BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 70 (74), Numărul 3, 2024
Secţia
CONSTRUCŢII DE MAŞINI
DOI:10.2478/bipcm-2024-0015

# OVERVIEW OF DATA-DRIVEN METHODS FOR DISTRICT HEATING SYSTEMS DIAGNOSIS

BY

**ALEXANDRU CEBOTARI and DANIELA POPESCU***

“Gheorghe Asachi” Technical University of Iaşi, Department of Fluid Mechanics, Fluid Machinery and Fluid Power Systems, Iaşi, Romania

Received: December 1, 2024
Accepted for publication: December 16, 2024

**Abstract.** District heating systems are essential for efficient and sustainable urban energy management, offering significant energy savings and environmental benefits. This paper presents some key data-driven methodologies, including advanced data analytics, machine learning, artificial neural networks and other modern methods to evaluate and optimize the design and operation of district heating networks. Several application areas are discussed: demand forecasting, design optimization of the network, fault detection and diagnosis. Recommendations regarding the use of Big Data and AI-driven insights combined with traditional thermal-hydraulic analysis to address challenges such as load variability, energy losses, and operational inefficiencies are formulated. Key challenges and limitations are highlighted, such as data quality and availability, algorithm choice, scalability, etc. The paper aims to provide insights into the potential of data-driven methods to transform classic district heating systems into smarter and sustainable systems towards wide implementation of the 4GDH.

**Keywords:** district heating, data-driven methods, machine learning, demand forecasting, fault detection.

*Corresponding author; *e-mail*: daniela.popescu@academic.tuiasi.ro

## 1. Introduction

District heating systems (DHS) are important in enabling efficient energy utilization and reducing carbon emissions in urban settings (Werner, 2017). As cities grow and energy demands are variable parameters, the need for efficient, adaptable, and intelligent systems has become part of energy policies. The United Nations estimated in its report *"World Population Prospects"* (2024) that in the next 30 years, the total population of the world is expected to grow by 2 billion, while in the "World *Urbanization Prospects*" (2018) report it was projected that urbanization level will increase by 13%, meaning that approximately 68% of world's population will reside in urban areas. In Europe these projections are even higher, being estimated to reach an 83.7% urbanization level by 2050. Considering the European Commission's long-term strategy of ensuring a climate-neutral society by 2050, an efficient DHS might be a pivotal factor in achieving such an ambitious goal.

Traditionally, the design and the optimization of the DHS rely on physics-based deterministic models. These models often struggle with complexities of real-world conditions, including non-linear relationships, dynamic environmental factors and renewable energy integration challenges. Advances in data-driven methodologies have revolutionized the field by proposing new models to address the challenges regarding different applications, such as demand forecasting, optimization, fault detection, etc.

In recent decades, there has been a significant increase in the number of publications focusing on ML methods applied to DHS. In attempt to conduct a systematic literature review, Ntakolia's *et al*. (2021) identified and analyzed a total of 74 relevant articles. The authors proposed to categorize the papers into two major groups: a) heat load/demand prediction and b) design, maintenance and scheduling. Figure 1 clearly illustrates this growing trend, while also highlighting the specific sub-areas of DHS where ML methods were applied.

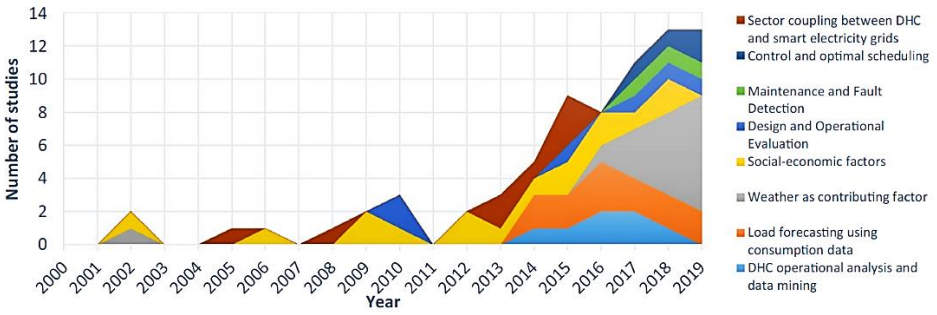


Fig. 1 – Trends in publications related to "ML applied to DHS" topic (Ntakolia *et al*., 2021).

Next sections of this paper present a review of a few key papers in the domain, exemplifying different application areas and approaches to be used for the appropriate design and control of the Fourth Generation District Heating Systems (4GDH).

## 2. Demand Forecasting Methods

Demand forecasting refers to the process of predicting the future energy requirements of consumers supplied from the DHS with thermal energy, the so-called thermal load.

Currently, it is one of the most relevant research topics in the field, as its importance cannot be underestimated. The precise forecasts can have multiple positive implications on the design and optimization of the system:

- Operational efficiency – reduce energy losses and fuel consumption;
- Cost management – help avoiding overproduction and associated costs;
- Customer experience – help meeting better customer's needs;
- Environmental sustainability – reduce gas emissions by minimizing excess production.

Forecasting the heat demand is inherently complex task due to several challenges as:

- Weather dependency – heat demand is highly sensitive to external factors (temperature, humidity, wind speed, etc.);
- Consumer behavior – variability in consumer's patters adds uncertainty;
- Urban particularities – mix of new/old buildings, uneven density distribution (disparity).

Broadly, thermal load forecasting methods can be classified in two main types: forward methods and data-driven methods (Sakkas and Abang, 2021).

Forward methods – known as physics-based approaches, simulate the thermal behavior of a system by mathematically describing the underlying physical principles, heat transfer, thermodynamics, system-specific characteristics, etc. (Fumo, 2014). These methods often require the detailed description of input parameters: building envelope properties, insulation levels, weather condition, etc. The main advantages of using forward methods are that the models can be easily interpreted and that they can be used even whether the input data are limited or incomplete. However, they suffer from several limitations, having a limited adaptability to dynamic, non-linear factors, such as weather conditions, user behavior, infrastructure ageing, etc. Moreover, these

models are considered to have a static nature, as they cannot adapt to the anomalies in the system, and it is difficult to improve them over time.

Data-driven methods – these methods predict the output by means of regression models applied to historical data collected by monitoring systems of the DHS. In this case, the model aims to identify the best-fit function that maps the provided input data to the observed output. Two sub-types of data-driven methods can be distinguished in the scientific literature, the *statistical* and the *machine learning* (ML) methodologies. For the statistical methodologies, the accuracy of the outputs is usually predetermined by the choice of regression model, while ML methodologies can describe complex models because they are learned adaptively by the algorithm itself (Geysen *et al*., 2017). Data-driven methods have been preferred in the last years, since they are based on a modern and transformative approach. In the following, results of some works are presented.

Song *et al*. (2021) proposed a novel approach for predicting hourly thermal load in DHS. The proposed architecture for the prediction model is based on the combination between Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) algorithm. CNNs excel at parallel processing and are able to deal with nonlinear features. Applications regarding DHS can identify the spatial characteristics, more exactly, temperature and pressure. LSTM represents a type of recurrent neural network, able to handle time series data. Since LSTM can remember the information for long periods, it can model better heat load changes over time, capturing temporal factors, such as weather or user demand. This allowed the model to cope with difficult to model phenomena, as the thermal inertia delay. By combining the two above mentioned algorithms into one CNN-LSTM model led to better understanding of spatial and temporal aspect of heating load data. A total of 9 input features were selected, the first 8 being measured in real-time and stored in a database:

- Supply primary temperature and pressure
- Return primary temperature and pressure
- Supply secondary temperature and pressure
- Return secondary temperature and pressure
- Outdoor temperature

The model has been evaluated using several performance metrics, the corresponding results for optimal case are centralized in Table 1. Furthermore, the authors compared CNN-LSTM performance against other more simplistic algorithms, such as Support Vector Machines, Extra Tree Regression, Random Forest Regression, Gradient Boosting Regression, etc. The accuracy of hybrid model was found to be very good, as results presented in the Table 1 are encouraging.

**Table 1**
*Performance metrics for CNN-LSTM prediction algorithm (Song's et al., 2021)*

| Performance metric | Result |
|---|---|
| Root mean square error (RMSE) | 0.026 |
| Coefficient of variation of root mean square error (CVRMSE) | 0.050 |
| Mean absolute error (MAE) | 0.011 |
| Mean absolute percentage error (MAPE) | 3.6% |

Liu *et al.* (2021) studied the possibility to predict the energy consumption of buildings considering a range of buildings with different types of envelopes. The study is based on a real-world case: specifically, best design for a new University. A total number of 6 design parameters (see Table 2*)* have been considered for modelling the thermal characteristics of the building envelope. The adopted methodology represents a hybrid approach. Initially, the building information was modelled using *Revit* software (Autodesk, 2024) and the result is imported into *DesignBuilder* software (DesignBuilder, 2024) for energy consumption analysis. Later, by performing testing and consumption simulation models a dataset has been generated based on the design parameters, and used as input for a ML-based algorithm. To accurately predict the energy consumption, the correlation and impact of each design parameter was analyzed. Random Forests (RF) algorithm was chosen as the main model, and Pearson function for evaluation of correlations. The model performance has been compared to predictions of other methods, such as ANN and Support Vector Machines (SVM). The RF model demonstrated best results. The study concluded that the parameters heat transfer coefficient of exterior walls (A1), heat transfer coefficient of outer windows (A5) and window-to-wall ratio (A6) have the strongest influence on the building energy efficiency performance. The outcome of the paper can have dual applicability, by offering insights for the design decision process or upgrades to existing buildings, and by proposing a novel prediction methodology.

**Table 2**
*Design parameters of building envelope (Liu et al., 2021)*

| Nr. | Parameter name |
|---|---|
| *A1* | Heat transfer coefficient of exterior walls |
| *A2* | Solar radiation absorption coefficient of exterior walls |
| *A3* | Heat transfer coefficient of the roof |
| *A4* | Solar radiation absorption coefficient of the roof |
| *A5* | Heat transfer coefficient of outer windows |
| *A6* | Window-to-wall ratio |

## 3. Fault Detection and Diagnosis Methods

Fault diagnosis and detection (FDD) represents an important aspect for the efficient operation and the reliability of DHS's. The role of FDD is to mitigate the negative impacts of operational faults, such as increased return water temperatures, energy losses, reduced hydraulic capacity, and customer discomfort. Faults in substations, pipes, or heat exchangers compromise the overall efficiency of the system, increase the operational costs and the environmental emissions.

Traditionally, field inspections and monitoring technics are employed to detect faults in the infrastructure, after they occur. Such methods are time-consuming, prone to human error and unsuitable for large-scale and dynamic nature of modern networks. As DH networks grew in complexity, incorporating multiple substations and serving thousands of customers, the transition towards digitized DHS was done, which implies implementation of smart meters, extensive sensors coverage and advanced monitoring technologies, specific to the third and fourth generation of DHS.

Palasz and Przysowa (2019) studied the potential of data-driven methods for a very specific and common issue – the failure of heat meters. The study analyzed an extensive 10-year dataset, which included detailed information about heat meter's installation, operation and replacement. After initial statistical analysis, the authors trimmed the dataset, choosing a total of 16 features (parameters) as being relevant for the study (see Table 3).

**Table 3**

*Features used for ML modelling (Palasz and Przysowa, 2019)*

| Feature Name | Feature Description |
|---|---|
| *Age* | Life of meter in months from the moment of installation |
| *ZIP* | Postal code of flat/office in which the meter is installed |
| *Floor* | Floor where the meter is installed |
| *Flat Type* | Type of usable area |
| *Room Type* | Type of area |
| *Acc Consumption* | Consumption from the moment of registration |
| *Current Consumption* | Consumption in the last settlement period |
| *Comm Type* | Type of communication with the meter |
| *Producer* | Name of meter's manufacturer |
| *Rating Factor* | Equalization factor |
| *Billing No* | Next number on settlement period |
| *Current Value* | Last recorded value of the meter |
| *Avg Consumption* | Average consumption in all settlement periods |
| *Max Consumption* | Maximum consumption in the settlement period |
| *Min Consumption* | Minimum consumption in the settlement period |
| *Calculated Age* | Calculated age of the meter in months |

The study found that the risk of monitoring device failures, usually follows Weibull probability distribution. A particularly interesting observation derived from Exploratory Data Analysis (EDA): the intensity of heat meters failures showed only a weak dependence on the *usage time* parameter, meaning that failures occur mostly as external random events. This led to the assumption that the state variables can be sufficient for prediction. A total number of three algorithms were used for ML modeling: ANN, SVM with Radial Basis Function (RBF), and Bagging Decision Trees (BDT). The performance of the models has been evaluated with two metrics: Area Under ROC Curve (AUC) and Matthews Correlation Coefficient. Additionally, other standard metrics as accuracy, precision, recall and $f_1$ score were analyzed. The individual comparison of algorithms revealed the BDT as being the most accurate. However, all 3 models showcased the same deficiency: being good at predicting meter's survival (True Negative case) and bad at detecting meter's failure (True Positive case). This behavior has been slightly improved through hyperparameter optimization, more specifically by applying Sequential Model-Based Optimization model and choosing AUC metric as objective function, which improved the models by 3-5%. Finally, Palasz *et al.* developed an Ensemble Model, which is essentially a meta-classifier that combines the 3 independent models mentioned earlier into a single model. Such classifier demonstrated even better results than individual fine-tuned models, claimed to achieve >95% when evaluated at AUC metric.

Wang's *et al.* (2021) paper explored the application of ML techniques to improve fault detection in a DHS from China. The proposed model detects faults (e.g., sensor faults, actuator faults, and element faults) based on deviations in flow rate, pressure, and temperature data. The methodology represents a combination of *Regression analysis*, augmented with *Exponential Smoothing Step Average* method, which improves prediction accuracy and minimizes noise in fault detection. $R^2$ multi-score analysis was used for evaluating the regression, the model showing an impressive $R^2$ score of 0.96 for specific fault scenarios. In addition to this, the authors applied SVM to classify the states as either normal or faulty. Such classifier demonstrated promising results, being evaluated at precision, recall and F1 scores (see Table 4).

**Table 4**

*Precision value, recall, F1-measure for Wang's et al. model*

| Faults | Precision | Recall | F1-measure |
|---|---|---|---|
| *Sensor faults* | *0.98* | *0.78* | *0.92* |
| *Actuator faults* | *0.77* | *0.82* | *0.96* |
| *Elements faults* | *0.81* | *0.88* | *0.89* |

## 4. Design Optimisation Methods

One key area where design optimization methods of DH networks can be important is the reduction of *heat losses*, which has been a challenge ever since DHS were built. The range of heat losses is influenced by several interdependent factors, ranging from technical, environmental and operational factors:

- Pipeline characteristics:
  - Insulation quality
  - Pipe diameter
  - Pipe material
- Operating conditions:
  - Operating temperature
  - Flow rate
  - System pressure
- Network layout and ageing:
  - Network length
  - Depth of burial
  - Age of infrastructure
- Environmental conditions:
  - Ambient temperature
  - Soil properties
  - Wind exposure

As can be observed from the above list, the phenomena of heat losses exhibit a complicated relationship between multiple factors. At the design phase, the traditional approach of thermal losses estimation relied mainly on predefined standards and guidelines, such as EN 13941 in Europe, that provides some empirical loss coefficients, standard tables and charts for typical designs. For the existing systems, the losses are usually estimated by the analysis of the energy balance, by simple comparing the quantity of thermal energy delivered by the power plant ($Q_{input}$) to the thermal energy delivered to end-users ($Q_{output}$)

$$Q_{loss} = Q_{input} - Q_{output} \quad (1)$$

Obviously, these approaches lack accuracy and flexibility to capture the particularities of each individual DHS and cannot provide information on the exact localization of the losses in the network. The topic was extensively investigated lately, when historical and real-time operational data became available.

Chen *et al.* (2022) studied the potential of hybrid models applied in the field of DH networks design optimization, considering both economical and operational aspects. A novel optimization model was proposed, able to optimize the dimensioning of pipe diameter and insulation thickness, with the final goal of

lowering heat losses in the network, while offering the most economical setup. The initial theoretical heat-transfer model has been used as a basis for heat losses estimation, thus augmented with ANN model based on measured data. This hybrid model allowed to reduce the differences between theoretical and measured heat losses. The application has been validated using the measured heat loss data from 3 different DHS located in Zhejiang, China.

## 5. Challenges and Limitations

Although data-driven methods were massively adopted in recent years by the researchers, the underlying challenges and limitations need to be mentioned.

Perhaps the most common challenge is data availability and quality, which is an essential aspect for acceptable performance. Normally, DHS-related datasets are not available in open-source and establishing industry collaborations is not always easy, as commercial secret or data privacy reasons may be invoked. However, some efforts were made in this regard lately, for example the project "*AI for Failure Detection and Heat Demand and Production Forecast*", fully-funded by International Energy Agency (IEA DHC, 2023), which made datasets available on a public platform (Kaggle). Thus, the new opportunities for generation of synthetic (augmented) datasets are open (Vallee *et al*., 2023).

Concerning poor bad data quality, the reasons can be multiple: poorly calibrated sensors, entry errors, communication failures, etc. Addressing data quality challenge requires robust data preprocessing techniques, such as outlier detection, data interpolation, and validation mechanisms.

Apart from that, choosing a specific data-driven algorithm is always a difficult task, due to the wide range of available options and the diversity of case studies. Typically, the main algorithm is selected according to the recommendations from the scientific literature, and subsequently tested against other potential alternatives. The state of the art presents combined methodologies based on several algorithms or on hybrid models (forward + data-driven methods).

Further work on the topic of design and optimization of DHS should have in view several issues: the lack of real-world data, the simulation of the non-linear characteristic of the data used for heat load forecasting, biased input data, as dataset came from a very specific climatic zone.

## 6. Conclusions

Data-driven methods are highly recommended for the analysis of DHS. Given the objective factors as the digitalization of DHS, the growing amount of data, advancements in ML techniques, increasing computing power and stringent climatic policies, the relevance of data-driven methods is expected to grow even further.

Successful adoption of data-driven solutions in DHS depends not only on the effectiveness of algorithms, but also on factors such as: integration with existing infrastructure, deployment and computational costs, type of computational resource (lightweight localized application vs cloud-based platform for large-scale system), desired response time (real-time vs offline computation), etc.

Additionally, we would like to emphasize a point of view that is highlighted by many other researchers, e.g. the work of Mbiydzenyuy *et al*. (2021), namely the advantages for industry players to make benchmarking datasets available in open-source. Besides, establishing a stronger collaboration between network operators and academia will help aligning better the research directions and real-world challenges faced by the industry.

Overall, we consider that data-driven methods have potential to become a paradigm-changer and a catalyst for transitioning towards 4GDH.

## REFERENCES

Autodesk Inc., https://www.autodesk.com/products/revit, Retrieved Dec 2024.

Chen K., Hu J., Yu L., Zheng M., Sun S., He D., Lin J., *A Data-driven Model of Pipe Diameter and Insulation Thickness Optimization for District Heating Systems*, Journal of Physics: Conference Series, Volume 2166, International Conference on Frontiers of Electrical Power & Energy Systems, 2022.

DesignBuilder Software Ltd, https://designbuilder.co.uk, Retrieved Nov 2024.

European Commission, *2050 Long-Term Strategy,* (https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en), Retrieved Nov 2024.

Fumo N., *A review on the basics of building energy estimation*, Renewable and Sustainable Energy Reviews, Volume 31, 2014, Pages 53-60.

Geysen D, De Somer O., Johansson C., Brage J., Vanhoudt D., *Operational thermal load forecasting in district heating networks using machine learning and expert advice*, Energy and Buildings, Volume 162, 2018, Pages 144-153.

IEA DHC, *Annex XIII Project 03, 2023* (https://www.iea-dhc.org/the-research/annexes/annex-xiii/annex-xiii-project-03)

Kaggle, dataset: *Fault Detection and Diagnosis in District Heating*, (https://www.kaggle.com/datasets/mathieuvallee/ai-dhc/data)

Liu Y., Chen H., Zhang L., Feng Z., *Enhancing building energy efficiency using a random forest model: A hybrid prediction approach*, Energy Reports, Volume 7, 2021, Pages 5003-5012.

Mbiydzenyuy G., Nowaczyk S., Knuttson H., Vanhoudt D., Brage J., Calikus E., *Opportunities for Machine Learning in District Heating*, Applied Sciences. 2021; 11(13):6112.

Ntakolia C., Anagnostis A., Moustakidis S. *et al*., *Machine learning applied on the district heating and cooling sector: a review*, Energy Systems, 2022; Volume 13, Pages 1-30.

Palasz P., Przysowa R., *Using Different ML Algorithms and Hyperparameter Optimization to Predict Heat Meters' Failures*, Applied Sciences. 2019; 9(18):3719.

Sakkas N.P., Abang R., *Thermal load prediction of communal district heating systems by applying data-driven machine learning methods*, Energy Reports, Volume 8, 2022, Pages 1883-1895.

Song J., Zhang L., Xue G., Ma Y., Gao S., Jiang Q., *Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model*, Energy and Buildings, Volume 243, 2021, 110998.

United Nations, *World Urbanization Prospects*, 2018, (https://www.un.org/development/desa/pd/news/world-urbanization-prospects-2018)

United Nations, *World Population Prospects*, 2024, (https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2024_wpp_2024_advance_unedited_0.pdf).

Vallee M., Wissocq T., Gaoua Y., Lamaison N., *Generation and evaluation of a synthetic dataset to improve fault detection in district heating and cooling systems*, Energy, Volume 283, 2023, 128387.

Wang P., Poovendran P., Manokaran K.B., *Fault detection and control in integrated energy system using machine learning*, Sustainable Energy Technologies and Assessments, Volume 47, 101366.

Werner S., *International review of district heating and cooling*, Energy, Volume 137, 2017, Pages 617-631.

## ANALIZA SISTEMELOR DE TERMOFICARE PRIN METODE BAZATE PE DATE DIN SISTEMELE DE MONITORIZARE

(Rezumat)

Sistemele de termoficare sunt esențiale pentru o gestionare eficientă și durabilă a energiei, oferind economii semnificative de energie și beneficii pentru mediu, cu precădere în zone urbane. Acest articol analizează principalele metodologii bazate pe analiza datelor, inclusiv analize avansate de date, învățare automată, rețele neuronale artificiale, și alte metode moderne de evaluare și optimizare a rețelelor de termoficare din cadrul SACET (sisteme de alimentare centralizată cu energie termică). Sunt prezentate mai multe domenii de aplicare ale acestor metode: prognoza cererii de căldură, optimizarea rețelei, detectarea defecțiunilor și mentenanța predictivă. În plus, se evidențiază modul în care metodele bazate pe volume mari de date și IA, combinate cu analiza termică-hidraulică tradițională, pot aborda provocări precum variabilitatea sarcinii, pierderile de energie și ineficiențele operaționale. În final, sunt discutate principalele provocări și limitări, precum calitatea și disponibilitatea datelor, alegerea algoritmilor sau scalabilitatea. Articolul își propune să ofere perspective asupra potențialului oferit de metodele bazate pe date de a transforma sistemele de transport și distribuție a căldurii în sisteme inteligente și sustenabile.