Sustainable Cities and Society 99 (2023) 104892

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

A hybrid machine learning approach for the load prediction in the sustainable transition of district heating networks

Mustapha Habib [a], Thomas Ohlson Timoudas [b], Yiyu Ding [c,*], Natasa Nord [c], Shuqin Chen [d], Qian Wang [a,e]

[a] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm 10044, Sweden
[b] RISE Research Institutes of Sweden, Division Digital Systems, Computer Science, Isafjordsgatan 22, Kista 164 40, Sweden
[c] Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim 7491, Norway
[d] College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China
[e] Uponor AB, Hackstavägen 1, Västerås 721 32, Sweden

ARTICLE INFO

Keywords:
District heating
Time-series clustering
Heat load prediction
Artificial neural networks
K-means

ABSTRACT

Current district heating networks are undergoing a sustainable transition towards the 4th and 5th generation of district heating networks, characterized by the integration of different types of renewable energy sources (RES) and low operational temperatures, i.e., 55 °C or lower. Due to the lower temperature difference between supply and return, it is necessary to develop novel methods to understand the loads accurately and provide operation scenarios to anticipate demand peaks and increase flexibility in the energy network, both for long- and short-term horizons. In this study, a hybrid machine-learning (ML) method is developed, combining a clustering pre-processing step with a multi-input artificial neural network (ANN) model to predict heat loads in buildings cluster-wise. Specifically, the impact of time-series data clustering, as a pre-processing step, on the performance of ML models was investigated. It was found that data clustering contributes effectively to the reduction of data training costs by limiting the training processes to representative clusters only instead of all datasets. Additionally, low-quality data, including outliers and large measurement gaps, are excluded from the training to enhance the overall prediction performance of the models.

1. Introduction

1.1. Background

Energy usage in buildings accounts for up to 40% of the total energy usage in the European Union (EU) (Pérez-Lombard, Ortiz and Pout, 2008). With this in mind, increasing the energy efficiency of buildings is one of the key objectives of the EU strategy for the decarbonization of the economy (Directive (EU), Directive). Current district heating (DH) networks are responsible for covering around 13% of the total thermal energy demand in the EU (Werner, 2017). The evolution of DH networks over the years has reduced the supply temperatures with the progressive implementation of the so-called 4th and 5th generation district heating (4GDH, 5GDH) (Lund et al., 2014, Li and Nord, 2018), which supply heat at temperatures around 45°C–55°C and below. This has enabled an increased integration of low-grade energy sources such as distributed renewable energy systems (RES) (Lumbreras and Garay, 2020) or waste heat streams (Wahlroos, Pärssinen, Manner and Syri, 2017, Ziemele et al., 2018) in the heating network.

Although the transition toward 5GDH has not yet been completed and considerable developments are still ongoing in this direction, However, there is a strong need to understand the energy demand trends on the end-user side and the interaction with the energy network, along with eventual connected RESs. This requires the introduction of accurate operating strategies to adapt local heat production and demand in the network. To do so, accurate understandings and characterization methods for heat loads are the first essential step, so the available local RES sources can be correctly operated concerning external variables such as weather and pricing models (Fitó et al., 2020). However, this task requires deep knowledge and understanding of the energy flow and heating patterns in buildings. Therefore, considerable efforts have been made nowadays for building energy assessments using metered data. Subsequently, the collected data can be potentially used for deep data analysis and optimization, and feed data-driven models for different load predictions (Frei et al., 2021, Ahmad and Chen, 2019).

* Corresponding author.
E-mail address: yiyu.ding@ntnu.no (Y. Ding).

https://doi.org/10.1016/j.scs.2023.104892
Received 27 February 2023; Received in revised form 17 August 2023; Accepted 21 August 2023
Available online 22 August 2023

Nomenclature

4GDH	4th Generation District Heating
5GDH	5th Generation District Heating
AI	Artificial intelligence
ANN	Artificial neural networks
DH	District heating
DHW	Domestic hot water
DSM	Demand side management
DTW	Dynamic Time Warping
ENOVA	Norwegian Energy Efficiency Agency
EU	European Union
HP	Heat pump
IoT	Internet of Things
LBFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine learning
MLP	Multi-layer perceptron
MSE	Mean Squared Error
NARX	Nonlinear autoregressive neural network with external input
RES	Renewable energy source
SH	Space heating

Additionally, given the integration of HP potentials loads become more sensitive as they are also influenced by other variables and disturbances that have not been conventionally considered, e.g., building heating system operating modes, building time constant, storage effects, and occupant schedules (Calikus et al., 2019). Moreover, unlike space heating (SH) loads, loads of domestic hot water (DHW) are commonly case/building type-specific, with different performance patterns compared to SH loads. Therefore, how to develop a holistic machine learning (ML) framework that can fill the above gaps is not an easy task (Ding, Brattebø and Nord, 2021). In this study, we focus on filling this knowledge gap by developing hybrid load prediction methods that can be essential for any optimal energy operation strategy, particularly, for RES-integrated district heating systems with such as heat pumps (HPs).

1.2. Previous studies

In contrast to electricity load analysis, prediction methods applied to heat loads are relatively complex, and this research field is yet to be consolidated (Lumbreras et al., 2022). In this regard, different approaches and frameworks have been investigated, such as white-box models, which are purely based on prior knowledge of building physics (Klein et al., 2017). However, the model development and calibration process via metered data is time- and resource-intensive and hard to replicate at community levels (Lumbreras et al., 2022). Another alternative is provided by data-driven models that are partially or fully based on energy meter data (Ding et al., 2022). A wide variety of data-driven models exist, ranging from black-box models, in which no prior knowledge of the building's physics is assumed, up to gray-box models formulated using differential equations calibrated with metered data. Concerning monthly and yearly load forecasts for energy planning and operation, the energy signature method is a broadly utilized data-driven methodology that expresses heating energy use as a function of weather variables (Eriksson, Akander and Moshfegh, 2020). In this context, authors in (Sha et al., 2019, Nielsen and Madsen, 2006, Wang, Lu and Li, 2019) studied the dependency between weather variables (e.g., outdoor temperature, wind speed, air humidity, solar irradiation) and the heating demand in buildings. The targeted prediction horizons were monthly, weekly, and daily. Here, it has been proven that outdoor temperature was considered to be the most dominant factor in the studied cases (Timoudasa). Some results showed that multiple regression models are applicable to short-term heat load prediction, however, these approaches need to be improved more by considering external factors, such as occupancy and indoor conditions. Furthermore, energy signature approaches are commonly valid for low-resolution predictions, such as weekly or monthly accumulated energies. However, operation strategies require high-resolution predictions (from daily to hourly), which leads to the obligation of developing accurate models. For this purpose, different ML approaches have shown strengths (Dagdougui, Bagheri, Le and Dessaint, 2019, Sandberg, Wallin, Li and Azaza, 2019). As an ML method, artificial neural networks (ANN), have recently been applied for the heat load prediction problem in buildings. In (Dagdougui, Bagheri, Le and Dessaint, 2019), a study of several ANN architectures is presented to predict heat loads in a residential building in the very short-term (hourly) and short-term (daily) horizons. The obtained results vary from a mean absolute percentage error (MAPE) of 3% to 4%. In the same context, authors in (Sandberg, Wallin, Li and Azaza, 2019) presented a nonlinear autoregressive neural network with external input (NARX) to perform heat demand forecasts. Thirteen input variables, including weather, energy, and social behavior parameters, have been considered to predict the hourly heat demand of a commercial building. In this case, a prediction error of 3.2 % is obtained. The result revealed that these input parameters can predict the building heat demand with up to 96 % accuracy on an hourly basis for a whole year. Nevertheless, as a drawback, existing ML approaches need large amounts of high-resolution data, which is not always available from the current DH substation measuring systems. This fact imposes additional duties on engineers and data scientists by requiring extensive cleaning and pre-processing procedures, which leads to considerable data analysis costs and time.

Regardless of the effectiveness of ML and data-driven approaches in regression problems, as stated earlier, such techniques are highly dependent on data resolution, quality, and availability. Data quality has been a major concern, leading to the introduction of data standards and quality frameworks (Biessmann et al., 2021, Breck et al., 2019, Gupta et al., 2021, Schelter et al., 2018, Gudivada, Apon and Ding, 2017). Recent advances in artificial intelligence (AI) have brought data quality back into the spotlight, and researchers have pointed out the impact of data processing methods on the performance of AI/ML models (Gudivada, Apon and Ding, 2017). In (Budach), a massive analysis, sustained by empirical results, is carried out to evaluate the effect of six data quality dimensions. It has been found that, in cases of regression problems, the presence of missing values or inaccurate features in the test data leads to undesired model performance. For building energy management systems, collected data, either from a legacy system or from the internet of things (IoT)-enabled sensors, is not an exception. In this particular case, additional uncertain factors, mainly human and environmental-related, can take place and make interpreting load data even more challenging (Li, Hong and Sofos, 2019, Gram-Hanssen, 2013). Given such data challenges, the classification of heat demand profiles in buildings, using time-series data clustering, has been an active research topic recently, either regarding different timeslots (Gianniou et al., 2018) or according to building occupancy activities (Carbonare, Pflug and Wagner, 2018).

1.3. Role of the proposed methods in the building energy operation framework

Below, the role of the proposed data analysis and modeling approach, in the framework of building energy management, is detailed and explained. This framework consists of four stages that should be

accomplished in sequence. A simple flowchart, illustrating this framework is displayed in Fig. 1. These four stages are:

1 Data cleaning and processing: it is common that "empty" values can be stored due to a temporary malfunction of the data transfer mechanism from sensors to databases. Additionally, noisy and faulty measurements can lead to outlier data that needs to be filtered out.
2 Data clustering: even after data cleaning, null values, and irregular patterns can be present in the collected data profiles, and here comes the role of data clustering. As energy consumption data is mostly considered here, time-series clustering is the right approach to dealing with time-based datasets. This operation regroups similar profiles into separate groups or clusters and filters out uncommon data patterns and measurement gaps.
3 Data-driven modeling: developing ANN models through training and validation processes using historical data. The models are subject to periodic updates by performing new training to enhance their performance using newly generated data.
4 Potential operation strategies: based on accurate prediction models, optimization algorithms can anticipate the heat load flows in a building or a district of buildings and find out the optimal operation strategy accordingly.

1.4. Objective of the study

Objective of this study is to develop and validate a hybrid ML approach for DH load prediction in the context of solving energy data quality problems applied to DH load profiles. This study includes data mining using unsupervised ML, followed by data-driven modeling using supervised ML. For data mining, time-series clustering using two different clustering algorithms is applied. For data-driven modeling, a multi-feature ANN regression model is developed and validated. As mentioned previously, due to the nature of regression problem, an MLP architecture with multiple hidden layers is adopted. Such an architecture is needed to deal with the complexity of relationships between different input parameters. The proposed approaches are validated on real datasets of DH networks involving different building types in the Nordic climate.

2. Methods

This section explains the methodology adopted for data processing and mining and the development of the ANN modeling approach. In this context, two techniques for time-series clustering are formulated for the classification problem of DH load datasets. The classified data can then be subjected to ANN training separately to generate, in the end, MLP prediction models.

2.1. Building and energy data inventory

In this article, hourly-based resolution datasets for 20 nursing homes and 31 schools in Trondheim, Norway, have been used for modeling and evaluation. The heated floor areas of the observed buildings range from 1350 to 10940 m^2 for nursing homes and from 1822 to 8996 m^2 for school buildings. Information on the observed buildings is summarized in Table 1, in which, the energy labeling scheme goes from A (best building energy performance) to G (weakest performance), and the labeling explanation is provided by the Norwegian Energy Efficiency Agency (ENOVA) (Enova Offentlig søk etter energiattester). The analysis was performed on the specific DH load of each building (W/m^2), which

Table 1
List of the observed buildings' information.

	Nursing homes						
Energy labeling	A	B	C	D	E	F, G	No info.
No. of buildings	/	4	6	6	3	/	1
Construction year	Before 1950	1950-1979	1980-1999	2000-2010	After 2010		No info.
No. of buildings	/	/	7	9	3		1
Label names	/	B1-B4	C1-C6	D1-D6	E1-E3	/	NI1
	School buildings						
No. of buildings	1	3	5	13	5	1	3
Construction year	Before 1950	1950-1979	1980-1999	2000-2010	After 2010		No info.
No. of buildings	2	5	3	18	/		3
Label names	A1	B1-B3	C1-C5	D1-D13	E1-E5	F1	NI1-NI3

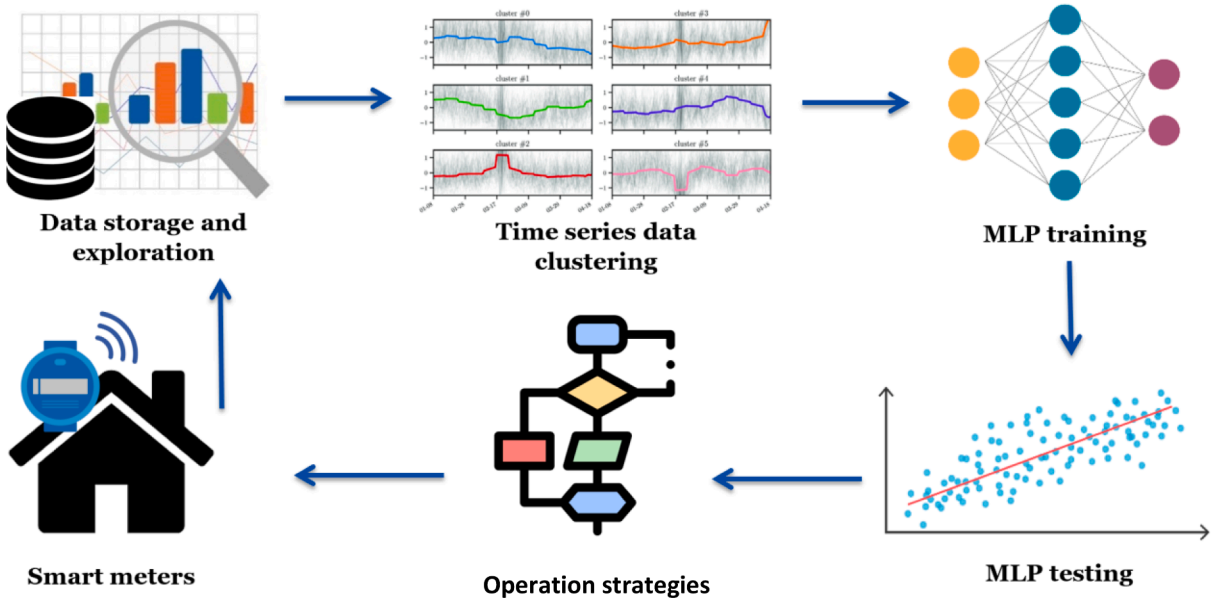

Fig. 1. Data transfer and analysis flowchart.

was meant to focus on the energy density of each building regardless of the building size and floor area.

2.2. Data preprocessing and normalization

To address the similarities of load patterns over the load magnitudes, data normalization was applied using the min-max normalization formulated in Eq. (1). In this study, data mining, modelling, and validation will be executed solely on normalized data, as the shape and patterns of the data are considered against data value ranges. The normalization is given as:

$$\overline{x_i} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where x_i refers to the DH load at *i-th* hour [kW],$\overline{x_i}$ refers to the normalized DH load at *i-th* hour, x_{min} and x_{max} refer to the minimum and maximum DH load value of the year [kW].

2.3. Time-series data clustering

Clustering is the task of organizing data in such a way that similar objects are placed into related or homogeneous groups without prior knowledge of the groups' definitions. Time-series clustering is one form of cluster analysis that has been used in many scientific areas to discover interesting patterns in time-series datasets, such as smart meter datasets (Aghabozorgi, Shirkhorshidi and Wah, 2015). As a data mining approach, datasets with similar characteristics are regrouped into one cluster, and each cluster can be studied separately afterwards, which improves the efficiency of the data analysis. In this study, the objectives of this approach are:

- Identifying low-quality datasets, related mainly to measurement gaps and outliers Such datasets will be later eliminated from the training process to improve the performance of ANN prediction.
- Regrouping datasets with irregular patterns within one specific cluster Those datasets are typically generated with uncommon load profiles, which are usually related to transitional changes in occupancy or in building energy performance.
- The rest of the datasets should have higher chances of being representative of the studied building type, so they are arranged in one separate cluster. The main advantage here is that one prediction model can be valid for all buildings belonging to this cluster.

Most of the approaches for time-series clustering are built based on two major design criteria: the clustering algorithm and the distance measure. The choice of clustering algorithms may depend on the strategy used to maximize intra-group similarity and minimize inter-group similarity (Javed, Lee and Rizzo, 2020). In this paper, k-means is chosen as a clustering method because of its popularity. The k-means algorithm generates spherical clusters that are similar in size and optimizes clustering by minimizing the distance between each cluster center (centroid) and the data points within that cluster (Jesper, Pag, Vajen and Jordan, 2021). The k-means algorithm requires that one input parameter be specified: the number of clusters (k). Given k, the algorithm iterates over two phases: (1) calculating centroids, and (2) assigning data points to their closest centroid, until a certain termination condition (e.g., number of iterations or convergence) is met. The initial centroids are chosen randomly, making the algorithm non-deterministic; all subsequent centroids are calculated to minimize the distance to all other data points within the given cluster.

For distance measurement, two different methods are adopted in this study: Euclidean and Dynamic Time Warping (DTW), generating two different clustering approaches.

2.3.1. Euclidean distance measure

The Euclidean distance between time-series is straightforward and widely adopted as a distance measure for k-means algorithm (Zhang, Tang, Huo and Zhou, 2014). It calculates the distance between measurement points in two different time-series datasets X and Y using Eq. (2):

$$EUC(X,Y) = \delta(x,y) = \sum_{i=1}^{T} \sqrt{(x_i - y_i)^2} \tag{2}$$

where $\delta(x,y)$is the distance between data points inside different time-series. $x_i$and y_i are two points from the two time-series X and Y respectively.

2.3.2. DTW distance measure

The DTW algorithm has earned its popularity by being extremely efficient, as it minimizes the effects of profile distortion in time by allowing "elastic" transformation of time series. This feature makes it possible to detect similar shapes with different phases (Zhang, Tang, Huo and Zhou, 2014). In contrast to Euclidean distance, which chooses the most straightforward way for aligning (Fig. 2 (a)), DTW distance chooses a more flexible alignment rule, it can find the best global alignment to achieve the minimum accumulated distance and handle time-series with different lengths, as shown in Fig. 2 (b). The distance measure according to DTW is formulated in Eq. (3).

$$DTW(X_i, Y_i) = \sum_{i=1}^{T} \delta(xi, yi) + \min \left\{ \begin{matrix} DTW(X_i, Y_{i-1}) \\ DTW(X_{i-1}, Y_i) \\ DTW(X_{i-1}, Y_{i-1}) \end{matrix} \right\} \tag{3}$$

With the following boundary conditions:

$$\begin{matrix} DTW(0,0) = 0 \\ DTW(X_i, 0) = \infty \\ DTW(0, Y_i) = \infty \end{matrix} \tag{4}$$

where $\delta$is the same as Euclidean distance Eq. (2) andX_iY_i are two time-series datasets.

2.4. Heat load prediction

In this article, prediction models, using multi-input ANN, were created to predict the DH load for any given hour. In addition to the simple implementation, the reason behind choosing ANN is its ability to model nonlinear phenomena and handle multi-feature regression problems efficiently (Abiodun et al., 2018). The different behaviors of SH and DHW about the outdoor temperature, lead to the fact that total heat demand in buildings is nonlinear, with which classical linear regression models are no longer efficient.

One model is to be created for each cluster, therefore, a total of four models were created for the two building types (nursing homes and schools). In addition to outdoor weather conditions, heat energy demand in buildings can be affected by many other factors. In this work, the chosen data inputs, for each hour, are listed below:

- The forecasted outdoor temperature (in °C), is typically accessible via an online service.
- Hour of the day: from 1 representing "1:00-2:00" to 24 representing "23:00-00:00",
- Day of the week: an integer, from 1 for Monday to 7 for Sunday,
- Month of the year: an integer, from 1 for January to 12 for December,
- Holiday indicator: a Boolean, 1 for a holiday and 0 for a working day.

Those parameters are set basically to give an estimation of the heat load trend, which changes essentially according to different timeslots. They can also give an approximate assessment of the occupancy rate and human activities related to the building type (Alam, Bao, Zou and Sanjaya, 2017).

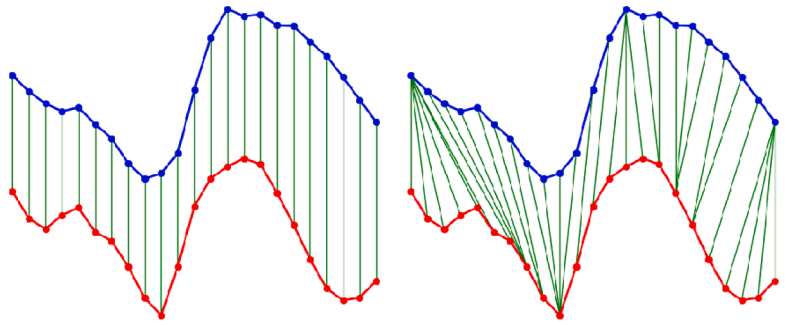

Fig. 2. Different aligning rules for (a) Euclidean distance and (b) DTW distance matching (Zhang, Tang, Huo and Zhou, 2014).

As the analyzed buildings are not residential, it can be possible to find some indicators having correlations with the occupancy rate with an accepted approximation, which is the reason behind the choice of the last four ANN inputs listed above. In this context, "hour of the day" can give an approximate guess of the daily scheduled work time for schools and nursing homes. In the same way, "day of the week" has been introduced to differentiate between working days and weekends, which is more significant in the case of schools as nursing homes still have partial work activities on weekends. With the same impact, the inputs "month of the year" and "holidays indicator" have been introduced to highlight vacation periods.

2.4.1. *Multi-layer perceptron*

A Multi-Layer Perceptron is a type of feedforward artificial neural network that consists of multiple layers of interconnected nodes, called neurons or units. It is a foundational and widely used architecture in the field of deep learning (Abiodun et al., 2018, Alam, Bao, Zou and Sanjaya, 2017). The term "perceptron" refers to a basic building block that models a single neuron, and "multi-layer" indicates that multiple layers of these perceptrons are stacked together.

In an MLP, the neurons are organized into layers, typically consisting of an input layer, one or more hidden layers, and an output layer. The input layer receives the input data, and the output layer produces the final output of the model. The hidden layers are intermediate layers between the input and output layers, responsible for processing and transforming the input information through a series of non-linear operations. Fig. 3 shows an illustrative architecture of an MLP with an input layer of five neurons; three hidden layers with 12 neurons each and one output layer with one single neuron.

Each neuron in the MLP is associated with a set of learnable parameters, including weights and biases. The weights determine the strength of the connections between neurons, while the biases introduce an offset or bias term to the neuron's output. These parameters are adjusted during the training process to optimize the performance of the MLP on a given task.

The functioning of an MLP involves two main steps: forward propagation and backward propagation. In forward propagation, the input data is fed through the network, and the activations of each neuron are computed layer by layer, ultimately producing the output. Backpropagation involves computing the gradients of the network's parameters concerning a chosen loss function, allowing for the adjustment of the weights and biases in a direction that minimizes the loss. This iterative process of forward and backward propagation is repeated until the model converges to an optimal state. Eq. (5) formulates the forward process through MLP architecture.

$$f^{(L)}(X) = a\left(w_0^{(L)} + \sum_{i=1}^{U_{L-1}} w_i^{(L)} f_i^{(L-1)}(X)\right) \tag{5}$$

where:

L is the layer number.

$f^{(L)}$ is the L function layer.

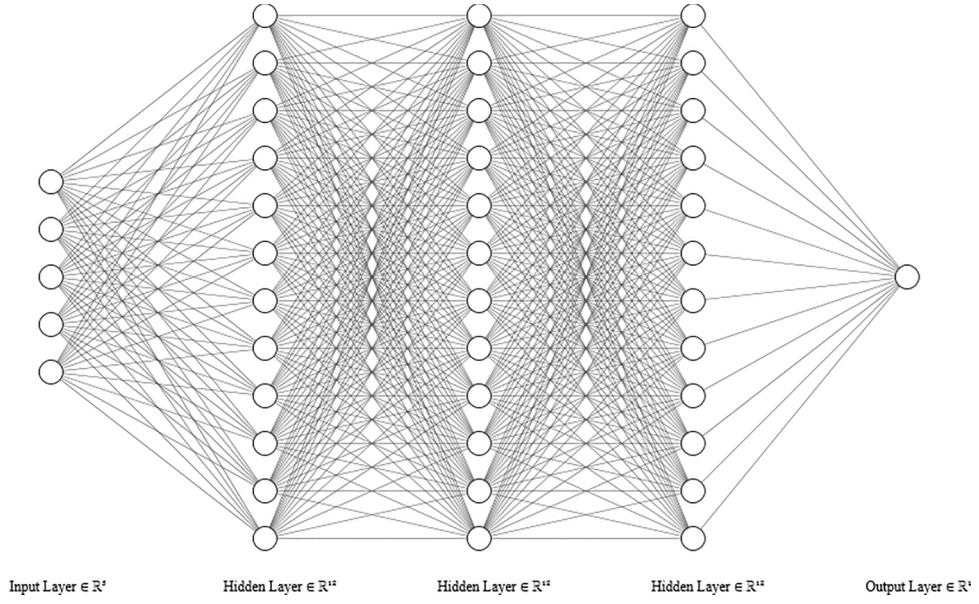

Fig. 3. Example of an MLP model structure.

$w_0^{(L)}$ and $w_i^{(L)}$ are respectively bias and weights vectors corresponding to layer L.

X is the input vector.

a is the activation function (it can represent different activation functions between different layers).

U_{L-1} is the number of previous layers of L.

The chosen activation function is the Sigmoid function, it is formulated as:

$$a(x) = \frac{1}{1+e^{-x}} \tag{6}$$

2.4.2. Training

During the training phase, an optimization algorithm is used to update the weights and biases of the MLP to minimize a loss function, which is formulated in Eq. (7). As an optimization method, this study addresses the use of Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS). It is an optimization algorithm used for solving unconstrained optimization problems. The algorithm combines the BFGS method with limited memory storage, allowing it to approximate the inverse Hessian matrix without explicitly computing or storing it. The limited-memory aspect of LBFGS makes it memory-efficient and suitable for large-scale optimization problems. However, the detailed mathematical description of LBFGS is outside the scope of this article.

$$MSE = \frac{1}{N_P}\sum_{i=1}^{N_P}(\overline{y_i} - \widehat{y}_i)^2 \tag{7}$$

where $\overline{y_i}$ and $\widehat{y}_i$ are the actual and predicted DH load respectively at the hour i, N_P is the number of data samples.

The method "*neural_network.MLPRegressor*", which is a part of the Python package "*sklearn*", was chosen as an implementation tool. The training parameters are set in Table 2, and those parameters are reached after some trial-and-error tests. As training data, two datasets for hourly DH heat consumption for 2017 and 2018 are used. As stated earlier, the training will be solely executed on datasets showing a high correlation with the ANN model input parameters, which should be gathered in Cluster 1 and Cluster 2. Low-quality data, having measurement gaps and outliers, is excluded as it represents anomalies. The corresponding datasets, in this case, are gathered in Cluster 3.

The pseudocode for a trained MLP is summarized below:

```
Initializing w_0^(1) and w_i^(1)
  Acquiring training data X
  While (training stopping criteria is not met){
    Calculating the loss function
  Updating weights and bias for all layers
  MLP feedforward}
```

2.4.3. Performance evaluation

Datasets for DH load in 2019 are preserved for ANN validation. The performance of the prediction model was evaluated using Eq. (7) and mean absolute error (MAE) formulated in Eq. (8). For this purpose, two datasets, arbitrarily taken from Cluster 1 and Cluster 2, are the subject of model testing. The motivation behind this strategy is to prove how efficient data analytics and modeling in cluster-wise instead of the classical approach where all datasets are addressed.

Table 2
MLP training parameters.

Number of neurons in the input layer	5
Number of hidden layers	5
Number of neurons in one hidden layer	50
number of neurons in the output layer	1
Activation function	Sigmoid
Optimizer	LBFGS
Learning rate	Constant (0.1)
Loss function	MSE

$$MAE = \frac{1}{N_P}\sum_{i=1}^{N_P}|\overline{y_i} - \widehat{y}_i| \tag{8}$$

3. Results

In this section, simulation results for the energy data processing and modeling explained previously, are presented. Section 3.1 shows data clustering results for the whole studied datasets. Section 3.2 is the stage for load prediction validation for different clusters, while Section 3.3 is a performance assessment study in comparison with a similar approach.

3.1. Time-series clustering

In this section, the outcomes of data clustering are presented for nursing homes and schools separately. The goal is to show different data patterns that may influence the modeling process and, therefore, the expected load prediction performance. For this purpose, only 10 datasets have been randomly selected for each building type. The reason behind limiting the study to this dataset is related mainly to the computing time restriction.

3.1.1. Nursing home data clustering

For nursing home data, Fig. 4 represents the clustering results using Euclidean k-means. The different clusters contain various dataset numbers and show different load patterns. In this regard, Cluster 1, which gathers the most regular load profiles, has 5 datasets. Cluster 2, which gathers relatively uncommon load data profiles, has 4 datasets, while Cluster 3 has only one dataset, characterized by tiny load variation and multiple data gaps and outliers (zones highlighted by black dashed rectangles). Large constant values, that appear in Cluster 3, which can keep appearing for months, are due to the chosen data interpolation policy in the preprocessing method adopted locally. This policy is based on keeping the same last measured value if data transfer from smart sensors is lost. This cluster, in particular, is to be eliminated from the ANN training.

Using the DTW k-means technique, Fig. 5 shows the dataset classification corresponding to the three clusters. In this case, Cluster 1 has 4 datasets, Cluster 2 has 5 datasets, and Cluster 3 has the same unique dataset shown previously (Cluster 3 in Fig. 5). In the case of nursing homes, Euclidean and DTW k-means algorithms show comparable results with slight differences in dataset numbers between clusters. It is worth mentioning that the computing time of DTW k-means is substantially larger compared to that of Euclidean k-means. About two hours of computing time have been recorded for DTW k-means, while only a few seconds have been recorded for Euclidean k-means using the same datasets.

3.1.2. Schools data clustering

Similarly, time-series clustering is applied to school datasets, and the results are displayed in Fig. 6 and Fig. 7 using, respectively, Euclidean and DTW k-means techniques. As a first general overview, the figures show that the seasonal variation of heat load for schools looks noticeably more fluctuating compared to nursing homes. The reason is related essentially to differences in occupancy and building geometry and physics. In this situation, the difference between Euclidean and DTW k-means in clustering performance looks more visible. For the first technique, Cluster 1 gathered 5 datasets, Cluster 2 gathered only 1 dataset, and 4 datasets are grouped in Cluster 3. Using DTW k-means, the dataset's classification is more meaningful, as Fig 7 shows. In this context, 2 datasets with regular patterns are grouped in Cluster 1, and 7 datasets are grouped in Cluster 2, some of which are having large load fluctuations in winter 2019, as highlighted in the dashed rectangles. One specific dataset could be identified in Cluster 3, which is considered to have a tiny load variation and a large measurement gap, highlighted by the dashed black box.

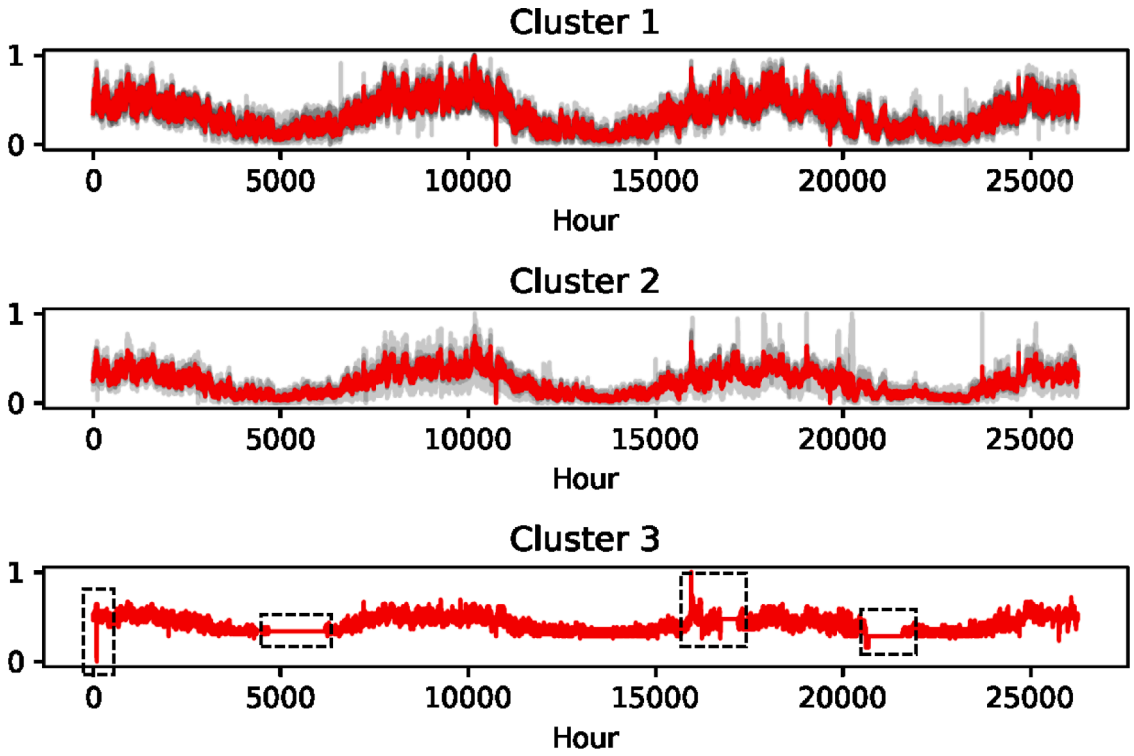

Fig. 4. Clustered datasets for nursing homes for the years 2017, 2018, and 2019 using Euclidean k-means.

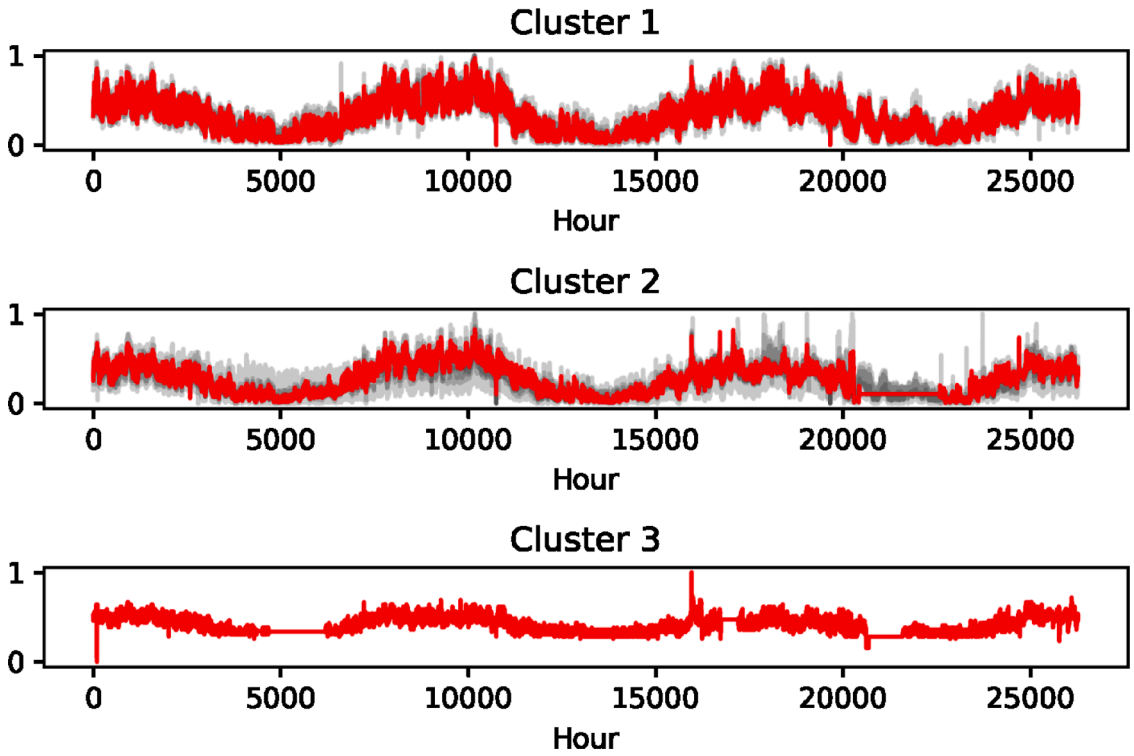

Fig. 5. Clustered datasets for nursing homes for the years 2017, 2018, and 2019 using DTW k-means.

For more quantified information, further statistical analysis has been summarized in Table 3, in which, average values for "mean" and "standard deviation" are calculated for all datasets existing in the generated clusters. As a general overview, the ratio of standard deviation to mean is relatively higher for schools than for nursing homes. This explains the fluctuating pattern of annual load profiles for schools. Additionally, similar values are calculated for different clusters separately. The results point to the fact that datasets in Cluster 2 are smoother compared to those in Cluster 1, with slight differences between the used clustering algorithms. It is worthy of mention that there is no need to calculate maximum and minimum values for datasets as they are all normalized.

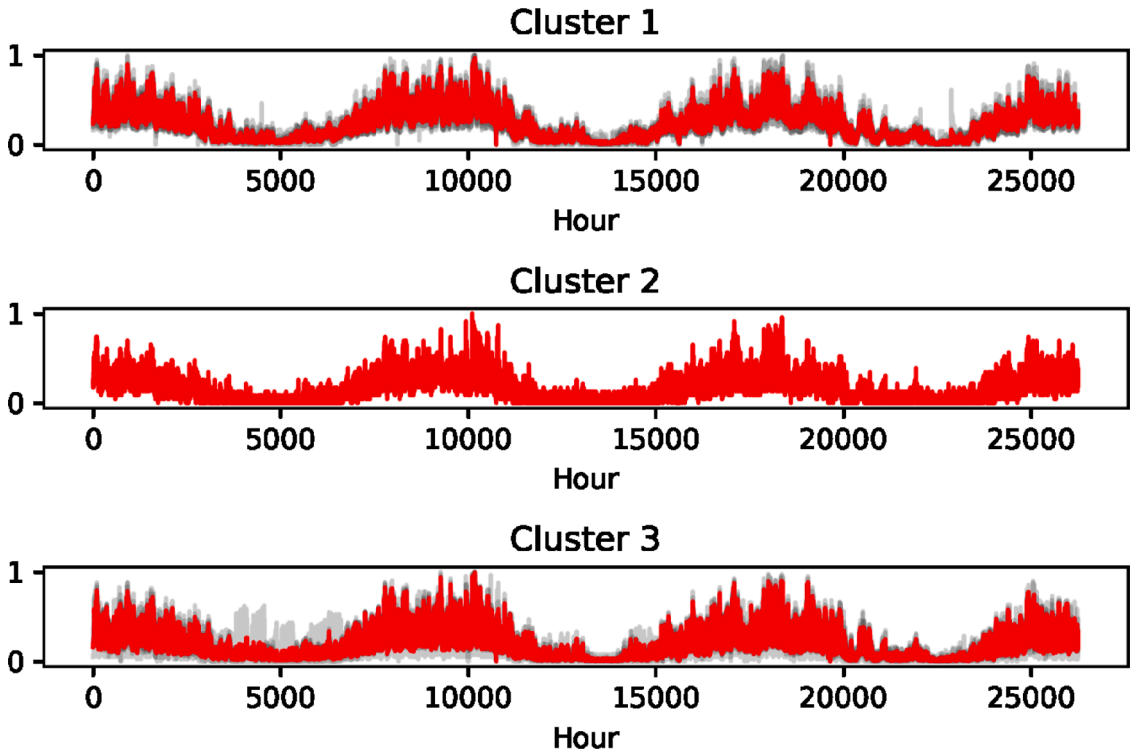

Fig. 6. Clustered datasets for schools for the years 2017, 2018, and 2019 using Euclidean k-means.

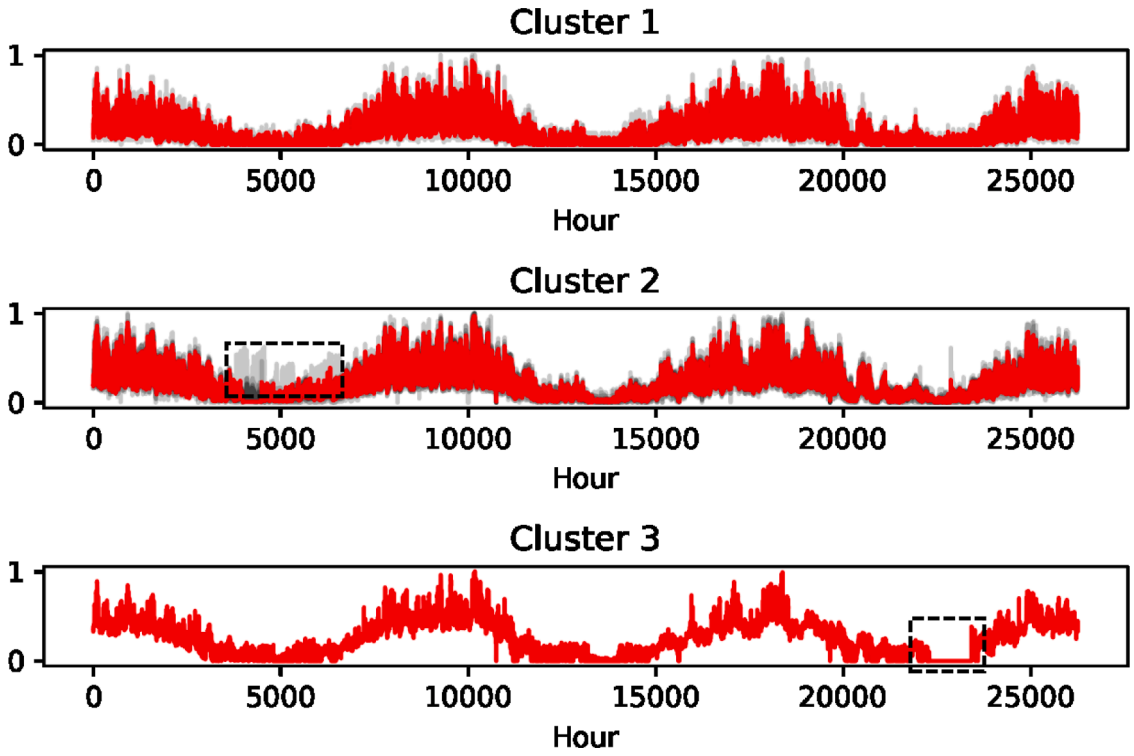

Fig. 7. Clustered datasets for schools for the years 2017, 2018, and 2019 using DTW k-means.

3.2. Load prediction

For each building type, two datasets, arbitrarily selected from clusters 1 and 2 separately, are the subject of a prediction performance evaluation using MLP. The goal is to verify and quantify the prediction accuracy cluster-wise for the two studied building types.

3.2.1. Long-term assessment

As a general overview of load prediction performance, annual assessments of the MLP model response over the validation data for the two building types are displayed. As the performance of data clustering varies slightly between the two techniques, Euclidean and DTW k-means, the prediction performance changes accordingly too. e.g., in

Table 3
Statistic measures of the generated normalized datasets in cluster-wise.

School datasets	Mean	Standard deviation	Nursing home datasets	Mean	Standard deviation
Cluster 1 (Euc k-means)	0.2558	0.1831	Cluster 1 (Euc k-means)	0.3368	0.1473
Cluster 1 (DTW k-means)	0.1214	0.1584	Cluster 1 (DTW k-means)	0.3235	0.1927
Cluster 2 (Euc k-means)	0.1492	0.1396	Cluster 2 (Euc k-means)	0.1337	0.0988
Cluster 2 (DTW k-means)	0.2185	0.1787	Cluster 2 (DTW k-means)	0.2328	0.1330

Fig. 8, the load prediction is conducted moderately differently between the two datasets due to the difference in load fluctuation patterns. The datasets extracted from Cluster 2 for schools, generated by both clustering methods, show relatively less modeling capability, which is translated to the average prediction performance summarized in Fig. 9.

Mainly due to the different occupants' activity schedules, the heat load annual profile for nursing homes is less fluctuating, as Fig. 10 shows, compared to schools. This is because DH load profiles for schools are much sharper, with distinct night setbacks and weekday/weekend shifts than those for nursing homes. Moreover, heat loads in summer for schools are moderately lower than those for nursing homes, which is highlighted by the black dashed rectangles in Fig 8 and Fig 9. The reason is that no official scheduled activities are carried out during the summer in schools, which is not the case for nursing homes.

Cluster 2, issued by the two studied clustering techniques for nursing home data, shows pretty different patterns compared to Cluster 1 (see Fig. 11). In this case, the annual heat load for the corresponding buildings is considerably lower. Low SH demands for these particular buildings can represent one potential justification. Moreover, Cluster 2 also includes some data anomalies, such as the outlier value in Fig. 11 (a) and the measurement gap in Fig. 11 (b), which are both highlighted by black dashed rectangles.

3.2.2. Seasonal assessment

To gain a deep insight into the prediction process in different seasonal periods, three-week scenarios have been selected: cold, warm, and mild. Below, more explanations, along with some graphs, are presented for this context. Datasets, chosen arbitrarily from Cluster 1, generated by DTW k-means for each building type, have been selected.

In Fig. 12, the ANN model shows a high prediction capability for schools in the winter compared to the summer scenario. The reason is that in winter, heat load peaks appear clearly during working days daily, while in summer, this load pattern is negligible. The heat demand is substantially lower during the weekend in winter, while in summer this difference is not significant. This matter is highlighted by the black dashed rectangles in Fig 12 (a) and Fig 12 (b). Peak energy demands in winter are linked mainly with the need for SH and the noticeable occupancy rate. In the summer, SH demands are insignificant, and occupancy is almost null due to the annual vacation.

Concerning nursing homes, the ANN model shows clearly low prediction capabilities, especially in summer (see Fig. 13 (b)), as in this case, the outdoor temperature factor is less dominant for DH load (SH demand is significantly low). In this particular case, the DHW load represents the main energy demand portion, which is related to factors other than weather data. This effect is less dominant in the case of schools due to the impact of other important aspects, mainly related to the occupant's activities.

For load prediction assessment in mild seasons, the fact that outdoor temperature is no longer a dominant parameter for determining the heat load leads to inaccurate predictions for nursing homes (see Fig. 14 (a)), however, ANN is still showing relatively better results when it comes to school heat load prediction.

3.3. Prediction performance evaluation

Table 4 lists MSE and MAE values for the prediction performance of the proposed ANN regression method for all clusters, compared to a reference method developed in (Ohlson Timoudas). The authors in (Ohlson Timoudas) developed an ANN approach that includes, along with the historical outdoor temperatures, historical heat loads with different time lengths. It has been found that better results were reached using historical values of outdoor temperature and load up to 24 hours. This approach has been validated on the same datasets used in this

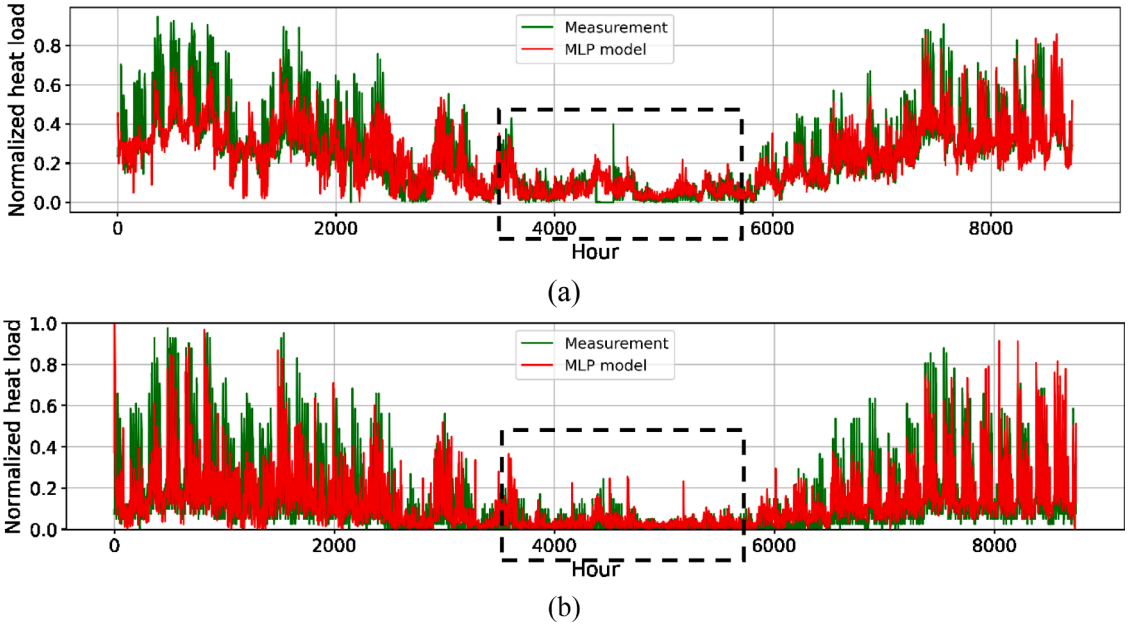

Fig. 8. Annual prediction assessment using a school dataset extracted from Cluster 1 (a) using Euclidean k-means (b) using DTW k-means.

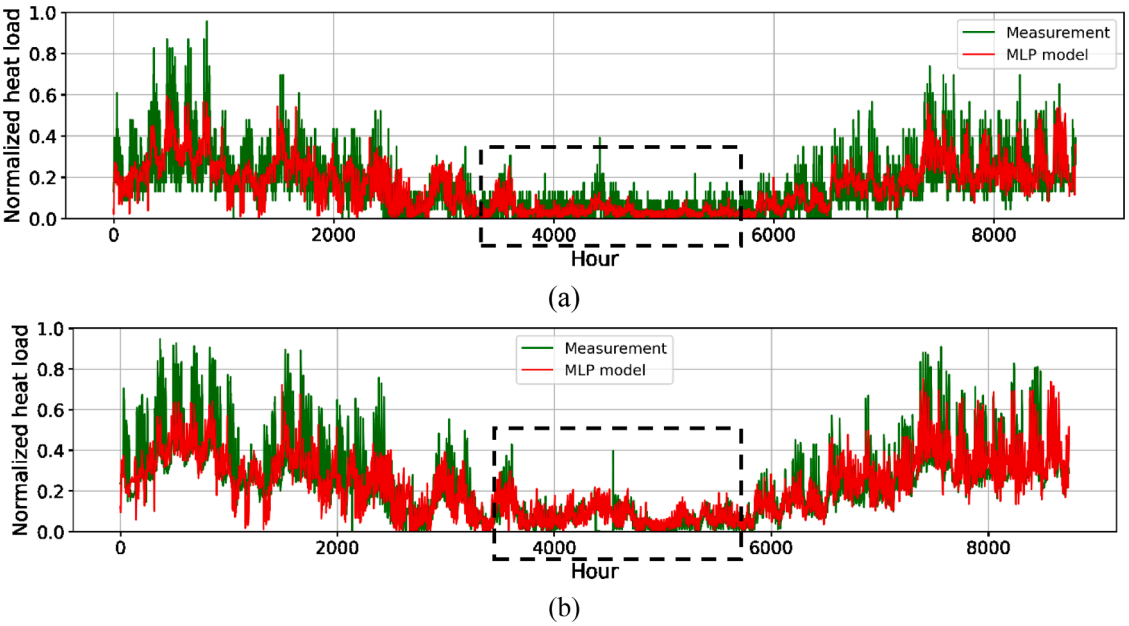

Fig. 9. Annual prediction assessment using a school dataset extracted from Cluster 2 (a) using Euclidean k-means (b) using DTW k-means.

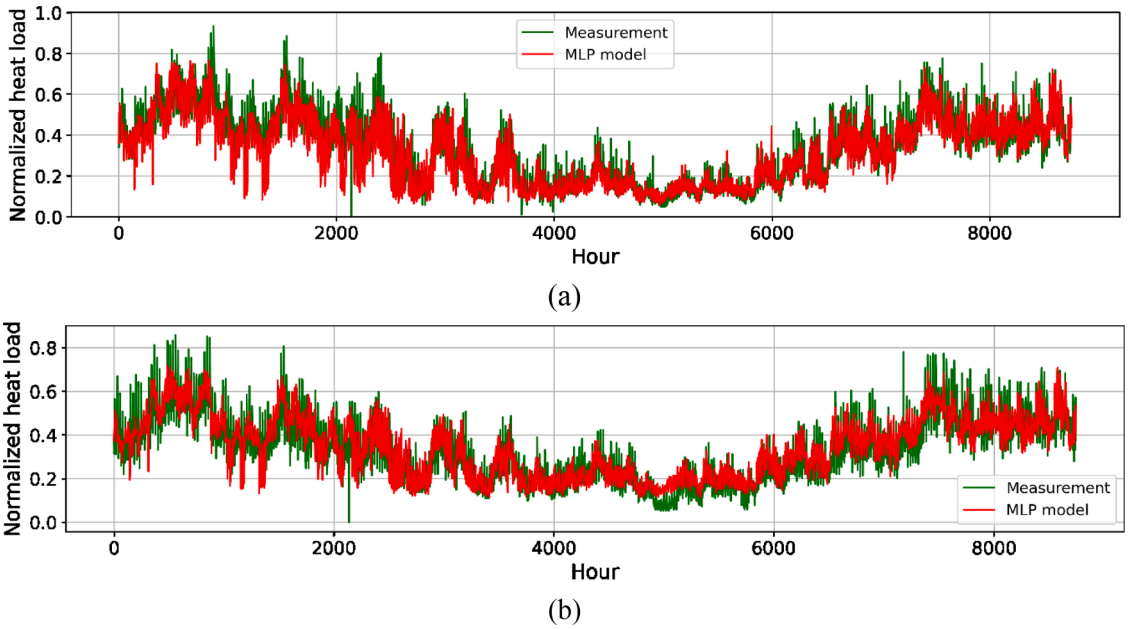

Fig. 10. Annual prediction assessment using a nursing home dataset extracted from Cluster 1 (a) using Euclidean k-means (b) using DTW k-means.

article, either for training or testing. Better results have generally been obtained using Cluster 2 datasets using both clustering techniques (Euclidean and DTW k-means). One justification for that is that Cluster 2 generally gathers less fluctuating load profiles, as shown in Fig. 11. This makes it easier for MLP to track smoother load curves appearing in the winter as they reflect SH demands. Such load curves are highly dependent on weather data, mainly, outdoor temperatures, which explains the high prediction performance in this case. Fig. 10 shows more fluctuating curves, which can be affected by other additional factors, such as occupancy change rates. For all clusters, the table shows clearly that the comparison is in favor of the MLP model, presented in this study, compared to the ANN baseline method.

4. Discussions and future study

4.1. Clustering

It has been verified that load data tendency has an impact on the expected ANN prediction performance. Inappropriate data quality may appear during the measurement process, typically due to a temporary malfunctioning of energy meters, possible noise sources, or data transfer interruptions. In this study, such datasets are gathered in Cluster 3 for each building type and eliminated from the modeling stage. This filtering option can be significantly important during ANN training by providing only datasets with higher correlations with the input variables. Additionally, time-series clustering also serves in the identification of uncommon load patterns, which generally arise from differences in occupancy rate or the change in energy performance between

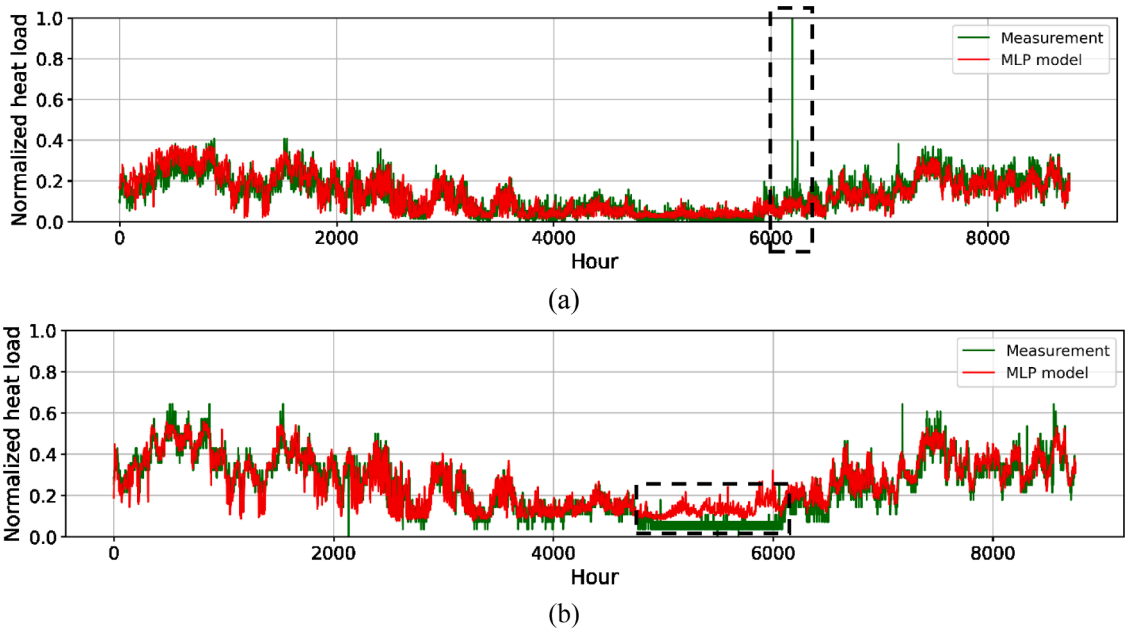

Fig. 11. Annual prediction assessment using a nursing home dataset extracted from Cluster 2 (a) using Euclidean k-means (b) using DTW k-means.

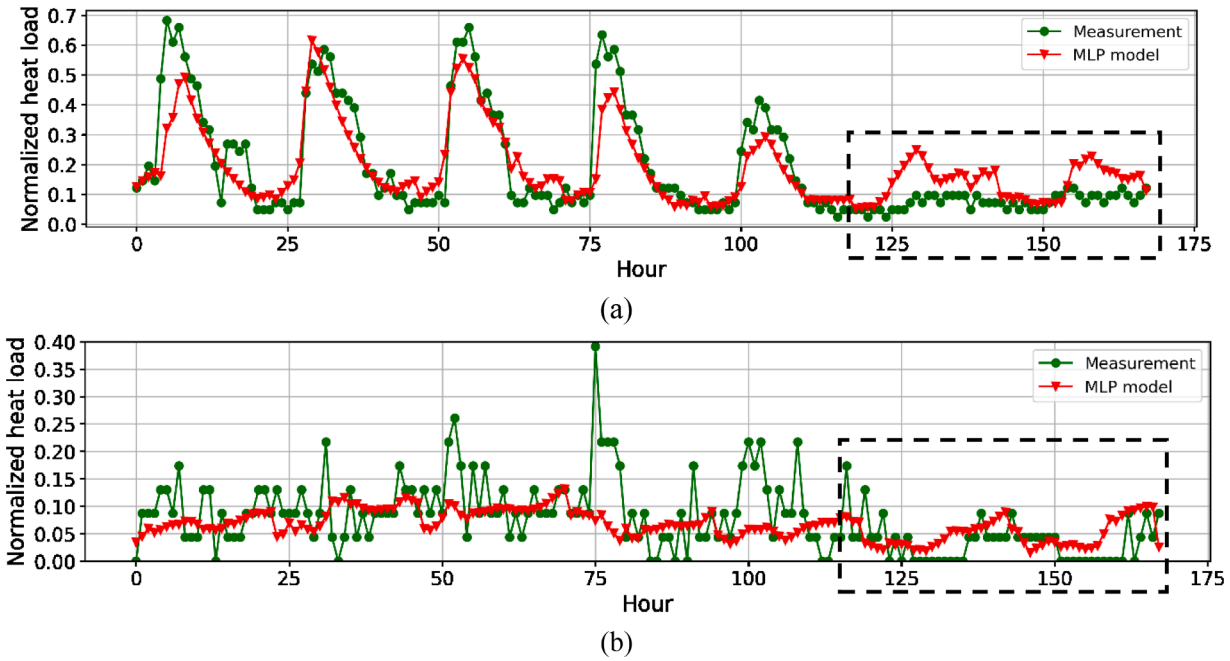

Fig. 12. Prediction performance for a school dataset (a) in winter (b) in summer.

buildings, which is mostly linked with building physics. Such datasets are grouped in Cluster 2. One main advantage of data clustering as a preprocessing step is that one prediction model is valid for all buildings classified in the same cluster. This can, significantly, reduce data analysis time and cost, and improve the expected modeling performances by selecting clusters with regular patterns only.

It should be noted that the number of datasets in each cluster for a particular building type may be subject to continuous change over the years due to changes in indoor conditions such as occupancy rate and building energy performance degradation. Outdoor conditions, mainly related to extreme weather conditions, can also have an impact in this regard.

4.2. Prediction

Particularly for the studied datasets, summer tests for all buildings and all clusters show relatively lower prediction performance compared to winter tests. In summer, occupants in nursing homes are present permanently, continuously utilizing DHW along with small fractions of SH. The prediction in this case can be challenging because of the lack of peak loads that typically characterize SH demands, whereas the nature of the school's occupancy is different, as no study activities are carried out in July. Schools have relatively more fluctuating patterns than nursing homes, and during the daytime, the peak load in schools can be significant. However, schools can be rented out outside of working hours, therefore, a few high heating demands in the evenings can be

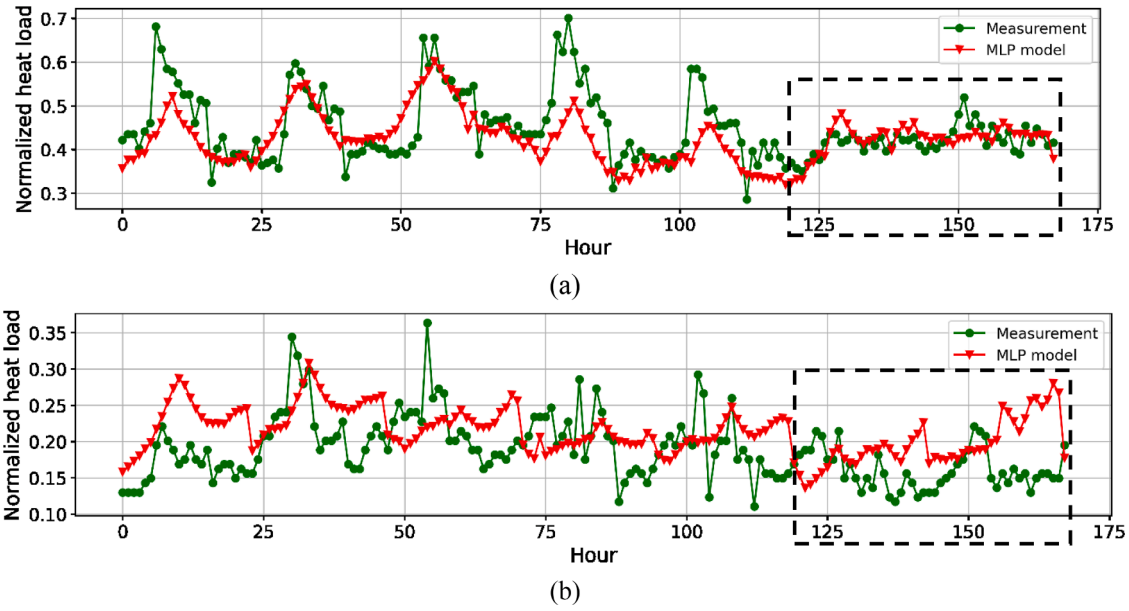

Fig. 13. Prediction performance for a nursing home dataset (a) in winter (b) in summer.

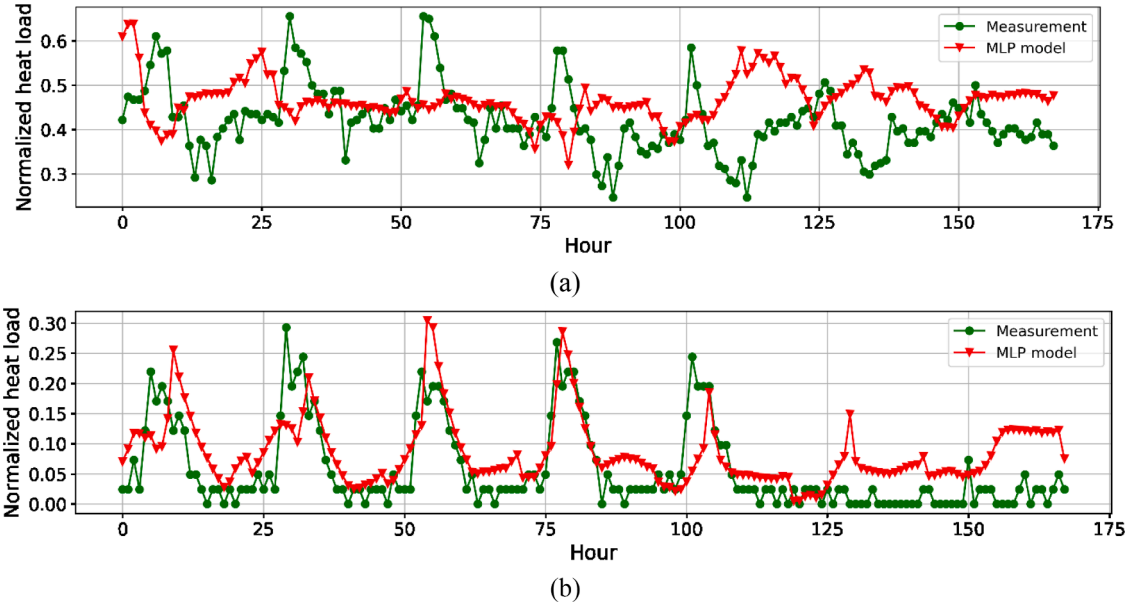

Fig. 14. Prediction performance in mild-season (a) for a nursing home (b) for a school.

Table 4
MSE and MAE evaluation using nursing homes normalized data for the year 2019.

Evaluation criteria	Proposed MLP		ANN proposed in [42]
MSE	Cluster 1 (Euc k-means)	0.0153	
	Cluster 1 (DTW k-means)	0.0154	0.0219
	Cluster 2 (Euc k-means)	0.0073	
	Cluster 2 (DTW k-means)	0.0072	
MAE	Cluster 1 (Euc k-means)	0.0975	
	Cluster 1 (DTW k-means)	0.0977	0.1106
	Cluster 2 (Euc k-means)	0.0639	
	Cluster 2 (DTW k-means)	0.0638	

present, which was reflected by a few outliers in some periods. Nursing homes are occupied all the time, so the DH load is smoother with fewer fluctuations during the day, between working days, and on weekends. In the mild season, the outdoor temperature is not a dominant factor in the heat energy demand, and therefore, weather data alone is not sufficient to have a better prediction performance. Though the proposed ANN model with the additional data inputs could overcome this challenge,

4.3. Limitations

Regardless of the high prediction performance of the proposed ANN regression model, data availability for building occupancy can be a challenge, especially in residential buildings. Contrary to nursing homes and schools, in which, the occupant's activities and habits are relatively predictable, occupancy in residential buildings is stochastic and irregular to some extent. Therefore, the input data used in this article for building occupancy estimation may not be effective for load prediction in normal residential buildings. One potential solution to deal with such an issue is the installation of IoT-enabled sensors, designed to detect the

occupancy rate and model the entire building. Along with other comfort sensors, those sensing points should be built within an optimal and cost-effective architecture (Deb and Schlueter, 2021). Though, data privacy issues in residential buildings should be solved in advance.

Furthermore, the simulation tests displayed in this article are carried out assuming a perfect prediction of the outdoor temperature. The ANN model is designed to take advantage of the available online services for weather forecasts. Therefore, another study should be carried out, investigating the impact of the online weather forecast error margin on the prediction performance and, consequently, on the energy operation strategy in buildings.

4.4. Future work

In contrast to the accurate performance of MLP models for school heat load prediction, especially in winter, the problem of DHW load prediction in nursing homes is still not solved yet. As stated earlier in this article, DHW needs are not highly correlated with weather data. As nursing homes represent a specific use case, the suggested ANN inputs, for occupancy estimation, are not accurate enough in this context. Adding historical data for heat consumption itself, as additional ANN input can partially solve this issue since DHW load should have a regular periodic pattern related to the previous usage. In future work, the MLP model will be validated on the same datasets using this new ANN architecture, and with additional input, which is historical data of the load itself with different time windows, e.g., up to 12 hours, 24 hours, or 72 hours.

The impact of load prediction accuracy on energy operation strategies in buildings should be evaluated and quantified. This depends highly on the chosen integration solution of the district heating network in building substations and on the available energy flexibilities. Energy storage systems and HPs can represent effective solutions in this context. Thus, different optimization-based operation policies can be validated, taking advantage of the flexible control capacity of HP and any potential energy storage systems.

Instead of dealing with data quality, time-series clustering can also be exploited to identify and classify building types and categories based on data profiles. This preprocessing phase can offer additional information to be included in ANN training data as a supplementary indicator. With this modeling architecture, the heat load prediction problem in buildings can be solved with a more generalized approach and improved prediction accuracy.

5. Conclusion

Heat load prediction in buildings is a complex task due to the variety of involved factors. Outdoor temperature is an essential parameter for prediction, particularly for SH power demands. However, more indicators are needed for the development of accurate load prediction models. In this article, a data-driven approach using multi-input ANN is essentially targeted. This study opens pathways toward optimal integration of prediction models into potential operation strategies, and therefore, into the overall transition toward modern DH networks. The proposed methods have been validated on real load datasets for a DH network in a Nordic climate. The main findings are summarized as follows:

- Data mining, as a pre-processing step, is suggested to deal with insufficient data quality. In this regard, time-series clustering can efficiently fulfill this mission, it gathers datasets with similar patterns into specific clusters, allowing the identification of noisy and intermitted measurements, as well as abnormal patterns.
- Data cluster-wise assessment can significantly reduce the analysis and modeling costs by limiting ANN training and validation to representative datasets only.
- Ordinary regression approaches, relying on weather data, may have some limitations since partial heat loads, mainly related to DHW, can be independent of weather conditions. This is observable in summer and mid-season. In this regard, MLP with additional input features can make a valuable contribution to solving this issue.
- The advantage of multi-layer ANN architecture (MLP), to deal with the complexity associated with the relationships between different data inputs and the heat energy use, has been verified. However, there is a slight difference between a good model and an overfitted one, when it comes to the selection of the optimal number of hidden layers and the number of neurons in each layer. This task can usually be accomplished by trial and error. In the same context, computing time during training is another parameter to consider.

Declaration of Competing Interest

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

Data availability

The authors do not have permission to share data.

Acknowledgment

This study is funded by the Swedish Energy Agency (grant number 51544-1), the European Union's H2020 programme (grant agreement number 101036656), and the Research Council of Norway (grant number 268248). Special thanks go to Trondheim Municipality.

References

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings, 40*(3), 394–398. https://doi.org/10.1016/j.enbuild.2007.03.007

Directive (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency.

Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC Text with EEA relevance OJ L 315, 14.11.2012, p.1-56.

Werner, S. (2017). International review of district heating and cooling". *Energy, 137*, 617–631. https://doi.org/10.1016/j.energy.2017.04.045

Lund, H., Werner, S., Wiltshire, R., Svendsen, S., Thorsen, J. E., Hvelplund, F., & Mathiesen, B. V. (2014). 4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems". *Energy, 68*, 1–11. https://doi.org/10.1016/j.energy.2014.02.089

Li, H., & Nord, N. (2018). Transition to the 4th generation district heating - possibilities, bottlenecks, and challenges. *Energy Procedia, 149*, 483–498. https://doi.org/10.1016/j.egypro.2018.08.213

Lumbreras, M., & Garay, R. (2020). Energy & economic assessment of façade-integrated solar thermal systems combined with ultra-low temperature district-heating". *Renewable Energy, 159*, 1000–1014. https://doi.org/10.1016/j.renene.2020.06.019

Wahlroos, M., Pärssinen, M., Manner, J., & Syri, S. (2017). Utilizing data center waste heat in district heating Impacts on energy efficiency and prospects for low-temperature district heating networks". *Energy, 140*, 1228–1238. https://doi.org/10.1016/j.energy.2017.08.078. Part 1.

Ziemele, J., Kalnins, R., Vigants, G., Vigants, E., & Veidenbergs, I. (2018). Evaluation of the industrial waste heat potential for its recovery and integration into a fourth generation district heating system. *Energy Procedia, 147*, 315–321. https://doi.org/10.1016/j.egypro.2018.07.098

Fitó, J., Hodencq, S., Ramousse, J., Wurtz, F., Stutz, B., Debray, F., & Vincent, B. (2020). Energy- and exergy-based optimal designs of a low-temperature industrial waste heat recovery system in district heating". *Energy Conversion and Management, 211*. https://doi.org/10.1016/j.enconman.2020.112753

Frei, M., Deb, C., Nagy, Z., Hischier, I., & Schlueter, A. (2021). Building energy performance assessment using an easily deployable sensor kit: Process, risks, and lessons learned. *Frontiers in Built Environment, 6*. https://doi.org/10.3389/fbuil.2020.609877

Ahmad, T., & Chen, H. (2019). Nonlinear autoregressive and random forest approaches to forecasting electricity load for utility energy management systems. *Sustainable Cities and Society, 45*, 460–473. https://doi.org/10.1016/j.scs.2018.12.013

Calikus, E., Nowaczyk, S., Sant'Anna, A., Gadd, H., & Werner, S. (2019). A data-driven approach for discovering heat load patterns in district heating. *Applied Energy, 252*, Article 113409. https://doi.org/10.1016/j.apenergy.2019.113409

Ding, Y., Brattebø, H., & Nord, N. (2021). A systematic approach for data analysis and prediction methods for annual energy profiles: An example for school buildings in Norway. *Energy and Buildings, 247*, Article 111160. https://doi.org/10.1016/j.enbuild.2021.111160

Lumbreras, M., Garay, R., Arregi, B., Diarce, G., Martin, K., Hagu, I., & Raud, M. (2022). Data-driven model for heat load prediction in buildings connected to District Heating by using smart heat meters. *Energy, 239*, Article 1223183. https://doi.org/10.1016/j.energy.2021.122318

Klein, S. A., et al. (2017). *TRNSYS 18: a transient system simulation program.* Madison, USA: Solar Energy Laboratory, University of Wisconsin. http://sel.me.wisc.edu/trnsys.

Ding, Y., Timoudas, T. O., Wang, Q., Chen, S., Brattebø, H., & Nord, N. (2022). A study on data-driven hybrid heating load prediction methods in low-temperature district heating: An example for nursing homes in Nordic countries". *Energy Conversion and Management, 269*, Article 116163. https://doi.org/10.1016/j.enconman.2022.116163

Eriksson, M., Akander, J., & Moshfegh, B. (2020). Development and validation of energy signature method – A case study on a multi-family building in Sweden before and after deep renovation. *Energy and Buildings, 210*, Article 109756. https://doi.org/10.1016/j.enbuild.2020.109756

Sha, H., Xu, P., Hu, C., Li, Z., Chen, Y., & Chen, Z. (2019). A simplified HVAC energy prediction method based on degree-day". *Sustainable Cities and Society, 51*, Article 101698. https://doi.org/10.1016/j.scs.2019.101698

Nielsen, H. A., & Madsen, H. (2006). Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings, 38*(1), 63–71. https://doi.org/10.1016/j.enbuild.2005.05.002

Wang, R., Lu, S., & Li, Q. (2019). Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities and Society, 49*, Article 101623. https://doi.org/10.1016/j.scs.2019.101623

T. O. Timoudasa, Y. Ding and Q. Wang, "A novel machine learning approach to predict short-term energy load for future low-temperature district heating", REHVA 14th HVAC Congress, 22nd-25th May, Rotterdam, the Netherlands.

Dagdougui, H., Bagheri, F., Le, H., & Dessaint, L. (2019). Neural network model for short-term and very-short-term load forecasting in district buildings. *Energy and Buildings, 203*, Article 109408. https://doi.org/10.1016/j.enbuild.2019.109408

Sandberg, A., Wallin, F., Li, H., & Azaza, M. (2019). An analyze of long-term hourly district heat demand forecasting of a commercial building using neural networks". *Energy Procedia, 105*, 3784–3790. https://doi.org/10.1016/j.egypro.2017.03.884

Biessmann, F., Golebiowski, J., Rukat, T., Lange, D., & Schmidt, P. (2021). Automated data validation in machine learning systems. *IEEE Data Engineering Bulletin, 44*(1), 51–65.

Breck, E., Polyzotis, N., Roy, S., Whang, S., & Zinkevich, M. (2019). Data validation for machine learning. In *Proceedings of the machine learning and systems*. Stanford, CA, USA: MLSys 2019. mlsys.org.

Gupta, N., Patel, H., Afzal, S., Panwar, N., Mittal, R. S., Guttula, S., Jain, A., Nagalapatti, L., Mehta, S., Hans, S., Lohia, P., Aggarwal, A., & Saha, D. (2021). Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *CoRR*. abs/2108.05935arXiv:2108.05935 https://arxiv.org/abs/2108.05935.

Schelter, S., Bießmann, F., Januschowski, T., Salinas, D., Seufert, S., & Szarvas, G. (2018). On challenges in machine learning model management. *IEEE Data Engineering Bulletin, 41*(4), 5–15. http://sites.computer.org/debull/A18dec/p5.pdf.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software, 10*(1), 1–20.

L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N.S. Noack, H. Patzlaff, H. Harmouch and F. Naumann, "The Effects of Data Quality on Machine Learning Performance", Proceedings of the VLDB Endowment, Vol. 15, No. 11 ISSN 2150-8097.

Li, H., Hong, T., & Sofos, M. (2019). An inverse approach to solving zone air infiltration rate and people count using indoor environmental sensor data. *Energy and Buildings, 198*, 228–242. https://doi.org/10.1016/j.enbuild.2019.06.008

Gram-Hanssen, K. (2013). Efficient technologies or user behaviour, which is the more important when reducing households' energy consumption? *Energy Efficiency, 6*, 447–457.

Gianniou, P., Liu, X., Heller, A., Nielsen, P. S., & Rode, C. (2018). Clustering-based analysis for residential district heating data". *Energy Conversion and Management, 165*, 840–850. https://doi.org/10.1016/j.enconman.2018.03.015

Carbonare, N., Pflug, T., & Wagner, A. (2018). Clustering the occupant behavior in residential buildings: A method comparison. *Modellierung des Nutzerverhaltens In Gebäuden. BAuSIM*.

"Enova Offentlig søk etter energiattester." https://attest.energimerking.no/(accessed May 10, 2021).

Aghabozorgi, S., Shirkhorshidi, AS, & Wah, TY (2015). Time-series clustering–a decade review. *Inf Syst, 53*, 16–38.

Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications, 1*, Article 100001. https://doi.org/10.1016/j.mlwa.2020.100001

Jesper, M., Pag, F., Vajen, K., & Jordan, U. (2021). Annual industrial and commercial heat load profiles: modeling based on k-means clustering and regression analysis. *Energy Conversion and Management: X, 10*, Article 100085. https://doi.org/10.1016/j.ecmx.2021.100085

Zhang, Zheng, Tang, Ping, Huo, Lianzhi, & Zhou, Zengguang (2014). MODIS NDVI time series clustering under dynamic time warping. *International Journal of Wavelets, Multiresolution and Information Processing HYPERLINK, 12*(05). https://doi.org/10.1142/S0219691314610116. https://www.worldscientific.com/toc/ijwmip/12/05

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*, 00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Alam, M. M., Bao, H., Zou, P. X. W., & Sanjaya, J. (2017). Behavior change of building users and energy consumption. *Encyclopedia of Sustainable Technologies*, 189–196. https://doi.org/10.1016/B978-0-12-409548-9.10193-9

Thomas OhlsonTimoudas, Yiyu Ding, Qian Wang. A novel machine learning approach to predict short-term energy load for future low-temperature district heating. REHVA 14th HVAC World Congress 22 –25 May, Rotterdam, The Netherlands. DOI: https://doi.org/10.34641/clima.2022.319.

Deb, C., & Schlueter, A. (2021). Review of data-driven energy modelling techniques for building retrofit. *Renewable and Sustainable Energy Reviews, 144*, Article 110990. https://doi.org/10.1016/j.rser.2021.110990