Energy Informatics

RESEARCH Open Access

Climatization and luminosity optimization of buildings using genetic algorithm, random forest, and regression models

Bruno Mota[1,2], Miguel Albergaria[1], Helder Pereira[1,2], José Silva[1], Luis Gomes[1,2], Zita Vale[1*] and Carlos Ramos[1,2]

From 1st Energy Informatics.Academy Conference Asia
Beijing, China . 29-30 May 2021

* Correspondence: zav@isep.ipp.pt
[1]Institute of Engineering - Polytechnic of Porto, Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto, Portugal
Full list of author information is available at the end of the article

Abstract

With the rise in popularity of artificial intelligence, coupled with the growing concern over the environment, there has been a surge in the use of intelligent energy management systems. Additionally, as more buildings transition into the smart grid and, consequently, more energy and environmental data is gathered, there has been a significant increase in the number of data-driven approaches for building management systems. This paper proposes a methodology that aims to optimize the climatization and luminosity inside a building, using a genetic algorithm, a random forest, and two polynomial models. The proposed methodology enables the real-time management of the building taking into account the user needs and preferences. Air conditioner units and light systems are optimized to minimize energy costs, while also improving the air quality and considering the users' temperature and luminosity preferences. This paper shows the results achieved, by the proposed solution, in an office building case study. The promising results demonstrate the possibility of minimizing energy costs while maximizing the users' comfort.

Keywords: Building energy management systems, Genetic algorithm, Polynomial regression, Random Forest, Air quality

Introduction

Globally, the industry and household sectors, combined, account for slightly over 50% of the total final energy consumption (Energy statistics, n.d.). And, with much of the energy coming from the burning of fossil fuels (IPCC 2014, 2014), leading to greenhouse gas emissions, the transition to smarter energy management systems is becoming increasingly more urgent (Vale et al., 2010). However, enhancing the energy efficiency in buildings, both, industrial and residential, requires in-depth knowledge of the underlying performance. Thus, the gathering of energy and environmental conditions data, coupled with the use of smart management, has become an essential step to achieve significant energy consumption reductions (Faria et al., 2015; Abrishambaf et al., n.d.).

A Genetic Algorithm (GA) is a metaheuristic search-based optimization algorithm inspired by natural selection, that belongs to the class of evolutionary algorithms (Genetic Algorithm, n.d.). Having been used for the optimization of energy consumption as far back as 1997 (Huang & Lam, 1997), it continues to be a highly used approach to the problem of energy consumption reduction, as seen in (Mota et al., 2021), where it is proposed a model for the management of loads in an industrial production line using a GA, and (Nguyen & Nassif, 2016), where a model-based optimization process for Heating, Ventilating and Air Conditioning (HVAC) systems using a GA is presented.

Random Forest Classifiers or Random Decision Forest Classifiers are an ensemble learning method for classification (Yiu, 2019), meaning that they resort to multiple machine learning models, in the Random Forest case, Decision Trees, in order to obtain better predictive performance than a sole model could obtain (Ramzai, 2019). One drawback associated with Decision Tree Classifiers is their high variance, due to the significant change a tree can suffer from a small variance in the training data, therefore, the Random Forest methodology was invented, so that tree classification could be more stable (Azar et al., 2014). The use of this technique for energy consumption reduction is also not new, as seen in (Ahmad et al., 2017), where it is used for the prediction of a hotel's HVAC energy consumption, and (Chen et al., 2019), where it is used for energy load consumption forecasting of a large hypermarket.

Regression analysis is a prediction technique that entrusts the relationship between variables, in order to obtain the best-fit regression equation, that can be used to make predictions (Pant, 2019). Unlike, Random Forest Classifiers, which are used for class predictions, these models aim for the prediction of continuous numerical values, such as of monthly heating demand for residential buildings, as seen in (Catalina et al., 2008).

The premise of this paper is to present a data-driven approach to the optimization problem that is the climatization and luminosity management of a building, using a Genetic Algorithm, a Random Forest model, and two Polynomial Regression models and taking into consideration the temperature, luminosity, air quality, energy cost, and occupant's comfort. In an initial phase, through a Genetic Algorithm, a dataset is generated that aims to minimize energy cost and maximize user comfort and health, by optimizing the actions to take on each equipment in the room (i.e., air conditioner, artificial lighting, motorized blinds, windows, and doors). Afterward, the generated dataset is used to train a Random Forest and two Polynomial Regression models. As a result, the main goal of the Random Forest and the two Polynomial Regression models is to replicate the Genetic Algorithm optimization, in real-time. The Random Forest controls the equipment status, by turning on/off the air conditioner and artificial lighting, and opening/closing the motorized blinds, door(s), and window(s). One of the two Polynomial Regressions is used to determine the specified temperature for the air conditioner, while the other is used to set the artificial lighting luminosity.

This article is divided into six main sections. The first contextualizes the objective of the paper and presents the adopted structure. The second section presents several state of art related projects, whose theme is the use of the mentioned machine learning techniques for building energy management systems. The third section presents the proposed methodology approach for the system. The fourth section presents the case study of the proposed methodology, along with the obtained results. The fifth section presents the conclusion of this paper, coupled with possible future development paths, and the sixth, and final section, presents the availability of data and materials, competing interests, funding, authors' contributions, and acknowledgments.

Related works

The use of artificial intelligence methods for building energy management systems is nothing new, as seen by the large number of studies done. However, improving this type of system is not one solution only type of problem. Therefore, throughout all the research and applications currently available, the computational techniques used, ranging from control and diagnosis to prediction, and optimization, with each category having its own set of recommended implementation algorithms. For prediction, the most used technique is Artificial Neural Networks (ANNs) (Ahmad et al., 2017), for optimization, Genetic Algorithms (Nguyen & Nassif, 2016), and for control and diagnosis, Fuzzy Logic (Ali & Kim, 2015) and Expert Systems (Faia et al., 2017; Ahmad et al., 2016). While on one hand, Genetic Algorithms are already widely used in the building energy domain, on the other hand, the use of Decision Trees (DTs) and Regression models are still very limited.

In 2011, Fernandes et al. proposed a Genetic Algorithm methodology to manage the consumption of typical house loads, while considering consumers' preferences for each load and context (Fernandes et al., 2011). The proposed approach aims to manage the energy consumption by cutting or reducing certain loads whenever the consumption is higher than the setpoint, and taking into consideration the load's weights, which are based on the preferences for the current day and time. Furthermore, this paper also presents a Mixed Integer Nonlinear Programming approach to the problem, with both solutions able to achieve positive results with, and without, consumers' actions (Fernandes et al., 2011). More recently, in 2013, Nguyen and Nassif proposed another model-based optimization process for HVAC systems using a Genetic Algorithm, relying on real measured data, and aiming for a decrease in energy costs while maintaining or improving indoor environmental conditions (Nguyen & Nassif, 2016). The presented evolutionary algorithm works by taking in the supply air temperature, duct static pressure, and outdoor airflow, and then trying to reduce the fitness, which is the energy consumption value, using the previous 15 min data interval. Additionally, the used algorithm was improved by the use of constraints, stochastic universal sampling (SUS), and elite-preserving operator, having achieved a 26% total energy consumption reduction when compared to traditional operating strategies (Nguyen & Nassif, 2016).

Ali and Kim, in 2015, also proposed an approach to building energy management systems using a Genetic Algorithm but taking into consideration, both, the energy consumption and the occupant's comfort index (Ali & Kim, 2015). For this, the proposed technique is able to integrate the users' comfort index, which is calculated using the thermal comfort (temperature), visual comfort (illumination) and air quality (CO2 concentration), and the corresponding energy usage into the fitness function, therefore, targeting the satisfaction of the occupants' requirements while using minimal energy consumption. In more detail, this system works by optimizing the comfort parameters, through the Genetic Algorithm, and then inputting the difference between the optimal parameters and the real environmental parameters, which are obtained through sensors, into Fuzzy controllers (one for each of the 3 parameters). Additionally, the proposed system also implements the Kalman Filter in order to predict the energy consumption, while taking in, as input, the actual consume energy. With the proposed GA based model, the authors were able to achieve an improvement in the occupants' comfort index and reduction of the energy consumption when compared to Particle Swarm Optimization (PSO) based systems, while also enabling its integration with SCADA software of buildings for real-world applications (Ali & Kim, 2015).

As seen, the improvement of building energy management systems can be accomplished through optimization. However, prediction approaches are also valid. In 2016, Ahmad, Mourshed, and Rezgui used Random Forest models to predict a Madrid hotel HVAC energy consumption (Ahmad et al., 2017). The dataset used consisted in 5 min historical values of HVAC electricity consumption for the studied building, total daily number of guests and rooms booked, 30 min outdoor weather conditions, which included air temperature, dew point temperature, wind speed, and relative humidity, and the time, consisting in the hour, day, and month. In total, after removing outliers and missing values, there were 10,972 data samples. Having the necessary data, the authors, then, trained the Random Forest models according to three studies they performed, which consisted in obtaining the optimal depth of the tree, optimal number of features, and importance of features. The first study consisted of training several models, using all features, but different depths, concluding an optimal depth value of 10. The second study consisted of evaluating several Random Forest models, all with a depth of 10 and each with a different number of features, which were randomly selected, obtaining 5 as the ideal number of parameters. The third study, consisted of replacing, in turn, each input variable for random noise and analyzing the deterioration of the performance of the model, allowing the measurement of the importance of each variable and concluding that the previous hour's electricity consumption is by far the most important feature. With these results, the authors were then able to evaluate two Random Forest models of depth 10, one created using the model with all features and the other with just the most important ones, achieving better performance using all features. Additionally, a Feed-Forward Back-Propagation Artificial Neural Network was also developed using the available data, resulting in slightly better results than both Decision Trees. Nonetheless, it was concluded that all models can be feasible and effective for predicting hourly HVAC electricity consumption (Ahmad et al., 2017).

Chen, Piedad Jr., and Kuo, in 2019, also used a Random Forest for energy consumption load forecasting (Chen et al., 2019). In this approach, the authors propose the use of a level-based methodology, contrary to the conventional value-based methodology approach, that works by training the model with the pre-processed dataset values and then converting the results into consumer-preferred levels (e.g., low, average, high). The proposed approach consists of converting the dataset values into levels during the pre-processing phase, instead of the post-processing phase, allowing the direct prediction of the desired levels using simpler classifier models without undergoing regression. Using a 12-month dataset of a large hypermarket, consisting in hourly energy consumption and temperature, as well, as 10-time cross-validation, it was concluded that the proposed approach performs better for all tests done, which were of 3, 5, and 7 levels, then the conventional way, however, the performance of the conventional classifier was also concluded to be able to approach the proposed method in terms of classification accuracy at the expense of computation time (Chen et al., 2019).

In 2008, Catalina, Virgone, and Blanco proposed the use of Regression models for the prediction of monthly heating demand for residential buildings (Catalina et al., 2008). For the models, the inputs consisted in the building shape factor, the building envelope U-value, which is the thermal transmittance in the envelope of a building (Franco, 2018), the window to floor area ratio, the building time constant, and the climate, which is defined by the difference between the heating set-point temperature and the monthly average sol-air temperature of the considered city. With the inputs defined, the authors resorted to simulations to obtain the necessary data to train the models, having found that quadratic (second order) polynomial models are the most appropriate solution. Additionally, the authors validated the obtained models through 270 different scenarios, having achieved an overall positive outcome, with a maximum deviation of 5.1% and an average error of 2%. It was, also, concluded that the shape of the building and the energy consumption has a good correlation, that the thermal inertia has a significant impact on the energy demand, and that the proposed approach to summarize the climate is efficient (Catalina et al., 2008).

This paper aims to fill the gaps of the above-cited works, by considering both user comfort and health with cost-effective optimization, by taking into account the usage of locally generated energy, volatile energy market prices, equipment configuration (e.g., availability and power consumption), in a real-time application. It proposes a methodology capable of retaining the Genetic Algorithm's robust searching solution, to minimize energy cost and maximize user comfort and health, while also delivering fast (i.e., real-time) optimizations, through the usage of a Random Forest and two Polynomial Regression models, that replicate the Genetic Algorithm optimization results. Additionally, it proposes the inclusion of windows and doors as factors for the optimization, that is, to minimize energy cost and maximize user comfort and health, by recommending the user for the opening or closing of windows and doors.

Proposed methodology

The proposed solution aims to minimize energy cost and maximize user comfort and health in a given room in smart buildings and smart homes, through a real-time application of sensor data acquisition and optimization of equipment configuration. It achieves these results by adjusting the air conditioner temperature, artificial lighting luminosity, motorized blinds, and by advising the opening of doors and/or windows.

The solution proposed in this paper uses the combination of artificial intelligence techniques of Random Forest, Polynomial Regression, and Genetic Algorithm, to achieve a system capable of delivering fast solutions with high precision and accuracy. The Random Forest and Polynomial Regressions are used for real-time application, while the Genetic Algorithm, being slower, is used for offline training of the previous models. Figure 1 shows the proposed solution structure.

In an initial phase, described by Fig. 1 a), before the system is applied for real-time application on smart buildings and smart homes, an initial dataset, containing different scenarios, is provided to a Genetic Algorithm. Then, the Genetic Algorithm is used to minimize energy cost and maximize user comfort and health by optimizing the actions to take on the equipment, that is, the air conditioner temperature, artificial lighting

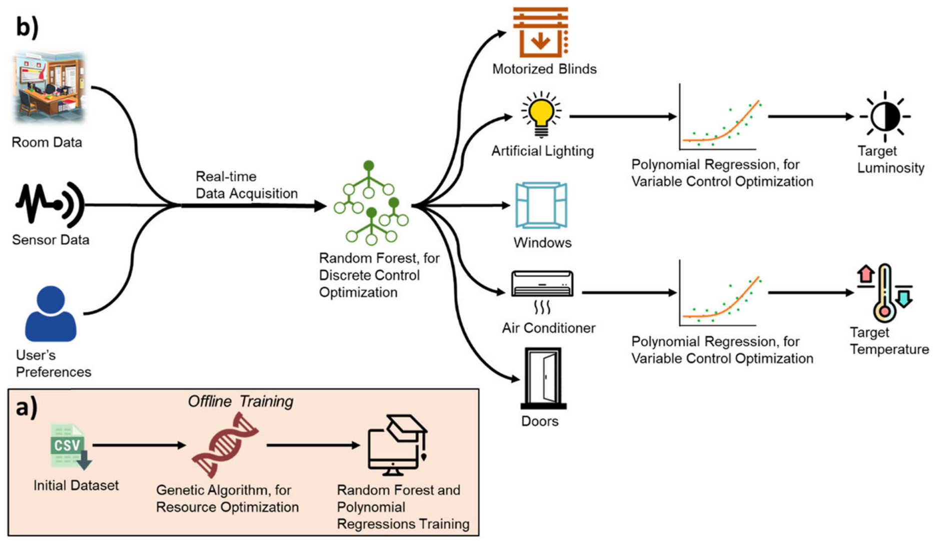
b)
Room Data
Sensor Data
User's Preferences
Real-time Data Acquisition
Random Forest, for Discrete Control Optimization
Motorized Blinds
Artificial Lighting
Windows
Air Conditioner
Doors
Polynomial Regression, for Variable Control Optimization
Target Luminosity
Polynomial Regression, for Variable Control Optimization
Target Temperature
a)
Offline Training
CSV
Initial Dataset
Genetic Algorithm, for Resource Optimization
Random Forest and Polynomial Regressions Training

Fig. 1 Flowchart of the proposed solution: a) Initial training, b) Real-time execution

luminosity, opening or closing of the remote blinds, and by suggesting/advising the opening of doors and/or windows. From the Genetic Algorithm, a dataset is generated, containing the optimized actions to take on each equipment for each scenario, which is then used to train a Random Forest and two Polynomial Regressions models, in order for them to replicate the Genetic Algorithm's optimizations. Therefore, the genetic algorithm is used exclusively for complex resource optimization, to train the Random Forest and two Polynomial Regressions, in an offline environment. The Random Forest, also described in this paper as a real-time discrete control optimization, is used to control only the equipment status (i.e., turn on/off and open/close) for the air conditioner, artificial lighting, motorized blinds, door(s), and window(s). From the two Polynomial Regressions, described in the paper as a real-time variable control optimization, one of them is used to determine the optimized temperature for the air conditioner, while the other is used for the optimized artificial lighting luminosity value. All of these techniques are developed in Python.

The application in real-time of the proposed solution, described by Fig. 1 b), begins 5 min before each passing hour (e.g., 15:55, 16:55, 17:55, …), with the acquisition of all the data needed to predict the optimized actions to be taken in a room. Such data is characterized by room data (e.g., air conditioner power, artificial lighting power, and the existence of motorized blinds), user's preferences (e.g., air conditioner target temperature, and artificial lighting target luminosity), and sensor data (e.g., outside temperature, the room's inside temperature, outside luminosity, the room's air quality, available PV, …). Afterward, a discrete control optimization is done to equipment status for the air conditioner, artificial lighting, motorized blinds, door(s), and window(s). Then, a variable control optimization is used to predict the temperature of the air conditioner and artificial lighting luminosity. Finally, the actions are sent to an equipment management system capable of executing said actions in the room.

The proposed solution, applied to each room of the building, focuses on the minimization of four equations and the balance between them: energy cost, temperature deviation, luminosity deviation, and Volatile Organic Compound.

The energy cost, to be minimized, considers the consumption of air conditioner units and artificial lighting. To further minimize energy costs, the solution considers locally renewable-based generation. The energy cost of a room is represented by:

$$Energy_{cost} = [E_{Gen} - (P_{AC} \cdot |T_{AC} - T_{Inside}| \cdot t + P_{AL} \cdot L_{AL} \cdot t)] * E_{Price} \tag{1}$$

where E_{Gen} represents the available generated energy, E_{Price} the energy price per unit of power, and t the time, in hours, the equipment is switched on (i.e., the time interval of the optimization). The power per air conditioner temperature degree difference from the room's temperature is represented by P_{AC}, T_{AC} portrays the air conditioner target temperature, and T_{Inside} the room's current inside temperature. The power per artificial lighting luminosity is portrayed by P_{AL}, and L_{AL} is the luminosity from the artificial lighting.

The difference between the intended temperature and the room's temperature, also known as the intended temperature deviation, represents the user's comfort regarding temperature. The closer to zero, the better the room is appropriated to the user, being zero the perfect scenario. The following equation portrays this metric:

$$Temperature\ Deviation = |[T_{AC} \cdot W_{AC} + T_{Blinds} \cdot W_{Blinds} + T_{Door} \cdot W_{Door} + T_{Window} \cdot W_{Window} + (1 - W_{AC} - W_{Blinds} - W_{Door} - W_{Window}) \cdot T_{Inside}] - T_{Intended}| \tag{2}$$

where T_{AC}, T_{Blinds}, T_{Door}, and T_{Window} represent, respectively, the air conditioner temperature, temperature by opening the blinds, door, and windows. The same applies for W_{AC}, W_{Blinds}, W_{Door}, and W_{Window}, which portrays the temperature weight in changing the overall temperature, for the air conditioner, by opening the blinds, door, and windows, respectively. The variable T_{Inside} represents the room's temperature, and $T_{Intended}$ the intended room temperature.

The intended lighting luminosity deviation follows the same concept as the intended temperature deviation. It is a calculation of the difference between the intended lighting luminosity and the room's luminosity. In case the difference is zero, the maximum comfort for lighting luminosity is reached, if not, the closer to zero, the better. The next equation represents this calculus:

$$Luminosity\ Deviation = |[L_{AL} + (L_{Blinds} \cdot S_{Exterior}) - L_{Intended}]| \tag{3}$$

where L_{AL}, L_{Blinds}, and $L_{Intended}$ portray the artificial lighting luminosity, blinds luminosity, and intended luminosity, respectively. The variable $S_{Exterior}$ represents the exterior sensor, which takes a value of one during the day and zero at night.

The Volatile Organic Compound (VOC) in a room is correlated with a room's air quality. Lower VOC values represent higher air quality, thus decreasing the chance of long-term health problems for users. The proposed solution considers that when an air conditioner is turned on, the circulation mode is always active, which reduces the VOC. Also, opening doors and windows increase the circulation of air, and consequently, it increases air quality. The equation to estimate the VOC is the following:

$$VOC \cong VOC_{Inside} \cdot (1 - W_{AC} - W_{Door} - W_{Window}) \tag{4}$$

where VOC_{Inside} portrays the room's VOC, obtained through a sensor. The variables W_{AC}, W_{Door}, and W_{Window}, correspond to the weight the air conditioner switched on, opening a door, and opening the windows, respectively, have in reducing the VOC.

Complex resource optimization using a genetic algorithm

A Genetic Algorithm is proposed for complex resource optimization, that aims to minimize energy costs and maximize user comfort, by adjusting the air conditioner temperature, artificial lighting luminosity, remote blinds, and by advising the opening of doors and/or windows. The proposed Genetic Algorithm not only allows energy cost and user comfort optimizations, but it also takes into account the user's health by achieving low values of VOC. Figure 2**.** Flowchart of the proposed complex resource optimizer.represents the flowchart of the proposed complex resource optimizer.

The Genetic Algorithm begins by creating an initial random population where each individual is characterized by: air conditioner temperature, artificial lighting luminosity, blinds state, door state, and window state.

The crossover performed is of the uniform type, where each optimization parameter is chosen from either parent 1 (individual 1) or parent 2 (individual 2), with equal probability, to be inherited to the child (offspring). The mutation procedure is done using a random value to determine whether or not the mutation will be applied to an individual. If the mutation is to be applied, then it is based on defining a new value for one of the optimization parameters. The selected value is randomly chosen, within the defined limits of the parameter (e.g., the parameter of door status, can only be either "Open" or "Close"), and is different from the old value (i.e., value before the mutation).

The selection phase begins with the union of the crossed and mutated population with the initial population of the previous generation. Also, repetitions of individuals are eliminated. Then, each individual is evaluated through (1), (2), (3), and (4), which correspond to the calculus of the energy cost, intended temperature deviation, intended lighting luminosity deviation, and VOC, respectively. Afterward, the values are normalized, using a Min-Max approach, and each individual is evaluated using the following fitness equation:

$$Maximize\ Fitness \cong \frac{1}{EcNorm} \cdot ecW + \frac{1}{TdNorm} \cdot rdW + \frac{1}{LdNorm} \cdot ldW + \frac{1}{VocNorm} \cdot vocW \qquad (5)$$

where $EcNorm$, $TdNorm$, $LdNorm$, and $VocNorm$, represent the normalized energy cost, normalized intended temperature deviation, normalized intended lighting luminosity deviation, and normalized VOC, respectively. Also, ecW, rdW, ldW, and $vocW$, correspond to the weights of the energy cost, intended temperature deviation, intended

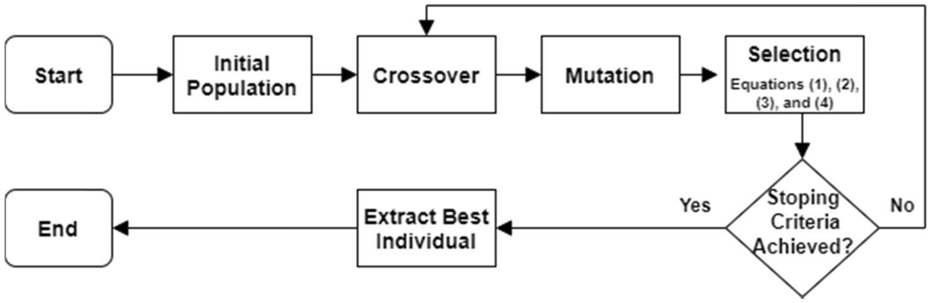

Fig. 2 Flowchart of the proposed complex resource optimizer

lighting luminosity deviation, and VOC, respectively. In the fitness equation, the inversion is applied for each evaluation factor (i.e., *EcNorm*, *TdNorm*, *LdNorm*, and *VocNorm*), since the lower the value of these factors, the better the solution.

Afterward, the selection of the n best individuals is done, based on the input data. The remaining individuals participate in non-elite tournaments. Each tournament consists of two randomly selected individuals, competing on the basis of their fitness ratings, in a probability approach. The algorithm calculates the probability of individual 1 winning the tournament, through the following equation:

$$Individual^{1}_{probability} = \frac{fit1}{fit1 + fit2} \tag{6}$$

where $fit1$ and $fit2$ represent the fitness of individual 1 and individual 2, respectively.

Then, a random decimal number between 0 and 1 is generated. If the generated decimal is lower than the probability of individual 1 winning, (6), then individual 1 is declared the winner. Otherwise, individual 2 leaves victorious. Therefore, the individual with the highest fitness is the one most likely to be chosen.

Discrete control optimization using random Forest

For the discrete control optimization, a Random Forest Classifier is proposed with the objective of learning how to predict the actions to take, in each possible scenario of an office or room, considering temperature, lighting, and air quality.

The first step is to split the data into two groups, training and testing, using the Holdout Method (Sharada, 2020). In the proposed solution a percentual value of 80 was used for training and 20 for testing. A good practice when using Holdout is to shuffle the data before splitting, to avoid dependencies between all testing scenarios, as being all of the same room.

Next, was performed a tuning to the hyper-parameters of the Random Forest. The technique used was RandomizedSearchCV (RandomizedSearchCV, n.d.) that randomly chooses one of the possible values for each one of the hyper-parameters and scores the estimator. The best estimator will be used in the model. Table 1 presents the hyper-parameters, their possible values, and their optimal values, result of the execution of the algorithm.

The used classifier is the RandomForestClassifier from Scikit-Learn. In the training process, the classifier will create the rules to achieve the minimum error relative to the training classes. After the model is correctly fitted with training data, it is prepared to do its predictions.

Variable control optimization using polynomial regressions

A variable control optimization is proposed using Polynomial Regressions, which aims to predict the air conditioner temperature, and artificial lighting luminosity, for each scenario that uses (i.e., turns on) the air conditioner and/or artificial lighting. The prediction is made for a one-hour execution time of the equipment (i.e., air conditioner and artificial lighting), however, it can be adjusted to other time intervals. Also, it predicts 5 min before each hour, using the average sensor data from the hour. The 5-min delay can be lowered by the user at the cost of compromising time consistency (i.e., the prediction could come after the hour has passed), but increasing accuracy (i.e., more

Table 1 Hyper-parameters possible and optimal values

Hyper-parameter	Possible Values	Optimal Value
Number of Trees	200 to 2000	400
Max Features	Auto, Sqrt or Log2	Sqrt
Criterion	Gini or Entropy	Gini
Max Depth	10 to 32	26
Min Samples to Split	2, 5 or 10	10
Min Samples in a Leaf	1,2 or 4	1
Bootstrap	True or False	False

recent data). Two Polynomial Regression models were created, using the Holdout Method, with 80% of the data set apart for training, and the rest for testing. All the data for training and testing was normalized, beforehand, using a Min-Max approach. Both models were trained using a technique that explores exhaustively the training data, within specified parameters, for the best polynomial equation; available through the Scikit-Learn library using the function GridSearchCV (GridSearchCV, n.d.). The technique was used with a 10 fold cross-validation, searching through polynomial equations of degree 1 to 10, and the metric Root-Mean-Square Error (RMSE) was used to evaluate the performance of the cross-validated model. The dependent variables chosen for each model were based on correlation maps, tests, and knowledge of the Genetic Algorithm equations (i.e., (1), (2), (3), and (4)). Also, the performance of each model was evaluated using the metrics: Mean Absolute Error (MAE), Root-Mean-Square Error (RMSE), Coefficient of Determination (R^2), and Adjusted Coefficient of Determination (Adjusted R^2).

The best air conditioner temperature prediction model obtained is of degree 6, with the following dependent variables:

- temperature sensor – inside room temperature.
- weighted doors temperature – temperature when opening room’s door(s), multiplied by its weight in affecting the overall temperature of the room.
- weighted windows temperature – temperature when opening room’s window(s), multiplied by its weight in affecting the overall temperature of the room.
- intended temperature – mean intended temperature from all the users in the room.
- air quality sensor – room’s VOC.
- available energy – available locally generated energy, in Wh.

The dependent variable available energy was chosen since it is the variable that most influences the final energy cost, (1). Therefore, this variable, greatly affects the air conditioner temperature, because the higher the difference in room temperature and air conditioner temperature, the higher the energy consumption, and so the energy costs. The energy market cost was not considered, in this case, because it showed to have a low correlation when predicting the air conditioner temperature. The variables temperature sensor, weighted doors temperature, weighted windows temperature, and intended temperature are used, due to being the main variables that affect the difference in room temperature and intended temperature, (2). Also, the variable air quality sensor was used since it has a high correlation with predicting the air conditioner temperature, and results showed that the inclusion of said variable, greatly decreased the error.

Other variables were tested, such as the difference between inside and hall temperatures, the difference between inside and intended temperatures, and the difference of current and previous intended temperatures. However, these variables did not yield good results, despite having a high correlation to the air conditioner temperature.

The best artificial lighting luminosity prediction model obtained is of degree 5, with the following dependent variables:

- light sensor – Inside room luminosity.
- blinds light – luminosity given by opening the blinds, at night the luminosity is always zero.
- blinds status – specifies if the blinds are open.
- intended luminosity – mean intended luminosity from all the users in the room.
- energy market cost – energy cost in EUR/Wh.
- available energy – Available locally generated energy, in Wh.

The dependent variable available energy was chosen for the same reason as in the air conditioner model because it affects greatly the energy cost, (1). Also, the more brightness (i.e., higher luminosity) the higher energy consumption, and so the energy costs. In this case, the energy market cost was included since it has a high correlation with predicting the artificial lighting luminosity, and prediction results improved with its addition. The variables light sensor, blinds light, blinds status, and intended temperature are used because they influence the difference in room luminosity and intended luminosity, (3).

Case study

The initial dataset, which, through the complex resource optimization, was processed for the training of the Random Forest and Regression models consists of real data from the GECAD' building, located in Porto, Portugal. It was considered five rooms from the building, all of them with temperature and lighting sensors, but just two of them with air quality sensors. The data that was read from the sensors has an interval of 10 s, which were then converted to hour averages, to get considerable changes between periods. Besides that, it contains the rooms' users' preferences, considering the temperature and lighting to each period of the day. Besides the rooms' sensors, it was also used exterior sensors, which provide information about exterior temperature and lighting.

As such, the complete dataset has information about the temperature in the room, hall and outside, luminosity levels indoor and outdoor, room's VOC level, air conditioner, and artificial lighting power, available renewable energy, and if a room has motorized blinds.

Additionally, it is used the energy market cost from *Mercado Ibérico de Eletricidade* (MIBEL), and the intended temperature and luminosity, which were randomly created according to valid intervals defined by the studied building users. These values were defined by room and time of the day. In total, the dataset has data relative to 4 rooms, with each having 8760 hourly data samples, totaling 35,040 data samples, which span a 1-year interval.

Offline complex resource optimization

The case study for the resource optimization uses real sensor data collected from the 1st of May 2019 to the 30th of April 2020. To this data set, it was added the energy market prices from MIBEL. Also, all the data is provided in intervals of one hour. For this case study, the Genetic Algorithm was executed for 30 s for each scenario. A scenario represents all the data from a room needed to predict the air conditioner temperature, artificial lighting luminosity, blinds state, door state, and window state, in a given timestamp (i.e., each row of the dataset). Furthermore, each optimization component (i.e., energy cost, intended temperature deviation, intended lighting luminosity deviation, and VOC) was balanced with a weight of 25%, thus having equal importance during the optimization. The algorithm was executed on a computer with an Intel® Core™ i3-7020U processor running at 2.30 GHz using 8 GB of RAM, running Windows 10 Home version 2004.

Figure 3 shows the average monthly energy costs of the optimization, considering a dynamic energy price and locally generated PV. The local energy generation, provided by photovoltaic panels, was used as much as possible by the algorithm, reducing the necessity to resort to energy retailers, thus reducing energy costs. For example, during October of 2019, even though there was a shortage of solar energy (i.e., available PV), and high energy prices, the algorithm was capable of maintaining low energy costs, by partially compromising the user's comfort. The maximum monthly average total energy cost obtained, in the case study, is 0.03245 EUR, and a minimum of 0.00579 EUR.

Figure 4 represents a comparison between the four main equations that the Genetic Algorithm tries to balance: energy cost, intended temperature deviation, intended lighting luminosity deviation, and VOC. The components that most fluctuate are the intended luminosity deviation and the total energy costs.

The balance for user comfort can be seen clearly in Fig. 4 with the average monthly temperature and luminosity deviations. The algorithm always tries to maximize both comfort components when possible, this can be observed by comparing Fig. 4 with Fig. 3, since when more PV is available both deviations are closer to one another, and the opposite can be seen as well. The monthly average temperature deviation, in the case study, had a maximum value of 5.2 °C deviation and a minimum of 1.8 °C. The

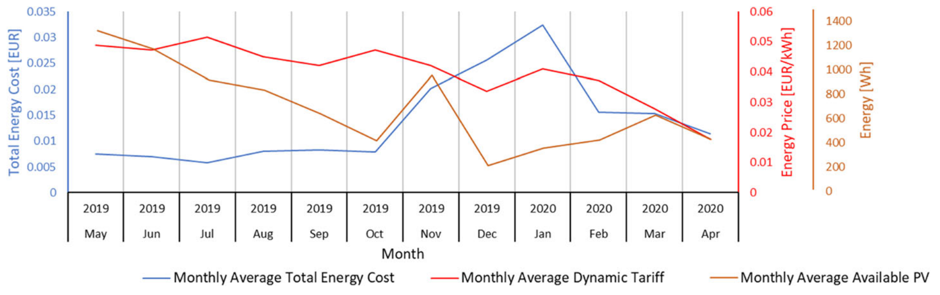

Fig. 3 Energy costs considering retailer prices, and available PV

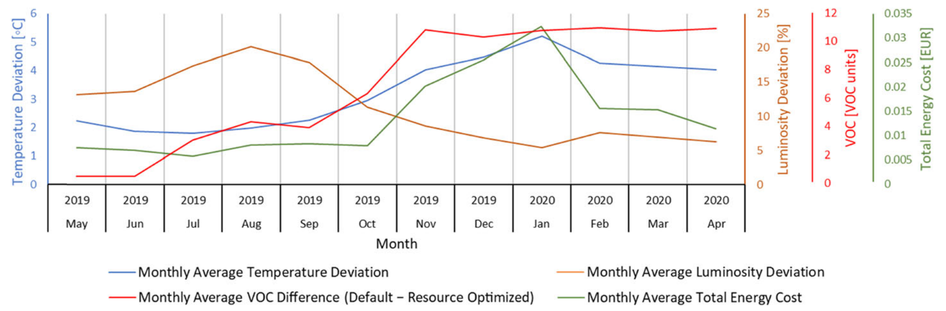

Fig. 4 Comparison between the four main equations: temperature deviation, luminosity deviation, VOC difference, and total energy cost

maximum monthly average luminosity deviation, for the case study, is 20.2%, and a minimum of 5.4%.

The optimization done to the VOC (i.e. air quality improvement done by the algorithm) is portrayed in Fig. 4 by the VOC difference, which represents the difference between the default VOC (i.e. room's VOC without any action taken, such as turning on the air conditioner, opening doors and windows) and the VOC obtained through the resource optimization. One of the crucial points the solution proposed in this paper focuses on is the user's health, and through the difference in VOC, it is clear that the algorithm delivers those results. The difference in VOC has a maximum monthly average value of 10.923 units of VOC, and a minimum of 0.544 units of VOC, in the case study for the proposed resource optimization.

Real-time discrete control optimization

In this case study, for the multiple output classification problem, three models were trained using the dataset generated by the offline complex resource optimization's Genetic Algorithm.

Regarding data preprocessing, besides what is proposed in the methodology section, it is also needed to encode one of the categorical values. The variable 'OpenBlinds' can take three different values: 'True', when the blinds are automatically opened, 'False', when the blinds are automatically closed, or 'None', when the blinds in the room are not motorized. Considering these possible values, the variable is binary encoded, and, consequently, split into two variables: 'OpenBlinds_1' and 'OpenBlinds_2'. Table 2 shows the original value and the corresponding encoded values of the new variables.

The first model trained was a Random Forest, which is an ensemble method, more specifically a bagging technique. The second model trained was a Decision Tree, due to being the Random Forest base estimator, and the third was a Gradient Boosting Classifier, which is an example of a boosting technique that also uses Decision Trees as a base estimator. Table 3 presents the precision values of the three trained models.

As presented in Table 3, all models were able to predict all targets with high precision, with the Random Forest achieving the lowest precision of 92% on the target 'OpenDoors' and overall precision of 70.28%. Using the Decision tree model as the predictive model it was achieved an overall precision of 63.58%, with the lowest precision of 91% also on the target 'OpenDoors'. And for the Gradient Boosting model, a final precision of 73.40% was achieved, with the lowest precision being 94% on the targets 'AirConditionerTemp', 'OpenDoors' and 'OpenBlinds_1'. We can observe that the Random Forest model outperformed its base estimator, with an overall improvement of 6.7%. This notable decrease in performance indicates that the Random Forest model, as expected, is better to handle this large quantity of data. The Random Forest creates a set of unpruned trees that are very diverse from each other, handling overfitting much better than a single Decision Tree. Comparing with the other models, Gradient Boosting outperformed both of them, with an overall precision improvement of 3.12%. Gradient Boosting also improved or matched the precision of every target and predicted 'ArtificialLightingLight' almost perfectly.

To better understand what each model did differently, Table 4 shows the three most important features of each model, coupled with their feature importance values.

From Table 4, it is observable that the Random Forest algorithm is focusing mainly on the remotely controllable blinds flag and the air quality sensor. The Decision Tree, similarly to the Random Forest algorithm, is also focusing on the remotely controllable blinds flag. In this model, the difference between the current and the intended temperature has more importance than the AirQualitySensor. Gradient Boosting focuses on the same features as the other models, however, the difference between the current and the intended temperature is now the most important feature, and the remotely controllable blinds flag and air quality sensor have the same importance.

Concluding this sub-section, boosting proved to be a slightly better solution over bagging. The way boosting improved this problem was due to the use of a Genetic Algorithm that optimizes the data, reducing a lot of the noise in the data. The noise reduction provided by the Genetic Algorithm has a greater impact on the gradient boosting algorithm since this technique is known to overfit easily with noisy data when comparing with the Random Forest algorithm. The use of the Genetic Algorithm may also increase slightly the bias, which has a greater negative impact on the Random Forest algorithm since bagging tries to reduce the error by reducing the variance while boosting techniques try to reduce the error by reducing bias. This leads to boosting not being impacted so much with this problem while being able to take more advantage of the benefits of the optimization phase.

Real-time variable control optimization

Using the same training and test data set, multiple models were created with different dependent variables to search for the highest performance polynomial equation. The polynomial equation degree is not taken into account since the function GridSearchCV (GridSearchCV, n.d.), already provides the highest performance degree between 1 and 10.

Table 2 Open Blinds variable encoding

True value	OpenBlinds_1	OpenBlinds_2
False	0	1
True	1	0
None	1	1

Table 3 Target precision of the trained classification models

Target/Model	Random Forest	Decision Tree	Gradient Boosting
AirConditionerTemp	0.93	0.95	0.94
ArtificialLightningLight	0.98	0.99	0.99
OpenDoors	0.92	0.91	0.94
OpenWindows	0.94	0.94	0.95
OpenBlinds_1	0.93	0.93	0.94
OpenBlinds_2	0.96	0.92	0.96

The performance of the best air conditioner temperature prediction model obtained, using a validation set (i.e., different subset than the one used for training) has an MAE value of 0.16892 and RMSE of 0.23377, which indicates that the error is low to medium. Also, R^2 with 0.57131 and Adjusted R^2 of 0.57092 imply that the model has acceptable results in predicting with a high number of dependent variables. These results are expected, since the temperature deviation, represented by eq. (2), is complex, which makes it harder to create a function that fits a model capable of predicting reliably the air conditioner temperature.

Figure 5 represents the mean absolute error (i.e., MAE) between the resource optimized air conditioner temperature, provided by the complex resource optimization, and the control optimized air conditioner temperature, obtained through the Polynomial Regression model. The proposed Polynomial regression model for the air conditioner temperature performs close to the resource optimization values, having a maximum monthly average absolute error of 0.214 °C, and a minimum of 0.024 °C.

The performance for the best artificial lighting luminosity prediction model obtained, using a validation set, has very low levels of error, with an MAE value of 0.01855 and an RMSE of 0.05835. Contrary to the air conditioner model, the prediction of the artificial lighting luminosity shows that the model is suited almost perfectly. The metrics R^2 of value 0.95977 and Adjusted R^2 of 0.95973 show that the function has high accuracy with the number of chosen dependent variables, having less than the number of dependent variables from the air conditioner model. Since the luminosity deviation equation is much simpler, (3), the Polynomial Regression is able to more easily adapt to a function capable of predicting the artificial lighting luminosity with high accuracy.

The MAE between the resource optimized artificial lighting luminosity, provided by the resource optimization, and the control optimized artificial lighting luminosity, obtained through the Polynomial Regression model, is represented in Fig. *5*. Comparing with the air conditioner model, the artificial lighting luminosity model seems less adaptable (e.g., from May to September the control optimized absolute error values are very high, compared with the rest of the year), however, we have to take into account

Table 4 The three most important features of classification models

Model/Feature	Random Forest	Decision Tree	Gradient Boosting
Most important feature	ControllableBlinds (0.210)	ControllableBlinds (0.240)	DifBetweenCurrentAndIntendedTemp (0.175)
Second most important feature	AirQualitySensor (0.180)	DifBetweenCurrentAndIntendedTemp (0.180)	AirQualitySensor (0.120)
Third most important feature	DifBetweenCurrentAndIntendedTemp (0.95)	AirQualitySensor(0.90)	ControllableBlinds (0.120)

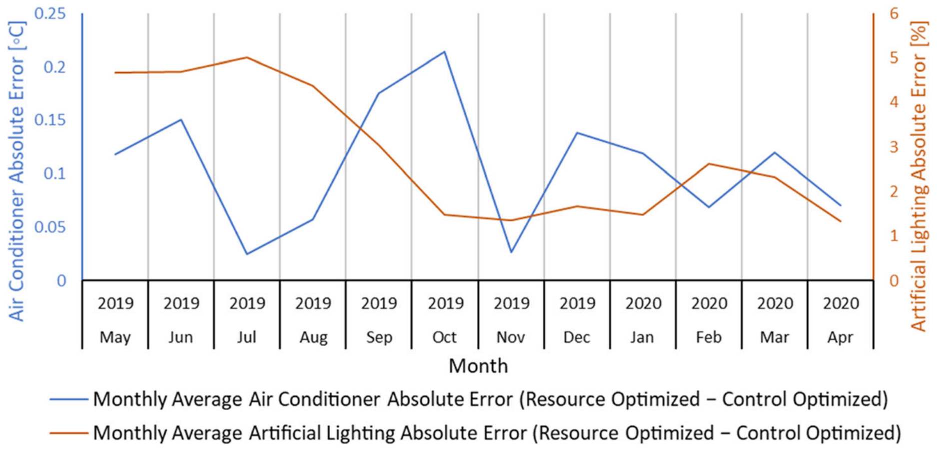

Fig. 5 Air conditioner and artificial lighting absolute errors

that the range for the luminosity is much wider than the temperature, with 0% to 100% and 18 °C to 28 °C, respectively. The maximum monthly average absolute error, for the artificial lighting luminosity, is 5.020%, and a minimum of 1.325%.

Conclusion

This paper proposes a methodology, using a Genetic Algorithm, classification models, and Polynomial Regression models, to manage the climatization and luminosity of rooms in a building. The proposed methodology is able to minimize energy costs while taking into account its users' preferences and improving indoor air quality. The consideration of users' preferences enables the solution to manage energy loads and resources avoiding the generation of negative impact to users, contributing to their acceptance and engagement.

The results showed that the proposed methodology is able to achieve high-quality results in real-time, using Random Forest for discrete control and Polynomial Regressions for variable control, with the actions taken by the proposed solution contributing to building management systems (BMS) and energy management systems (EMS), promoting a real-time optimization and management of resources. The real-time optimization results were compared to the results provided by the offline optimization that used a Genetic Algorithm, demonstrating the real-time capabilities.

Regarding future work, the main development path would be the execution of further testing of the proposed methodology in real-world scenarios, through the integration of the achieved solution onto the GECAD' energy management system. Additionally, the proposed methodology also presents some limitations in order to become commercially viable, such as not being able to process missing or corrupted data (e.g., no data or corrupted data provided by a faulty temperature sensor). Also, the proposed solution demands the building to have controllable resources, such as air conditioner units, and lighting, and have sensors that can measure temperature, clarity, and air quality. Such limitations could easily be surpassed by implementing systems capable of detecting faulty sensor data and correct it accordingly, in case of missing resources, the system could be used to provide notifications for the users to interact manually with the resources. However, the proposed system needs the installation of sensors in the building.

The results of the proposed solution allow the validation of real-time optimizations inside a smart building considering the minimization of energy costs while maximizing the user comfort and health promoted through the maintenance or air quality.

Abbreviations

ANNs: Artificial Neural Networks; BMS: Building Management System; DTs: Decision Trees; EMS: Energy Management System; GA: Genetic Algorithm; HVAC: Heating, Ventilating and Air Conditioning; MAE: Mean Absolute Error; MIBEL: Iberian Electricity Market; PSO: Particle Swarm Optimization; RMSE: Root-Mean-Square Error; SUS: Stochastic Universal Sampling; VOC: Volatile Organic Compound

Acknowledgments

Not applicable.

About this supplement

This article has been published as part of Energy Informatics Volume 4, Supplement 2 2021: Proceedings of the Energy Informatics.Academy Conference Asia 2021. The full contents of the supplement are available at https://energyinformatics.springeropen.com/articles/supplements/volume-4-supplement-2.

Authors' contributions

All the authors made contributions to the conception of the proposed solution. The system architecture design and software developments were mainly done by BM, MA, HP, and JS. Data acquisition and analysis were done by BM, MA, HP, JS, and LG. The interpretation of data was done by all the authors. The first draft was written by BM, and MA, while all other authors contributed to the final version of the paper. All authors read and approved the final manuscript.

Funding

This work has received funding from FEDER Funds through COMPETE program and from National Funds through (FCT) under the projects UIDB/00760/2020 and MAS-Society (PTDC/EEI-EEE/28954/2017).

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Engineering - Polytechnic of Porto, Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto, Portugal. [2]GECAD – Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development, Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto, Portugal.

Published: 24 September 2021

References

Abrishambaf O, Faria P, Gomes L, Spínola J, Vale Z, Corchado JM (2017) Implementation of a Real-Time Microgrid Simulation Platform Based on Centralized and Distributed Management. Energies 10:806. https://doi.org/10.3390/en10060806

Ahmad MW, Mourshed M, Rezgui Y (2017) Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build 147:77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Ahmad MW, Mourshed M, Yuce B, Rezgui Y (2016) Computational intelligence techniques for HVAC systems: A review. Build Simul 9(4):359–398. [cited 2021 Mar 3]. https://doi.org/10.1007/s12273-016-0285-4

Ali S, Kim DH (2015) Optimized Power Control Methodology Using Genetic Algorithm. Wirel Pers Commun 83(1):493–505 [cited 2021 Mar 3]. Available from: https://link.springer.com/article/10.1007/s11277-015-2405-3

Azar AT, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. Comput Methods Programs Biomed. 2014;113(2):465–473 [cited 2021 Mar 3]. Available from: https://doi.org/10.1016/j.cmpb.2013.11.004

Catalina T, Virgone J, Blanco E (2008) Development and validation of regression models to predict monthly heating demand for residential buildings. Energy Build 40(10):1825–1832. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378778808000844. https://doi.org/10.1016/j.enbuild.2008.04.001

Chen YT, Piedad E, Kuo CC (2019) Energy consumption load forecasting using a level-based random forest classifier. Symmetry (Basel) 11(8):1–9

Energy statistics - an overview - Statistics Explained. (n.d.) [cited 2021 Jan 10]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview#Final_energy_consumption

Faia R, Pinto T, Abrishambaf O, Fernandes F, Vale Z, Corchado JM (2017 Nov 15) Case based reasoning with expert system and swarm intelligence to determine energy reduction in buildings energy management. Energy Build. 155:269–281. https://doi.org/10.1016/j.enbuild.2017.09.020

Faria P, Vale Z, Baptista J (2015 Mar 15) Constrained consumption shifting management in the distributed energy resources scheduling considering demand response. Energy Convers Manag 93:309–320. https://doi.org/10.1016/j.enconman.2015.01.028

Fernandes F, Sousa T, Silva M, Morais H, Vale Z, Faria P. Genetic algorithm methodology applied to intelligent house control. IEEE SSCI 2011 - Symp Ser Comput Intell - CIASG 2011 2011 IEEE Symp Comput Intell Appl Smart Grid. 2011;(April):139–146

Franco JT. How to Calculate the Thermal Transmittance (U-Value) in the Envelope of a Building. 2018 [cited 2021 Jan 20]. Available from: https://www.archdaily.com/898843/how-to-calculate-the-thermal-transmittance-u-value-in-the-envelope-of-a-building

Xin-She Yang (2021) Chapter 6 - Genetic Algorithms, Editors: Xin-She Yang, Nature-Inspired Optimization Algorithms (Second Edition), Academic Press 91-100. https://doi.org/10.1016/B978-0-12-821986-7.00013-5

GridSearchCV. (n.d.)[cited 2021 Mar 3]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Huang W, Lam HN (1997) Using genetic algorithms to optimize controller parameters for HVAC systems. Energy Build 26(3): 277–282. https://doi.org/10.1016/S0378-7788(97)00008-X

IPCC 2014 (2014) Climate Change 2014: Synthesis Report. In: Team CW, Pachauri RK, Meyer LA (eds) Report of the Intergovernmental Panel on Climate Change, Geneva [cited 2021 Mar 3]. Available ferom: http://www.ipcc.ch

Mota B, Gomes L, Faria P, Ramos C, Vale Z, Correia R (2021) Production Line Optimization to Minimize Energy Cost and Participate in Demand Response Events. Energies 14:462. https://doi.org/10.3390/en14020462

Nguyen T, Nassif N (2016) Optimization of HVAC Systems Using Genetic Algorithm. In: Proceedings of the 2013 National Conference on Advances in Environmental Science and Technology. Springer International Publishing [cited 2021 Mar 3]. p. 203–9. Available from: https://link.springer.com/chapter/10.1007/978-3-319-19923-8_21

Pant A. Introduction to Linear Regression and Polynomial Regression. 2019 [cited 2021 Jan 20]. Available from: https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

Ramzai J. Simple guide for ensemble learning methods. 2019 [cited 2021 Mar 3]. Available from: https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2

RandomizedSearchCV. (n.d.)[cited 2021 Jan 31]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Sharada R. Introduction of Holdout Method - GeeksforGeeks. 2020 [cited 2021 Jan 31]. Available from: https://www.geeksforgeeks.org/introduction-of-holdout-method/

Vale Z, Morais H, Faria P, Khodr H, Ferreira J Kadar P (2010) "Distributed energy resources management with cyber-physical SCADA in the context of future smart grids," Melecon 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference 431-436. https://doi.org/10.1109/MELCON.2010.5476239

Yiu T. Understanding Random Forest. How the Algorithm Works and Why it Is So Effective. 2019 [cited 2021 Jan 10]. Available from: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.