Predicting Energy Consumption of Florida Single-Family Houses using ANNs

Rita Elias, ritaelias@ufl.edu
Rinker School of Construction Management, University of Florida, Gainesville, FL, USA

Raja R. A. Issa, raymond-issa@ufl.edu
Rinker School of Construction Management, University of Florida, Gainesville, FL, USA

Abstract
Building energy-efficient residences throughout all of Florida plays an important role in reducing greenhouse gas emissions and fossil fuel consumption. The owner's energy efficiency goals are best achieved when they are taken into account right from the beginning of the conceptual design phase. Conventional engineering modeling methods have proved to be an extremely time-consuming process and lacking in terms of energy analysis. Machine learning (ML) methods have recently been shown to effectively forecast the buildings' energy consumption and to replace such energy modeling tools. This study established an Artificial Neural Network (ANN) model to predict the heating and cooling loads on detached houses based on a big dataset of more than 18 thousand newly constructed single-family houses in Florida between the years 2009 and 2019. The ANN will assist designers in decision-making regarding the conceptual design of houses by investigating different design alternatives using generative design for energy efficiency purposes.

Keywords: Energy prediction, Energy efficiency, Cooling Load, Heating Load, Single-Family Houses, Supervised learning, Machine learning, Artificial intelligence, Artificial neural network

1 Introduction

The buildings and buildings construction sectors account for more than one-third of the worldwide delivered energy consumption and about 40% of total Carbon dioxide (CO2) emissions. According to the U.S. Energy Information Administration (2019), energy needs are anticipated to keep on increasing about 1.3% each year from 2018 to 2050. The residential sector in Florida expends 1,197 trillion Btu representing around 27% of the total primary energy available (U.S. Energy Information Administration EIA 2019). A population growth of around 350,000 people per year is the main cause behind the rise in housing demands in Florida. Building energy-efficient dwellings throughout all of Florida plays an important role in reducing greenhouse gas emissions and fossil fuel consumption.

The subject of building energy performance prediction has caught the attention of many researchers in the last 30 years. They have suggested several tools to predict the energy loads on buildings. These tools can be split into three categories: engineering, Artificial Intelligence (AI) based, and hybrid methods. The engineering methods calculate and analyze building energy loads and thermal dynamics using the physical features and characteristics of building elements and materials (Sun et al. 2020). The internal logic of these engineering methods is known, classifying them as the white-box approach. TRNSYS, eQUEST, DOE-2, DeST, and EnergyPlus are examples of energy simulation software. In contrast, AI-based methods are deemed a black-box approach, as they predict and assess the building energy loads without examining the internal logical

relationships. Combining the white-box and black-box approaches, the hybrid method eliminates the shortcomings of both methods, and is thus known as the grey-box approach.
Alwisy et al. (2018) ranked the buildings' elements for their environmental impact (e.g., land use, water user, and energy use) by extensively reviewing the literature. They proposed a work process for designers to meet the owner's energy efficiency goals, taking into consideration the rank of green building design factors. They determined that the building envelope consisting of exterior walls, windows, roof, and floors, has a greater environmental impact than the heating, ventilation, and air conditioning (HVAC) system. Numerous design factors impact the buildings' energy performance following a nonlinear relationship (Cebrat & Nowak 2018). An early assessment of these factors right from the beginning of the conceptual design phase is essential to improving the energy efficiency of buildings. Traditional engineering modeling techniques are limited and have proved to be cumbersome and tedious to implement, considering the non-linear effect of design features on the energy performance of detached residences (Cebrat & Nowak 2018; Wang & Srinivasan 2015). Also, an advanced level of expertise is required to adequately use energy simulation programs and calibrate the design based on the internal logical relationships (Wang & Srinivasan 2015).

Artificial intelligence (AI), and particularly machine learning (ML) approaches, have lately been replacing such modeling methods, successfully forecasting the energy loads on buildings without the need for proficiency in energy calculations (Seyedzadeh et al. 2018; Sun et al. 2020). This research aimed at modeling an Artificial Neural Network (ANN) based on a big dataset of design features of newly constructed single-detached dwellings in Florida. It will potentially help designers optimize the energy efficiency of houses, accounting for different design alternatives. The sections of this paper are arranged as follows. First, the most relevant literature about the applications of machine learning to predict the buildings' energy loads along with some of their limitations and advantages, is presented. Secondly, the development of the ANN and its assessment based on its ability to forecast the heating and cooling loads on Florida's detached residences are discussed in the methodology section. In the Results and Discussion section, the best possible ANN structure is specified. Lastly, conclusions and future research work are presented.

2 Literature Review

2.1 ANN inspiration, benefits, and drawbacks.

Artificial Neural Networks (ANNs) are nonlinear statistical learning techniques inspired by the brain's neurobiological structure. ANNs are similar to the biological brain, in that they consist of artificial neurons linked to each other forming a network. These links are made via activation functions such as sigmoid, linear, and hard-limit functions (Sun et al. 2020). The simplest ANN structure consists of three layers connected to each other: input, hidden, and output layers (Wang & Srinivasan 2015).

The earliest ANN models developed were Feed Forward Networks (FFNs), in which the information goes in one direction starting with the input layer all the way to the output layer and passing through the hidden ones. Recurrent Neural Networks (RNNs), Radial Basis Function Networks (RBFNs), and other more advanced ANN structures exist. RNNs use the internal memory to learn from prior experiences creating loops between output and input layers. Recursive, Long Short-Term Memory (LSTM), and fully connected are examples of the different structures and architectures of RNNs. RNNs are useful for solving very deep learning problems, where more than one thousand layers are required. In this respect, the ANN is considered a deep learning ANN. On the other hand, RBFNs consist of only one hidden layer and employ radial basis activation functions to find the output. According to Seyedzadeh et al. (2018), RBFNs are very useful for time series estimations. The ANN architecture should be selected based upon the problem at hand.

The process for the development of AI-based energy prediction models typically begins with the collection of the dataset consisting of inputs and outputs associated with each other. The

independent variables, which constitute the input layer of the ANN, can incorporate weather data, occupants' information, global heat loss coefficient, indoor environmental factors, time index, socioeconomic and historical data, as well as buildings' materials (Sun et al. 2020; Wang & Srinivasan 2015). The output layer comprising the dependent variables, includes information relating to building energy consumption such as chilled and hot water consumption, electricity expenditure, air conditioning loads, and gas consumption. After being collected, the raw data is arranged and formatted using data interpolation, data transformation, data normalization, and other data pre-processing methods to improve the quality of data. Subsequently, the training of the ANN starts using the training dataset to learn the bias, weights, the number of hidden layers, the number of neurons in each layer, as well as the appropriate activation functions needed to make good predictions. The prediction performance of the model is then tested utilizing the testing dataset which is a different independent dataset. The different tools and criteria used for testing the ANN model are discussed in a later section.

ANNs are suitable for solving complex problems involving nonlinear relationships between the inputs and the outputs of the model. They do not require an advanced level of expertise and proficiency (Sun et al. 2020; Wang & Srinivasan 2015). Wang & Srinivasan (2015) stated that accurate prediction results are only achieved when training the ANN model using appropriate hyperparameters. Moreover, ANNs provide designers with a time-efficient and cost-effective method to conveniently forecast the energy loads on buildings, even with a preliminary conceptual design of the physical building features which should be finalized and detailed for energy analysis and simulations performed using engineering modeling methods. Turhan et al. (2014) predicted the heating loads on buildings using both ANNs and the energy simulation software "KEP-IYTE-ESS" and concluded that energy simulation software are more sophisticated and less efficient than ANNs.

However, ANNs still suffer from a lot of drawbacks, such as the long run-time needed to train the data and to develop a prediction model with good quality and high accuracy (Sun et al. 2020). Also, ANNs encounter the curse of dimensionality problem, as the number of training data needed exponentially grows, by linearly increasing the number of independent variables included in the model. Furthermore, ANNs are almost impossible to be extended to include additional independent variables if the building's design or operation changes. This is mainly due to the lack of knowledge of the internal relationships existing between the input and the output layers (Wang & Srinivasan 2015). In this scenario, the ANN needs to be re-trained from scratch.

2.2 Applications of Machine Learning.

Khayatian et al. (2016) utilized ANNs to forecast the heat demand indicators based on the characteristics of residential buildings using the Italian CENED software and database as training data. Sholahudin et al. (2016) modeled an ANN to predict the air conditioning loads on buildings taking into account eight independent variables including building's orientation, wall area, relative compactness, floor area, roof area, and glazing area. They used energy simulation data for 768 residential buildings published in the literature to serve as the training dataset. Ascinoe et al. (2017) performed energy simulations on EnergyPlus software to gather data needed for the training and testing of the ANN model. The intended purpose of their ANN was to assess the thermal comfort of the buildings' tenants and the energy performance of these buildings. They utilized the MATLAB platform to develop the ANN model. Kerdan & Gálvez (2020) relied on a genetic algorithm that generates and assesses thousands of ANN architectures based on their performance in predicting the building exergy destructions, thermal comfort of occupants, and life cycle costs. They simulated 3,000 building energy models located in the various climatic regions of Mexico utilizing ExRET-Opt which is an EnergyPlus-based tool. The independent variables included in the energy models along with the results of the energy simulations served as the training and testing datasets needed for model development. Cebrat & Nowak (2018) studied the existing relationships between the design parameters of buildings utilizing self-organizing maps which are a type of ANNs which employ an unsupervised learning algorithm. They randomly generated around 5,000 detached houses to train the model. Lee et al. (2019) developed an artificial neural network using around 5,000 single-person-dwellings in addition to

their EnergyPlus simulations and their corresponding occupant's hourly schedule. They revealed the existence of a correlation between energy use and occupants' characteristics. Moradzadeh et al. (2020) used Support Vector Regression (SVR) and Multilayer Perceptron (MLP) to predict the cooling and heating loads on residential buildings including eight input variables in the analysis. Around 770 buildings assumed to be in Greece, Athens were simulated using Autodesk Ecotect Analysis software and were utilized to train, validate, and test the model. Both SVR and MLP techniques were able to successfully predict the air conditioning loads with correlation coefficients of 0.9878 and 0.993, respectively (Moradzadeh et al. 2020).

Very few researchers have developed ANN energy consumption prediction models based on large real datasets. Most of them either utilized energy analysis software programs to simulate building energy models, or used data from the literature, or randomly generated datasets (Cebrat & Nowak 2018) for training, testing, and validating purposes. Additionally, they mainly used a dataset consisting of at most 5,000 samples, for the development of ML models capable of predicting the air conditioning loads on residential buildings. Moreover, no research was proposed to predict the heating and cooling loads on detached residences in Florida using AI methods. This study fills these gaps in research by modeling an artificial neural network based upon a large dataset of more than 18 thousand newly constructed detached dwellings in Florida between the years 2009 and 2019. Furthermore, this dataset contained actual data collected from local building permit departments in counties all over the state of Florida and used no randomly generated data.

2.3 Tools to test the prediction accuracy of the ANN model.

The model training's prediction accuracy should be compared with the model testing's prediction accuracy, to avoid overfitting and underfitting. Overfitting occurs when the validating dataset's accuracy is much lesser than the training dataset's accuracy, indicating that the model not only fits the training dataset but also fits the noise and outlier. Underfitting can be detected when erroneous training and testing results are attained suggesting that the established ANN model cannot successfully forecast the energy loads on buildings. Hence, the ANN model cannot generalize the relationship between the independent and the dependent variables, when either underfitting or overfitting occurs (Sun et al. 2020). The accuracy of the ANN in predicting the energy performance of buildings can be measured using various statistical parameters including Coefficient of Variance (CV), R square (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Bias Error (MBE), Root Mean Squared Error (RMSE), Mean Squared Percentage error (MSPE), and Normalized MBE (NMBE). Seyedzadeh et al. (2018) and Sun et al. (2020) described each of these fitness metrics in their research. In this study, the performance of the ANN model was assessed using MSE which can evaluate both bias and variance of the projected energy loads to the actual loads. R square was employed to validate the ANN model.

3 Methodology

Six steps were employed to make sure of the usefulness and validity of the developed ANN model (Figure 1).

Step 1, *Strategizing*, involved defining the study's goals. The architecture of the ANN model including the number of hidden layers and the number of neurons per layer, was specified. The purposes of the ANN model included examining the existing relationships between the design parameters of single-family houses, forecasting the energy loads taking into consideration the design elements and features, and enhancing the energy performance of residences. Designers can use the ANN model to predict the heating and cooling loads on different design options and make proper decisions right from the start of the design phase avoiding extra costs and charges that may arise as a result of changes at an advanced stage. The input and output spaces, along with the respective shape and scope of their problem and solution domains were studied. The input space of the ANN model was comprised of 10 independent variables which were the main data provided on the energy forms collected. The independent variables included Glass to Floor

Area, Windows' Area Weighted Average U-value, Windows' Area Weighted Average Solar Heat Gain Coefficient (SHGC), Conditioned Floor Area, Total Area of Walls, Walls' Area Weighted Average R-value, Roof's R-value, Roof's Area, and Ducts' R-value. In addition, the Florida Climate Zone was included in the inputs of the ANN, as Florida mainly has two climate zones, the humid tropical and the subtropical sub-humid mesothermal climate zones (Elias & Issa 2019) and the climate zone has a strong influence on the building's energy loads (de Rubeis et al. 2020). The ANN's outputs were heating and cooling loads. Supervised Learning was the training algorithm adopted.

Step 2, *Data Collection and Evaluation,* was the second step in the process. Data were collected from energy forms prepared between the years 2009 and 2019 as part of the permit application process at local building permit departments in various counties all over Florida. To ensure the quality of data, data were first evaluated and cleaned. Outlier examples suggesting incorrect data were excluded from the analysis, bringing about 18,317 usable data. Examples of faulty data included an SHGC value for windows exceeding 1.0, a U value for windows greater than 1.2, and a glass-to-floor area ratio larger than 1.0. Consequently, the energy efficiency of houses was explored utilizing a data set of 18,317 newly constructed detached dwellings in Florida. The dataset was divided into three sets. The first one was the training dataset which consisted of 80% of the data available. Ten percent of the data was used to decide when to stop training and to pick the optimal model (Validating dataset). The rest of the data (10%) served as the Querying dataset used to make a final assessment of the chosen model. Data were also evaluated to make sure that they represented the problem at hand, signaling that the training samples were well distributed across the problem domain, as they somehow created an envelope for all other testing and querying samples.

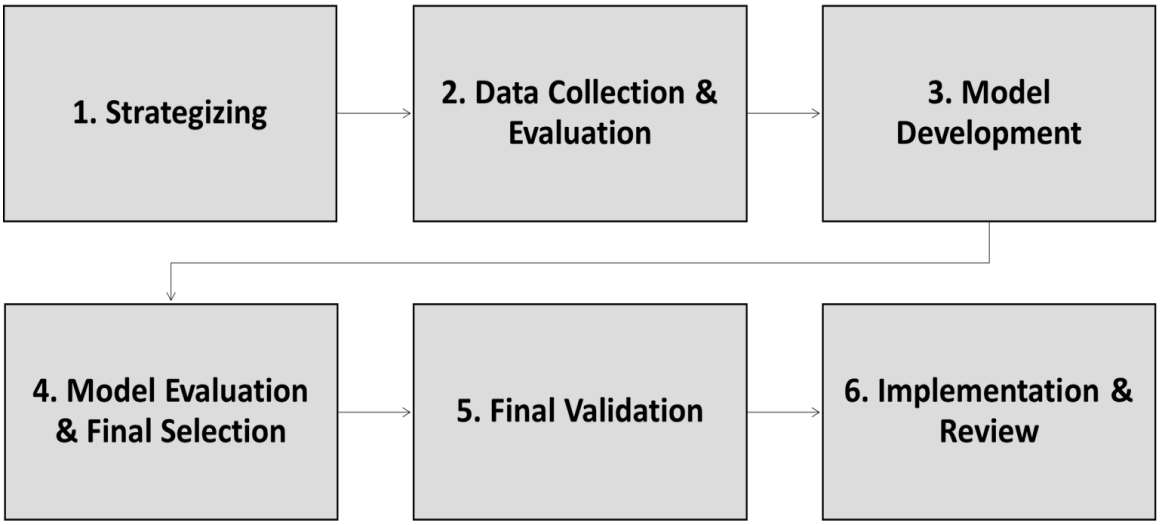

Figure 1: Methodology of Work

Step 3, Model Development, consisted of developing the ANN structure, writing and running a programming code utilizing the TensorFlow platform for machine learning and Python-based Keras libraries. Data were first imported. Then, in the second portion of the code, "climate zone" which is a categorical feature, was encoded employing the scikit-learn preprocessing package. Dummy variables were created to prevent prioritization between climate zones. Additionally, feature scaling was performed on datasets independently to prevent "data leakage" which is a machine learning problem that commonly happens when establishing predictive models (Singh et al. 2020). The ANN was subsequently developed using the training dataset, progressively altering the architecture of the ANN, including the number of hidden layers and the number of hidden neurons in a layer, and examining the effect of these modifications on the ANN performance, looking for the best design.

Step 4, *Model Evaluation and Final Selection*, is aimed at evaluating the accuracy of the model in relation to the system being represented using the Validating dataset. The Validating dataset helped in monitoring and terminating training and in comparing the performance of all design alternatives. The performance was assessed based upon the overall objective function (mean squared error - Validating dataset) and the consistency in performance across the problem domain. The ANN structure was chosen accordingly and is described in the next section.

Step 5, Final Validation, makes the final validation and assesses the performance of the chosen ANN model using a third and independent dataset (querying dataset). This Querying dataset helps in avoiding biased results. This step is aimed, therefore, at accurately evaluating the model's performance and at checking whether further development may be necessary.

Step 6, Implementation and Review, is the last step and has not been completed at the time of this submission. It consists of implementing the ANN-based model to solve real-life problems and gathering feedback to continue its improvement.

4 Results and discussion

The findings of Model Development, Model Evaluation and Final Selection, in addition to Final Validation (Steps 3, 4, and 5, respectively) are discussed in this section. The model development/evaluation had an iterative aspect, as multiple ANN architectures and various hyperparameters were employed and assessed. The best ANN model determined at the time of this publication was comprised of two hidden layers of 13 and 19 neurons, respectively. The model was manually tuned to determine the optimal number of epochs which turned out to be 100 at which the training stopped. "Adam" optimizer was used to develop the ANN model. "Adam" optimizer is capable of handling sparse gradients on noisy problems and optimizes the properties of "AdaGrad" and "RMSProp" algorithms. As shown in Figure 2, the performances of the training and validating datasets were assessed based on the mean squared error for both the training and validating datasets (the overall objective function). The best number of epochs was determined to be 100, at which the validating mean squared error achieved a value of 0.32.

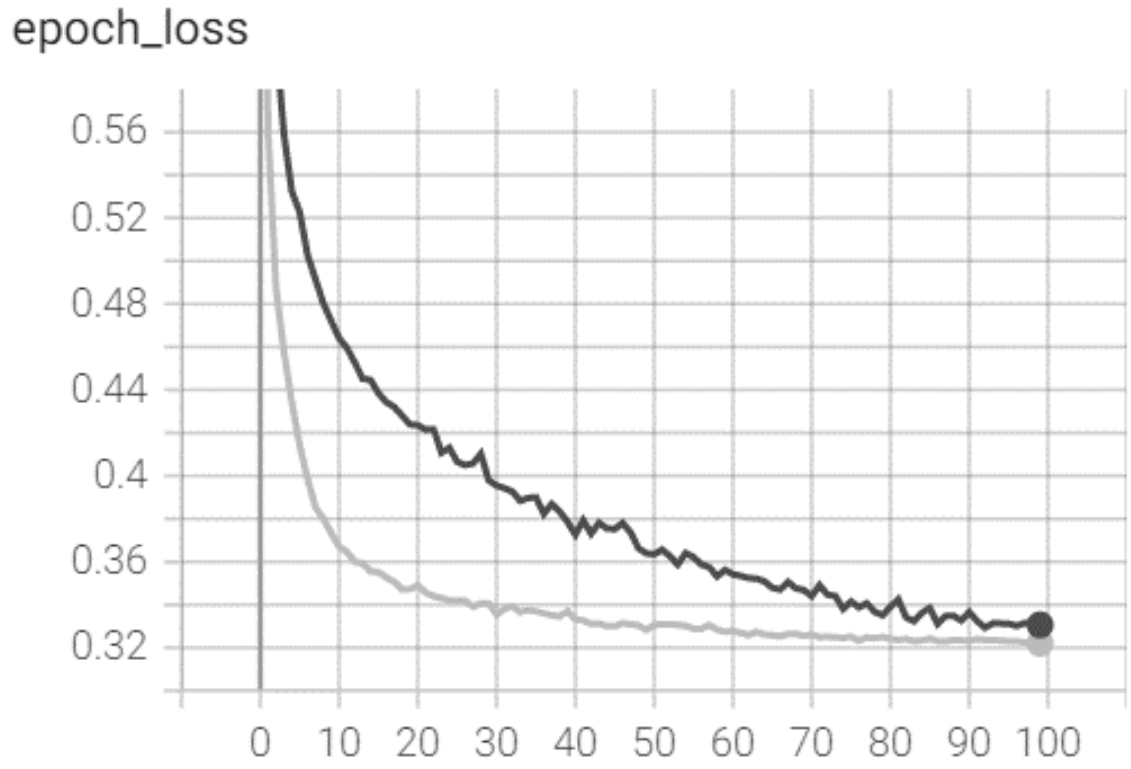

Figure 2: Mean squared error for both training (in black) and validating (in grey) datasets vs. epochs.

The selected ANN model took around two minutes of training time. Moreover, multiple dropout rates were considered to prevent overfitting. Dropout rates relate to the probability of ignoring neurons during the training stage. After many iterations, a dropout rate of zero was chosen. Furthermore, the hyperbolic tangent activation function (tanh) was utilized to provide the non-linear sophisticated functional relationship that exists between inputs and outputs. The best batch size was specified to be 10 causing a low error. During the Final Validation step (step 5), the Querying dataset was utilized to measure the performance of the selected model. This was performed by forecasting the cooling and heating loads and comparing them with the actual values utilizing the R square (R^2) metric. A value of 70.3% was reached for R^2, implying that further development is needed until attaining an R^2 value of at least 95%.

5 Conclusions and Future Research

This research developed an ANN model to predict the heating and cooling loads on a detached house within a computer-based simulation environment based upon a large dataset exceeding 18,000 newly constructed detached houses in Florida between the years 2009 and 2019. Ten independent variables were included in the input layer of the ANN. The energy load projections performed utilizing the ANN model showed that further development is required to accomplish better and more accurate results. This study will assist designers in making effective decisions right from the start of the conceptual design stage of residential projects. Additionally, this research will pave the road for generative design considering the independent variables incorporated in the ANN model, since machine learning models can be applied to investigate and generate various design options to optimize the energy performance of buildings. However, this study has limitations. Although several design factors influence residential buildings' energy performance, the developed ANN only comprised 10 input variables excluding the occupancy conditions and factors that greatly affect the energy use of buildings. Furthermore, this study was limited to training the ANN using newly constructed residences located in the state of Florida which is mainly characterized by humid tropical and subtropical sub-humid mesothermal climate zones. Future studies must incorporate the occupants' behavior and factors to assess their impact on the prediction performance of the ANN. Furthermore, further studies are needed to develop ANN models able to predict the energy loads on residential buildings not only in Florida, but in all the United States. Also, future research work will consist of pruning the ANN in terms of its inputs to scrutinize its sensitivity to each of the input variables and determine whether some of them can be omitted from the ANN.

References

Ascione, F., Bianco, N., De Stasio, C., Mauro, G., & Vanoli, G. (2017). "Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach". Energy, 118, 999-1017.

Alwisy, A., BuHamdan, S., & Gül, M. (2018). Criteria-based ranking of green building design factors according to leading rating systems. Energy and Buildings, 178, pp. 347-359.

Cebrat, K., & Nowak, Ł. (2018). Revealing the relationships between the energy parameters of single-family buildings with the use of Self-Organizing Maps. Energy and Buildings, 178, pp. 61–70.

de Rubeis, T., Falasca, S., Curci, G., Paoletti, D., and Ambrosini, D. (2020). "Sensitivity of heating performance of an energy self-sufficient building to climate zone, climate change and HVAC system solutions". Sustainable Cities and Society, 61, 102300.

Elias, R., & R. A. Issa, R. (2019). "Big Data: A Decade of Energy Characteristics of Single-Family Homes in Florida". In Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization (UrbSys'19). Association for Computing Machinery, New York, NY, USA, 101–111. DOI: https://doi.org/10.1145/3363459.3363533

García Kerdan, I., & Morillón Gálvez, D. (2020). "Artificial neural network structure optimisation for accurately prediction of exergy, comfort and life cycle cost performance of a low energy building". Applied Energy, 280, 2-19.

Khayatian, F., Sarto, L., & Dall'O', G. (2016). "Application of neural networks for evaluating energy performance certificates of residential buildings". Energy and Buildings, 125, 45-54.

Lee, S., Jung, S., & Lee, J. (2019). "Prediction Model Based on an Artificial Neural Network for User-Based Building Energy Consumption in South Korea". Energies, 12(4), 608.

Moradzadeh, A., Mansour-Saatloo, A., Mohammadi-Ivatloo, B., & Anvari-Moghaddam, A. (2020). "Performance Evaluation of Two Machine Learning Techniques in Heating and Cooling Loads Forecasting of Residential Buildings". Applied Sciences, 10(11), 3829.

Seyedzadeh, S., Rahimian, F., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering, 6(1).

Sholahudin, Alam, A. G., Baek, C. I., & Han, H. (2016). "Prediction and Analysis of Building Energy Efficiency Using Artificial Neural Network and Design of Experiments". Applied Mechanics and Materials, 819, 541–545.

Singh, A., Gupta, I., Verma, R., Gautam, V., & Yadav, C. (2020). "A Survey on Data Leakage Detection and Prevention". SSRN Electronic Journal.

Sun, Y., Haghighat, F., & Fung, B. (2020). "A review of the-state-of-the-art in data-driven approaches for building energy prediction". Energy and Buildings, 221, 110022.

Turhan, C., Kazanasmaz, T., Uygun, I., Ekmen, K., & Akkurt, G. (2014). "Comparative study of a building energy performance software (KEP-IYTE-ESS) and ANN-based building heat load estimation". Energy and Buildings, 85, 115-125.

U.S. Energy Information Administration. (2019). International Energy Outlook 2019. U.S. Energy Information Administration, Washington, DC, 52.

Wang, Z., & Srinivasan, R. (2015). "A Review of Artificial Intelligence Based Building Energy Prediction with a Focus on Ensemble Prediction Models". Winter Simulation Conference, IEEE 2015, Piscataway, NJ, 3438-3445.