Available online at www.sciencedirect.com
ScienceDirect
Energy Reports 7 (2021) 380–388
www.elsevier.com/locate/egyr

The International Symposium on District Heating and Cooling, Nottingham Trent University, 17th Edition, 06–09 September 2021, Nottingham, UK

Heat demand prediction: A real-life data model vs simulated data model comparison

Kevin Naik*, Anton Ianakiev[1]

Nottingham Trent University, Nottingham, UK

Received 27 July 2021; accepted 7 August 2021

Abstract

In the recent years machine learning algorithms have developed further and various applications are taking advantage of this advancement. Modern machine learning is now used in district heating for more precise and realistic heat demand prediction. Machine learning methods like Artificial Neural Network (ANN), Linear Regression (LR), and Decision Tree (DT) are commonly adopted in heat demand prediction to produce more accurate results. This research paper compares the performance of several machine learning methods on different datasets generated by the combination of simulations and real-life data collected from a local district heating site in Nottingham. The result shows that Linear Regression generates better prediction than Artificial Neural Network and Decision Tree, for dataset generated using simulator, whereas Decision Tree performs best for real-life data.

Peer-review under responsibility of the scientific committee of the International Symposium on District Heating and Cooling, Nottingham Trent University, 2021.

Keywords: Heat demand; Prediction; Machine learning; District heating; Data models

1. Introduction

District heating (DH) system plays vital role to achieve low carbon emission by 2050. The UK government with private sector is supporting the development of District Heating by investing GBP 1351M [1]. In 2018, the report recorded 14,000 heat networks in UK, out of which only 2000 heat networks are DH. Moreover, estimated turnover from heat networks is around 300 million pound per annum. Additionally, over the next decade a greater number of heat networks are expected to be installed [2].

An overview of the DH systems shows that increasingly more investment is made towards research and development to make DH systems more efficient and sustainable. By 2050, half of the buildings in UK are targeted to be connected to District Heating Networks. Moreover, the cost evaluation of DH vs traditional heating shows

* Corresponding author.
E-mail addresses: kevin.naik@ntu.ac.uk (K. Naik), anton.ianakiev@ntu.ac.uk (A. Ianakiev).
[1] Professor in Sustainable Energy Systems.

https://doi.org/10.1016/j.egyr.2021.08.093

Peer-review under responsibility of the scientific committee of the International Symposium on District Heating and Cooling, Nottingham Trent University, 2021.

saving of 4 pence per kWh per customer. Use of DH has shown reduction in CO2 emission from 240 kg/MWh to 60 kg/MWh [3]. As DH has potential to grow in next 30 years, the scope for technological enhancement is significant.

This paper is relevant mainly to 4th Generation Low Temperature District Heating (LTDH), which is the most efficient DH system, but at the same time it requires precise control of the heat generation and for that reason accurate and realistic heat demand prediction is essential. LTDH systems can achieve a higher degree of independence by using renewable energy sources and heat storages. Due to the lower feeding temperature in LTDH, sources of heat supply can be diversified by using renewables and recycling of local excess heat. The application of LTDH will require realistic and more accurate designs with respect to heat demand, costs, and operating conditions. It will also require heating systems with low temperature demands, no short-circuit flows in the distribution networks.

The application of different machine learning algorithms for heat demand prediction is presented in various literature sources and it illustrates the importance of this approach. Three most used machine learning algorithms are Artificial Neural Network, Decision Tree and Linear Regression. The paper compares the performance of several machine learning algorithms applied to two different datasets [4].

2. Background and related work

The modern heat demand prediction is more than 2 decades old. The use of statistical method as a kind of machine learning model for prediction of loads in district-heating systems is first described in the Simple model based on social behavior and outdoor temperature by Dotzauer [4]. Many researchers started to explore heat prediction for district heating using time-series with the inclusion of outdoor temperature [5].

After 2009, the increase of computational power made the application of machine learning algorithm a real possibility. Various researchers started using the machine learning algorithms in District heating for heat demand prediction [6,7].

Talebi et al. [8] evaluates various heat prediction techniques used in the past. It describes the complexity level by four parameters viz. number of technologies utilised, number of users, temporal profile, and spatial concerns. The paper also classifies district heating based on geographical conditions, scale, heat density and end-user demand. Various predictive models have been used, namely regression, artificial intelligence algorithm (ANN) and fuzzy Artificial Neural Network. The Artificial Neural Network is more widely used for load prediction compared to other methods. Also, support vector machine (SVM) is used where small datasets are available.

In the last five years, two branches of artificial intelligence have dominated, machine learning and deep learning. Both methods rely on the data size; more data available means the algorithms would perform better. Different research groups were keen on exploring the forward and the data-driven models for heat prediction. Generally heating depends on many factors like the type of building and its physical condition, etc. for which complex energy simulation programs have been written. The forward simulation models are very well developed and are replicated in software programs like EnergyPlus and TRNSYS [8]. The data-driven models use statistical methods to predict thermal/heat load. The operational thermal load forecasting was carried out mainly using machine learning algorithms viz. Linear Regression, Artificial Neural Networks, support vector machine and extremely randomised trees regressor, were applied on data from various district heating system in Sweden [9]. Similar data-driven models are used for short-term load forecasting of the solar community in [10]. Few literature sources demonstrate the use of deep learning models for load forecasting of demand [11]. The deep learning techniques used are, polynomial Linear Regression, Ridge Regressor, Lasso Regressor and Deep Neural net. Suryanarayana et al. investigated in [12] heat prediction investigated Long Short-Term Memory (LSTM) model, and Feature Fusion Long Short-Term Memory (FFLSTM). It was compared with other models like back propagation, support vector regression, regression tree, random forest regression, gradient boosting regression, and extra trees regression [12–14].

The literature review shows the importance of heat prediction and data-driven models [15]. It can be concluded that Linear Regression, Decision Tree and Artificial Neural Network are the most popular methods used in heat prediction. In this paper these three methods would be used on the different datasets for heat prediction.

3. Datasets

The datasets discussed in this section, are from two different projects: the SHARING CITIES project and the REMOURBAN project. Both datasets are from residential homes. The SHARING CITIES project aims to develop

open-source smart city solution. The SHARING CITIES project data are generated by using EnergyPlus and Ptolemy II energy simulation software based on building characteristics and environmental variables. The data generated by the simulators are from 95 homes located in Ernest State, London. The data generated consist of 3 variables that are: timestamp, outdoor temperature, and heat demand for each home. The data generated is for 3 years at frequency of 15 min.

REMOURBAN project aims to leverage Information Communication Technology, Energy, Society, Mobility and Sustainability. The data acquisition system is setup in 30 Homes for a period of 1 year at a frequency of 1 min in Sneinton, Nottingham, UK. Parallel to data acquisition , Weather station was also setup and weather data were collected for the same period at a frequency of 5 min.

For this paper, to compare the data, both datasets are converted to frequency of 15 min. Moreover, aggregate heat demand is used for prediction. The data in the London dataset was generated from the simulator as mentioned above whereas the data in the REMOURBAN dataset is real-life data. Each dataset shows high correlation with outdoor temperature. The same conclusion can be drawn from the scatter plots of the overall dataset, are shown in Fig. 1 (dataset 1) and Fig. 2 (dataset 2).

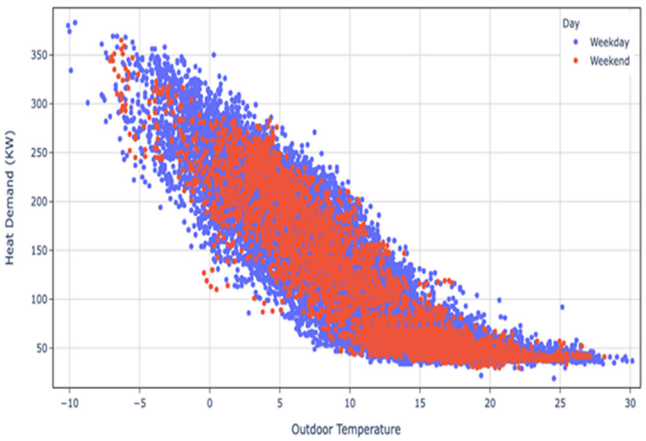

Fig. 1. Representation of Simulated data Heat Demand vs Outdoor Temperature.

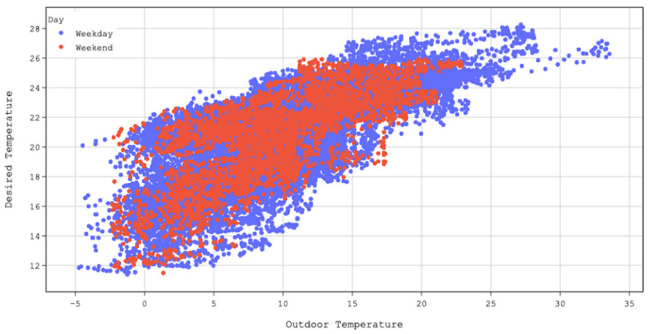

Fig. 2. Representation of collected data Desired Temperature vs Outdoor Temperature.

Before moving to the analysis section, it is important to understand the features used for developing heat demand prediction. An important feature which is converted from single to multiple is timestamp. Timestamp is converted to hour, minute, day, weekday, weekend, and month. Apart from these the SHARING CITIES project has features like outdoor temperature and heat demand. In REMOURBAN project the features apart from timestamp are temperature, humidity, pressure, sunset time, sunrise time, weather type, and wind speed. The features and their values are summarized in Table 1.

Table 1. Features for building model.

Project	Feature	Value units
REMOURBAN & SHARING CITIES project	Hour	0 to 23
	Minute	15, 30, 45, 60
	Day	1 to 31
	Weekday	Monday to Sunday
	Month	1 to 12
	Outdoor Temperature	− to + in degrees
REMOURBAN	Heat Demand	KW
	Pressure	hPa
	Humidity	0 to 100%
	Wind Speed	m/s
	Sunset time	In seconds
	Sunrise time	In seconds
	Weather Type	Clear, Clouds, Drizzle, Fog, Haze, Mist, Rain, Snow, Thunderstorm, Tornado

4. Analysis

In this section, the analysis of data collected is described as well as its use is explained. After, the data are collected or generated, the first step is to clean the data. The outliers are removed from both datasets. The simulated dataset for a period of 3 years is divided into a dataset of the first 2 years data for training the models and the second dataset contains 1-year data for testing the models. The real-life dataset (REMOURBAN project), the weather data and the data from the monitored residential homes are stitched together using timestamp. It is then randomly divided into 2/3 data for training models and 1/3 data are used for testing. Once the datasets are split three models i.e. Linear Regression, Decision Tree and Artificial Neural Network are defined. For comparison, the configuration of all models is kept the same as hidden layer, solver, iteration, leaf size etc. Before training the models to avoid overfitting and under-fitting of data the correlation matrix for feature selection is determined. The training dataset is used to build the models and then test dataset is used to predict the value.

Figs. 3 and 4 are a graphical representation of the correlations calculated for both datasets. The values represented are in percentage, so the range is from −100 to +100, the sign '+' or '−' shows whether the correlation is positive or negative. Also, the float is ignored for clear representation. From Fig. 4 it can be seen that, the correlation value is low for minute, day, and weekday therefore these features can be dropped from the building model. The correlation shows that outdoor temperature has a major impact on heat demand. Month shows the seasonal impact on heat demand. Hour and weekday are not dropped even though the correlation percentage is small because weekday or weekend result in the consumption changes which can be seen in Fig. 1. Moreover, Hour allows us to understand the working hours and non-working hours.

Similar analysis is carried out on REMOURBAN project data (Fig. 3). The percentage of correlation hour, minute, pressure, weed speed and weekday is close to zero. But as reasoned before, hour and weekday are not dropped. Other variables show the correlations which can be used in Decision Tree and Artificial Neural Network. The performance of models is evaluated using different methods: for Linear Regression R2 Score is used and for other methods confusion matrix is used.

Figs. 5 and 6 shows the overall prediction for SHARING CITIES and REMOURBAN projects using Linear Regression, Artificial Neural Network and Decision Tree. In REMOURBAN project the data are randomly divided in a ratio of 70:30 for training and testing model. So, the x-axis is not timestamp but is a number. In the SHARING CITIES project prediction plot x-axis is timestamp. For the REMOURBAN project the plot trend is not evident but zoomed analysis is discussed in the result section. Overview of the SHARING CITIES plot (Fig. 6) shows that trend of actual heat demand is followed by the output of prediction algorithms.

5. Results

The correlation analysis shows that a greater number of variables can be used to train the model for REMOURBAN project than for the SHARING CITIES project because the generated and collected data are from two different

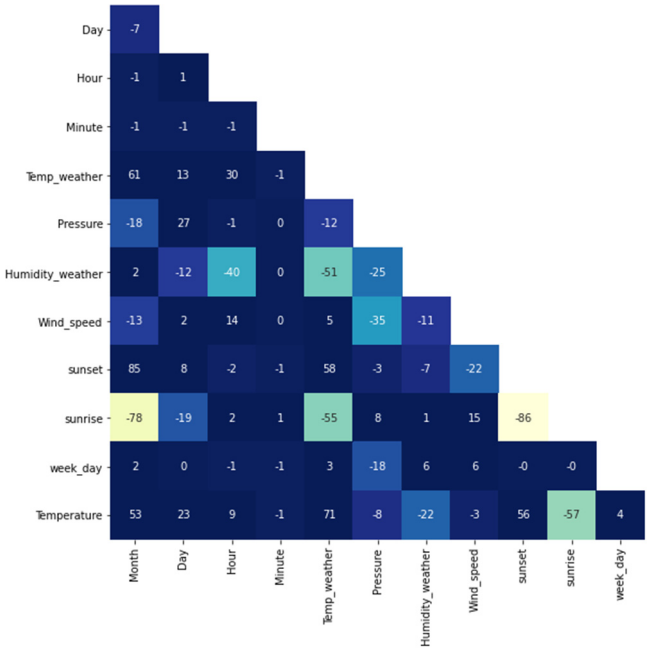

Fig. 3. Correlation Matrix representation of REMOURBAN Project.

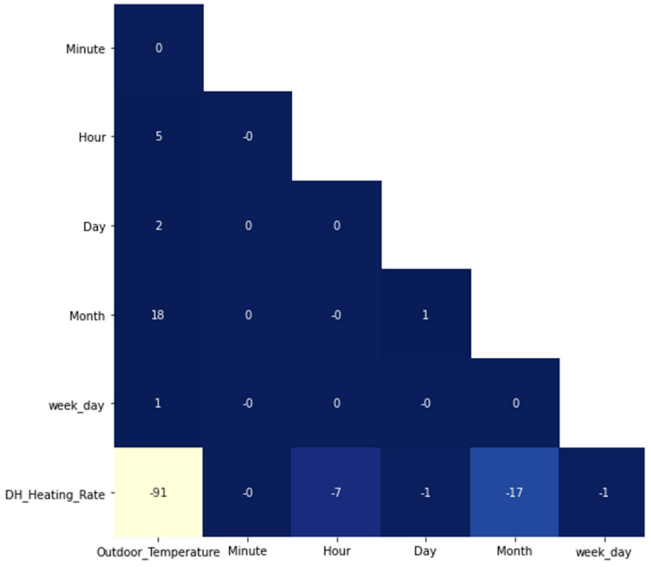

Fig. 4. Correlation Matrix representation of Simulated Data.

processes. Eight features are used for training the models of the REMOURBAN project whereas only 4 features are used for SHARING CITIES project. The advantage of having a greater number of features can be leveraged by using machine learning algorithms like Decision Tree and Artificial Neural Network. The disadvantage of features with high correlation is that they will overpower the models by having more influence, or in other terms they will generate biased prediction.

The reason for Linear Regression working very well on simulated data is that data generation is based on the physical characteristics of building and weather data. The advantage of real-life customer behaviour is that they can reflect customer behavior, which is very important factor to consider, and it is reflected in the heat consumption

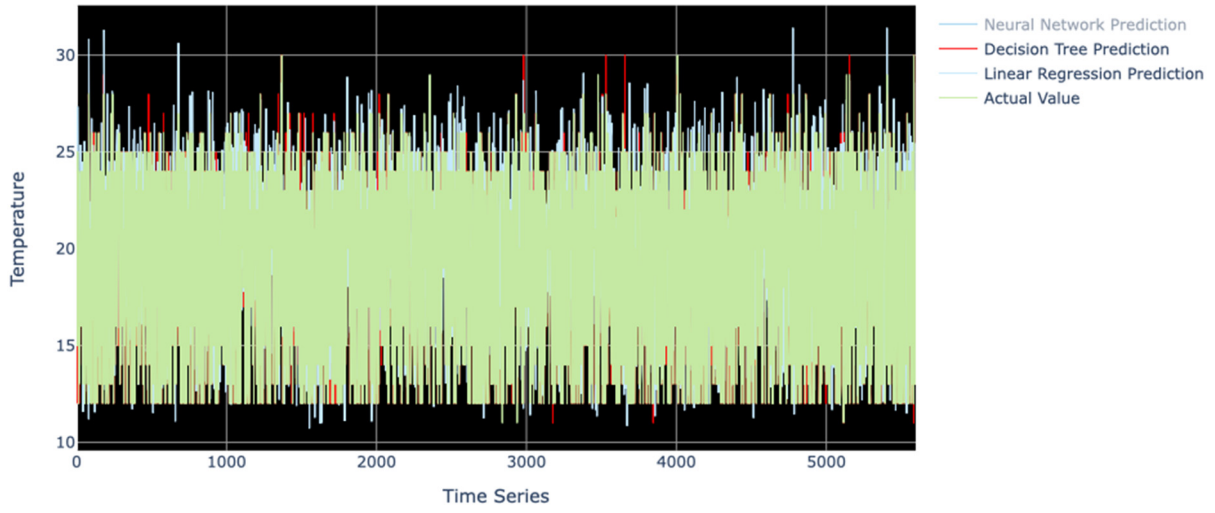

Fig. 5. Prediction of REMOURBAN Project.

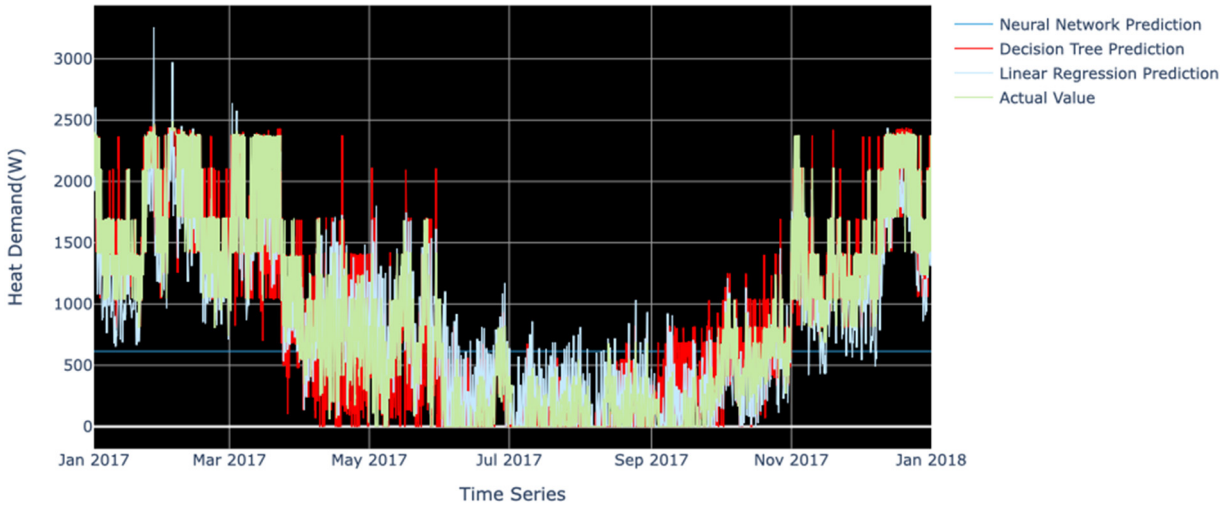

Fig. 6. Prediction of Sharing Cities.

pattern. This brings a bit of complexity in the analysis as customer behavior is dependent on non-tangible factors like mood, affordability etc. However, over a long period the consumption pattern would be consistent.

From Table 2, the conclusion is that the high R2 Score of the Linear Regression 81.53% means that the test dataset and the predicted values for it are close when compared to each point by 81.53%. The accuracy of Decision Tree and Artificial Neural Network is less than 25%. So, for the SHARING CITIES project the best performance algorithm out of the three is Linear Regression. However, for the REMOURBAN project the best accuracy is from Decision Tree which is 87.22% a much better performance compared to Artificial Neural Network and Linear Regression.

The factor explaining the result achieved for the SHARING CITIES dataset is the 91% correlation with outdoor temperature, which means one dominant feature is driving the prediction. There are not enough features available for Artificial Neural Network to develop learning pattern and improve the performance. Linear Regression works better for one of the datasets but not for the other one. It shows that there cannot be one method of heat prediction suitable for all the district heating network. Moreover, the data simulated, and the real-life data collected have different parameters affecting performance. Decision Tree outperforms in terms of accuracy in the REMOURBAN

Table 2. Summary of result comparison model.

Models	SHARING CITIES project	REMOURBAN project	Evaluation
Linear regression	81.53%	54.77%	R score
Decision tree	24.75%	87.22%	Accuracy
Artificial neural network	21.00%	73.12%	Accuracy

dataset but shows low level of accuracy for SHARING CITIES project due to the smaller number of features and the output is dependent on the dominant feature—the outdoor temperature. Artificial Neural Network is showing also low level of accuracy for SHARING CITIES project also due to the smaller number of features and after 500 iterations the algorithm is not able to converge, whereas the bigger number of features have different effect in the case of REMOURBAN project. In the REMOURBAN project the Artificial Neural Network converges before 100 iterations and shows decent accuracy though it will improve if amount of collected data continues to grow. Fig. 7 compares the prediction of Linear Regression and Decision Tree with actual value. Decision Tree prediction is very close to the actual value highlighted in red for the REMOURBAN data. Fig. 8 shows comparison between Linear Regression prediction and actual heat demand for SHARING CITIES. The prediction is close to actual value and two peaks per day can be seen.

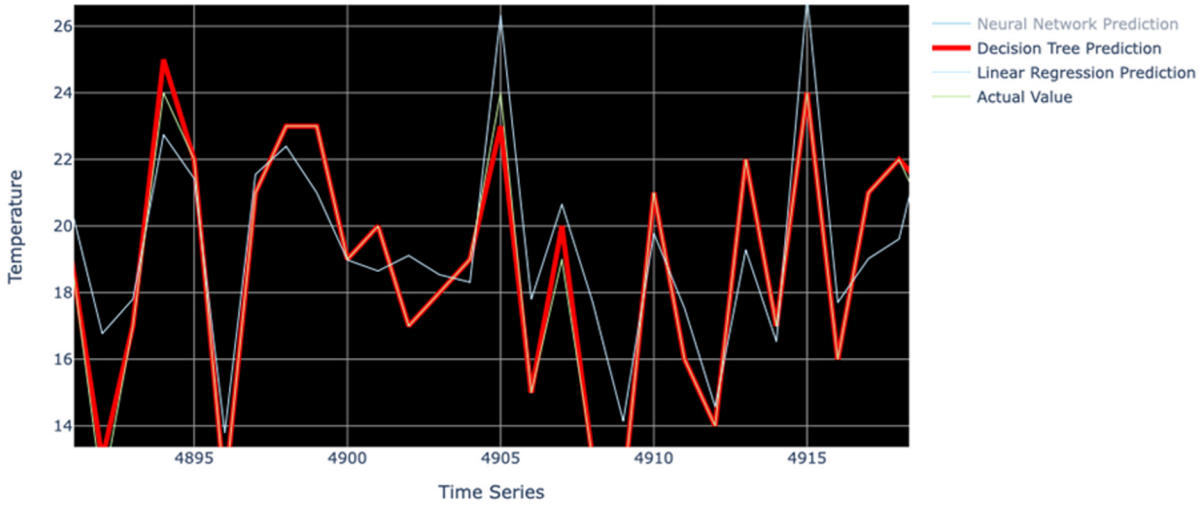

Fig. 7. Close look at prediction of REMOURBAN Project. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

The 4th Generation Low Temperature District Heating (LTDH) is the most efficient but at the same time it requires precise control of heat generation and for that reason accurate and realistic heat demand prediction is essential. The heat demand prediction of district heating system using data-driven approach that has been presented in this paper is aimed to be used for LTDH systems, but older generation district heating systems will also benefit. The heat demand prediction for residential building of data generated and data collected are studied. The results show that for simulated data and data collected, different algorithms work better. Linear Regression performs well compared to Decision Tree and Artificial Neural Network in case of simulated data. For real-life data collected, the Decision Tree outperforms Artificial Neural Network and Linear Regression. The outdoor temperature shows major impact on heat demand prediction. The different weather features can be used to improve the performance of the algorithms and predict heat demand. Moreover, Artificial Neural Network which has shown good results in various literature, has shown disappointing results in the case of simulated data (SHARING CITIES) as the amount of data collected is small and the number of features is also small for the algorithm to develop its strength. The

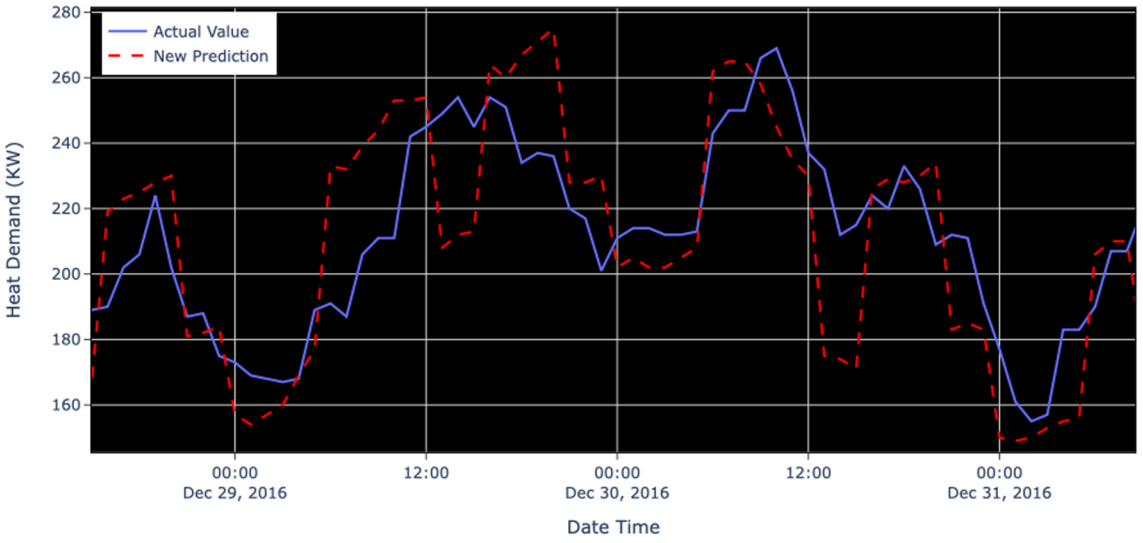

Fig. 8. A sample day prediction of Sharing Cities.

accuracy of the Artificial Neural Network would improve once data points with increased number of features are collected. It can also be concluded that for datasets with smaller amount of data Decision Tree is highly accurate. The approach proposed and used in this paper can also be used for predicting short term or long-term heat demand using weather forecast data.

CRediT authorship contribution statement

Kevin Naik: Conception and design, Analysis and interpretation of the data, Drafting the article or revising it critically for important intellectual content. **Anton Ianakiev:** Conception and design, Analysis and interpretation of the data, Drafting the article or revising it critically for important intellectual content.

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.

Acknowledgments

We gratefully acknowledge the support of Core Control and Siemens. We would also like to thank Sean Jones of Siemens for insightful discussion. Dr. Edward O'Dwyer of Imperial College of London for Sha his contribution to SHARING CITIES. All authors approved the final version of the manuscript.

References

[1] Department for Business Energy & Industrial Strategy. Heat networks. 2019, [Online]. Available: https://www.gov.uk/government/publications/hndu-pipeline. [Accessed 14 March 2020].
[2] © Crown copyright. In: Heat networks market study. Competition and Markets Authority; 2018.
[3] Multipipe Ltd. 5 simple statistics that prove district heating is the future of UK heat distribution. 2018, [Online]. Available: https://www.multipipe.co.uk/5-simple-statistics-that-prove-district-heating-is-the-future-of-uk-heat-distribution/. [Accessed 14 March 2020].
[4] Dotzauer E. Simple model for prediction of loads in district-heating systems. Appl. Energy 2002;73(3–4):277–84.
[5] Chramcov B, Dostál P, Balátě J. Forecast model of heat demand. In: The 29th annual international symposium on forecasting. Hong Kong: 2009.
[6] Provatas S. An online machine learning algorithm for heat load forecasting in district heating systems. Karlskrona, Sweden: Blekinge Institute of Technology; 2014.
[7] Johansson C, Bergkvist M, Geysen D, Somar OD, Lavesson N, Vanhoudt D. Operational demand forecasting in district heating systems using ensembles of online machine learning algorithms. Energy Procedia 2017;116:208–16.

[8] Talebi B, Mirzaei P, Bastani A, Haghighat F. A review of district heating systems: modeling and optimization. Front Built Environ 2016;2:22.
[9] Arce IH, López SH, Perez SL, Rämä M, Klobut K, Febres JA. Models for fast modelling of district heating and cooling networks. Renew. Sustain. Energy Rev. 2018;82:1863–73.
[10] 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 Build. 2018;162:144–53.
[11] Saloux E, Candanedo JA. Forecasting district heating demand using machine learning algorithms. Energy Procedia 2018;149:59–68.
[12] Suryanarayana G, Lago J, Geysen D, Aleksiejuk P, Johansson C. Thermal load forecasting in district heating networks using deep learning and advanced feature selection methods. Energy 2018;157:141–9.
[13] Xue G, Pan Y, Lin T, Song J, Qi C, Wang Z. District heating load prediction algorithm based on feature fusion LSTM model. Energies 2019;12(11):2122.
[14] Molitor C, Groß S, Zeitz J, Monti A. MESCOS—A multienergy system cosimulator for city district energy systems. IEEE Trans. Ind. Inf. 2014;10(4).
[15] Shaikh PH, MohdNor NB, Nallagownden P, Elamvazuthi I, Ibrahim T. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. Renew. Sustain. Energy Rev. 2014;34:409–29.