Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Determine the heat demand of existing buildings with machine learning

To cite this article: Joachim Werner Hofmann *et al* 2023 *J. Phys.: Conf. Ser.* **2600** 032013

Determine the heat demand of existing buildings with machine learning

Joachim Werner Hofmann, Christian Amoser, Achim Geissler, Monika Hall

Institute for Sustainability and Energy in Construction (INEB) University of Applied Science and Arts Northwestern Switzerland, Hofackerstrasse 30, CH 4132 Muttenz

E-mail: joachim.hofmann@fhnw.ch

Abstract. The renovation rate of existing buildings plays a major role in the Swiss Energy Strategy 2050+. To increase this rate, there must be a simple and cost-effective method to determine the heat demand of existing buildings. In this paper, the generation of such a method, based on the Swiss cantonal building energy certificate (GEAK) database with the help of machine learning (ML), is studied. The aim of the project was to develop a ML model which allows the heat demand of existing buildings to be determined quickly with a minimal set of parameters. The comparison of the GEAK building envelope class for single family houses calculated with the new ML model and the original GEAK classes shows that approximately 62 % have the same class, 32 % differ by one class and 6 % by two classes. The ML model is a good starting point for further refinements and developments.

1. Introduction

The Swiss Federal Office of Energy (SFOE) is aligning its vision with the net-zero basic variant of the energy perspectives 2050+ [1]. The vision of the SFOE for the Swiss building stock can be explained with the term ROSES. ROSES stands for Reduction, Optimization, Substitution, Renewable Energy and Sustainability. The project is based on the topics of reduction, optimization and substitution. To significantly increase the renovation rate in the coming years, easy-to-use planning methods must be developed to determine the heat demand of buildings. In these calculations, the determination of the U-values and the areas are work intensive. For this reason, the aim of the project was to develop a ML method to quickly determine the heat demand of existing buildings. The potential for using this ML method is diverse:

- Reduction of the effort necessary to calculate the heat demand for existing buildings as a basis for planning energy renewal measures. Fast and yet sufficiently accurate recording of the status, even if the planning basis is missing or incomplete.
- Simple and yet sufficiently precise determination of the heat demand as a basis for the design of heating systems in existing buildings, especially for heat pump systems.
- Determination of the heat demand of districts as a basis for the design of local heating networks or energy policy decisions.
- The ML method also has the potential for the simplified creation of energy or CO_2 labels.
- There is still a lack of experts in the energy sector. Efforts to increase efficiency in their work can help to increase both the renewal rate and the assessment quality of existing buildings.

2. Methodology

2.1. Overview

Methods using machine learning are suitable for solving complex technical problems. The methodology of the procedure (process steps) is shown in Figure 1.

Figure 1. Workflow for machine learning [2,3].

The workflow is divided into two parts. The first part (green) is the statistical and building physical approach with a strong emphasis on feature selection. The second part (blue) is the development of the desired machine learning model.

2.2. Statistical and building physical approach.

2.2.1. Requirements. The GEAK [4] is used to determine, among other things, the heat demand of a building. The calculation is based on SIA 380/1:2016 [5]. This calculation is complex and therefore a simplified model with fewer parameters was developed in this project. Data from the GEAK database was used for this. The GEAK database contains different building types but for this study only residential and office buildings were used.

2.2.2. Data Collection and Data Cleaning of GEAK database. Each GEAK contains building data such as a building ID, year of construction, energy reference area, U-values, building orientation, number of storeys, building type and more. The data provided by the GEAK association were available in the form of various Excel spreadsheets. A uniform data set was created from the tables using the GEAK numbers. With this step, the data set was massive reduced. Many GEAK numbers only appeared on one of the Excel spreadsheets, so information was missing and therefore had to be discarded. After these, descriptive statistics were used, to examine the remaining parameters for mean, standard deviation, distribution of the values and their ranges. In the end 28,000 of over 100,000 GEAKs were left out. The entry for each building contains 32 parameters, which were classified to be potentially important for the model development.

2.2.3. Data Analysis. The aim of the project was to find a mathematical model with a minimal parameter set with which the heating demand calculation in the GEAK – Tools could be executed. Using the parameters described in SIA 380/1:2016 [5], as described in the requirements, 11 physically sensible parameters were found. Spearman's is a rank correlation which is used to determine the correlation between the 11 parameters and the heat demand. The Spearman's rank correlation is monotonic/non-linear [6,7,8], so we do not filter out the non-linear relationships. This is useful when using DNN (Deep Neural Networks) in the model generation process. A DNN is suitable for dealing with non-linear relationships in the data. The output of these analyses is shown in Figure 2, where for example, the U-values have the strongest correlation to heat demand of the 11 parameters in all categories.

2.2.4. Feature selection. To clarify which of the 11 parameters are useful for the ML algorithm due to their correlation with the heat demand, they are discussed below.

- U-values: the U-values have a strong correlation with the heat demand. Therefore, U-values are a suitable feature for the ML algorithm. The determination of U-values is often time consuming and difficult, however.
- Energy reference area (ERA): The ERA shows a medium correlation for all building types. Therefore, it is also a suitable feature for the ML algorithm. The ERA can usually be extracted from building plans.

- Year of construction: The year of construction has a medium to strong correlation with the heat demand. This mirrors the correlation with the U-values. This can be explained by the fact that newer buildings are better insulated and accordingly have better U-values. It should be noted, however, that many buildings have been renovated since they were built, so the year of construction can lead to incorrect conclusions about the U-values. It would be preferable to know the renovation year of older buildings (this information is currently not available in the GEAK database). Nevertheless, the construction/renovation year is a suitable feature for the ML algorithm. The year of construction or the year of renovation is easy to determine.

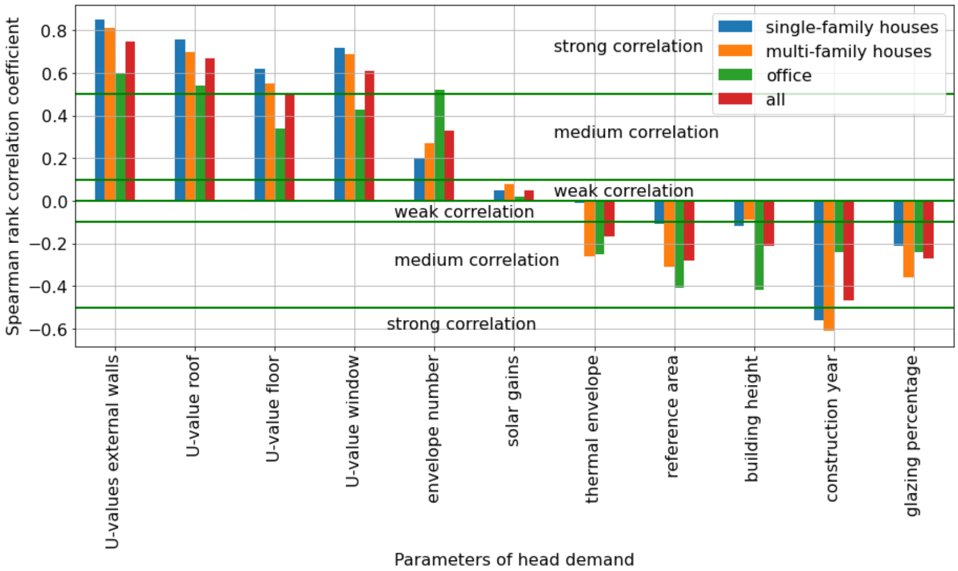

Figure 2. Spearman's correlation between selected parameters in relation to heat demand (±1: maximal correlation, 0: no correlation, positive/negative values: generates low/high heat demand, all: residential and office buildings).

- Solar heat gains: The solar heat gains have only a weak correlation with the heat demand. Therefore, solar heat gains are not a suitable feature for the ML algorithm.
- Other parameters: Other parameters that describe the building geometry, such as the thermal building envelope, the building height, and the proportion of glazing show mainly medium correlations with the heat demand. As the parameters are not given in a useful form for the ML algorithm, these parameters were not used. The building envelope was estimated via ERA and the number of stories. Building height and the proportion of glazing were not used.

For the calculation of the heat demand, the outside temperature is needed. The annual mean outside temperature is derived from SIA 2028 [9] for the building location. All nine features chosen for the ML algorithm are given in table 1.

Table 1. Features chosen for the ML algorithm.

Name	Unit
Energy reference area	m^2
Construction year	-
Number of storeys	-
Building orientation	-
Mean annual outside temperature	°C
U-values: external wall, floor, roof, window	$W/(m^2\,K)$

2.3. Development of a machine learning model to determine the heat demand.
For the development of the ML algorithm, Tensorflow [10] is used. Tensorflow is an open-source machine learning and artificial intelligence library for high performance numerical computation [11]. For the project, one of the official Tensorflow examples was adapted.

2.3.1. Model Training. The available 28,000 data sets were divided into three separate data sets for training, validation and testing. 70 % were used for the training, 10 % for validation and 20 % for testing. The training data set was used to train and create the base model. The validation data set was used to fine-tune the model's hyperparameters to refine the model. Finally, the test data set was used to evaluate the refined model. The adjustment was done with the mean absolute error (MAE) [11] as in equation (1) where y_i is the prediction, x_i the true value and n number of sample size.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \quad (1)$$

Hyperparameters control the algorithm creations for ML model building. The algorithm used in this project has three hyperparameters: the number of epochs (iteration), the number of neurons (single calculation nodes) in the hidden layers (calculation core) and the learning rate (adjustment limit between two iterations). The learning curve shows the progress of learning over the epochs [12]. In this project, this is shown by the decreasing MAE over the number of epochs.

Data standardization [11] plays an important role in the machine learning process. The data are standardized to ensure that the size of the values and their scatter do not falsify the model and that each feature has an equally large influence. This is done with z-standardization.

2.3.2. Selection of ML algorithm. Four different algorithms (single feature linear regression, multiple feature linear regression, single feature DNN and multiple feature DNN were used. The accuracy of the generated models is expressed by MAE. Table 2 shows that single feature algorithms perform poorly compared to the multiple features algorithms. This is to be expected, since calculating the heat demand is a complex task that depends on many variables. The DNN performs better than the linear regression. Therefore, the multiple-feature DNN was chosen for the ML model development.

Table 2. Comparison of the MAE for single-family houses and multi-family houses calculated with different algorithms.

name	single-family houses	multi-family houses	unit
single-feature linear regression	30.1	26.4	kWh/(m^2 a)
multiple-feature linear regression	19.9	17.8	kWh/(m^2 a)
single-feature DNN	29.1	26.7	kWh/(m^2 a)
multiple-feature DNN	17.8	14.1	kWh/(m^2 a)

The DNN algorithm creates a model consisting of four layers. The input layer consists of the standardized data of the nine defined features (Table 1). The model also consists of two hidden layers, each defined here with 16 neurons. Finally, there is the output layer, which provides the predicted heat demand. The final set of hyperparameters is shown in Table 3 and Figure 3. The starting ML model for single-family houses achieved a mean absolute error of 17.8 kWh/(m^2 a) and for multi-family houses 14.1 kWh/(m^2 a). This corresponds to the relative average deviation of around 15.5 % for both building types [10].

Table 3. Hyperparameters of the initial model.

Hyperparameter	Selected Values
number of input layers	1
Number of hidden layers	2
Number of output layers	1
Neurons per hidden layer	16
Number of trainable parameters	449
learning rate	0.001
number of epochs	200

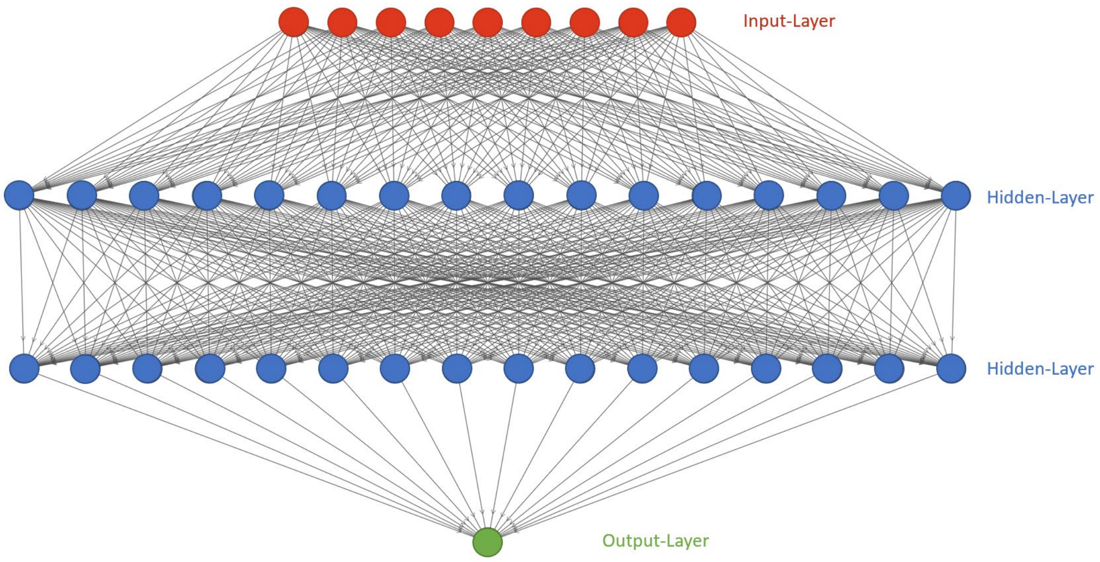

Figure 3. Visualization of the initial model: input layer with nine input features, two hidden layers with 16 neurons and output layer with one neuron.

3. Results

The aim of the project was to predict the energy demand of a building with a minimal set of features. To generate the machine learning model, regression was used. With the predicted energy demand and the standard calculation [13,5], the building classification was achieved. This result was then compared with the result derived from the GEAK tool. The comparison was done with 1816 single-family house test data sets and 974 multi-family house test data sets. Table 4 summarises the differences found between the GEAK class predicted by the ML model and the class predicted by the GEAK tool [2].

Table 4. Comparison of ML model and GEAK tool for single-family houses and multi-family houses.

	single-family houses		multi-family houses	
GEAK class	numbers	%	numbers	%
Equal to GEAK tool	1134	62.4	639	65.6
1 class next to it	567	31.5	301	30.9
2 class next to it	108	6.1	34	3.5

In the case of single-family houses, around 62 % of all buildings achieved the same GEAK class as the class assigned by the GEAK tool. Around 31 % of the buildings missed the GEAK class by one class and approximately 6 % missed the GEAK class by two classes. Similar results were also found for the multi-family houses.

4. Conclusions

It was shown that it is possible to predict the heat demand with a fair degree of accuracy using a model generated by machine learning. Only nine features were used. The mean absolute error was 17.8 kWh/(m^2 a) for single-family houses and 14.1 kWh/(m^2 a) for multi-family houses which, in relative terms, means an average deviation of around 15.5 %. The GEAK class could be correctly predicted in almost two thirds of all cases. These results of the first version of an ML model are promising, but further work is necessary:

- The feature selection should be refined.
- The hyperparameters (including hidden layers) of the ML model can likely be further refined.
- Additional algorithms can be tested and compared.

Of course, an integration of such an ML model into the GEAK tool is to be discussed.

Acknowledgments

We would thank the “Verein GEAK” which has provided us with the large number of input data. Also, we would thank our undergraduate students: Aline Augustin and Daniel Hügli, for carrying out a huge part of the data analysis and ML simulations.

References

[1] Bundesamt Für Energie «Energieperspektiven 2050+», https://www.bfe.admin.ch/bfe/de/home/politik/energieperspektiven-2050-plus.html.
[2] J.W. Hofmann et al, «Use of machine-learning for monitoring solar thermal plants”, 2021 J. Phys.: Conf. Ser. 2042 012007
[3] S. Amershi et al, «Software Engineering for Machine Learning: A case Study”, ICSE-SEIP '19: Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice May 2019 Pages 291–300https://doi.org/10.1109/ICSE-SEIP.2019.00042
[4] Verein GEAK https://www.geak.ch/verein/der-verein/
[5] SIA 380.1:2016 «Thermische Energie im Hochbau»
[6] A. Augustin, «Datenaufbereitung für die Anwendung neuronaler Netze zur Bestimmung des Heizwärmebedarfs», EUT-P6-22FS FHNW, Bachelortheis
[7] Spearman’s Rangkorrelationskoeffizienten bestimmen und interpretieren Scribbr, Jul. 07, 2020. https://www.scribbr.de/statistik/rangkorrelationskoeffizient/
[8] Spearman's rank correlation coefficient, Jul. 07, 2023, https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
[9] SIA 2028:2010 «Klimadaten für Bauphysik, Energie- und Gebäudetechnik»
[10] D. Hügli, «Anwenden von „Machine-Learning“ zur Bestimmung des Heizwärmebedarfs eines Gebäudes», EUT-P6-22FS FHNW, Bachelortheis
[11] https://www.tensorflow.org/tutorials/keras/regression
[12] A. Geron, «Hands-on Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to build Intelligent Systems, O’REILY, 2019
[13] https://www.geak.ch/media/rnorm_normierung_230330_geak.pdf