifting factor, and the function $\psi(x)$ is denoted as mother's wavelet. Wavelet transformation has been many times proved as a very effective method for various data smoothing, features extraction, and other related tasks. One of the most significant limitations in Wavelet transformation is the parameters settings to select a proper wavelet. Meaning that a wavelet setting that approximates the predicted $CO_2$ signals with minimal differences when comparing with gold standard. In this sense, based on the experimental testing it is obvious that selection of mother's wavelet $\psi(x)$ and decomposition level $(a)$ play essential roles for achieving a high correlation with gold standard and improving the whole system accuracy. Unfortunately, this is not a trivial task. We have to consider a plenty wavelet families for testing, including Daubechies, Biorthogonal, Coiflets, Symlets and many others. Also, various settings of decomposition levels. Such settings offer a wide spectrum of combinations, which would be complicated to analyze. Therefore, our aim is to propose a versatile approach for wavelet features selection, which will indicate and recommend appropriate wavelet settings to be used for $CO_2$ signal prediction optimization.

#### 3.7.1. *Spatial wavelet response for $CO_2$ prediction*

A selection of a proper wavelet settings can be perceived as a complex challenging task. Meaning that we need to gradually analyze a lot of combinations of mother's wavelets and selected decomposition levels. For instance, when we have Daubechies (Db) wavelets, which contains 45 mother's wavelets, and when we use n different level decompositions, we would test $45 \cdot n$ different settings which should be evaluated to find out which one the best approximate a $CO_2$ signal in a contrast of a gold standard. This is our motivation to propose a scheme, which simultaneously evaluate an appropriateness of respective wavelet settings. The proposed recommendation system for optimal wavelet settings is based on so-called spatial mapping of wavelet response. This scheme represents a distribution of wavelet settings in the form of evaluating coefficients, which give an objective information about the respective wavelet effectiveness. Firstly, we present the evaluation coefficients: index of correlation (Corr) and Mean Square Error (MSE), which have the potential to objectively evaluate a level of similarity (Corr) and difference (MSE).

**Index of Corelation** provides a level of linear dependence between the gold standard and predicted $CO_2$ signal. We evaluate this parameter in the range: $[-1;1]$, where $-1$, respectively 1 stands for a perfect linear dependence in the decreasing $(-1)$, respectively increasing $(1)$ sense, contrarily 0 stands for the zero linear dependence. We use this parameter to evaluate a level of similarity for predicted $CO_2$ signals.

**MSE** represents an evaluation parameter which evaluate an average quadratic difference between a gold standard and predicted $CO_2$ signal. Instead of Corr where higher values represent a better wavelet settings, MSE represents a level of difference it means that lower values represent better agreement between a gold standard and predicted $CO_2$ signal. MSE formulation is given by the following expression, see equation (3), where $X_i$ represents $i^{th}$ sample of a gold standard, and $Y_i$ stands for $i^{th}$ sample of predicted $CO_2$ signal. In the ideal case $(MSE = 0)$, we would achieve the zero difference (full agreement) between a gold standard and prediction.

$$MSE = \frac{1}{m} \cdot \Sigma_{i=1}^{m} \left[X_i - Y_i\right]^2 \tag{3}$$

A proposed spatial wavelet response represents a distribution of Corr, respectively MSE for various wavelet settings (selected mother's wavelets and decomposition levels). This scheme provides two important benefits for objective evaluation. Firstly, this spatial mapping provides simultaneous computing the wavelet response instead of gradual selecting various wavelet settings. Nevertheless, the main benefit of this method is simultaneous spatial $(2D)$ distribution of evaluation coefficients for various wavelet settings, which can be arbitrarily selected by user. The following pseudo algorithm represents a generalized structure of the spatial modeling.

**Algorithm 1** Generalized algorithm of the spatial modeling.

```
Step 1: Load input $CO_2$ signal
Step 2: Selection of mother's wavelets (n)
Step 3: Selection of level of decomposition (k)
Step 4: Definition of zeros matrix (dimension: k x n) for storing the results
Step 5: for 1:k
        Step 6: for 1:n
                Step 7: Compute wavelet-based $CO_2$ signal smoothing
                Step 8: For each settings (k, n) compute index of correlation (Corr(k, n))
                Step 9: For each settings (k, n) compute index of correlation (MSE(k, n))
        end
end
Step 10: Generate color mapping for Corr(k, n) MSE(k, n) spatial response
```

In order to properly identify differences in various settings, we use an artificial color mapping of spatial response to create a color spatial map for each evaluating parameter. This approach apparently has the benefit in simultaneous visualization of wavelet response for multiple wavelet settings. Practically, we need to operate with this spatial model to select the wavelet settings which will be finally used for prediction. In the simplest way, the settings with the highest correlation, or the lowest MSE can be selected. Nevertheless, this is only one setting, which may be impractical because for different $CO_2$ signals, we can achieve different such evaluated settings. For this reason, a robust prediction system, which would classify based on the correlation and MSE distribution, wavelet settings into appropriate, neutral and inappropriate would significantly contribute to this complex recommendation model. When such classification procedure is applied for each $CO_2$ signal and we receive a set of appropriate wavelets settings, then we perform an intersection among these sets and obtain a mutual set of wavelet settings, which the best fit to analysed predictions. In order to perform this task, we propose a decomposition scheme, which classify the spatial wavelet response into three subsets.

### *3.7.2. A decomposition scheme for optimal wavelet selection*

In this section, we introduce a novel approach for modeling of the Wavelet settings response for $CO_2$ signal. As we discussed earlier, one of the main unclear issues in the Wavelet-based denoising is selecting the most effective wavelet to be used for the data enhancement. From the general point of the view, we have the particular wavelet family, containing a finite number of the mother's wavelets, which can be applied with using various decomposition levels to data, which we are going to filter out. In this context, we tackle with the optimization problem, which is formulated as the follows:

$$\Omega_{opt}\left(\psi_i(N,\tau), N_j\right) = \min_{\forall \psi_i(s,t), N_j \in \Omega} \left(y_m, y_{w,m}\right),\ \Omega_{opt} \subseteq \Omega,\ i,j = \left\{1, 2, \ldots, \psi_{max}, N_{max}\right\} \tag{4}$$

Here, we search for a finite domain $\Omega_{opt}$, which is represented by a set of $\psi_i(N,\tau)$, where $N$ stands for the wavelet-based decomposition and $\tau$ stands for the time shift of the wavelet. To solve such optimization problems, we search for the cost function (eq. (4)), which minimizes the difference between ideal $CO_2$ signal (gold standard) $y_m$ and $y_{w,m}$. The task can theoretically be accomplished in two ways. We can find the global maximum of similarity parameters (for example, the correlation index) or the global minimum of difference parameters (for example, the MSE) for all wavelets. In such a procedure, only one wavelet setting is selected. As we discussed in the previous section, different wavelet settings can be recommended for various $CO_2$ signals, so we are focusing on developing a robust recommendation system that will provide a finite set of wavelet settings for smoothing $CO_2$ signals in order to improve prediction accuracy. In light of these assumptions, we aim to build a robust recommendation system that is capable of selecting a finite number of wavelets and their settings within $\Omega_{opt}$ as the most suitable solution for particular wavelet families.

### 3.7.3. Wavelet-based recommendation system

Here, we are focused on the proposal of the mathematical model for selection of a set of the most suitable Wavelet settings to be used for the $CO_2$ data smoothing. This classification procedure is based on the evaluation matrix decomposition. We suppose that the evaluation wavelet-based matrix represents a domain of distribution of respective evaluation parameter, such is the correlation coefficient or MSE for each analysed wavelet and it's the decomposition level. Such matrix is decomposed into several finite disjunctive subdomains, corresponding with grouped values of the evaluation coefficients. Among such groups, we can easily select such one, corresponding with the most suitable wavelet settings. For the building of the classification procedure, we take advantage the concept of the thresholding. In conventional hard thresholding, individual values can be grouped via a system of the threshold values. In this context, a selection of suitable threshold is a challenging issue. Many concepts utilize so called hard thresholding, which firmly determine the threshold values. Such procedure is less effective, when supposing threshold placement variations, caused by various noise level and artefacts. This assumption points out on the limitation of the hard thresholding would not be effective in such case. Therefore, we propose so called soft thresholding procedure based on the sequence of the fuzzy sets, approximating behavior of the evaluation parameters distribution. Such sequence of the fuzzy sets can be theoretically distributed over the interval of the evaluation parameters by multiple ways. In order to predict the optimized scheme of the fuzzy set's distribution, we employ a modified version of the ABC (Artificial Bee Colony) algorithm for selection the optimized configuration of fuzzy sets-based soft thresholding model. This decomposition model classifies the evaluation values into predefined classes, corresponding with grouped values of these parameters. By selecting respective class, we select a group of the wavelets, appearing as the most optimal compromise for given $CO_2$ data and noise suppression. This complex model represents the minimization procedure, as we depicted in equation (4). This model represents a recommendation system for an autonomous wavelet selection for $CO_2$ signal prediction improvement.

### 3.7.4. A design of decomposition model for wavelet recommendation

In this section, we introduce the proposed model for the evaluation matrix decomposition. We suppose the input evaluation matrix in the following form, presented in the eq. (5).

$$EP(N,\psi) = f(N,\psi) \tag{5}$$

Where $EP$ stands for the respective evaluation parameters, such is the correlation index or MSE, $N$ stands for the level of the decomposition, and $\psi$ is a Wavelet, belonging to a respective wavelet class, which is tested. The proposed classification scheme does the matrix decomposition via using $L$ decomposition classes, grouping such $EP(N,\psi)$, which have similar features. Generally, this decomposition scheme is given by the matrix $M(r)$ in the equation (6):

$$M(N,\psi) = g_s\{EP(N,\psi)\} \tag{6}$$

In this configuration defined by eq. (6), $gs\{.\}$ stands for the decomposition method, transforming individual levels of $EP$ into $L$ finite classes of the decomposition model. When supposing that the evaluation matrix $EP(N,\psi)$ contains $n$ elements, then we have: $L < n$.

Based on this decomposition scheme, it is supposed that each element of the evaluation matrix belongs to unique decomposition class. The main task in such decomposition procedure is a system of the rules, determining classification values of EP into respective class. In order to this task, we propose a decomposition scheme based on the fuzzy classification driven by the ABC evolutionary algorithm as defined in equation (7). For the following text, we use the notation for fuzzy set:

$$\mu_l(N,\psi),\ l = 1,\ 2,\ \ldots,\ L. \tag{7}$$

### 3.7.5. Definition of fuzzy logic-based decomposition model

In this part, we introduce the principle of the decomposition model, which is based on the system of fuzzy sets theory. The main idea of this method approximates a finite group of the evaluation parameter values by fuzzy set, which classify these values from other in the evaluation matrix. Such task can be in principle done by using histogram approximation based on some statistical distribution, for instance a sequence of Gaussian functions can be adopted for building such decomposition model. Such approach would be linked with the approximation procedure to find and optimize parameters of respective Gaussian distribution. Furthermore, for different situations we would have to recalculate this optimization procedure. Also, it is not ensured that the data behavior well corresponds with the shape of some predefined probability functions. Bases on such reasons, we build this decomposition scheme on the sequence of the trapezoidal fuzzy functions, where their the most suitable distribution is selected on a general fitness function in the optimization evolutionary ABC algorithm.

Fuzzy based decomposition algorithm utilizes the histogram of the evaluation matrix, which is decomposed into $L$ classes. We define the membership function for each region. This membership function determines a level of the membership for each wavelet and respective decomposition level $N$ in each class. This feature is taken as the classification rule for the respective wavelet settings. In the case of the triangular-shaped fuzzy function, we have the following rule for the membership functions: $\sum_{i=1}^{L} \mu_l(N,\psi) = 1$. Based on fuzzy-based classification, we classify respective Wavelet settings into the class with the highest membership value by the formulation in eq. (8).

$$M(N,\psi) = argmax_l\{\mu_l(N,\psi)\} \tag{8}$$

This procedure represents a classification of the Wavelet settings into individual classes based on the soft thresholding. The substantial advantage of such procedure is the fact that the membership function determines a degree of each wavelet to each class in the contrast in the hard thresholding methods, which gives a hard threshold. In this context, this method determines a certain membership in each class, and consequently we just need to set a rule for the final classification as it is depicted in eq. (8). Based on such multiple membership assignment, we compute the membership values for each Wavelet settings by the following way, see equation (9):

$$\mu(EP(N,\psi)) = \left[\mu_1(EP(N,\psi)) \quad \mu_2(EP(N,\psi))\ldots \quad \mu_L(EP(N,\psi))\right] \tag{9}$$

In the case of the triangular function only two elements are non-zero. The decomposition model based on the triangular fuzzy functions respects the following rules:

- **Complete division**: $\forall x,\ \exists\mu_l(x),\ 1 \le l \le L$, so that $\mu_l(x) > 0$
- **Consistency:** if $\mu_l\left(x_0\right) = 1$, then $\mu_k\left(x_0\right) = 0,\ \forall k \ne l$
- **Normality:** $\max\left(\mu_l(x)\right) = 1$
- **Intersection** between adjacent classes: $\mu_l\left(x_0\right) = \mu_{l+1}\left(x_0\right) = 0.5$

The main issue of such soft-thresholding approach is its construction. Each sub fuzzy set is constructed (in geometric sense) by determining its vertex, when utilizing the rule, describing the relationship between the vertex of the class ($\mu_l\left(x_0\right)$) and adjacent foot of triangle ($\mu_k\left(x_0\right)$): $\mu_l\left(x_0\right) = 1$, then $\mu_k\left(x_0\right) = 0,\ \forall k \ne l$. It practically means that we need to find a sequence of the vertex function: $V = \left\{V_1,\ V_2, \ldots, V_L\right\}$, where $V_L$ stands for the vertex of Lth decomposition class. After performing this vertex function, we obtain the decomposition model based on the sequence of the fuzzy triangular sets.

### 3.7.6. Prediction model for estimation of vertex function

In this section, we introduce the model, which is aimed on the prediction of the vertex locations to be consequently used for the fuzzy based classification. For such task various methods may be adopted. One of the simplest ways can be histogram maxima detection, alternatively, methods based on the clustering may be utilized. Such methods are capable of grouping the evaluation values with similar features into groups, and by consequent computing of the centroids of such clusters we can obtain the vertexes for the fuzzy classes. The main argument, which deteriorates the quality of such methods is producing one unique solution of the vertex vector, which is depended on the initial placement of centroids, in the case of clustering algorithm. In the case of the maxima detection, the histogram does not have to have appropriate shape. Thus, identification of the maxima values is depended on the shape of the histogram. Based on such arguments, we employ ABC algorithm, which is based on the evolutionary computing. This method represents an optimization method, which iteratively search for the best solution based on the given criteria. An essential advantage of this method is the fact that we do not need any maxima detector, this proposed method analyzes the class distribution based on the entropy function to find concentrated distribution of the evaluation parameter in respective class, and in the contrast of the clustering, the optimized solution selected from various possibilities based on the optimization process. Here, we describe ABC algorithm for optimization of vertex locations.

ABC evolutionary algorithm is an optimization algorithm, which is inspired by a swarm of bees, searching for the food. In the contrast of natural reality, this method utilizes the synthetic bees, which are classified into four categories: employed bees ($EB$), onlooker bees ($OB$) and scouts ($SB$). The whole population is composed from the same number of $EB$ and $OB$. It is supposed that each $EB$ has one food source. Such food source represents one possible combination of the vertexes. The most optimal vertex location is perceived as the optimization problem: $V_i = V_{i,1},\ V_{i,2}, \ldots V_{i,p}$, where: $V_i$ represents $i^{th}$ possible solution of the vertex combination with p-dimensional parameter, which is optimized. In this situation $p$ stands for the number of the classes, in which the evaluation matrix is decomposed.

In the first part of ABC algorithm (Employed Bees), the initial population of $V_i$ is generated. For such purpose, we propose a special scheme for definition of the initial population of food sources, respective possible initial combinations of vertex functions $V_i$. Since the centroid-based clustering algorithms have been frequently proved as a reliable and powerful methods for the data grouping, therefore we are aimed to use such principle to find the initial clusters of evaluation parameters based on the clustering. The evaluation matrix is initially decomposed based on the FCM (Fuzzy C-means) algorithm. From this algorithm, we just take the centroids of individual clusters, for $L$ decomposition classes we have the vector of centroids: $C = \left\{C_1,\ C_2, \ldots,\ C_L\right\}$. Consequently, we use the generator of normal numbers with the Gaussian distribution probability to generate the initial population of the food sources, where the mean value of this distribution is taken centroid of each class, computed by FCM method. This principle is better explained by the following pseudo algorithm.

Based on this procedure, we generate a system of random vertex locations $V_N$, containing N-combination of the vertexes, which are normally distributed.

Consequently, in the $EB$ phase, for each vertex solution in
$V_N = \left\{V_{N,1},\ V_{N,2}, \ldots, V_{N,N}\right\}$, we generate the alternative solution $X_N$, which is given by the eq. (10).

$$X_{ik} = V_{ik} + \phi_{ik} \times (V_{ik} - V_{jk}) \tag{10}$$

**Algorithm 2** Pseudo-code for the initialization and computation of centroids.

**Initialization:** definition of variables:
$N$ (number of food sources), $\sigma^2$ (variance of Gaussian distribution),
$L$ (number of decomposition classes), $EP$ (wavelet matrix of evaluation parameter)
**Output $\mathbf{V_N}$:** (initial combination of food sources), $C$ (vector of initial centroids)

```
C = compute FCM(EP, L)
for each element in C do
    compute $\mathbf{V_N} = rand(N_{PDF}(\mu = C,\ \sigma^2))$
end for
```

In this formulation, $V_{jk}$ represents a random candidate solution $(i \neq j)$, $k$ stands for the random index: $k \in \{1,\ 2, \ldots, L\}$, and function $\phi_{ik}$ represents a generator of random numbers. As soon as $X_N$ is generated, it is approached to the selection based on the fitness function $(fit_N)$ between the solution $X_N$ and $V_N$. In the case when $fit_{X_N} > fit_{V_N}$, then $X_N$ is stored as a better solution. Otherwise a new random solution $X_N$ is generated and selection is repeated. This repetition process is limited by the selection limit $L_v$ (we experimentally use $L_v = 10$). After reaching $L_v$, the respective food source $X_N$ is perceived as exhausted and is eliminated from the memory.

The second phase of the methods tackle with onlooker bees (OB), which globally evaluate individual vertex solutions. This procedure is based on the roulette selection, which is perceived as a probabilistic selection mechanism and given by the term in eq. (11).

$$P_i = \frac{fit_i}{\sum_{j=1}^{N} fit_j} \tag{11}$$

In this selection mechanism, $fit_i$ stands for the fitness function of $i^{th}$ solution. As well as in the EB phase, the higher value of the fitness function is achieved, the better solution is then indicated. Roulette selection is performed iteratively, where number of the iterations is controlled by the roulette limit $R_l$ (we experimentally use $R_l = 20\%\ of\ V_N$).

The last phase of ABC algorithm tackle with the scouts (SB). Each scout searches for a new food source instead of exhausted one. When denoting the exhausted food source as $V_{exh}$, then scout discovers a new one food source instead of the exhausted. Consequently, the whole evaluation process in ABC algorithm is repeated within predefined number of the iteration cycles (NC). A new food source in the scout phase is given by the eq. (12).

$$V_{ik} = lb_j + rand(0, 0.1) \times (ub_j - lb_j) \tag{12}$$

Where $rand(0, 0.1)$ represents a random value from the range $[0; 0.1]$. The parameters $lb$, $ub$ stand for the lower and upper limit of the parameter space, for instance in the case of the correlation coefficient we use the range: $[0; 1]$. The final output of the ABC optimization is such $V_i$, for which we can state: $\max\left(fit_{V_i}\right)$, $\forall i \in N$.

### 3.7.7. Definition of fitness function

In the previous sections, we described the optimization methodology based on ABC algorithm with the goal to find a decomposition scheme of evaluation matrix, representing a spatial distribution of evaluation coefficients for the most optimal Wavelet setting selection. An essential element of this evolutionary approach is a fitness function. This function serves for the evaluation of each possible combination of the triangular fuzzy sets, consisting the decomposition model. This fitness function is aimed on bringing a global information about the quality of each configuration of fuzzy triangular sets. We use the fitness function, which is based on the distribution of the parameter values in each triangular fuzzy set. The essential idea is forming parameters values inside each fuzzy set close to the shape of Gaussian distribution. Such presumption lead to decomposition regions with concentrated parameter values, with minimal outliers. In order to do this task, we firstly calculate the probability of each evaluation parameter value (eq. (13)).

$$p_k = \frac{par(k)}{\sum_{k=0}^{L-1} par(k)} \tag{13}$$

In this approach, we suppose that the evaluation matrix contains L parameter values (number of Wavelet settings), where: $k = 0, 1, 2,\ \ldots, L-1$, and $p_k$ stands for the probability of the parameter value $par(k)$. Each of the decomposition class is represented by the three parameters: two marginal thresholds t and the centroid $V$, such expression can be written as: $\left\{t_{i,1}\ V_i,\ t_{i,2}\right\}$ for $i^{th}$ decomposition class. Supposing the decomposition model contains $p$ classes, then we construct the definition vector for triangular fuzzy sets system (eq. (14)).

$$T = \left[\left\{t_{1,1}\ V_1,\ t_{1,2}\right\},\ \left\{t_{2,1}\ V_2,\ t_{2,2}\right\}, \ldots, \left\{t_{p,1}\ V_p,\ t_{p,2}\right\}\right] \tag{14}$$

Gaussian distribution has the ability to maximize entropy. Since we are aimed to approximate individual decomposition classes with Gaussian distribution, we use the concept of entropy maximization as the main criteria for the fitness function. In this context, Kapur's entropy enables measuring the compactness and separability of the decomposition classes. Supposing that we construct p-decomposition model, we use the following system of Kapur's entropy functions for all the decomposition classes (eqs. (15) (16) (17) (18)).

**Table 1**

Grid-Search Model Evaluation. **RoD** - Range of Days. **Accuracy (Acc)** - Model accuracy with the test set. **Accuracy CV** - Model accuracy with cross-validation. **MSE** - Mean squared error of the model with the test set. **MSE CV** - MSE of the model with cross-validation. **RMSE** - Root-mean-square error of the model with the test set. **RMSE CV** - RMSE value of the model with cross-validation.

| RoD | Acc. | Acc. CV | MSE | MSE CV | RMSE | RMSE CV | Month |
|---|---|---|---|---|---|---|---|
| 1 | **0.9932** | 0.9825 | 0.0067 | -0.0007 | 0.0260 | -0.0367 | April |
| 3 | 0.9622 | 0.9647 | 0.0026 | -0.0025 | 0.0507 | -0.0490 | April |
| 7 | 0.9036 | 0.8929 | 0.0059 | -0.0067 | 0.0766 | -0.0490 | April |
| 14 | 0.8827 | 0.8834 | 0.0061 | -0.0065 | 0.0780 | -0.0813 | April |
| 30 | 0.8263 | 0.8029 | 0.0096 | -0.0105 | 0.0981 | -0.1025 | April |
| 1 | 0.9863 | **0.9863** | **0.0005** | **-0.0005** | **0.0231** | **-0.0239** | March |
| 3 | 0.9525 | 0.9424 | 0.0019 | -0.0022 | 0.0431 | -0.0467 | March |
| 7 | 0.9036 | 0.9044 | 0.0040 | -0.0039 | 0.0632 | -0.0622 | March |
| 14 | 0.8990 | 0.9312 | 0.0041 | -0.0028 | 0.0637 | -0.0527 | March |
| 30 | 0.8708 | 0.8823 | 0.0060 | -0.0053 | 0.0778 | -0.0727 | March |

$$H_1 = -\sum_{i=t_{1,1}}^{t_{1,2}} \frac{p_i}{\omega_1} \ln\left(\frac{p_i}{\omega_1}\right), \omega_1 = \sum_{i=t_{1,1}}^{t_{1,2}} p_i \tag{15}$$

$$H_2 = -\sum_{i=t_{2,1}}^{t_{2,2}} \frac{p_i}{\omega_2} \ln\left(\frac{p_i}{\omega_2}\right), \omega_2 = \sum_{i=t_{2,1}}^{t_{2,2}} p_i \tag{16}$$

$$H_{p-1} = -\sum_{i=t_{p-1,1}}^{t_{p-1,2}} \frac{p_i}{\omega_{p-1}} \ln\left(\frac{p_i}{\omega_{p-1}}\right), \omega_{p-1} = \sum_{i=t_{p-1,1}}^{t_{p-1,2}} p_i \tag{17}$$

$$H_p = -\sum_{i=t_{p,1}}^{t_{p,2}} \frac{p_i}{\omega_p} \ln\left(\frac{p_i}{\omega_p}\right), \omega_p = \sum_{i=t_{p,1}}^{t_{p,2}} p_i \tag{18}$$

Based on the Kapur's entropy, individual fuzzy sets in p-decomposition model are specified. When taking advantage the fact to maximize the entropy function, we are searching such solution within the ABC optimization process which satisfy the following fitness function (eq. (19)). Based on the stated presumptions, the higher values of the fitness function we achieve, the better solution we obtain.

$$fit_i = \arg max\left(\sum_{k=1}^{p} H_k\right) \tag{19}$$

## 4. Results

### 4.1. Model performance

After completing the tests, the results are presented in both tabular and graphical formats. This enables easy visualization of the model's performance through comparisons between predicted outcomes and actual values. Below are the results obtained using the Grid-Search method using the first set of hyperparameters (Kernel = Radial basis function (RBF); C = 32; Gamma = 8; Epsilon = 0.0313), shown below in the following Table 1. It is important to note that the negative values for MSE and RMSE obtained through cross-validation are due to the scoring function used. In this case, a combined score is applied, which is always maximized. Therefore, when a test result must be minimized, as in this case, the score is negated for the scoring function to work correctly. Consequently, the returned score is negated when it must be minimized, while it is maximized.

Proceeding with the tests, they have been performed using Random-Search Model Evaluation with the set hyperparameters (Kernel = RBF; C = 86.8032; Gamma = 46.0693; Epsilon = 0.0313), see Table 2.

It can be seen how the precision of the model varies considerably using both methods. Thus, it is much more accurate when predicting intervals between one day and three days. However, as the number of days increases, the model begins to decrease the precision. The best values obtained have been for the second set of hyperparameters, in which we can observe the best values obtained for the set of days.

### 4.2. Implementation of the practical experiments

All the measurement were conducted using operating sensors with specific characteristics. For indoor $CO_2$ measurements, the sensors offer a range of 0 to 2000 $ppm$, and accuracy of less than $50ppm + 2\%$ at 25 °C and 1013 $mbar$, and a temperature dependency of 2 $ppm$ $CO_2/°C$. In terms of indoor and outdoor temperature measurements, the sensors have an accuracy at 20 °C and a range of 0 to 50 °C. Lastly, for indoor relative humidity measurements, capacitance-type sensors were used, which provide a range of 0 to 100% $RH$, and an accuracy of $3\%RH$ at 20 °C. Implementation, construction, and verification of the SVR model are carried out as follows:

**Table 2**
Random-Search Model Evaluation. **RoD** - Range of Days. **Accuracy (Acc)** - Model accuracy with the test set. **Accuracy CV** - Model accuracy with cross-validation. **MSE** - Mean squared error of the model with the test set. **MSE CV** - MSE of the model with cross-validation. **RMSE** - Root-mean-square error of the model with the test set. **RMSE CV** - RMSE value of the model with cross-validation.

| RoD | Acc. | Acc. CV | MSE | MSE CV | RMSE | RMSE CV | **Month** |
|---|---|---|---|---|---|---|---|
| 1 | **0.9939** | **0.9824** | **0.0006** | **-0.0015** | **0.0245** | **-0.0366** | April |
| 3 | 0.9646 | 0.9690 | 0.0024 | -0.0022 | 0.0490 | -0.0461 | April |
| 7 | 0.9333 | 0.9045 | 0.0041 | -0.0059 | 0.0637 | -0.0768 | April |
| 14 | 0.9069 | 0.8988 | 0.0048 | -0.0057 | 0.0695 | -0.0750 | April |
| 30 | 0.8611 | 0.8285 | 0.0077 | -0.0092 | 0.0877 | -0.0955 | April |
| 1 | 0.9604 | 0.9480 | 0.0015 | -0.0022 | 0.0391 | -0.0466 | March |
| 3 | 0.9055 | 0.9236 | 0.0037 | -0.0029 | 0.0608 | -0.0535 | March |
| 7 | 0.9060 | 0.8809 | 0.0039 | -0.0049 | 0.0623 | -0.0692 | March |
| 14 | 0.9120 | 0.9369 | 0.0035 | -0.0026 | 0.0595 | -0.0507 | March |
| 30 | 0.9127 | 0.9096 | 0.0041 | -0.0041 | 0.0639 | -0.0636 | March |

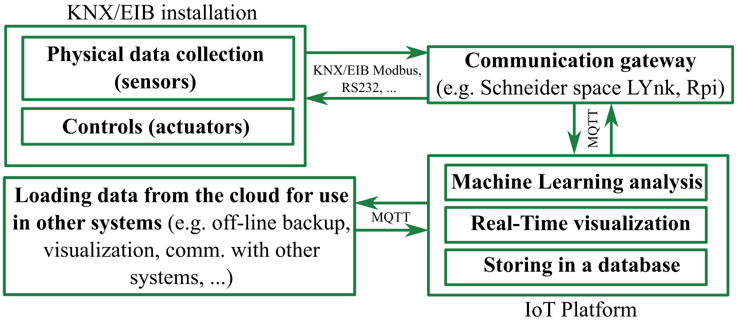

KNX/EIB installation
Physical data collection (sensors)
Controls (actuators)
KNX/EIB Modbus, RS232, ...
Communication gateway (e.g. Schneider space LYnk, Rpi)
MQTT
Machine Learning analysis
Real-Time visualization
Storing in a database
IoT Platform
MQTT
Loading data from the cloud for use in other systems (e.g. off-line backup, visualization, comm. with other systems, ...)

**Fig. 2.** Block scheme of data collection within Intelligent building automation.

Step 1: Measurement of non-electrical variables (Fig. 2) Temp1, Temp2, rH, and $CO_2$ at the Ostrava Faculty of Electrical Engineering and Computer Science (FEI) during two periods: March 4 to March 31, 2019, and April 1 to April 25, 2019.

Step 2: Data preprocessing by removing correlated elements, outliers, missing values, normalization, and duplicate patterns.

Step 3: Compare the model to be implemented along with other algorithms to compare its performance.

Step 4: Optimization of hyperparameters of the model to be built to obtain higher performance.

Step 5: Training of the SVR model, for the parameters obtained through the search methods explained in previous sections.

Step 6: Evaluation of the model using the described metrics, as well as through the use of cross-validation.

Step 7: Evaluation of the results achieved.

### 4.3. SVR 1

The tested data (spanning the same timeframe) for March 4th to March 5th, 2019 is shown in Table 2. The best results were obtained with the predicted course of $CO_2$ for SVR using Grid-Search hyperparameters, as indicated in the image and employing a five-fold cross-validation, which yielded an $MSE = 0.023\ ppm$ and an $R^2 = 0.98$.

### 4.4. SVR 2

From April 1st to April 2nd, 2019, the tested data (spanning the same time frame) is shown in Table 1 and Table 2. The best results were obtained with the predicted course of $CO_2$ for SVR using Random-Search hyperparameters, as indicated in the image and employing a five-fold cross-validation, which yielded an $MSE = 0.0365\ ppm$ and an $R^2 = 0.982$.

### 4.5. Days prediction

**Predicting one day:** The accuracy of both models for single-day predictions is 0.99 with a test set and 0.98 with cross-validation. Both models have $RMSE$ values of 0.02 with cross-validation in April. Nevertheless, the results for March are lower, with Grid-Search achieving 0.98 compared to Random-Search 0.94, and $RMSE$ values of 0.02 with Grid-Search and 0.04 with Random-Search.

**Predicting three days:** The three-day predictions show a similar pattern to the previous model, with both achieving an accuracy of 0.96 using test set and cross-validation. Additionally, the $RMSE$ values of both models are comparable, with 0.04 in cross-validation, indicating a good fit to the data. The results have been better for April, regardless of the search method used. The best outcome achieved an $R^2 = 0.96$ and an $RMSE = 0.04\ ppm$. In contrast, for March, the optimal result was an $R^2 = 0.92$ and $RMSE = 0.05\ ppm$, obtained using the hyperparameters derived from Random-Search.

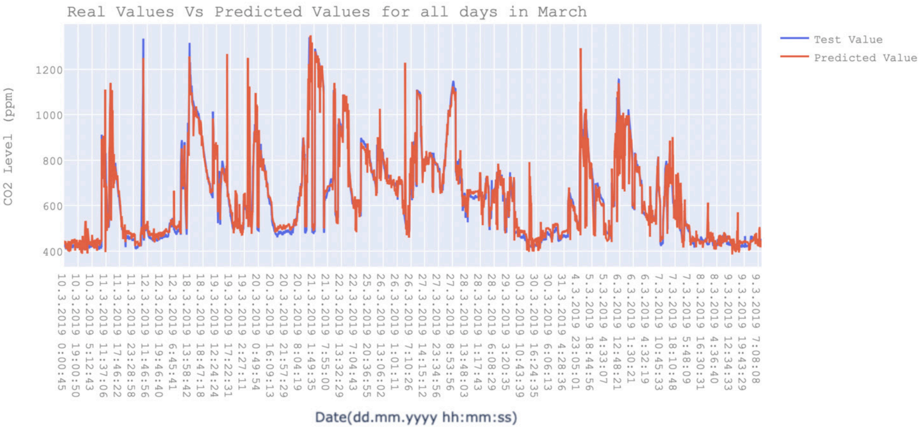


**Fig. 3.** Prediction obtained for March.

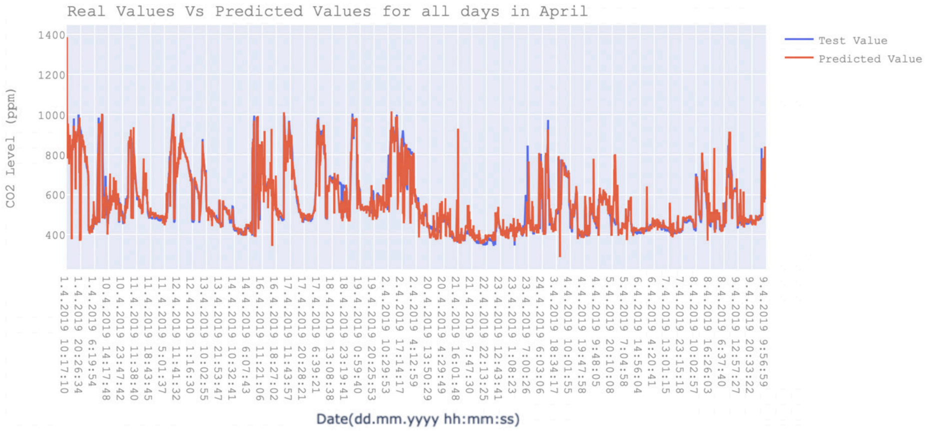


**Fig. 4.** Prediction obtained for April.

**Predicting seven days:** In the seven-day prediction, we found a difference between the different methods, since the parameters obtained through Random-Search, have achieved a higher precision than that of Grid-Search, being this of 0.904, as well as a lower $RMSE$ than that of Grid-Search with a value of 0.07. For April, the best prediction obtained was $R^2 = 0.90$ and $RMSE = 0.06/ppm$. By using the hyperparameters obtained through Grid-Search, the best result obtained for March was $R^2 = 0.90$ and $RMSE = 0.06/ppm$.

**Predicting fourteen days:** In the fourteen-day prediction range, Random-Search achieved similar results through cross-validation, attaining an $R^2$ of 0.936 and an $RMSE$ of 0.05 for March. These results are comparable to the values obtained with Grid-Search for the same month. However, for April, the outcomes have decreased to 0.89 with an $RMSE$ of 0.07.

**Predicting a month:** For the prediction of a full month, the best result has been obtained with Random-Search, with a validation precision of 0.9; however, the $RMSE$ committed is very low, at 0.06. For March, the predictions were better as shown in Fig. 3.

For April, the best result obtained was $R2 = 0.82$ and $RMSE = 0.09$ $ppm$, by using the hyperparameters obtained with Random-Search, as is shown in the next Fig. 4.

### 4.6. Results summary

This study describes an approach applying support vector machines to monitor the occupation of the designated space, in this case, the Ostrava Faculty of Electrical Engineering and Computer Science (FEI). The first part of the study explains the mathematical procedure on which the operation of SVM is based, as well as the implementation of the KNX system in an IoT environment since the data processed were obtained using the KNX system and the use of sensors temperature $\circ C$, relative humidity (% RH) and $CO_2$ sensors ($ppm$) (Fig. 2). As previously mentioned, the majority of the time spent in this study was dedicated to data processing. The reason for this is that dealing with repeated values, missing values, outliers, and attributes on different scales requires substantial effort in order to achieve favorable results. Our next step was to optimize the SVR algorithm by searching for the optimal hyperparameters for our problem. Furthermore, we trained our model using 80% of the patterns from the April dataset. This dataset was chosen because it yielded the best results in the search and subsequent tests. A standard scaling has been used. However, in the initial tests, the robust Scaler provided better results. However, the computation time for obtaining the hyperparameters, as well as the execution of the SVR algorithm increases dramatically.

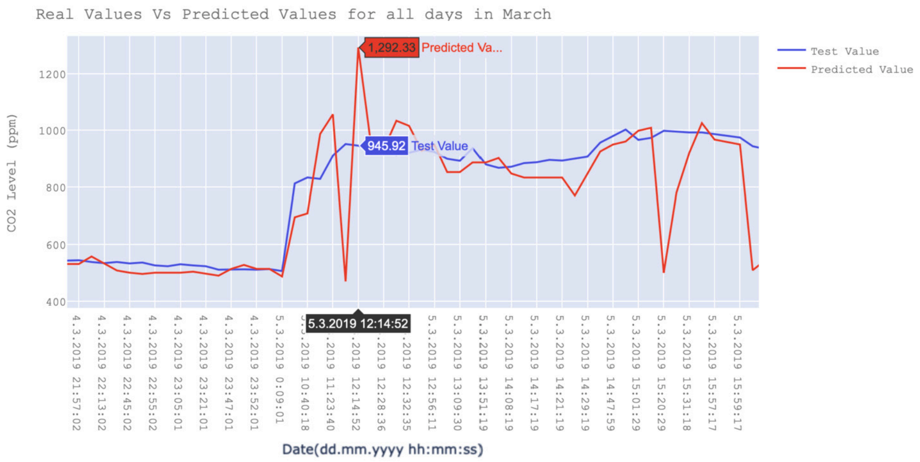


**Fig. 5.** Prediction Peaks.

Following optimization, as demonstrated in Table 1 and Table 2, it has been possible to improve the initial result from 0.67 to a precision close to and even greater than 0.90. Additionally, the Root Mean Squared Error (RMSE) metric was employed using cross-validation to verify the effectiveness of our model with previously unseen data. Lastly, the predicted ppm values were plotted against the actual values, resulting in a graph with denormalized values, which helped us understand the real values that were predicted. It can be seen how at the most extreme peak, as shown in Fig. 5, it corresponds to March 5 at nine and 12:14, obtaining a high predicted $CO_2$ level with a value of 1292 $ppm$; however, the real value is 945 $ppm$. So it may be a time when the windows and doors are closed, with the ventilation system turned off. It can be seen how, before the abrupt ascent to the maximum peak since 9:00, the $CO_2$ levels are deficient, being around 450 to 500 $ppm$. To conclude, this increase is likely due to the presence of people in the room as well as the closed and off ventilation systems, as previously mentioned. The concentration of $CO_2$ is medium before the peak. The presence of people in the room may remain constant for the next few hours, before slowly decreasing again. After that, it rises again. We assume that the room will remain empty until 14:00, with ventilation systems turned off and windows closed.

Moreover, To properly control HVAC (Heating, Ventilation, and Air Conditioning), three indoor air parameters must be measured - temperature, relative humidity, and $CO_2$ concentration. As a result of breathing, every individual releases $CO_2$. By measuring $CO_2$ concentrations, it is possible to obtain reasonably accurate information regarding the presence of people in an enclosed space or the time of their arrival and departure from the monitored space (Fig. 6). $CO_2$x is a natural gaseous component of atmospheric air and is odorless [32]. Higher concentrations of $CO_2$ may result in drowsiness, fatigue, headaches, and nausea, as well as a loss of concentration. It is therefore necessary to supply fresh air regularly by ventilating the living space [33]. In addition to volatile organic compounds (VOCs), dust, microorganisms, water vapor, and radon, other factors also affect indoor air quality. However, measuring all of these components would be relatively uneconomical. Therefore, it is possible to obtain information about the measured quantities indirectly. As an example, the article presents a newly proposed method for predicting $CO_2$ concentration from measurements of relative humidity and indoor temperature using a wavelet transform to remove additive noise from the predicted $CO_2$ waveform [34]. Information about the occupancy of the monitored spaces (Fig. 6) can be analyzed retrospectively to set the optimal HVAC (heating, ventilation, and air conditioning) control in the living space [35]. Additionally, occupancy information can be used to monitor the activity of the occupants of the apartment (opening a window causes a sharp drop in $CO_2$ concentrations, increasing $CO_2$ concentrations means a greater number of people in the monitored space, etc.) [36].

### 4.7. *Results of wavelet-based recommendation system*

In this section, we present the experimental results of the proposed recommendation system for $CO_2$ signal prediction. For the testing of the proposed system, we used historical data of the predicted $CO_2$ signals. We used the predictions from April and March 2019, where we tested the predictions for 1, 3, 7, 14 days and the whole month. So, we provide testing together for 10 $CO_2$ signals. As an example Fig. 6, we provide a comparison between the gold standard $CO_2$ signal and predicted signal based on SVM for one day. We can gather information about the arrival and departure of individuals in the monitored area by analyzing the predicted $CO_2$ values and the $CO_2$ values obtained through wavelet filtering for additive noise removal. Moreover, we can determine the time intervals during which people were present in the area, denoted as $\Delta t\ a = (t1 - t2)$, $\Delta t\ b = (t3 - t4)$, and $\Delta t\ c = (t5 - t6)$. Time stamps are as follows: $T1$ (arrival), $T2$ (departure), $T3$ (arrival), $T4$ (departure), $T5$ (arrival), $T6$ (departure), $T7$ (arrival), and $T8$ (departure).

### 4.8. *Analysis of spatial maps*

In this part, we present experimental results of the spatial $(2D)$ mapping of Wavelet distributions. As we described in the previous text, this approach enables to perform a complex distribution of multiple wavelet response for given $CO_2$ signal, or group of the signals. Another important fact is a versatility of the proposed scheme in the context of any number of wavelets from one or multiple

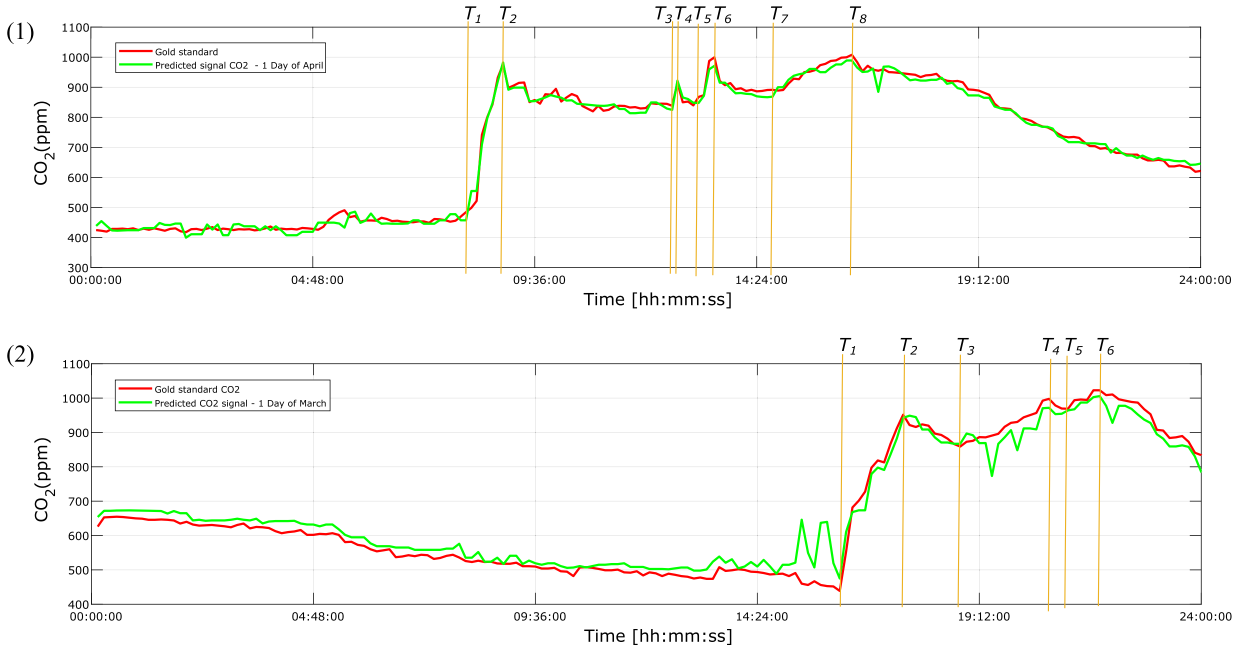
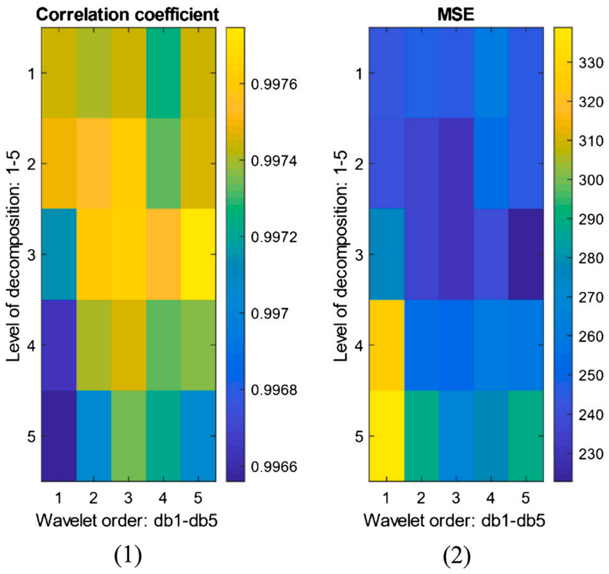


**Fig. 6.** Comparison of predicted one-day $CO_2$ signal from April (1) and March (2) with the gold standard signals.



(1) (2)

**Fig. 7.** Spatial modeling of wavelet response for one-day April prediction for wavelets db1-db5: corelation spatial matrix (1) and MSE spatial matrix (2).

wavelet's families and levels of decomposition. We determine the wavelet response distribution by using spatial color mapping, which allows us to distinguish individual levels of the response. Fig. 7 shows experimental results of the correlation coefficient and mean square error (MSE) for selected wavelets in the Daubechies family (db): db1-db5 in Fig. 7, and db1-db20 in Fig. 8. These results are shown for both one-day and one-month predictions. When compared to other families, the Db family demonstrates better performance in terms of lower MSE values and higher correlation indices between the gold standard $CO_2$ signal and the respective prediction from SVM. Analyzing the spatial maps Fig. 8, we found comparatively different MSE values from shorter (one-day) prediction and a longer time period (month-prediction), where MSE values are approximately ten-times higher. This fact predetermines substantial differences also in effectivity of SVM prediction for various length of $CO_2$ signal prediction.

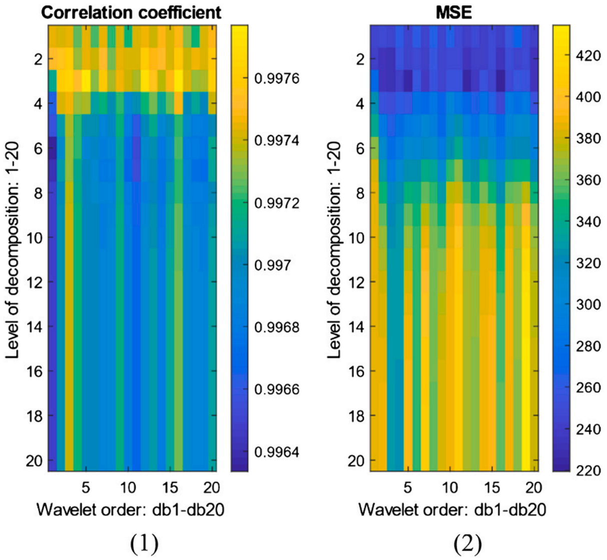


**Fig. 8.** Spatial modeling of wavelet response for one-day April prediction for wavelets Db1-Db20: corelation spatial matrix (1) and MSE spatial matrix (2).

### *4.9. Features analysis from spatial map*

Based on the spatial maps, we can evaluate the $CO_2$ signal distribution in arbitrary $2D$ domain. Such spatial maps allow subjective evaluation regarding particular evaluation parameter as a level of quality prediction. In order to provide a robust quantitative analysis, we extracted statistical features, which can evaluate trend and distribution of wavelet features among various settings, signal length and individual tested families. Based on the experimental testing, we found that Db family appears as effective for $CO_2$ signal prediction – we provide later a justification of this statement. Thus, we firstly provide a comparison of histograms for the tested predictions from April and March. We provide a comparison of histograms, representing the distribution of MSE values for the wavelets: Db1-Db5, with decomposition levels: 1-5 (Fig. 9), and Db1-dB20, with decomposition levels 1-20. Based on the results, we can conclude that individual distribution for the same time period is separate. That means that we receive a significantly different MSE values for various tested months. This fact is better observable in shorter predictions. In the case of one-month prediction for April, both histograms are partially overlaid. Thus, in this case, there are slighter difference between both predictions. These distributions also well illustrate the difference in MSE values in the dependence on the time of prediction. It is notable that for shorter predictions we receive smaller MSE values when comparing with for instance one-month prediction. This fact brings a modeling of the SVM accuracy for different time-period of prediction.

Lastly, we provide the characteristic (Fig. 10) of the globally best wavelet settings among all the tested wavelet families. This part of the analysis we consider as the most important from the view of appropriate wavelet selection. Here, we must note that this characteristic brings the information about the comparison among individual tested wavelet families, and not particular wavelet settings, appearing as the best for particular $CO_2$ signal. In this comparison, we constructed the spatial wavelet distributions for the whole families: Daubechies, Symlet, Coiflet, Biorthogonal, Reverse Biorthogonal, Shannon and Gaussian. For each of these settings, we selected the minimal MSE values, representing the global best evaluation of respective family. When analyzing the results in Fig. 10, it is also obvious an increasing trend of MSE values for all the tested wavelet settings. Here, we can objectively justify that Db family (red mark) is comparatively better to others. Therefore, it is obvious this family would be the most suitable for the $CO_2$ signal prediction, judging by MSE evaluation.

### *4.10. Classification of wavelet response*

In the previous analysis of the results of the enhancement of $CO_2$ signal prediction, we studied the behavior of individual wavelet families and the prediction period on the signal prediction accuracy. As we report earlier, Db family appears as the most optimal based on the spatial map and consequent feature extractions. On the other hand, such analysis does not bring the information which particular wavelet settings (mother's wavelet and decomposition level) appears as the best to be used in the proposed prediction system. The major problem is we can expect various the best wavelet settings for different $CO_2$ signals. That is a complication in the context of selection unique wavelet settings for arbitrarily $CO_2$ signal prediction. In order to build a robust prediction system, we employed a classification procedure to select a set of the most suitable wavelet settings, satisfying any $CO_2$ signal, regardless its time prediction.

This classification procedure performs the matrix decomposition based on the fuzzy soft thresholding driven by evolutionary ABC algorithm. This decomposition procedure classifies the spatial wavelet response for any evaluation parameter into predefined number of classes, representing the quality of wavelet response. In this study, we use number of classes $(L = 3)$. This configuration performs

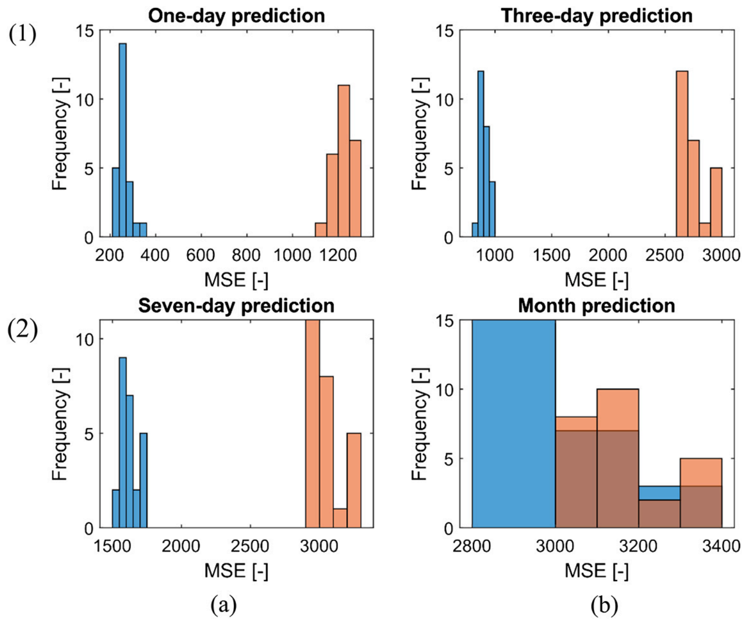


**Fig. 9.** Comparison of histograms for April (blue) and March (red) of MSE distributions for the wavelet settings: Db1-Db5 and level of decomposition: 1-5. One-day prediciton (1a), three-day prediction (1b), seven-day prediction (2a), and month prediction (2b).

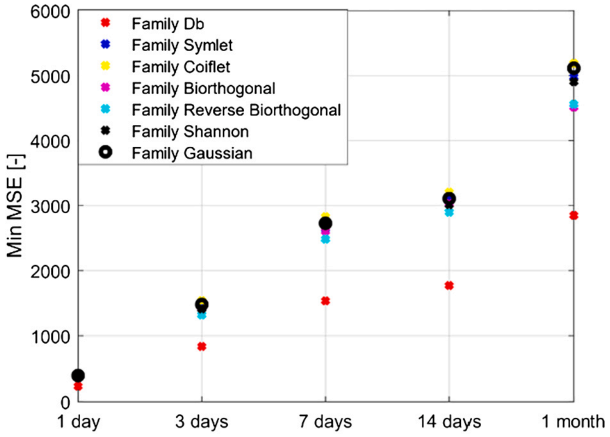


**Fig. 10.** A comparison of individual tested families for various time prediction, where data from April and March were averaged.

the decomposition into the least suitable, neutral and the most suitable settings. We tested the proposed system for the number of food sources ($N = \{100,\ 200, 500, 1000, 2000, 5000\}$) and the number of iterations ($NC = \{100,\ 500, 1000, 5000\}$). Fig. 11 represents an example of the decomposition procedure for the spatial map of MSE parameter see Fig. 11 (1). This decomposition performs the spatial map labeling, according to the classification of wavelet settings into individual classes, an unique color is selected for a respective class to visually distinguish individual settings. Blue mark indicates the class with the most suitable wavelet settings (the lowest values of MSE), green neutral settings and blue the least effective settings. Consequently, we present a binary classification (see Fig. 11 (2)), where we selected the class with the lowest MSE values as logical one, and others are suppressed as logical zero.

ABC optimization algorithm is linked with initial settings of the parameters $NC$ and $N$, as we mentioned earlier. Thus, it is needed to evaluate the most suitable settings for using ABC algorithm for finding the optimal wavelet settings. ABC algorithm classifies the most suitable wavelet sets for each tested $CO_2$ signal (1, 3, 7, 14-day and month prediction). For each such signal, we receive a finite set of classified mother's wavelets and decomposition levels.

Since we need to achieve a unified set of these wavelet settings, we compute the intersection of these best wavelet settings for to find mutual wavelet settings, suiting all the signals periods. This operation is done for all the ABC settings. Quality of each such ABC settings is evaluated based on the average MSE of classified wavelet with the gold standard. Logically, the most suitable wavelet settings are that, which minimize MSE function. In order to evaluate these settings, we constructed spatial evaluation matrix: $ABC_{MSE}(N, NC)$, representing the MSE distribution for all the testing settings. Based on the experimental results, we select: $N = 1000$ and $NC = 500$ as the best compromise, minimizing the MSE function.

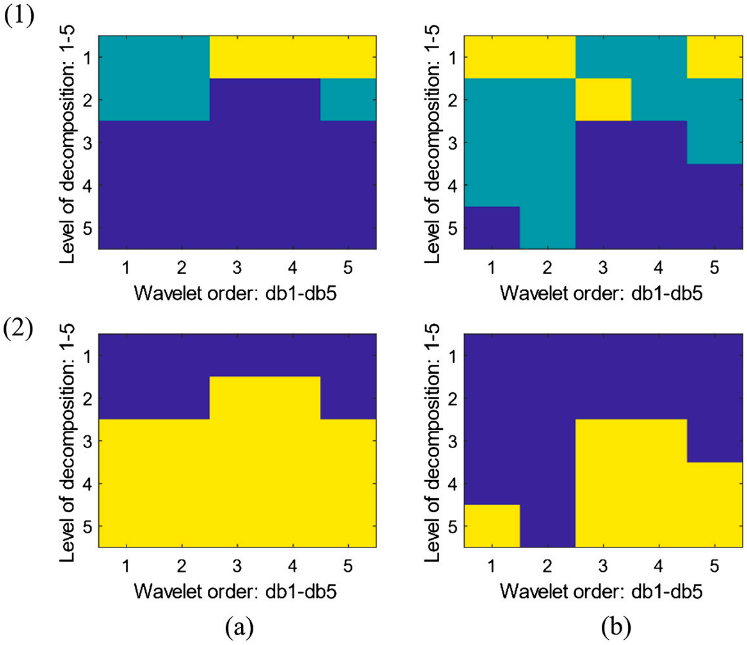


**Fig. 11.** An example of decomposition matrix for wavelet response classification for 14-day prediction in April (a) and 14-day in March (b) based on MSE – upper row (1), and the binary classification of the wavelet response (2).

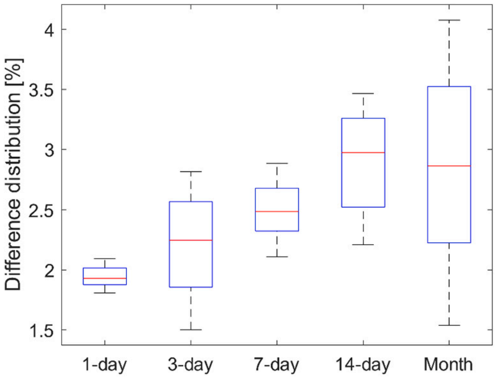


**Fig. 12.** Difference distribution for individual prediction periods for wavelet settings: Db5 and level 4.

For this configuration, we found wavelets Db5 with decomposition level 4 as the most effective with minimal MSE function against the gold standard $CO_2$ signal. As we selected the most suitable wavelet settings, we objectivize a difference distribution between individual smooth $CO_2$ signals (Db5, level 4) and its gold standards. We constructed these difference distributions for all the samples in both signals: wavelet prediction and gold standards (Fig. 12). Judging by the results, comparable differences can be noted for various time predictions. It is obvious that the lowest difference is achieved for one-day prediction, contrarily the month prediction exhibits higher differences. Completely, all the difference distributions show differences below 4%, which can be considered as satisfactory results.

As an example of the smoothing procedure, we provide the example (Fig. 13) of testing the wavelet smoothing prediction on real $CO_2$ signal. For the comparison, we show the gold standard signals from April prediction for one day and the whole month. We noted only neglectable difference between both signals, leading to improving the prediction accuracy.

Lastly, we provide a comparative analysis (Fig. 14) of the MSE evaluation between the original predictions for all the tested time periods and the smoothing procedures (Db5 and decomposition level 4). Here, we provide a percentual difference between individual predictions before and after the smoothing procedure. We can notice slightly different results between April and march predictions. That points out relatively robust wavelet settings for various datasets. For the predictions $1-14$-period, we obtained a decreasing tendency. Nevertheless, the interesting fact is a relatively high difference in month prediction (around 9%) where the original prediction was the most inaccurate compared with other periods, and the smoothing procedure was able to significantly improve the prediction accuracy.

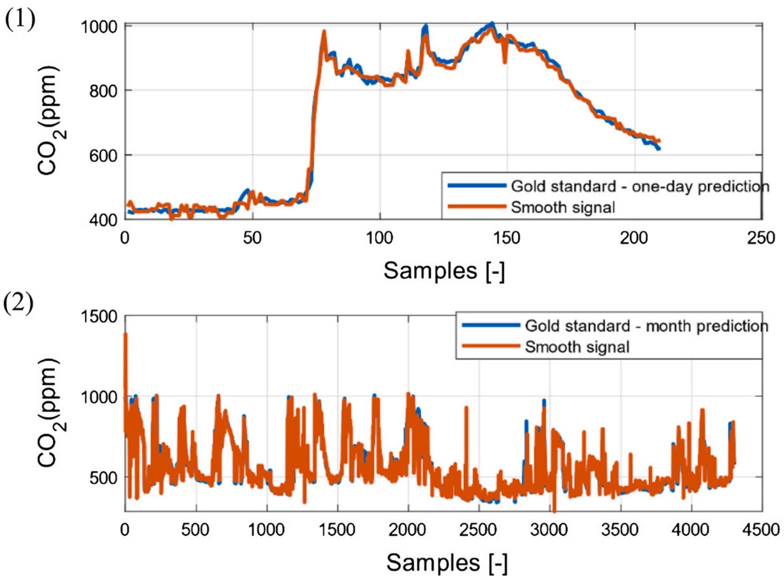


**Fig. 13.** An example of the gold standard $CO_2$ signal and predicted signal with wavelet-based smoothing with Db5 and decomposition level 4 for one-day prediction (1) and month prediction (2).

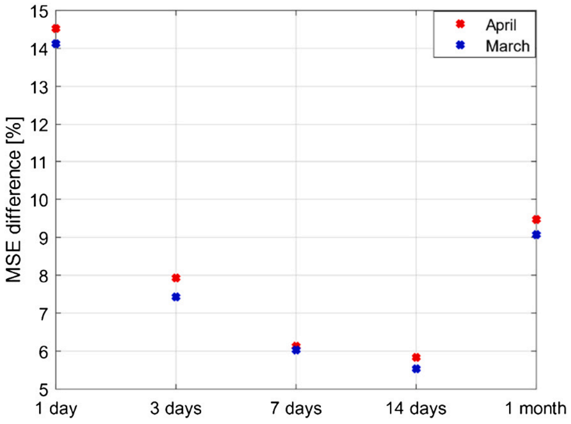


**Fig. 14.** Representation of MSE distribution for various time-predictions between original data and smooth signals (Db5 and decomposition level 4).

## 5. Discussion

In this paper, we proposed a combined method for improving and analyze the SVM prediction of $CO_2$ signals in the concept of the smart home. We used the tested data for different time periods, where for all the records we use the gold standard to evaluate the quality of the prediction. We use 1, 3, 7, 14-day, and month predictions for testing the proposed prediction system. In this prediction system, we utilize the features of Wavelet transformation for data smoothing. Wavelet transformation is a versatile approach, offering a lot of settings especially the selection of mother's wavelets and decomposition levels. These parameters are essential for the effect of wavelet-based smoothing.

Instead of randomly selecting combinations of wavelet settings to be used for the prediction enhancement, we proposed a versatile scheme, representing a spatial 2D distribution of wavelet response of arbitrarily wavelet settings based on selected evaluation parameters (in this study, we use correlation index and MSE). These spatial characteristics allow for simultaneously analyze multiple wavelet features and decide between more and less suitable wavelet settings. Based on the features extraction of such spatial wavelet distributions, we found that Daubechies wavelets appear as the most suitable for $CO_2$ signals prediction enhancement, judging by the lowest MSE values and the highest correlation index when comparing with other wavelet families. Therefore, for further analysis, we only use Db family settings for signal smoothing.

The main task in this analysis is the selection of single, or multiple wavelet settings, which approximate the $CO_2$ signal trend with minimal error functions against the gold standard signals. This task is done based on the feature extraction of the proposed spatial maps. In order to extract the most suitable wavelet settings, we employ the classification procedure based on the matrix decomposition via fuzzy soft thresholding driven by ABC evolutionary optimization procedure. We use three decomposition classes,

**Table 3**
Comparison with state-of-the-art approaches for occupancy estimation and $CO_2$ sensing.

| Topic of the Article | Observations | Accuracy (%) |
|---|---|---|
| Hidden Markov models (HMM) in occupancy estimation: A novel methodology applied to the study case of occupancy detection [38] | Occupancy estimation and detection, HMM | 97.8 |
| A Hybrid of Interactive Learning and Predictive Modeling for Occupancy Estimation in Smart Buildings [39] | Occupancy estimation, classification - predictive distribution, generalized Dirichlet (GD), set of small and simple nonintrusive sensors | 87.5 |
| Human activity recognition (HAR), activities of daily living (ADL) [42] | motion detectors [PIR] and door entry sensors, novel method based on different entropy measures | 99.1 |
| Occupancy-Driven Energy-Efficient Buildings Using Audio Processing with Background Sound Cancellation [43] | occupancy information in real time, environmental sounds on the recorded voice sounds of humans, background sound cancellation algorithm | 80.8 |
| Accurate people counting towards energy efficient buildings [40] | People counting, IoT, Lattice iCE40-HX1K stick FPGA boards and Raspberry Pi modules, direction of human movement | 97.0 |
| Neural Embedding Singular Value Decomposition for Collaborative Filtering [44] | Filtering, Singular value decomposition (SVD), recommender systems (RSs), neural network | – |
| Measuring Indoor Occupancy through Environmental Sensors [37] | $CO_2$ sensing as the main environmental parameter and RH as priority measures, data processing for occupancy detection - preferred Machine Learning algorithms such as SVM, RF and ANN | – |
| Our study | $CO_2$ sensing, occupancy estimation in intelligent building, big data analysis, support vector machine (SVM), wavelet transformations | 87.1 – 99.3 |

recognizing wavelet settings as the most suitable, neutral, and the least suitable for $CO_2$ signals. This recommendation system is capable of identifying the wavelet settings, which should minimize the error function for respective $CO_2$ signals against its gold standard.

The proposed approach has several advantages as summarized below:

- Use of operational sensors measuring the quality of the indoor environment ($CO_2$, Temperature, relative humidity) [37] to determine the prediction of the course of the $CO_2$ concentration as part of the occupancy detection of the monitored spaces in the Smart Home.
- Using an indirect method (without violating privacy, e.g. a camera) to determine the occupancy of monitored spaces [38,39] within the IoT [40,41].
- Using the SVM method to predict the course of the $CO_2$ concentration with great accuracy for the measured data in an interval of 1 day (accuracy was better than 98%) see Table 1 and Table 2 [40,37].

The results of the proposed method are comparable with the state-of-the-art approaches, as summarized in Table 3.

Future research should focus on the cost-effective sensor placement and exploration of fusion modules [37] that can reduce data redundancy while correlating subject measurements from multiple points [45]. The further approach for the $CO_2$ prediction in real time [43] utilizes other methods for own $CO_2$ prediction such as HMM [38], SVD [44] for HAR and ADL with the possibility of using other sensors, e.g. (motion detectors [PIR] and door entry sensors) [42] and using new methods [44] to additive noise canceling in the predicted course of the $CO_2$ signal. Moreover, as suggested in [46], the classification can be further improved by projecting the training instances into the low-dimensional singular subspace; the SVM can train the classification model on it while not violating the privacy requirements for the training data. The main advantage such approach is that singular value decomposition does not need to calculate the matrix of covariance, such as in the case of other methods, for example eigenvalue decomposition. The method also protects privacy of the training instances before training the classification model.

Nevertheless, we are aware of certain limitations in this approach that could be summarized as follows:

- Additive noise appeared in the prediction of the course of $CO_2$ concentration [44].
- For the measured data within a one-week and two-week interval (Table 1, Table 2), the accuracy of the used method of predicting the course of $CO_2$ decreased to values of around 80%.
- A lack of direct comparisons with other relevant algorithms. By not including these comparisons, we may not be able to fully assess the effectiveness of our chosen model in relation to alternative approaches. This could potentially limit the generalizability and applicability of our findings.

To eliminate the disadvantages of the proposed method, the Wavelet method was used to additive noise canceling in the predicted course of the $CO_2$ signal. It is needed to mention that the proposed system utilizes the mother's wavelet and the decomposition level as the wavelet-based features, which are evaluated for the prediction enhancement. Looking at the wavelet-based smoothing, there are further parameters that may influence the smoothing effect. Since this procedure utilizes the wavelet coefficient thresholding, we select the threshold selection rule and type of thresholding. In this study, we used the principle of Stein's Unbiased Risk as a thresholding rule and soft thresholding. In the future study, it would be worth studying the significance of wavelet response differences between soft and hard thresholding, and individual thresholding rules. Thus, the impact of these settings on prediction accuracy. Also, we should mention that we use two predictions from April and March for modeling the wavelet response, where we can observe the significantly different wavelet responses, especially for the different time periods. To better prove the robustness of the proposed method, we are going to perform testing on the extensive dataset to verify the achieved results from this study.

## 6. Conclusion

The presented work can be used for the area of technological support of independent housing of seniors, elderly and disabled persons in buildings where automation of the operational and technical functions (SH) is implemented with regard to the needs of the inhabitants with the possibility of monitoring daily life activities.

The article describes the design and verification of the indirect method of predicting the course of $CO_2$ concentration (ppm) using SVR with parameter optimization techniques for monitoring the presence of people in the Ostrava Faculty of Electrotechnics and Computer Science (FEI). The article further describes the method of Support Vector Regression (SVR) for predicting the course of $CO_2$ from the values measured by the relative temperature and humidity sensors. For estimating $CO_2$ concentration in the air in order to obtain information on the occupancy of individual rooms (arrival time, departure time, number of people). There are much better methods for predicting human occupation by measuring the level of $CO_2$ (ppm), such as neural networks [47], as well as methods for optimizing model parameters, both regression and classification.

As it has already been commented in the development of this research, due to the computation time of the Grid-Search method with a large data set, it is possible that the performance of the optimization of these parameters can be improved by another class of algorithms [48,49], such as genetic algorithms [50], Particle Swarm Optimization (PSO) [51], possible improvements in Grid-Search, algorithms based on sine-cosine, and multi-objective optimization [52]. In this way, it is possible to achieve optimal hyperparameters in reasonable computation time, thus trying to improve the prediction model to a level of reliability higher than that obtained in the development of this study.

## CRediT authorship contribution statement

**Jan Vanus, Jan Kubicek:** Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

**Dominik Vilimek:** Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

**Marek Penhaker:** Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

**Petr Bilik:** Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data included in article/supp.material/referenced in article.

## Acknowledgement

This work was supported by the Student Grant System of VSB Technical University of Ostrava, grant number SP2023/090 and SP2023/028.

## Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e16114.

## References

[1] Sreeranga Rajan, W.V. Ginkel, N. Sundaresan, Anant Bardhan, Y. Chen, A. Fuchs, A. Kapre, A. Lane, Rongxing Lu, Pratyusa Manadhata, J. Molina, A.C. Mora, P. Murthy, Arnab Roy, Shiju Sathyadevan, Nrupak Shah, Cloud Security Alliance Report on the Top Ten Challenges in Big Data Privacy and Security, https://doi.org/10.13140/RG.2.1.1744.1127, 2013.

[2] S. Spiegel, Optimization of in-house energy demand, in: F. Hopfgartner (Ed.), Smart Information Systems, Springer International Publishing, Cham, 2015, pp. 271–289.

[3] L. Brett, R. Love, L. Harvey, Big Data: Time for a lean approach in financial services, a Deloitte Analytics paper.

[4] Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are.

[5] B. Alotaibi, Utilizing blockchain to overcome cyber security concerns in the Internet of things: a review, IEEE Sens. J. 19 (23) (2019) 10953–10971, https://doi.org/10.1109/JSEN.2019.2935035.

[6] S.N. Mohanty, K. Ramya, S.S. Rani, D. Gupta, K. Shankar, S. Lakshmanaprabu, A. Khanna, An efficient lightweight integrated blockchain (ELIB) model for IoT security and privacy, Future Gener. Comput. Syst. 102 (2020) 1027–1037, https://doi.org/10.1016/j.future.2019.09.050.

[7] S. Pešić, M. Tošić, O. Iković, M. Radovanović, M. Ivanović, D. Bošković, BLEMAT: data analytics and machine learning for smart building occupancy detection and prediction, Int. J. Artif. Intell. Tools 28 (06) (2019) 1960005, https://doi.org/10.1142/S0218213019600054.

[8] K. Kim, A. Jalal, M. Mahmood, Vision-based human activity recognition system using depth silhouettes: a smart home system for monitoring the residents, J. Electr. Eng. Technol. 14 (6) (2019) 2567–2573, https://doi.org/10.1007/s42835-019-00278-8.

[9] E. Longo, A.E. Redondi, M. Cesana, Accurate occupancy estimation with WiFi and bluetooth/BLE packet capture, Comput. Netw. 163 (2019) 106876, https://doi.org/10.1016/j.comnet.2019.106876.

[10] F. Wang, Q. Feng, Z. Chen, Q. Zhao, Z. Cheng, J. Zou, Y. Zhang, J. Mai, Y. Li, H. Reeve, Predictive control of indoor environment using occupant number detected by video data and CO 2 concentration, Energy Build. 145 (2017) 155–162, https://doi.org/10.1016/j.enbuild.2017.04.014.

[11] D. Ioannidis, P. Tropios, S. Krinidis, D. Tzovaras, S. Likothanassis, Building multi-occupancy analysis and visualization through data intensive processing, in: L. Iliadis, I. Maglogiannis (Eds.), Artificial Intelligence Applications and Innovations, vol. 475, Springer International Publishing, Cham, 2016, pp. 587–599.

[12] K. Akkaya, I. Guvenc, R. Aygun, N. Pala, A. Kadri, IoT-based occupancy monitoring techniques for energy-efficient smart buildings, in: 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), IEEE, New Orleans, LA, USA, 2015, pp. 58–63.

[13] G. Demiris, Privacy and social implications of distinct sensing approaches to implementing smart homes for older adults, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Minneapolis, MN, 2009, pp. 4311–4314.

[14] Q. Ni, A. García Hernando, I. de la Cruz, The Elderly's independent living in smart homes: a characterization of activities and sensing infrastructure survey to facilitate services development, Sensors 15 (5) (2015) 11312–11362, https://doi.org/10.3390/s150511312.

[15] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO 2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39, https://doi.org/10.1016/j.enbuild.2015.11.071.

[16] B.W. Hobson, D. Lowcay, H.B. Gunay, A. Ashouri, G.R. Newsham, Opportunistic occupancy-count estimation using sensor fusion: a case study, Build. Environ. 159 (2019) 106154, https://doi.org/10.1016/j.buildenv.2019.05.032.

[17] K. Arendt, A. Johansen, B.N. Jørgensen, M.B. Kjærgaard, C.G. Mattera, F.C. Sangogboye, J.H. Schwee, C.T. Veje, Room-level occupant counts, airflow and CO $_2$ data from an office building, in: Proceedings of the First Workshop on Data Acquisition to Analysis, ACM, Shenzhen China, 2018, pp. 13–14.

[18] F. Lachhab, M. Bakhouya, R. Ouladsine, M. Essaaidi, Context-driven monitoring and control of buildings ventilation systems using big data and Internet of things–based technologies, proceedings of the institution of mechanical engineers, Part I, J. Syst. Control Eng. 233 (3) (2019) 276–288, https://doi.org/10.1177/0959651818791406.

[19] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, N. Villar, PreHeat: controlling home heating using occupancy prediction, in: Proceedings of the 13th International Conference on Ubiquitous Computing - UbiComp '11, ACM Press, Beijing, China, 2011, p. 281.

[20] W. Wang, X. Xu, H.-H. Wei, B. Ren, J. Chen, Modeling occupancy distribution in large building spaces for HVAC energy efficiency, Energy Proc. 152 (2018) 1230–1235, https://doi.org/10.1016/j.egypro.2018.09.174.

[21] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework, Energy Build. 88 (2015) 395–408, https://doi.org/10.1016/j.enbuild.2014.11.065.

[22] A. Crivello, F. Mavilia, P. Barsocchi, E. Ferro, F. Palumbo, Detecting occupancy and social interaction via energy and environmental monitoring, Int. J. Sens. Netw. 27 (1) (2018) 61, https://doi.org/10.1504/IJSNET.2018.092136.

[23] N. Noury, M. Berenguer, H. Teyssier, M.-J. Bouzid, M. Giordani, Building an index of activity of inhabitants from their activity on the residential electrical power line, IEEE Trans. Inf. Technol. Biomed. 15 (5) (2011) 758–766, https://doi.org/10.1109/TITB.2011.2138149.

[24] D. Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, San Francisco, Calif, 1999.

[25] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Intelligent Systems Reference Library, vol. 72, Springer International Publishing, Cham, 2015.

[26] S. García, J. Luengo, F. Herrera, Data Preparation Basic Models, vol. 72, Springer International Publishing, Cham, 2015, pp. 39–57.

[27] T. Eitrich, B. Lang, Efficient optimization of support vector machine learning parameters for unbalanced datasets, Int. J. Comput. Appl. Math. 196 (2) (2006) 425–436, https://doi.org/10.1016/j.cam.2005.09.009.

[28] B. Farahani, F. Firouzi, K. Chakrabarty, Healthcare IoT, in: F. Firouzi, K. Chakrabarty, S. Nassif (Eds.), Intelligent Internet of Things, Springer International Publishing, Cham, 2020, pp. 515–545.

[29] Y. Ho, D. Pepyne, Simple explanation of the no-free-lunch theorem and its implications, J. Optim. Theory Appl. 115 (3) (2002) 549–570, https://doi.org/10.1023/A:1021251113462.

[30] D. Wolpert, W. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (1997) 67–82, https://doi.org/10.1109/4235.585893.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg Scikit-learn, Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[32] J. Vanus, O.M. Gorjani, P. Bilik, Novel proposal for prediction of CO2 course and occupancy recognition in intelligent buildings within IoT, Energies 12 (23) (2019) 4541, https://doi.org/10.3390/en12234541.

[33] J. Vanus, R. Martinek, L. Danys, J. Nedoma, P. Bilik, Occupancy detection in smart home space using interoperable building automation technologies, Hum.-Cent. Comput. Inf. Sci. 12 (1) (2022) 616–632, https://doi.org/10.22967/HCIS.2022.12.047.

[34] J. Vanus, O. Majidzadeh Gorjani, P. Dvoracek, P. Bilik, J. Koziorek, Application of a new CO$_2$ prediction method within family house occupancy monitoring, IEEE Access 9 (2021) 158760–158772, https://doi.org/10.1109/ACCESS.2021.3130216.

[35] J. Vanus, K. Fiedorova, J. Kubicek, O.M. Gorjani, M. Augustynek, Wavelet-based filtration procedure for denoising the predicted CO2 waveforms in smart home within the Internet of things, Sensors 20 (3) (2020) 620, https://doi.org/10.3390/s20030620.

[36] J. Vanus, J. Kubicek, O.M. Gorjani, J. Koziorek, Using the IBM SPSS SW tool with wavelet transformation for CO2 prediction within IoT in smart home care, Sensors 19 (6) (2019) 1407, https://doi.org/10.3390/s19061407.

[37] A.R. Mena, H.G. Ceballos, J. Alvarado-Uribe, Measuring indoor occupancy through environmental sensors: a systematic review on sensor deployment, Sensors 22 (10) (2022) 3770, https://doi.org/10.3390/s22103770.

[38] S. Ali, N. Bouguila, Towards scalable deployment of hidden Markov models in occupancy estimation: a novel methodology applied to the study case of occupancy detection, Energy Build. 254 (2022) 111594, https://doi.org/10.1016/j.enbuild.2021.111594.

[39] J. Guo, M. Amayri, N. Bouguila, W. Fan, A hybrid of interactive learning and predictive modeling for occupancy estimation in smart buildings, IEEE Trans. Consum. Electron. 67 (4) (2021) 285–293, https://doi.org/10.1109/TCE.2021.3131943.

[40] Q. Huang, K. Rodriguez, N. Whetstone, S. Habel, Rapid Internet of things (IoT) prototype for accurate people counting towards energy efficient buildings, Electron. J. Inf. Tech. Constr. 24 (2019) 1–13, https://doi.org/10.36680/j.itcon.2019.001.

[41] K.P. Shirsat, G.P. Bhole, An empirical study on the occupancy detection techniques based on context-aware IoT system, in: S. Shakya, V.E. Balas, W. Haoxiang, Z. Baig (Eds.), Proceedings of International Conference on Sustainable Expert Systems, vol. 176, Springer Singapore, Singapore, 2021, pp. 95–105.

[42] A. Howedi, A. Lotfi, A. Pourabdollah, Employing entropy measures to identify visitors in multi-occupancy environments, J. Ambient Intell. Humaniz. Comput. 13 (2) (2022) 1093–1106, https://doi.org/10.1007/s12652-020-02824-z.

[43] Q. Huang, Occupancy-driven energy-efficient buildings using audio processing with background sound cancellation, Buildings 8 (6) (2018) 78, https://doi.org/10.3390/buildings8060078.

[44] T. Huang, R. Zhao, L. Bi, D. Zhang, C. Lu, Neural embedding singular value decomposition for collaborative filtering, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–9, https://doi.org/10.1109/TNNLS.2021.3070853.

[45] A. Das, R. Gupta, S. Chakraborty, A study on real-time edge computed occupancy estimation in an indoor environment, in: 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), IEEE, Bengaluru, India, 2020, pp. 527–530.

[46] Z. Sun, J. Yang, X. Li, Differentially private singular value decomposition for training support vector machines, Comput. Intell. Neurosci. 2022 (2022) 1–11, https://doi.org/10.1155/2022/2935975.

[47] J.P. Skön, M. Johansson, M. Raatikainen, K. Leiviskä, M. Kolehmainen, Modelling indoor air carbon dioxide (CO2) concentration using neural network, Methods 14 (15) (2012) 16.

[48] Qiujun Huang, Jingli Mao, Yong Liu, An improved grid search algorithm of SVR parameters optimization, in: 2012 IEEE 14th International Conference on Communication Technology, IEEE, Chengdu, China, 2012, pp. 1022–1026.

[49] S. Li, H. Fang, X. Liu, Parameter optimization of support vector regression based on sine cosine algorithm, Expert Syst. Appl. 91 (2018) 63–77, https://doi.org/10.1016/j.eswa.2017.08.038.

[50] D. Zhang, W. Liu, A. Wang, H. Jin, Parameter optimization for SVR based on genetic algorithm and simplex method, in: 2010 Chinese Conference on Pattern Recognition (CCPR), IEEE, Chongqing, China, 2010, pp. 1–6.

[51] Q. Pan Kongzhi-Lilun-Zhuanye-Weiyuanhui (Ed.), 2013 32nd Chinese Control Conference, (CCC 2013): Xi'an, China, 26 - 28 July 2013, IEEE, Piscataway, NJ, 2013.

[52] L. Pasolli, C. Notarnicola, L. Bruzzone, Multi-objective parameter optimization in support vector regression: general formulation and application to the retrieval of soil moisture from remote sensing data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (5) (2012) 1495–1508, https://doi.org/10.1109/JSTARS.2012.2197178.