On The Evaluation of District Heating Load Predictions

Herman Hansson

Thesis for the degree of Master of Science in Engineering
Division of Efficient Energy Systems
Department of Energy Sciences
Faculty of Engineering | Lund University

On The Evaluation of District Heating Load Predictions

by Herman Hansson

LUND
UNIVERSITY

Thesis for the degree of Master of Science
Thesis advisors: Dr. Sara Månsson, Senior Lecturers. Kerstin Sernhed and Marcus Thern

To be presented, with the permission of the Faculty of Engineering of Lund University, for public criticism on the meeting at the Department of Energy Sciences on Friday, the 2nd of June 2023 at 15:45.

This degree project for the degree of Master of Science in Engineering has been conducted at the Division of Efficient Energy Systems, Department of Energy Sciences, Faculty of Engineering, Lund University.

Supervisors at the Department of Energy Sciences were Senior Lecturers Kerstin Sernhed and Marcus Thern. Supervisor at Utilifeed was Dr. Sara Månsson.

Examiner at Lund University was Associate Professor Martin Andersson

The project was carried out in cooperation with Gothenburg-based Utilifeed, a software provider for District Heating companies.

Department of Energy Sciences
Faculty of Engineering
Lund University

ISSN: <0282-1990>
LUTMDN/TMHP-23/5524-SE

Typeset in LaTeX
Lund 2023

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ANN Artificial Neural Network

DH District Heating

CVRMSE Coefficient of Variance of Root Mean Squared Error

DHN District Heating Network

DT Decision Tree

FFNN Feed-Forward Neural Network

GBDT Gradient Boosted Decision Tree

LR Linear Regression

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MLP Multi-Layered Perceptron

NMBE Normalized Mean Bias Error

OLS Ordinary Least Squares

RBF Radial Basis Function

RF Random Forest

RMSE Root Mean Squared Error

RN_RMSE Range Normalized Root Mean Squared Error

SVR Supported Vector Regressor

Sammanfattning

Fjärrvärme är en teknologi som har stor potential att bidra till att möjliggöra ett fossilfritt samhälle. För att uppnå denna potential, finns det dock en del förbättringar som måste göras för att förbättra fjärrvärmeverksamheten i stort, minska förlusterna från systemen och därmed öka konkurrenskraften hos fjärrvärme som teknologi. Ett flertal studier har visat att ett sätt att göra detta, och därmed öka fjärrvärmens effektivitet, är att använda sig av databaserade metoder som maskininlärning.

Det har ifrågasatts huruvida de studier där maskininlärning har implementerats på fjärrvärmedata avspeglar faktiska behov hos fjärrvärmeleverantörer. Därför var det denna rapports syfte att studera hur maskininlärningsmodeller, implementerade på fjärrvärmedata för att prediktera last, kan utvärderas på ett sätt som överensstämmer med hur de används i faktisk fjärrvärmeverksamhet. Lastprediktioner har en mängd olika användningsområden hos fjärrvärmeleverantörer, exempelvis kan de användas för att planera och optimera värmeproduktion.

Rapporten delades in i tre olika undersökningar. En enkätstudie för fjärrvärmeleverantörer genomfördes för att ge insikt i hur de använder lastprediktionsmodeller samt hur de utvärderar dessa modeller. Baserat på teori kring fjärrvärme, maskininlärning och statistiska avvikelsemått, undersöktes hur lastprediktionsmodeller ändamålsenligt kan valideras. Undersökningen resulterade i ett förslag på utvärderingsramverk. Slutligen utvärderades Energy Predict, en lastprediktionsmodell utvecklad av mjukvaruföretaget Utilifeed, mot detta utvärderingsramverk.

Ett antal slutsatser kunde dras från studierna. Enligt resultaten från enkäten används lastprediktioner hos fjärrvärmeleverantörer för planering av produktion och försäljning, dimensionering av utrustning/infrastruktur/produktion samt som ett steg i feldetektering. Det verkade dock som om noggrannheten hos dessa modeller generellt sett inte utvärderas hos fjärrvärmeleverantörer. Från teoridelen drogs slutsatsen att studier kring lastprediktioner har varit begränsad till lastprognoser med kort tidshorisont, som används i produktionsplanering, samt som ett sätt att detektera fel. Baserat på detta föreslogs två utvärderingsramverk: ett ramverk för att utvärdera lastprediktionsmodeller med syfte att dimensionera, och ett för modeller med syfte att planera försäljning. Det fastslogs att de två olika syftena, dimensionering och försäljningsplanering, kräver olika typer av pricksäkerhet och därför olika utvärderingsramverk. För dimensionering ansågs det vara värdefullt att kunna prediktera toppar i värmelast och för försäljningsplanering ansågs det vara värdefullt att kunna prediktera ackumulerade summor av värmelaster.

Insikterna som samlats in angående statistiska mått och utvärderingsperioder användes för att utveckla de två utvärderingsramverken. Ramverken använde måtten Coefficient of Variance of Root Mean Squared Error och Normalized Mean Bias Error, samt olika uppdelningar av valideringsdata för att bedöma vanliga brister hos maskininlärningsbaserade lastprediktionsmodeller. För att visa på hur de två utvärderingsramverken skulle genomföras utvärderades fyra olika lastprediktionsmodeller: Energy Predict, energisignatur, Supported Vector Regressor och XGBoost. Energy Predict uppvisade bäst pricksäkerhet av alla fyra modeller mot båda ramverken.

Dessa ramverk, och diskussionen som gavs med dem, är framtagna för fjärrvärmeleverantörer och andra utvecklare av lastprediktionsmodeller. Då olika modeller jämförs med varandra, eller då ett mått på pricksäkerheten behöver kvantifieras, kan dessa ramverk vara av värde.

Abstract

District Heating is a technology with the potential to enable a fossil-free society. However, to realize this potential, some improvements need to be made in order to improve District Heating operation at large, decrease losses in the systems, and thus increase the competitiveness of District Heating as a technology. Several have shown that a possible solution, to increase the efficiency of District Heating, is to utilize data-based methods such as Machine Learning.

Questions have been raised regarding studies on Machine Learning implemented on District Heating data, stating that the research does not reflect the actual needs of District Heating utilities. Therefore, the aim of this report was to investigate how Machine Learning models, implemented to predict heat load, could be evaluated in a way that aligns with how they are used in District Heating operation. Heat load predictions have a number of use cases, for example as a way to plan and optimize heat production.

The report was divided into three different investigations. A survey study for District Heating utilities to give insight into how they currently use heat load prediction models as well as how they evaluate these models. Based on theory regarding District Heating, Machine Learning, and statistical error measures, it was investigated how heat load predictive models could be evaluated in a suitable way. The investigation resulted in the proposal of an evaluation framework. Lastly, the heat load prediction model Energy Predict, developed by Utilifeed, a software provider to District Heating utilities, was evaluated against this framework.

Some conclusions could be drawn from the studies. According to the results from the survey, heat load predictions are used in District Heating utilities for planning production and sales, dimensioning equipment/infrastructure/production, and as a step in fault detection. However, it seemed as if the accuracy of these models is generally not evaluated in District Heating utilities today. From the theory section, it was concluded that research on heat load predictions has been limited to short-term load forecasts, used for planning production, and as a step in detecting faults. As a result, two evaluation frameworks were proposed, evaluating the predictive performance of heat load prediction models used for dimensioning and sales planning. It was concluded that the two purposes, dimensioning and sales planning, require different kinds of accuracy and thus also different kinds of evaluation frameworks. For dimensioning, it was considered valuable to predict peaks in heat load, and for sales planning, it was considered valuable to predict accumulated sums of heat loads.

The insights gathered regarding evaluation metrics and periods were used when proposing the two evaluation frameworks. The frameworks utilized the error measures Coefficient of Variance of Root Mean Squared Error and Normalized Mean Bias Error, as well as different sectionings of the validation data, assessing the common flaws of heat load prediction models based on Machine Learning. The two evaluation frameworks were showcased by evaluating the predictive performance of four different load prediction models: Energy Predict, heat load signature, Supported Vector Regressor, and XGBoost. Energy Predict showed the best performance of all four models on both frameworks.

These frameworks, and the discussion provided with them, are developed for District Heating utilities and other users and developers of heat load prediction models. As different models are compared with each other, or when measures of accuracy need to be quantified, these frameworks may be found valuable.

Chapter 1

Introduction

District Heating (DH) could prove to be a technology enabling sustainability in society. IEA states that there are over 6 000 District Heating Networks (DHNs) in Europe, accounting for approximately 11% of the European heat demand [1]. These networks are supplied with heat from different energy sources, including both fossil fuels and renewables. Werner and Frederiksen, however, argue that a fundamental idea of DH, and which to some degree is already realized, is to utilize heat sources that would otherwise have gone to waste - such as excess heat from industrial processes [2]. This resource-efficient way of implementing DH results in a heating method with lower emission rates of greenhouse gases compared to conventional heating methods such as oil- or gas burners. DH could stand out as a potential heating method in a sustainable society, so much so that it has been identified at the IEEE International Conference on Power and Energy as a key technology in decarbonizing the heating sector [3].

For DH to gain attraction against other ways of heating, Gaballo, Nielsen, Khan, and Heller acknowledge the importance of both maintaining and improving the efficiency of DHNs [3]. In the report, different scenarios for the European heating sector in 2050 are studied and it is concluded that ”A reduction of the losses from 0.10 to 0.05 results in about 16 % more DH expansion”. A reduction of losses lowers the levelized cost of energy compared to other heating technologies, increasing the expected uptake of DH.

With the amount of data available in a DHN, it is suggested that data-based methods such as Machine Learning (ML) can provide insights for utilities, for example by predicting the heat load (i.e. the amount of heat produced and distributed). This insight could help DH utilities to maintain and improve the efficiency in their DHNs [3].

The ability to predict heat load, both short- and long-term, can lead to increased efficiency in DHNs. On a short time scale of a few hours, an accurate prediction of the heat demand enables a DH utility to produce a sufficient amount of heat so that all customer needs are met, while still not getting excess heat, higher temperatures, back in the returning pipes [4]. Higher return temperatures cause a lower efficiency in a DHN, this will be further described in the theory section of this report.

However, at times it can be beneficial to temporarily overshoot the heat demand in a

DHN for the cause of production efficiency. A way of explaining this dynamic is through marginal costs for heat in different production units of a DHN. For example, a big plant driven by low-cost fuels such as residual waste or wood chips often has higher capital costs and lower marginal costs than a peak oil burner. This is a result of higher cost per energy unit in oil compared to wood chips, increasing the marginal cost, but smaller plant capacity, lowering the need for invested capital. The cost structure incentivizes the DH utility to utilize the bigger plants more, which may make it desirable to produce a temporary excess of heat, making a buffer of heat in the DHN so that the peak burner does not have to start. In this case, an accurate prediction is beneficial from both an environmental and the utility's financial perspective, as oil combustion increase costs and the climate impact. Further, it can be said that this dynamic is relevant for both low, but specifically high overall heat loads in a DHN since it is in those scenarios where peak burners potentially need to be used [2].

On longer periods of months to several years, accurate predictions on both accumulated and peak heat demand can provide valuable insight to DH utilities to plan capacity. Forecasts on total heat demand over a period can guide the need for stored fuel. Forecasts on peak demand can help the DH utility to ensure that the capacity of all production units is sufficient. Predicting the heat load on a longer time scale is thus beneficial so that DH utilities do not have to pay fixed costs for capacity not needed [5].

The predicted heat load on longer periods can also provide insight when projecting different scenarios. For example, financial insight can be provided by predicting the income from certain different weather conditions and price models. Many DH utilities have a price model where the heat load is charged differently depending on the time of day and year and/or charged upon other measures than the total heat load in kWh. Invoicing based on not only heat load but also peak power demand or flow can incentivize customers to change their consumption in a way that is beneficial for overall DHN efficiency. Projecting heat load in a DHN can also provide insight regarding the development of a DHN and its infrastructure, for example when combining a long-term heat load prediction with a model of a DHN and simulating the network when, for example, connecting a new suburb or production facility [5].

Dimensioning of equipment on a substation or cluster (an accumulated set of substations) level is another use case for long-term heat load predictions. Measuring instruments are often more expensive the bigger the heat load, and need to be replaced more often. But since the heat load pattern of substations (i.e. the behavior of customers) is hard to predict, large safety margins are often applied, increasing costs. The same dynamic can be seen on a cluster level, analyzing the need for investments in infrastructure, (i.e. whether to increase the flow capacity or not) to a neighborhood as more customers are connected. [2].

Regarding predictions of DH load, and using these for, among else, production planning and/or dimensioning, there is generally an asymmetry of how costly the cases of over- and underestimation are. In the case of underestimation, the DH utility has failed in its

commitment to satisfy customer needs. If the temperature goes below a certain threshold value, it may not be hot enough for certain industrial processes and/or for the needs of residential customers the furthest away from the heat source (the further, the more temperature loss). If the flow goes below a certain threshold value, it is because of a too-low pressure gradient in the network, meaning that when the pressure gradient decreases on the grid, it will not be high enough for the outer parts to receive the required flow, and thus the needed heat. These two cases are both severe in the sense that the DH utility has failed in its commitment to its customers. The costs related to an equally sized overestimation, for example the cost of oversized infrastructure, equipment, or distribution losses, are not as high. Thus, there is often a margin of safety added to heat load predictions, meaning that the DH utility and therefore its customers potentially need to pay the costs for superfluously produced heat and installed equipment. With better-performing predictive models, however, these margins can be lowered and lower the costs for DH utilities and their customers [2].

As developers and users of predictive models validate predictions, it is necessary to have frameworks for evaluating the predictive performance in terms of accuracy, comparing the predicted heat load to the actual outcome. Without a developed evaluation framework, it is impossible to compare the predictive performance when developing models, as well as to assess if the accuracy is sufficient.

Studies have been made where ML has been utilized and predicted the heat load. The ways that these studies evaluate their predictive models can be used as examples when evaluating models in an industrial setting. However, questions have been raised regarding the gap between academic research on the subject (ML implementations on DH data) and actual DH operation, questioning that the research is limited to solely a few use cases [4]. For those use cases where the research is limited, the development of evaluation frameworks for those use cases is assumingly also limited. This lack of development constitutes a problem for model developers.

Another question that has been raised is that academic research is limited to what models could be improved by more advanced ML algorithms but not focused on the actual needs of DH utilities [4]. As model developers choose what to focus their development on, the development must be aligned with the needs of DH utilities. For example, it may not be as valuable to further develop models that are already validated as sufficiently accurate according to DH utilities.

To study predictive models and how they are used, this research project has been in collaboration with Utilifeed, a company developing and providing software to DH utilities. Their software tool Energy Predict is used among DH utilities to give them insight into their operation. Among else, it utilizes ML algorithms to predict heat load, on several different time horizons and for both substations, clusters of substations, or whole DHNs. The knowledge that Utilifeed has acquired regarding the use of heat load prediction models, specifically Energy Predict, has been utilized in the project.

1.1 Objective

This thesis aims to shine a light on promising ML implementations where DH load is predicted. Further, this thesis will study how the performance of these implementations is evaluated. In the cases where the development of evaluation frameworks is deemed insufficient, new ones will be proposed. These proposed frameworks will aspire to evaluate heat load prediction models in a way that aligns with how they are and/or could be used in DH operation. Evaluation frameworks are important in the further development of ML implementations, concerning DH but also beyond.

This will be done by answering the following questions, in the report, these questions will be referred to as Research Questions Number 1, 2, 3, and 4:

1. Investigated through a survey study, for what purposes are heat load predictions used in Swedish DH utilities?
2. Investigated through the same survey study, are these heat load prediction models validated as sufficiently accurate according to the DH utilities?
3. Based on theory regarding DH, ML, and statistic error measures, how can heat load predictive models suitably be evaluated?
4. With an evaluation framework developed in line with the conclusions drawn from Research Question Number 3, how does Energy Predict perform against this evaluation framework?

Research Question Number 1 and 2 will be answered by a survey study with DH utilities as respondents. Research Question Number 3 will be investigated by a discussion, based on theory regarding DH, ML, and statistic error measures. The concluding answer will be in the form of a proposed evaluation framework, evaluating those ways of predicting load where the development of evaluation frameworks is deemed insufficient. Research Question Number 4 will be answered by utilizing the proposed evaluation framework in Research Question 3, comparing Energy Predict to other commonly used load prediction models discussed in the theory section.

1.2 Constraints

Heat load predictions in DHNs is the topic of this report. It will be investigated how they are calculated, how they are used, and how they are evaluated. Other metrics that can be predicted in DH operation are temperature and flow but these predictions are not regarded as within the scope of the thesis.

The questions of the report which require contact with DH utilities will be constrained to Swedish utilities. While there, assumingly, are similarities between DH operation in

different countries, the conclusion that will be drawn concerning Research Questions Number 1 and 2 will be drawn for Swedish conditions.

A heat load forecast is often dependent on an underlying weather forecast. Studying the importance of accurate weather forecasts, as well as comparing the performance of different weather forecasts for DH load prediction, is complex and considered beyond the scope of the report.

When discussing different ML algorithms on which to base heat load predictions, there is a possibility to combine different models. Via learning which ML algorithms perform best in which regimes, there is potential for better predictive capability. However, because of the increasing complexity of these so-called ensemble models, they are considered beyond the scope of the report.

For some ML algorithms and inputs, preprocessing of the input features can prove to be beneficial. There are many ways to preprocess and some of these ways have been implemented on DH data. However, as is the case for ensemble models, the complexity increases if different ways of preprocessing would be analyzed together with different ML algorithms. Thus, it is considered beyond the scope of the report.

One of the algorithms that will be discussed in this report is Energy Predict by Utilifeed. It will be discussed how it can provide value for DH utilities as well as how its performance compares to other ML algorithms. However, how Energy Predict is implemented and what ML algorithms it is based upon will not be revealed due to non-disclosure reasons.

1.3 Outline of the thesis

The report is not structured in a common way of Introduction, Method, Results, and Discussion. An alternative structure was chosen as the proposal of the evaluation frameworks using insights gathered from the results and analysis of the survey study.

To gather insights that can be used when proposing an evaluation framework, to answer Research Question Number 3, a theory section will be provided. The theory chapter, Chapter 2, is divided into four sections. The first section will describe DH, what it is and the factors affecting its efficiency and competitiveness. The industry standard of making heat load predictions in DH, heat load signature, will also be described in this section. The second section will provide a brief theory regarding the commonly used ML algorithms in DH load predictions. The third subsection will investigate how ML models have been applied to DH data and predicted heat load, as well as to what degree these models have a developed evaluating framework associated with them. Insight into how ML is implemented to predict heat load will be used when proposing an evaluation framework. The fourth, and last, section will describe how different evaluation metrics are calculated and compare what they aim to measure.

Subsequently, a chapter will be provided regarding the survey study that was made for DH utilities. The aim will be to answer Research Questions Number 1 and 2, how heat load predictions are used, i.e. their purpose, and if they are validated as sufficiently accurate in Swedish DH utilities.

Followingly in Chapter 4, Energy Predict by Utilifeed will be described. More specifically, it will be described how it can be used to predict heat load for a DH utility. Showcasing the use cases specified in this chapter makes it possible to further in the report propose suitable evaluation frameworks.

In Chapter 5, an evaluation framework will be proposed for those ways of heat load prediction where the current ways of evaluation are regarded as insufficient. It will be assumed that the heat load predictions are implemented and used in a way that a user of Energy Predict would.

To showcase the evaluation framework and how Energy Predict and other heat load prediction models perform against it, Chapter 6 will be provided.

A discussion will be held in Chapter 7. Insights beyond the concluded answers to the Research Questions will be discussed, both regarding the present operation of DH utilities, the previous studies on the subject, and the evaluation framework. Further, subjects for future work to investigate will be suggested.

Lastly, concluding remarks will be given in Chapter 8.

1.4 Clarifying note

In this report, there will be a distinction between the words *forecast* and *prediction*. Forecasts are strictly made forward in time, based on backward-looking data with one or more features. Predictions can both be made forward in time and in hindsight, the input can both be based on backward-looking data as well as present data (in relation to the point of time where the prediction is made).

Chapter 2

Theory

2.1 District Heating

A theory section of DH is provided in order to understand what needs DH utilities have and how heat load predictions can be used to fulfill these needs. Further, an overview will be given of how these heat load predictions are commonly calculated according to the industry standard *heat load signature*. Werner and Frederiksen have constituted the foundation of DH theory for this thesis and the interested reader is encouraged to find further theory in the textbook *District Heating and Cooling* [2].

2.1.1 District Heating Distribution

In essence, DH is a technology where heat is distributed to customers from available heat sources. Heat demands suitably satisfied by DH include space heating, preparation of domestic hot water, and low-temperature industrial heating demands. The distribution is done through a network of pipes, insulated to decrease heat loss and filled with a heat-storing medium, often pressurized water. The medium deploys the heat, either in a heat exchanger or directly in the customers' heating system, eventually circulating back to the supply source to reheat. In order to be a competitive energy service, the network can not have unlimited reach since that increases both heat loss during distribution and pipe-related capital investments. A District Heating Network (DHN) is thus often limited to a certain region or city - hence the prefix 'District' in 'District Heating'. A simplified illustration of a DHN is shown in Figure 2.1 [2].

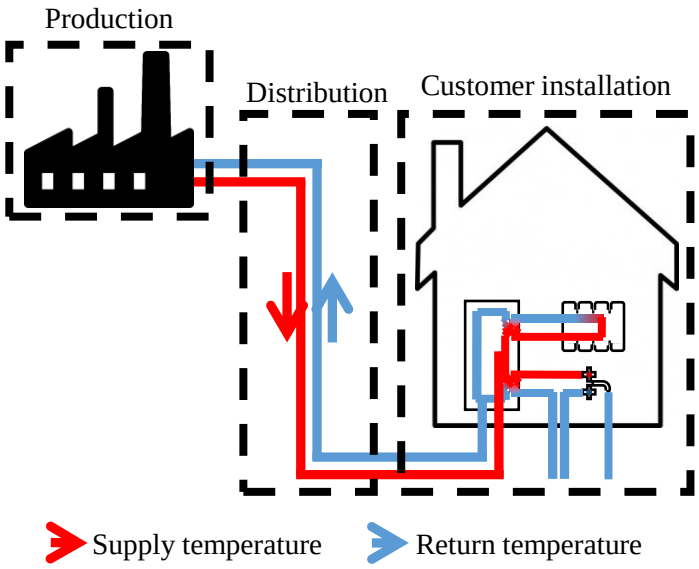

Figure 2.1: A simplified illustration of a District Heating Network, including Production, Distribution, and Consumption (Customer installation)

Substations typically consist of heat exchangers, pumps, control valves, and sensors. The heat exchangers are responsible for transferring heat from the DHN to the building's heating system, while the pumps and valves help to regulate the flow of hot water through the system. The sensors help to monitor and control the temperature and pressure of the hot water flowing through the system, also acting as metering instruments for billing DH load. Metering is a standard requirement in DHNs as it supports fair and objective invoicing to DH customers. Via using two thermometers, one before and one after the heat exchanger, along with a flow sensor, the amount of heat delivered to a customer can be calculated and then invoiced [2].

Further, this DH data can be collected and analyzed employing data-based methods, such as ML, which will be seen in this report.

A simplified illustration of a substation can be seen in Figure 2.2.

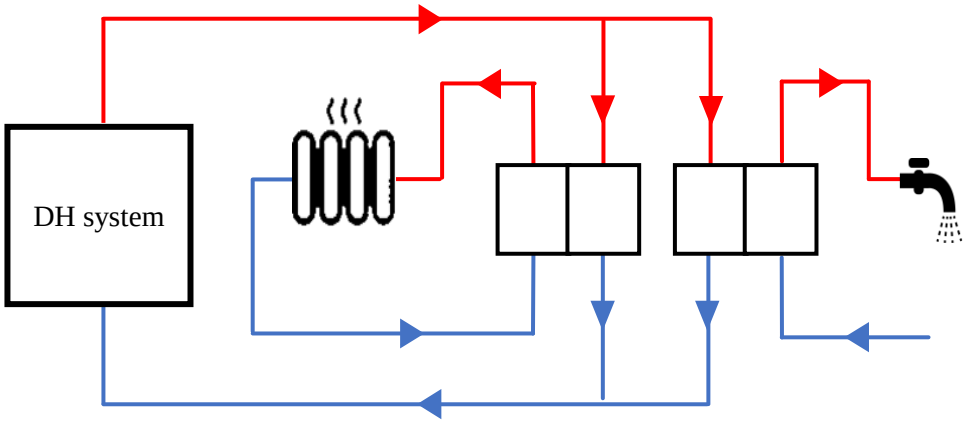

Figure 2.2: A simplified illustration of a substation with indirect connection in both space heating and domestic hot water systems, a common way to create hydraulic separation in the substation.

Increasing the efficiency in DHNs is closely related to lowering the return temperature, i.e. the temperature of the medium returning to the heat source. A low return temperature makes it among else possible to extract more heat from supply sources, possible to integrate other lower-grade heat sources, and increases the coefficient of performance for any heat pumps in the network. Additionally, since the heat gradient is linear to temperature difference, lower overall temperatures in the network pipes decrease heat loss in distribution. In principle, lower supply temperatures or flow would cause the return temperature to decrease. However, these metrics must be kept high enough to satisfy the needs of every customer in the grid. For example, a lower pressure gradient at a specific substation will lead to a lower possible flow, which may make it impossible to supply the heat needed [2].

2.1.2 Heat Load Signature

The industry standard of predicting heat load in DHNs today is by determining a heat load signature (also commonly mentioned as *energy signature*) [5]. It is done under the assumption that, for each substation, the outdoor-temperature-dependant heat load (e.g. space heating) has a negative linear correlation with outdoor temperature until it reaches zero at the *balance temperature*, specific for a certain DHN/substation/building. At temperatures above the balance temperature, the heat load that remains is the heat

load not dependent on outdoor temperature (e.g. tap water and industrial processes) as it is assumed to be constant. These two categories of heat load added together result in a hockey stick-shaped relationship between heating demand and outdoor temperature, breaking at the balance temperature [2].

The modeled relationship of heat load to outdoor temperature can be used in both the short- and long-term. Short-term temperature forecasts can be used as input for short-term heat load forecasts, enabling production planning. Long-term minimum outdoor temperatures can be used to yield a dimensioning, maximum heat load - commonly mentioned as *Design Load*. The Design Load value can be used when dimensioning production facilities, the network, and/or equipment. The outdoor temperature used as input for dimensioning purposes is often called *design outdoor temperature* and has been specified by the standard CEN 2004 as the mean temperature for the 0.08 % coldest hours during a year. Previous standards have had lower percentage rates (0.05 %) and therefore lower temperatures, but as the insulation capacity in buildings generally has increased, recent standards use higher percentage rates [2]. In Figure 2.3 it is shown how a low design temperature can provide a DH utility with a sufficiently high heat load value to dimension upon.

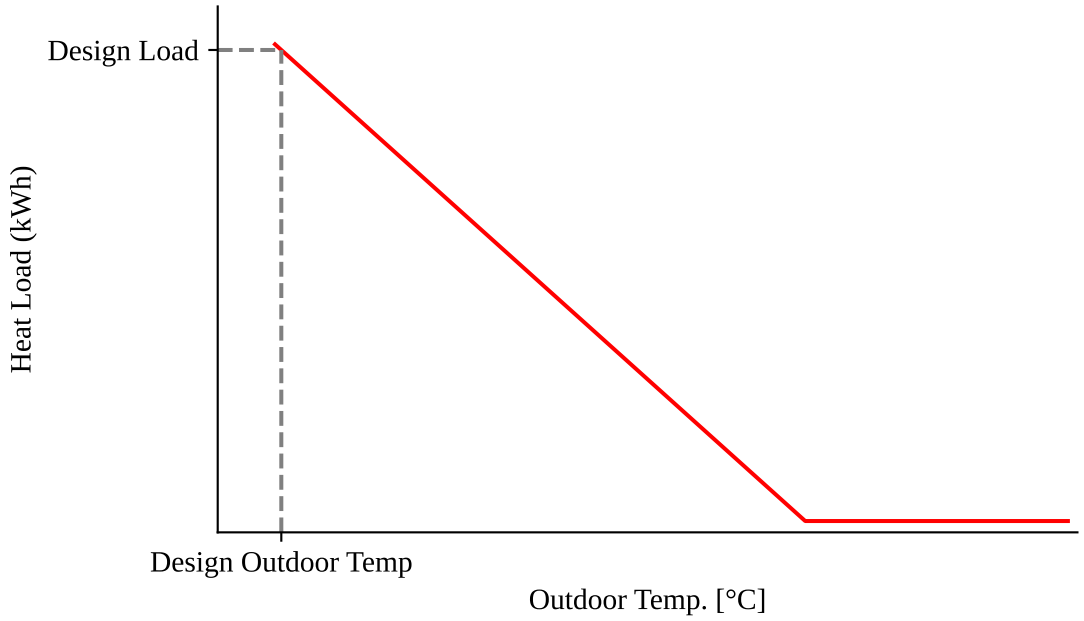

Figure 2.3: A design outdoor temperature and a heat load signature (red line) can be used to derive a sufficiently high heat load value (Design Load) to dimension upon

The heat load signature is often calculated by implementing a Linear Regression (LR) for heat load versus outdoor temperature, applied to heat load measurements that are made below the balance temperature. The linear relationship, i.e. the constants $\hat{\beta}_0$ and $\hat{\beta}_1$ in the equation $y = \hat{\beta}_1 x + \hat{\beta}_0$ is then specified where the sum of squared residuals between

the line and input data finds its minima, meaning calculating the relationships in Eq. 2.2 and 2.1, then calculating Eq. 2.3 and 2.4. This line, together with the mean heat load of those data points over the balance temperature, constitutes the heat load signature [5].

$$SS_{xx} = \sum_{i=1}^{n} x^2 - \frac{1}{n}(\sum_{i=1}^{n} x)^2 \tag{2.1}$$

$$SS_{xy} = \sum_{i=1}^{n} xy - \frac{1}{n}(\sum_{i=1}^{n} x)(\sum_{i=1}^{n} y) \tag{2.2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.3}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{2.4}$$

Where x is all x-values in the data set of size n, y all y-values, and $\bar{x}$, $\bar{y}$ the mean of x- and y-values in the data set respectively.

The heat load signature is widely used partly because of its simplicity. However, the assumption that the heat load is linear with outdoor temperature up until a certain balance temperature is a serious assumption, not taking into consideration that solar irradiation may lower the heat load demand or that high wind speeds may cause wind chill and thus increase demand. Furthermore, societal behavior is not taken into consideration [2].

Below, two figures are shown, the former of which where the heat load signature can be considered a relatively good approximation (Figure 2.4), and the latter showcasing how it may not be as good of an approximation (Figure 2.5). This is often a result of the former being a substation connected to one or more apartment buildings, and the other to an office building or industrial customer. The heat load patterns of office buildings and industrial customers tend to differ from apartment buildings as the heat demand increases specifically on weekdays and working hours, when there are people present. This heat load pattern can not be seen in apartment buildings and is not captured by a heat load signature [2].

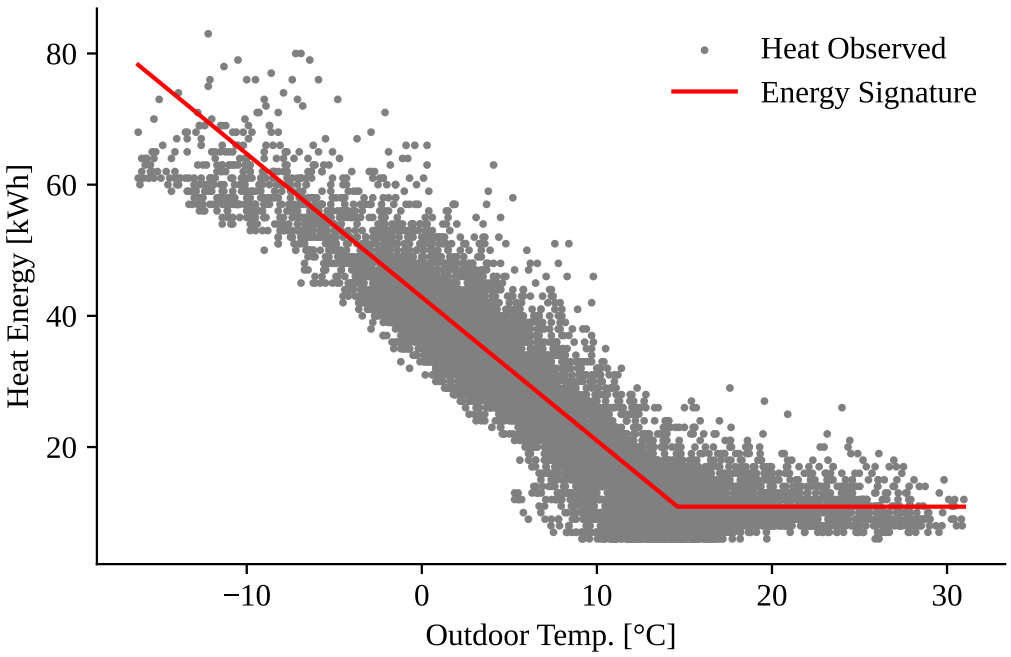

Figure 2.4: The industry standard for heat load predictions, heat load signature, fitted to observed values. The model follows the assumption that heat load is linear with outdoor temperature up until a certain balance temperature. The picture showcases how the heat load signature at times can resemble the actual behavior relatively well. To be compared with Figure 2.5

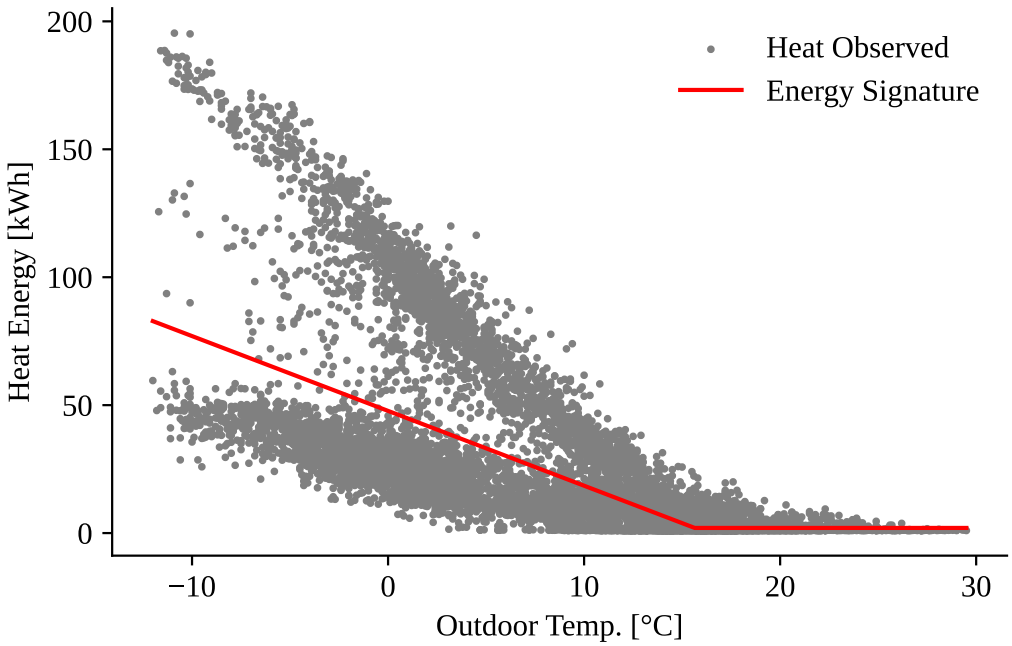

Figure 2.5: The industry standard for heat load predictions, heat load signature, fitted to observed values. The model follows the assumption that heat load is linear with outdoor temperature up until a certain balance temperature. The picture showcases how the heat load signature at times can resemble the actual behavior relatively bad. To be compared with Figure 2.4

2.2 Machine Learning

By using models that can learn from different features, as well as being able to reproduce non-linear relationships, a heat load prediction's accuracy can be significantly improved. Below, in Figure 2.6 and Figure 2.7, the same substations are forecasted upon, but with an ML-based algorithm (more specifically the model Energy Predict, provided by Utilifeed) making the predictions. The accuracy is improved in both cases, but especially in the latter. Both models (heat load signature and Energy Predict) and substations are trained on the one-and-a-half-year period 2021-06-01 to 2022-12-31 and then tested on the one-and-a-half-year period 2020-01-01 to 2021-05-31.

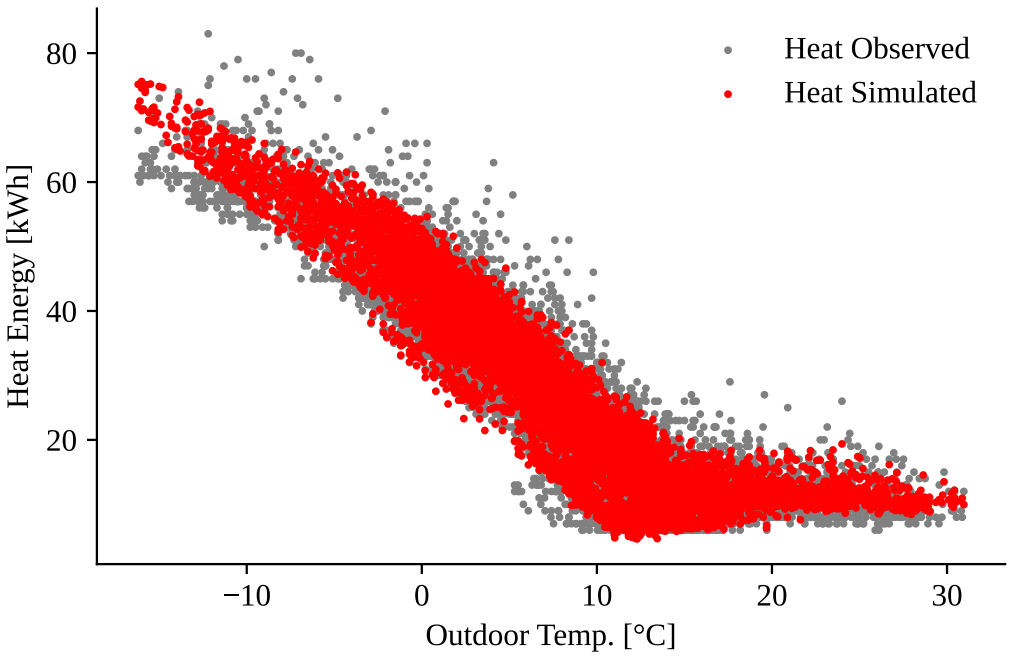

Figure 2.6: The same data set as in Figure 2.4 but predicted by the ML model Energy Predict, provided by Utilifeed. Illustrating improved accuracy graphically.

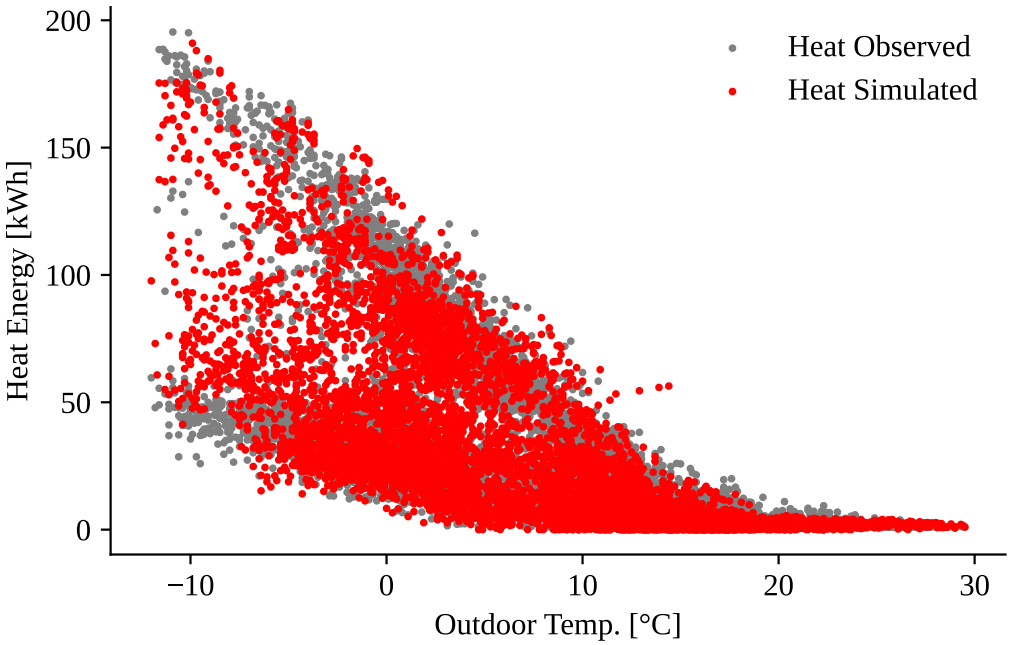

Figure 2.7: The same data set as in Figure 2.5 but predicted by the ML model Energy Predict, provided by Utilifeed. Illustrating improved accuracy graphically.

To evaluate promising ML models, in a way that resembles how these models are trained and implemented, insight into ML is needed. A brief theory section on ML will therefore be provided.

ML has a wide range of applications but the main principle, as Müller and Guido put it, is about extracting knowledge from data [6]. With large amounts of data available, it has become an increasingly important tool for solving complex problems and making predictions in a variety of fields. A common division of ML methods is Supervised Learning and Unsupervised Learning, this report will focus on the former as it is the deployed method when predicting heat load. Supervised ML is a type of ML where the goal is clear, i.e. the insight that one wants to extract from data is already specified. It involves providing a model with labeled data in order to learn from it and make predictions on new data input [6].

To evaluate the performance of a supervised learning model, the labeled data is divided into a training set and a test set. The training set is used to train the model by providing examples of inputs and their corresponding outputs. The test set is used to evaluate the performance of the trained model by measuring how accurately it can predict the correct output [6]. The workflow associated with supervised learning is illustrated in Figure 2.8.

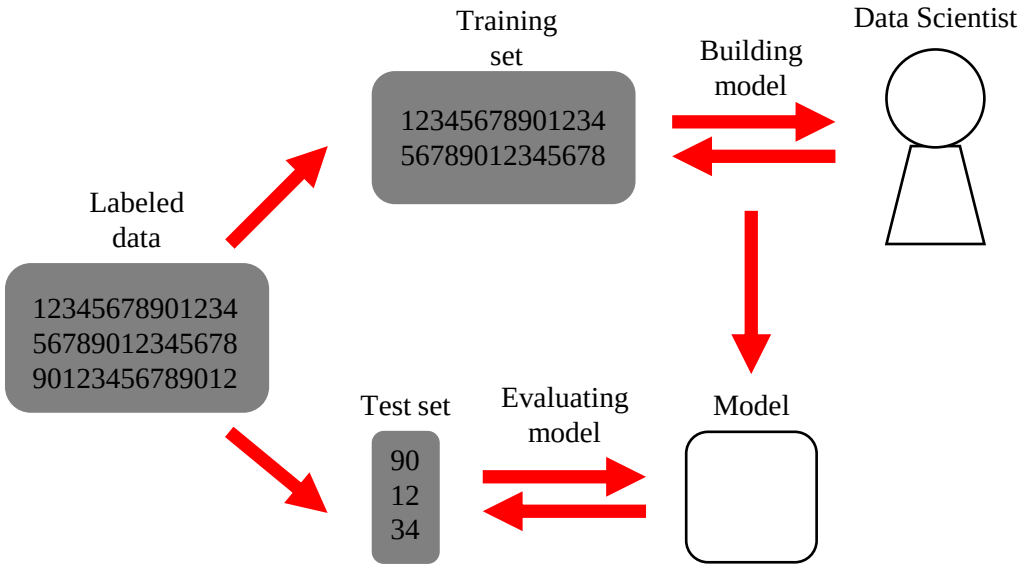

Figure 2.8: Simple graphical illustration of the workflow associated with supervised learning

Supervised learning can be used for a wide variety of tasks, including classification and regression. Predicting heat load in the way that a heat load signature and other predictive models do is a regression task since a distinct value of DH load is requested. Thus, only models used for regression will be discussed. Different algorithms can be used for

the task, each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the specific problem being solved and the nature of the data. Below, a short introduction will be given regarding the most promising algorithms when applying ML to DH data.

2.2.1 Artificial Neural Networks

The concept of Artificial Neural Networks (ANNs) is shown graphically in Figure 2.9. An ANN, also commonly mentioned as a Deep Learning algorithm, is composed of one or more layers of interconnected nodes (neurons) that process data and perform computations. In this report, only the relatively basic implementations of Deep Learning will be mentioned, namely, Multilayer Perceptrons (MLP) also known as Feed-Forward Neural Networks (FFNN) or just Neural Networks. The basic unit of an ANN is a neuron, which receives weighted input from data, or other neurons, and produces an output signal. The output of one neuron can, in much the same way, be used as the input to another neuron, allowing for the formation of complex networks of interconnected neurons. During the training process, the weights of the connections between neurons in the network are adjusted based on the input data and the desired output. Proposed weights will be evaluated by doing a prediction on the training set that is compared to the actual outcome, assessing the difference between the two to update the weights [6].

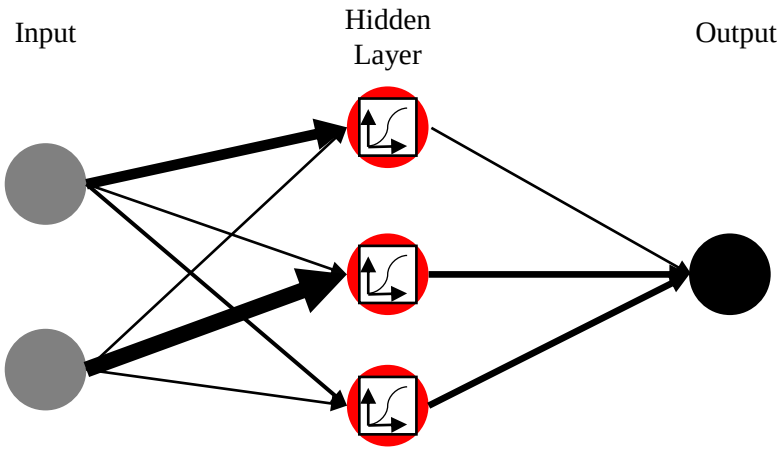

Figure 2.9: Graphical showcasing of a simple ANN. The width of the arrows exemplifies different weighting factors.

A series of weighted sums, however, will not be more powerful than other LR models. What makes ANNs powerful is the non-linear function applied to the weighted sum in a neuron, making it possible for the model to replicate non-linear relationships between input and output. Examples of these non-linear functions include the step function, the tangens hyperbolicus, or the rectified linear unit [6].

Although ANNs are seen as state-of-the-art algorithms due to their ability to capture complex information in large data sets, they can require a lot of computational time as well as careful tuning of parameters. One should also keep in mind that an ANN performs better on a data set where all the features vary similarly, ideally with a mean of 0 and a variance of 1. In data sets where this is not the case, preprocessing may be needed [6].

2.2.2 Supported Vector Regressors

A Supported Vector Regressor (SVR) is another ML algorithm used for regression tasks. SVRs are based on the concept of finding a hyperplane that can map high-dimensional inputs to an output. There are two main types of SVRs: linear and non-linear. Linear SVRs are equivalent to multidimensional LRs and assume that the relationship between the input and output variables is linear and can be represented by a hyperplane. Non-linear SVRs, on the other hand, use a specific kernel function to calculate the distance between data points and the hyperplane, enabling non-linear characteristics [6].

A common kernel function that is often seen in ML implementations, for example on DH data, is the Radial Basis Function (RBF), measuring the distance using a Gaussian distribution with the width given by a hyperparameter gamma. It is a universal kernel, meaning that any continuous function can be approximated with arbitrary accuracy. No further discussion will be provided on SVR or the kernel functions, but from here on, all SVRs mentioned are non-linear with RBF as kernel function [6].

SVRs have several advantages over traditional regression methods. They are less susceptible to outliers and can handle non-linear relationships between input and output variables. On the downside, they require preprocessing of the data, with all features varying on a similar scale, and are heavily dependent on appropriately adjusted parameters (such as the aforementioned gamma) [6].

2.2.3 Decision Trees

A Decision Tree (DT) is an ML algorithm that models a decision-making process using a tree-like structure, where each internal node represents a decision based on a feature, each branch represents the possible outcomes of that decision, and each leaf node represents a final decision or outcome. A schematic illustration of a simple DT is also shown in Figure 2.10.

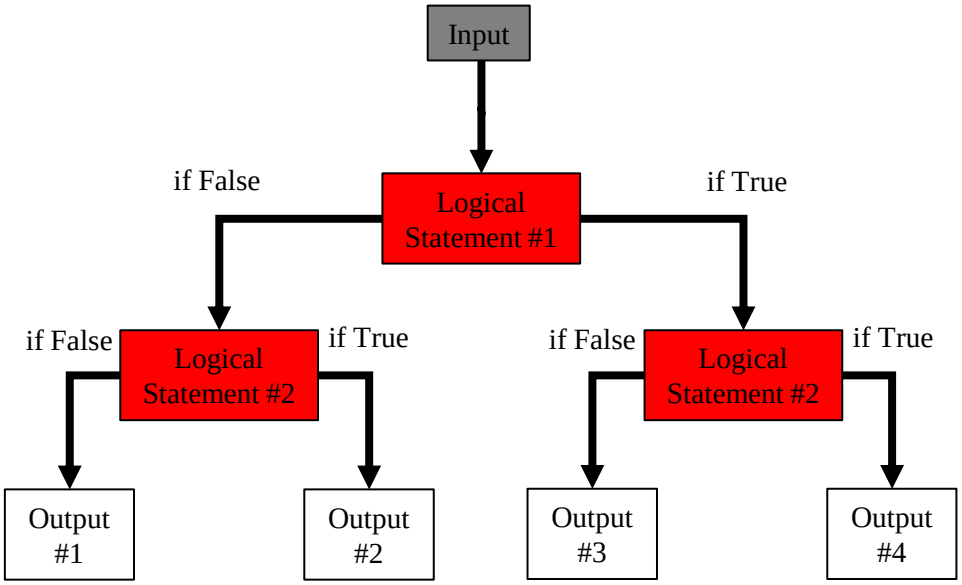

Figure 2.10: Graphical illustration of a simple DT

Once the tree is built, it can be used to make predictions for new, unseen data by following the path from the root node to a leaf node based on the values of the input features. The value of the target variable at the leaf node is then used as the prediction [6].

Random Forests

Random forests (RFs) is an ML algorithm based on DTs. It works by building a large number of DTs, each of which is trained on a random subset of the training data and a random subset of the features. After the training stage, the actual prediction is made (in the case of regression) by the average of the predictions made by all the trees in the forest [6].

Gradient Boosted Decision Trees

Another ML algorithm based on DTs is the Gradient Boosted Decision Trees (GBDTs). Gradient boosting is a technique used to improve the performance of a model by sequentially training weak learners (models that perform slightly better than random guessing). Instead of making many full DTs (as in the case of RF), the additional shallow trees try to correct the mistakes of the previous trees [6].

Algorithms based on DTs, both RFs and GBDTs, are widely appreciated in both industry-

and academic settings. They perform well, both in terms of accuracy and computational time, on big data sets and can handle a mixture of both binary and continuous features. One should keep in mind, however, that a DT is unable to extrapolate data. It is therefore crucial to train the model on a data set where the upper and lower output boundaries are set. On a data set with non-stationarities, or structural trends, this can be achieved by preprocessing the data [6].

2.3 Machine Learning implemented to predict heat load

The large amounts of data that can be generated from DH customers' heat load consumption in a DHN make it suitable to be investigated utilizing data processing and ML. As a consequence, the number of studies where ML is applied to DH has in recent years surged. For example, in a review by Ntakolia, Anagnostis, Moustakidis, and Karcanias, the authors not only consider ML to be promising but also needed to develop models with functionalities such as "economic evaluation of operations, monitoring of environmental indicators, diagnosis of abnormal operations and development of dynamic models for decentralized on-line control" [7]. However, it has been argued, for instance by Mbiydzenyuy et al., that there is a mismatch between current ML research and the real challenges of DH utilities today, questioning that the research is limited to solely a few use cases [4].

Theory regarding how ML has been implemented to predict DH load will be provided. The basis for this theory section is provided by the two reviews *Machine learning applied on the district heating and cooling sector: A review* by Ntakolia et al. and *Opportunities for machine learning in district heating* by Mbiydzenyuy et al. as they take a holistic approach to ML-based implementations on DH data [4] [7]. The purpose will first be to identify if there are ways of using heat load predictions that lack coverage in the literature. Further, it will be investigated if and how different ML models are implemented and evaluated, providing theory for Research Question Number 3. It will also be investigated which ML algorithms have shown promising results compared to other models. The insights from this investigation will be valuable when implementing models that can be used as benchmarks for evaluating Energy Predict, Research Question 4.

2.3.1 General comments regarding the literature

As stated both by Ntakolia et al. and Mbiydzenyuy et al., studies of ML applied to DH have so far mostly focused on short-term, often 24 h, heat load forecasts with an hourly granularity [4] [7]. Often, a supervised learning approach is taken, with historical heat load data as well as weather data and weather forecasts as input.

As stated in the introduction, heat load predictions can be used in an operational setting to align production with demand. DH utilities often have methods for optimizing heat production given a forecasted heat load [8]. It has been stated that a MAPE of 3 % to 5 % for heat load forecasts a couple of hours away is sufficient to steer operations in a way that increases efficiency in terms of energy demand by "as much as a few percent" [9]. It should be noted, however, that long lead times of several hours between the time heat is produced to the heat being available at a given location in the DHN pose a constraint for planning [4].

2.3.2 Input features of heat load prediction models

Some consensus has already been reached on which input variables are beneficial for heat load forecasting models. The most important feature in the weather data, for example, is the outdoor temperature but it is also common to include other parameters such as humidity, wind speed, and solar irradiation [10] [11] [12]. Other commonly used features are time-related ones such as holiday-, and weekday labels - changing both the pattern in residential hot water usage and industrial needs [7].

Moreover, it has been concluded that developing a model from granular heat load data, from each substation in a DHN, and summing the heat loads up for a forecast on a network level, is desirable. It enables ML models to notice distinguished user patterns and shows better results compared to basing a model on an aggregated sum of heat load data. Alternatively, clusters of substations with similar use patterns can be used, saving computing power without losing too much accuracy [13].

Some studies evaluate models where the DH load forecast is based on, among others, a weather forecast without managing the uncertainty involved with these. Some studies, e.g. [14] [15], use weather input as if forecasts were 100 % correct. In these cases, one should bear in mind that in a real scenario, a forecast is made based on a weather forecast, meaning that the uncertainty of the weather forecast is inherited by the heat load forecast. The effect was however studied by Dahl, Brun, Kirsebom, and Andresen, where predictions were made for the same data sets with both weather forecasts and actual weather outcomes. With their circumstances, the change to 100 % accurate forecasts led to a decrease of the total absolute error measure RMSE with 14 % [11].

2.3.3 Evaluation periods of DH load prediction models

In a study by Kurek et al., the evaluation of short-term (both 24 h and 72 h) forecasting models is divided into three seasons - winter, intermediate, and summer [16]. The models compared are among else an ANN, with different transformational preprocessing of input data. The study concluded that the models showed better performance during winter and summer since the dependency between heat load and outdoor temperature is

more fixed. During the winter, space heating is dependent on the outdoor temperature, and during the summer, no dependency can be seen. In the intermediate season, the dependency varies, making it harder for the models to predict the heat load.

In a study by Kristensen, Hedegaard, and Petersen, it is concluded that the model has specifically low accuracy during summer months, especially in an unusually warm July 2018. In the study, a long-term prediction model is trained on one year and evaluated on another one year, thus making the model weigh in the accuracy for different seasons equally. Furthermore, from the same study, it can be concluded as important to have representative training and test periods, as could be seen in the study where the summer season of the evaluation period may not be representative of a common summer - making the models come out as unrightfully inaccurate [17].

It can thus be concluded that different models can show different patterns and accuracy in different seasons. It is therefore important to have all seasons represented and equally weighted in the training and test set.

Since a cold spell can happen during different times in a winter season, certain calendar years can have two cold spells and other calendar years may not have any. Therefore the training should be whole heating seasons (June 1st to May 31st). Partly because nonlinear behavior in the winter can be seen, and if the model is based upon one or multiple DTs, which are unable to extrapolate, it is even more important [5].

2.3.4 Evaluating different ML algorithms

When evaluating Energy Predict and other heat load predictions, it is beneficial to have benchmark models to compare the predictions against. Heat load signature is one benchmark, as it is commonly used in DH utilities. But it could also be valuable to compare models to the ones that have shown promising results in academic research. To investigate which algorithms these could be, the studies mentioned by Mbiydzenyuy et al., where different ML algorithms have been compared, have been investigated. The conclusions from these studies are summarized in Table 2.1.

Table 2.1: Summary of studies where different ML algorithms' performance on heat load forecasting has been compared

Ref.	Time Horizon	Models	Eval. Metrics	Best Model
[18]	24 h	SVR, FFNN, LR	MAPE	SVR
[11]	15 h to 38 h	SVR, MLP, OLS	MAPE	SVR
[10]	24 h	SVR, RF, XGB, MLP etc.	RMSE	XGB
[19]	24 h	LR, RF, SVR, ANN	MAPE	ANN

In words, Idowu, Saguna, Ahlund, and Schelén showed promising results for SVR compared to FFNN and LR when looking at the performance measure MAPE. The

evaluation was done on several substations with a horizon of up to 24 h ahead [18]. Similar results were found by Dahl et al., who compared SVR, MLP, and OLS on 15 h to 38 h forecasts [11]. These results were seen again when Ziqing et al. compared, among others, SVR, RF, XGBoost (a certain implementation of GBDT), and MLP on 24 h forecasts and concluded that SVR performed the best when looking at the MAPE. However, it was outperformed by XGBoost when looking at the error measure RMSE [10]. There are studies where other models have shown better results than the SVR. Geysen, De Somer, Johansson, Brage, and Vanhoudt compared LR, RF, SVR, and ANN, concluding that ANN was the best performing of the set when looking at MAPE and on a 24 h horizon [19].

2.3.5 Long-term heat load prediction

The evaluating studies mentioned above have all been on models with forecast horizons of up to a few days. Nothing is said about how these models can be expected to perform on longer time horizons. Mbiydzenyuy et al. further state that there is no indication of any ML-driven analysis targeting a longer time scale - even though it could provide value when planning infrastructure and longer-term fuel stock [4].

In a study by Kristensen, Hedegaard, and Petersen, a prediction of building-specific DH load was made over one year with an hourly granularity [17]. Beyond the common weather input features mentioned earlier in this theory section, building properties such as U-value (a measure of the rate of transfer of heat through a structure divided by the difference in temperature across that structure) were used as input. With this input, a physics-based, contrary to a statistical ML-based, model was proposed. Even though these building-specific features, and the model not based on ML, are not within the scope of this thesis, some learnings from that study can be applied in this work. The overall performance, measured according to the guidelines on building energy models in ASHRAE 2014-14 was deemed both better than previous physics-based models and suitable for general energy planning purposes. It is stated, however, that the model did not perform better on short-term heat load forecasts compared to ML-based models. The explanation was that an ML approach can learn from training data in a way that includes non-physical aspects, such as societal behavior. In the physics-based model, for example, no consideration was given to which days were working days or not. The strength of such a physics-based model, however, is that it can provide predictions on heat load in a building that is yet to be built, or any other case where the amount of training data is insufficient.

There is also a possibility to use the insight from the prediction to later use as a projection. This distinction is made by Nateghi and Mukherjee, in an article where energy demand - and climate data is used to train a statistical model [20]. Contrary to previously mentioned models that utilize data with an hourly granularity, the training was done with many years worth of annual energy demand data and monthly weather data. This model is then

used together with different climate forecasts, projecting the heat demand in the future. However, no further discussion was provided on how this method can be used to provide benefits for energy companies or society at large. The evaluation that is done here is on historical data, where it is compared to the so-called mean-only model, i.e. that the energy demand for a year is equal to the mean energy demand on historical data.

2.3.6 Concluding Remarks

To summarize, even though many agree that the potential for ML implemented on DH data is higher, research on the subject of heat load prediction is limited to short-term heat load forecasts. More work is needed regarding the potential for ML-driven methods fulfilling other needs of DH utilities.

As the research is limited to short-term load forecasts, the development of evaluation frameworks for ML-driven methods predicting heat load is also limited to short-term load forecasts. However, by combining insights from other evaluation frameworks used for long-term predictions with insights regarding ML models, an evaluation framework can be developed. This approach will be taken in this report.

2.4 Evaluation Metrics

This report includes an assessment of how heat load prediction models are and could be evaluated. Therefore, the following theory section includes a description of different evaluation metrics that can be used for this purpose.

2.4.1 RMSE

As has been pointed out, for example by Mbiydzenyuy et al., Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) are the most commonly used metrics for evaluating prediction models for DH load [4]. This can also be seen in the studies mentioned above where different ML algorithms have been evaluated and only RMSE and/or MAPE have been used [10] [11] [18] [19]. The way to calculate RMSE can be seen in Eq. 2.5:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i^2)} \tag{2.5}$$

e_i being the residual between simulated and observed load (observed value minus simulated value) number i, and n the number of data points in the data set.

2.4.2 MAE

Related to RMSE, meaning that they are both absolute error measures, is the Mean Absolute Error (MAE). MAE is calculated as in Eq. 2.6:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{2.6}$$

e_i being the residual between simulated and observed heat load (observed value minus simulated value) number i, and n the number of data points in the data set.

Optimizing a model to perform on an MAE basis is equivalent to making the model search for the median in a distribution, contrary to the metric RMSE which instead optimizes for the mean. When forecasting symmetric distributions, the median and mean are equivalent, but in the not-so-uncommon case of forecasting skewed distributions, it is not [21].

2.4.3 MAPE

The other commonly used metric aside from RMSE is MAPE, calculated according to Eq. 2.7. The metric is calculated by averaging all errors on an absolute and percentage-wise (with the observed value as the denominator) basis.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}(100 \cdot \frac{|e_i|}{y_i}) \tag{2.7}$$

y_i being the heat load data point number i, e_i being the residual between simulated and observed heat load (observed value minus simulated value), and n the number of data points in the data set.

2.4.4 NMBE

The Normalized Mean Biased Error (NMBE) is calculated as Eq. 2.8 and is mentioned by Kristensen, Hedegaard, and Petersen as a way of calculating a model's bias, i.e. if the model under- or over-predicts on a general basis [17]. The measure generates a percentage-based value, with a value of zero if no bias can be seen, a positive value if the model under-predicts on a general basis, and a negative value if it instead over-predicts.

$$NMBE = \frac{1}{\bar{m}} \cdot \frac{\sum_{i=1}^{n}(e_i)}{n} \cdot 100(\%) \tag{2.8}$$

$\bar{m}$ being the mean observed heat load in the data set, e_i being the residual between simulated and observed heat load (observed value minus simulated value) number i, and n the number of data points in the data set.

This way of evaluating bias comes from the ASHRAE (American Society of Heating, Refrigerating, and Air-Conditioning Engineers) Guideline 14 regarding "Measurement of Energy and Demand Savings.", with the most recent version from 2014, from the related American FEMP (Federal Energy Management Program) and the IPMVP (International Performance Measurement and Verification Protocol). The protocols aimed to reduce energy - and water consumption by quantifying both the statistical performance of energy-saving tools as well as the actual energy savings. Further, the intention of ASHRAE Guideline 14 is to provide guidance on minimum acceptable levels of performance for determining energy and demand savings, using measurements". It is more technical than the other documents, and as a result, the majority of the scientific community uses this document in their research [22].

The evaluation metrics ASHRAE Guideline 14 suggests for forecasting models are the Coefficient of Variation of the Root Mean Squared Error (CVRMSE, which will be further discussed below), NMBE, and Coefficient of Determination (R^2, which will also be further discussed below). The baseline performance that is needed, as a minimum acceptable level for building energy models, is for the NMBE, ± 5 % for monthly heat load values and ± 10 % for hourly heat load values [22].

2.4.5 CVRMSE

As stated above, CVRMSE is the second evaluation metric that is mentioned in ASHRAE Guidelines 14. In the documents, it is stated that CVRMSE is a measure of how much the errors vary and that it "gives an indication of the model's ability to verify the accuracy of the model.". The baseline performance that is needed, as a minimum acceptable level for building energy models is, 15 % for monthly heat load values and ± 30 % for hourly heat load values [22].

In words, it uses the aforementioned, and widely known, measure RMSE but divided by the mean of observed values, resulting in a strictly positive percentage-based value. Lower values of this metric are desirable. It is also commonly mentioned as Normalized Root Mean Squared Error.

$$CVRMSE = \frac{1}{\bar{m}} \sqrt{\frac{\sum_{i=1}^{n} (e_i)^2}{n}} \cdot 100(\%) \tag{2.9}$$

$\bar{m}$ being the mean observed heat load in the data set, e_i being the residual between simulated and observed heat load (observed value minus simulated value) number i, and n the number of data points in the data set.

2.4.6 RN_RMSE

RN_RMSE is calculated as in Eq. 2.10 and is just like CVRMSE based upon the RMSE value. However, instead of the mean observed value as a denominator, it uses the range of observed values (meaning the difference between the highest and lowest observed value).

$$RN\_RMSE = \frac{1}{y_{max} - y_{min}} \sqrt{\frac{\sum_{i=1}^{n} (e_i)^2}{n}} \cdot 100(\%) \tag{2.10}$$

y_{max} and y_{min} being the highest and lowest observed heat load, respectively, in the data set, e_i being the residual between simulated and observed heat load (observed value minus simulated value) number i, and n the number of data points in the data set.

2.4.7 R^2

The last evaluation metric that will be mentioned is the coefficient of determination R^2. It is calculated as in Eq.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (e_i)^2}{\sum_{i=1}^{n} (y_i - \bar{m})^2} \tag{2.11}$$

$\bar{m}$ being the mean observed heat load in the data set, e_i being the residual between simulated and observed heat load (observed value minus simulated value) number i, n the number of data points in the data set, and y_i being the heat load data point number i.

The baseline performance that is needed, as a minimum acceptable level for building energy models is 0.75 [22].

2.4.8 Comparing evaluation metrics

Different evaluation metrics have been mentioned, both how they are defined and how they are calculated. In this subsection, these evaluation metrics will be compared to one another, discussing what strengths and weaknesses they imply when incorporated into an analytical and evaluating framework. RMSE and MAE will be compared to each other as they both are dependent on the scale and absolute. MAPE will be discussed separately. At last, the pair of CVRMSE and NMBE will be compared to the pair of RN_RMSE and R^2 together as they both stem from the purpose of evaluating building energy models.

As a first remark, the metrics discussed below are symmetric, meaning that over- and underestimations are equally valued in the total error metric. However, as mentioned

above, the consequences of forecast errors are not equally severe for a DH utility. A loss function, based on the asymmetry of costs linked to over- and underestimations, could be proposed as a way of handling the dilemma. But since cost conditions assumingly vary between DHNs, it would need to be a network-specific weight function. This is regarded as beyond the scope of the report.

It should be mentioned, however, that the definition of a "symmetric" evaluation metric seems rather unclear. In the work by Hewamalage, Ackermann, and Bergmeir, which will be commonly cited in this theory subsection, an evaluation metric is regarded as symmetric if the value of the observation y_i and simulation $\hat{y}_i$ can switch and still yield the same resulting performance [21]. For example, RMSE is regarded as symmetric since the squared difference between y_i and $\hat{y}_i$ is the same, even if values are switched. On the contrary, MAPE is regarded as asymmetric since every residual is divided by the observed value to render a percentage-based error value. In the article, some evaluation metrics are provided which both have the relative, independent of scale, aspects of MAPE, making the performance comparable between different data sets, and the symmetry of the RMSE. Additionally, some of these metrics also have the dynamic of not valuing over- and underestimations equally. However, it is not explained what the benefits of symmetric evaluation metrics are, and most importantly, as is mentioned in the article, many of the evaluation metrics that are mentioned in the article lack interpretability for a real-world business application. Thus, these metrics will not be discussed in the report.

RMSE and MAE

A common criticism of using RMSE as an error measure is that it has no real-world meaning, in contrast to the MAE which is more easily interpreted as the average size of the residual. However, RMSE being less useful than MAE is necessarily not true as the RMSE is equivalent to the standard deviation which can be used in other calculations and frameworks. For example, if there is reason to believe that the errors follow a known distribution, it is possible to calculate a prediction interval, which has been mentioned as desirable.

Additionally, optimizing a model to perform on an RMSE basis is equivalent to making the model search for the mean in a distribution, which in many cases is desirable compared to the MAE, which instead optimizes for the median. When optimizing for the mean, a model which generally performs well can end up being the worst due to large errors when predicting outliers. Due to considering the square of the error, a metric like RMSE is more susceptible to outliers than an absolute error metric such as MAE. Therefore, if the outliers are of interest and capturing them is important, using the squared operator before aggregating the errors is the better option. If they should not be regarded in the error metric, however, a metric like MAE is more suitable [21].

On the subject of optimizing the model, it has been argued that models using an L2 loss function, i.e. minimizing the squared residual in the training set, are made to perform

better on metrics such as RMSE and should thus be evaluated on those metrics [21]. However, an evaluation metric is not for the sake of what is most ”fair” to the model it is used upon but instead what is most beneficial for the actual use case. One should begin with the need addressed by the model when assessing what evaluation metrics to use. It is, however, a good remark for further work when applying evaluation frameworks to ML models, that one should keep in mind what error metric different models strive to minimize and evaluate them accordingly.

Another downside to RMSE as an evaluation metric, adding to those already mentioned, is the scale dependency. It is generally not possible to compare RMSE between data sets and substations since the overall heat load size correlates with how big the errors are. This is troublesome when evaluating models on several data sets with different scales. A calculated mean RMSE will in that case make it more important to predict the data sets with higher heat load correctly. That may not always be desired [21].

MAPE

MAPE lacks in the way that an equally sized error is deemed worse on low load values than on high ones. As previously mentioned, in the use case of DH it is not necessarily so that a forecasting model needs to be accurate on a relative basis. It may be more important to be accurate on high heat loads than on lower ones since that is when the use of peak burners comes into question.

While MAPE is less susceptible to high-value outliers than RMSE, both because of the non-squared dynamic but also because of all errors being divided by the observed value, it is very susceptible to low-value outliers. In the extreme case of zero-value heat loads, which is not uncommon when looking at, for example, specific substations in a DHN, the MAPE value reaches infinity. This error does therefore not only ruin the error metric at the specific data point but propagates through the error metric calculation of the whole data set, making the metric not as robust as others.

A strength of MAPE is when predicting a non-stationary process, i.e. a process with a change in mean and standard deviation over time [21]. With structural and seasonal trends in DH, the heat load can hardly be regarded as a stationary process. Trends can both be seen long-term (several years) both from more efficient buildings and climate change, medium-term because of weather, which has a higher variance in winter than in summer, and short-term (peak demand in morning hours). It can be argued that these non-stationarities imply that there should be different weighting of an error depending on where in the time series it is. In MAPE, this is done automatically since every residual is divided by the observed heat load.

CVRMSE and NMBE vs RN_RMSE and R^2

The combination of NMBE and CVRMSE is a widely appreciated way of evaluating building energy models [23]. Further, it is also mentioned in the works by Hewamalage, Ackermann, and Bergmeir (although it is there named Normalized Root Mean Squared Error) for time-series forecasting. In the report, CVRMSE is deemed suitable for use cases where comparability as well as optimizing for the mean is important, being both independent of scale and resembling the L2 - loss function [21].

In an article by Chakraborty and Elzarka, it is argued that NMBE and CVRMSE are insufficient as metrics when assessing the behavior of a heat load model on a system level [24] and that the RN_RMSE is a more robust alternative to CVRMSE. Both metrics, CVRMSE and RN_RMSE can handle multiplicative differences but CVRMSE changes with an additive difference between data sets (assuming constant variance within the data set), which is not true for RN_RMSE. It is argued that the difficulty of predicting the heat load is not dependent on the absolute heat load values but instead on the variance these values show. The variance is, in turn, more connected to the range of the observed heat load value than the absolute value.

Chakraborty and Elzarka further argue that R^2, and not NMBE, should be used as a complement to RN_RMSE [24]. An evaluation approach is suggested in the sense that when the RN_RMSE value is low even though the R^2 value is poor, it can be seen as a sign of non-linear relationships between observed and simulated values. It is explained that bad performance in terms of R^2 value generally results from high bias, high variance, or non-linear relationships between observed and simulated values. If the RN_RMSE value is low, the two former alternatives can be ruled out and thus a conclusion can be drawn, a conclusion that can provide further information on what could be improved in the model. However, it should be mentioned that even though the proposed metrics and framework can provide information that the metrics from ASHRAE Guideline 14 can not, the opposite is also true since the framework is unable to recognize, for example, biases in the model.

Chapter 3

Present Heat Load Predictions used by District Heating utilities

3.1 Method

To be able to answer the questions *"For what purposes are heat load predictions used in DH utilities?"* and *"Are these heat load prediction models validated as sufficiently accurate according to the DH utilities?"*, a survey study was carried out with DH utilities in Sweden. A study in this form enabled the information gathering to include more participants than if an interview study was chosen.

3.1.1 Regarding the selection

The selection of participants was constrained to Swedish DH utilities, more specifically the DH companies registered by the Swedish energy industry organization *Energiföretagen Sverige*.

For every company, the e-mail address for the one in charge of DH operation was searched for on the company's website. In case that was not available, the e-mail address for the CEO or, lastly, a general contact e-mail was used in order to reach out to the DH utilities. For some companies, an e-mail address was not found. The total number of invited DH utilities reached 103.

While 103 DH utilities were invited to participate, 28 answered the survey. Additionally, 4 respondents answered via e-mail saying that they either managed a too small DHN (2 respondents), did not have the ability to adjust production since it solely derived from industrial waste heat (1 respondent), or had constant overproduction (1 respondent).

3.1.2 Regarding the questions

The survey consisted of seven questions. All but the first of these questions were left open-ended, making it possible for the respondents to fill in a free text answer. This was chosen to better capture the respondents' own experiences and attitudes. At the same time, it left room for interpretation by the researcher when the answers were to be analyzed and potentially categorized into different groups of answers.

In Appendix A, one can see the complete questionnaire in Swedish as well as an interpretation in English. Further, in the English version, not only are the questions presented but also accompanied by the reasoning as to why they were asked.

3.1.3 Sources of errors

The different sources of errors, summing up to a total survey error, can be divided into four main types: sampling error, coverage error, non-response error, and measurement error.

The sampling error occurs since survey participants represent only a part, or a sample, of the population of interest. Thus, participants' answers to the questionnaire represent only a part of the population's answers, not the answers of the entire population. There are therefore possible answers in the population that are not seen in the answers to the questionnaire.

Coverage error occurs when the participants of the survey and their distribution is not fully representative of the distribution of the population. As a result, the results from the survey may be unrightfully skewed.

The non-response error occurs when invited participants do not answer the survey, resulting, once again, in the population being not fully represented by the participants of the survey. The non-response error can refer to both item non-responses or unit non-responses. Item non-responses occur when questions in the questionnaire are not answered by participants, unit non-responses when participants refuse the survey as a whole.

At last, the measurement error occurs when the participants answer the questionnaire in a way that is different from the truth. This creates a discrepancy between the truth and the results estimated by the surveyor.

The sampling error and coverage errors of the survey conducted in this report may be considered small since the survey was sent to almost the entire population of Swedish DH utilities. There are approximately 130 DH companies, members of Energiföretagen Sverige. These members constitute approximately 99 % of all DH income in Sweden. Thus, the coverage of the survey is nearly equivalent to the whole population of interest.

The non-response error may be considered more significant, since approximately 70 % (75 out of 103) of the utilities did not answer the questionnaire. It could be suspected that the DH utilities who do not use heat load predictions in their operation did not feel targeted by the questionnaire and are thus over-represented in the non-responding group. This should be taken into account when assessing the results of the survey.

At last, the measurement error should also be taken into account when analyzing the results of the survey. Especially since the definitions of certain words are assumingly easy to misinterpret. The risk of misconceptions was addressed by providing the respondents with an introductory section, defining words that were assumingly easy to interpret. Further, different questions were asked in the questionnaire to validate the answers given, and for every question, there was room for the respondents to clarify their answers if needed.

3.1.4 Analysis

The first question of the questionnaire was asked in order to answer the question of *"For what purposes are heat load predictions used in DH utilities?"*. The answers collected to this question were thus not analyzed in a quantitative way (How often or how usual are the different purposes?) but rather it was analyzed which purposes were mentioned by the DH utility, with no respect to how often they were mentioned. This is in line with how the Research Question is formulated.

The second and third questions were analyzed in a validating way so that the participants' way of using the word *heat load predictions* corresponded to the way that it is used in this research study. These answers will not be disclosed in the Results section.

The fourth question was analyzed in the way that all answers were divided into the categories *"Insufficient"*, *"Sufficient"*, *"Sufficient as long as the input weather parameters are correct"*. The categorization was done to answer the question *"Are these heat load prediction models validated as sufficiently accurate according to the DH utilities?"*.

The answers to the fifth question were categorized as *"No answer as the utility did not use heat load predictions"*, *"Yes, the accuracy has been evaluated"*, *"No, the accuracy has not been evaluated"*, and *"We trust our system provider to validate the model"*.

The answers to the sixth question were categorized as *"No answer"*, *"Yes, higher accuracy would increase the number of use cases for the heat load prediction"*, *"No, higher accuracy would not increase the number of use cases for the heat load prediction"*, and *"Maybe, we don't know"*.

The seventh question's answers will not be disclosed in the Result section as the question was only asked to validate previous answers.

3.2 Results

The results of the first question can be concluded as:

- Production Planning
- Sales Planning (budgeting was included in this category)
- Dimensioning of equipment
- Dimensioning of network
- Dimensioning of production facilities
- As a step in fault detection

The results of the fourth, fifth, and sixth questions can be seen in Figures 3.1, 3.2, and 3.3 respectively.

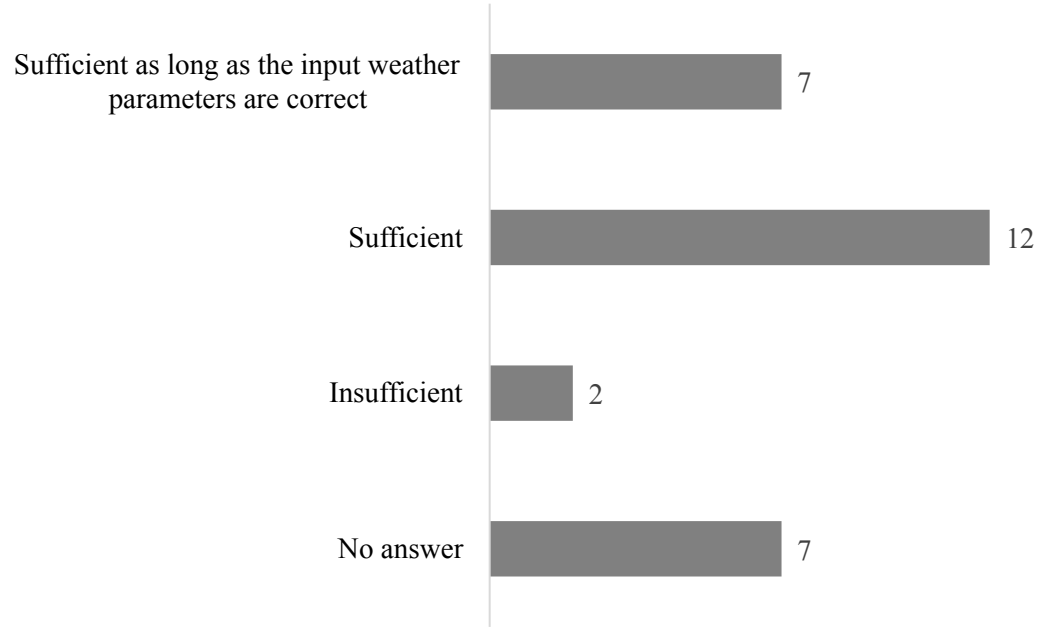

Figure 3.1: Categorized answers to the fourth questionnaire question: *"Do you feel like the accuracy (how predictions compare to the outcome) is sufficient for the use case of the heat load prediction? Feel free to motivate"*

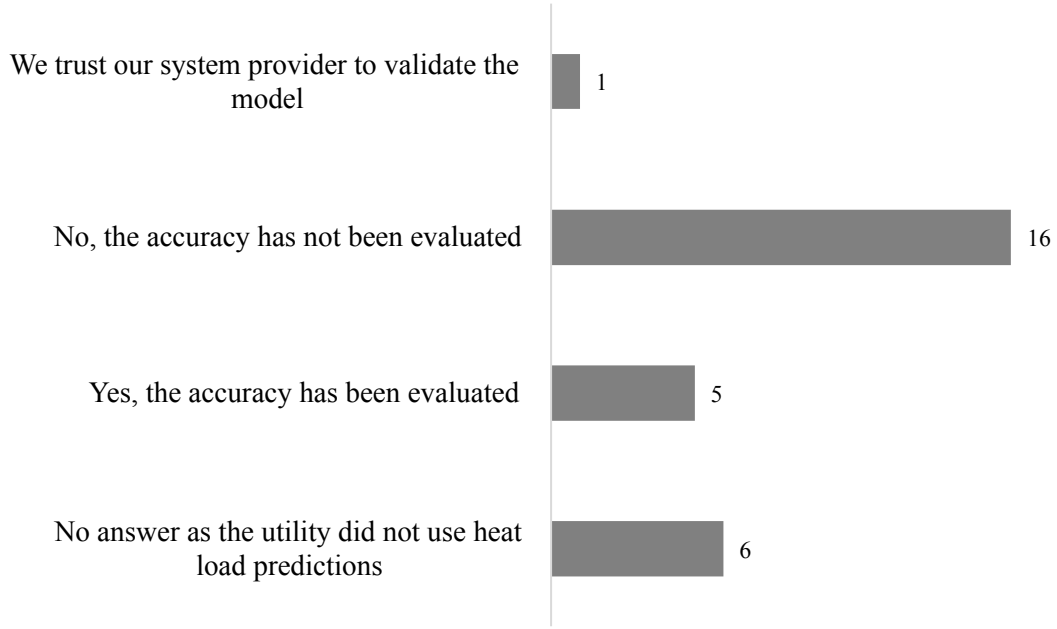

Figure 3.2: Categorized answers to the fifth questionnaire question: *"Have you evaluated the accuracy of the heat load predictions? In that case, how?"*

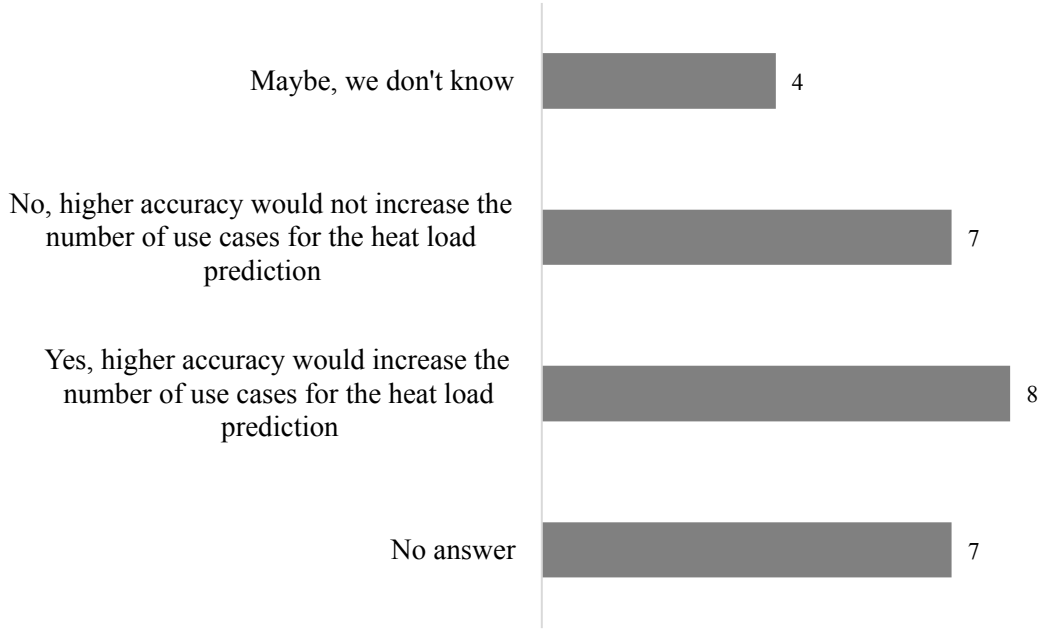

Figure 3.3: Categorized answers to the sixth questionnaire question: *"Had an increased accuracy enabled additional use cases than those you have today? In that case, which and why?*

3.3 Conclusion

As can be seen in Section 3.2, the answer to Research Question 1 is that the survey respondents are currently using heat load forecasts for one, or a combination, of the six different purposes mentioned in the section. However, the survey also showed that the majority of the participants had not evaluated the performance of the model they are currently using. This means that the answer to Research Question Number 2 (Are these heat load prediction models validated as sufficiently accurate according to the DH utilities), according to the survey, is that they are not. It is not necessarily so that the heat load prediction models are insufficiently accurate, no conclusions will be drawn in that regard, but rather that there is no evaluation framework developed.

As concluded in the answers to Research Question Number 1, heat load predictions can provide DH utilities with a way to dimension as well as to plan sales. Yet no studies have been seen on these matters and no consensus has been developed on appropriate evaluation frameworks, according to the studies by Mbyidzenyuy et al. and Ntakolia et al. [4] [7]. Moving forward in this report, the focus will be on implementing an evaluation framework for heat load predictions with purposes of dimensioning and sales planning.

Chapter 4

Energy Predict by Utilifeed

Utilifeed is a software provider, offering a so-called Business Intelligence tool to DH companies. In the software platform, a series of different analyzing tools are provided, visualizing data in different ways. Adding to the data visualization, there is a data-based ML algorithm called Energy Predict that is used for different purposes in the platform. These purposes include:

- Short-Term Load Forecasts with Prediction Interval
- Design Load
- Normal Year Projections
- Sales Projections
- Fault Detection

In this chapter, the functions of Energy Predict will be discussed. At the end of the chapter, a concluding remark will be given on how these functions relate to the purposes of dimensioning and sales planning. It is the aim to develop evaluation frameworks for these purposes, as concluded in Section 3.3.

However, how Energy Predict is implemented and what ML algorithms it is based upon will not be discussed due to non-disclosure reasons.

4.1 Short-Term Load Forecasts with Prediction Interval

The Short-Term Load Forecasts powered by Energy Predict is an ML algorithm that, in the standard case, has a training period consisting of the last 18 months and a simulation period of seven days ahead. The training- and simulation period are, however, customizable so that the user can evaluate the model on data that has been seen before - or as a way of filling in missing data points when invoicing their customers. Input parameters for the algorithm include both weather and calendar parameters, using actual

weather outcomes when predicting in hindsight and weather forecasts (gathered from weather institutes) when not.

The function enables DH utilities to plan their short-term production. As mentioned in the theory section, an accurate prediction makes the DH utility able to align heat production with demand, causing lower temperatures and therefore higher efficiency. Production planning is also the case of temporarily overshooting the demand, storing heat in the network so that peak boilers with high marginal costs don't need to be used.

Along with the ordinary forecast, Energy Predict also provides a prediction interval, i.e. a probabilistic forecast, addressing the uncertainty of the prediction. With the costs associated with overshooting demand being lower than the ones associated with not fulfilling the needs of the customers, a cost-benefit analysis is needed when planning production - a probabilistic forecast is an enabler for such analysis.

4.2 Design Load

When dimensioning infrastructure and production capability in a DHN, one must take into consideration that the network must fulfill the needs of customers even during peak heat load values. Energy Predict has a function called Design Load, which trains on a certain calendar year and is then simulated over the twenty years between 2000 and 2019, setting the highest heat load value simulated as the dimensioning heat load. The heat load can be simulated for substations, clusters of substations, as well as the whole network. The heat load can then be used when dimensioning measuring instruments, pipe capacity, and/or production capacity.

4.3 Normalization

When comparing the total heat load in a DHN for two different years, the difference between them can generally be described by two factors: changes in buildings (or the DHN as a whole) and differences in weather. Energy Predict has a tool called Normalization that can be used to separate these two effects and analyze them separately. It is done by training on the period you want to normalize and simulate the model on a reference period. The reference period is a normalized year that is based on the twenty years between 2000 and 2019. This means that a normalized yearly value, for a certain year, can be interpreted as what the yearly average value would be for the period 2000 to 2019 given that the substation (or substations) is behaving as it did during the training period. The difference between the measured total heat load for the specific year and the normalized total heat load for the year can then be interpreted as the effect the weather had on the heat load, at least compared to an average year between 2000 and 2019. This

insight about substations and a DHN as a whole can be useful when planning further capacity, as well as for general DHN analyses.

4.4 Sales Projections

Different pricing models can be used to incentivize customers as desired by a DH utility. One can see a trend that DH utilities not only charge their customers for the heat load but also (or in some cases solely) flow. This model incentivizes the customers to use as much heat as possible from the flow that they get, leading to lower return temperatures and thus higher efficiency in the DHN. Another pricing model is to charge more for heat load during peak hours (mornings) or peak seasons (winter) as a way to incentivize customers to steer their demand away from critical hours in the DHN.

When analyzing different price models, projecting how different price models would affect the total income, Energy Predict has a function called Sales Projections. It is done by training a model on hourly data over one year, simulating for another one-year period or a normalized year. In this way, DH companies can predict how their income will be affected by implementing different price models - enabling financial planning.

4.5 Fault Detection

The possibility to detect faults via Energy Predict is a rather new function, released during the writing of this report. Although very interesting, it will not be addressed further in this report as it is not predicting any load.

4.6 Concluding remarks

Energy Predict has functions that are yet to be researched in an academic setting. The Design Load is one example as it corresponds to the purpose of dimensioning. Normalization and Pricing are two others, as they correspond to sales planning.

Chapter 5

Proposed Performance Evaluation Framework

Below, an evaluation framework for the functions Design Load, Normalization, and Sales Projections, described in Chapter 4, will be proposed. These functions were chosen as they correspond to the purposes of dimensioning and sales planning, concluded in Chapter 3 to be purposes that are both used in DH utilities but lacking a developed evaluation framework. The proposal will be based on insights from the theory in Chapter 2.

5.1 Framework for evaluating Sales Projections and Normal Year Projections

Due to similarities between the functions Sales Projections and Normal Year Projections, the proposed evaluation framework will not handle them separately. Both functions attempt to learn the behavior of a network/substation and then apply that behavior to another test period. In the case of Sales Projection, the test set is a specific year and in the case of Normal Year Projections, it is an imaginary ”normal” year. In both functions, the most important metric to perform upon is the accumulated heat load over one year as that is the most contributing factor to DH utility income [5]. However, it is still necessary to have an hourly granularity as price models are not based on a constant charging rate per heat load value. As the DH utility's income also can depend on power output and the distribution of heat load during the year, it is important to have a model that resembles the hourly variation of heat load. Since the function of Sales Projections is equivalent to calculating the income, and Normal Year Projections can be used to calculate a ”normal year income” they both need to be implemented on data sets with an hourly granularity.

Over time, DHNs change as the number of substations and their behavior change. When using the functions Sales Projections and Normal Year Projections, the model's purpose is to catch that behavior and project it on a different time period. Thus, admittedly somewhat abstractly explained, it is the model's ability to catch DH heat load behavior

that needs to be evaluated. But as previously stated, a DHN's behavior changes over time, making it hard to know if the model's error when predicting over a year is due to its inability to catch the behavior or if the behavior has changed. To combat this, it is proposed that models should, in the case of evaluating long-term predictions, be evaluated by calculating the error metric NMBE twice per substation/DHN. The first time, the NMBE will be calculated on a certain test period with a certain training period. The second time, the training - and test periods are switched. Thus, the change in behavior between different periods will negate each other and the average error metric will thus be the actual bias of the model.

The mean NMBE and the mean CVRMSE on a large number of predictions done on substation DH load are proposed as evaluation metrics. These are chosen since the overall bias is of interest, as it is a metric resembling the model's ability to predict the accumulated heat load. The proposal is in line with the ASHRAE 2014-14 standard but also with the works by Hewamalage, Ackermann, and Bergmeir as well as Kristensen, Hedegaard, and Petersen [21] [17]. The most important metric for the models to perform upon, in order to fulfill their purpose, is the NMBE. But since the hourly variation of the heat load is important as well, a CVRMSE below the value given by the standard ASHRAE 2014-14 (30 %) is proposed to be a limitation for the heat load prediction model.

Firstly, before assessing a model's performance on CVRMSE, it is however proposed that the NMBE on different seasons (winter, spring, summer, autumn), is calculated. The proposal is in line with the way that has been done by Kristensen, Hedegaard, and Petersen as well as (to some degree) Kurek et al. [17] [16], namely calculating an error measure for every three months corresponding to a season (December to February is winter, March to May is spring, etc.). If there are different biases during different seasons, it is proposed that an NMBE calculation is based on whole years of simulation as well as one year of training, the periods should not include fractions of a year.

In the theory section regarding DH and heat load signature, it was shown that different models show different capabilities of resembling a heat load-to-outdoor temperature relationship that does not follow the likes of a heat load signature. To address this robustness, it is proposed that substations in the data set are divided into two categories - those whose heat load pattern follows a heat load signature and those who do not. By calculating the mean CVRMSE on these two categories separately, further insight into a model's robustness in this matter can be provided. The mean NMBE is assumed to not differ as much.

If a model shows poor performance, either according to the ASHRAE standard or the specific needs of the DH utility, improvements should be considered. Potential improvements can be made by, for example, changing the ML algorithm or adding more input features to the model. Poor performance can also be a result of some substations having changed behavior from the training period to the test period. It is therefore advised to, as a last step in the evaluation framework, plot the distribution of NMBEs

for different heat load predictions as a histogram. This process can potentially make it possible to identify outliers where the model is unrightfully deemed inaccurate.

5.2 Framework for evaluating Peak predictions

As the function Design Load is implemented via projecting a hypothetical high heat load between 2000 and 2019, a comparison between the predicted and actual heat load at the time is problematic. Not only would the comparison need twenty years worth of data, but it would also assume that the behavior of the DHN or substation had not changed over the course of twenty years. This is considered unviable and instead of the evaluation framework aiming to evaluate a model's ability to project a heat load twenty years in hindsight, it will aim to evaluate a model's general performance on peaks. The ability to predict peaks is not only applicable when predicting the dimensional heat load but also in general production planning as the need for production planning increase with the heat load value (see the introductory subsection with the title *The need for Heat Load Predictions*).

The definition of *peaks* can be considered somewhat fluid. As has been mentioned, the dimensional outdoor temperature is standardized as the average temperature of the 0.08 % coldest temperatures during a year with hourly granularity. To be in line with this standard, somewhat enabling comparability between the two predictions, the 0.08 % highest heat loads will be considered as peaks in this proposed evaluation framework.

In much the same way that measures independent of scale were needed in the framework for evaluating Sales Projections and Normal Year Projections, enabling an average error measure to be calculated for all substation predictions, it is also needed in this framework. Furthermore, since the outliers are of interest and capturing them is the essence of predicting peaks, a squared error measure is argued to be more suitable than the means of MAPE or CVMAE. At last, since the peak heat loads in a data set can have values close together, RN_RMSE would not be as robust as the CVRMSE. Therefore, the evaluation metric CVRMSE is proposed for this evaluation framework.

As certain models are expected to lack the ability to extrapolate (models built upon DTs), it becomes increasingly important making sure that the training set includes cold spells where the heat load is assumingly high. A cold spell can happen during different times in a winter season, some winters in December and some winters in February. It is therefore proposed in this evaluation framework that the training set should be a set of broken calendar years, broken outside of the heating season. In this report, a broken calendar from the 1st of June to the 31st of May will be used.

To assess the risks of not including a sufficiently cold outdoor temperature (and thus a high heat load value) in the training set, it is proposed to identify substations where there is a distinct difference between the maximum heat load observed (correlating with the

minimum temperature observed) on the different time periods. By calculating the mean CVRMSE for the peaks in those cases where the test period has a higher maximum heat load observed than in the training period, and vice versa, the model's robustness to high maximum heat loads in the test period can be assessed.

In similarity with the other framework, the capability of resembling different heat load-to-outdoor temperature patterns will be assessed. Substations will be categorized by their heat load pattern, those where the heat load diverges from the heat load signature when temperature decreases, and those where the heat load is aligned with the heat load signature. By calculating the mean CVRMSE for the peaks on these two categories separately, further insight into a model's robustness in this matter can be provided.

If a model shows poor performance according to this evaluation framework, improvements can again potentially be made by either changing the ML algorithm or adding more input features to the model. Another possibility is implementing a model specifically designated to only train and simulate peaks. How that would be done, however, is beyond the scope of the report.

Chapter 6

Evaluating different Machine Learning models

To showcase the evaluation framework proposed in Chapter 5, different ML algorithms were evaluated, eventually providing further insight into the subject of long-term DH load predictions.

6.1 Method

6.1.1 Model Implementation

Four algorithms were chosen for the evaluation. The first was Energy Predict by Utilifeed as it is known to be used in the operation of different DH utilities. The second was heat load signature according to the DH theory section where it was stated that the model is the industry standard in DH utilities. The third and fourth were SVR and XGBoost (an implementation of a GBDT) as they are two algorithms that, in the theory section were mentioned as promising algorithms (albeit for short-term heat load forecasts and not long-term heat load predictions).

Energy Predict

The implementation of the algorithm powering Energy Predict will not be discussed in this thesis due to non-disclosure reasons.

Heat Load Signature

The heat load signature was for every training set implemented according to how it was described in the theory section. However, since the balance temperature is not known, no heat loads occurring between outdoor temperatures of 12.5 °C and 17.5 °C were included

in the calculations. Instead, the heat loads that occurred above an outdoor temperature of 17.5 °C were considered for the calculated constant heat load. Respectively, the heat loads that occurred below an outdoor temperature of 12.5 °C were considered for the calculated linear negative correlation between heat load and outdoor temperature. By intersecting these two load relationships, a heat load signature was derived.

SVR and XGBoost

As the implementation and comparison of different ML algorithms is not the main question for the report, but instead the evaluation framework for these models, only a limited description will be given regarding the implementation of SVR and XGBoost. Further information on the concepts *gridsearch* and *cross-validation* and setting the parameters right can be found in the works by Müller and Guido [6].

The SVR was implemented with the scikit-learn package in Python. As a first step, the data was preprocessed by using the method StandardScaler. After the preprocessing, the model was fitted to the training data, using gridsearch cross-validation to fit the parameters in every prediction. The parameter grid was implemented in line with what has been done by Ziqing et al., Dahl et al., and Idowu et al., with an overlapping range for the parameters C (range of 1 to 5), γ (range of 0.0001 to 0.01), and ϵ (range of 0.01 to 0.05) [10] [11] [18].

The XGBoost was also implemented with the xgboost package in Python. The parameters used were the preset ones, aside from the number of estimators being 1000 and the learning rate being 0.1.

6.1.2 Data gathering

Heat load data was provided by Utilifeed's customers and other data on input features was provided by Utilifeed. In total, 36 different substations from four different utilities were studied. For every substation, a prediction was made in two ways. The first prediction was made with the period 2021-06-01 to 2022-05-31 as test period and the period 2020-06-01 to 2021-05-31 as training period. The second prediction was made oppositely, predicting the period 2020-06-01 to 2021-05-31 and training on the period 2021-06-01 to 2022-05-31.

For all substations, the heat load pattern against outdoor temperature was analyzed. Based on this analysis, all substations were divided into two categories - those where the heat load diverged from a heat load signature as outdoor temperature decreased and those that did not. 9 substations were assigned to the former category and 27 to the latter.

The same input data was used for the algorithms Energy Predict, SVR, and XGBoost. These correspond to the ones mentioned as beneficial in the literature study.

6.2 Results from the evaluation framework

6.2.1 Sales Projections and Normal Year Projections

As a first step, the mean NMBE was calculated for different seasons and the results can be seen in Table 6.1.

Table 6.1: Mean bias, measured as NMBE, for different predicting models for different seasons

Seasons	Energy Predict	Heat Load Signature	XGBoost	SVR
Winter	0.29 %	5.25 %	−0.31 %	2.55 %
Spring	−1.97 %	−7.94 %	−3.90 %	0.75 %
Summer	−7.83 %	−70.81 %	17.47 %	4.65 %
Autumn	−0.48 %	2.59 %	−0.76 %	2.89 %

As there was a considerable difference between the NMBE of different seasons, at least for models such as the XGBoost and specifically the heat load signature, it was considered appropriate to always train and test on whole-year periods. The mean NMBE and - CVRMSE can be seen in Table 6.2.

Table 6.2: Mean bias, measured as NMBE, and mean error, measured as CVRMSE, for different predicting models

Metric	Energy Predict	Heat Load Signature	XGBoost	SVR
Mean NMBE	−0.35 %	−0.84 %	−1.57 %	2.68 %
Mean CVRMSE	32.23 %	43.18 %	33.46 %	35.44 %

It can be seen in Table 6.2 that the bias measured as the NMBE of the different models is not considered high for any model, although the lowest (in absolute terms) for Energy Predict. Furthermore, none of the models has an accuracy, in terms of CVRMSE, below the threshold limit as stated by ASHRAE 2014-14. However, since Energy Predict is the best-performing model according to both error measures, it comes out as the best-performing model according to this framework.

The models' performance on substations with different heat load patterns against outdoor temperature can be seen in Table 6.3.

Table 6.3: Mean CVRMSE for different predicting models, for substations with heat load pattern corresponding to the one of a heat load signature (A) and for those not corresponding (B)

Case	Energy Predict	Heat Load Signature	XGBoost	SVR
A	28.98 %	36.20 %	30.40 %	30.29 %
B	41.97 %	64.13 %	42.62 %	50.88 %

As seen in Table 6.3, Energy Predict and XGBoost show more robustness compared to the other models when evaluated on substations with a heat load pattern not corresponding to the one of a heat load signature. Compared to SVR and heat load signature, their performance did not change as remarkably. However, the increased CVRMSE between cases A and B is considerable for all models.

To illustrate what was stated in the section above regarding behavioral errors negating each other between time periods a histogram of the NMBEs for all predictions is shown in Figure 6.1, n. In Figure 6.2 and 6.3, a specific substation has been isolated in both cases, illustrating how a high bias towards a certain direction is oppositely high in the other direction when switching the training and test period, assumingly a sign of behavioral change.

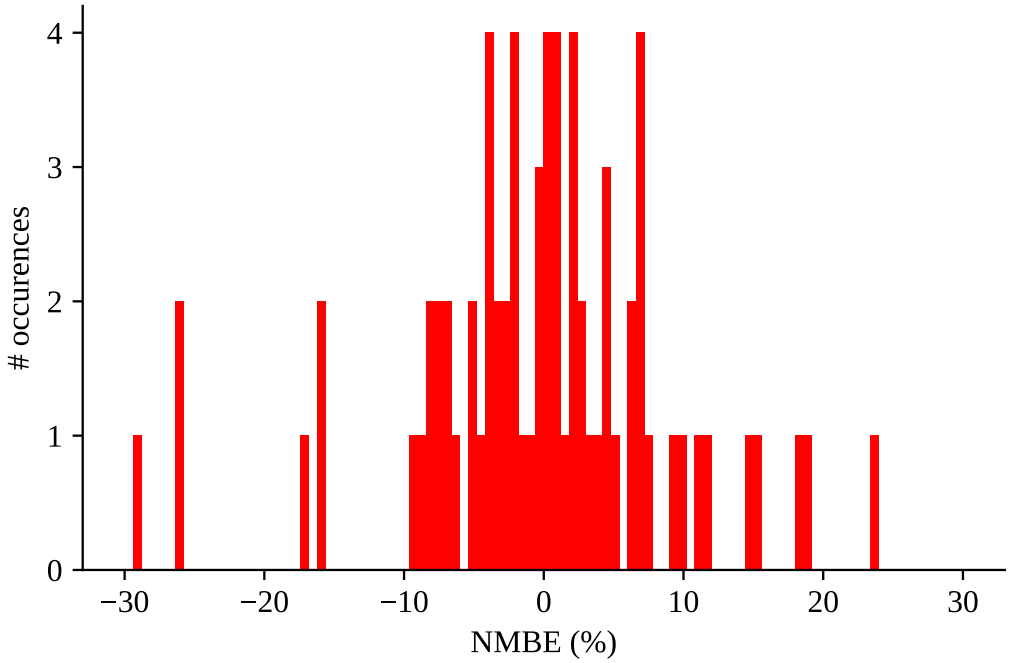

Figure 6.1: Histogram over calculated NMBEs

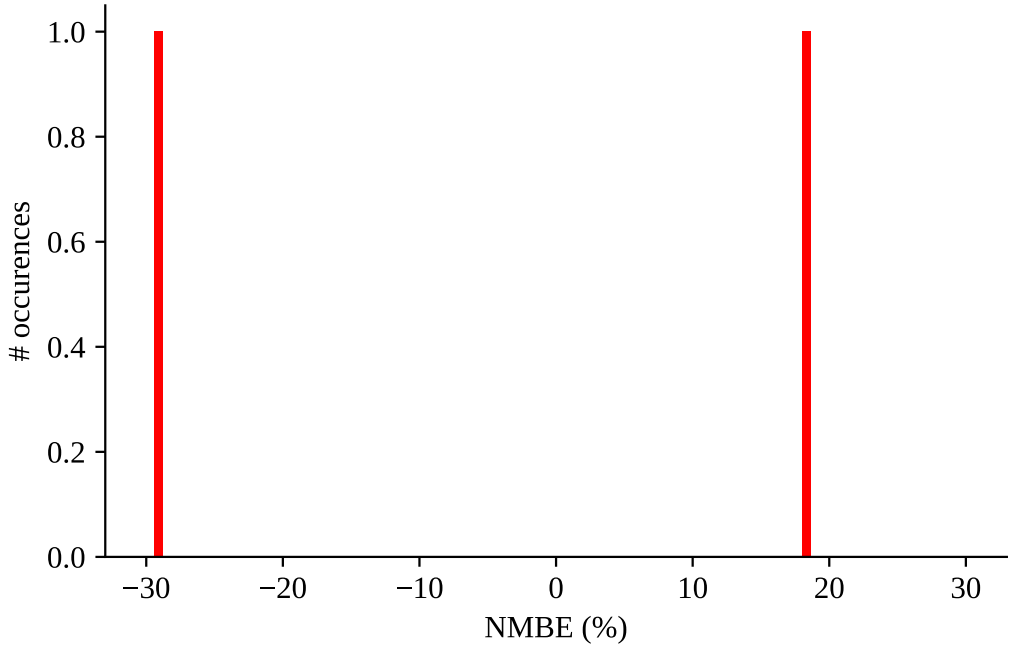

Figure 6.2: The NMBEs calculated twice for a substation, with training and test period switched. To be compared with Figure 6.1

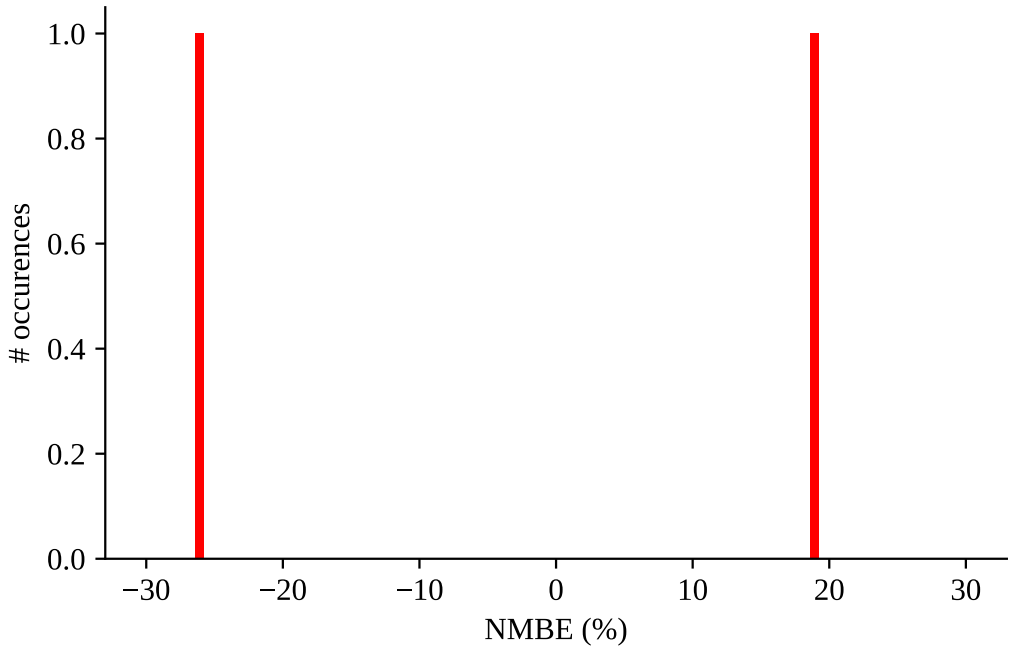

Figure 6.3: The NMBEs calculated twice for another substation than in Figure 6.2, but which shows a similar change in bias when training and test period are switched. To be compared with Figure 6.1

6.2.2 Peak prediction

The mean CVRMSE was calculated for different models as seen in Table 6.4. Furthermore, for two of the chosen utilities (accounting for 18 of the 36 substations in the dataset), one could see a distinct difference between the lowest observed temperatures, and therefore also the highest observed heat load for the different periods 2020-06-01 to 2021-05-31 and 2021-06-01 to 2022-05-31. For these 18 substations, two separate mean CVRMSEs were calculated, adding to the mean CVRMSE calculated for all substations and test periods, being one mean CVRMSE where the test period was the period with a higher maximum heat load value (case ii) and one mean CVRMSE where the training period was the period with a higher maximum heat load value (case iii). These CVRMSEs can also be seen in Table 6.4. The peaks in the test period for case ii were, on average, 10.6 % higher than the peaks in the test period for case iii.

Table 6.4: Mean CVRMSE for different models predicting peaks, both in the case of all substations (case i) and test-training-setups but also specifically for those cases where the test set had a considerably higher (case ii) / lower (case iii) maximum heat load value than in the training set.

	Energy Predict	Heat Load Signature	XGBoost	SVR
Case i	28.38 %	36.26 %	30.30 %	36.49 %
Case ii	32.01 %	36.28 %	36.82 %	40.05 %
Case iii	31.91 %	35.96 %	32.29 %	35.91 %

The models' performance on substations with different heat load patterns against outdoor temperature can be seen in Table 6.5.

Table 6.5: Mean CVRMSE for different models predicting peaks, for substations with heat load pattern corresponding to the one of a heat load signature (A) and for those not corresponding (B)

Case	Energy Predict	Heat Load Signature	XGBoost	SVR
A	27.72 %	31.87 %	30.38 %	32.93 %
B	30.36 %	49.44 %	30.03 %	47.14 %

According to this evaluation framework, Energy Predict stands out as the highest-performing Ml algorithm for peak prediction. It is also the most robust algorithm when the test period has a higher maximum heat load than the training period, only decreasing performance with 0.10 % units. This is lower than 0.32 % units, 4.53 % units, and 4.14 % units for the algorithms heat load signature, XGBoost, and SVR, respectively.

Further, Energy Predict and XGBoost show robustness when evaluated on substations with a heat load pattern not corresponding to the one of a heat load signature. Compared to SVR and heat load signature, their performance did not decrease as much.

If the performance of Energy Predict is still not satisfactory for a DH utility, alternatives include either changing algorithms, adding relevant features or designating a model that is specifically implemented for predicting peaks.

Chapter 7

Discussion and future work

In this section, a discussion will be provided. Insights beyond the concluded answers to the Research Questions will be discussed, both regarding the present operation of DH utilities, the previous studies on the subject, and the evaluation framework. Further, subjects for future work to investigate will be suggested.

7.1 Regarding present District Heating operation

From the survey study, it can be concluded that even though DH utilities generally use heat load predictions in their operation, the majority of these utilities do not measure the accuracy of these predictions. As there are costs associated with inaccuracy, DH utilities should measure it, assessing the risk of their heat load prediction model over- and underpredicting. Higher costs for the DH utility trickle down to the end-user, such as private households, and decrease the competitiveness of DH as a heating method.

In the survey study, the respondents were asked if an increased accuracy of heat load predictions would enable additional use cases. The most common answer was that a higher accuracy would increase the potential use cases. Thus, higher accuracy would not only limit the costs of over- and underpredictions but would also enable DH utilities to develop the operation in other ways, further increasing competitiveness. When assessing if a heat load prediction model's accuracy is sufficiently accurate for new use cases, evaluation frameworks such as the ones proposed in this report can be of value.

A common answer among those respondents who stated that the accuracy of the model was sufficient, was that the potential inaccuracies were compensated for by an accumulator tank. In the case of an overprediction, the excess heat is stored in the tank, and in the case of an underprediction, heat is extracted from the tank. However, an accumulator tank leads to increased capital costs for the DH utility and although it is used to store heat, heat losses are still present. Thus, a heat load prediction model accurate enough for the DH utility to not need an accumulator tank would most certainly result in lower costs. Followingly, an accumulator tank may lower the need for accuracy of short-term heat load forecasts, but for other purposes (e.g. dimensioning and sales planning), the need

for accuracy remain the same. The majority of respondents saying that the accuracy is sufficient may thus rather be reflecting the lack of competence in the subject and not that the accuracy actually is sufficient.

An additional sign of lack of competence could be seen in the sixth questionnaire question *"Had an increased accuracy enabled additional use cases than those you have today? In that case, which and why?"*. There was a considerable number of respondents who answered that they did not know if an increased accuracy would enable more use cases for them. DH utilities not knowing if they need improved models constitutes a potential hurdle for heat load prediction model developers.

There was a considerable number of respondents saying that they did not use heat load predictions in their operation. Among these respondents, there was a general conception that their DHN was too small for heat load predictions to be useful. How the use of heat load predictions differs among DH utilities of different sizes, is a question that was not captured in the scope of this thesis. But there is assumingly some need for dimensioning and planning for all DH utilities, regardless of size. Furthermore, some DH utilities argued that they utilized the experience of certain employees to fulfill the purposes that a load prediction model otherwise could, such as planning production or dimensioning equipment. How model-based heat load predictions compare to these judgments, both in terms of accuracy and other measures, would be interesting questions for future studies to investigate.

To summarize, there is both a need for improved awareness of heat load predictions as a whole, and improved awareness regarding the accuracy of these predictions. Initiatives to increase the overall competence on the matter among DH utilities are therefore encouraged.

A few matters should be taken into consideration when analyzing the results of the survey study. It showed that, out of 28 DH utility respondents, 22 of those were using heat load predictions, of which 16 have not evaluated the accuracy of those predictions. In the survey, there was no way to tell if those who answered that they had evaluated the accuracy, had done so for all of their ways to predict heat load. It could be, for example, that these respondents had only evaluated their model for short-term heat load forecasts and not for dimensioning or sales projections. By asking the respondents *which* heat load predictive model they had evaluated, further insight would be provided on the matter.

Further, the ratio of DH utilities that do not use heat load predictions in their operation, could potentially be higher than what was seen in the survey study. Among the 75 out of 103 utilities that did not answer the survey, it is a possibility that DH utilities not using load predictions are over-represented as those may not have felt targeted by the scope of the survey. However, there are many possible reasons to why a utility would not answer the survey. Nonetheless, the non-response error must be taken into consideration when analyzing the results of the study.

As the two purposes of dimensioning and planning, i.e. predicting peaks and a yearly

accumulated load, have not been covered in literature, there is no evaluation framework developed for these predictions. Thus, a choice was made to develop and propose such an evaluation framework. However, it was not investigated how common it is among DH utilities to have processes for evaluating models and how these potential processes in which case are implemented. The respondents in the survey study were asked the two questions *"Have you evaluated the accuracy of the heat load predictions? In that case, how?"* but the purpose of the second question was solely to validate the first. The answers to the second question were often in the likes of "We store the heat load predictions and compare those to actual outcome" which gives no further insight into how the evaluation framework is developed in terms of evaluation periods and evaluation metrics, only that an evaluation is done. To investigate these evaluations further, it could potentially have been beneficial to do an interview study, instead of a survey, since it would enable the researcher and respondent to cooperate in coming to a conclusion. An interview study on the subject of evaluation frameworks is therefore suggested for future studies.

7.2 Regarding previous work on the subject

It was concluded that research on the subject of ML applied to DH data has been limited to short-term forecasts and fault detection. The other use cases for heat load predictions that were found in the survey study, dimensioning and planning, have not been investigated as much, even though they are used in DH utilities today. As stated in the proposal section, these use cases differ from the use cases of short-term forecasts and fault detection as they are implemented by predicting an imaginary heat load, for a year or under specific circumstances, and not an actual outcome. Thus, there is no elementary way of assessing the accuracy of these predictions and the proposed evaluation framework in this report should not be seen as the reproducible result from the studies but rather a proposal, based on the studies. The difficulties associated with evaluating a model that predicts an imaginary heat load may be a contributing factor to why the research was found to be limited on the subject.

In the study by Dahl et al., it was stated that a change to 100 % accurate weather forecasts led to a decrease of the total absolute error measure RMSE with 14 %, given that their forecasts were made in 15 h to 38 h in advance [11]. However, how the error decrease of 14 % changes with different weather forecast providers or as the time horizon is extended to several days to a week, is left for future work. As weather forecasts are important features for heat load forecasts, their own accuracy is important to assess. The importance of accurate weather forecasts was further addressed in the survey study as a third of those who answered the fourth questionnaire question answered that their heat load prediction model is sufficiently accurate as long as the input weather parameters are correct.

7.3 Regarding the evaluation framework and its results

The insights gathered regarding evaluation metrics and - periods were used when proposing the two evaluation frameworks. The frameworks utilized the error measures CVRMSE and NMBE, as well as different sectionings of the validation data, assessing the common flaws of ML-based heat load prediction models.

As could be seen in the results from the evaluation framework, and as expected, different predicting models show different biases for different seasons. Further, the normalized error, measured as CVRMSE, was lower when predicting datasets where the heat load pattern corresponded to the heat load signature. This was the case, both for the framework evaluating the ability to predict a yearly load and for the one evaluating the ability to predict peaks. For the framework evaluating peaks, it could also be seen that it was generally unfavorable for models to predict a test set with a higher maximum observed heat load than in the training set.

Based on the dataset used in this report, Energy Predict showed the best performance for both frameworks. However, to make a rigorous conclusion regarding which model performs the best on a general basis, more datasets may be needed as well as a test to see if the difference between models is statistically significant or not. This step was excluded from the report as it was not the aim of the study to draw conclusions in that regard.

One could ask how valuable a model is if it has a CVRMSE of approximately 30 %. A calculation of the MAPE could be made for the different models on the different evaluation periods and compare the error value with what was stated by Wojdyga, mentioned in the theory section (MAPE of 3 % to 5 % for heat load forecasts a couple of hours away is sufficient to steer operations in a way that increases efficiency in terms of energy demand by ”as much as a few percent”)[9]. However, one should have in mind that the mean CVRMSEs that were calculated for the different models were done so for individual substations, contrary to a whole DHN. The stochastic behavior of a single substation for one or a few households makes the heat load harder to predict, while the heat load on an aggregate level in a DHN is more true to a model. However, evaluating models to predict an aggregate DHN load would eliminate the possibility to assess the models' robustness in predicting substations with different behaviors.

In this report, it has been argued that not only a heat load prediction but also a prediction interval could provide value for DH utilities. A way of implementing such an interval would be to, in the evaluation framework, analyze the distribution of errors in order to conclude if they follow any generalizable distribution. However, that would require insight into how to assess the different distributions of errors for different ML algorithms. This was considered beyond the scope of the thesis but is, however, a suggestion for future studies to investigate.

Chapter 8

Conclusion

For Swedish utilities, the different purposes that heat load predictions could be used for are, according to the results of the survey study:

- Production Planning
- Sales Planning
- Dimensioning of equipment
- Dimensioning of network
- Dimensioning of production facilities
- As a step in fault detection

Further, the models implemented for these purposes are in general not validated by the DH utilities as sufficiently accurate. Not necessarily that the heat load prediction models are insufficiently accurate (in that regard no conclusion was drawn) but rather that the DH utilities have not implemented any evaluation framework for the models.

It was concluded that, even though many agree that the potential use cases for load predictions are more, research on the subject is limited to short-term heat load forecasts used for production planning. Since studies on ML-based heat load predictions for dimensioning and sales planning purposes were not found, no developed evaluation framework was found either. Therefore, two evaluation frameworks were proposed in the thesis, aiming to evaluate a model's ability to predict heat load for the purposes of dimensioning and sales planning. A choice was made to rephrase that into evaluating a model's ability to predict a yearly heat load with an hourly granularity, as well as to predict peaks.

The evaluation framework will not be thoroughly described in this concluding section. Briefly, it used the evaluation metrics CVRMSE and NMBE for evaluating the ability to predict the yearly heat load and CVRMSE for evaluating the ability to predict peaks. It assessed the models' robustness for different seasons, if the load pattern corresponded to the one of a heat load signature or not, as well as if the training set had a considerably

higher or lower maximum heat load value than the test set.

The two evaluation frameworks were showcased by evaluating the predictive performance of the different load prediction models heat load signature, Energy Predict, SVR, and XGBoost. Energy Predict showed the best performance of all four models on both frameworks.

The evaluation framework can be used by model developers when developing new models, aligned with how heat load predictions are used in DH utilities. It can also be used by DH utilities as a way to measure the accuracy of current evaluation methods and thus address the need for more advanced data-based ML methods. Thus, this evaluation framework could help in bridging the gap between DH utilities and ML model developers.

Bibliography

[1] IEA, "Energy efficiency 2022", International Energy Agency (IEA), Tech. Rep., 2022.

[2] S. Frederiksen and S. Werner, *District Heating and District Cooling*. Lund: Studentlitteratur, 2013.

[3] F. Gaballo, P. S. Nielsen, B. S. Khan and A. Heller, "The role of district heating in the future european energy system", in *2022 IEEE International Conference on Power and Energy (PECon)*, 2022, pp. 420–425. DOI: 10.1109/PECon54459.2022.9988781.

[4] G. Mbiydzenyuy, S. Nowaczyk, H. Knutsson, D. Vanhoudt, J. Brage and E. Calikus, "Opportunities for machine learning in district heating", *Applied Sciences*, vol. 11, no. 13, 2021, ISSN: 2076-3417. DOI: 10.3390/app11136112. [Online]. Available: https://www.mdpi.com/2076-3417/11/13/6112.

[5] product managers and system developers at Utilifeed, "Gothenburg", Interview 7.2.2023.

[6] A. Müller and S. Guido, *Introduction to Machine Learning with Python*. Sebastopol, CA, USA: O'Reilly Media, 2017.

[7] C. Ntakolia, A. Anagnostis, S. Moustakidis and N. Karcanias, "Machine learning applied on the district heating and cooling sector: A review.", *Energy Systems*, vol. 13, no. 1, pp. 1 –30, 2022. [Online]. Available: https://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true\&AuthType=ip,uid\&db=inh\&AN=22145388\&site=eds-live\&scope=site.

[8] S. Moustakidis, I. Meintanis, G. Halikias and N. Karcanias, "An innovative control framework for district heating systems: Conceptualisation and preliminary results", *Resources*, vol. 8, no. 1, 2019, ISSN: 2079-9276. DOI: 10.3390/resources8010027. [Online]. Available: https://www.mdpi.com/2079-9276/8/1/27.

[9] K. Wojdyga, "Predicting heat demand for a district heating systems", *International Journal of Energy and Power Engineering*, vol. 3, no. 5, pp. 237–244, 2014.

[10] Z. Wei, T. Zhang, B. Yue *et al.*, "Prediction of residential district heating load based on machine learning: A case study", *Energy*, vol. 231, p. 120 950, 2021, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2021.120950. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544221011981.

[11] M. Dahl, A. Brun, O. S. Kirsebom and G. B. Andresen, "Improving short-term heat load forecasts with calendar and holiday data", *Energies*, vol. 11, no. 7, 2018, ISSN: 1996-1073. DOI: 10.3390/en11071678. [Online]. Available: https://www.mdpi.com/1996-1073/11/7/1678.

[12] N. P. Sakkas and R. Abang, "Thermal load prediction of communal district heating systems by applying data-driven machine learning methods", *Energy Reports*, vol. 8, pp. 1883–1895, 2022, ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2021.12.082. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484721015213.

[13] E. Ogliari, P. Eleftheriadis, A. Nespoli, M. Polenghi and S. Leva, "Machine learning methods for clustering and day-ahead thermal load forecasting of an existing district heating", in *2022 2nd International Conference on Energy Transition in the Mediterranean Area (SyNERGY MED)*, 2022, pp. 1–6. DOI: 10.1109/SyNERGYMED55767.2022.9941387.

[14] E. Dotzauer, "Simple model for prediction of loads in district-heating systems", *Applied Energy*, vol. 73, no. 3, pp. 277–284, 2002, ISSN: 0306-2619. DOI: https://doi.org/10.1016/S0306-2619(02)00078-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261902000788.

[15] T. Fang and R. Lahdelma, "Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system", *Applied Energy*, vol. 179, pp. 544–552, 2016, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2016.06.133. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261916309217.

[16] T. Kurek, A. Bielecki, K. Świrski *et al.*, "Heat demand forecasting algorithm for a warsaw district heating network", *Energy*, vol. 217, p. 119 347, 2021, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2020.119347. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544220324543.

[17] M. H. Kristensen, R. E. Hedegaard and S. Petersen, "Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling", *Energy*, vol. 201, p. 117 687, 2020, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2020.117687. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544220307945.

[18] S. Idowu, S. Saguna, C. Åhlund and O. Schelén, "Forecasting heat load for smart district heating systems: A machine learning approach", in *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2014, pp. 554–559. DOI: 10.1109/SmartGridComm.2014.7007705.

[19] D. Geysen, O. De Somer, C. Johansson, J. Brage and D. Vanhoudt, "Operational thermal load forecasting in district heating networks using machine learning and expert advice", *Energy and Buildings*, vol. 162, pp. 144–153, 2018, ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.12.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778817312070.

[20] R. Nateghi and S. Mukherjee, "A multi-paradigm framework to assess the impacts of climate change on end-use energy demand", *PLoS ONE*, vol. 12(11), e0188033, 2017. DOI: https://doi.org/10.1371/journal.pone.0188033. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188033.

[21] H. Hewamalage, K. Ackermann and C. Bergmeir, "Forecast evaluation for data scientists: Common pitfalls and best practices.", *Data Mining and Knowledge Discovery*, pp. 1 –45, 2022, ISSN: 1384-5810. [Online]. Available: https://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true\&AuthType=ip,uid\&db=edssjs\&AN=edssjs.AEE6D3FA\&site=eds-live\&scope=site.

[22] G. R. Ruiz and C. F. Bandera, "Validation of calibrated energy models: Common errors.", *Energies (19961073)*, vol. 10, no. 10, p. 1587, 2017, ISSN: 19961073. [Online]. Available: https://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true\&AuthType=ip,uid\&db=a9h\&AN=125994985\&site=eds-live\&scope=site.

[23] J. Granderson, S. Touzani, C. Custodio, M. Sohn, S. Fernandes and D. Jump, "Assessment of automated measurement and verification (m&v) methods:.", 2015. [Online]. Available: https://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true\&AuthType=ip,uid\&db=edssch\&AN=edssch.oai\%3aescholarship.org\%2fark\%3a\%2f13030\%2fqt636424jc\&site=eds-live\&scope=site.

[24] D. Chakraborty and H. Elzarka, "Performance testing of energy models: Are we using the right statistical metrics?.", *Journal of Building Performance Simulation*, vol. 11, no. 4, pp. 433 –448, 2018. [Online]. Available: https://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true\&AuthType=ip,uid\&db=inh\&AN=18020390\&site=eds-live\&scope=site.

[25] R. A. Peterson, *Constructing Effective Questionnaires.* Thousand Oaks, CA, USA: SAGE Publications, Inc, 2000.

Chapter 9

Questionnaire

9.1 Interpretation in English

At first, the participants were provided with an introductory section. Since there was a risk of misconception with the concept of load predictions, being more than solely forecasts, the word was defined in this section before answering the questionnaire. It was stated that: (translated from Swedish)

"Heat load predictions can both be forward- and backward-looking. Generally, a method with one or more features (for example outdoor temperature or time of day) is used in order to calculate a heat load value. Examples of heat load prediction methods include energy signature or data-based machine learning. The calculated heat load value can be aggregated on a basis of hours, days, etc, and both at a substation- or network level."

Progressively, the participants were provided a text section where it was stated that if the reader felt that there was another person at their utility that had more insight into the questions asked, they should not hesitate to forward the questionnaire to that person. This section was provided to ensure that the participants were in the right position to answer the questions asked in the questionnaire.

The defining section, as well as the section ensuring that participants had the appropriate position, was made in line with what Peterson states as "a researcher must try to ensure that all study participants are educated equally" [25].

Regarding the questions, the first one related to Research Question Number 1 and a direct question, *"For what purposes are you using heat load predictions today?"* with the purpose of answering Research Question Number 1. The question was supported with the predetermined answers:

- Production Planning
- Sales Planning
- Dimensioning of equipment

- Dimensioning of network
- Dimensioning of production facilities
- As a step in fault detection

As well as a blank option for the participants to fill in if desired, in case there were more purposes that were not captured by the predetermined answers. The reason for predetermining potential answers was again in order to prevent the misconception of heat load predictions only referring to heat load forecasts. By providing examples of what heat load predictions can be used for, aside from production planning, participants were given a chance to realize that they may use heat load predictions in more ways than they initially thought.

To further prevent the misconception of heat load predictions only being heat load forecasts, some participants, the ones that either stated that they only used heat load predictions for production planning or not at all, received a follow-up e-mail. In the e-mail, the participants were asked what method they use when dimensioning their network and/or equipment. The answer to this question enabled the surveyor to validate that the participants were right in stating that heat load predictions (as it has been defined in this report) were not used in their operation.

The second question was *"Can you describe how these heat load predictions are calculated? (which system/algorithm/method)"*. The question was asked in case there was a commonly used model that could be used when showcasing the proposed evaluation framework further down in this report.

The third question was *"With what time horizon are these predictions made (a couple of hours, days, months, years)"* to further validate that the different heat load predictions were done in a way that corresponds to how they are mentioned in this report. The question may not be applicable for dimensioning purposes, which is why the participants were not obligated to answer the question (nor any of the others, but increasingly important for this question).

The fourth question was *"Do you feel like the accuracy (how predictions compare to the outcome) is sufficient for the use case of the heat load prediction? Feel free to motivate"*. The purpose was to obtain an indication of how the need for better heat load prediction models is experienced by the actual users. The question could potentially provide a discussion topic.

The fifth question was *"Have you evaluated the accuracy of the heat load predictions? In that case, how?"*. It was the intention to set this question after the previous one as a way to know their unbiased opinion about the accuracy before actually stating if the accuracy had been evaluated. Nevertheless, the question validates the answer to the fourth question, to say that the heat load prediction models are sufficiently accurate without being evaluated. Asking the participants how they evaluate the accuracy was done in order to ensure that they actually evaluate the heat load prediction model (contrary to,

for example, the customer satisfaction)

The sixth question was *"Had an increased accuracy enabled additional use cases than those you have today? In that case, which and why?*, further validating the answer to Research Question Number 2. If increased accuracy had led to more use cases, the accuracy is, in a way, insufficient.

As a last question, the participants were asked if they had something to add to their previous answers. The question was asked in case the participants felt like they had input, relevant to the project, that they had not had the opportunity to share earlier in the questionnaire.

9.2 Survey in Swedish

Hej, du får denna enkät i samband med ett projekt mellan Utilifeed och Lunds Tekniska Högskola (LTH) där vi undersöker hur fjärrvärmebolag använder sig av lastprediktioner och hur pricksäkerheten i dessa prediktioner utvärderas.

Lastprediktioner kan vara både framåt- och bakåtblickande. Generellt används en metod med en eller flera parametrar (exempelvis utomhustemperatur eller tid på dagen) för att på så sätt få fram ett värde på lasten. Exempel på lastprediktionsmetod är energisignatur eller databaserad maskininlärning. Lastvärdet kan i sin tur avse olika tidshorisonter (timme, dag, etc.) på både undercentral- eller nätnivå.

Om du känner att det finns en annan person på ditt företag som har större insikt i frågorna som ställs nedan så är vi väldigt tacksamma om du vill vidarebefordra enkäten till denne. Vi blir väldigt tacksamma om ni väljer att medverka!

// Sara Månsson (Utilifeed) och Herman Hansson (LTH) telefon: +46707975405 mail: he3857ha-s@student.lu.se

Namn på person(er) som fyller i enkäten (required):

Short text answer

Email-adress till person(er) som fyller i enkäten (required):

Short text answer

Bolag (required):

Short text answer

Samtycke till att lagra personuppgifter (required):

Checkbox:: Härmed samtycker jag till att Lunds universitet får spara och lagra mina uppgifter. Detta för att de ska kunna kontakta mig angående förtydligande av de svar

som ges i denna enkät.

Får vi kontakta dig utifall vi skulle vilja ställa ytterligare frågor?

Multiple choices

- Ja
- Nej

Vad använder ni lastprediktioner till idag?

Multiple choices

- Produktionsplanering
- Försäljningsplanering
- Dimensionering av utrustning
- Dimensionering av nät
- Dimensionering av produktionsanläggningar
- Som ett led i feldetektering
- *Other, short text answer*

Kan du beskriva hur dessa lastprediktioner tas fram? (vilket system/algoritm/metod?)

Long text answer

Hur långt in i framtiden görs dessa prediktioner? (Några timmar, dagar, månader, år framåt?)

Long text answer

Upplever ni att pricksäkerheten (hur prediktionerna förhåller sig till utfall) är tillräcklig för det användningsområde som lastprediktionerna används till? Motivera gärna

Long text answer

Har ni utvärderat pricksäkerheten i prediktionerna och isåfall hur?

Long text answer

Hade en ökad pricksäkerhet möjliggjort användningsområden utöver de ni har idag? Isåfall, vilka och varför?

Long text answer

Har du något att tillägga?

Long text answer