www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(1), Jan-April 2025, 258-265. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i1.1057

MASTER'S THESIS SUMMARY

Optimizing Energy Consumption in Smart Buildings by Using Machine Learning Algorithms

Student*: Luis Arturo Ortiz-Suarez*
Universidad Politécnica de Pachuca
lortizsuarez@gmail.com

Advisor: *Jorge A. Ruiz-Vanoye*
Universidad Politécnica de Pachuca
jorgeruiz@upp.edu.mx

Advisor: *Francisco Rafael Trejo-Macotela*
Universidad Politécnica de Pachuca
trejo_macotela@upp.edu.mx

Abstract. In this article, an approach for optimizing energy consumption in smart buildings using machine learning algorithms is presented. Utilizing TDSP, data on climatic conditions, occupancy, and energy consumption obtained from EnergyPlus software are integrated. Feature selection and feature importance techniques, as well as statistical analyses, are implemented to select variables that are used to train machine learning models such as MLP neural networks, support vector machines, random forest, and XGBRegressor for predicting energy consumption, with accuracy evaluated using RMSE. It was demonstrated that models based on neural networks offer better accuracy, thereby enabling measures to achieve energy optimization.

Keywords: Energy optimization with AI, energy optimization in smart buildings, energy consumption prediction.

Article Info
Received February 20, 2025
Accepted February 22, 2025

1 Introduction

The increase in energy demand has made it evident that there is a need to find ways to optimize its use, and in turn, it has driven the development of technologies that enable a more efficient management of energy consumption in smart buildings. The integration of various devices and advancements in machine learning algorithms open opportunities to develop systems that predict energy consumption and, thereby, take actions that contribute to optimizing the use of electrical energy.

This work focuses on developing and evaluating machine learning models that help implement measures to optimize energy consumption. The first section discusses the issues related to energy consumption as well as the objectives to be achieved with this research, highlighting how global population growth similarly impacts the demand for resources needed to meet requirements, particularly in buildings.

Section 2 presents an analysis of various studies that focus on optimizing energy consumption using machine learning algorithms and the results they have achieved. It describes the algorithms used, the variables involved, and the approach through which their efforts are directed.

The following sections describe the proposed solution, briefly outlining how the TDSP methodology is implemented, the processes followed, and the results obtained. Comparisons are made among the various developed models and how the selection of the dataset affects the achievable outcomes. The document concludes with brief conclusions and observations for the future of this work.

1.1 Motivations

The need to implement methods that help address the growing energy demand and curb the pollution generated in its production is the primary reason why developing solutions like the one proposed in this document is important.

Furthermore, there are various circumstances in which optimizing the use of electrical energy is crucial. In some cases, access to energy is limited by technology, infrastructure, or resources, making it essential to develop new ways to maximize the resource.

With advancements in current information and communication technologies and artificial intelligence, significant solutions have been achieved in various fields. Applying these advancements to energy consumption offers the opportunity to develop new tools that support decision-making and action.

1.2 Description of the research problem

There are three major issues related to energy and buildings: the growing energy demand, the costs generated by its consumption, and the pollution produced in the generation of electrical energy. The following paragraphs describe and analyze these problems.

In general, the disadvantage of centralized electrical systems, according to Chinchero & Alonso (2020), is the lack of communication among the components of the traditional network, the limitation on system expansion, and large-scale failures of the entire system due to failures in a small part of the module.

One of the main problems in electricity production is the emission of greenhouse gases produced by power plants. Deb & Gao (2022) indicate that the energy crisis, the low air quality index, and global warming have been some of the main concerns over the past decade.

Bastidas Paz & Chinchero Villacís (2023) state that most of the energy for daily consumption is generated by plants that utilize fossil fuels, nuclear reactions, and hydroelectric power; however, these have an environmental impact due to their greenhouse gas emissions.

The figures provided by La et al. (2016) offer an example of the role buildings play in electricity consumption around the world, indicating that, of the generated energy, non-residential or commercial buildings consumed 31% in Singapore in 2005 and around 35% in the United States in 2010. These figures are expected to increase given the global population growth and urbanization rates.

Energy consumption remains the primary cause of climate change, as it accounts for around 60% of global greenhouse gas emissions (United Nations, n.d.).

Between 50% and 70% of the energy demand in smart buildings is constituted by ventilation, air conditioning, and heating systems required to maintain comfort in terms of temperature, humidity, and air purification for occupants (Montalvo García, 2020).

Traditional lighting systems do not optimize energy use and do not adapt to the specific needs of the spaces they occupy (Bastidas Paz & Chinchero Villacís, 2023).

Molla et al. (2018) explain that due to the ever-increasing demand for electrical energy, the percentage of energy generation increases considerably along with rising energy prices. The most common pricing schemes include real-time pricing (RTP), time-of-use pricing (TOUP), and critical peak pricing (CPP).

Another problem with energy consumption in buildings is the lack of adaptability of energy management systems to the usage patterns of different areas and the changes in spaces required to adapt to the climate.

By integrating renewable energy sources such as solar or wind power, we face the uncertainty these sources offer, as electricity generation depends on climatic conditions, which results in a reliance on traditional energy sources (Chen et al., 2015).

Centeno & Ramírez (2024) mention that there is a lack of research on the integration of AI in hybrid systems that combine multiple renewable energy sources. They also discuss the exponential growth in energy demand, driven by economic development and population growth.

1.3 Objectives of the tesis

- Identify the most influential variables in energy consumption through the analysis of energy consumption data from smart buildings in order to develop prediction algorithms.
- Create machine learning models to predict energy consumption using the identified variables and parameters.
- Compare the machine learning model with traditional methods using accuracy metrics to validate the model's performance.

1.4 Brief description of the contribution of the thesis

Through the simulation of a building, a dataset is obtained that encompasses variables related to electrical consumption, climate, and occupancy that affect energy consumption, which is then statistically analyzed to observe their impact on energy usage.

Once this is done, various feature selection techniques are applied to obtain different datasets, so that it can later be observed how each one influences the results produced by machine learning models for predicting a building's energy consumption. In this way, a comparison is made between the feature selection techniques and the machine learning models. At this point, a comparison is made between the machine learning methods and traditional statistical methods.

Finally, it is proposed to develop a machine learning algorithm that helps optimize real-time energy consumption. This algorithm would function as a controller in a smart grid of a smart building, making decisions regarding the switching on and off various electrical devices—mainly the HVAC system—as well as managing the distribution of energy obtained from various sources.

2 Background

In this section, the context of smart buildings is addressed, along with how artificial intelligence serves as a tool to solve energy consumption issues within them. Additionally, various studies are analyzed to aid in understanding both the problem and its potential solutions.

2.1 Important Issue 1

A smart building is one that integrates technology to monitor both the interactions that occur within it and the resources being used, thereby enabling informed decisions regarding various systems such as HVAC, lighting, and electrical devices for proper management.

Studies such as that by Naug et al. (2019) demonstrate that by applying deep reinforcement learning techniques to real-time temperature control systems, an average energy savings of 3% can be achieved. Additionally, Castillo De La Barrera (2023) found that by applying AI methods in HVAC systems, energy consumption reductions between 0.9% and 40% can be achieved.

2.2 Major Theme 2

Machine learning is a tool that allows us to analyze large amounts of data to detect complex patterns. In the context of buildings, techniques such as ANN, SVR, RF, GBRT, and CNN have been applied for predicting energy demand (Bendaoud et al., 2022), while metaheuristic algorithms like gray wolf optimization, genetic algorithms, bee colony, ant colony, etc., have been used to find configurations for electronic devices that help reduce energy consumption (Molla et al., 2018).

Deep learning methods face challenges regarding the availability of high-quality data with enough diversity to analyze the various aspects encountered by buildings, which would enable the algorithms to better perform the tasks for which they were developed.

2.3 Related work

In this chapter, various studies and works are presented in which different artificial intelligence techniques have been applied to optimize electrical energy—covering aspects such as energy consumption prediction, optimal energy distribution, and the search for configurations in various devices to achieve energy savings.

The study by Naug & Biswas (2018) focuses on predicting and reducing energy consumption in large buildings by combining supervised and unsupervised learning techniques, primarily targeting HVAC systems, which are the largest energy consumers. Machine learning algorithms such as AdaBoost, Random Forest, and Support Vector Regression (SVR) are used and compared for prediction, while the Stochastic Gradient Descent (SGD) method is employed for optimization. Using data from the Vanderbilt Alumni Hall, a 12% energy savings is demonstrated.

In the study conducted by Le et al. (2019), four new artificial intelligence (AI) techniques are proposed to forecast the thermal load for the energy efficiency of buildings, based on the potential of artificial neural networks (ANN) and metaheuristic algorithms. These include optimization via artificial bee colony (ABC), particle swarm optimization (PSO), imperialist competitive algorithm (ICA), and genetic algorithm (GA). The aforementioned metaheuristic algorithms are used to calculate the connection weight biases and the biases of the artificial neural network used for making predictions.

Zhong et al. (2019) propose a novel support vector regression model based on vector fields (SVR) to predict energy consumption in buildings. The objective is to improve the accuracy, robustness, and generalization capacity of the predictions. A comparison is made with other models such as multiple linear regression (MLR), backpropagation neural networks (BPNN), standard SVR, extreme deep learning (EDL), and gradient boosting regression (GBR). The model was applied to predict the cooling load of a building in the city of Tianjin, China.

In the project carried out by Montalvo García (2020), predictive models based on ARIMA, TBATS, and RNN are studied and developed, with a focus on LSTM networks selected for their ability to detect short- and long-term patterns in the data. The study is complemented with classical statistical models for time series prediction, providing a baseline against which the results obtained with the machine learning models can be compared.

Yu et al. (2021), with the aim of improving prediction accuracy in electrical demand, combine three techniques: the ARIMA model, GAN, and the Wavelet transform. The implemented methodology has three main stages: capturing the linear component of the demand, decomposing the non-linear residue into sub-frequencies, and finally predicting the sub-frequencies and summing the ARIMA and GAN results into a final prediction. This approach demonstrates a significant improvement compared to methods such as ARIMA, ANN, and SVR.

Bendaoud et al. (2022) develop an approach to predict electrical energy demand in the Algeria region using load profiles (LP) generated for different time periods (annually, weekly, and daily) with machine learning techniques. They apply ANN, SVR, RF, GBRT, and CNN—with a particular focus on CNN—and found that CNNs exhibit high predictive accuracy; however, when using LPs, the accuracy was lower.

Deepanraj et al. (2022) propose a short-term electrical load prediction using neural networks by implementing the wild geese algorithm to optimize hyperparameters. They achieve better results with fewer errors compared to methods such as SVR, LSTM, and SVR-LSTM. The predictions are based on domestic and commercial load profiles.

Kavitha et al. (2022) apply a hybrid deep learning model to predict the heating and cooling electrical load. The model combines a convolutional neural network (CNN), long short-term memory (LSTM), and bidirectional long short-term memory (BiLSTM), and it uses the Harris Hawks algorithm to adjust the hyperparameters. They utilize structural variables of the building, and the methodology consists of data preprocessing, model development, and hyperparameter tuning.

Cai et al. (2023) apply support vector regression (SVR) combined with metaheuristic optimization algorithms to predict the load that heating or cooling systems in buildings will use. Their approach is based on the structural characteristics of the buildings, and the optimization algorithms are used to fine-tune the hyperparameters of the SVR model.

The study conducted by Freire et al. (2023) aims to develop prediction systems that outperform traditional methods in terms of electrical generation at the Illuchi hydroelectric power plant. A dataset covering the period from 2008 to 2020 from the power plant is used; these data are split into 70% for training and 30% for validation. The study implements and compares the use of artificial neural networks (ANN) and gated recurrent units (GRU) using the ADAM optimizer. It is found that LSTM achieves slightly better prediction accuracy compared to GRU.

3 Proposed Solution Approach

The methodology used is based on Microsoft's Team Data Science Process (TDSP), which we will apply in the context of energy optimization in smart buildings, since it "combines elements of data science, software engineering, and agile processes" (Hotz, 2018). It begins with an understanding of the problem by defining the objective of predicting and optimizing energy consumption. It then moves on to data understanding, which involves acquiring data through the simulation of a building to obtain consumption, climatic, and occupancy variables, allowing for data preprocessing and feature selection. Next, the modeling stage is carried out, during which machine learning models—such as artificial neural networks for consumption prediction and reinforcement learning techniques for real-time optimization—are developed and compared using a simulated environment of a smart grid in a smart building. The process concludes with the integration of the model into the simulated environment for validation and feedback.

With the proposed methodology, the iterative nature of its phases is leveraged to apply statistical techniques, feature selection, and feature importance for variable selection. In this way, we can train and validate machine learning algorithms using the obtained datasets and assess the impact of these variables on the model outcomes.

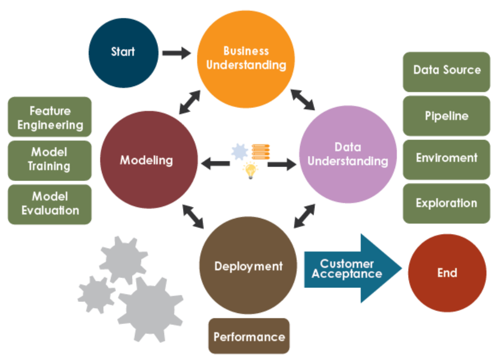

Figure 1. Microsoft TDSP methodology process (Hotz, 2018).

Within the algorithms to be evaluated are neural networks, support vector machines, XGBRegressor, Keras Regressor, and Random Forest Regressor, in addition to applying an ensemble learning approach that combines machine learning models to improve accuracy and robustness (Mohammed & Kora, 2023).

Regarding the prediction of energy consumption, accuracy will be measured using the root mean square error (RMSE), and optimization will be verified in terms of kW/h differences relative to the simulation data.

The dataset is obtained by simulating a building using EnergyPlus software version 24.2, as it allows for modeling various components and generating energy consumption reports (EnergyPlus, n.d.). The variables used for the analysis are:

- Date/Time
- InteriorEquipment:Electricity [J](Hourly)
- InteriorLights:Electricity [J](Hourly)
- Electricity:HVAC [J](Hourly)
- Fans:Electricity [J](Hourly)
- Refrigeration:Electricity [J](Hourly)
- ElectricityProduced:Facility [J](Hourly)
- Photovoltaic:ElectricityProduced [J](Hourly)
- ExteriorLights:Electricity [J](Hourly)
- Cooling:Electricity [J](Hourly)
- Electricity:Facility [J](Hourly)
- Environment:Site Outdoor Air Humidity Ratio [kgWater/kgDryAir](Hourly).

- Environment:Site Wind Speed [m/s](Hourly).
- Environment:Site Wind Direction [deg](Hourly).
- Environment:Site Sky Temperature [C](Hourly).
- Environment:Site Horizontal Infrared Radiation Rate per Area [W/m2](Hourly).
- People occupant count: entrega un reporte del número de personas en cada habitación.

In the data understanding phase, the following processing steps are carried out to perform the time series analysis:

- A single file is generated that combines all the records from both obtained reports.
- A conversion of the consumption metrics from the data obtained in the reports is performed.
    - Conversion from Julios to kW/h.
    - Conversion from Watts to kW/h.
- A column is generated with the total number of occupants in the building.
- A year is added to the date/time column to comply with the date format and enable time series analysis.

4 Experimental results

A statistical analysis of the available variables was conducted, initially applying a correlation analysis using Spearman's coefficient. This analysis revealed that the most important variables for energy consumption, without interdependency, are HVAC, total occupancy, lighting, and interior equipment. Additionally, the most critical climatic variable is temperature, as shown in the heatmap in Figure 1.

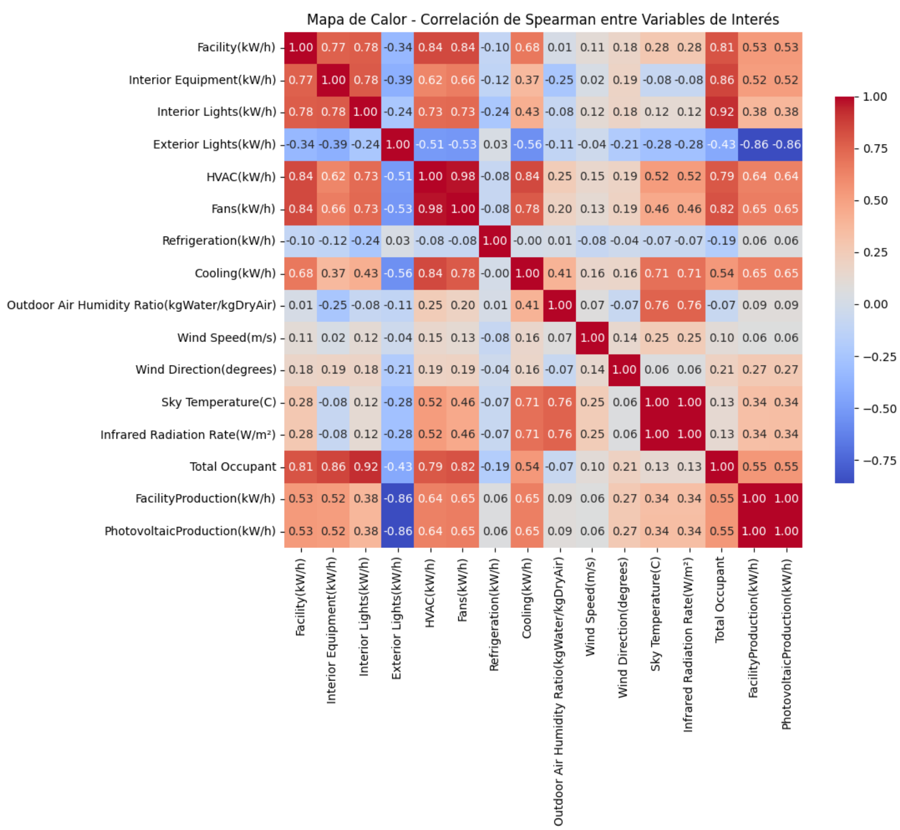

Figure 2. Spearman coefficient heat map. (Own elaboration, 2025)

Within this analysis, measures of central tendency, data distribution analysis, percentile analysis, and seasonal analysis are conducted, resulting in a dataset without outlier records.

In the modeling phase, the accuracy of the Keras Regressor, Neural Network (MLP), Random Forest Regressor, Support Vector Machine (SVR), and XGBRegressor models are evaluated using RMSE. Table 1 presents these results, applying the datasets obtained through feature selection techniques.

Table 1. RMSE with feature selection.

	Backward Elimination	Mutual Information	Recursive Feature Elimination (RFE)	Variance Threshold
Keras Regressor	1.275428848	1.335675642	1.678062718	1.195154758
Neural Network (MLP)	0.699321023	0.711946748	0.91663323	0.671586502
Random Forest Regressor	3.870392689	3.723547237	3.704117998	3.885881702
Support Vector Machine (SVR)	28.68902389	23.8256986	26.52675301	29.13542355
XGBRegressor	3.987850317	3.97702206	3.522636004	3.955199328

The following table shows the RMSE of the machine learning models' accuracy when these models are trained with the datasets obtained through the application of feature importance techniques.

Table 2.RMSE with feature importance

	Decision Tree	Extra Trees	Random Forest	LASSO	Greedy Selection
Keras Regressor	1.69834858	0.95076538	1.119732593	1.12139474	2.25581072
Neural Network (MLP)	0.90851014	0.82108634	0.792092153	0.71370946	2.11893499
Random Forest Regressor	3.72209417	3.66911008	3.712471535	3.58525127	3.45318088
Support Vector Machine (SVR)	26.6863709	21.3416018	23.8256986	25.2648513	22.4777267
XGBRegressor	3.93959529	4.00906769	3.982473177	3.87915773	3.11526544

Finally, Table 3 shows the accuracy of the different selected models using the dataset obtained from the statistical analysis.

Table 3. RMSE with statistical analysis.

Model	Precision
Neural Network (MLP)	12.5498995
Support Vector Machine (SVR)	25.6832274
XGBRegressor	8.74757473
Random Forest Regressor	8.87646993
Keras Regressor	14.8418609

5 Conclusions

Variable selection methods impact the prediction outcomes of the models, which is why it is important to conduct an appropriate data science process to generate higher-quality datasets that provide relevant information to the learning models. In this regard, it is also important to mention that the adoption of the TDSP methodology allows complex problems such as energy optimization in smart buildings to be addressed effectively, as it enables interaction between the different phases when modifications are necessary. Up to this point, it has been identified that the machine learning model with the best performance—regardless of the dataset used—is the MLP (Multi-Layer Perceptron), a deep learning technique. Furthermore, this model performs better when feature selection techniques are applied.

An important aspect to consider is that when obtaining datasets through an energy consumption simulation, there is a possibility that the models capture the data generation patterns of the software. Therefore, they must be evaluated using data from other sources or different datasets to validate their robustness.

6 References

Bastidas Paz, C. H., & Chinchero Villacís, H. F. (2023). El Sistema de Gestión de Energía para Edificios Inteligentes utilizando Generación Distribuida. *Latin-American Journal of Computing*, *10*(1), Article 1.

Bendaoud, N. M. M., Farah, N., & Ben Ahmed, S. (2022). Applying load profiles propagation to machine learning based electrical energy forecasting. *Electric Power Systems Research*, *203*, 107635. https://doi.org/10.1016/j.epsr.2021.107635

Cai, W., Wen, X., Li, C., Shao, J., & Xu, J. (2023). Predicting the energy consumption in buildings using the optimized support vector regression model. *Energy*, *273*, 127188. https://doi.org/10.1016/j.energy.2023.127188

Castillo De La Barrera, J. C. (2023). *Metodologías de ahorro energético aplicadas a los sistemas HVAC utilizando inteligencia artificial: Una revisión del estado del arte.* https://repositorio.itm.edu.co/handle/20.500.12622/6023

Centeno, M. A. C., & Ramírez, L. G. B. (2024). Smart energy soportado por Inteligencia Artificial: Un futuro sostenible y eficiente. *Encuentro Internacional de Educación en Ingeniería*. https://doi.org/10.26507/paper.3926

Chen, G., Lewis, F. L., Feng, E. N., & Song, Y. (2015). Distributed Optimal Active Power Control of Multiple Generation Systems. *IEEE Transactions on Industrial Electronics*, *62*(11), 7079–7090. IEEE Transactions on Industrial Electronics. https://doi.org/10.1109/TIE.2015.2431631

Chinchero, H. F., & Alonso, J. M. (2020). A Review on Energy Management Methodologies for LED Lighting Systems in Smart Buildings. *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*, 1–6. https://doi.org/10.1109/EEEIC/ICPSEurope49358.2020.9160796

Deb, S., & Gao, X.-Z. (2022). A hybrid ant lion optimization chicken swarm optimization algorithm for charger placement problem. *Complex & Intelligent Systems*, *8*(4), Article 4. https://doi.org/10.1007/s40747-021-00510-x

Deepanraj, B., Senthilkumar, N., Jarin, T., Gurel, A. E., Sundar, L. S., & Anand, A. V. (2022). Intelligent wild geese algorithm with deep learning driven short term load forecasting for sustainable energy management in microgrids. *Sustainable Computing: Informatics and Systems*, *36*, 100813. https://doi.org/10.1016/j.suscom.2022.100813

EnergyPlus. (n.d.). *EnergyPlus*. Retrieved December 11, 2024, from https://energyplus.net/

Freire, F. S. B., Fiallos, J. N. C., Caiza, C. I. Q., & Peña Herrera, H. R. R. (2023). Unidades recurrentes cerradas (GRU) vs redes neuronales artificiales en la predicción de la generación eléctrica de la Central Hidroeléctrica Illuchi. *AlfaPublicaciones*, *5*(3), Article 3. https://doi.org/10.33262/ap.v5i3.395

Hotz, N. (2018, September 11). *What is TDSP?* Data Science PM. https://www.datascience-pm.com/tdsp/

Kavitha, R. J., Thiagarajan, C., Priya, P. I., Anand, A. V., Al-Ammar, E. A., Santhamoorthy, M., & Chandramohan, P. (2022). Improved Harris Hawks Optimization with Hybrid Deep Learning Based Heating and Cooling Load Prediction on residential buildings. *Chemosphere*, *309*, 136525. https://doi.org/10.1016/j.chemosphere.2022.136525

La, Q. D., Chan, Y. W. E., & Soong, B.-H. (2016). Power Management of Intelligent Buildings Facilitated by Smart Grid: A Market Approach. *IEEE Transactions on Smart Grid*, *7*(3), 1389–1400. IEEE Transactions on Smart Grid. https://doi.org/10.1109/TSG.2015.2477852

Le, L. T., Nguyen, H., Dou, J., & Zhou, J. (2019). A Comparative Study of PSO-ANN, GA-ANN, ICA-ANN, and ABC-ANN in Estimating the Heating Load of Buildings' Energy Efficiency for Smart City Planning. *Applied Sciences*, *9*(13), Article 13. https://doi.org/10.3390/app9132630

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774. https://doi.org/10.1016/j.jksuci.2023.01.014

Molla, T., Khan, B., & Singh, P. (2018). A comprehensive analysis of smart home energy management system optimization techniques. *Journal of Autonomous Intelligence*, *1*(1), Article 1. https://doi.org/10.32629/jai.v1i1.14

Montalvo García, D. (2020). *Construcción de algoritmos de aprendizaje automático para predicción de consumos energéticos en edificios inteligentes* [Universidad de Valladolid]. https://uvadoc.uva.es/handle/10324/42713

Naug, A., Ahmed, I., & Biswas, G. (2019). Online Energy Management in Commercial Buildings using Deep Reinforcement Learning. *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 249–257. https://doi.org/10.1109/SMARTCOMP.2019.00060

Naug, A., & Biswas, G. (2018). Data Driven Methods for Energy Reduction in Large Buildings. *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, 131–138. https://doi.org/10.1109/SMARTCOMP.2018.00083

Organización de las Naciones Unidas, M. (n.d.). Objetivos de desarrollo sustentable. *Desarrollo Sostenible*. Retrieved August 2, 2024, from https://www.un.org/sustainabledevelopment/es/energy/

Yu, F., Yue, Q., Yunianta, A., & Aljahdali, H. M. A. (2021). A novel hybrid deep correction approach for electrical load demand prediction. *Sustainable Cities and Society*, *74*, 103161. https://doi.org/10.1016/j.scs.2021.103161

Zhong, H., Wang, J., Jia, H., Mu, Y., & Lv, S. (2019). Vector field-based support vector regression for building energy consumption prediction. *Applied Energy*, *242*, 403–414. https://doi.org/10.1016/j.apenergy.2019.03.078