Information Sciences 662 (2024) 120211

Contents lists available at ScienceDirect

Information Sciences

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

Energy disaggregation risk resilience through microaggregation and discrete Fourier transform

Kayode S. Adewole [b,a,*], Vicenç Torra [c]

[a] Department of Computer Science and Media Technology, Malmö University, Sweden
[b] Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[c] Department of Computing Science, Umea University, Sweden

ARTICLE INFO

Keywords:
Smart meters
Smart grid
Disclosure risk
Non-intrusive load monitoring
Data privacy
Microaggregation
Discrete Fourier transform

ABSTRACT

Progress in the field of Non-Intrusive Load Monitoring (NILM) has been attributed to the rise in the application of artificial intelligence. Nevertheless, the ability of energy disaggregation algorithms to disaggregate different appliance signatures from aggregated smart grid data poses some privacy issues. This paper introduces a new notion of disclosure risk termed energy disaggregation risk. The performance of Sequence-to-Sequence (Seq2Seq) NILM deep learning algorithm along with three activation extraction methods are studied using two publicly available datasets. To understand the extent of disclosure, we study three inference attacks on aggregated data. The results show that Variance Sensitive Thresholding (VST) event detection method outperformed the other two methods in revealing households' lifestyles based on the signature of the appliances. To reduce energy disaggregation risk, we investigate the performance of two privacy-preserving mechanisms based on microaggregation and Discrete Fourier Transform (DFT). Empirically, for the first scenario of inference attack on UK-DALE, VST produces disaggregation risks of 99%, 100%, 89% and 99% for fridge, dish washer, microwave, and kettle respectively. For washing machine, Activation Time Extraction (ATE) method produces a disaggregation risk of 87%. We obtain similar results for other inference attack scenarios and the risk reduces using the two privacy-protection mechanisms.

1. Introduction

The realization and sustainability of smart cities have witnessed immense growth over the last few years due to the advancement in AI, smart meters, Internet-of-things and smart grid technologies. Immense research efforts to develop technological solutions that tend toward energy conservation, smart grid resource availability, and improving the well-being of societies, in general, have been on the rise [1–3]. Energy conservation addresses effective utilization of energy resources for actualizing self-sustainability in energy management. This can be achieved through energy control and proper monitoring of energy demand for better optimization to minimize load consumption requirement of individual households [4]. A fine-grained monitoring of energy demand targeting household-level consumption will assist in minimizing energy wastage. Significant research efforts have been made toward devel-

* Corresponding author at: Department of Computer Science and Media Technology, Malmö University, Sweden.
E-mail addresses: kayode.adewole@mau.se, adewole.ks@unilorin.edu.ng (K.S. Adewole), vtorra@cs.umu.se (V. Torra).

https://doi.org/10.1016/j.ins.2024.120211
Received 8 April 2023; Received in revised form 27 November 2023; Accepted 22 January 2024

Available online 26 January 2024
0020-0255/Â© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

oping methodologies for effective energy demand monitoring and control [2]. Therefore, providing consumers with prior insights regarding their fine-grained energy consumption will reduce the heavy burden on smart grid resources and this can eventually reduce energy wastage. Non-intrusive load monitoring (NILM) or energy disaggregation is one of the methodologies that aims to achieve this objective. Energy disaggregation is defined as the task of separating household load consumption recorded at the aggregate level into the constituting loads of the appliances that are used in the household [5].

Two approaches to achieving energy disaggregation are Intrusive Load Monitoring (ILM) and NILM [6]. ILM involves the use of low-end smart meter to measure energy consumption of one or more appliances with one sensor device per appliance. Conversely, NILM involves the use of a single smart meter to monitor energy consumption of a single household or building. NILM offers several advantages over ILM as it reduces the cost of maintaining multiple smart meters in a building. NILM guarantees real-time feedback per appliance energy demand by disaggregating and analyzing the aggregated load consumption recorded by the mains smart meter attached to the building. The benefits of NILM include the provision of real-time feedback to consumers, fault detection, anomaly detection, appliance activation events detection, and encouraging energy-saving behaviors [7,8].

There are two major approaches to developing NILM solutions in the research community, which are classification and regression. The two approaches focus on developing appliance-level event activation (ON/OFF) as well as predicting the load of individual appliances in the aggregated signals. NILM classification system can infer whether a given appliance is in operation during the day with the help of classification algorithms. Different methods for extracting appliance activations based on appliance-level energy consumption data have been investigated. For example, Laviron et al. [9] studied three activation extraction methods, which are ValmA, SimBA and Cartesio for extracting individual appliance states from appliance-level data. Desai et al. [10] studied Variance-Sensitive Thresholding (VST) aiming to extend Middle-Point Thresholding (MPT) approach [7]. Kelly et al. [11] presented activation time extraction (ATE) method that was tuned using UK-DALE NILM dataset. Activation event data extracted using any activation extraction method serve as useful information for developing machine learning models. On the other hand, the regression model takes aggregated household energy consumption data to produce the different appliance load signatures [5,11,12]. Classification and regression models have produced promising results in NILM domain to develop solutions that monitor energy consumption per appliance.

Nevertheless, the capability of NILM algorithm to disaggregate individual appliance loads data present in the aggregated signals has raised a serious privacy concern. Fine-grained electricity consumption data are privacy-sensitive as they are capable of revealing consumers' households lifestyles based on their energy consumption patterns [13,14]. Additionally, a majority of the publicly available smart grid datasets available in NILM research domain have their associated meta-data that can provide background information regarding the data collection processes. This information serves as external background knowledge for attackers to explore and coupled with the inferences made from NILM models to reveal household identities. This background information is useful to third parties, such as criminals, marketers and law enforcement agents. For example, critical cyber-attack cases on smart grid infrastructure leading to the disruption of energy systems, which supply heat and light to many households in Ukraine have been reported in 2015, 2016 and 2017 [15,16]. Therefore, hiding the individual appliance signatures in the masked aggregated data is safer before releasing smart grid data to the public. This enforces a new privacy requirement for protecting smart grid data which is a major challenge to existing privacy protection mechanisms as none of these approaches investigates this new disclosure risk. Although several privacy protection mechanisms for smart grid data exist, which include data anonymization [15,17–21] and differential privacy [19,22]. For instance, [18] studied the performance of K-ward microaggregation to protect building occupancy and smart grid data. The authors focused on establishing a method to learn utility-specific applications that are of concern to data user. [20] also investigated the performance of microaggregation algorithm and discussed advantages and disadvantages of other privacy protection methods such as data permutation, time slicing, random noise, data transformation, scope aggregation and differential privacy. In [19] the performance of Long Short-term Memory (LSTM) with differential privacy for energy forecasting has been studied. Nevertheless, each of the existing studies pays less attention to investigating energy disaggregation risk and to hiding different appliance signatures in the aggregated masked data.

Therefore, in this paper, a first attempt is made to investigate the level of disclosure risk relating to the different appliance signatures in smart grid data that are published in their original form. To achieve this aim, machine learning models based on Seq2Seq NILM deep learning algorithm are developed to ascertain the ability of NILM algorithm in detecting the signatures of individual appliances that constitute the aggregated data. These predictive loads are subjected to VST, MPT and ATE activation extraction techniques to ascertain the ability of Seq2Seq NILM algorithm in predicting the correct activation state of each appliance from their load signatures. Thereafter, we compute the disclosure risk probabilities for the individual appliances that are present in the aggregated signals based on the three attack scenarios we investigated. This enables us to concertize a new disclosure risk metric called *disaggregation risk*. To prevent energy disaggregation risk, we propose two privacy protection mechanisms based on microaggregation and DFT. These two algorithms have been recently studied in our work [23]. We evaluate the proposed approach based on two datasets for energy disaggregation in NILM domain. These datasets are publicly available. More importantly, this paper contributes in the following ways:

- we introduce a new notion of disclosure risk called energy disaggregation risk and study how it can be empirically computed.
- we investigate the performance of Seq2Seq NILM algorithm and three event detection methods in revealing households' lifestyles based on the signature of appliances used. We consider two publicly available datasets (UK-DALE and REFIT).
- we simulate three inference attack scenarios to better understand the level of privacy violation on the individual households' lifestyles.
- we investigate the performance of two privacy-protection mechanisms to reduce energy disaggregation risk.

Table 1
List of abbreviation/acronym.

Abbreviation/Acronym	Definition
AI	Artificial Intelligence
ATE	Activation Time Extraction
BLH	Battery-based Load Hiding
CNN	Convolutional Neural Networks
CO	Combinatorial Optimization
dAE	Denoising Autoencoders
DFT	Discrete Fourier Transform
DP	Differential Privacy
DR	Disaggregation Risk
FMDAV	Fast Maximum Distance to Average Vector
FHMM	Factorial Hidden Markov Model
GAN	Generative Adversarial Networks
GMDAV	Grey Maximum Distance to Average Vector
HF	High Frequency
HMM	Hidden Markov Model
ILM	Intrusive Load Monitoring
LSTM	Long Short-term Memory
MAE	Mean Absolute Error
MDAV	Maximum Distance to Average Vector
MPT	Middle Point Thresholding
MSE	Mean Absolute Error
NILM	Non-Intrusive Load Monitoring
RMSE	Root Mean Squared Error
SAE	Signal Aggregate Error
Seq2Seq	Sequence-to-Sequence
VST	Variance Sensitive Thresholding

The remaining parts of this paper are organized as follows: Section 2 summarizes related works in NILM energy disaggregation as well as privacy preservation methods developed specifically for smart grid data. Section 3 presents the proposed methodology to investigate disclosure risk associated with energy disaggregation as well as the propose privacy-preserving mechanisms to reduce this risk. Section 4 focuses on the experimental settings adopted for the three inference attack scenarios investigated in this study as well as experimental setup for the privacy-preserving approaches. Section 5 summarizes the results obtained from the different experiments conducted and finally, Section 6 concludes the paper. Table 1 provides the list of acronyms used in our paper with their descriptions.

2. Related work

Research trend in NILM energy disaggregation as well as privacy-preserving smart grid data has recently witnessed rapid growth due to the availability of many AI technologies and machine learning approaches. This section summarizes the research development in the two domains in relation to smart grid data.

2.1. Non-intrusive load monitoring

The domain of Non-Intrusive Load Monitoring (NILM) has been in existence since the late 80's due to the noticeable research of [24]. NILM focuses on monitoring energy consumed at the appliance-level by disaggregating individual appliance loads from aggregated data. NILM is often referred to as a single point of measurement since one smart meter is used to monitor the energy consumption of the entire household [6]. This is necessary because it is almost impossible to sub-meter every appliance in the household or building. NILM provides demand-response service and real-time feedback to consumers based on their load consumption. In addition, NILM can predict the state of individual appliances from aggregated signals. Hart et al. [24] present a load identification and energy disaggregation method that is based on clustering analysis. This method uses steady-state features extracted from aggregated load consumption. The extracted features were compared with the appliance-level data during the training phase to predict the signatures of appliances present in the aggregated signal. The author evaluated the proposed approach using steady-state signals only and focused on identifying simple appliances with finite states (on/off) achieving an accuracy of 85%. Apart from the traditional features extraction techniques that focus on steady and transient state features, non-traditional features such as time of the day, peak time usage, light sensing and temperature, appliance usage frequency, eigenvalues of current signal and many more have been investigated for load disaggregation task [2]. Gopinath et al. [2] presented a comprehensive survey of the state-of-the-art techniques for energy management based on NILM. In their work, the authors categorized the features used by the existing traditional NILM techniques into steady-state, transient state, and non-traditional features, and then presented different features under each category. To further show the research trends in NILM domain, techniques that focused on deep learning approaches were also presented. Similarly, [6] also presented a review paper that focused on the research trend in NILM.

The recent advancement in deep learning domain has provided improvement in relying on automated feature extraction methods for energy disaggregation tasks rather than focusing on hand-engineering feature extraction methods from the aggregated power

data [11,6]. These features provide comparable and promising results over the algorithms that rely on hand-engineering feature extraction. For instance, Kelly et al. [11] adapted three deep learning architectures to NILM research, which are based on LSTM, Denoising Autoencoders (dAE) and Rectangle architecture for extracting individual appliance consumption loads from aggregated data. The target appliance signal was reconstructed based on dAE while treating the aggregated consumption as noisy input. The results obtained have shown the significance of deep learning algorithms on load disaggregation for the five appliances considered over the existing NILM approaches like Factorial Hidden Markov Model (FHMM) and Combinatorial Optimization (CO). Particularly, dAE and Rectangle architectures were reported to perform well when disaggregating unseen houses.

Wang et al. [25] also developed deep learning models based on LSTM and dAE using five selected appliances. The authors established that the proposed deep learning models outperformed Hidden Markov Model (HMM). Empirically, F1-score of 0.53, 0.985, 0.53, 0.746 and 0.382 were achieved for Heat pump, Cloth dryer, Washing machine, Dish washer, and Fridge respectively. To address the problem of designing effective sliding windows to handle long sequences of power signals and combine predictions from different sliding windows, [12] studied sequence-to-point (Seq2Point) and Seq2Seq deep learning algorithms where the input is a window of the mains meter and the output represents the consumption of the target appliance. The proposed Seq2Point algorithm reduces Mean Absolute Error (MAE) and Signal Aggregate Error (SAE) by 84% and 92% respectively when compared with the approach in [26]. [27] adopted CNN and Seq2Seq algorithms for load consumption optimization. In their study, CNN and Seq2Seq model produced better results when compared with CNN and LSTM. While their work also focused on the adoption of Seq2Seq, privacy-preservation of household consumption was not investigated. Similarly, [28] proposed a lightweight Seq2Seq algorithm that can be implemented on the edge devices for constrained-based equipments. The model provides real-time appliance load monitoring usage and optimization since the data are being processed on the edge devices. In comparison to the cloud computing, edge computing provides better privacy since the computation is done on the edge. While this study provides interesting concepts, our goal in this research focuses on privacy-preserving data publishing where utility company is interested in publishing energy consumption without violating the privacy of the individual households. In [29], a multiscale residual network based on Seq2Seq model has been studied. The authors proposed a new CNN model that can target the extraction of more features for appliance load disaggregation. It was observed that this architecture provides better F1-score and MAE for appliance load monitoring.

Dash and Shaoo [30] introduced a multitask deep learning model that is evaluated on UK-DALE and REFIT datasets. The model employed low-frequency energy data from the two datasets for simultaneous appliance state detection and energy disaggregation. The authors claimed that the model achieved superior performance and provided generalizability and transferability properties when compared with the state-of-the-art models. The proposed model achieved accuracy of 96.27%, 96.97%, 94.70%, 63.78% and 99.20% respectively for microwave, dish washer, washing machine, fridge and kettle. One of the research problems in NILM domain is how to find high-quality labeled samples to develop the disaggregation models. To solve this problem, [31] proposed a transfer learning framework that is based on active learning. The framework can improve the performance of NILM deep learning model as it learns from a small amount of data in the new environment. Using REFIT dataset, the framework achieved accuracy-labeling trade-off with only 5 to 15% of the query pool labeled, indicating that the labeling effort could be reduced by as much as 85% as reported in the study.

A number of studies have addressed the big data challenges in smart grid. For instance, [32] proposed a framework that is based on cloud computing and fog computing to manage smart grid data among different agents in the cloud. The authors proposed a hybrid gray wolf differential evolution optimization algorithm that combined gray wolf optimization and improved differential evolution. Results show that the proposed hybrid approach is 54 ms, 82.1 ms, and 81.6 ms faster than particle swarm optimization, differential evolution, and gray wolf optimization. The hybrid model also achieved processing time is 53 ms, 81.2 ms, and 80.6 ms faster than the three models respectively. A relaxed consensus plus innovation-based negotiation technique to foster energy cooperation between smart grid and microgrid has been proposed in [33]. The paper analyzed the effect of uncertainty parameters within the system on the effectiveness of the proposed negotiation approach.

Although there have been significant progress in NILM and smart grid research domain, our paper explores a different research dimension. More specifically, we investigate the use of NILM as a tool to perform inference attacks on smart grid data. To achieve this, we propose a new disclosure risk measure called energy disaggregation risk. We empirically quantify this disclosure risk measure, and investigate how this risk can be reduced to prevent privacy attacks on individual households' energy consumption.

2.2. Smart grid privacy preserving data publishing

Research efforts have been intensified to provide mechanisms for privacy preservation of smart grid data publishing. The techniques used in privacy preservation of smart grid data publishing include data anonymization with Battery-based Load Hiding (BLH) [17], k-anonymity [34,18,20,23], Generative Adversarial Network (GAN) with some correlated noise [35,36], and differential privacy (DP) [19,22] just to mention a few.

BLH research aims to provide a rechargeable battery situated at the consumer end and capable of being charged and discharged. This feature makes the smart meter incapable of accurately collecting the real energy consumed by the appliances. The privacy guarantee of BLH masking approach is yet to be empirically validated as this masking method is mainly theoretic. Investigating the real-world applications of BLH with a focus on efficient privacy guarantees for smart grid data requires further analysis [17].

K-anonymity was originally proposed in [37] and provided improved privacy guarantees for the protected data. While k-anonymity is not a privacy mechanism on its own, it is a condition that enforces group-based anonymization on the protected data. The goal of k-anonymity is to ensure that each individual in protected data cannot be identified within a set of k individuals. This means that the original data is partitioned into a set of at least k indistinguishable records. Different methods exist in the liter-

ature that satisfy k-anonymity condition. One of such method is microaggregation [38], which builds small microclusters and then replaces the original data in each cluster with the cluster representative.

K-ward microaggregation algorithm was proposed in [18] for protecting building occupancy and smart grid data. The k-ward algorithm leverages agglomerative clustering to generate k-partition which clusters the data into group sizes of at least k records. A substitution step then perturbs the data by replacing the true values using the group centroid. The goal of the authors is to extend their work in [34] to provide a nonlinear feature representation mapping for their data publishing system. The authors claimed that the proposed privacy-preserving system can achieve better utility under reasonable protection. However, the main limitation of this approach is scalability as data users will have to manually define the similarity of their data points based on the specific data utility application. [20] also investigated the performance of microaggregation algorithm and discussed the advantages and disadvantages of other privacy protection methods such as data permutation, time slicing, random noise, data transformation, scope aggregation and differential privacy. The authors established that no single data anonymization technique that can fit all data utility applications. They show that there is a significant data utility loss between 4- and 8-anonymization approaches being used to evaluate the microaggregation algorithm in their study. Authors in [39] and [40] focused on extending MDAV microaggregation algorithm. For instance, [39] proposed Grey Maximum Distance to Average Vector (GMDAV) which target how to consider the importance of each quasi-identifiers. To achieve this, authors proposed weighted Euclidean distance based method for gray relational analysis and another metric for information loss model. Results show that GMDAV achieved better complexity. Similarly, authors in [40] also proposed Fast Maximum Distance to Average Vector (FMDAV) to address big data challenge. Adewole and Torra [23] investigated the performance of microaggregation and DFT for protecting daily energy consumption data. In their work, the authors establish adversarial scenarios that involve attackers who are interested in launching interval disclosure risk and distance-based record linkage on energy data. They proposed microaggregation algorithms to reduce these disclosure risks. The utility of the protected energy data was established using four approaches, which are based on information loss metric, classification, clustering, and time series forecasting methods. The results show that their proposed methods can achieve accuracy, Silhouette score, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) that are close to the values obtained for the original data to be published.

The ability to model uncertainties of the original data is one of the benefits of exploring Generative Adversarial Network (GAN). This model can then be used to generate new data. The goal involves training two deep neural networks. The first neural network (i.e. Generator) is trained to produce more realistic data that resemble the original data and the second neural network (i.e. Discriminator) is trained to estimate the probability that the input originates from the real data. Based on this assumption, a novel solution for smart grid data generation is offered. However, the capability of GAN to prevent disclosure risk attacks such as membership inference is still an open research problem [41]. A de-facto standard for privacy-preserving mechanism called Differential privacy (DP), guarantees $\epsilon-DP$ for each record that is present in the protected data. For instance, [19,22] studied DP algorithms for protecting smart grid data.

In [42], the problem of smart grid outages that may be associated with cyber-attack is extensively studied. In this work, the authors proposed a two-layer framework that determines network vulnerability points due to physical faults and cyber-attacks. The first layer of the framework proposed blockchain technique while the second layer studied reinforcement learning method. These layers collaboratively work together to determine the vulnerable points through monitoring of the smart grid data. The finding from this study shows a significant reduction in the network vulnerability indices owing to the cooperating microgrid. The study also determined the severity of the threat level of attacks that may be linked to the power outages in smart grid.

Nevertheless, existing studies on privacy-preserving smart grid data publishing have paid less attention to investigating the disclosure risk concerned with disaggregating different appliance load signatures from aggregated power data and providing privacy mechanisms for reducing this disclosure risk. Therefore, it is worth investigating inference attacks on smart grid data based on energy disaggregation risk as well as how this inference attack can be minimized. This is the major contribution of our paper. We established a new disclosure risk and empirically show how it can be used as a measure of statistical disclosure that is based on inferencing. We further investigate how this disclosure risk can be reduced using two privacy-preserving mechanisms.

3. Proposed approach

In this section, we present disaggregation risk as defined in our previous work [43] and discuss the proposed mechanisms to reduce energy disaggregation risk in smart grid domain. Fig. 1 shows the components of the proposed framework. The first part of the framework presents the methodology used to investigate the level of disclosure risk that is associated with disaggregating individual appliance load signatures from aggregated signals. The second part of the framework shows how the proposed privacy-preserving mechanisms have been applied to reduce energy disaggregation risk.

3.1. Disaggregation risk

Energy disaggregation risk is the ability of an attacker to reveal household consumption behavior by performing an inference attack on household aggregate consumption based on the signatures of the appliances that they use in the household. We provide a formal definition as follows.

Definition 1. Disaggregation risk [43] is the probability of predicting the load signature of appliance (ℓ) and its associated ON events from aggregated energy data within the specified time period using NILM algorithm.

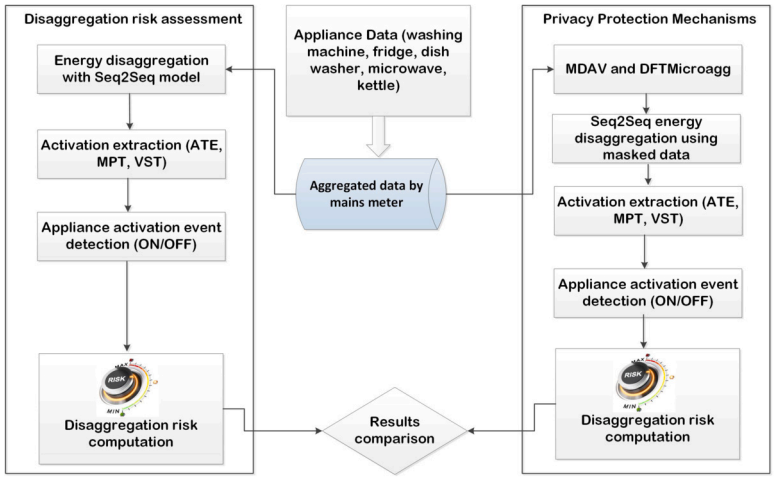


Fig. 1. Proposed framework for reducing energy disaggregation risk.

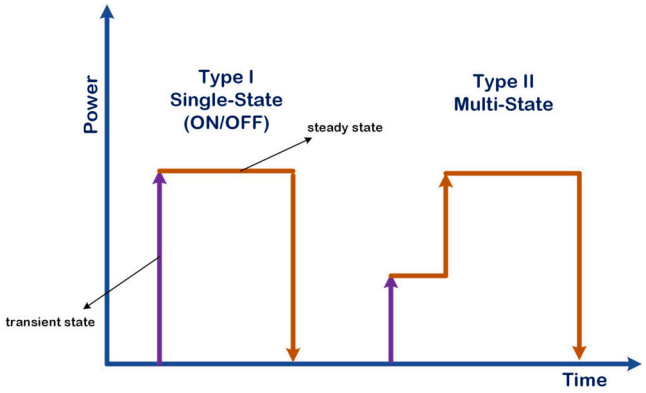


Fig. 2. Operating states of Type I and Type II appliances.

This is formalized based on Eqn. (1).

$$DR^{(\ell)} = TP^{(\ell)}/(TP^{(\ell)} + FN^{(\ell)}) \tag{1}$$

where $TP^{(\ell)}$ represents the number of correct predictions of ON events of appliance ℓ, $FN^{(\ell)}$ represents the number of ON events of appliance ℓ which are wrongly predicted as OFF events. The disaggregation risk ($DR^{(\ell)}$) of appliance ℓ has a value in the interval [0,1] such that the higher this value is, the higher the disclosure risk associated with the appliance.

Therefore, the performance of Seq2Seq deep learning algorithm is assessed based on this disaggregation risk metric.

3.2. Proposed method for energy disaggregation risk assessment

In this section, we discuss our proposed method to investigate the level of disclosure risk that is associated with disaggregating individual appliance loads in the aggregated data. As shown in the first part of Fig. 1, Seq2Seq deep learning model is trained using individual appliance data. This enables us to build a trained model for each appliance signature, which can be used to test the capability of Seq2Seq NILM algorithm for energy disaggregation tasks. The disaggregated load signatures of individual appliances are subjected to MPT, VST and ATE event detection algorithms. We discuss the threshold methods in the subsequent sections. The results from the three activation methods are used to compute the disaggregation risk according to eqn. (1). We adapted Seq2Seq NILM deep learning architecture due to its effectiveness as demonstrated in [5,12].

For appliance selection, we focus on Type I and Type II appliances for the investigative study. Type I appliances have two states of operation (ON/OFF) and examples include kettle, toaster, light bulb, and lamps (see Fig. 2). Appliances in this category consumed energy when they are turned ON. Multi-state or finite state appliances are Type II appliances with a finite number of operating states which are possible to be executed repeatedly. Rising and falling edges of energy consumed within a period are used to detect the appliance transition states. Examples of Type II appliances are dish washers, washing machines, refrigerators and stove burners. Type I and Type II appliances are commonly studied in NILM domain and for this reason, we selected five appliances that belong to these two categories. The selected appliances in this study are dish washer, washing machine, fridge, kettle and microwave. The second

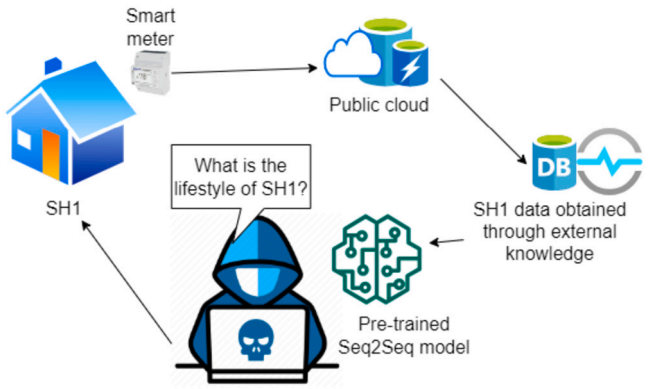


Fig. 3. Inference attack scenario 1. SH represents smart home.

reason for this choice of appliance selection is that these appliances have been utilized in not less than two households from the two datasets investigated in this study. This helps us to simulate three inference attack cases that focus on investigating appliance usage patterns in the same household and across different households.

Typically, an aggregated energy consumption is provided as input to NILM systems to predict the load of individual appliances. The aggregated power consumption P_t generated at time t represents the sum of all appliance loads according to Eqn. (2):

$$P_t = \sum_{\ell=1}^{L} P_t^{(\ell)} + e_t \tag{2}$$

where $P_t^{(\ell)}$ represents the power of appliance ℓ at time t, L represents the total number of appliances in the household or building, and e_t is the residual load. The residual load can be considered as a noise which may originate from the mains electric meter (i.e. smart meter used for the aggregate reading) or at the appliance sub-meter. The residual load (noise) from the mains electric meter is an unstructured noise while the residual load (noise) from the sub-meter represents the structured noise. Because it is impossible to provide sub-meters for all the appliances in the building, appliances that are not sub-metered usually generate this structured noise.

3.2.1. *Inference attack simulations*

Three inference attack scenarios were simulated to provide a better understanding of the disclosure risk associated with disaggregating individual appliance loads from the aggregated signals. The first inference attack is termed *attack on the same building data*. In this inference attack scenario, an attacker has trained a Seq2Seq model that is based on some time-series data of a particular household and then he wishes to use this model to disaggregate an aggregated signal from the same household data. The goal is to reveal if a specific appliance signature is present in the aggregated energy data. The second inference attack scenario is termed *attack on different households selected from the same dataset*. Considering this type of attack, an attacker has a pre-trained Seq2Seq model using the data from one building and then he wishes to use this model to disaggregate an aggregated signal from another household in the same dataset. If the attack is successful, then the attacker has been able to infer useful information regarding the target household lifestyle and subsequently use the background information in his possession for malicious purposes. The third inference attack is termed *attack on different household from different datasets*. With this type of inference attack, an attacker has a trained model based on specific household data in one dataset and then he wishes to disaggregate the aggregated signal from the target household in another dataset. The success of this inference attack can lead to privacy violations in the target household. We present the three inference attack scenarios in Fig. 3, 4, and 5 respectively.

Therefore, for each of the inference attack scenarios, we evaluate the performance of the Seq2Seq energy disaggregation algorithm to ascertain the extent of privacy violation on the individual target household based on appliance usage patterns. We propose two privacy protection mechanisms to reduce these inference attacks. Section 3.3 presents the privacy-preserving methods for energy disaggregation risk reduction.

3.2.2. *Seq2Seq disaggregation algorithm*

There are several NILM algorithms in the literature, however, in this study, we adapted deep learning Seq2Seq NILM algorithm [12,26]. This algorithm uses the architecture of deep neural networks based on different Convolutional Neural Networks (CNNs) layers as presented in Fig. 6. The algorithm takes as input the individual appliance loads and the aggregated signal during the training phase, which are used to learn the signature of the different appliances present in the aggregated data. During the testing phase, only the aggregated signal is provided to the trained Seq2Seq model and the model disaggregate this signal to produce the constituting appliance loads. The performance of the algorithm for energy disaggregation and event detection is then evaluated.

Formally, let F_s represent a deep neural network which takes the input sequence consisting of the sliding windows $Y_{t:t+W-1}$ that correspond to the aggregated power consumption from the mains meter and maps it to the respective windows $X_{t:t+W-1}$ representing the load sequence of the target appliance power. We define a regression equation such that $X_{t:t+W-1} = F_s(Y_{t:t+W-1}, \theta_s) + \epsilon$, where ϵ represents W-dimensional Gaussian random noise, and the parameters of the deep neural network F_s are represented as θ_s.

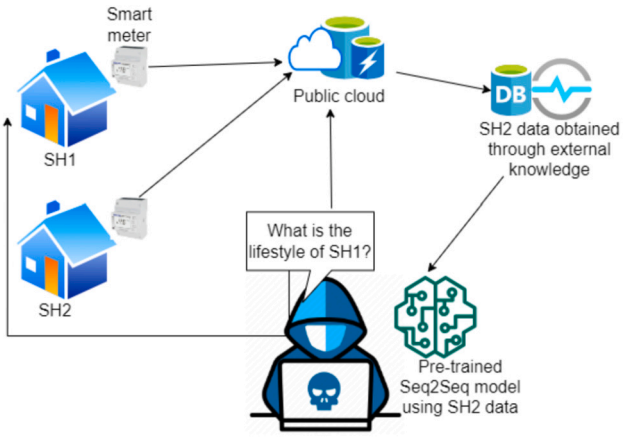


Fig. 4. Inference attack scenario 2. SH represents smart home.

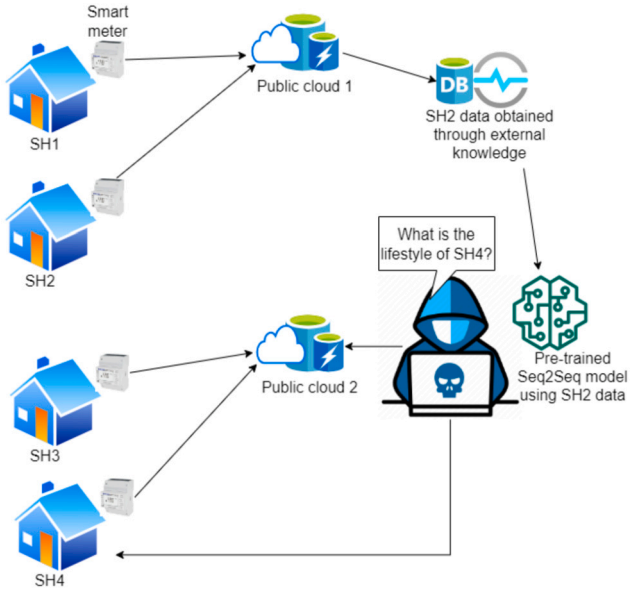


Fig. 5. Inference attack scenario 3. SH represents smart home.

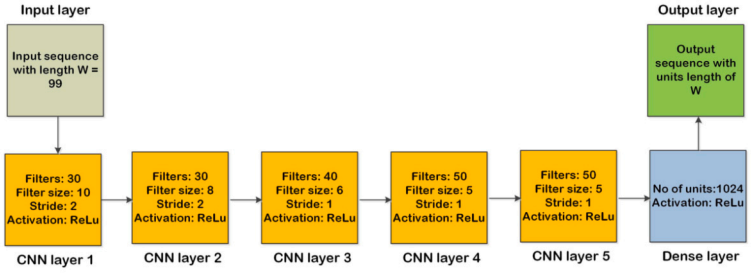


Fig. 6. Architecture of the adapted Seq2Seq NILM algorithm.

The full architecture of Seq2Seq algorithm is depicted in Fig. 6 and the training procedure is presented in Algorithm 1. For the sake of clarity, we added a Dropout layer with probability 0.2 between CNN layer 4 and 5. This Dropout layer is also added between CNN layer 5 and the fully connected layer. Also, between the fully connected layer and the output layer. This Dropout layer helps to prevent overfitting of the algorithm during the training phase. In addition, we flatten the result from CNN layer 5 prior to the application of the fully connected layer. This architecture enables the algorithm to learn the most discriminative features that depict the signature of each individual appliances that are present in the aggregated signal.

Algorithm 1: Seq2Seq training procedure.

Input: sequence_length, n_epochs, batch_size, train_main, train_appliances
Output: training_model for each appliance in train_appliances
begin
  $train\_main = train\_main.values.reshape(-1, sequence\_length, 1)$;
  $new\_train\_appliances = []$;
  foreach $app\_name, app\_df s \in train\_appliances$ **do**
    $app\_df\_values = app\_df.values.reshape(-1, sequence\_length)$;
    $new\_train\_appliances.append((app\_name, app\_df\_values))$;
  $train\_appliances = new\_train\_appliances$;
  foreach $appliance\_name, power \in train\_appliances$ **do**
    /* create a Sequential model in Tensorflow for appliance_name using the architecture in Fig. 6 */
    $training\_model[appliance\_name] = SequentialModel(model\_architecture)$;
    /* fit the model for appliance_name */
    $training\_model[appliance\_name].fit(train\_main, power, epochs = n\_epochs, batch\_size = batch\_size)$;
    /* save the trained model for appliance_name to file */
    $training\_model[appliance\_name].save(appfilePath)$;
  return $training\_model$;

3.2.3. Event detection using thresholding methods

This enables us to determine the switching ON/OFF of individual appliances from their representational loads. To detect the ON event of a specific appliance ℓ, we need to compute the threshold $\lambda^{(\ell)}$. Thereafter, Eqn. (3) is applied to map appliance load consumption $P_t^{(\ell)}$ to its activation state at time t based on the computed threshold $\lambda^{(\ell)}$. This equation represents the ON event of appliance ℓ within a specified timestamp t. Formally,

$$s_t^{(\ell)} = I(P_t^{(\ell)} \geq \lambda^{(\ell)}) \tag{3}$$

where $P_t^{(\ell)}$ is the load consumption of appliance ℓ within the specified period t and $s_t^{(\ell)}$ represents the ON event/state of appliance ℓ within this period t.

To compute the threshold $\lambda^{(\ell)}$, we adopted three thresholding methods as earlier stated, which are MPT, VST and ATE.

3.2.4. Middle-point thresholding

Threshold $\lambda^{(\ell)}$ is computed using MPT method [7] according to Eqn. (4). The method generates two clusters by splitting the training data using a clustering algorithm. The centroid of each cluster is then considered. Any clustering algorithm can be used for this purpose, however, we utilized K-means clustering algorithm due to its scalability and wide acceptance. These two centroids obtained using K-means clustering algorithm are represented as $m_0^{(\ell)}$ and $m_1^{(\ell)}$ for OFF and ON states respectively. Thus, $\lambda^{(\ell)}$ is fixed between these two values such that,

$$\lambda^{(\ell)} = \frac{m_0^{(\ell)} + m_1^{(\ell)}}{2} \tag{4}$$

3.2.5. Variance-sensitive thresholding

This thresholding method attempts to offer an improvement over the MPT method through the incorporation of standard deviation parameter $\sigma_k^{(\ell)}$ obtained from the data points of each cluster. This extension, for the case when $\sigma_1 > \sigma_0$, ensures that the data points in cluster 1, which represents the ON event and which are far away from m_1 centroid are correctly classified. Hence, the constraint ensures that the threshold shift in the direction of m_0. The threshold $\lambda^{(\ell)}$ is computed according to as follows:

$$\begin{aligned} d &= \frac{\sigma_0^{(\ell)}}{\sigma_0^{(\ell)} + \sigma_1^{(\ell)}} \\ \lambda^{(\ell)} &= (1-d)m_0^{(\ell)} + d m_1^{(\ell)} \end{aligned} \tag{5}$$

3.2.6. Activation time extraction

The two thresholding methods discussed earlier (i.e. MPT and VST) fix the threshold $\lambda^{(\ell)}$ for appliance ℓ using the distribution of power measurements. In most cases, especially for multi-state appliances, due to the noise generated by the smart meters, several power measurements may be absent during short time intervals when the device is operating or may produce irregular peaks when the device is in the OFF state. Considering this behavior, [11] introduced ATE algorithm, which considered both power and time thresholds. The algorithm was specifically tuned for UK-DALE NILM dataset.

3.3. Proposed methods for privacy preservation

This section presents the two privacy protection mechanisms proposed in this study to reduce energy disaggregation. The first approach is based on microaggregation using Maximum Distance to Average Vector (MDAV) [38]. The second approach is based on

the hybridization of DFT and microaggregation (DFTMicroagg) where DFT is used to offer an additional layer of perturbation for energy data. These two algorithms have been recently studied in our work [23] to protect the privacy of individual daily energy consumption data.

Suppose there is a time series dataset of aggregated signals from individual appliance consumptions with smart meter numbers from each household, timestamps of energy usage and the exact aggregated loads from the appliances. This data is equivalent to the aggregated data that is disaggregated to constitute appliance signatures by Seq2Seq NILM algorithm. As stated earlier, this high-frequency data (HF) from the mains smart meter are capable of revealing the consumption patterns of the households' lifestyles through the signature of the individual appliances that have been aggregated. The goal is to protect this aggregated data in order to reduce the probability that any NILM disaggregation algorithm can correctly predict the appliance load signatures. In other words, we want to reduce the disaggregation risk that may be associated with disaggregating the individual appliance signatures (i.e. to reduce inference attack on energy data). To achieve this goal, we investigate the performance of MDAV and DFTMicroagg algorithms as privacy protection mechanisms.

3.3.1. MDAV microaggregation

One of the microaggregation algorithms that satisfies k-anonymity [37] is MDAV. This algorithm enforces k-anonymity condition by ensuring that the privacy of individual records in protected data is guaranteed. To do this, each record in protected data is masked such that this record cannot be identified within a set of k individuals records. This study adopted MDAV because of its privacy-preservation effectiveness as reported in the previous studies [44,23]. The procedures used by MDAV algorithm are presented in Algorithm 2.

Algorithm 2: MDAV.

Input: Original dataset X, anonymity level k
Output: Masked dataset $\hat{X}$ satisfying k-anonymity
begin
  $C = \emptyset$;
  while $|X| \geq 3k$ **do**
    $\bar{x} \leftarrow$ mean of all records present in X;
    $x_r \leftarrow$ closest record to $\bar{x}$;
    $x_s \leftarrow$ closest record to x_r;
    $C_r \leftarrow$ perform clustering with x_r as centroid (using x_r and $k-1$ most distant records from x_r);
    $C_s \leftarrow$ perform clustering with x_s as centroid (using x_s and $k-1$ most distant records from x_s);
    delete entries in C_r and C_s from X;
    $C = C \cup \{C_r, C_s\}$;
  if $|X| \geq 2k$ **then**
    $\bar{x} \leftarrow$ mean of all records present in X;
    $x_r \leftarrow$ closest record to $\bar{x}$;
    $C_r \leftarrow$ perform clustering with x_r as centroid (using x_r and $k-1$ most distant records from x_r);
    $C_s \leftarrow X \setminus C_r$ (use the rest of the records to form another cluster);
    $C = C \cup \{C_r, C_s\}$;
  else
    $C = C \cup \{X\}$;
  return (C);

3.3.2. DFTMicroagg

Generally, given a real sequence of numbers, discrete Fourier transform (DFT) produces another sequence of complex numbers with the same length. DFT ensures that an equally-spaced finite sequence sample of a function are transformed to the same length of equally-spaced sequence of coefficients of a finite combination of complex-valued function of frequency. An inverse DFT is a Fourier series that uses the DFT samples as coefficients of complex sinusoids at the corresponding DFT frequencies. An inverse DFT produces the same sample values corresponding to the original input sequence. Thus, DFT is generally referred to as the frequency domain representation of the original input values. DFTMicroagg (see Algorithm 3 and [23]) benefits from microaggregation and DFT to provide an additional layer of perturbation to microaggregation procedures.

The number of coefficients for DFTMicroagg is obtained by,

$$coeff = \frac{T}{i} \tag{6}$$

where T is the number of time stamps based on the standard representation format of the aggregated dataset, i represents a constant that enforces privacy, which is chosen by the utility company to protect the aggregated data.

4. Experimental setup

All methods discussed are implemented in Python. NILMTK and NILMTK-Contrib API [5] were used for energy disaggregation. We use Intel(R) Core^{TM} i9-8950H CPU Dell Laptop with @2.90 GHz 1TB HDD 32 GB RAM and GeForce GTX 1050 Ti with Max-Q Design, and CUDA version 11.2. The batch size and number of epochs for Seq2Seq algorithm are set to 32 and 50 respectively. UK-DALE

Algorithm 3: DFTMicroagg.

Input: Original dataset X, anonymity level k,
$coeff$:− an integer value for the number of coefficients of DFT to retain.
Output: Masked dataset $\hat{X}$ satisfying k-anonymity
begin
$totalts \leftarrow$ total time stamps from X;
if $is\text{-}even(coeff)$ **then**
$Real\text{-}indices = sequence(1, coeff, 2)$;
$Imag\text{-}indices = sequence(2, coeff, 2)$;
$dft\text{-}fft \leftarrow$ compute DFT on X based on $Real\text{-}indices$ and $Imag\text{-}indices$;
else
$Real\text{-}indices = sequence(1, coeff, 2)$;
$Imag\text{-}indices = sequence(2, coeff + 1, 2)$;
$dft\text{-}fft \leftarrow$ compute DFT on X based on $Real\text{-}indices$ and $Imag\text{-}indices$;
$X\text{-}modified \leftarrow$ compute inverse FFT using $dft\text{-}fft$ and $totalts$;
$\hat{X} \leftarrow MDAV(X\text{-}modified, k)$;
return $(\hat{X})$;

and REFIT NILM datasets have been considered. Section 4.1 described the detail of the two datasets. We re-sampled both appliance and aggregated data using 1 min (60 s) period during the energy disaggregation experiment. This enables proper alignment of both aggregated and appliance-level signals during energy disaggregation phase. The same sampling frequency was used to compute the threshold values from the appliance data. Section 4.2 presents the training and testing configurations used during the three attack scenarios discussed in Section 3.2.1.

During microaggregation and DFTMicroagg implementation, the value of K was set to 5 and DFT coefficient was 14,400 for the two datasets, where the value of i has been fixed to 6 (see Eqn. (6)). However, due to the high sampling frequency of UK-DALE and REFIT datasets, subset-based microaggregation was performed when running MDAV and DFTMicroagg to protect individual appliance load signatures. Therefore, during microaggregation phase, each dataset has been partitioned into two subsets based on 12 hours sampling.

4.1. Datasets

This study considered two datasets that are publicly available. They are widely used in NILM energy disaggregation domain. These datasets are UK-DALE [26] and REFIT [45]. The first dataset, UK-DALE, contains five (5) households data. In UK-DALE dataset, aggregate apparent mains power was recorded by the appliance sub-meters for each of the households. This was sampled every six (6) seconds. In addition, Household 1, 2 and 5 in UK-DALE dataset also have per-second measurements for both active and reactive mains power. Conversely, REFIT dataset has 20 households with 8 seconds of sampled data for both aggregate and appliance-level meters.

4.2. Training and testing periods

Three experimental settings were configured for simulating the three attack cases earlier discussed. The first inference attack case involves the use of household 2 data based on active power for both aggregate and appliance-level smart meters. For this scenario, 4 months data was used for training covering a period from 20/05/2013 to 20/09/2013, and the testing was done based on about 1 month data covering a period from 21/09/2013 to 10/10/2013 using UK-DALE dataset. Household 2 in REFIT dataset was also utilized in this scenario based on active power for both mains and appliance smart meters. For this dataset, the training period starts on 17/09/2013 and ends on 17/01/2014. The testing period starts from 01/03/2014 and ends on 01/04/2014.

The second case of the attack simulation utilizes household 2 data to develop the Seq2Seq model, which was then tested using household 1 data. For this case, the training covers a period from 20/05/2013 to 20/09/2013, and the testing covers a period from 21/09/2013 to 10/10/2013 using UK-DALE dataset. Conversely, for REFIT dataset, Seq2Seq algorithm was trained and tested using household 2 and 5 respectively. The training starts on 17/09/2013 and ends on 17/01/2014, and the testing period starts on 01/03/2014 and ends on 01/04/2014.

In the third inference attack scenario, the training of the Seq2Seq algorithm was based on UK-DALE household 2 data and the resulting model was tested based on REFIT household 2 data. In this case, the training starts on 20/05/2013 and ends on 20/09/2013. The testing starts on 01/03/2014 and ends on 01/04/2014.

4.3. Threshold computation

This section presents the threshold results computed using appliance-level training data. The value of the threshold for ATE in [11] was used for REFIT and UK-DALE datasets. This helps us to test the efficacy of the threshold method in [11] on the two datasets. Nevertheless, the ON power threshold values were calculated for MPT and VST methods using individual appliance-level data as shown in Table 2 for both datasets.

Table 2
Computed threshold values for ON power in Watt based on UK-DALE and REFIT.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	20.0	864.49	219.8	20.0	1028.20	592.29
Fridge	50.0	47.85	18.73	50.0	44.73	3.85
Dish washer	10.0	1054.8	146.22	10.0	1100.99	669.62
Microwave	200.0	562.56	72.30	200.0	555.67	67.51
Kettle	2000.0	1059.66	117.34	2000.0	1359.92	241.59

Table 3
Disaggregation risk results for attack scenario 1 for each appliance based on UK-DALE and REFIT datasets before applying privacy protection mechanisms.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	0.87	0.48	0.62	0.56	0.44	0.70
Fridge	0.96	0.96	0.99	0.75	0.80	0.99
Dish washer	0.98	0.98	1.00	0.93	0.64	0.81
Microwave	0.83	0.62	0.89	0.00	0.00	0.00
Kettle	0.90	0.95	0.99	0.41	0.55	0.72

5. Results and discussion

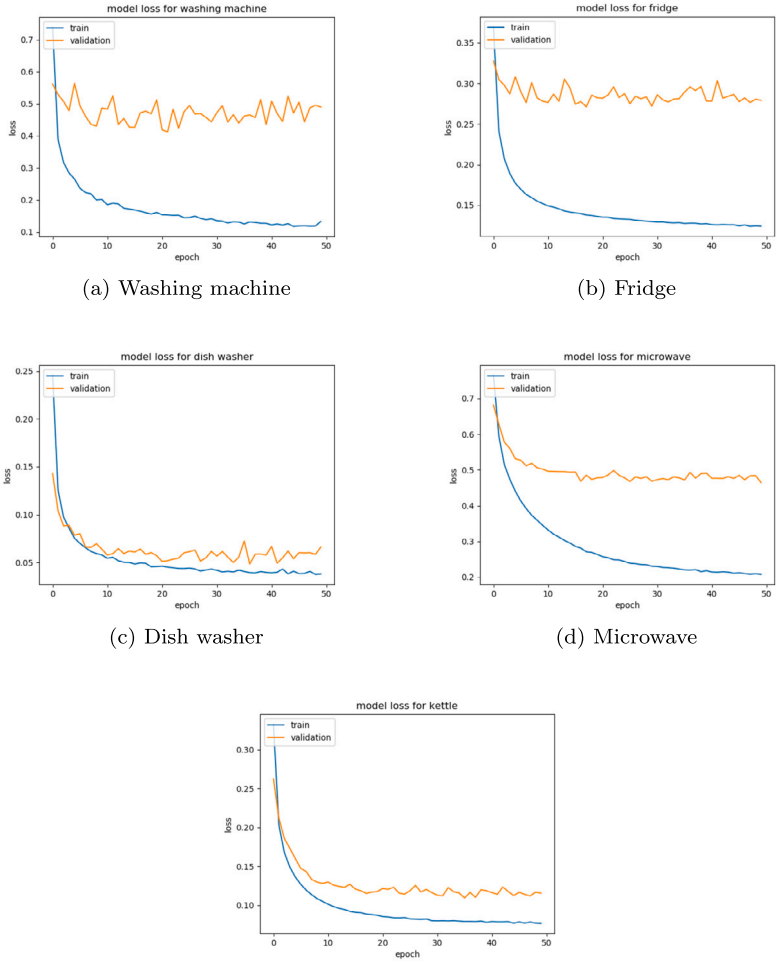
Here, we present the disclosure risk results for the individual appliances based on the three cases of inference attack that we investigated. This section also presents the equivalent results obtained when the proposed privacy-preserving mechanisms (MDAV and DFTMicroagg) were applied for hiding individual appliance load signatures that are present in the aggregated signals.

5.1. Inference attack simulation results

5.1.1. Scenario 1: inference attack on the same household data

In this attack scenario, Seq2Seq model was trained and tested using data from the same household in the same dataset. Fig. 7 presents sample performance of the Seq2Seq deep learning model during the training process. After the training process is completed and the testing has been performed, events detection using the three thresholding methods are then employed to enable us compute the disclosure risk probability. Table 3 presents the disclosure risk of individual appliances using Seq2Seq energy disaggregation model for individual appliance loads disaggregation and the three threshold methods for event detection. The privacy leakage is computed based on the disaggregation risk for each appliance due to the success rate of Seq2Seq model in predicting the signature of each appliance before applying the proposed privacy protection methods. The disaggregation risk results of the three threshold techniques are very close, confirming that the three methods have the tendency of revealing the ON power states of the individual appliances. However, for this scenario, there is a challenge with the three methods to accurately predict the positive states/events of microwave appliances on REFIT dataset. Considering UK-DALE dataset, the results obtained in this table show that VST outperformed the other two methods by revealing the signatures of four out of five appliances with the highest probability. This probability implies the level of disaggregation risk as previously emphasized. For REFIT dataset, VST also outperformed the other three methods for event detection. In the majority of cases, the results show the ability of Seq2Seq NILM algorithm in disaggregating each appliance signature. In this scenario, we observed a high probability showing that attacker can successfully disaggregate individual appliance signatures using a pre-trained Seq2Seq model on the same household data.

Table 4 and Table 5 show the equivalent results when MDAV and DFTMicroagg were implemented as privacy protection methods. It can be seen that DFTMicroagg offers slight improvement (not in all cases) as compared to the results produced if MDAV algorithm is applied as a standalone protection method. This can be seen in the case of the washing machine for ATE threshold method. Similarly, for fridge appliance on REFIT with ATE and MPT methods. However, both MDAV and DFTMicroagg exhibit close performance and they reduce the disaggregation risk of publishing the aggregated signals in their original form without applying any privacy protection mechanism (see Table 3 for comparison). For dish washer and washing machine in both datasets ATE gave the highest disclosure risk after applying the two privacy protection methods. In case of fridge, the highest disaggregation risk can be seen with the result produced by VST method. For microwave, VST also gave the highest disclosure risk. Although the disaggregation risks of the appliances were reduced after applying the two privacy protection methods, the results still confirmed that VST and ATE tried to reveal the behavioral patterns of the appliances. We observed that the disaggregation risk of fridge is still on the high side even after applying the proposed privacy protection mechanisms. This can be attributed to the usage patterns of fridge appliances because this device is always in the ON state in the majority of the timestamps in the two datasets, making it difficult to completely lower the disaggregation risk for this particular appliance.



(a) Washing machine (b) Fridge

(c) Dish washer (d) Microwave

(e) Kettle

Fig. 7. Sample training runs of Seq2Seq for scenario 1 using UK-DALE.

Table 4
Disaggregation risk results for attack scenario 1 for each appliance based on UK-DALE and REFIT datasets after MDAV microaggregation.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	0.12	0.04	0.04	0.03	0.00	0.00
Fridge	0.80	0.82	0.98	0.29	0.35	0.99
Dish washer	0.34	0.03	0.12	0.05	0.00	0.00
Microwave	0.05	0.01	0.14	0.00	0.00	0.00
Kettle	0.00	0.00	0.00	0.00	0.00	0.01

Table 5
Disaggregation risk results for attack scenario 1 for each appliance based on UK-DALE and REFIT datasets after DFTMicroagg.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	0.10	0.05	0.04	0.01	0.00	0.00
Fridge	0.79	0.82	0.98	0.16	0.18	0.99
Dish washer	0.34	0.04	0.14	0.02	0.00	0.00
Microwave	0.07	0.00	0.11	0.00	0.00	0.01
Kettle	0.00	0.00	0.00	0.00	0.00	0.00

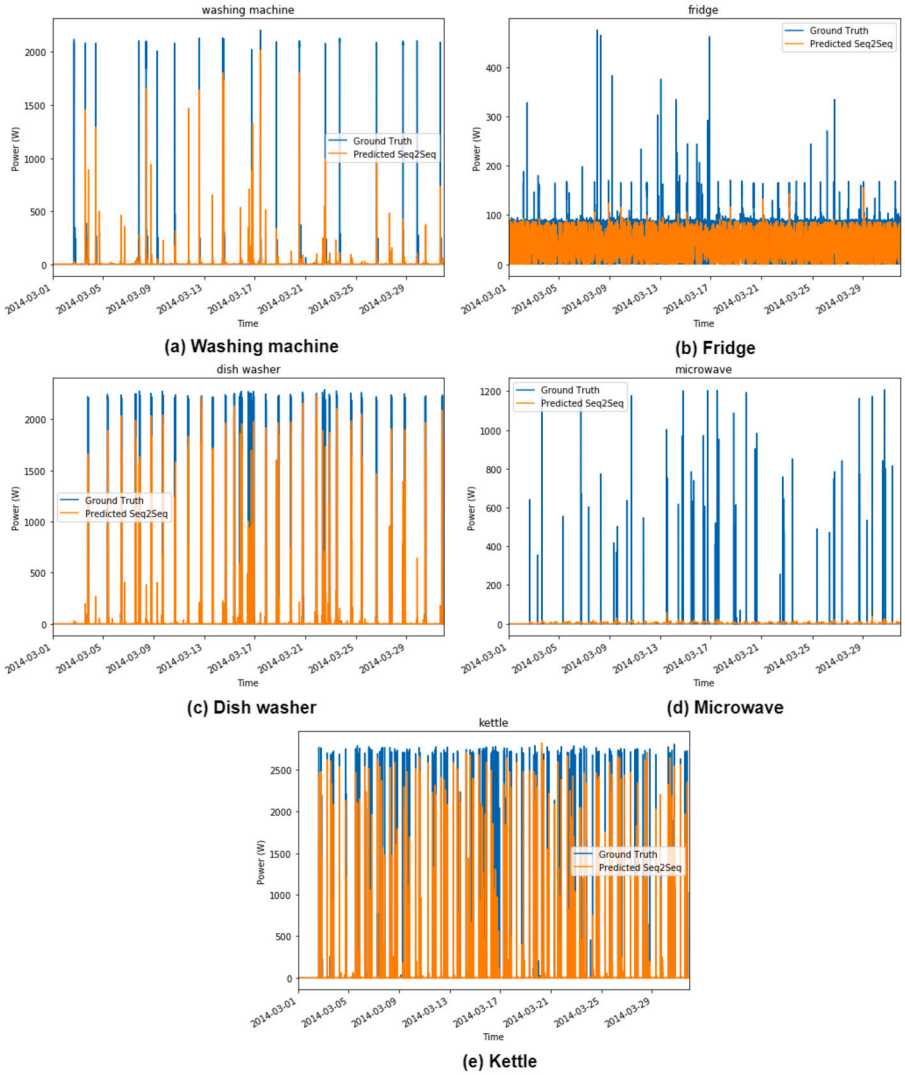


Fig. 8. Scenario 1 - Ground truth and Predicted load consumptions of the Seq2Seq NILM algorithm on REFIT dataset for the appliances before applying privacy protection methods.

Comparatively, Fig. 9 and 10 further show how the proposed privacy protection mechanisms have been able to prevent Seq2Seq energy disaggregation algorithm from accurately predicting the individual appliance load signatures when compared with the results in Fig. 8 where the privacy protection method was not implemented. We show the results for REFIT dataset only for this scenario because of the space constraint. The figures also show the reason why the disaggregation risk of the fridge appliance is still on the high side due to the frequency of usage of this appliance.

5.1.2. Scenario 2: inference attack on different households in the same dataset

Table 6 presents the results of the disclosure risk based on inference attack on different households using the same dataset. Comparing these results with the first scenario of inference attack, it was observed that the disclosure risk of microwave and washing machine reduces. Although the results of the threshold methods are close as observed in the previous scenario, however, ATE and VST outperformed MPT except in the case of fridge and kettle where MPT slightly outperformed ATE. VST gave the highest disaggregation risk when considering fridge appliance. In fact, VST produced 100% disaggregation risk, which confirmed that this method can reveal the signature of appliances with frequent usage patterns as well as appliances with ON/OFF state as in the case of kettle for the two datasets when compared with the other methods. Nevertheless, it can be seen that the probability of revealing individual appliance signatures is observable and this poses privacy risk to the individual households' lifestyles. In all cases, the results further confirmed that Seq2Seq NILM algorithm is a good candidate deep learning algorithm for energy disaggregation as we can observe a high level of disaggregation risk for the different appliances. We further check in the subsequent paragraph if the application of the two privacy protection methods can lower this disclosure risk.

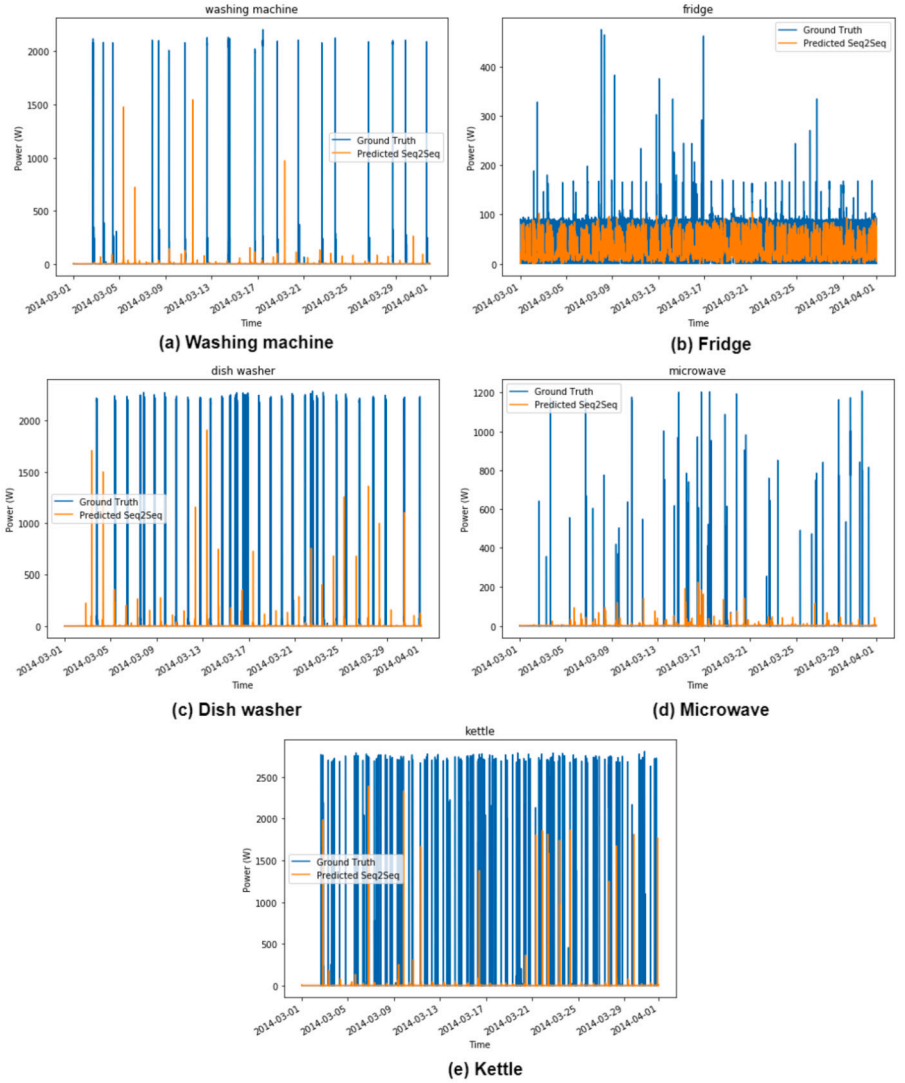


Fig. 9. Scenario 1 - Ground truth and Predicted load consumptions of the Seq2Seq NILM algorithm on REFIT dataset for the appliances after applying MDAV microaggregation algorithm.

Table 6
Disaggregation risk for attack scenario 2 for each appliance based on UK-DALE and REFIT datasets before applying privacy protection mechanisms.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	0.26	0.13	0.15	0.41	0.01	0.10
Fridge	0.88	0.89	0.99	0.55	0.64	0.99
Dish washer	0.93	0.70	1.00	0.92	0.10	0.44
Microwave	0.40	0.19	0.57	0.00	0.00	0.03
Kettle	0.51	0.73	0.85	0.36	0.56	0.77

Similarly, Table 7 and Table 8 show the equivalent results when MDAV and DFTMicroagg were implemented as privacy protection methods against inference attack scenario 2. Similar results to the one obtained in Scenario 1 were observed. This confirmed the consistency of the proposed mechanisms. We also observed from the results in these tables that VST and ATE compete with each other in revealing the signatures of the appliances. In the case of fridge, VST produced the highest energy disaggregation risk. Furthermore, both privacy protection methods produced promising results by lowering the disaggregation risks associated with disaggregating the individual appliances in the aggregated signals as compared to the results in Table 6 when privacy protection method has not been applied.

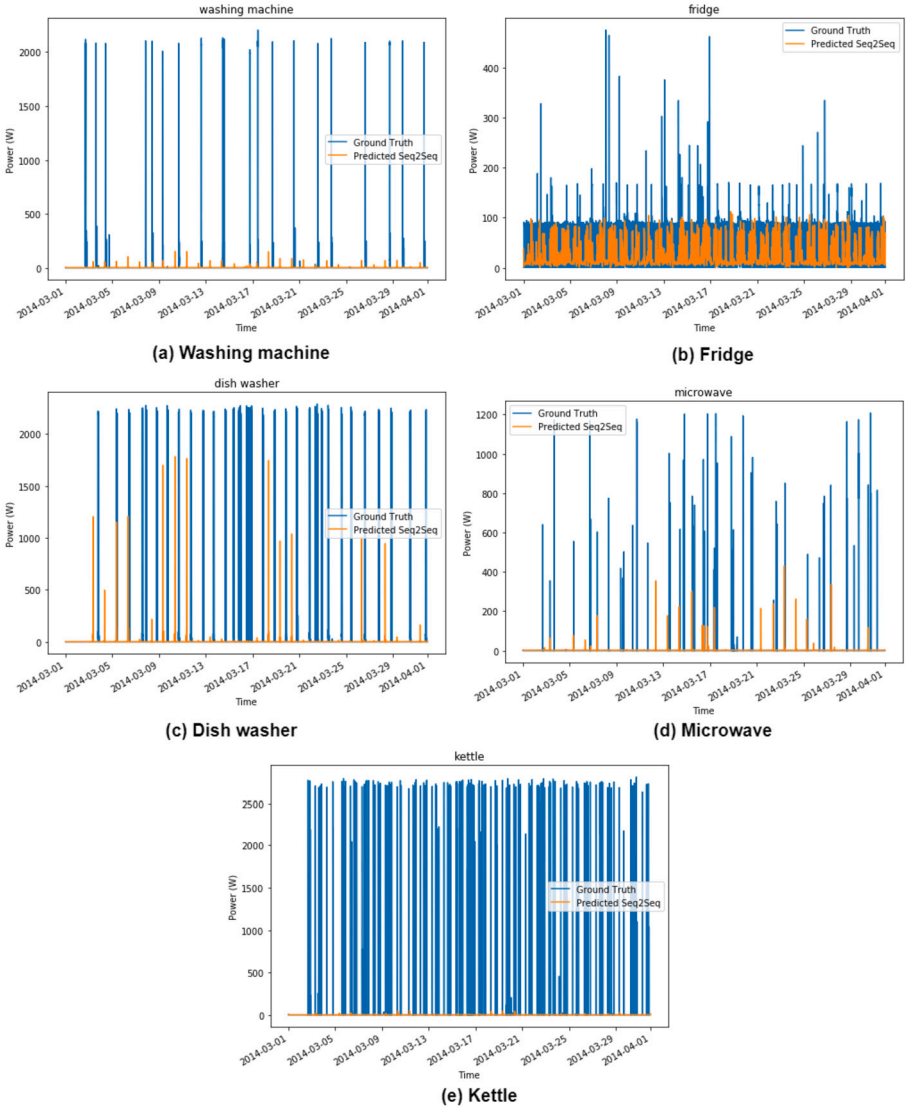


Fig. 10. Scenario 1 - Ground truth and Prediction load consumptions of the Seq2Seq NILM algorithm on REFIT dataset for the appliances after applying DFTMicroagg algorithm.

Table 7
Disaggregation risk for attack scenario 2 for each appliance based on UK-DALE and REFIT datasets after applying MDAV microaggregation.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	0.19	0.01	0.10	0.02	0.00	0.00
Fridge	0.77	0.79	0.98	0.36	0.42	0.96
Dish washer	0.31	0.19	0.26	0.01	0.00	0.00
Microwave	0.01	0.00	0.03	0.02	0.00	0.07
Kettle	0.00	0.00	0.00	0.00	0.00	0.00

5.1.3. Scenario 3: inference attack on different households from different datasets

The results of this experiment are shown in Table 9. The table presents the disclosure risk of each appliance when Seq2Seq was trained with UK-DALE data and tested with REFIT data. In this case, the attacker is targeting to obtain useful information regarding appliance consumption patterns in the households in REFIT dataset. We observed a high success rate of the NILM algorithms despite the differences in the data utilized for both training and testing (with the exception of washing machine). We noticed that the algorithms successfully disaggregate microwave signatures in REFIT datasets using a pre-trained model with UK-DALE data during

Table 8
Disaggregation risk for attack scenario 2 for each appliance based on UK-DALE and REFIT datasets after applying DFTMicroagg.

Appliance	UK-DALE			REFIT		
	ATE	MPT	VST	ATE	MPT	VST
Washing machine	0.18	0.00	0.11	0.00	0.00	0.00
Fridge	0.74	0.77	0.97	0.23	0.27	0.92
Dish washer	0.31	0.19	0.25	0.00	0.00	0.00
Microwave	0.04	0.01	0.05	0.00	0.00	0.00
Kettle	0.00	0.00	0.01	0.00	0.00	0.00

Table 9
Disaggregation risk for attack scenario 3 for each appliance in UK-DALE and REFIT datasets before applying privacy protection methods.

Appliance	UK-DALE and REFIT		
	ATE	MPT	VST
Washing machine	0.12	0.01	0.04
Fridge	0.81	0.83	0.99
Dish washer	0.99	0.83	0.99
Microwave	0.44	0.22	0.61
Kettle	0.54	0.73	0.87

Table 10
Disaggregation risk for attack scenario 3 for each appliance in UK-DALE and REFIT datasets after applying MDAV algorithm.

Appliance	UK-DALE and REFIT		
	ATE	MPT	VST
Washing machine	0.05	0.00	0.02
Fridge	0.66	0.69	0.95
Dish washer	0.09	0.01	0.03
Microwave	0.00	0.00	0.01
Kettle	0.00	0.00	0.00

Table 11
Disaggregation risk for attack scenario 3 for each appliance in UK-DALE and REFIT datasets after applying DFTMicroagg algorithm.

Appliance	UK-DALE and REFIT		
	ATE	MPT	VST
Washing machine	0.01	0.00	0.00
Fridge	0.78	0.79	0.95
Dish washer	0.01	0.00	0.00
Microwave	0.00	0.00	0.00
Kettle	0.00	0.00	0.00

the inference attack simulation to infer useful information from the target households. The findings further confirmed that Seq2Seq NILM algorithms can successfully reveal the load consumption patterns of individual appliances in aggregated signals. For this scenario, VST produced the highest performance except in the case of washing machine where ATE outperformed the other two event detection methods.

Nevertheless, the results obtained after applying the proposed privacy protection mechanisms provide promising privacy guarantees for individual households based on their energy usage patterns. Table 10 and 11 also confirmed the efficacy of MDAV and DFTMicroagg algorithms respectively as effective privacy protection mechanisms to reduce energy disaggregation risk. These algorithms reduced the ability of Seq2Seq disaggregation algorithm to accurately predict the energy consumed by the individual appliance for this inference attack scenario. Promising results were achieved across the three threshold methods that have been considered in this study. For instance, in the case of ATE, washing machine disaggregation risk was reduced from 12% to 1%. Dish washer disaggregation risk was reduced from 99% to 9% for MDAV and to 1% for DFTMicroagg. Disaggregation risk for Microwave and Kettle was reduced from 44% and 54% respectively to 0% when MDAV and DFTMicroagg algorithms were used. Similar results were achieved for MPT and VST where MDAV and DFTMicroagg were able to significantly reduce the disaggregation risk associated with

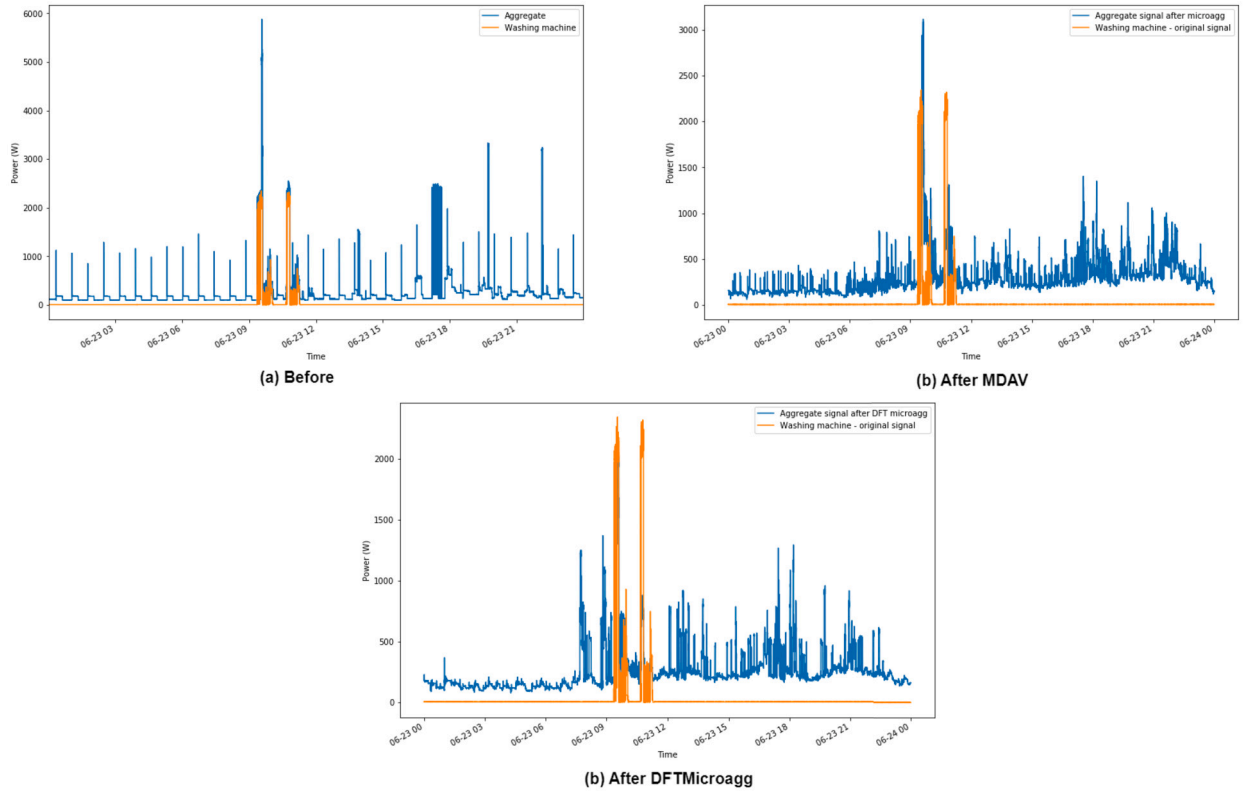


Fig. 11. Washing machine energy consumption patterns in the aggregated signal before and after applying privacy protection mechanisms.

disaggregating the individual household lifestyles. However, we noticed a similar result in the case of fridge as previously observed in other attack scenarios. This is because the activation period for fridge spans almost the entire timestamps of the datasets and this does not present a clear pattern of usage as to when the appliance is in the OFF state. This accounts for the increase in the disaggregation risk for fridge as compared with the results obtained based on the other appliances.

5.2. Impact on signal structure

Fig. 11 shows that both MDAV and DFTMicroagg algorithms have introduced some distortions in the aggregated signal as compared to the original signal before applying privacy protection mechanisms. It can be seen that the aggregated signal resulting from the application of MDAV and DFTMicroagg cannot be clearly linked to the original aggregated signal (e.g. washing machine signature as a sample case). This result further confirmed the applicability and usefulness of the two privacy-preserving methods for smart grid data publishing. From these figures, it can also be seen that the patterns of other appliances in the background have also been successfully hidden after applying the proposed mechanisms.

5.3. Comparison with existing NILM studies

In this section, we compare the performance of our adapted Seq2Seq model with existing NILM studies. While existing studies in NILM focus on energy disaggregation and optimization, our objective with the NILM part of this research is to create Seq2Seq deep learning model that can generate attack models for inference attacks on the individual households as presented in the different scenarios. The disaggregation risk in Eqn. (1) helps in correctly modeling the power of an adversary in terms of the probability of predicting which load is in operation at a specific time of the day. It also helps in avoiding class imbalance problem that is associated with energy data. However, to compare our approach with existing studies in NILM, we conducted experiment where metrics such as F1-score and MAE are computed. We compute specific metric as used in each of the studies. The result of our Seq2Seq model for F1-score is based on VST thresholding. It is also important to mention that this comparison is based on the first scenario of our study as this is the case that is closely related to the existing studies that we used for comparison. Table 12 provides the detail of the results for UK-DALE dataset. As observed from this table, our proposed Seq2Seq for the attack modeling provides comparable results with the state-of-the-arts. We achieved the highest F1-score for Washing machine, Microwave and Kettle. The model in [30] produced the highest F1-score for Fridge and Dish washer. The F1-score of our proposed model is closer to the one in [30]. In terms of the MAE, our proposed model produced lowest error for Washing machine, Dish washer, Microwave and Kettle while that of [30] achieved the lowest MAE for Fridge.

Table 12
Comparison of the results of our adapted Seq2Seq architecture with existing NILM studies based on UK-DALE.

Metrics	Methods	Washing M.	Fridge	Dish W.	Microw.	Kettle
F1-score	Luan et al. [28]	52.30	82.90	48.90	35.7	90.20
	Zhou et al. [29]	84.80	74.00	86.80	43.40	87.30
	Dash et al. [30]	97.30	**88.00**	**99.28**	94.20	96.06
	Proposed	**99.00**	72.00	99.00	**99.10**	**98.01**
MAE	Kelly et al. [11]	163.47	38.45	237.96	14.56	13.00
	Dash et al. [30]	42.48	**10.78**	17.47	20.47	50.65
	Zhang et al. [12]	10.15	20.89	27.704	8.66	7.44
	Proposed	**6.32**	14.66	**7.32**	**3.93**	**4.36**

Table 13
Comparison of the results of our adapted Seq2Seq architecture with existing NILM studies based on REFIT.

Metrics	Methods	Washing M.	Fridge	Dish W.	Microw.	Kettle
F1-score	Luan et al. [28]	53.60	**73.00**	38.5	67.80	49.8
	Dash et al. [30]	94.70	63.78	96.97	96.27	**99.20**
	Proposed	**98.89**	66.00	**99.00**	**98.90**	99.00
MAE	Luan et al. [28]	**18.60**	22.10	**11.20**	9.50	**15.50**
	Dash et al. [30]	78.48	21.49	37.45	13.56	16.92
	Proposed	20.98	**18.23**	39.55	**5.32**	17.41

Similarly, Table 13 presents the comparison of our approach with existing NILM studies based REFIT dataset. It can be seen that the proposed model achieved the highest F1-score for Washing machine, Dish washer and Microwave, as well as the lowest MAE for Fridge and Microwave. The model in [28] produced the lowest MAE for the remaining three appliances. Nevertheless, our proposed model produced promising MAE for REFIT dataset that is comparable to the state-of-the-art methods.

6. Conclusion

In this paper, we establish a new measure of disclosure risk called energy disaggregation risk. We demonstrate the capability of Seq2Seq deep learning NILM algorithm in predicting the load signatures of individual appliances that constitute aggregated power signals. Seq2Seq NILM algorithm produced significant results for energy disaggregation tasks. The ability of deep learning NILM algorithms to automatically learn the load signature of appliances has a significant impact on the load disaggregation results. Additionally, we consider three threshold methods for event detection, which are used to detect the signature of each appliance based on the energy consumption predicted by Seq2Seq NILM algorithm. The results obtained show that VST and ATE event detection methods produced high disaggregation risk on the two publicly available datasets that we considered. This implies that the methods can successfully detect appliance signature with high level of confidence. Therefore, this study revealed that publishing smart grid data without the application of data anonymization violates individual household lifestyles.

To further confirm the level of disclosure risk that is associated with energy disaggregation, we simulated three inference attack scenarios. The findings show that the possibility of executing successful inference attacks on smart grid data is on the high side. Particularly, out of the three event detection methods that we studied, VST produced the highest disaggregation risk confirming that this method has the highest probability of inferring household lifestyle. This result is followed by the ATE method. For instance, for the first scenario of inference attack on UK-DALE dataset, VST produced disaggregation risk of 99%, 100%, 89% and 99% for fridge, dish washer, microwave, kettle respectively. Similar pattern of result was obtained in the case of REFIT dataset. The implication of this result is that Seq2Seq in combination with VST method for event detection can successfully reveal household lifestyles. VST produced similar pattern of result in the case of second and third scenarios of inference attacks. The results empirically validate our notion of disaggregation risk. In all cases, the results confirmed the efficacy of Seq2Seq NILM algorithm and the possibility of launching inference attack on smart grid data.

To prevent infringement on the privacy of individual household lifestyles, we investigated the performance of two privacy protection mechanisms. The results, after applying these mechanisms, show that DFTMicroagg offered slight improvement over MDAV algorithm for smart grid data anonymization. These algorithms lower the disclosure risk associated with each appliance. Particularly, for UK-DALE dataset the disaggregation risk results of VST for the first scenario were reduced to 98%, 14%, 11% and 0% for fridge, dish washer, microwave, kettle respectively. The REFIT disaggregation risk result for this scenario was reduced to 0% except in the case of fridge. We observed similar pattern of results for other scenarios. We also observed a specific case for the fridge disaggregation results where the two mechanisms failed to reduce the disaggregation risk. The reason can be attributed to the fact that fridge is mostly in the ON state in the majority of the timestamps as observed in the two datasets and based on their real energy usage patterns.

Nevertheless, to improve the privacy of smart grid data before publishing, the proposed approach can be employed to hide individual appliance signatures that constitute aggregated power data. This will prevent attackers from inferring load signatures of appliances and consequently protect individual households' lifestyles. These mechanisms prevent NILM algorithms from accurately predicting the load consumption of individual appliances and their ON events based on the established disaggregation risk metric.

Furthermore, the findings from this paper also confirmed that the proposed mechanisms have introduced some distortions in the aggregated signal making it difficult to link original signals with the protected signals. This prevents attackers from directly inferring a specific appliance signature from the aggregated signal. Future work can focus on hierarchical protection techniques for multilevel privacy protection in smart grids. Additionally, future research can also consider reducing the computational requirements of microaggregation for high-frequency smart grid data with a large number of households and evaluating different values of parameter k for sensitivity analysis. Measuring uncertainty of the proposed approach can be considered in the future work.

CRediT authorship contribution statement

Kayode S. Adewole: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Vicenç Torra:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Kayode S. Adewole reports financial support was provided by Synergy Project of the Internet of Things and People (IoTaP) Research Centre, Malmo University, Sweden (Grant number 20220087). Kayode S. Adewole also got support from Kempe Foundation. Vicenc Torra reports financial support was provided by Knut and Alice Wallenberg Foundation, and the Swedish Research Council under the project Privacy for complex data (VR 2022-04645).

Data availability

The data used in the research are publicly available. We have shared the links.

Acknowledgement

This research has been partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation. Support also received from the Swedish Research Council under the project Privacy for complex data (VR 2022-04645). The first author has received support from Kempe Foundation and the Synergy Project of the Internet of Things and People (IoTaP) Research Centre, Malmo University, Sweden (Grant number 20220087).

References

[1] M. Ibrahim, A. El-Zaart, C. Adams, Smart sustainable cities roadmap: readiness for transformation towards urban sustainability, Sustain. Cities Soc. 37 (2018) 530–540.
[2] R. Gopinath, M. Kumar, C.P.C. Joshua, K. Srinivas, Energy management using non-intrusive load monitoring techniques-state-of-the-art and future research directions, Sustain. Cities Soc. 62 (2020) (2020) 102411.
[3] J. Parra-Arnau, Pay-per-tracking: a collaborative masking model for web browsing, Inf. Sci. 385 (2017) 96–124.
[4] A. Janik, A. Ryszko, M. Szafraniec, Scientific landscape of smart and sustainable cities literature: a bibliometric analysis, Sustainability 12 (3) (2020) 779.
[5] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, O. Parson, Towards reproducible state-of-the-art energy disaggregation, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, ACM, 2019, pp. 193–202.
[6] A. Verma, A. Anwar, M. Mahmud, M. Ahmed, A. Kouzani, A comprehensive review on the NILM algorithms for energy disaggregation, arXiv preprint, arXiv: 2102.12578, 2021.
[7] D. Precioso, D. Gomez-Ullate, Nilm as a regression versus classification problem: the importance of thresholding, arXiv preprint, arXiv:2010.16050, 2020.
[8] B. Wang, H. Ma, F. Wang, U. Dampage, M. Al-Dhaifallah, Z.M. Ali, M.A. Mohamed, An iot-enabled stochastic operation management framework for smart grids, IEEE Trans. Intell. Transp. Syst. 24 (1) (2022) 1025–1034.
[9] P. Laviron, X. Dai, B. Huquet, T. Palpanas, Electricity demand activation extraction: from known to unknown signatures, using similarity search, in: Proceedings of the ACM International Conference on Future Energy Systems, e-Energy, ACM, 2021.
[10] S. Desai, R. Alhadad, A. Mahmood, N. Chilamkurti, S. Rho, Multi-state energy classifier to evaluate the performance of the NILM algorithm, Sensors 19 (23) (2019) 5236.
[11] J. Kelly, W. Knottenbelt, Neural NILM: deep neural networks applied to energy disaggregation, in: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, ACM, 2015, pp. 55–64.
[12] C. Zhang, M. Zhong, Z. Wang, N. Goddard, C. Sutton, Sequence-to-Point Learning with Neural Networks for Non-intrusive Load Monitoring, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, AAAI, 2018.
[13] D. Mashima, A. Serikova, Y. Cheng, B. Chen, Towards quantitative evaluation of privacy protection schemes for electricity usage data sharing, ICT Express 4 (1) (2018) 35–41.
[14] V. Tudor, M. Almgren, M. Papatriantafilou, A study on data de-pseudonymization in the smart grid, in: Proceedings of the Eighth European Workshop on System Security, 2015, pp. 1–6.
[15] S. Armoogum, V. Bassoo, Privacy of energy consumption data of a household in a smart grid, in: Q. Yang, T. Yang, W. Li (Eds.), Smart Power Distribution Systems, Academic Press, 2019, pp. 163–177.

[16] BBCNews, Ukraine power cut 'was cyber-attack', https://www.bbc.com/news/technology-38573074, 2017.
[17] J.-X. Chin, T.T. De Rubira, G. Hug, Privacy-protecting energy management unit through model-distribution predictive control, IEEE Trans. Smart Grid 8 (6) (2017) 3084–3093.
[18] F.C. Sangogboye, R. Jia, T. Hong, C. Spanos, M.B. Kjærgaard, A framework for privacy-preserving data publishing with enhanced utility for cyber-physical systems, ACM Trans. Sens. Netw. (TOSN) 14 (3–4) (2018) 1–22.
[19] E.U. Soykan, Z. Bilgin, M.A. Ersoy, E. Tomur, Differentially private deep learning for load forecasting on smart grid, in: 2019 IEEE Globecom Workshops (GC Wkshps), IEEE, 2019, pp. 1–6.
[20] V. Thouvenot, D. Nogues, C. Gouttas, Data-driven anonymization process applied to time series, in: SIMBig, 2017, pp. 80–90.
[21] J. Parra-Arnau, Optimized, direct sale of privacy in personal data marketplaces, Inf. Sci. 424 (2018) 354–384.
[22] F. Fioretto, P. Van Hentenryck, Differential private stream processing of energy consumption, arXiv preprint, arXiv:1808.01949, 2018.
[23] K.S. Adewole, V. Torra, Dftmicroagg: a dual-level anonymization algorithm for smart grid data, Int. J. Inf. Secur. 21 (6) (2022) 1299–1321.
[24] G.W. Hart, E.C. Kern Jr., F.C. Schweppe, Non-intrusive appliance monitor apparatus, Tech. Rep., Massachusetts Inst. of Technology (MIT), Cambridge, MA (United States), 1989.
[25] T. Wang, T. Ji, M. Li, A new approach for supervised power disaggregation by using a denoising autoencoder and recurrent lstm network, in: 2019 IEEE 12th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED), IEEE, 2019, pp. 507–512.
[26] J. Kelly, W. Knottenbelt, The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes, Sci. Data 2 (1) (2015) 1–14.
[27] W. Lian, T. Wu, Y. He, Z. Shan, G. Si, A convolutional neural network and sequence-to-sequence model based energy disaggregation algorithm for non-intrusive load monitoring, 2021.
[28] W. Luan, R. Zhang, B. Liu, B. Zhao, Y. Yu, Leveraging sequence-to-sequence learning for online non-intrusive load monitoring in edge device, Int. J. Electr. Power Energy Syst. 148 (2023) 108910.
[29] G. Zhou, Z. Li, M. Fu, Y. Feng, X. Wang, C. Huang, Sequence-to-sequence load disaggregation using multiscale residual neural network, IEEE Trans. Instrum. Meas. 70 (2020) 1–10.
[30] S. Dash, N. Sahoo, Attention based multi-task probabilistic network for non-intrusive appliance load monitoring, IEEE Trans. Instrum. Meas. 72 (2023) 2513412.
[31] T. Todic, V. Stankovic, L. Stankovic, An active learning framework for the low-frequency non-intrusive load monitoring problem, Appl. Energy 341 (2023) 121078.
[32] F. Alsokhiry, A. Annuk, M.A. Mohamed, M. Marinho, An innovative cloud-fog-based smart grid scheme for efficient resource utilization, Sensors 23 (4) (2023) 1752.
[33] M.A. Mohamed, A relaxed consensus plus innovation based effective negotiation approach for energy cooperation between smart grid and microgrid, Energy 252 (2022) 123996.
[34] R. Jia, F.C. Sangogboye, T. Hong, C. Spanos, M.B. Kjærgaard, Pad: protecting anonymity in publishing building related datasets, in: Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments, 2017, pp. 1–10.
[35] X. Feng, J. Lan, Z. Peng, Z. Huang, Q. Guo, A novel privacy protection framework for power generation data based on generative adversarial networks, in: 2019 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), IEEE, 2019, pp. 1–5.
[36] A.S. Khwaja, A. Anpalagan, M. Naeem, B. Venkatesh, Smart meter data obfuscation using correlated noise, IEEE Int. Things J. 7 (8) (2020) 7250–7264.
[37] P. Samarati, Protecting respondents identities in microdata release, IEEE Trans. Knowl. Data Eng. 13 (6) (2001) 1010–1027.
[38] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, Data Min. Knowl. Discov. 11 (2) (2005) 195–212.
[39] H. Liu, Q. Zhang, K. Guo, Y. Wu, Grey maximum distance to average vector based on quasi-identifier attribute, J. Grey Syst. 30 (1) (2018) 21–31.
[40] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A.M. Mezher, J. Parra-Arnau, J. Forné, The fast maximum distance to average vector (F-MDAV): an algorithm for k-anonymous microaggregation in big data, Eng. Appl. Artif. Intell. 90 (2020) 103531.
[41] M. Azadmanesh, B.S. Ghahfarokhi, M.A. Talouki, A white-box generator membership inference attack against generative models, in: 2021 18th International ISC Conference on Information Security and Cryptology (ISCISC), IEEE, 2021, pp. 13–17.
[42] J. Chen, M.A. Mohamed, U. Dampage, M. Rezaei, S.H. Salmen, S.A. Obaid, A. Annuk, A multi-layer security scheme for mitigating smart grid vulnerability against faults and cyber-attacks, Appl. Sci. 11 (21) (2021) 9972.
[43] K.S. Adewole, V. Torra, Privacy issues in smart grid data: from energy disaggregation to disclosure risk, in: International Conference on Database and Expert Systems Applications, Springer, 2022, pp. 71–84.
[44] J. Alarte Aleixandre, Application of clustering techniques to privacy protection, Thesis, Universitat Oberta de Catalunya, 2018.
[45] D. Murray, L. Stankovic, V. Stankovic, Refit: electrical load measurements (cleaned), https://pureportal.strath.ac.uk/en/datasets/refit-electrical-load-measurements-cleaned, 2016.