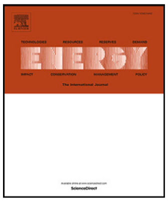
Energy 313 (2024) 134016

Contents lists available at ScienceDirect

Energy

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



Creating a labeled district heating data set: From anomaly detection towards fault detection

Dominik Stecher [a], Martin Neumayer [b], Adithya Ramachandran [c], Anastasia Hort [a], Andreas Maier [c], Dominikus Bücker [d,b], Jochen Schmidt [a,*]

[a] *Rosenheim Technical University of Applied Sciences, Department of Computer Science, Hochschulstr. 1, 83024 Rosenheim, Germany*
[b] *Institut für nachhaltige Energieversorgung GmbH (INEV), Eduard-Rüber-Str. 7, 83022 Rosenheim, Germany*
[c] *Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany*
[d] *Rosenheim Technical University of Applied Sciences, Department of Engineering Sciences, Hochschulstr. 1, 83024 Rosenheim, Germany*

ARTICLE INFO

Keywords:
Data set creation
Data augmentation
District heating
Fault detection
Machine learning

ABSTRACT

For an efficient operation of district heating systems, being able to detect anomalies and faults at an early stage is highly desirable. Here, data-driven machine learning methods can be a cornerstone, particularly for fault detection in district heating substations, where the availability of heat meter data keeps increasing. However, the creation of data sets suitable for training such machine learning models poses challenges to researchers and practitioners alike. To address this problem, we propose a systematic and domain-specific process for data set creation for fault detection in the form of practical guidelines. This process concretizes the data science and data mining cross-industry standard CRISP-DM for the district heating domain and focuses on the process steps of goal definition, data acquisition and understanding, and data curation. We aim to enable researchers and practitioners to create data sets for fault detection in the district heating domain and therefore also enable the creation or improvement of machine learning models in this domain. In addition, we propose a minimum viable feature set for fault detection in district heating networks with the goal of enabling better cooperation between researchers and easier transfer of the resulting machine learning models, to better proliferate new progress in the field.

1. Motivation

District heating systems (DHS) can provide sustainable heating if operated efficiently and powered by renewable energy sources. Therefore, DHS are a crucial part of future heating systems and decarbonizing the heating sector [1–3]. Yet previous studies [4] found that many substations operate sub-optimal due to faults, resulting in high return temperatures [5]. Generally, higher supply temperature levels lead to increased heat losses in the distribution system and reduce the range of potential usable heat sources [4–9]. Using renewable heat sources and low-temperature grids are also key concepts of fourth generation of district heating systems (4GDH) [10]. Detection and elimination of faults are therefore essential to achieve lower return temperatures, ensure efficient operation, and pave the way towards 4GDH.

With progressing digitization and legal requirements to provide customers with information on their consumption, more and more substation data are becoming available for fault detection [9]. With the availability of sufficient data and advances in fault detection and predictive maintenance in other domains, interest in fault detection in DHS rises [11]. However, typically fault detection is done via supervised learning, i.e., an algorithm is trained with previously annotated examples, so-called labeled data. In the context of DHS, labeled data should accurately describe what type of fault is detected, and at what point in time [12] alongside heat meter data and contextual information, e. g., environmental data [11]. To create such a labeled data set, utility companies might have to collect the various data from different systems, assess the data quality, merge, and annotate the data, i. e., curate the data. As data curation is a mostly manual process and requires expertise in the data and the district heating domain, creating a labeled data set poses a major challenge. On the one hand, cross-industry standards like CRISP-DM [13] provide orientation and a high-level understanding of how to approach data-driven projects, previous research has broadened the understanding of faults and the organizational context surrounding them [7,12,14]. On the other hand, articles like [15] provide guidance on typical data curation steps, yet domain-specific processes and guidelines that combine both aspects for data set creation are missing. We argue that the lack of processes

[*] Corresponding author.
E-mail address: jochen.schmidt@th-rosenheim.de (J. Schmidt).

https://doi.org/10.1016/j.energy.2024.134016
Received 20 December 2023; Received in revised form 4 November 2024; Accepted 26 November 2024
Available online 3 December 2024

and guidelines also results in a lack of publicly available labeled data sets for fault detection in DHS [9]. Instead, authors often rely on unsupervised learning methods on unlabeled data [9,11]. In this paper, we show both network- and substation-based approaches for fault detection, and address the gap in the existing literature regarding the creation of the necessary data sets. To do so, we propose a systematic process for data set creation, including an overview of typical data sources and data curation techniques, addressing the following research questions:

1. How can a systematic process for the creation of a data set for fault detection look like? (Section 2)
2. What data sources can be used to create a data set? (Section 3)
3. What data curation techniques can be used? (Section 4)
4. How does the process work in a real-world example? (Section 5)
5. What are current issues for DH data set creation and data democratization? (Section 6)
6. What is our concise recommendation for a DH fault detection data set? (Section 7)

By answering these questions, we aim to empower researchers and practitioners to develop more robust and accurate machine learning (ML) models.

2. Data set creation process

Our proposed data set creation process is based on CRISP-DM, a well-established process for data set creation shown in Fig. 1 and has been simplified for the sake of clarity: Jumping back to a previous process step is always possible if necessary, but arrows pointing back are mostly omitted in Fig. 2. A detailed and adapted version of the process we used to create our data set as described in the case study in Section 5 is shown in Fig. 2. It also shows the progression of the data set as well as decision criteria for where to go back in the process if an issue is encountered.

Further, both Figs. 1 and 2 highlight the scope and the main contribution of this paper: We focus on the process of data set creation, including the goal definition, data acquisition, and understanding, as well as data curation. The subsequent prototyping and evaluation to check the feasibility of the goals are considered briefly at the end of this section as well as in the case study. The implementation of a more complex model, its tuning, and deployment, are not the focus of this article. Neither is the publication of a complete data set.

2.1. Goal definition phase

The data set creation process starts with defining one or more goals. During the goal definition phase, it is important to understand how goals influence the data required and applicable algorithms. For example, if a time threshold is set, e. g., the ML model must detect a fault after 1 h, the currently typical sampling rate of 1 h for substation parameters will not be enough as it gives the model only a single set of measurements to detect the fault. If a faster sampling rate is not possible or feasible, the goal must be redefined.

Another often overlooked part is that data sets can be reused to solve other, future problems, or to improve on existing solutions as technological progress increases the quality of ML models. By looking at current research efforts and considering other potential use cases, two primary changes can be made. One, to include additional data sources, which, while not relevant to the current goal, may prove useful in the future. Two, to improve the quality of the recorded data, for example by increasing the sampling rate from hourly to 5-minute or even 1-minute intervals or by adding additional categories for the fault root cause. It is important to consider that district heating works on a yearly cycle, so data sets often require a full year of data. In addition, some faults occur very rarely. Recording these data now can significantly reduce future projects by cutting down or completely eliminating the time to record a new, goal-specific data set. The cost incurred for this preemptive recording of data is often paid in terms of storage. However, recording a single measurement such as inlet temperature at 64-bit precision for 1000 households over a 3-year period, requires about 200 MB of uncompressed storage, when done hourly, compared to about 12 GB when done at 1-minute resolution. While the increase by two orders of magnitude is significant, the overall volume of data is minuscule by today's standards.

2.2. Data acquisition and understanding

After defining the goals, the necessary data have to be acquired and understood. Data acquisition refers to searching for, gaining access to and collecting relevant data (see Section 3 for data sources that are relevant for fault detection in district heating substations).

An important part of data understanding is exploratory analysis, in which the data are visualized, and analyzed manually and statistically with appropriate tools. In this way, correlations between attributes and relationships between different types of data can be found. Exemplary use cases can be found in Sections 4.1.2 and 5.1

Data understanding should also include understanding problems in data quality, e. g., missing or imprecise data, as they can prevent the objectives from being achieved or must be remedied using appropriate methods. One example of a typical data quality issue we encountered was a high occurrence of single-hour missing values in supply and return temperature readings of individual substations raising questions and leading to the discovery of a flawed calculation if measurements arrived too late. Subsequent changes to the process removed the majority of gaps in the time series.

Another important source of information for understanding the data are the technicians and engineers currently analyzing faults and anomalies, and the data itself. Technicians and engineers are often already performing the fault detection or classification process and thus can help to determine relevant features for the data set as they already know those relevant to their work. Their knowledge is also essential for any kind of review and analysis of already collected data.

Insights during the data acquisition and understanding phase might require changes in the goal definition. For example, determining the time difference between specific faults occurring and the customer calling the service hotline can confirm or change some requirements made during the goal definition phase, e. g., by relaxing or shortening the reaction time the future ML model must achieve to be of practical use.

2.3. Data curation

Data curation brings together different data sources, deals with problems in data quality, and prepares data for use in ML models (see Section 4 for typically used data curation methods). As a first step data from different sources are collected in one place, joined, and relationships are made explicit, e. g., by connecting the heat meter time series with the timestamp of a customer call for a given fault incident. Next, problems in data quality can be remedied, e. g., by removing wrong or implausible values, imputing missing values, or fixing structural discontinuities in GIS data. Lastly, the resulting data can be prepared to be used in ML models; e. g., neural networks are typically trained on vectors of floating-point numbers. Therefore this is a good opportunity to consider suitable numerical representation of other data types such as one-hot encoding or using functions.

It should be noted, however, that any manipulation should ideally be reviewed by someone familiar with the original data to verify the quality of the gap-filling method, for example. It is also advisable to run some tests early in the collection process, e. g., to determine whether the gaps in the recorded time series data can be filled or if improvements in the collection process must be made. For this, a short collection test run can be done and different issues such as varying gap lengths can be simulated.

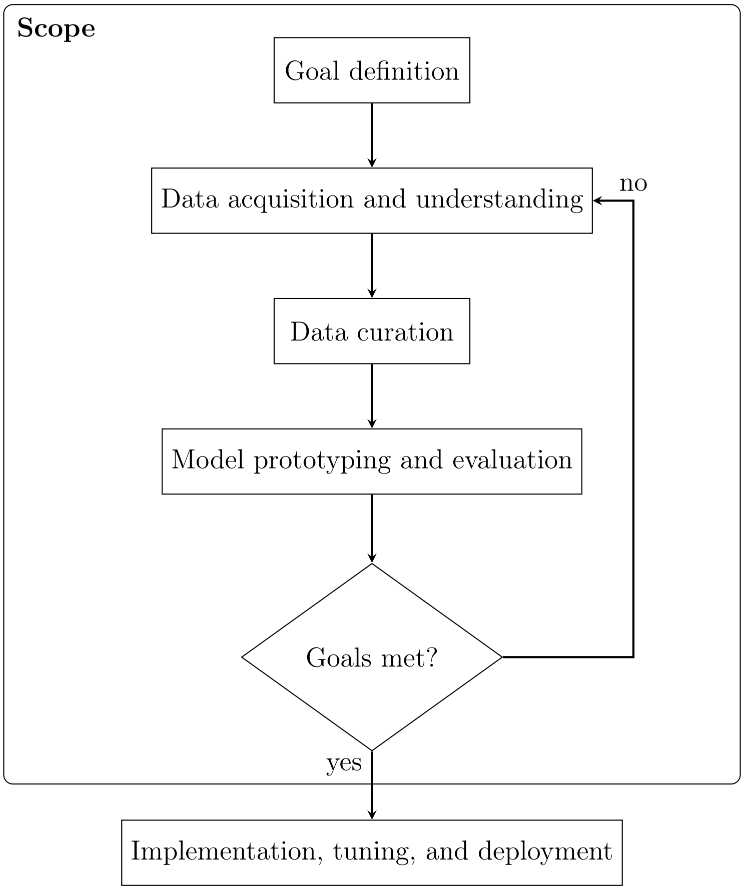

Scope
Goal definition
Data acquisition and understanding
no
Data curation
Model prototyping and evaluation
Goals met?
yes
Implementation, tuning, and deployment

Fig. 1. Simplified overview.
Source: Adapted from [16].

2.4. Model prototyping and evaluation

As the data curation process is progressing or at the latest once the data set has been completed, it is recommended to train and evaluate ML prototypes on the already collected data to detect faulty specifications and make changes early in the process. These ML prototypes should give an impression of whether the goals defined can be met with a more complex model or if one has to take a step back in the process.

3. Data sources, types, and aggregation

In a DHS, various data are available from different sources, which are described in the following sections. An overview of typical sources and the data types and content provided is shown in Table 1.

3.1. District heating data

District heating data encompasses information about DH infrastructure such as pipe networks and substations, related measurements from these infrastructure components, as well as customer or consumer information. This data is typically already available at DH companies but may need to be collected from different data bases, e. g., technical and billing departments.

3.1.1. Substation data

Substation data can be divided into static information and heat meter readings. Static information might include the type of substation used, its connection conditions, e. g., the substation's connection value, its position, and the associated customer.

Heat meter readings are frequently used in the context of fault and anomaly detection. The available heat meter data depend on the installed station, but some meter readings are required for billing and are therefore typically available: Flow meters measure the flow rate of the heat transfer medium. Temperature sensors are employed to monitor the supply and return temperature on the customer side, i. e., the secondary side; in some cases, they are also employed on the grid side, i. e., the primary side. Based on the flow rate, and supply and return temperature, further metrics such as temperature differences or heat delivered can be derived. Depending on the substation, pressure sensors might be in place to monitor pressure differences within the substation. Further, some substations allow the measurement of valve positions. Heat meter readings are collected cyclically, e. g. in an hourly interval. In a previous work [9], we found that most authors reported meter readings in an hourly interval, however, some authors used data with a higher temporal resolution of 3, 5, or 10 min. Data beyond what is needed for billing such as valve positions and data from the secondary side of the heat exchanger are not typically available. Heat meter readings may be considered personal data and fall under local data protection laws.

3.1.2. Network topology

An intricately interconnected network of insulated pipes, in combination with suppliers and consumers, constitutes the foundation of a DH supply chain. At the core of the DH network are virtual entities known as nodes, serving as fundamental building blocks. In addition to holding vital meta-information, they also play a crucial role in shaping the network's physical infrastructure, including the configuration of pipes. Each pipe within the network is associated with a specific starting and ending node, accompanied by a sequential array

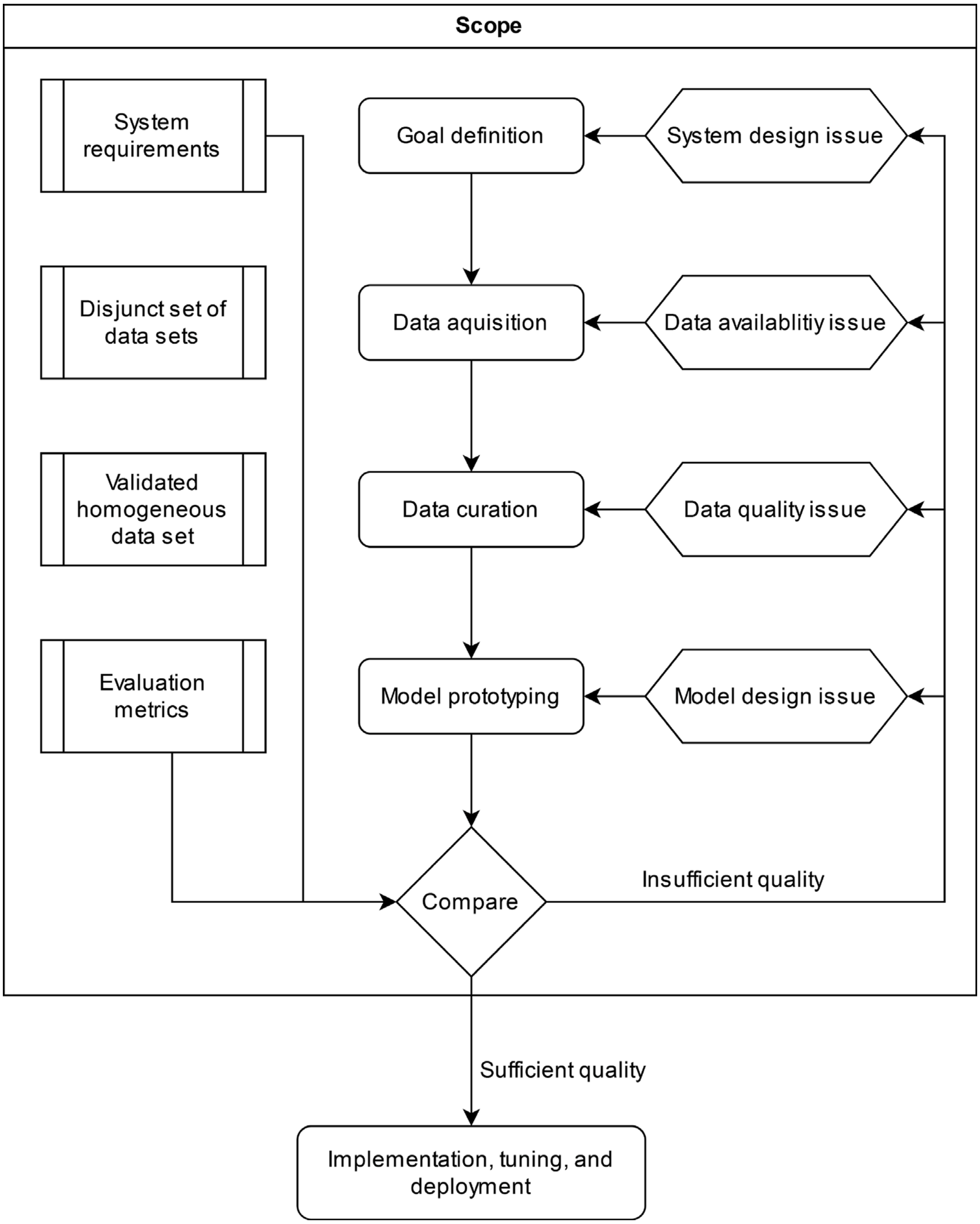


Fig. 2. Detailed overview of the adapted version of the CRISP-DM process with data set descriptions and shortcuts depending on problem type.

of nodes that render the pipe's geographical path, encapsulating its geometric attributes. The spatial data of DHS adheres to the standards of Geographic Information System (GIS) and are represented either as geometric shape files or as encoded geometric features in relational databases.

In our previous literature survey [9], we found that location information is rarely used for fault detection. However, topological data can provide valuable insight about observed faults such as heat or pressure loss due to leakage, or the geographic influence on expected heat meter data of a substation.

3.1.3. Consumer information

Consumer information can give essential context for ML algorithms when it comes to evaluating the condition of individual substations. We distinguish between data related to the heated building and data related to the customer itself. Data related to the customer may be available to DH utility companies for the purpose of installing properly specified substations. The type of customer, e. g., industrial, residential, office, or retail, provides context for the amount of heat consumed as well as information on when to expect loads. We would not expect office or retail consumption on public holidays, for example. Information related to the building may be obtained from government agencies, e. g. geodata agencies, such as the Danish Cadastre from the Danish Geodata Agency.

However, there are potential alternative data sources: For example, the heated floor space area, which can be estimated using the building footprint and the number of floors, also provides context for the expected heat consumption. The building footprint can be obtained from city models, building cadastres, or derived from satellite or aerial photographs, accounting for a smaller footprint due to protruding roofs. If such information is not available, number of rooms, bedrooms, or bathrooms can also provide additional context regarding building size. Together with its outside surface area, a building's insulation defines the heat loss to the environment. A lower bound for the insulation

Table 1
Overview of typical data sources, the data types they contain, and exemplary contents.

Data source	Typical types of data	Typical content
Substation data	Time series, Technical specifications	• type of substation • connection conditions • supply and return temperature • mass flow
Grid information	Technical specifications, GIS data	• network topology • type of pipes • insulation
Consumer information	Socio-economic data	• heated area • number of residents • year of construction
Meteorological data	Time series	• outside temperature • humidity • solar irradiation • wind speed
Calendrical data	Time series, categorical	• weekday • calendar week • public holidays • special events
Fault logs	Categorical	• timestamps • customer report • repair technician report • bill of materials • root cause

quality can be derived from the building's age. In certain cases, this can be determined by looking at historically newly zoned residential areas, which can define the age and thus minimum building standards for large clusters of buildings. Similarly, building height can be highly uniform over large areas of a community.

3.2. Meteorological data

Meteorological data is an essential part of ML and statistical applications for heat-related tasks – be that heat forecasting, supervised or unsupervised fault detection. As such, it is important to include a copy of the local weather records in the full data set. The importance of outdoor temperature as well as the uncertainty of weather forecasts and solar irradiation has been analyzed in [17] with the key finding that accurate outdoor temperature prediction is the most influential parameter. While using actual outdoor temperature records instead of a weather forecast when working on historical data, the benefit is minor compared to including any form of outdoor temperature in the first place. Solar irradiation produces only minor benefits, although its importance may vary depending on latitude and season. Further research is needed for a definitive answer to this question. Similarly, [11] showed that outdoor temperature is an important factor for fault detection. Additionally, wind speed combined with a social component capturing the weekly cycle of human activity yields a good heat load predictor according to [18], which can in turn be used as a fault detection component as shown by [19].

Historical meteorological measurements are available at national weather services. It can also be advantageous to record weather forecasts for more realistic training scenarios and thus more robust ML models.

3.3. Calendrical data

Human activity changes in daily and weekly cycles as does heat consumption. For example, [20] showed different weekly usage patterns depending on building use. These patterns even allow conclusions on the building use such as the distinct drop in heat demand on Wednesdays and Sundays in public administration offices in Sweden with a peak on Saturdays, showing the deviation from the normal Monday to Friday work pattern. In addition, [18] showed that removing information about daily and weekly behavior patterns from the training data leads to significantly worse outcomes for heat load prediction. While information about hour, day, and week is typically available in time series data sets, regionally specific information about school holidays, public holidays, and special occasions with changed or untypical behavior. As outside researchers may not be familiar with all of these, such data should be added as early as possible in the data set creation process.

3.4. Fault logs

Data on faults can be obtained from two main sources. First, from maintenance logs, and second, from customer calls. The level of detail found in maintenance logs varies depending on the DH provider. However, with a focus on technical components, faulty batches of certain components can be identified, and spare repair capacity assigned to preemptively replace such components. Analyzing and utilizing maintenance logs must be handled with care as some components may be replaced on a fixed schedule following local regulations or internal guidelines. Therefore, internal communication and documentation of such processes is essential. Customer calls, complaints, or fault reports provide a first reference when a problem actually impacted a customer. This is typically the time of the call and ticket creation and provides a cut-off time for taking action before the customer notices. A second timestamp often found is when the repair technician closed the ticket after repairing the fault. This timestamp may be delayed by several hours depending on mobile connectivity or if the ticket is closed at the end of the shift. In some cases, tickets remain open for extended periods of time, either due to delayed spare parts or because the ticket was never closed. In combination, customer complaint time and ticket closure time – after a manual review – can provide data on reaction and repair time. Extreme cases can help optimize internal processes and spare parts management.

4. Data curation

In the data curation process, a variety of issues within the raw data set are resolved. These range from missing or implausible values caused by interruptions in the transmission or sensor glitches to establishing connections between different data sources. In this section, we explain how a variety of issues can be fixed depending on the type of data, which pitfalls to avoid, and show some exemplary tools to achieve a smooth and efficient workflow.

4.1. Time series curation methods

Time series data is defined by the individual points of data forming a temporal sequence, which is a property we can exploit to detect and fix issues with the data set. For the content of this section, we consider time series data to consist of momentary values, not cumulative values.

4.1.1. Removing wrong or implausible values

Wrong or implausible values can be caused by hard- and software issues ranging from faulty connections, corrosion, wear and tear, to transmission or conversion faults, and data storage corruption. This can result in physically impossible or implausible values such as a temperature of −32768 °C, which is the minimum value for a 16-bit signed integer. For each sensor, a set of physically possible or under the current use case plausible limits for the value range can be defined. For example, the water in a specific pipe should never freeze or boil. We can now define a range of −5 °C to +105 °C, which the sensor values – considering tolerance – cannot exceed. Values outside the accepted range can either be set to be the limit or be treated as a gap as shown in Section 4.1.2. That leaves values that fall inside the limits, but are implausible due to adjacent values. For example, we would not expect the outside temperature to fluctuate by ±10 °C each hour at night time. Such faults can be found either by looking at the first derivative of the time series and imposing limits there or by employing advanced ML methods which can detect such values such as [21]. A practical example of limit-checking substation data can be found in [15].

4.1.2. Imputation of missing values

Missing values are a common issue with time series data. While there are many simple solutions such as zero-padding, repeating the last value, linear or polynomial interpolation, we recommend using one of two more advanced methods.

The first requires one or more highly correlated time series without gaps at the relevant time. This data can then be used to train a simple ML model, e. g. a support vector machine (SVM) or a fully connected neural network, to predict values for the time series containing the gap and then be used to fill in the gap. An example of this approach can be found in Section 5.1.

The second option uses a similar approach but without other correlated time series. This removes the data requirements from the first option but requires a stronger ML model performance to generate realistic values, with requirements growing depending on the gap length. An example for one such ML model is TimesNet [21], which is capable of capturing both short- and long-term patterns in the time series.

4.1.3. Temporal alignment

The temporal alignment of time series faces three primary challenges. First, there is the issue of differing sampling rates, e. g., one time series being sampled at 5-minute intervals whereas another only uses hourly measurements. This can be solved by up- or downsampling the lower or higher sampling frequency data, although downsampling inherently comes with a loss of information.

Next, we face misaligned time series data, e. g., one sensor is read every full hour whereas another is read 20 min after each full hour. Depending on the sensor type, this can either be fixed by shifting one time series by the necessary amount, or by using interpolation to obtain values at the sampling points for the other time series.

Finally, there is the issue of time zones and daylight saving time. To avoid issues we recommend using either universal coordinated time (UTC) or the default time zone offset for the specific location for all time series data, ignoring daylight saving time. Data available with daylight saving time causes a gap and a duplicate value or two gaps at known points in time, depending on the implementation while also shifting between time zones. Time zones result in time series being shifted relative to each other, weakening or breaking correlation between individual features. Instead, time zone and daylight saving time, i. e. information about the time of day as perceived by a customer, should be encoded as part of the calendrical data. This avoids misalignment and discontinuity issues with sensor data while providing such information explicitly to the ML model.

4.2. Categorical curation methods

Categorical or qualitative data such as fault type or pipe diameter lacks the temporal sequence aspect, which limits the methods by which the data set can be augmented before use. The methods discussed here require more human input compared to the time series curation methods.

4.2.1. Imputation of missing values

Missing values in categorical data are filled in two major ways. The first is to either determine the most similar complete sample in the data set according to the remaining features and fill the gap using its value. A more complex approach is to use for example a clustering algorithm to create groups from the complete data set according to all other features, then determine the cluster for samples with the missing feature. Once the cluster has been determined, the other samples from the cluster can be evaluated and an according value be assigned – e. g., assigning the most represented value.

4.3. Fault data curation

To understand the entirety of fault data, it is important to understand the different aspects of an individual fault. In general, there is fault data, e. g., substation ID, fault location or affected parts of the hardware, fault cause, and there is temporal information in the time series data of the substation. There are four specific moments in time that are of interest. First, there is the time of failure (ToF), which is either defined by the onset of a spontaneous fault or by the time by which it is financially advisable to fix a creeping fault such as limescale buildup, i. e. the time by which we expect to be notified of a fault. Second, there is the time of customer complaint (ToCC), determining the detection time frame for a fault detection algorithm. As customer feedback is typically used to detect faults, this gives a human baseline to compare algorithms to. Third is the time of repair (ToR), which is obtained from the repair technician marking a repair job as complete. This also defines the reaction time in general and for individual fault causes. Fourth and last is the time of (re)normalization (ToN), the time by which the behavior of the substation has returned to normal. At this time, the repair can be considered successful and the fault has come to an end.

4.3.1. Verification and connection of existing data

Fault data may not be directly connected to the correct time series data from the affected substation. For example, a renting customer call will be registered to the customer and their address, the corresponding substation, however, can be associated with the landlord, or, in the case of multiple nearby apartment buildings, be registered under a different address or be one of multiple substations at the same address. Due to this complexity, a manual review and selection of the correct substation currently suffering a fault is necessary. During this process, other existing fault-related data entries can be verified, e. g., do the assigned root cause, replaced parts, and time series match?

4.3.2. Adding actual fault occurrence and repair timestamp

DH systems are slow to react to a substation fault until the customer detects a noticeable decrease in indoor temperature. In addition, only two out of the previously defined four timestamps are typically available. First, the ToCC – either by opening a ticket or calling a hotline – and the ToR when the technician confirms that the substation has been repaired and resumes normal operation. The ToF and ToN must then be added later by manually reviewing each fault event and the corresponding time series data, for example by DH experts. A suitable tool to review each existing fault is shown in 4.5.1.

Table 2
DH Network – Key meta-data describing the nodes that build the network.

Feature	Description
Node ID	Unique ID of the node – alpha numeral code
Type	Classification of node
Activity Status	What is the status of the node? – exist/discontinued
Pipe ID	ID of the pipe the node is part of – alpha numeral code
X, Y	Geolocation of the node – UTM/WGS-84 projection
Address	Address the node is present in

4.3.3. Manual review of time series data

While existing fault logs capture faults that impact the customer experience, there are other faults that may not be reported by the customer but negatively impact the DH network. An example would be a customer reporting a valve that is stuck in the closed position due to a lack of heating. Contrary, a valve stuck in the fully open position would not cause the customer to log a complaint immediately as heating performance is not negatively impacted. The same goes for limescale buildup in a heat exchanger, which is removed at regular maintenance intervals but may be more economical to remove on a shorter timescale. For this reason, we recommend a manual review of the time series data. During the review, additional faults may be discovered, which impacts data set quality as well as ML model training and evaluation. Furthermore, additional flags for abnormal behavior can be added for timeframes that experts deem relevant for an ML algorithm to detect, yet do not fall into the fault category. A suitable tool for reviewing and labeling large sections of time series data is described in 4.5.2.

4.4. Topological data curation

GIS data that depicts the network topology of the DHS predominantly comprises data as nodes and pipes. The data characterizing the network can be categorized into two main types: continuous data, such as geographical location details, and categorical data, encompassing pipe types, diameter values, pipe status, and more. A subset of the most useful features are described in Table 2 for nodes and Table 3 for pipes. The curation of topological data poses significant challenges, given that only a sparse portion of physical entities remain visible, offering limited opportunities to cross-validate obscure elements.

4.4.1. Missing or mislabeled categorical metadata

Initiating data curation by leveraging open-source tools like Open Street Maps to reference factual elements within the meta-data, such as address values, proves instrumental in augmenting missing attributes and validating existing data. The adoption of a standardized address schema for network entities aids in identifying erroneous mappings within the native database. Extending the data from a tabular to a graphical representation enables graph-based algorithms to redress discrepancies in the data. Attribute propagation facilitated by graphs enables the transfer of metadata from one element to its connected components to address gaps or inaccuracies in the data. DHS benefit immensely from such features, as they are analogous to tree-like structures with branches and sub-branches. A potential benefit is realized, when flow through a pipe is suspended through valves, the meta-data indicating the status of the pipe switches from used to unused. As a consequence, sequential pipes adjoining the unused pipe are also discontinued. Manual labeling, in such instances, can be prone to mislabeling meta-data.

4.4.2. Structural discontinuities

Considering network design influencing factors associated with heat suppliers, consumers, and network geography, the network may exist as one or more independent networks. When represented as graphical networks these independent networks are known as sub-graphs. However, the transition from physical to digital representation may introduce challenges due to measurement and digitization inaccuracies, resulting in physically connected entities appearing separate in the digital space, with occasional translational errors or missing components. Based on node attributes (type, address, etc.), pipe attributes (address, status, etc.), and the geographic distance between related network components in the neighborhood, analytical formulation encourages a feasible solution. The nature of the analytical solution addressing discontinuities in a couple of instances is visualized in Fig. 3.

In addition, the presence of nodes with unique IDs in the same location, connected to different pipe sequences, can lead to disruptions, causing an increase in sub-networks. To address redundant nodes, a virtual connection ensures the preservation of network integrity without introducing geometric disruptions. Furthermore, addressing the absence of mapping between the DH network and smart meters, which often provide geographic coordinates, becomes crucial. An example is shown in Fig. 4. Leveraging property cadastre data and employing analytical methodologies, a comprehensive mapping based on the positions of pipes and meters can be systematically formulated, enhancing precision and completeness in the network representation.

4.5. Software tools

In this section, we present two types of tools used for data curation. The two present opposite approaches to data curation while solving tasks that the other is not suited for. Of note is the potential for private data contained in the data set, especially when using event-based tools, such as names, addresses, and private phone numbers. Therefore, this kind of data must be handled with care and consideration for local privacy policies such as the General Data Protection Regulation (GDPR) in the EU, especially when using online services that may be located outside the EU.

4.5.1. Event-based tools

First, an event-based tool is used to review the existing stock of fault reports, join them with the correct time series data, e. g. substation sensor data, and add additional information such as ToF and ToN. It is, however, limited to working with existing events. A streamlined workflow can proceed as follows:

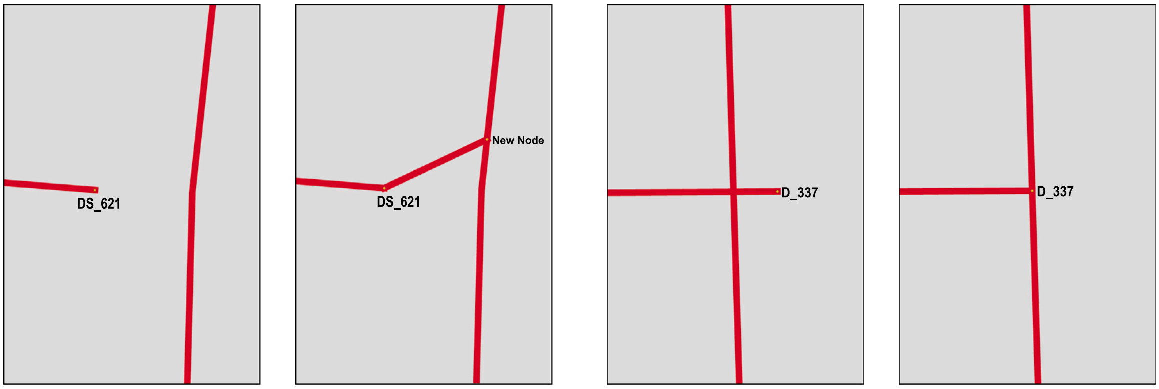
First, a fault is selected from a list of existing faults. The user is shown existing information such as incident ID, customer report, and repair technician report. Judging from these, a fault can either be classified as irrelevant for ML training, e. g., it was a false alarm, or it is joined with the correct time series data by selecting the correct substation (cf. Section 4.3.1). Once a substation is selected, the time series data are plotted and displayed. To show only relevant data and avoid confusion, the ToCC and ToR timestamps are used to narrow down the timeframe. Depending on the specifics of the DH network, an offset is determined to show 5–14 days prior to and after those timestamps. This prevents flooding the user with unnecessary information but provides context for the next task, which is assigning the ToF and ToN. For this, the ToCC and ToR are shown in the graph alongside the time series data. The user can now select the last time the graph data looks normal and the first time, normal behavior has been re-established. Finally, experts familiar with this kind of data can verify certain root causes as listed in the repair log.

4.5.2. Time series-based tools

Contrary to the event-based workflow, a time series-based workflow is designed to review the time series data of individual substations while allowing fast labeling of intervals. This is necessary because not all faults are noticed by the customer and thus reported. To avoid doing the same work twice, we recommend first using event-based tools and only switching to time series-based tools once the event-based work is complete. Faults already reviewed can then be imported and highlighted in the time series.

Table 3
DH Network – Key meta-data describing the pipes present in the network.

Feature	Description
From Node ID	Node ID of the starting node of the pipe – alpha numeral code
To Node ID	Node ID of the ending node of the pipe – alpha numeral code
Type	Classification of pipe type based on supply hierarchy – branch/supply/transmission/main pipe
Material	Material of the pipe
Activity Status	Is the pipe active? – exist/discontinued
Pipe ID	Unique ID of the pipe (alphanumeric code)
Geometry	Geometry of the pipe delineated through a set of sequential geo-coordinates (UTM/WGS-84)
Address, Road Code	Address the pipe is present in, identifier for roads



(a) Disconnected pipes (b) Fixed disconnection (c) Disconnected pipes through overshooting (d) Fixed disconnection

Fig. 3. (a) Digital representation of an instance where pipes are not connected due to measurement errors. (b) The fixed network after preprocessing the digitized data. (c) Digital representation of an instance where pipes are not connected due to overshooting. (d) The fixed network after preprocessing the digitized data.

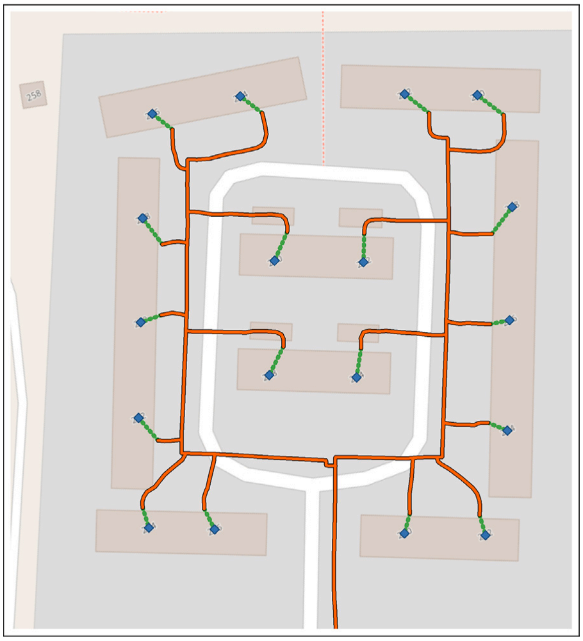


Fig. 4. A virtual mapping between heat meters and pipes, to have an end-to-end representation. The orange lines and blue points depict pipes and meters respectively, and the dotted green line represents the mapping.

Due to the large workload of reviewing long periods of time for a large number of substations, a fast and efficient workflow is critical. For both kinds of tools, this is mainly regarding the process of moving along the time axis, showing a sufficiently high resolution to detect problems, while at the same time not getting slowed down by minute details, as well as creating new labeled sections in the time series using hotkeys. Two examples of software suitable for this purpose are Label Studio [22] and Kili [23]. An example using Label Studio can be found in 5.4.

5. Case study: ILSE project

Project ILSE [24] aims to use individual DH substation data to detect faults using supervised machine learning methods. It is a co-operative project including researchers from the Technical University of Applied Sciences Rosenheim, and INEV Institut für nachhaltige Energieversorgung GmbH, as well as the local district heating company Stadtwerke Rosenheim GmbH & Co. KG and the district heating industry association AGFW. It is funded by the German Federal Ministry for Economic Affairs and Climate Action. For this purpose, DH substation sensor data is combined with customer fault reports, after which the time-series data undergoes a manual review process.

For the ILSE project, two primary data sources are available from a German DH network in a city with a population of about 60k. The first data source contains historic smart meter readings consisting of supply and return temperature as well as flow rate as raw measurements and the heat energy drawn from the network calculated from those. These values are available at an hourly sampling rate. The second data source is the event log of customer calls concerning district heating issues. This includes the ToCC and ToR, customer address, customer fault description, and, for more recent reports, information from the repair technician about the root cause selected from a list. A graphical overview of the data curation process described in the following sections can be found in Fig. 5.

5.1. Additional data sources

Other data beyond substation sensor data can help experts interpret the raw sensor data and thus should be made available to ML algorithms.

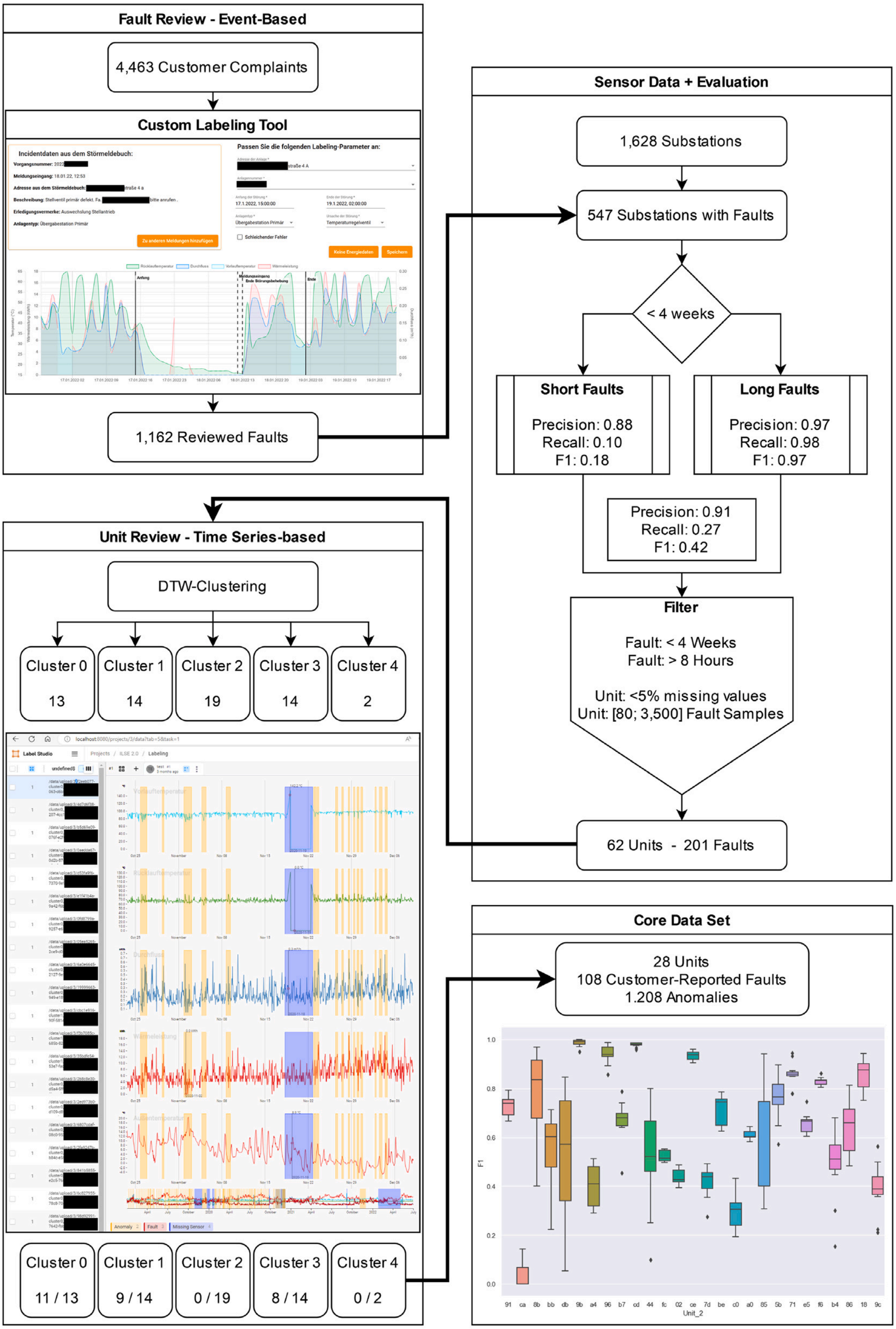


Fig. 5. Overview of the data set creation in the ILSE project.

First, heat demand is strongly correlated with the outside temperature. For this reason, we obtained a copy of hourly temperature measurements from a public data repository managed by the German meteorological service “Deutscher Wetterdienst” (DWD). To fill in missing values, three surrounding weather station records were also obtained, and a Support Vector Machine (SVM) was trained to predict local values from surrounding stations. This yielded significantly better results compared to polynomial or spline interpolation.

In addition, heat demand also follows human activity, as can be seen in [20]. To make this information available, we use several encodings such as decimal weekday, one-hot weekday, hour of the day, week of the year, and one-hot public holiday for the measurement and fault data period.

The time axis for both resulting data sets uses hourly resolution, making them easy to join with the sensor and fault data.

5.2. Joining sensor and fault data, fault review

During this first step of data augmentation, we aim to combine DH substation sensor data with customer reports while also reviewing the report data. For this, we face three key issues, which must be accommodated by the labeling software.

1. Each fault must be assigned to a substation with only the customer-reported address to identify the affected substation.
2. While ToCC and ToR are known, ToF and ToN must be defined manually.
3. Information from the customer report as well as technician repair remarks must be available during the labeling process to properly perform the task.

To solve these issues, we developed a custom browser-based labeling tool, which can be seen in Fig. 5 in the Custom Labeling Tool-box. Our event-based Labeling Tool will be made public via Github [25], however, at the point of writing is not yet available. Initially, the user is presented with the customer and repair technician reports. If possible, the address is matched to one from the known substation addresses and must only be confirmed by the user. Otherwise, the user is offered a search bar. Once an address has been confirmed, all substations located there can be selected from a dropdown list. Upon selection of a substation, sensor data from the time frame defined by ToCC and ToR with a margin of 2 weeks is displayed below. ToCC and ToR are shown in the graph as visual cues. If multiple substations are available or the wrong address has been selected, the lack of a visible fault signature speeds up this part of the labeling process. With the correct substation selected, ToF and ToN can be defined by selecting points in the time series graph. Labelers were asked to only accept customer reports with visually recognizable deviations in the sensor data.

Out of 4463 available customer complaints, this process yielded 1162 faults affecting 547 out of 1628 substations. Out of these, 661 faults occurred or ended between 01.01.2019 and 30.06.2022, which is the core time frame used for further studies as the data quality for both sensors and customer reports decline in earlier years.

5.3. Intermediate evaluation

Using this preliminary data set, several different ML models were implemented, using only substations with labeled faults. The task is to classify each hour as fault or no-fault, using historical time series data covering between 8 and 168 h. For ground truth, we labeled every hour between ToF and ToN as a fault. For evaluation, we used precision, recall, and F1-score. With the initial results being very low, we further decided to split the data set by fault duration at 4 weeks. In faults persisting for more than 4 weeks, we typically see limestone build-up or wrong parameters for the control system which reduce efficiency, but only become apparent to the customer in very cold conditions. Under 4 weeks, the data set contains short-term faults such as blockages, stuck valves, or broken pumps which have an immediate effect on heat supply. The evaluation showed very good performance on long-term faults, scoring above 0.95 in all three metrics. However, the data set is heavily skewed towards fault samples including substations that are labeled as faulty for 70%–90% of the duration. Conversely, the short-term faults are heavily skewed towards no-fault samples with typically 10–40 h of fault-samples per fault occurrence and typically 1–2 faults per unit over the 3.5 years in the data set. For these cases, we found an acceptable precision between 0.80 and 0.90 but a recall of around 0.1, indicating the loss of most faults. Across all faults and affected substations, we found a precision between 0.85 and 0.95, a recall between 0.25 and 0.30, and a resulting F1-score of about 0.4 in the best cases. As such values are well below any practical usability and several different ML methods showed the same results and limitations, we concluded that the problem must originate from the data set. This resulted in a second labeling process using different tools.

5.4. Time series manual review

With no further data sources to add to the data set and poor performance from the ML models, the focus shifted to data quality. For this, a selection of substations with the best sensor data quality and the best short-term fault data was created, ultimately consisting of 62 substations and 201 faults. For the selection, we only considered faults with a duration between 8 h and 4 weeks, as shorter faults consist of very few samples which makes them difficult to detect due to system inertia, and longer faults were already shown to work. Next, substations with more than 5% missing sensor values were removed. Finally, only substations with more than 80 and less than 3500 fault hours during a 3.5 year period were considered to filter out substations with very low fault hours as well as remove substations with both short and long-term faults overlapping. Finally, the 62 substations are clustered into 5 groups according to their heat consumption behavior using a similar method as presented by [20]. These substations were then loaded into the Label Studio software [22] for ease of use and provided to DH experts as shown in Fig. 5 in the Unit Review – Time Series-based box. Their task at this point is to label any abnormal-looking time series data, they deem important for an ML model to detect.

The resulting data set consists of 28 substations with 108 of the original customer-reported faults as well as 1208 newly labeled anomalies. Two clusters could not be labeled as the behavior of the included substations is too erratic to determine with the other clusters being partially labeled. Including anomalies with the other faults and limiting training and evaluation to the 28 substations yields an F1-score of 0.87 with a recall of 0.95 and a precision of 0.81 averaged over 11 experiments. An overview of the model performance across 11 runs can be seen in Fig. 5 in the Core Data Set box, showing a box plot of the F1-score for each individual substation in the data set.

6. Discussion

In this section, we address open points and practical concerns that are important to the task of data set creation but are not part of the actual process and thus do not fit into other sections. However, they are important enough to be mentioned as they can significantly impact the chance of success, reduce workload, or shift targets during the goal definition phase.

6.1. Data reduction

During the data set creation process, the majority of the initially recorded data may have to be discarded, and are thus unavailable for the ML process. The following figures were determined from the ILSE project (Section 5) and may give a general insight to set expectations for other researchers: First, out of 4463 recorded customer complaints, only 1162 were left after the initial review. This was partially due to focusing on recent fault data as sensor coverage and quality improved over the years. However, many faults were lost to the condition that faults must be visible to a human expert in the sensor data. This also removed faults for which no sensor data was available for experts to review. Second, those 1162 faults only affected 547 out of 1628 DH substations. Faults in robust systems are rare events and well-dimensioned substations with proper maintenance may not suffer a single fault over a 3–4 year period. Lastly, by focusing the time-consuming and thus expensive manual time series review on only the most promising substations with both high-quality sensor and fault data, the data set was reduced to 62 units affected by 201 reviewed faults. Of those, only 28 time series were suitable for a manual review to label anomalies. Overall, only 1.72% of the available substations and 2.42% of faults were included in our final data set shown in the case study in Section 5. This opens up a path for time savings by optimizing and filtering the faults and substations available to the reviewers to avoid losing already reviewed faults or substations in later process stages.

6.2. Automation

Automating the fault or time series review process at the current stage is very difficult. The current main objective in creating a labeled district heating fault data set is twofold. One is to create a ground truth to quantitatively evaluate existing fault and anomaly detection algorithms. The other is to create a labeled data set to unlock supervised ML methods to perform fault detection better than current unsupervised methods. A review of data sets and fault detection methods can be found in [9]. In short, there is currently no freely available labeled data set nor are comparable evaluation metrics used in recent efforts. There is also no established method how to label faults. Facing these challenges, any automation can only be very task-specific. Furthermore, any form of automated fault review is highly likely to solve the actual task of fault detection. This opens up the potential for a self-sustaining system once fault detection has been solved which can adapt to gradual changes in system behavior. Finally, there is potential to partially automate the review process using imperfect ML models to create fault suggestions that are faster to review and adjust. Another similar use case for imperfect fault detection models is as a second reviewer checking for missed faults in already reviewed files. Software suites such as Label Studio [22] support ML backend applications for such recommendations which can be updated regularly to include the most recent labels added or confirmed by human reviewers.

6.3. Fault types

In this work, we mention several different fault types explicitly as examples. We also cite several other works with exhaustive lists of potential faults. However, we do not intend this work to deal with specific fault types but rather encourage others to confer with local DH experts on the topic of faults. For our case study, we worked with already recorded fault types provided by the DH company and an “other faults” category for extremely rare faults to avoid being overly specific. In general, faults can be grouped in the following ways for practical purposes.

6.3.1. Faults and anomalies

Faults directly impact the customer to the point of triggering a customer complaint, whereas anomalies go unnoticed and are therefore still hidden in the data set after a fault review as described in Section 5.2 and can only be discovered as described in Section 5.4. However, depending on severity and situation, the same degree of limestone buildup can either be an anomaly on a warm spring day where a customer will not notice the loss in efficiency, or it can be a fault during especially cold and persistent winter conditions when the heat transfer capacity is no longer sufficient to heat the living space. Detection of anomalies is still of value to DH companies to improve efficiency and perform preventative maintenance. During our case study, we asked reviewers to go back and label the anomaly/fault start at the point, where a repair would be economically advisable as that would be the best point in time for an ML model to alert the user.

6.3.2. Immediate and creeping faults

While immediate faults are relatively easy to label and detect with several hours to days of historical data, creeping faults require long-term information to be detected. This can be done by directly providing historical data, e. g. from one year ago, to the ML model. However, this results in significantly larger input vectors and thus more complex models. Comparatively less complex would be providing information about, e. g., correlation between outside temperature and heat demand, as those variables are highly correlated. By tracking the development of this correlation, a gradual loss in efficiency may also be detected. However, we recommend separate models for immediate and creeping faults to optimize reaction time for immediate faults and historical context for creeping faults without compromising either performance.

6.4. Privacy issues, legal and market constraints

Besides technical problems, privacy issues and legal and market constraints often impede research in creating or sharing real-world data sets. Privacy concerns are often mentioned when smart meter data are collected [26,27]. Users of smart meters are concerned that the recorded data can be used to reconstruct their activities at home and violate their privacy. In [26], the authors list concerns ranging from stalking and spying on users of smart meters to commercial usage, such as targeted advertising. To deal with the aforementioned concerns several techniques are proposed in the literature, e. g., anonymization, pseudonymization, aggregation, or the addition of noise [27,28]. Anonymization is concerned with removing any personally identifiable information from a data set so that the natural person(s) the data relates to cannot be identified but stays anonymous. Pseudonymization, in contrast, is the process of replacing personally identifiable information from a data set with pseudonyms to be able to restore personally identifiable information later. As data resolution determines how much information is disclosed, authors also propose spatial and temporal aggregation of data [26]. Spatial aggregation could include combining data of multiple consumers in a neighborhood, while temporal aggregation might include decreasing the temporal resolution of data. Further, researchers add generated noise to data to protect users’ privacy [28].

As current research focuses mainly on metering data, such as mass flow, supply, and return temperature, and often leaves out potentially personally identifiable socio-economic data [9], we argue anonymization and pseudonymization should be preferred over aggregation or the addition of noise. Anonymization and pseudonymization allow to leave metering data at a high spatial and temporal resolution, which enables even short-lived faults and anomalies to be detected. However, since the socio-economic data of users has hardly been used in fault and anomaly detection, further research should clarify how socio-economic data can be used without violating users’ privacy and what influence it has on corresponding models.

While smart meter data are likely available in many existing district heating systems, using and publishing these data for research purposes can be subject to further legal restrictions. Depending on the jurisdiction, personal data collected for a specific purpose, e. g., heat meter data used for billing, cannot be used for a different purpose without a separate legitimation, i. e., the user’s consent [27]. Therefore, researchers should seek the explicit consent of customers and document it in writing. To obtain the customers’ consent, researchers have to provide information and transparency about the published data, the publication process, their rights, and the importance of publishing data sets in a scientific environment. Additionally, incentives can help to gain the consent of users.

Lastly, market constraints may impede researchers from publishing data: District heating operators might see smart meter data as a valuable source of information that has to be kept secret from competitors to stay ahead of the market. While we argue that public datasets accelerate the scientific progress process and improve comparability between different methods, there are technical ways to minimize

Table 4
Summary table of features and properties for a minimum viable substation-based fault detection and classification data set.

Minimum		
Source/Type	**Feature**	**Properties**
Substation	Supply/return temp., flowrate, prim. side	1 h sampling rate
Social	Working day	boolean
Social	Public holiday	boolean
Social	Local special events	boolean
Fault	Time of Failure	timestamp
Fault	Time of Customer Complaint	timestamp
Fault	Time of Repair	timestamp
Fault	Time of (re)Normalization	timestamp
Fault	Root cause	Limited list of causes (e. g. [14])
Meteorological	Outside temperature	1 h sampling rate
Recommended		
Meteorological	Temperature forecast	1 h sampling rate, 24–72 h period
Fault	Replacement parts	Bill of Materials
Substation	Supply/return temp., flowrate, prim. side	3–5 min sampling rate
Optional		
Meteorological	Solar irradiation	1 h sampling rate
Meteorological	Relative humidity	1 h sampling rate
Substation	Supply/return temp., flowrate, sec. side	1 h or better
Substation	Valve positions	1 h or better

information sharing in machine learning. In a federated learning setting [29], a service provider, such as a research institute, could provide an initially trained model without publishing the data set used in training. The initial model is then shared between other parties or clients, such as district heating operators or researchers, to continue training with their private data sets in their domain. The clients then report their adapted models to the service provider, who aggregates the results and starts a new iteration by sending an updated model to the clients. In this way, a shared model can be trained without sharing the underlying training data. While the idea of federated learning is promising and could also provide advantages for privacy protection, it also has major drawbacks: Besides technical requirements such as additional infrastructure or interoperability of data sets, the heterogeneity of data sets might prevent successful training and evaluation or induce unwanted bias.

7. Conclusion

In this paper, we address the problem of creating suitable data sets for data-driven fault detection in district heating substations. We propose a systematic and domain-specific process for data set creation for fault detection in the form of practical guidelines, which are based on the cross-industry standard CRISP-DM. The proposed process concretizes the cross-industry standard and thus simplifies application for the district heating domain. We focus on the process steps goal definition, data acquisition and understanding, and data curation. Therefore, we survey and discuss typical data sources that can be used to achieve fault detection, such as heat meter data, grid information, consumer information, and fault logs alongside calendrical and meteorological data. Further, we present and discuss typical data curation techniques and tools to bring the aforementioned data sources together, address data quality issues, and transform the resulting data for use in ML models. We answer the question of how to deal with poor data quality by proposing methods, e. g., for the removal of implausible values, the removal of structural discontinuities in GIS data, or the imputation of missing values. Data sets created and curated with the proposed process and methods can then be used to train and evaluate ML prototypes to give an impression of whether the goals defined can be met. More specifically, it leads to a standardized minimum feature set as described in Table 4, which accelerates development and promotes collaboration in the field while also showing a path towards more complex data sets for future use. With our work, we aim to enable and encourage researchers and practitioners to create data sets for fault detection in the district heating domain to also enable and improve ML models in this domain. Finally, we demonstrated the effectiveness of our proposed process in a case study where a labeled district heating data set was created by showing the incremental improvements in prediction quality associated with each iterative step.

CRediT authorship contribution statement

Dominik Stecher: Writing – original draft, Methodology, Investigation. **Martin Neumayer:** Writing – original draft, Methodology, Investigation. **Adithya Ramachandran:** Writing – original draft, Visualization, Methodology, Investigation. **Anastasia Hort:** Software, Methodology, Investigation. **Andreas Maier:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Dominikus Bücker:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jochen Schmidt:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT 3.5 in order to help improve the readability for rephrasing long and complex sentences in some parts of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

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

Acknowledgments

We thank our project partners at Stadtwerke Rosenheim Netze GmbH for providing district heating data and expertise during the labeling process. This work was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), grant 03EN3033, and supported by the Bayerische Wissenschaftsforum (BayWISS). The work is also funded by the Bavarian State Ministry of Economic Affairs and Media, Energy and Technology, grant DIK0325/01.

Data availability

No data was used for the research described in the article.

References

[1] Lund H, Möller B, Mathiesen B, Dyrelund A. The role of district heating in future renewable energy systems. Energy 2010;35(3):1381–90. http://dx.doi.org/10.1016/j.energy.2009.11.023, URL https://www.sciencedirect.com/science/article/pii/S036054420900512X.

[2] Johansen K, Werner S. Something is sustainable in the state of denmark: A review of the danish district heating sector. Renew Sustain Energy Rev 2022;158:112117. http://dx.doi.org/10.1016/j.rser.2022.112117, URL https://www.sciencedirect.com/science/article/pii/S1364032122000466.

[3] Jimenez-Navarro J-P, Kavvadias K, Filippidou F, Pavičević M, Quoilin S. Coupling the heating and power sectors: The role of centralised combined heat and power plants and district heat in a european decarbonised power system. Appl Energy 2020;270:115134. http://dx.doi.org/10.1016/j.apenergy.2020.115134, URL https://www.sciencedirect.com/science/article/pii/S0306261920306462.

[4] Gadd H, Werner S. Fault detection in district heating substations. Appl Energy 2015;157:51–9. http://dx.doi.org/10.1016/j.apenergy.2015.07.061.

[5] Østergaard DS, Smith KM, Tunzi M, Svendsen S. Low-temperature operation of heating systems to enable 4th generation district heating: A review. Energy 2022;248:123529. http://dx.doi.org/10.1016/j.energy.2022.123529.

[6] Averfalk H, Werner S. Essential improvements in future district heating systems. Energy Procedia 2017;116:217–25. http://dx.doi.org/10.1016/j.egypro.2017.05.069, 15th International Symposium on District Heating and Cooling, DHC15-2016, 4-7 2016, Seoul, South Korea.

[7] Månsson S, Johansson Kallioniemi P-O, Thern M, Van Oevelen T, Sernhed K. Faults in district heating customer installations and ways to approach them: Experiences from Swedish utilities. Energy 2019;180:163–74. http://dx.doi.org/10.1016/j.energy.2019.04.220.

[8] Leoni P, Geyer R, Schmidt R-R. Developing innovative business models for reducing return temperatures in district heating systems: Approach and first results. Energy 2020;195:116963. http://dx.doi.org/10.1016/j.energy.2020.116963, URL https://www.sciencedirect.com/science/article/pii/S0360544220300700.

[9] Neumayer M, Stecher D, Grimm S, Maier A, Bücker D, Schmidt J. Fault and anomaly detection in district heating substations: A survey on methodology and data sets. Energy 2023;276:127569. http://dx.doi.org/10.1016/j.energy.2023.127569, URL https://www.sciencedirect.com/science/article/pii/S0360544223009635.

[10] Lund H, Werner S, Wiltshire R, Svendsen S, Thorsen JE, Hvelplund F, et al. 4th generation district heating (4GDH): Integrating smart thermal grids into future sustainable energy systems. Energy 2014;68:1–11. http://dx.doi.org/10.1016/j.energy.2014.02.089.

[11] van Dreven J, Boeva V, Abghari S, Grahn H, Al Koussa J, Motoasca E. Intelligent approaches to fault detection and diagnosis in district heating: Current trends, challenges, and opportunities. Electronics 2023;12(6). http://dx.doi.org/10.3390/electronics12061448, URL https://www.mdpi.com/2079-9292/12/6/1448.

[12] Månsson S, Thern M, Johansson Kallioniemi P-O, Sernhed K. A fault handling process for faults in district heating customer installations. Energies 2021;14(11). http://dx.doi.org/10.3390/en14113169.

[13] Wirth R, Hipp J. Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, vol. 1. 2000, p. 29–39.

[14] Månsson S, Lundholm Benzi I, Thern M, Salenbien R, Sernhed K, Johansson Kallioniemi P-O. A taxonomy for labeling deviations in district heating customer data. Smart Energy 2021;2:100020. http://dx.doi.org/10.1016/j.segy.2021.100020.

[15] Sandin F, Gustafsson J, Delsing J. Fault detection with hourly district energy data : probabilistic methods and heuristics for automated detection and ranking of anomalies. Tech. rep., 2013.

[16] Oracle: Machine learning process. 2023, https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/21/mlsql/img/ml-process.jpg. [visited on 19 December 2023].

[17] Potočnik P, Škerl P, Govekar E. Machine-learning-based multi-step heat demand forecasting in a district heating system. Energy Build 2021;233:110673. http://dx.doi.org/10.1016/j.enbuild.2020.110673, URL https://www.sciencedirect.com/science/article/pii/S0378778820334599.

[18] Fang T, Lahdelma R. Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. Appl Energy 2016;179:544–52. http://dx.doi.org/10.1016/j.apenergy.2016.06.133, URL https://www.sciencedirect.com/science/article/pii/S0306261916309217.

[19] Theusch F, Klein P, Bergmann R, Wilke W, Bock W, Weber A. Fault detection and condition monitoring in district heating using smart meter data. In: PHM society European conference, vol. 6, no. 1. 2021, p. 11.

[20] Calikus E, Nowaczyk S, Sant'Anna A, Gadd H, Werner S. A data-driven approach for discovering heat load patterns in district heating. Appl Energy 2019;252:113409. http://dx.doi.org/10.1016/j.apenergy.2019.113409.

[21] Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M. Timesnet: Temporal 2D-variation modeling for general time series analysis. In: International conference on learning representations. 2023.

[22] Label studio. 2023, https://labelstud.io/. [Accessed 18 December 2023].

[23] Kili. 2023, https://kili-technology.com/. [Accessed 18 December 2023].

[24] Ilse research project. 2024, https://projekte.th-rosenheim.de/en/forschungsprojekt/761-ilse. [Accessed 01 September 2024].

[25] Labeling tool. 2024, https://github.com/ilse-thro. [Accessed 28 June 2024].

[26] McKenna E, Richardson I, Thomson M. Smart meter data: Balancing consumer privacy concerns with legitimate applications. Energy Policy 2012;41:807–14.

[27] Asghar MR, Dán G, Miorandi D, Chlamtac I. Smart meter data privacy: A survey. IEEE Commun Surv Tutor 2017;19(4):2820–35.

[28] Giaconi G, Gunduz D, Poor HV. Smart meter data privacy. 2020, arXiv preprint arXiv:2009.01364.

[29] Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. Found Trend Mach Learn 2021;14(1–2):1–210.