| IET Smart Grid

ORIGINAL RESEARCH

Fühler-im-Netz: A smart grid and power line communication data set

Christoph Balada[1] | Sheraz Ahmed[1] | Andreas Dengel[1] | Max Bondorf[2] | Nikolai Hopfer[2] | Markus Zdrallek[2]

[1]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany

[2]University of Wuppertal, Wuppertal, Germany

Correspondence

Christoph Balada, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, 67663, Germany.
Email: christoph.balada@dfki.de

Abstract

The increasing complexity of low-voltage networks poses a growing challenge for the reliable and fail-safe operation of electricity grids. The reasons for this include an increasingly decentralised energy generation (photovoltaic systems, wind power etc.) and the emergence of new types of consumers (e-mobility, domestic electricity storage etc.). At the same time, the low-voltage grid is largely unmonitored and local power failures are sometimes hard to detect. To overcome this, power line communication (PLC) has emerged as a potential solution for reliable monitoring of the low-voltage grid. In addition to establishing a communication infrastructure, PLC also offers the possibility of evaluating the cables themselves, as well as the connection quality between individual cable distributors based on their signal-to-noise ratio (SNR). The roll-out of a large-scale PLC infrastructure therefore not only ensures communication, but also introduces a tool for monitoring the entire network. To evaluate the potential of this data, we installed 38 PLC modems in three different areas of a German city with a population of about 150,000 as part of the Fühler-im-Netz (FiN) project. Over a period of 22 months, an SNR spectrum of each connection between adjacent PLC modems was generated every quarter of an hour. The availability of this real-world PLC data opens up new possibilities to react to the increasingly complex challenges in future smart grids. This paper provides a detailed analysis of the data generation and describes how the data was collected during normal operation of the electricity grid. In addition, we present common anomalies, effects, and trends that could be observed in the PLC data at daily, weekly, or seasonal levels. Finally, we discuss potential use cases and the remote inspection of a cable section is highlighted as an example.

KEYWORDS

data set, deep learning, IoT, machine learning, power line communication, smart grid

1 | INTRODUCTION

Despite of the advances in IoT and increasingly cheaper sensory hardware, low-voltage grids remain largely unmonitored, as this is uneconomical due to their high complexity [1]. However, the aspect of comprehensive monitoring has become more important due to the growing challenges arising from the transition to a higher share of renewable energy and more complex energy consumers. On the way to meet these challenges, power line communication (PLC) has proven to be an efficient way to connect households, cable distributors and grid operators [2, 3]. By providing such a mesh network, interconnecting all stakeholders, PLC is a good candidate to back the operation of many different smart devices like chargers, metres, photovoltaic systems (PV) etc. Besides this potential as a backbone for smart grid devices, we want to shed light on another set of applications that could also be realised by using data from such a PLC network. While the communication capability of PLC networks is already

receiving attention in research, the exploration of the benefits of the data collected in PLC networks remains largely unaddressed. By attaching sensors to PLC nodes, it is possible to collect a wide range of different types of data. In this paper, however, we want to focus on the signal-to-noise ratio (SNR) measured by the PLC modems during their normal operation and without additional sensors. One reason why this data is largely unconsidered in research is not least because data from the practical use of PLC networks has not yet been sufficiently available to the research community. In this work, we want to address this shortcoming by presenting the FiN data set and potential use cases. Within FiN 38 PLC modems were distributed in three different areas of a German city with population of about 150,000. Those 38 modems yield 68 1-to-1 connections and collected data over around two years. During this time every quarter of an hour an SNR assessment as described in Section 3.4 was performed. Thus, over 3.7 million data points were collected during the project period. Along with this SNR data, we publish an extensive collection of metadata on the cables in the network. To underline the benefits of the data, we also highlight a number of different applications.

1.1 | Main contributions

We present the FiN data set holding 3.7 million data points, where each data point represents the SNR profile of a PLC connection between two cable distributors in the low- and mid-voltage grid. Furthermore, we provide metadata on weather conditions and cable characteristics. All the data was recorded during the real-life usage of the PLC nodes. The signal quality is measured by assessing the SNR on 917 channels with different frequencies between 2 and 30 MHz. The data was collected over a period of 22 months and involves 38 different PLC nodes inside the cable distributors. FiN data set aims on a wide range of different potential applications:

- Anomaly detection (e.g. discovery of noise sources)
- Grid monitoring (e.g. detection of power outages, localisation of cable breaks)
- Asset management (e.g. lifetime estimation of cables, meta data estimation)
- Additional security layer (e.g. detection of attacks against the grid infrastructure, manipulations)
- Misc (e.g. Cloudiness estimation through PV)

This broad range of applications highlights the potential of this data, but on the same side, also the complexity of evaluating every use case in detail.

In this paper, we propose a novel data set consisting of over 3.7 million real-world PLC measurements (S-3). The data acquisition process is described in detail in S-3.1. In S-4, we provide a first analysis of trends and different types of interference in the FiN data set and S-5 discusses the aforementioned fields of application in detail. Our analysis highlights the uniqueness of the proposed PLC data set, and its

relevance for a variety of future applications. This paper should raise awareness and encourage the research community to further investigate the various possibilities of FiN and similar data sets.

2 | RELATED WORK

Electricity grids are an essential part of the economic and personal lives of most people on earth. However, the electricity grid is presently undergoing one of the biggest transformations since its invention [1]. Climate changes are unquestionable and with rising urgency for changes in the way we produce, consume, store and transmit electrical power, new approaches for our electrical grid become inevitable. On the other hand, digitalisation and automation are opening up new ways of dealing with these changes. To address both problems many different aspects of future grid systems, often called smart grids, need to be rethought.

A big number of research works concern especially forecasting of future demand in electrical power to optimise generation and demand [4–11]. Since it is a complex task to store significant amount of electrical power, it is the goal to achieve an equilibrium between generation and demand. Refs. [12–16] show different data sets and works concerning electrical load forecasting and address different scenarios on hourly or daily basis, for households, big buildings, residential buildings, cities or whole countries. Especially due to financial interests, the topic of consumption forecasting receives a lot of attention. In the context of the transformation towards a smart grid, however, this area only covers one aspect. Topics such as asset monitoring, grid automation or grid security benefit comparatively little from this financial interest. In contrast, the FiN data set attempts to address these less highlighted use cases in particular and its SNR data from practical application adds another important building block that will support research. In the following, an overview of other possible fields of application of SNR data is given, even though these were mainly carried out on simulated data and were not evaluated in practice.

- **Asset monitoring** deals with the monitoring and assessment of the grid assets. This includes, for example, the recording of the ageing process and partial discharges [17–20] or the detection of events such as a fuse failure [21]. As refs. [17–19, 21] show, SNR data is also suitable for this area. Section 5.1 shows an example of how the FiN data can be used to estimate the number of connecting joints installed without manual measurements on site.
- **Security** is also of increasing importance due to the increasing digitalisation of the electricity grid. More intelligent controlling also opens up more routes of attacking the electricity grid. Additional devices that supervise the correct operation of the grid and that act independently of the normal grid operation have been successfully tested in ref. [22]. The PLC infrastructure can be used analogously for decentralised, secure and independent monitoring of the

network status [23]. The FiN data can be used to enable research on real data.

- **Grid stability and automation** is a topic of major importance, especially due to the increasing complexity of the electricity grid. Many different works [24–28] deal with different aspects and techniques to ensure stability or automation. However, most of them have in common that they are significantly based on grid parameters, such as voltage or frequency. The FiN data set, together with the SNR data, provides an additional new foundation of data, which also opens up new scope for research in this area.

3 | PROPERTIES OF FIN

FiN consists of 3.7 million SNR assessments collected from 38 PLC nodes that form 34 connections, since each connection is measured in both directions, the data set shows overall 68 connections. A highlight feature of FiN is that the data was collected over a period of about 2 years and recorded entirely in a real-life environment. Therefore, the FiN data covers various weather conditions, seasonal effects, noise sources, trends and sudden changes on mid-voltage (mv) as well as on low-voltage (lv) level. The data was collected from 12th of April 2017 to 5th of February 2019 while 14 PLC modems are installed on mid-voltage level and 24 on low-voltage level. All cable lines examined are underground cables. Apart from the SNR measurements, metadata was collected for each PLC modem. Most important metadata are cable related data (length, type, the number of wires, cross-section, age, number of t-joints) and environmental related data (daytime, weather, characterisation of the surrounding area). See Table 1 for an overview of all available metadata information. This section provides a detailed overview of how the data was collected and Table 2 summarises the key facts.

TABLE 1 Summary of the available metadata

Data field	Description
year of installation	Year in which the cable/cable section was installed
year approximated	Whether an installation year was approximated or not
Cable section	Index of the respective cable section
Length	Length of the cable/cable section, in m
Number of conductors	Number of conductors
Cross section	Conductor cross-section, in mm^2
Voltage level	Voltage level, MV = mid voltage; LV = low voltage
t-joints	Number of t-joints within the respective cable section
Type	Type of the respective cable/cable section

3.1 | Device locations

Overall 38 PLC modems were installed in three different areas of a German city with around 150,000 inhabitants. Furthermore, the PLC modems are certified under the IEEE 1901 orthogonal frequency-division multiplexing fast Fourier transform Access standard. Figure 1 shows an example of how a FiN installation looks like. All PLC modems were installed in cable distribution cabinets in the case of lv and in substations

TABLE 2 A summary of the main characteristics of the data set

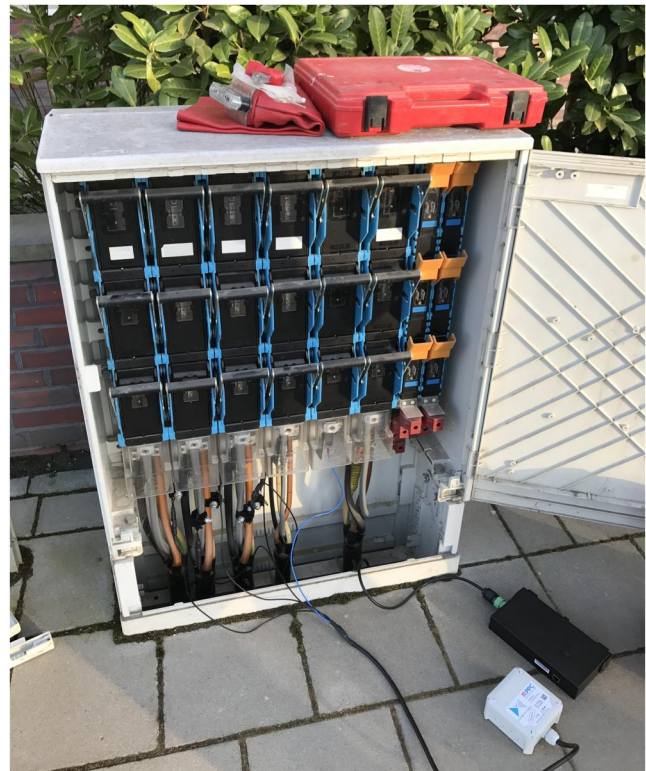| Data set key facts	
Number of nodes	38
Number of nodes on low voltage level	24
Number of nodes on mid voltage level	14
Number of 1-to-1 connections	68
Number of data points	3,747,610
Number of different locations	3
Measurement frequency	15 min interval
Measurement begin	Apr 2017
Measurement end	Feb 2019
Meta data	Weather and cable characteristics

FIGURE 1 A typical Fühler-im-Netz (FiN) installation. The power line communication (PLC) modem (black box) is connected to one phase and remains in the cable distribution cabinet.

in the case of mv. Each location shows different types of surroundings and covers different types of consumers. The following gives a short overview of the three locations and Figure 2 illustrates the network topology of all 38 PLC nodes.

3.1.1 | Location 1—low-voltage area

Residential area, small houses/single family houses, only few and small photovoltaic setups, adjacent to a big office complex and smaller industry buildings with a big photovoltaic setup.

3.1.2 | Location 2—mid-voltage area

Residential area, mainly big houses/apartment buildings, only few photovoltaic setups, adjacent to a residential area with many small photovoltaic setups.

3.1.3 | Location 3—low-voltage area

Residential area, single family houses and big apartment buildings, only few and small photovoltaic setups, adjacent to a swimming pool complex and multiple schools.

3.2 | Electrical and visible neighbours

Since all PLC modems share the cable as communication medium, they are able to talk not even to direct neighbours but also to neighbours that are multiple hops away. Therefore, it is necessary to distinguish between electrical neighbours, which are physically adjacent to each other and visible neighbours that can establish a PLC connection to each other. To keep the amount of data and redundancy low, we have decided to omit visible neighbours that are no physical neighbours. Since only electrical neighbours are shown in this data set (see Figure 2), they are simply referred to as neighbours in the following.
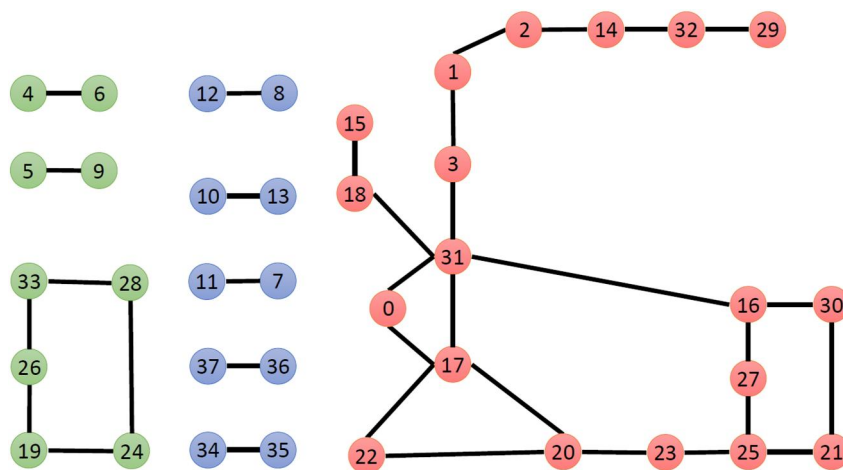
3.3 | Operation time and timeouts

Unless otherwise stated, PLC modems run 24/7. However even in case of a high quality connection of a PLC modem we observed that timeouts occur as part of the normal behaviour. The reasons for timeouts are manifold and will be discussed further in Section 4. However, if a PLC modem crashes it automatically reboots and continues its normal operation cycle, which is started with an initial SNR assessment. This initial assessment is one reason for measurements that occurred outside the 15-min slots.

3.4 | SNR assessment

Every quarter of an hour, PLC modems stop their normal communication task and perform an assessment of the connection quality by measuring the S. Therefore, each PLC modem sends a predefined signal while its neighbours are listening for this signal. Based on the discrepancy between sent and received signal the SNR is computed. Since PLC uses a range of frequencies, the SNR is computed for 917 channels used by the modem. The temporal progression of the SNR values over all frequencies is referred to as the SNR spectrum. Figure 3 shows an example of an SNR spectrum for 192 assessments that correspond to 2 days. Apart from the SNR values, PLC modems also capture the modulation level that was used to encode and decode the PLC signal. The modulation level is given in eight discrete steps and indicates the number of bits that are transmitted per signal word. While modulation level is captured in receiving and transmitting direction, the SNR values are only computed for the received signal. Similar to the SNR spectrum, the modulation level is recorded over a frequency range and is referred to as the tonemap. Figure 3 shows an example of a tonemap for 192 assessments that correspond to 2 days. In general, a higher SNR value and a higher modulation level correspond to a better connectivity in terms of data throughput.
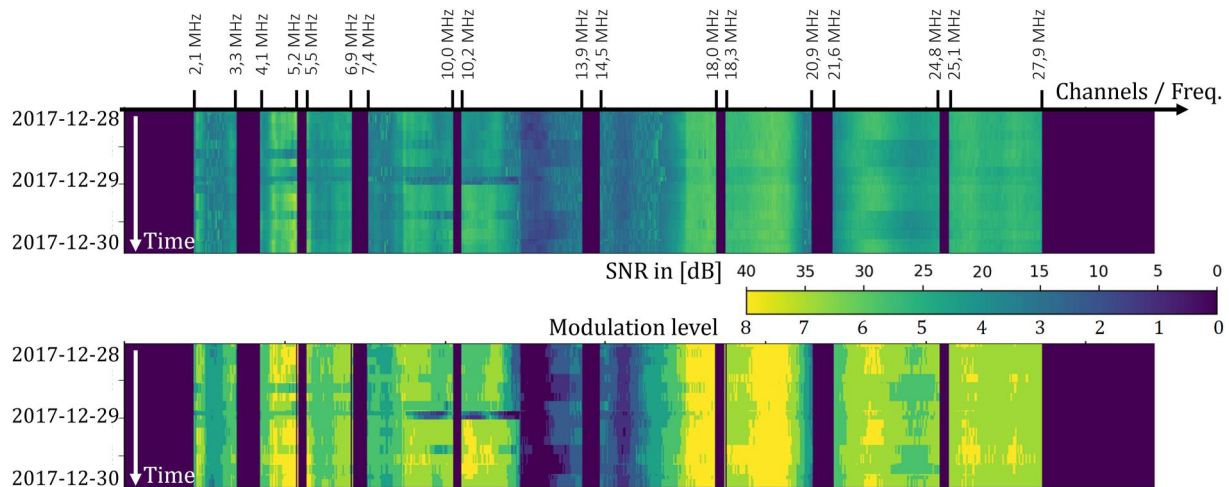

FIGURE 2 Grid graph of all three locations. Location 1 in green, location 2 in blue, location 3 in red. Topology that is formed by the power line communication (PLC) infrastructure. The numbers represent the node ID, which is also used within the data set. Since the signal-to-noise ratio (SNR) is recorded in the forward and backward directions, there are 68 measured 1-to-1 connections.

FIGURE 3 Top: Signal-to-noise ratio (SNR) spectrum for a single connection. The spectrum illustrates the SNR spectrum for 192 assessments that correspond to two days. Dead spaces in the spectrum are due to channels or frequencies that are not cleared for use in power line communication (PLC) communication. Of about 1500 channels, 917 remain useful channels. Bottom: Tonemap for a single connection. The tonemap illustrates the modulation level for 192 assessments.

3.5 | Metadata

In addition to the SNR assessment multiple metadata information were collected for each PLC modem. We distinguish between two types of metadata: cable related and environmental related metadata.

3.5.1 | Cable related metadata

Cable related metadata covers all information that were available from the cooperating grid operators or collected during the PLC modem installation.

Available metadata: *Age, length, number of wires, cross-section, voltage level, number of t-joints, cable type, and number of sections* (if a connection between two PLC modems splits up into multiple parts). Due to missing historical records, it was necessary to approximate some very old entries for the number of t-joints and the age. Oldest cable connections in this data set were estimated to be installed around 1955. Approximated values are indicated in the data set. See Table 1 for an overview of all available metadata information.

3.5.2 | Environmental related metadata

To investigate environmental influences on the PLC connection quality we provide weather data for each of the locations. Due to the limited spatial extent of each location (below 800 m) we assume a homogeneous weather inside each location. Provided weather data covers in an hourly fashion: temperature, humidity, wind speed, cloud coverage, and amount of rain. A full list of all available weather data fields and for a detailed description we refer to the documentation of OpenWeather [29] were the weather data was acquired.

4 | PATTERNS, TRENDS IN FIN DATA SET

During the full 21 months of collecting data many different trends, events, and pattern could be observed. In general, all events are classified into different groups based on two characteristics. The first is the periodicity, that is, whether an event occurs once or in a periodic scheme. The second is the time scope of an event, that is, whether an event takes place in the scope of hours, days, months or even if it lasts forever on. For example: Figure 4 shows the SNR pattern of a fuse failure that takes place once and has a lasting effect.

However, it is not possible to fully outline the reasons of anomalies, as it is not known what effects will be visible in the data. While the reason for a switching operation in the grid is obvious, most of the sources of interference remain unknown. Since even the operators of the electricity grid have no economic means to determine the causes, this information cannot be provided. Basic physical factors, like signal reflection, for example, at joints, as well as poorly shielded electronics are therefore assumed to be the main cause of the general connection quality and most interference sources. Contrary to our expectations, however, no significant influence due to the load could be noticed. Although there are fluctuations within a day, these do not correlate with the expected standard load profile. The following is an overview of different groups of visible behaviours.

4.1 | Spiking disturbers and timeouts

Spiking disturbances similar to salt and pepper noise, including complete timeouts, are part of the normal behaviour, even of nodes with an overall excellent signal quality. But especially in case of timeouts it is not possible to determine the timeout
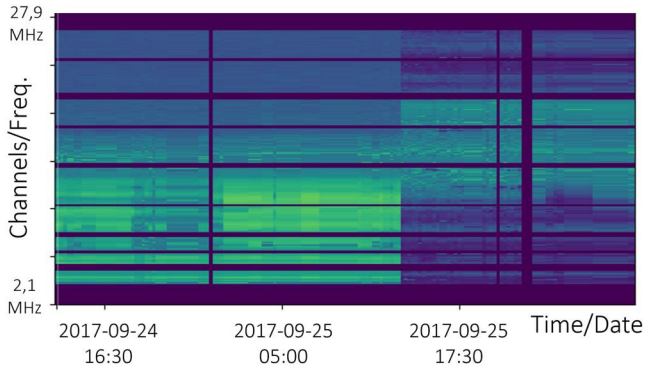
FIGURE 4 The signal-to-noise ratio (SNR) spectrum shows an imitated fuse failure. On 25th of September the SNR spectrum shows a significant break down due to an imitated fuse failure. The fuse failure was simulated by using a coupling capacitor. The effect in the SNR is similar to a high-pass filter.
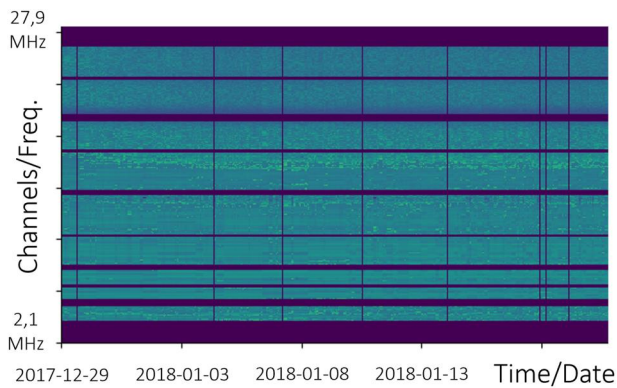

FIGURE 5 Spiking disturbances and timeouts. The spectrum shows the signal-to-noise ratio (SNR) course over time and channels. Horizontal dead spaces in the spectrum are due to channels or frequencies that are not cleared for use for power line communication (PLC) communications. The noisy course with occasional timeouts corresponds to the normal case for many nodes.

cause due to the absence of background information. Figure 5 shows an example of an SNR spectrum that shows spiking disturbances and timeouts.

4.2 | Periodic disturbances

When comparing the SNR profiles over time, a general periodicity becomes visible in most nodes. Since the sun and human behaviour are two major impact factors, this periodicity is particularly evident on a daily basis. Causes for periodic disturbances could be for example, poorly shielded hardware, broken inverters or street lighting. However, since it is not possible to track down the cause for observed anomalies the actual source stays unknown. Figure 6 shows an example with a significant SNR breakdown between 09:30am to 11:00pm. A manual evaluation of the area around the node showed that, the time window of the breakdown matches the opening hours of a neighbouring tanning studio. Nonetheless,
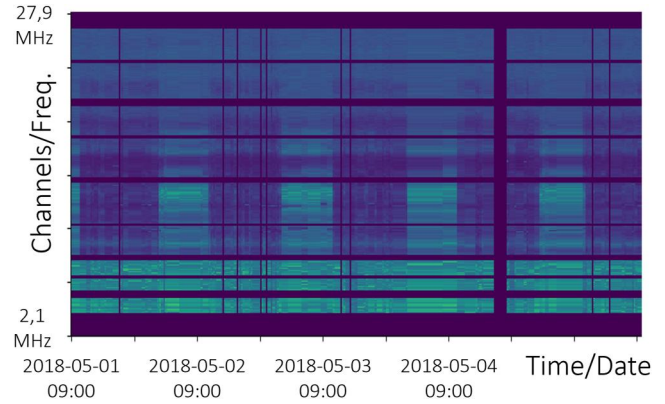

FIGURE 6 Periodic disturbances in the signal-to-noise ratio (SNR) spectrum of a connection. The time intervals of the disturbances match with the opening hours of an adjacent tanning salon.

it was not possible to prove this correlation during the project time, but the granularity that has to be dealt with becomes clear.

4.3 | Trends

Electricity grids are subject to constant change. As reported in refs. [17–19], changing cable properties due to cable age in particular can affect the SNR spectrum and result in trends on a long term perspective. Furthermore, other reasons for trends are seasons, changing grid topology or changing human behaviour like increasing energy demand. Figure 7 shows an example of an SNR spectrum under seasonal change.

4.4 | Misc

The last group covers all effects that are related to the normal operation of the grid. On the one hand side, this could be a switching operation as an expected behaviour, but on the other side also a fuse failure as an unexpected behaviour. Other examples are cable breakdowns, partial discharges, and the installation of new grid hardware like coupling capacitors. As an example for this group, Figure 4 shows the SNR pattern of a fuse failure imitated using a coupling capacitor.

5 | FIELDS OF APPLICATION

To highlight the novelty and utility of the FiN data set we want to shed light on different potential fields of application. Therefore, we present the estimation of the joint count within a cable section in more detail and give in addition an overview of other examples. Underpinning the field of asset monitoring with this joint count estimation we also want to discuss the grid monitoring in general and security aspects All experiments are based just on the SNR spectrum and do not involve any onsite
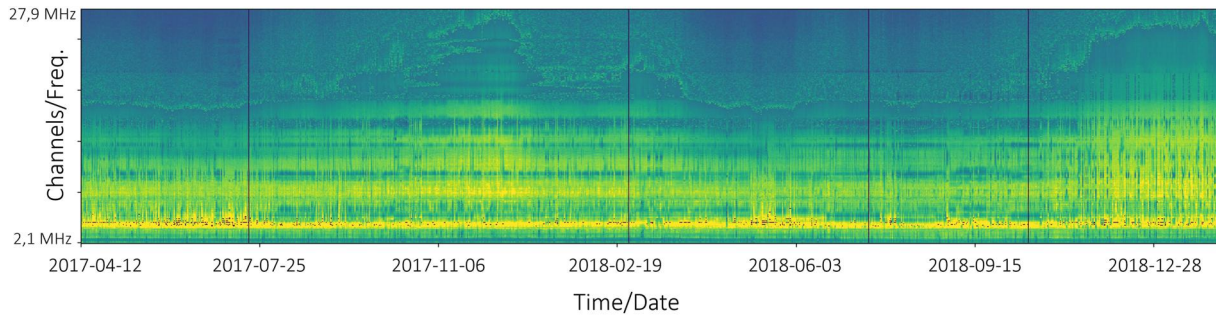
FIGURE 7 Seasonal change in signal-to-noise ratio (SNR). Especially high the higher frequencies respectively channels, show a seasonal fluctuation in the SNR. While high frequencies seem to be affected by interference in the summer months, the winter months show better signal quality.

measurements. We especially encourage the use of Machine Learning (ML) to make use of the vast amount of data that is included in FiN. However, this overview is not meant to be limited to the applications on it, but rather to encourage one to explore even more applications on top of it. Therefore, FiN is proposed as a new tool to generate new ways to deal with the manifold changes in electricity grids.

5.1 | Asset monitoring

Asset monitoring is an advantage of the wide use of PLC infrastructure. An example of these assets are the cables themselves, which can contain various numbers of joints. There are different types of joints, which either connect two different cables (transition joints), two identical cables (connecting joints) or a branch to a third end point, for example, for a house connection (t-joints). Due to the individual characteristics of the joints, or due to a different quality of installation, the joints on a cable section should have an influence on the SNR spectrum. To evaluate this, the FiN data set provides metadata of the cables. Based on the number of sections of a cable connection and the cable type, respectively, transition and connection joints can be determined. The t-joints are also explicitly indicated. To estimate the number of joints within a cable section a multilayer perceptron (MLP) is used as a baseline, which performs the regression task on the basis of individual SNR profiles. As an extended approach, a ResNet18 is utilised, which uses the SNR spectrum of one day as input.

5.1.1 | Data preparation

The FiN data set provides 38 nodes, which are randomly divided into training set with 27 nodes and test set holding 11 nodes. Since the SNR spectrum is highly correlated in the back and forth directions of two nodes, attention should be paid not to split them between the training and test data sets. Timeouts are not considered for training and no further data augmentation is performed. Finally, the training batches are scaled between 0 and 1.

5.1.2 | Baseline and ResNet18

A simple 3-layered MLP is used as a baseline, which receives a single SNR profile of a connection as input and uses this to determine the number of joints. Similarly, a ResNet18 is trained, which receives a section of the SNR spectrum as input instead of a single SNR profile. The section covers 96 time steps, which corresponds to a time span of 24 h. Furthermore, during the generation of the training batches, it is ensured that no timeouts are included in this section. In both cases, a mean absolute error (MAE) is used as loss, which has the advantage of being less sensitive to outliers. Over time, the SNR spectrum changes and may even fluctuate periodically based on days, weeks or seasons. Since these fluctuations can affect the predicted number of joints, it is primarily important that the mean or median prediction is correct. Outliers, on the other hand, which are caused by peaks or short-term noise, are therefore not problematic. The results are summarised in Table 3.

5.1.3 | Results

In our experiments, the ResNet18 was able to clearly outperform the MLP architecture. Nonetheless, a more detailed analysis of the results shows that, as expected, the accuracy of the results varies significantly. However, we assume that providing the SNR spectrum of a whole day as input for the ResNet allows the model to learn features that are more resilient to spiking noise. Figure 8 shows an example of the spectrum of a connection, the corresponding results from the ResNet, and an overlay that was created using Grad-CAM [30]. The overlay corresponds to the sensitivity in relation to the regression result of the model depending on the PLC channels involved. With high sensitivity, the given regions of the spectrum have a strong impact on the regression result. It is clearly visible that the regions involved in the sections in which an accurate determination of the number of joints is available deviate strongly from the regions that lead to other results. In the SNR spectrum it is also clearly visible that the parts for which a poor prediction was made deviate strongly from the remaining parts. We

assume that the data set as a whole still contains too few connections and thus variants of joint installations to train a model that can generalise without limitations. The influence

TABLE 3 Joint estimation results; cj = connecting joints; tj = t-joints; mae = mean absolute error; mse = mean squared error

	MAE_{train}	MSE_{train}	MAE_{val}	MSE_{val}
MLP_{tj}	0.49	0.91	4.16	70.97
MLP_{cj}	0.11	0.08	1.79	7.59
$ResNet18_{tj}$	1.33	12.79	2.35	19.69
$ResNet18_{cj}$	0.05	0.01	0.82	0.98

of each joint on the SNR spectrum is expected to vary over time and will also depend on the quality of the installation. Since both would require manual inspection on site, we use FiN to mediate this and exploit the results so far as a basis for further investigation into the aspect of how sleeves affect the SNR spectrum over time.

5.2 | Grid monitoring

The ongoing digitalisation and automation of the electricity grid is opening up new ways to use this data for monitoring. By using a decentralised infrastructure such as a PLC network, comprehensive monitoring of the electricity grid
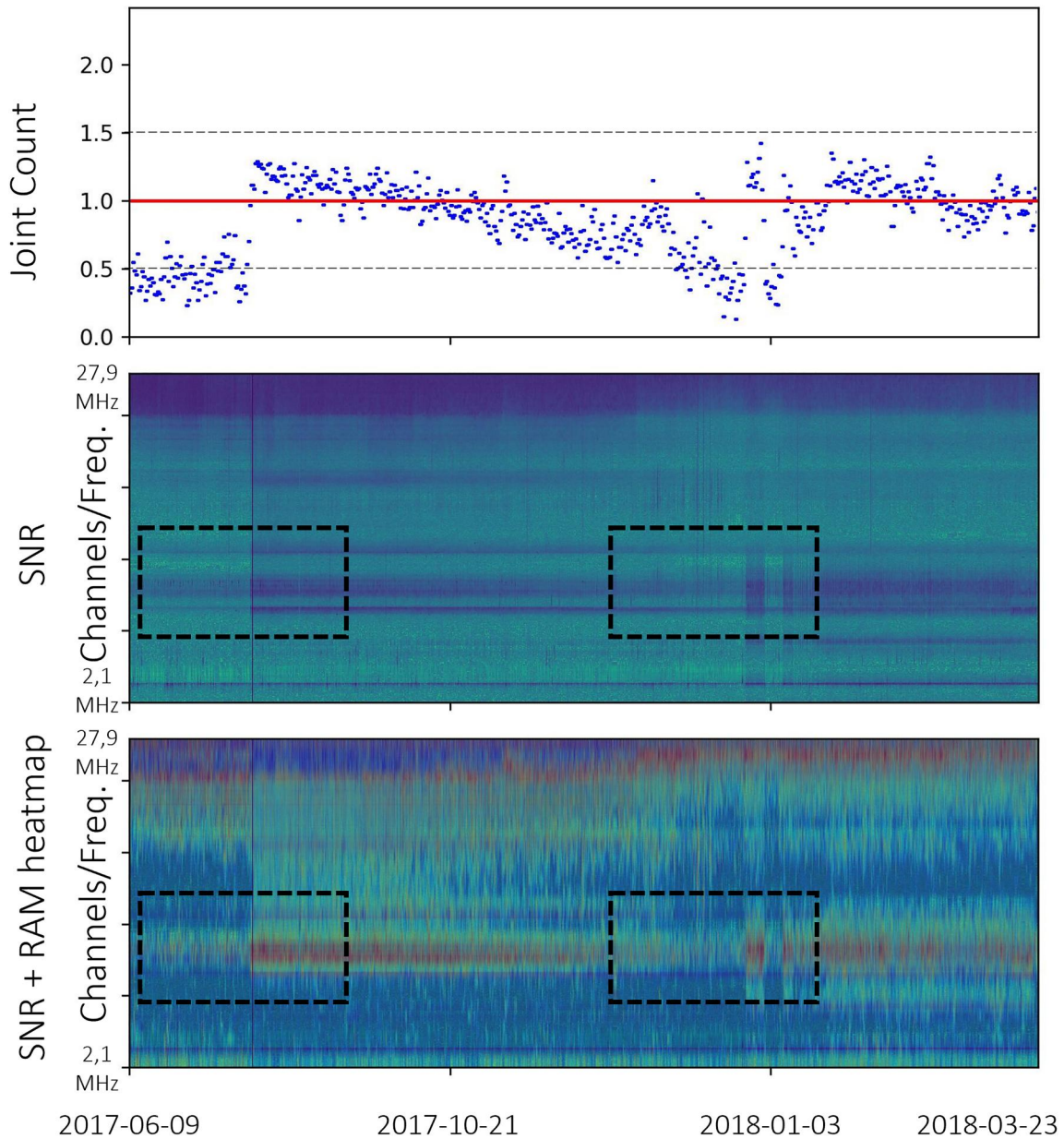

FIGURE 8 Connection joint estimation results for a joint over time. The left side shows the regression results and the corresponding signal-to-noise ratio (SNR) spectrum, while the right side overlays the regression activation in addition to the same SNR spectrum. Sudden changes in SNR, as indicated by dashed boxes, can significantly affect the regions considered for the regression task.

becomes feasible, even at the low-voltage level. Applications within grid monitoring are, for example, the localisation of disturbances and anomalies, the detection of fuse failures and partial discharges. To tackle these challenges the SNR spectrum was already proofed as a good candidate [21, 31]. However, we want to address the aspect of grid monitoring from a different direction. By using the complete FiN data set to find a suitable low dimensional representation and subsequently clusters within this representation, we aim on defining clusters of similar connection states. The SNR spectrum shows overall rich information concerning many different influences, however it also shows a lot of noise that does not affect the general state of a PLC connection. By merging similar SNR profiles into the same cluster, we want to support interpretability of the SNR spectrum and provide a good foundation for downstream tasks like anomaly detection. In the following, we show how we built a low dimensional representation of the SNR data and how we obtained cluster labels.

5.2.1 | Low-dimensional representation

A low-dimensional representation, especially in two dimensions, support computational efficiency as well as providing a good interpretability in terms of visualisation. While the former is important when thinking about a electricity grid that consists of many thousands of nodes that need to be monitored, the latter aspect is important for people during a manual inspection. In order to build a low-dimensional representation, t-distributed stochastic neighbor embedding (t-SNE) [32], uniform manifold approximation and projection (UMAP) [33] and multidimensional scaling (MDS) [34] were utilised and within the experiments it could be shown that all approaches have their limitations. While t-SNE and UMAP show good results when the number of samples is kept low (e.g. 150,000 samples), both approaches tend to break down when a million or even the entire data set is used. Figure 9 shows an example of different low-dimensional representations using t-SNE when using

different sample sizes. In contrast to t-SNE and UMAP, the MDS approach shows significant drawbacks due to its memory complexity of $O(N^2)$, wherefore MDS was not used for any further experiment. Due to these limitations, more sophisticated approaches such as ML based ([35, 36]) techniques are suggested. Nevertheless, the next two sections discuss how a low-dimensional representation, as shown in Figure 9, can be used to detect anomalies and monitor the connection states within the PLC network.

5.2.2 | Clustering

Providing low-dimensional data is not only valuable for downstream ML tasks but can also be used to distil information from the data. When operating an electricity grid, decisions (e.g. scheduling a, onsite inspection) have to be made on the basis of grid states and recognition of issues, which is why a comprehensible representation of the grid state is essential. One way to provide such a representation is a comprehensive visualisation. To construct a visualisation based on the raw SNR spectrum that can also be read by non-specialists, t-SNE is applied to a subset of 250,000 samples. Afterwards, density-based spatial clustering of applications with noise is utilised to obtain labels for the low-dimensional data. In the following, the cluster assignments are used for the visualisation instead of the SNR profiles. Figure 10 shows a visualisation of the cluster assignments in form of a radial chart. Since the course of the state of a PLC connection is strongly dependent on human processes, a visualisation on a 24-h basis is appropriate. By such a concentric arrangement, noise can be clearly distinguished from a trend, and changes that extend over several days can be easily recognised.

5.2.3 | Anomaly discovery

Anomaly discovery in SNR spectra is one of the most challenging applications in this field. Even when an anomaly
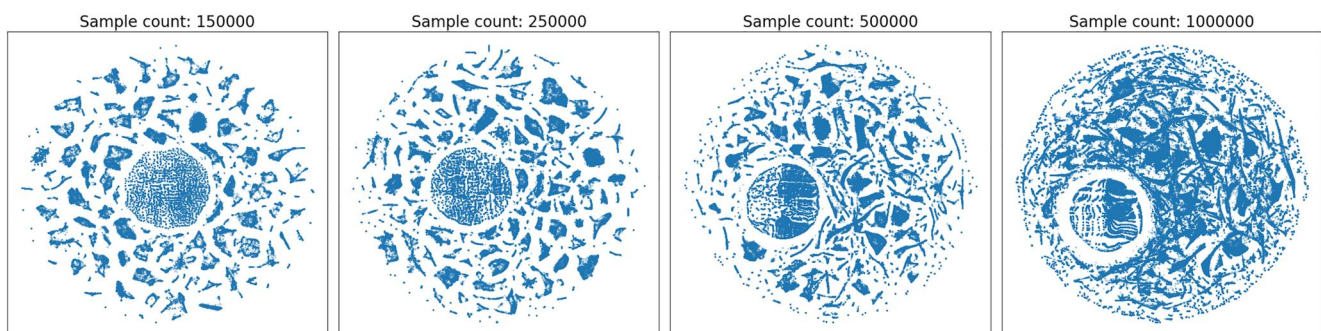

FIGURE 9 Visualisations of t-distributed stochastic neighbor embedding (t-SNE) embeddings at different samples sizes from the Fühler-im-Netz (FiN) data set.
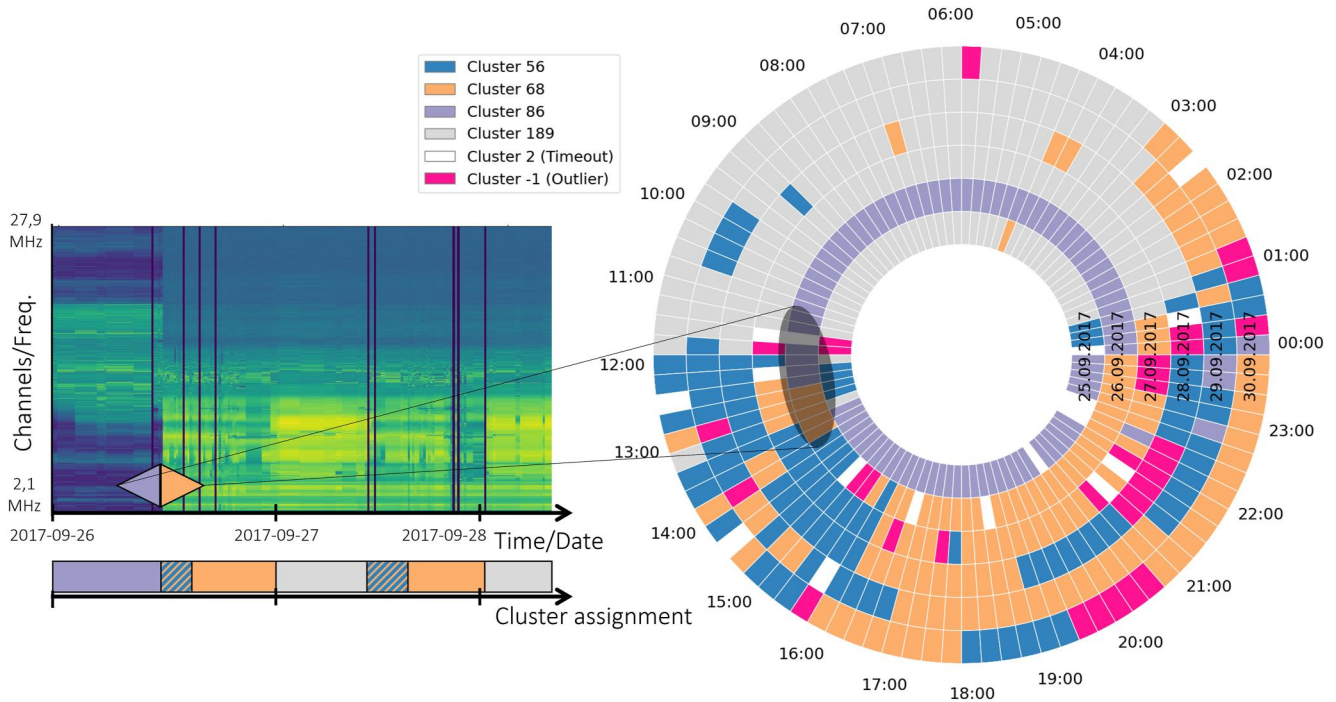
FIGURE 10 A excerpt from the signal-to-noise ratio (SNR) spectrum (left) and the corresponding concentric visualisation of cluster assignments (right). The time range shows an imitated fuse failure. Two triangles show the SNR spectrum before (purple) and after (orange) uninstalling a coupling capacitor used to simulate a fuse failure. The corresponding area in the concentric diagram is highlighted as well. Overall, the concentric arrangement of the diagram provides good access to insights about the data. At the same time, it is possible to distinguish at first glance between random timeouts and time-dependent trends.

is reliably discovered, the cause of the anomaly remains unknown, as it is usually very costly to find out. Therefore, because the origins are largely unknown and the anomalies usually have no direct impact on grid operations, it is not possible to determine anomalies of interest. In addition, a vast number of PLC connections show significant fluctuations that at first glance appear to be an anomaly. It exists a wide range of influences, which act as anomalies in the SNR spectrum, but have no influence on the electricity grid or the PLC. On the other hand, the anomalies carry a wide range of potential information. Figure 10 shows a fuse failure which could be within the SNR spectrum as well as in the circular visualisation. Since, connection states are described by discrete cluster labels we suggest approaches like Hidden Markov Models [37], Dynamic Time Warping [38] or Temporal Convolutional Networks [39] to detect anomalies. To provide a first glance at this topic, Figure 11 shows a transition graph representing the probability that a given PLC connection switches between two states at a given time. This graph could be used directly to calculate a certainty for the current state of being an anomaly or to estimate the same based on the last hours. Beyond that, nevertheless, the use of more sophisticated probabilistic approaches is recommended. However, this aspect is subject to further research and is briefly discussed in Section 8.
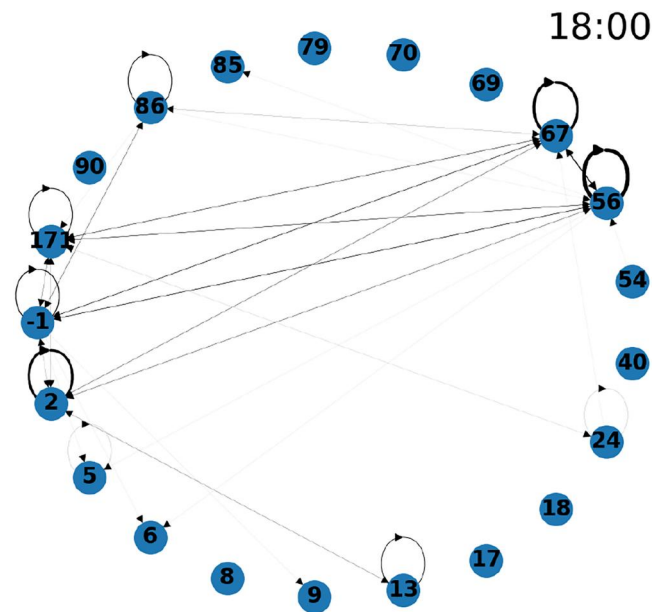

FIGURE 11 Transition graph at 6 am. Based on the past sequence of visited states, the transition graph illustrates the chance to switch from one state to another. As can be well illustrated in the Figure 10, the probability of being in a certain state changes over the course of a day. The strength of a connection indicates the probability of the specified transition.

5.3 | Security

We also want to shed light on the idea of using a PLC-based infrastructure as additional security layer, to monitor the electricity grid. Due to their distributed, independent and flexible nature, PLC systems are well suited to act as a redundant control mechanism to check against traditional grid monitoring. Since all nodes act independently, the network is hard to compromise and even in the event of single failures, the network is able to adapt if alternative routes are available. Furthermore, since each cable connection is measured from both sides (transmitting and receiving), adjacent nodes can check a signal for plausibility. Because of these properties, PLC-based monitoring systems might be a useful tool for the detection of attacks against critical electricity infrastructure.

6 | CONCLUSION

In this paper, we proposed and analysed a new real data set collected during practical operation of the power system. Our work emphasises the high value of the data in a variety of potential application domains, as well as the richness of its nature, consisting of a mixture of structured, dynamic and static information. The SNR data presented here open up applications that go beyond what previous data sets based on consumption and voltage values allow. The applications shown are intended to give a first impression of the wide range of possible applications of the data set. Furthermore, we were able to successfully show how the SNR data can be used to estimate the joint count within a cable section. The key advantage of our approach is that no additional knowledge about the infrastructure (cable length, type etc.) is necessary. Only the SNR spectrum was used for this task. We believe that further research on this topic will yield a variety of new research challenges as well as important developments in PLC-based monitoring of smart grids. The final Section 8 mentions further potential application possibilities which will be pursued in the future.

7 | FUTURE WORK

The digitalisation of electricity grids is currently picking up speed and is coupled with significant opportunities, risks and challenges that still need to be overcome. Possibilities and use cases that we have addressed so far can only shed light on a small aspect.

As mentioned earlier, the FiN data set represents a novel data set that can be used to address various challenges arising from the transition to a smart grid. Therefore, we set out to take a first glance at what is hidden in this previously widely unrecognised SNR data from PLC networks. However, the FiN data sets are only a first step in this area and represent a proof of concept in several directions. To advance this new area of SNR-based network assessment, we intend to address several open questions in our future research:

- PV: Ref. [41] suggests that noise from PV inverters can be detected within PLC networks, consequently it could be possible to find a correlation between noise in SNR and weather. Future work should further investigate these correlations and prove them with practical examples. In addition, an automatic detection of new PV systems connected to the grid will also be evaluated.
- Predictive maintenance: Encouraged by our results, we conclude that the SNR spectrum contains rich information about a wide range of cable properties. Challenges to be investigated in the future are, for example, to what extent artificial intelligence methods are suitable for predicting the remaining lifetime of cable installations or the extent to which a grid state identification can be implemented.
- Cable joints: During the collection of the data, it was noticed that a significant proportion of failures result in particular from failing joints. Further analysis of the joints and their PLC spectrum is therefore subject of further research.
- Topology estimation: As discussed in Section 3.2, PLC cannot provide information on the type of neighbourhood, whether an adjacent node is a direct or indirect neighbour. Therefore, we propose the SNR spectrum as a carrier for fingerprints of specific cable sections to determine whether two neighbours are directly or indirectly adjacent. Future research will address AI-based automatic detection of PLC network topology.
- Privacy and cyber-security: The risks associated with any digitalisation are software bugs and vulnerability to attacks. For this purpose, findings from measurements of the PLC infrastructure could also be used to plausibilise measurement data and network states of the legacy grid control systems on a second, independent level. In this regard, questions also arise concerning the privacy of customers who are forced to connect to such a system during a PLC rollout. Even if full-scale expansion is still a long way off, we would like to encourage discussion of these issues.
- Additional measurements: Since PLC networks provide a good platform to collect various measurements from the nodes, future research will address the extent to which other measurements, such as voltage, can contribute to more comprehensive predictions regarding the application areas presented.
- Robustness: Due to the comparatively small spatial extent of the area, we want to explore the possibility of transferring the models and findings obtained in FiN to other grid areas. These include, above all, the future research topics already presented, as well as the approaches shown in this work.

To address all the future work mentioned above, a second data set is currently being created that uses the previous results as a proof-of-concept and, in particular, increases the number of nodes involved to several thousand. The massively larger resulting data set is expected to yield more robust results in all directions.

AUTHOR CONTRIBUTIONS

Christoph Balada: Conceptualisation; Methodology; Software; Validation; Visualisation; Writing – original draft; Writing – review & editing. **Max Bondorf**: Writing – review & editing. **Sheraz Ahmed**: Supervision. **Nikolai Hopfer**: Supervision. **Andreas Dengel**: Supervision. **Markus Zdrallek**: Supervision.

ACKNOWLEDGEMENT

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at http://dx.doi.org/10.5281/zenodo.5948717.

ORCID

Christoph Balada https://orcid.org/0000-0003-0307-7866

REFERENCES

1. Miller, M., et al.: Status report on power system transformation: a 21st century power partnership report. In: Tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States) (2015)
2. Sendin, A., Peña, I., Angueira, P.: Strategies for power line communications smart metering network deployment. Energies 7(4), 2377–2420 (2014). https://doi.org/10.3390/en7042377
3. Dedé, A., et al.: Smart meters as part of a sensor network for monitoring the low voltage grid. In: 2015 IEEE Sensors Applications Symposium (SAS), pp. 1–6. IEEE (2015)
4. Rolnick, D., et al.: Tackling Climate Change with Machine Learning (2019). arXiv preprint arXiv:1906.05433
5. Hippert, H.S., Pedreira, C.E., Souza, R.C.: Neural networks for short-term load forecasting: a review and evaluation. IEEE Trans. Power Syst. 16(1), 44–55 (2001). https://doi.org/10.1109/59.910780
6. Soliman, S.A.-h., Al-Kandari, A.M.: Electrical Load Forecasting: Modeling and Model Construction. Elsevier (2010)
7. Hong, T., Fan, S.: Probabilistic electric load forecasting: a tutorial review. Int. J. Forecast. 32(3), 914–938 (2016). https://doi.org/10.1016/j.ijforecast.2015.11.011
8. Alzahrani, A., et al.: Solar irradiance forecasting using deep neural networks. Procedia Comput. Sci. 114, 304–313 (2017). https://doi.org/10.1016/j.procs.2017.09.045
9. Sun, Y., Szűcs, G., Brandt, A.R.: Solar pv output prediction from video streams using convolutional neural networks. Energy Environ. Sci. 11(7), 1811–1818 (2018). https://doi.org/10.1039/c7ee03420b
10. Wan, C., et al.: Photovoltaic and solar power forecasting for smart grid energy management. CSEE J. Power Energy Syst. 1(4), 38–46 (2015). https://doi.org/10.17775/cseejpes.2015.00046
11. Mathe, J., et al.: A Lrcn Architecture for Spatio-Temporal Photovoltaic Powerforecasting from Numerical Weather Prediction (2019). arXiv preprint arXiv:1902.01453

12. Candanedo, L.M., Feldheim, V., Deramaix, D.: Data driven prediction models of energy use of appliances in a low-energy house. Energy Build. 140, 81–97 (2017). https://doi.org/10.1016/j.enbuild.2017.01.083
13. United Nations Statistics Division: UN Energy Statistics Database (2021). URL http://data.un.org/Explorer.aspx
14. PJM Interconnection: PJM Energy Market (2021). URL https://www.pjm.com/markets-and-operations/energy
15. Tsanas, A., Xifara, A.: Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build. 49, 560–567 (2012). https://doi.org/10.1016/j.enbuild.2012.03.003
16. U.S. Energy Information Administration: Eia Open Data Access (2021). URL https://www.eia.gov/opendata/bulkfiles.php
17. Hopfer, N., et al.: New approach for medium voltage power cable assessment using broadband powerline communications. In: The 20th International Symposium on High Voltage Engineering (ISH 2017), Buenos Aires (2017)
18. Hopfer, N., et al.: Identification of power cable conditions based on broadband powerline communication characteristics in distribution grids. In: Stará Lesná (2017)
19. Rezaei, H., et al.: A new approach for online monitoring of power cable condition in distribution networks using broadband powerline communication and simple sensor solutions. In: 11th Conference on Electricity Distribution of Serbia (CIRED SERBIA 2018) (2018)
20. Förstel, L., Lampe, L.: Grid Diagnostics: Monitoring Cable Aging Using Power Line Transmission (2017)
21. Bondorf, M., et al.: Broadband power line communication and big-data-analytics for supporting grid operation. In: ETG Congress 2021, pp. 1–6. VDE (2021)
22. He, Z., et al.: Power-grid controller anomaly detection with enhanced temporal deep learning. In: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 160–167. IEEE (2019)
23. Jin, C., et al.: An efficient heterogeneous signcryption for smart grid. PLoS One 13(12), e0208311 (2018). https://doi.org/10.1371/journal.pone.0208311
24. Schäfer, B., et al.: Taming instabilities in power grid networks by decentralized control. Eur. Phys. J. Spec. Top. 225(3), 569–582 (2016). https://doi.org/10.1140/epjst/e2015-50136-y
25. Nord, S.: Prof. dr.-ing. habil. lutz hofmann and prof. dr. rer. nat. habil. michael sonnenschein (2015)
26. Arzamasov, V., Böhm, K., Jochem, P.: Towards concise models of grid stability. In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE (2018)
27. Kamps, K., et al.: Reliability of decentralized network automation systems and impacts on distribution network reliability. In: 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), pp. 1–6. IEEE (2020)
28. Ludwig, M., et al.: Multi-agent-based Grid Automation: Field Test Experiences of the Distributed Grid State Control (2019)
29. Openweather Ltd: OpenWeather History Bulk Download Documentation (2021). URL https://openweathermap.org/history-bulk/
30. Selvaraju, R.R., et al.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
31. Huo, Y., et al.: Power Line Communication Based Smart Grid Asset Monitoring Using Time Series Forecasting (2021). arXiv preprint arXiv:2110.10219
32. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. 9(11), 33–55 (2008). https://doi.org/10.1007/s10994-011-5273-4
33. McInnes, L., Healy, J., Melville, J.: Umap: Uniform Manifold Approximation and Projection for Dimension Reduction (2018). arXiv preprint arXiv:1802.03426

34. Kruskal, J.B.: Nonmetric multidimensional scaling: a numerical method. Psychometrika 29(2), 115–129 (1964). https://doi.org/10.1007/bf02289694

35. Tian, K., et al.: A general clustering framework based on deep learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 809–825. Springer (2017)

36. Yan, X., et al.: Clusterfit: improving generalization of visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6509–6518 (2020)

37. Eddy, S.R.: What is a hidden Markov model? Nat. Biotechnol. 22(10), 1315–1316 (2004). https://doi.org/10.1038/nbt1004-1315

38. Müller, M.: Dynamic Time Warping, Information Retrieval for Music and Motion, pp. 69–84 (2007)

39. Oord, A.v.d., et al.: Wavenet: A Generative Model for Raw Audio (2016). arXiv preprint arXiv:1609.03499

40. Balada, C.: Fin: a smart grid and powerline communication dataset, (2022). https://doi.org/10.5281/zenodo.5948717

41. López, G., et al.: Noise sources, effects and countermeasures in narrowband power-line communications networks: a practical approach. Energies 10(8), 1238 (2017). https://doi.org/10.3390/en10081238