energies

MDPI

Article

Low-Frequency Non-Intrusive Load Monitoring of Electric Vehicles in Houses with Solar Generation: Generalisability and Transferability

Apostolos Vavouris *, Benjamin Garside, Lina Stankovic and Vladimir Stankovic

Department of Electronic and Electrical Engineering, Faculty of Engineering, University of Strathclyde, Glasgow G1 1XW, UK; benjamin.garside.2017@uni.strath.ac.uk (B.G.); lina.stankovic@strath.ac.uk (L.S.); vladimir.stankovic@strath.ac.uk (V.S.)

* Correspondence: apostolos.vavouris@strath.ac.uk

Abstract: Electrification of transportation is gaining traction as a viable alternative to vehicles that use fossil-fuelled internal combustion engines, which are responsible for a major part of carbon dioxide emissions. This global turn towards electrification of transportation is leading to an exponential energy and power demand, especially during late-afternoon and early-evening hours, that can lead to great challenges that electricity grids need to face. Therefore, accurate estimation of Electric Vehicle (EV) charging loads and time of use is of utmost importance for different participants in the electricity markets. In this paper, a scalable methodology for detecting, from smart meter data, household EV charging events and their load consumption with robust evaluation, is proposed. This is achieved via a classifier based on Random Decision Forests (RF) with load reconstruction via novel post-processing and a regression approach based on sequence-to-subsequence Deep Neural Network (DNN) with conditional Generative Adversarial Network (GAN). Emphasis is placed on the generalisability of the approaches over similar houses and cross-domain transferability to different geographical regions and different EV charging profiles, as this is a requirement of any real-case scenario. Lastly, the effectiveness of different performance and generalisation loss metrics is discussed. Both the RF classifier with load reconstruction and the DNN, based on the sequence-to-subsequence model, can accurately estimate the energy consumption of EV charging events in unseen houses at scale solely from household aggregate smart meter measurements at 1–15 min resolutions.

Keywords: non-intrusive load monitoring (NILM); energy dissagregation; electric vehicles (EVs); deep neural networks; transfer learning

Citation: Vavouris, A.; Garside, B.; Stankovic, L.; Stankovic, V. Low-Frequency Non-Intrusive Load Monitoring of Electric Vehicles in Houses with Solar Generation: Generalisability and Transferability. *Energies* **2022**, *15*, 2200. https://doi.org/10.3390/en15062200

Academic Editors: Dimitrios I. Doukas, Antonios Marinopoulos and Abu-Siada Ahmed

Received: 25 February 2022
Accepted: 15 March 2022
Published: 17 March 2022

1. Introduction

Decarbonisation of transportation is a major activity worldwide towards "securing global net zero by mid-century and keeping 1.5 degrees within reach" [1]. Transportation was responsible for 27% of the United Kingdom (UK)'s carbon dioxide (CO_2) emissions in 2019, of which over 90%, or 111 Mega-tonnes (Mt), of CO_2 were a product of road transport vehicles. Cars, including taxis, have played a major role in these emissions as, combined, they produced 68 Mt of CO_2, corresponding to 61% of the road transport vehicle emissions or a staggering 15% of UK's annuals emissions [2]. As transportation is an essential part of our daily activities, and therefore cannot easily be reduced, electric vehicles (EVs) are a promising solution to tackle this challenge. The short- to medium-term aim is to replace vehicles that run with internal combustion engines with electric ones, especially if the electricity is produced by renewable energy sources.

Driven by global climate change goals, and transition to net-zero economies, many governments world-wide have provided attractive incentives to EV users leading to a tremendous boom in EV purchase for domestic and business use. Indeed, according to

the International Energy Agency [3], the share of EVs on roads in 2020 has exceeded the 10 million mark—a jump from 7 million in 2019 [4].

With the increasing penetration of EVs in the market, power grids face great challenges regarding the ability to supply, transfer and distribute power. Indeed, the exponential increase of electric car sales—both plug-in hybrid and fuel cell, which are fuelled from the grid—requires major changes in the energy markets and grid infrastructure as electrification of transportation poses numerous challenges for the existing power networks. In particular, modelling shows that large-scale residential charging of EVs could result in overloading of distribution networks during peak hours if infrastructure upgrades and smart grid management are not implemented [5,6].

Understanding where and when EVs are charging is important for uptake modelling, supply planning, and grid infrastructure reinforcement [7]. Knowledge of EV charging patterns is also required for smart grid solutions such as Demand Response (DR) [8] and Vehicle-to-Grid (V2G) [9]. Future energy polices and transport planning would also benefit from accurate information on EV charging patterns. In addition, information about household consumption as a result of EV charging could be useful for customers to manage running costs of EVs in a similar manner to fuel costs for petrol and diesel cars. Awareness of the financial and carbon footprint implications of EV charging at home may incentivise customers to charge at public points and places of work, or opt-in to DR-based tariffs and V2G programmes, alleviating the overloading of distribution systems during peak hours as a result and maximising charging from solar feed-in.

There are two options for monitoring the presence of EV charging on power networks. The first is Intrusive Load Monitoring, a popular approach which sub-meters the EV charger at charge point, requiring additional hardware installation and maintenance costs; while this may benefit the end user and car manufacturer who have access to the EV's charging consumption statistics, the data are often not readily available to utilities, grid operators and network operators for infrastructure planning and grid demand management. A preferred alternative is Non-Intrusive Load Monitoring (NILM), from which energy consumption and time-of-use of the EV chargers is obtained via advanced signal information processing of aggregate power data, collected at a single point of measurement, e.g., a smart meter. The aim of NILM estimation of consumption—a regression problem—and time of use—a classification problem—algorithms is to match the equivalent submetered readings, therefore acting as a virtual sub-meter.

NILM has been an active area of study for over 30 years, but with the ongoing roll-out of millions of smart electricity meters globally, the deployment of large-scale residential NILM systems is emerging. Early work in the area often assumed the availability of mid- to-high frequency power measurements in the region of 1 Hz and above, as well as current and voltage measurements. However, due to storage limitations and potential privacy concerns, current real-world smart meter readings are only available at 15 to 60 min intervals, and providing only aggregate consumed power. The Smart Meter Equipment Technical Specifications Version:2 (SMETS2) framework in the UK, for example, permits regular smart meter readings to be taken at a 30 min resolution [10]. This motivates the need for for low- (1–60 s) to very low-resolution (15–60 min) NILM algorithms operating on power measurements only [11]. Recent years have seen an explosion of low-frequency NILM approaches, mostly based on Deep Neural Networks (DNNs). Indeed, according to [12], which provides a thorough literature review of DNN approaches for NILM, there were 87 DNN-NILM publications in the period 2018 to 2020. However, these DNN-NILM approaches focus primarily on typical household appliances, excluding EVs, and do not report results with meaningful performance metrics to truly evaluate consumption estimation. This is partly because of the limited availability of EV charging consumption datasets and generic, non-application-specific regression and classification metrics for evaluating DNN approaches for benchmarking.

Besides DNN approaches, the low-frequency NILM problem has also been tackled via other supervised and unsupervised approaches over the years, the former requiring training

on labelled data unlike the latter—an up-to-date review can also be found in [13]. Examples of supervised NILM approaches are Graph Signal Processing (GSP) approaches [14], Support Vector Machines (SVM) [15], Decision Trees (DTs) [16] and k-Nearest Neighbour (kNN) [17]. Some unsupervised approaches include Combinational Optimisation, unsupervised GSP [11], Hidden Markov Models (HMM) [18,19] and Dynamic Time Warping [16]. Unsupervised methods have the advantage of not being limited by appliances available in training data, but achieving good performance is challenging. Supervised approaches could equally be viable for practical large-scale deployment as long as sufficient labelled training data are available, and generalisability to similar unseen data and cross-domain transferability to other data can be demonstrated [20]. Most supervised approaches have mainly focused on NILM on seen houses and, more recently, unseen houses on the same dataset, and even fewer on cross-domain transferability [12,13].

The most popular electrical measurements' datasets on which NILM approaches are generally validated in the literature are REDD [21], UK-DALE [22], REFIT [23] and Pecan Street Dataport [24]—see [13] for some other examples of commonly used datasets, none of which include EVs. Only Dataport includes EV sub-metering and aggregate meter readings for multiple houses for a few months. A thorough review of available EV load datasets including charging point locations, historical and real-time charging sessions which refer to the period of time an EV is charged, traffic counts, travel surveys and registered vehicles, is presented in [25] in order to improve EV load modelling. However, none of the vehicle-centric data contain actual consumption readings from charging points, but rather spatial and temporal EV charging sessions to artificially reconstruct synthetic house-level and aggregated load consumption. This is not used in this study since synthetic loads do not reflect true consumption from the grid, and are not integrated into the household overall mains metering with other interfering loads and prosumers.

While NILM models have been developed for disaggregation of most conventional household appliances, NILM techniques for the disaggregation of EV loads is still an emerging area of study. Although, at first glance, EV load disaggregation may seem a relatively simple problem due to its high power level and being a single state load, houses nowadays use many electric devices with complex electrical signals and high energy consumption that make the separation of the EV signal a challenge. These include households with electric heaters, heat pumps, electric showers, Air Conditioning (AC) and Heating, Ventilation and Air Conditioning (HVAC) units, or prosumers, i.e., consumers that also produce electricity through solar panels and/or other renewable energy sources—tend to have quite complex load signals.

In this article, a detailed and robust methodology for large-scale robust evaluation of EV load disaggregation from household smart meter data are presented, leveraging on prior NILM algorithms that have been shown to have excellent classification performance, namely the Random Decision Forest (RF) classifier as used for EV load classification in [26], and the sequence-to-subsequence DNN of [27], which was shortlisted in [12]'s review paper to have one of the best regression performance on standard household appliances. Sequence-to-sequence and sequence-to-point DNN are used to perform a sequence transformation and therefore are appropriate for identifying electrical load signatures. The sequence-to-subsequence network was chosen as a trade-off between convergence speed of sequence-to-sequence, and computational load of sequence-to-point, as the proposed methodology should be both accurate and computationally efficient for scalability. The main contributions of this paper are:

- A critical review of the emerging literature on disaggregation of EV loads using supervised and unsupervised NILM;
- Implementation of RF bagging—based on [26]—and a suite of boosting-based ensemble algorithms, namely AdaBoost (ADA), XGBoost (XGB), Light Gradient-Boosting Machine (LGBM) and CatBoot, for binary classification of EV charging load;

- Novel, low-complexity post-processing steps for mitigating false positives arising due to high load interference, and for accurately estimating the EV load based on RF classification output, using time information;
- Adapting sequence-to-subsequence DNN-based NILM from [27] to EV load estimation, providing full details to enable reproducibility of the work including hyper-parameter tuning and post-processing steps;
- Evaluation on 15 real houses from two geographical regions in the USA from the Dataport dataset, with 1 and 15 min resolution data containing high power interference from AC and different EV load profiles, where the EV charging power, duration of EV charging events, sparsity of charging events, and the relative noise or interference from unknown loads that could negatively affect disaggregation performance, are calculated and reported for each house;
- Rigorous evaluation of the above NILM approaches, with a focus on creating realistic test scenarios including generalisability on unseen households with EVs with similar EV load profile from Austin, Texas and cross-domain transferability evaluation on unseen houses with a different EV load profile from New York;
- Quantifying generalisability and cross-domain transferability of the proposed methods by adapting metrics of [28,29];
- Evaluation of meaningfulness of standard and NILM-specific metrics and recommendations for EV load disaggregation for network operators.

For NILM evaluation, the most popular metrics include the standard $F-$score for classification and the standard Mean Square Error (MSE) or Mean Absolute Error (MAE) for regression, as well as more meaningful NILM specific metrics such as Accuracy (Acc) and Match Rate (MR). These and other NILM metrics are reviewed in [30]. The choice of the dataset and how challenging it is to disaggregate loads of interest can be measured through the noisiness of the dataset metric of [28]. Additional metrics to calculate the generalisation loss that occurs when testing a NILM model on unseen houses is described in [29]. Given the potential impact of residential EV charging on the smart grid and the benefits of NILM for EV charging consumption and time-of-use for network operators and energy consumers, this paper will present and discuss results using all the above metrics.

The rest of the paper is organised as follows. In Section 2, supervised and unsupervised NILM approaches that are specifically designed for EVs are reviewed. In Section 3, a rigorous approach to evaluate the generalisability and transferability of ensemble methods and a sequence-to-subsequence DNN for classification of time of use of EVs and consumption estimation is proposed. In Section 4, the proposed methodology is evaluated using generic classification and regression metrics, as well as NILM specific consumption metrics. This is followed by Section 5, where observations are discussed in detail in relation to EV load estimation to inform grid demand. Lastly, conclusions are summarised in Section 6.

2. Background on NILM Approaches for EV Charging

This section, is focused on unsupervised and supervised NILM approaches developed and evaluated specifically for EV load disaggregation. For general NILM algorithms, we refer readers to recent review papers [12,13,30].

2.1. EV Load Disaggregation-Unsupervised Approaches

To the best of our knowledge, load disaggregation of EV charging is first tackled in [31], where a training-free approach, based on time-series signal thresholding, filtering and denoising, is proposed that uses knowledge of known appliance signatures to remove contributions from other loads and estimate power consumption of EVs. The approach is validated with over a year of 1 min data from Dataport [24] between 2012 and 2013, across 11 houses, randomly picked out of hundred of houses from Austin area. Monthly consumption error and MSE were used to evaluate the performance of the method and results were benchmarked against the HMM algorithm of [18]. Results outperformed HMM, which had difficulty distinguishing between EV loads and AC "spike trains", which

becomes particularly challenging in the summer months. However, the calculation of error in terms of monthly consumption is not as rigorous as the *Acc* and MR that have emerged more recently and are often calculated based on daily consumption estimates [11,30]. The authors of [31] do not provide the IDs of the households that were used, and therefore the results cannot be reproduced and compared.

Another unsupervised approach is proposed in [32], where Independent Component Analysis (ICA) is used to extract EV loads from aggregated signals. This is followed by a series of complex processing steps to remove interference from appliances with similar load characteristics and rebuild an estimated EV load profile. Validation of the method is carried out on 1-min Dataport [24] from 34 houses, and on 5 min resolution samples, obtained by re-sampling the measured 1 min readings. Results were evaluated using EV load reconstruction error, calculated sample-by-sample and a modified $F-$score that takes Accurate/Inaccurate True Positives into account as used in [33]. However, when tested on 5 min resolution data performance was significantly reduced. As with [31], the authors of [32], do not provide the IDs of the households that were used, hence the results cannot be reproduced.

2.2. EV Load Disaggregation-Supervised Approaches

In [17], a Mean Sliding Window algorithm is used to detect and extract features from ON/OFF events, i.e., an appliance switched ON and OFF—which are subsequently classified as AC and EV charging, using a kNN classifier. The method is validated on 1 min data collected by Dataport [24], from June to August 2014. The classifier was trained on 15 days of data collected at house 26 and tested on 4 days from the same house achieving $F-$scores of 83% and 91% for EV charging ON and OFF events, but $F-$scores fell to 86% and 75%, respectively, for 5 min data. Generalisability to unseen house 3036 in the Dataport dataset was attempted using a pre-trained model but optimal k-values were chosen based on misclassification error rate for each house individually; this requires labelled data for both houses and therefore fails to fully test generalisability to house 3036. Although classification results are promising, a testing period of only 8 days from 2 houses is inadequate to fully evaluate the effectiveness of the method. It is also unclear how the test days and houses were chosen. No energy consumption estimations were calculated from the classifications and hence no consumption-based metrics were used for evaluation.

Another low-complexity supervised method is proposed in [26] where active power data are split into overlapping windows that are fed into an RF classifier. Principal Component Analysis (PCA) is used for feature extraction. Once again, the method is validated on 1 min data from Dataport [24]—6 houses were considered over the period January 2016 to December 2017. The data for each house were split into 10-minute overlapping windows and used directly as input to the RF classifier achieving an $F-$score of 92.61%. In [26], PCA is applied to the windows and all 10 principal components (PCs) are used, resulting in a reported improvement in classification performance. However, the $F-$score was only changed by 0.08% which is far from a significant increase. The authors discuss the use of PCA for removing redundant information and show that over 95% of the variance in their dataset is explained by 2 PCs, but no attempt is made to reduce dimensionality. It is also unclear whether PCA was applied to the train and test datasets separately, which is important for ensuring that no bias is imparted on the training data through implicit knowledge of the test data. The application of PCA to the windows resulted in a small reduction in false negatives with marginally improved $F-$score of 92.69%. According to [26], this outperforms the ICA unsupervised approach of [32]. However, the direct comparison with [26] is hard to make for two reasons. Firstly, the RF classifier is given a balanced dataset for testing, i.e., 50% of windows contain EV charging and 50% only contain other household appliances which is achieved by random under-sampling. This does not represent the real proportion of EV charging vs. non-charging windows, which is reported to be 6% in the initial, unbalanced dataset. As a result, it is not demonstrated how robust the classifier is against false positives that may arise from interference from other

large loads. Secondly, labelled data from all 6 houses were used for training, and therefore generalisability on unseen houses is not demonstrated.

Building upon [26], in [34] RF is evaluated alongside kNN and Artificial Neural Network (ANN) approaches for EV load disaggregation. Models are trained and tested with a selection of 18 houses for a period of a month from Pecan Dataport [24]—however, which houses are used exactly for training and testing is not specified. Therefore, the results cannot be reproduced or compared. In the pre-processing step, since only one month was considered with insufficient EV charging events, the authors simulated additional EV load charging patterns instead of using real data from other days. Training and testing sets were created by selecting only one month of data from selected houses that included both EVs and PVs. Generic classification and regression metrics are presented, without taking into account NILM-specific metrics such as MR or *Acc* [30]. The RF model outperformed the other two models, with results presented only for two set-ups—classification ($F-$score = 93% and 75% for training and testing on a selection of houses and testing on one unseen house, respectively) and regression (MAE = 500 W and 630 W for training and testing on a selection of houses and testing on one unseen house, respectively).

2.3. Summary of State-of-the-Art

In summary, there is limited previous work on EV classification and load consumption estimation using a range of signal information processing methods. The main gaps are the lack of transparency in reporting sufficient details—such as specifics and number of houses and days used for training and testing—for reproducing and comparing results, lack of transparency in the choice of experimental data—including quality and quantity metrics. Furthermore, performance evaluation in current literature tends to be non-rigorous, especially on generalisability and cross-domain transferability of the methods, which is needed for practical deployment. This paper addresses all the aforementioned gaps, and proposes a novel rigorous methodology for EV load disaggregation and evaluation, leveraging upon supervised ensemble and DNN-based approaches that have already been shown to outperform other learning approaches for NILM classification and regression.

3. Methodology

The proposed methodology is described in order, except for Sections 3.4 and 3.5, where either ensemble classification with load reconstruction or regression based on DNN can be chosen, before proceeding with generalisability and cross domain transferability evaluation.

3.1. Experimental Data Selection and Preparation

The data acquisition process for the development of a supervised NILM methodology involves the selection and preparation of a dataset with aggregate and sub-metered appliance power measurements, sampled at low to very-low sampling rates, constituting "labelled data" for algorithm training and testing. As discussed previously, the Dataport [24] dataset has been primarily used for EV NILM evaluation in the past, as it contains many households in different areas of the USA with different appliances, including EVs, that have been monitored continuously for a long period of time. The data portal was accessed through a University research account, providing free access to 1 s, 1 and 15 min data collected from 73 houses across Austin, Texas, California and New York. Of these 73 houses, 8 from Austin, 7 from New York and 0 from California were listed as owning EVs. Available data for Austin houses span over a period of 12 months—from 1 January 2018 until 31 December 2018—and for New York houses over a period of 6 months—from 1 May 2019 until 31 October 2019.

Table 1 represents a summary of the Dataport houses used for the experiments, including metadata on the presence of Air Conditioning (AC) and solar generation, and the total amount of EV charging time per household as well as the sparsity of EV load. The latter information, i.e., the total amount of EV charging time and the sparsity of EV load—that is rarely stated in the literature—but provides an indication of the amount of

data available for training and testing. Houses 2335, 3517 and 5058 were omitted from Table 1 as either their EV sub-metering was found to be null or contained no EV charging activity. From the remaining houses, 3 were discarded as the data were faulty and/or scarce: House 3000 appeared to have erroneous data as the addition of mains reading, i.e., the amount of energy consumed in the household, and solar power generation, which is either consumed or fed back into the grid if the production is greater than usage, did not add up to the grid reading, i.e., the total energy equilibrium that is apparent from the connection point of the house to the grid; House 7719 included only 71 h of charging events throughout a period of 12 months, which was equivalent to only 10 activations of a charge and insufficient to produce meaningful results; House 9053 had noisy sub-metering, i.e., there were unusually long periods where the load pattern did not resemble a typical EV charging signal, probably because there was another load metered on that plug. Lastly, house 4767 appears to have changed the EV charger from 4 kW to 6.6 kW after 3 July 2018 and therefore, in the experimental results later on, results are presented for 4767-1 and 4767-2, to represent House 4767, before and after this change of the charge, respectively.

All 9 selected houses have solar panels and all but one have high power AC interference, making the NILM task challenging. The whole dataset of each of these 9 houses is used for the experimental tests, so that different energy usage patterns can be observed across different seasons of the year. Austin houses contain data that spread across all seasons in a year, whereas New York houses contain data from late Spring until early Autumn.

Table 1. Summary of EV charging in Dataport houses used for the experiments, including noisiness metrics [28] of the considered households in dataport Dataset. EV Sparsity is calculated as the charging duration divided by the total monitoring duration.

Area	House	EV Charging Power [kW]	Charging Duration [h]	EV Sparsity	AC/Solar	$NM^{(T)}$	$NM^{(EV)}$
Austin	661	3.3	781	8.92%	Yes/Yes	86.35%	39.34%
	1642	3.3	982	11.21%	Yes/Yes	82.37%	39.19%
	4373	3.3	1359	15.51%	Yes/Yes	75.91%	44.48%
	4767-1	4.0	485	10.98%	Yes/Yes	86.60%	36.49%
	4767-2	6.6	485	11.16%	Yes/Yes	85.65%	26.12%
	6139	3.3	622	7.10%	Yes/Yes	90.50%	40.31%
	8156	3.3	615	7.02%	Yes/Yes	90.67%	47.34%
NY	27	3.3	338	7.65%	Yes/Yes	74.99%	21.15%
	1222	6.6	139	3.15%	No/Yes	82.86%	25.54%
	5679	6.6	210	4.76%	Yes/Yes	76.76%	30.67%

3.2. Quantifying Interference

NILM is a source separation problem, where any consumption measurement other than target loads of interest are considered as interfering signals, or noise. Therefore, the noisier a dataset, i.e., the more unknown or non-submetered loads, the more challenging the classification and disaggregation problem, directly impacting the accuracy. To quantify the difficulty of successfully estimating individual loads from aggregate, Ref. [28] introduced a "noisiness" measure (*NM*) given by Equation (1):

$$\%-NM^{(T)}=\frac{\sum_{t=1}^{T}\left|y_t-\sum_{m=1}^{M}y_t^{(m)}\right|}{\sum_{t=1}^{T}y_t}, \tag{1}$$

where T is the total monitoring duration—in the number of samples—y_t is the aggregated load measured at time sample t and $y_t^{(m)}$ is the submetered measurement of load/appliance m at time sample t. The above measure assumes that there is an equal interest in estimating all M "targeted"/submetered loads. In the case discussed, and in Table 1, the noisiness metric for $M=1$ is presented where the only m of interest is the EV, and denotes $y_t^{(1)}=y_t^{(EV)}$. Since this research is interested only in disaggregating EV loads, all other loads contributing

to the aggregate would be considered as noise. Therefore, Equation (1) i slightly revised such that the noise, i.e., unknown loads, are only considered during EV charging times, to capture better their interfering effect:

$$\%-NM^{(EV)}=\sum_{t=1}^{T}c_t\left|1-\frac{y_t^{(EV)}}{y_t}\right|, \tag{2}$$

where an indicator $c_t = 1$ if EV charging is ON during time sample t, i.e., $y_t^{(EV)} > 0$, and zero, otherwise.

Table 1 includes the noise metrics $\%-NM^{(T)}$ and $\%-NM^{(EV)}$ from Equations (1) and (2), respectively, for all the households under consideration. The higher the noise metric the more interference from unknown loads, and therefore the more challenging to accurately estimate energy consumption. It can be observed that a lower $\%-NM^{(T)}$ metric does not always imply a lower $\%-NM^{(EV)}$ metric. Both metrics will be reviewed in an attempt to explain classification and regression performance.

3.3. Train–Test Split

The entire dataset is split into train and test datasets at the pre-processing stage. A rigorous approach to split training and testing data is proposed where a small number of days are randomly selected from each month to make up the test dataset. The number of days selected from each month is set such to obtain a train–test split ratio of around 70:30, resulting in 10 days of each month kept for testing purposes. Days are chosen at random to guarantee a natural distribution of EV charging vs. non-charging windows, and to demonstrate that days have not been "hand-picked". Selecting the same number of days from each month ensures that the method is tested equally across all seasonal variations in solar generation and appliance use, e.g., AC in summer, furnace in winter. Leaving the test windows in-order creates a more realistic simulation of a real-world NILM system and allows for complete EV loads to be reconstructed for visualisation and evaluation of performance in terms of consumption metrics. After the test dataset has been formed, the remaining windows are randomly under-sampled—by removing windows with no EV charging—to obtain a balanced train dataset that is randomly ordered to ensure no bias in the training of the classifier. This process is summarised in Figure 1 and repeated for each house.

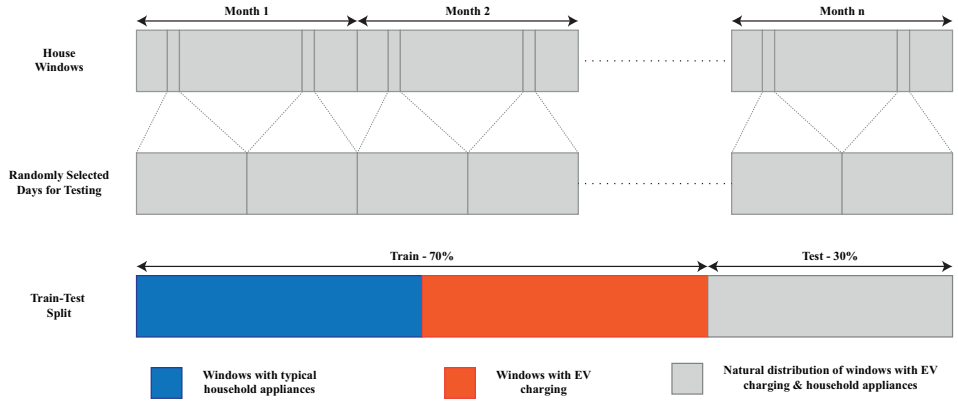

Figure 1. Visual demonstration of the proposed method for obtaining a balanced train dataset and a realistic in-order test dataset for each house, where n is the number of months in the data collection period for a given house. Note that in the train dataset, the blue block contains periods without EV charging and the orange block is formed from the periods with EV charging and other households appliances running in parallel.

3.4. Classification and Load Estimation via Ensemble Learning

Classification is performed by first adapting an RF classifier using sliding windows of raw power data and performing feature extraction using PCA, as per [26] for validation and comparison. RF is a bagging ensemble algorithm comprising multiple unpruned DTs, each trained on a randomly selected subset of the input features, and making a binary classification decision—EV charging or EV disconnected—based on majority voting or averaging of decision tree decisions. Four popular ensemble algorithms are also implemented based on boosting, which operate by combining many weak learners, e.g., a single decision tree—into strong learners, to determine the best ensemble-based classifier for EV NILM. These are ADA based on canonical boosting, and three gradient-boosting algorithms, namely XGB, LGBM and CatBoot.

3.4.1. Sliding Window Process

A sliding window is used to split the data into small, overlapping windows suitable for training and testing of the classifier. There are two variables that can be adjusted: window size, ω, and overlap, δ. Given the size of the windows ω and the overlap δ, both expressed in the number of samples, the number of windows, N, over a period of T [samples], is calculated as:

$$N(T) = \frac{T - \omega}{\omega - \delta} + 1, \tag{3}$$

where all variables are given in samples to generalise for 1 and 15 min data.

It is expected that increasing window size ω leads to improved robustness against interference from other large loads, due to relatively long duration of EV charging sessions. For example, some AC systems operate at a similar power to EV charging, but in general, AC spikes have a width of less than 30 min, compared to EV charges which typically last between 30 and 200 min [31]. A window that is larger than the typical duration of spikes caused by AC or other large loads is more likely to be correctly classified as "EV disconnected" in this circumstance, and therefore a reduced number of false positives would be expected. However, this comes at the cost of potential loss in accuracy when reconstructing the EV load profile as the exact start and end times of charging would harder to determine. Similarly, increasing the overlap is also expected to improve classification performance due to better monitoring of every transition between windows. An improvement may also be observed simply due to the fact that more windows would be generated for training the classifier according to Equation (3)—although the increased size of train and test sets will of course come with the cost of increased running times.

In the results section, the window size ω and overlap δ are varied experimentally, and final values are chosen that optimise performance for 1 and 15 min disaggregation.

3.4.2. Post-Processing

Given classification outcomes from the ensemble classifier, a low-complexity correction algorithm is proposed to reduce overall classification errors using knowledge about expected characteristics of EV charging loads. Since EV charging duration typically spans 30 to 200 min, a series of EV Charging classifications, i.e., EV charging is ON—for a duration of less than 30 min is likely to be false positives caused by interference from another large load such as AC. Any series of possible misclassifications that meet these criteria would be corrected to the class EV Disconnected by the proposed algorithm.

Similarly, a charging EV is unlikely to be unplugged and then reconnected within a short period of time. Therefore, a short series of EV Disconnected classifications in the middle of a charging block would be reclassified as EV Charging. The correction algorithm is described below and applied to each test day, where a state-change refers to a change from an EV Disconnected classification to EV Charging or vice versa, and the correction window is the maximum length of Charging/Disconnected blocks for which corrections will be applied. The steps of the correction algorithm are:

1. Iterate through the original classifications and record each state-change that is observed as well as the time index at which it occurs;
2. As every state change is observed, compare the current time index to the previous recorded index;
3. If the difference is less than or equal to the correction window, overwrite all the classifications between the previous and current time indices with the current state, remove the record of the previous state change, and continue.

The algorithm is designed to be computationally efficient and would be able to apply corrections as data are being collected in a real-world system, with a small delay equal to the size of the correction window. This would still therefore allow near real-time observations in a NILM system for smart grid management applications. The exact values for the correction window are optimised for 1 and 15 min data through a series of tests in the results section to follow.

3.4.3. EV Load Reconstruction Following Classification

Using the corrected classifications, as described in Section 3.4.2, that show when each EV charging started and ended, an EV load profile can be reconstructed if the EV model, or simply the charging power of the EV at a given house, is known. For the purposes of this study, the charging power of EVs at each house is estimated as shown in Table 1. In a real-world NILM application this information could be obtained through an appliance survey, a widely proposed concept in NILM research that would involve requesting appliance information from customers, or from manufacturer manual.

Estimated energy consumption over a period of T samples is given by Equation (4) as:

$$E_{est} = \frac{S}{60} \sum_{n=1}^{N(T)} C_n p(\omega - \delta) \quad [\text{kWh}], \tag{4}$$

where S is the sampling interval in [mins], i.e., $S \in \{1, 15\}$ mins, C_n is the classification outcome for the nth window, $C_n \in \{0, 1\}$—EV not charging/charging—p is the charging power of the EV expressed in kW (3.3, 4, or 6.6 in this case), ω is the window size and δ is the overlap—in the number of samples.

True energy consumption over the period of T samples is given by Equation (5) as:

$$E_{true} = \frac{S}{60} \sum_{t=1}^{T} P_t \quad [\text{kWh}], \tag{5}$$

where P_t is the power in kW, at time sample t of the EV ground truth signal. Both E_{est} and E_{true} are calculated for each day in the test dataset such that $T = \frac{1440}{S}$, the number of samples in a day, and $N(T)$ is the number of overlapping sliding windows per day computed as in Equation (3).

Besides total consumption, the number of charges is also recorded by counting the blocks of windows classified as EV charging. Each time that the end of a charge is detected, a small correction is made to E_{est} to account for the slightly over-estimation that would otherwise occur due to the classification of overlapping windows. Using the classifications and E_{est} directly, the consumption for each charge would be over-estimated by $\frac{S}{60} p \times \delta$, so this value is subtracted from E_{est}, every time the end of a charging block is detected.

3.5. Regression Based on DNN

Sequence-to-subsequence learning is adapted and optimised, with conditional Generative Adversarial Network (GAN), using publicly available code [35], to disaggregate EV loads. Sequence-to-subsequence network targets the middle part of a sequence, and therefore a shorter sequence compared to sequence-to-sequence DNN, making convergence faster. Additionally, since the network targets a subsequence instead of a point, as is the case with the sequence-to-point DNN architectures [36,37], training is faster and less

computationally expensive. For the disaggregation algorithm to work successfully, it is essential that the whole EV charging event is included in the targeted subsequence of the DNN. Thus, the optimal sequence-to-subsequence window size ω is set as:

$$\omega \gtrsim \frac{2 \times L}{S}, \omega \in \{2^0, 2^1, ..., 2^n, ...\}, \tag{6}$$

where L [in mins] is the usual length of EV charging period, and S [in mins] is again the resolution of the data samples—1 or 15 min. The above equation is proposed based on the following criteria: the subsequence has a width that is equal to half of the window size and the window size must be a power of 2—a requirement of the DNN architecture. Note that different window sizes will be used on 1 and 15 min data. The epoch number—the number of times that the learning algorithm worked through the entire training dataset—was also varied, and the number of epochs that maximise performance was used using early stopping criterion on the validation set. In addition, other hyperparameters, including the number of generator filters as well as the number of discriminator filters, were tuned using the validation set, and depend on the sampling rate. Lastly, the number of layers was adapted based on the window size. The selected values are reported in Section 4.

Given the produced sub-sequences of EV load, a simple correction procedure is applied. As with the classification discussed earlier, it was observed that the sequence-to-subsequence model produced some false positive results, mostly due to interference from other similar loads. All of these false positive sub-sequences had one feature in common: their maximum value appeared to be very small, around 5 W. Therefore, these values are simply zeroed. Additionally, negative values that were produced from the DNN were also zeroed as negative values have no physical significance.

3.6. Performance Evaluation Metrics

To evaluate classification performance, $Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}$, $Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$, $Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and $F-\mathrm{score} = \frac{2\times\mathrm{TP}}{2\times\mathrm{TP+FP+FN}}$ are used, where TP are the True Positives, TN are the True Negatives, FP are the False Positives and FN are the False Negatives. All these metrics are calculated on a sample-by-sample basis, e.g., TP is the number of samples when EV charging is used that were classified as EV changing. $F-$score captures both false positives and false negatives, while giving a more rigorous evaluation than *Accuracy*, particularly for datasets with a significant imbalance. For example, suppose data are split into windows for a given house and only 10% of windows are labelled as containing EV charging. A model that classifies every window as EV Disconnected would achieve an *Accuracy* of 90% but an $F-$score of 0%.

Typical regression metrics MAE (Equation (7)) and normalised Signal Aggregate Error (SAE) (Equation (8)) are used for evaluation of the network performance. Additionally, the following two NILM-specific metrics are used to estimate how well the load estimation from both classification and regression learning approaches estimate the energy consumed by EV charging: *Acc*—sometimes referred to as Total Energy Correctly Assigned (TECA) (Equation (9)) and MR (Equation (10)).

$$\mathrm{MAE} = \frac{\sum_{d=1}^{D} |E_{est_d} - E_{true_d}|}{D}, \tag{7}$$

$$\mathrm{SAE} = \frac{\left|\sum_{d=1}^{D} E_{est_d} - \sum_{d=1}^{D} E_{true_d}\right|}{\sum_{d=1}^{D} E_{true_d}}, \tag{8}$$

$$Acc = 1 - \frac{\sum_{d=1}^{D} |E_{est_d} - E_{true_d}|}{2 \times \sum_{d=1}^{D} E_{true_d}}, \tag{9}$$

$$\mathrm{MR} = \frac{\sum_{d=1}^{D} \min\{E_{est_d}, E_{true_d}\}}{\sum_{d=1}^{D} \max\{E_{est_d}, E_{true_d}\}}, \tag{10}$$

where E_{est_d} and E_{true_d} refer to estimated and ground truth consumption for day d, and D is the number of days in the test dataset. MR is generally considered to be a better load estimation metric [30] for the same reason $F-$score is considered a better measure of performance than classification *Accuracy*, as it can better indicate the match between the estimated and the true energy. When compared to MAE, MR is more robust and deviates less when an experiment is repeated. On the other hand, the MAE metric calculates the average of errors, i.e., a large error for one subsequence or point would significantly affect the value.

3.7. Generalisability and Transferability Evaluation

Since labeled data are scarce, hard to collect, and thus only available for a small portion of households, it is essential that the proposed methods are able to produce reliable results on unseen houses in a similar dataset—generalisability—and unseen houses, i.e., houses without any labelled data, from another domain—transferability. To this effect, generalisation loss is proposed in [29] to evaluate the performance of a NILM algorithm for both event detection from classification and load estimation on an unseen house. For event detection, Generalisation Loss, as a percentage, is given by Equation (11)—a comparison between the classification accuracy on unseen (Acc_{u}) and seen houses (Acc_{s}), where *Acc* can be the *Accuracy* or $F-$score metric. Similarly, for regression accuracy, Generalisation Loss, as a percentage, is given by Equation (12)—a comparison between the error on the unseen houses ($\mathrm{ERR_u}$) and seen houses ($\mathrm{ERR_s}$)—where ERR can be any of the standard regression metrics such as MAE, SAE, Root Mean Square Error (RMSE).

$$G_{\mathrm{loss}}^{\mathrm{class}} = \left(1 - \frac{Acc_{\mathrm{u}}}{Acc_{\mathrm{s}}}\right)\%, \tag{11}$$

$$G_{\mathrm{loss}}^{\mathrm{reg}} = \left(\frac{\mathrm{ERR_u}}{\mathrm{ERR_s}} - 1\right)\%. \tag{12}$$

For the experimental evaluation of generalisation loss of classification performance, $Acc = F-$score is used. For consumption accuracy, however, the use of Equation (13) instead of Equation (12) is proposed. This approach is preferred as a more accurate measurement of performance of consumption estimation, and is calculated based on the *Acc* metric (Equation (9)).

$$G_{\mathrm{loss}}^{\mathrm{energy}} = \left(1 - \frac{Acc_{\mathrm{u}}}{Acc_{\mathrm{s}}}\right)\%. \tag{13}$$

The same Generalisation Loss equations can be used to evaluate generalisability and transferability. In the former case, unseen houses would be those from similar datasets. In the case cross-domain transferability evaluation, unseen houses would be from another domain.

4. Experimental Results

This section, first describes the evaluation strategy, then present the experimental results for classification, regression and generalisability/transferability for EV NILM as per Sections 3.4, 3.5, and 3.7, respectively.

4.1. Evaluation Strategy

Firstly, to validate correct implementation of the RF classification methodology proposed in [26], the balanced test conditions were replicated. Following that, the additional stages of the methodology described in Section 3 are used for EV disaggregation from a realistic test dataset, with a natural imbalance, in-order windows and randomly selected days. The training set is split 60% for training, and 10% for cross-validation to determine the best set of parameters, which are then fixed for final results of testing on an unseen 30% of the samples. As explained in Section 3.3, the testing set used for presented results below comprised a total of 120 days for each Austin house and 60 days for each New York House

when testing and training on the same household. For generalisability and cross domain transferability tests, all data available from unseen houses, i.e., 12 months for each Austin house and 6 months for each New York house—were used as testing sets. Repeatability of experiments is also performed by repeating testing on the same conditions and data ten times.

Simulations were carried out such that classifiers are trained and tested on the same houses individually to observe how accurately EV consumption can be estimated under these ideal conditions. Generalisability tests were then carried out to fully evaluate both NILM approaches under conditions required for implementation in a real-world NILM system. This procedure involved training classifiers and regression models on a selected number of houses and testing on unseen houses belonging to the same geographic area and similar EV load signatures. Transferability tests were carried out on unseen houses in a different geographical area and different EV load signatures.

4.2. Classification and Load Reconstruction via Ensemble Algorithms: Experimental Setup

In this section, the classification results as well as results of load reconstruction based on EV detection events, obtained via ensemble algorithms, are discussed.

4.2.1. Replicating Balanced Train–Test Conditions

The RF approach of [26] is first replicated for validation, using the same experimental conditions reported. Note that [26] did not specify the house IDs used in the reported experimental results and therefore results cannot be fully reproduced. Data from all Dataport houses from Table 1 were split into ω =10 min windows with $\delta = 5$ min overlap and windows were randomly under-sampled to obtain balanced train and test datasets. A total of around 900,000 windows were generated, reduced to 144,000 after under-sampling and split into 115,000 windows for training and 29,000 for testing. In the first experiment, windows were used directly as feature input to the classifier.

For the next experiment, feature extraction via PCA on the same raw windows was carried out and all resulting 10 PCs were fed to the classifier as features, as per [26]. The same experiments were repeated with the effects of solar feed-in artificially removed by subtracting from the self-consumption—demand+solar—by the sub-metered PV generation data. All results were compared to results from [26], as shown in Table 2.

Table 2. Classification results achieved when RF classifier is trained and tested on balanced datasets, with comparison to results reported in [26]—raw windows (Raw) & PCs as features. Note that the houses used for training and testing are different since house IDs were not provided in [26].

Experimental Setup	*Accuracy* (Raw, PCs)	*Precision* (Raw, PCs)	*Recall* (Raw, PCs)	$F-$Score (Raw, PCs)
Solar	87.92%, 84.41%	87.86%, 83.16%	88.20%, 85.79%	88.02%, 84.45%
No-Solar	90.70%, 87.10%	89.22%, 84.37%	92.96%, 91.62%	91.05%, 87.85%
[26]	92.58%, 92.34%	92.33%, 92.66%	92.88%, 93.03%	92.61%, 92.69%

The RF model was built with 500 DTs estimators and *Entropy* was chosen for the splitting criterion to match hyperparameters selected in [26]. However, as it is found in [38], the criterion does not affect splitting performance in a meaningful way, so *Gini Impurity* was chosen for all results that follow, since the calculation of a logarithm is avoided. Bootstrapping is set to be *True*, and the maximum features considered by each DT is chosen to be the square root of the total feature set.

As can be seen from Table 2, overall, the results are comparable to those achieved in [26], where a larger dataset—2 years of data compared to 1 year—was used and a greater number of windows were therefore generated for training and testing. Removing the contributions of solar generation improved the performance—an expected improvement as with the removal of solar generation the stochastic element of PV production that may interfere with the load signal is eliminated—resulted in a near identical $F-$score when raw

data windows were used directly as input to the classifier. This may indicate that solar generation was also removed from—or simply not present in—the dataset used by [26]; however, small differences could equally be explained by the larger volume of training data or the investigation of different houses from Pecan Street Dataport [24]. Contrary to the findings of [26], the use of PCA reduced the $F-$score achieved under both test conditions. Once again, this might be the result of differences in the datasets used. The effectiveness of PCA is further investigated so as to decide whether the feature projection stage should be used as part of the methodology whatsoever. The shortcomings of this balanced testing strategy have been discussed in Section 2, and therefore the results presented above are only included for validation purposes. In the following, tests on a realistic unbalanced dataset are conducted.

4.2.2. Window Size Optimisation

The effect of varying window size, ω, on performance was heuristically studied. As ω is increased, classification performance—measure by $F-$score—gradually improves, while regression performance—measured by MR—increases at first but quickly converges and starts to decrease as window size increases. This behaviour is expected, as an increased window size constitutes more "features" for the classifier, and therefore improved robustness is achieved against interference from other loads which may operate at a similar power to EV charging but typically over shorter periods of times. These load characteristics may be learned by the RF classifier when larger windows are used, resulting in better classifications. Conversely, larger windows result in more ambiguity when determining start and end times for EV charging, and subsequent consumption estimates—regression performance—are less accurate.

Similarly, the effect of window overlap, δ, was heuristically studied. It was observed that both, classification and regression metrics were improved as overlap was increased. When overlap was maximised, the exact transition between every window is monitored and one window is generated for every minute of data; resulting classification, correction and EV load reconstruction are therefore more accurate. As discussed previously, part of the improvement in performance might also be a result of more training data for the classifier, since the number of windows in a given time period increases with overlap according to Equation (3). This of course comes with the cost of increase running times, but the additional computation is deemed to be acceptable given the subsequent improvements in performance and the low-complexity nature of the methodology as a whole.

To strike a balance between a larger feature set for the classifier and good time-resolution for reconstructing EV load, a window size of $\omega = 5$ and $\delta = 4$ are selected for 15 min data, and $\omega = 30$ and $\delta = 29$ samples for 1 min data.

4.2.3. PCA

For all houses in the dataset, and for both sampling rates, the $\omega = 5$- or 30-dimensional input data could be reduced to 2 dimensions with no significant impact on performance as two most significant PCs combined explained more than 90% variance. Indeed, increasing the number of PCs above 2 resulted in little or no improvement in classification performance. To assess the performance with PCA, the RF classifier described in the previous section is used, where PCA was applied to the same windows and all PCs were used to train and test the classifier.

It is clear from Figure 2 that PCA significantly reduces performance in terms of $F-$score and MR at both sampling rates when compared to using the windows directly. These findings are contrary to those in [26], where a marginal improvement is reported as a result of PCA, however, this was observed under balanced test conditions, i.e., a non-realistic scenario. The results above show that PCA has quite the opposite effect when testing on the naturally unbalanced data—mostly due to an increase in false positives. Clearly, good results in a balanced scenario do not necessarily translate to good performance on unbalanced data, especially when the imbalance is significant—in this case EV charging

windows make up only 5–15% of total windows in dataset, as indicated by the sparsity highlighted in Table 1.

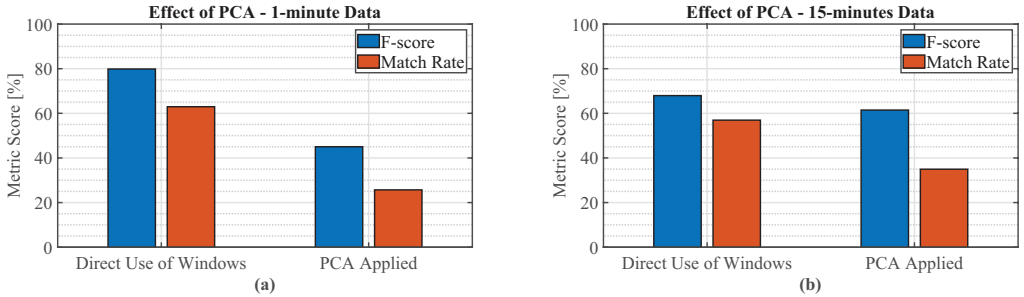

Figure 2. Effect of PCA on performance in terms of $F-$score and MR (**a**) 1-minute data. (**b**) 15 min data.

4.2.4. Classification Probability Threshold

When a trained DT ensemble classifier is given test data to classify, each input is evaluated by every DT in the model and all classifications can be averaged to calculate a probability for a sample to belong to the given class. By default, the threshold for making a classification decision is a probability greater than 0.5. In this experiment the probability threshold was gradually increased up to a maximum of 0.95 to observe the effect on performance. A higher probability is expected to decrease the number of FP, but also increase FN. At both data resolutions, a probability threshold of 0.7 appears to be the optimal balance between FPs and FNs, resulting in a 10% improvement in MR for 1 min data and over a 5% increase for 15 min data compared to the default value of 0.5. So, the value of 0.7 is selected for the following experiments.

4.2.5. Correction Window

As discussed in Section 3.4.2, any two state changes that occur in a time smaller than the correction window are corrected to the current state. It was experimentally verified that the optimal correction window length is 20-samples for 1 min granularity and 2 samples, i.e., 30 min—for 15 min data. Intuitively, these values seem reasonable given the goal of the correction algorithm—to remove short periods of false positives and false negatives—and knowledge of typical EV charge times—30 to 200 min. A correction window greater than 30 min may remove short EV charges or merge separate charges, resulting in under-or over-estimation of consumption, respectively.

4.2.6. Comparison with Other DT Ensemble Algorithms

The performance of RF was compared against other boosting ensemble algorithms, namely ADA, XGB, LGBM and CatBoost as well as a single DT. Figure 3 presents both $F-$score and MR of these algorithms, where classifiers were trained and tested on house 1642, as per the 70:30 train:test split described in Section 3.3. House 1642 was chosen because it had the second largest number of EV charging hours (see Table 1), providing sufficient data samples for training and testing. All classifiers used 500 estimators.

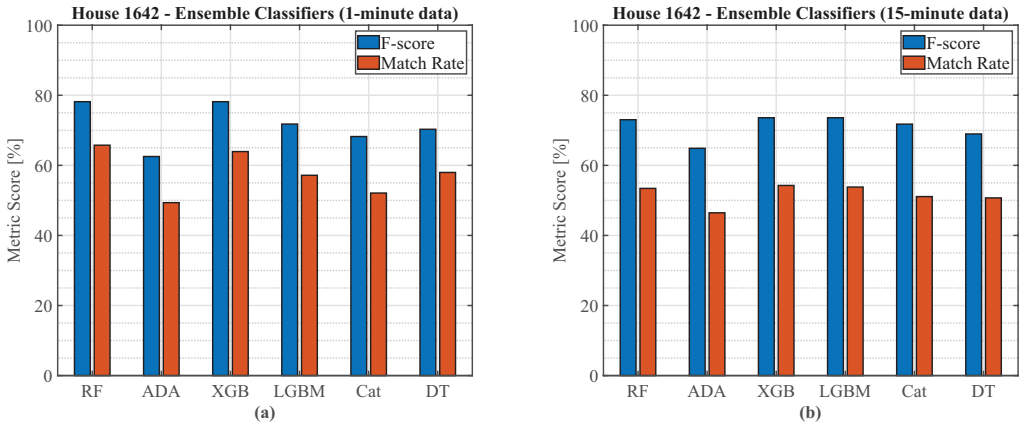

Figure 3. Ensemble classifiers $F-$score and MR performance on house 1642 for (**a**) 1 min data; (**b**) 15 min data.

The same ensemble algorithms were evaluated for transferability to observe how classifiers perform when tested on unseen houses (Figure 4). The same classifiers, trained on house 1642, were now tested on all remaining houses from Table 1 to find overall $F-$scores and MR. Transferability tests results are shown in Figure 4, and indicate a drop in $F-$score and MR performance across all algorithms compared to Figure 3.

As expected, all ensemble algorithms outperform the DT algorithm. Overall performance for all algorithms is slightly poorer on 15 min data compared to 1 min data. The two gradient-boosting algorithms, especially XGB, have similar performance to RF.

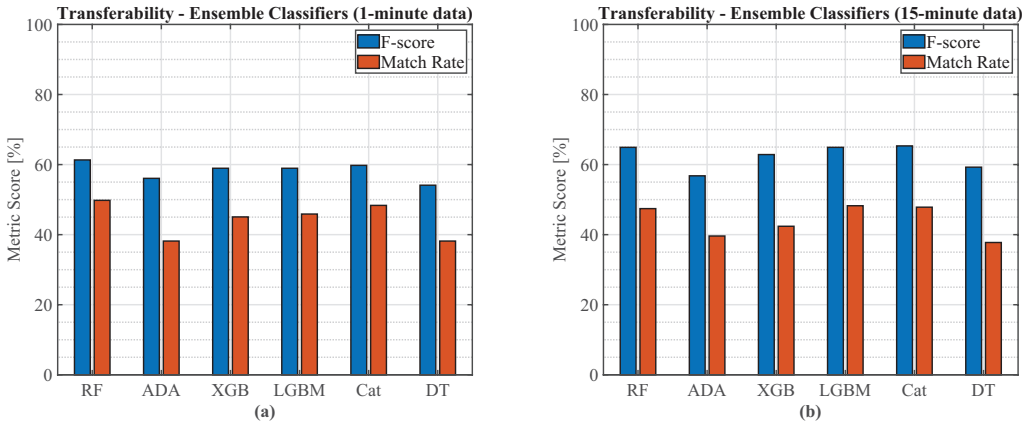

Figure 4. Ensemble classifiers $F-$score and MR performance trained on data from house 1642 and tested on all other houses for (**a**) 1 min data; (**b**) 15 min data.

4.3. Regression–Sequence-to-Subsequence: Simulation Setup

For the sequence-to-subsequence with conditional GAN network, the window size is set to 512 and 64 samples, for 1 and 15 min data, respectively, as these values led to the best performance on the validation set. L1 (Least Absolute Deviation) loss was used in both setups, whereas Stochastic Gradient Descent (SGD) and ADAM optimiser were used for discriminator and generator filters, respectively. The initial learning rate was 0.001 (for SGD) and 0.0005 (for ADAM). The momentum term of ADAM was equal to 0.5, and the weights on L1 and GAN term for the generator gradient were 100 and 1, respectively. For 1 min data, a total of 7 layers were used and thirty-two generator and discriminator filters in the first convolutional layer. For 15 min data, a total of 5 layers were used and 4 generator and discriminator filters in the first convolutional layer. Using early stopping criterion on the validation set, it was concluded that the best value for the number of epochs is 120.

4.4. Simulation Results

In this section, results are presented using the optimised experimental settings described in Sections 4.2 and 4.3 for the RF classifier and sequence-to-subsequence DNN, respectively.

First, classification and regression models are trained and tested on the same house, referred to as "observed" scenario, to demonstrate the ideal performance of the methodology under realistic test conditions, i.e., randomly selected days of in-order windows that have not been balanced. Secondly, generalisability and transferability are evaluated on unseen houses that are not part of the training set. The metrics that are presented, are the product of ten independent executions of the same algorithm to ensure experimental repeatability. Both the mean value and the standard deviation of each metric are presented, in order to further underline the meaningfulness of each metric. A metric with a low standard deviation is more robust to randomness that is introduced due to the random initialisation of algorithms. Therefore, these metrics are considered to produce much more concise results as they present a more accurate performance of the models.

4.4.1. Performance: Observed Scenario

Tables 3 and 4 present the results obtained by training and testing on the same household for 1 min resolution for RF-based classification and load reconstruction, and DNN-based regression, respectively. The standard deviation for all metrics over 10 runs of the experiment, in identical conditions, is less than 2%, which indicates that all the experiments are repeatable. However, in Table 4 it is observed that the MAE metric has a significantly larger standard deviation of 10–15%. This is a common observation of DNN-based regression networks and therefore why presenting results exclusively with the MAE metric, which is rife in recent NILM literature, can often be misleading especially if repeatability of experiments is not demonstrated.

Table 3. Assessment metrics, training and testing on the same household, using RF classifier and load reconstruction, 1 min data.

		Classification Metrics		**Consumption Metrics**	
Area	**House**	***Accuracy***	***F*−Score**	***Acc***	**MR**
Austin	661	92.32%	73.91%	82.54%	71.07%
	1642	94.61%	81.35%	84.94%	72.22%
	4373	95.30%	86.73%	92.24%	85.47%
	4767	95.72%	85.96%	78.63%	57.96%
	6139	94.74%	67.95%	75.03%	61.17%
	8156	94.71%	76.21%	77.26%	62.68%
NY	27	94.10%	77.41%	80.39%	61.45%
	1222	97.74%	87.92%	88.50%	80.78%
	5679	99.21%	96.41%	98.76%	97.55%

Comparing *Acc* and MR metrics between load reconstruction from classification vs. load estimation via regression, it can be seen, as expected, that the DNN network is more susceptible than RF to insufficient samples in the training set, as exemplified by house 1222 results, which has the fewest EV charging hours. For the same reason and the fact that house 4767 changed its EV charger, with a significantly different wattage, in the middle of the year—as discussed in Section 3.1—the regression network also performs relatively poorly. Otherwise, the *Acc* and MR performance measures are consistent for EV load estimation from both RF classification and load reconstruction with post-processing, and from the regression approach. Furthermore, House 6139 has relatively poorer classification and load estimation performance compared to others because it has the highest noisiness metric value as shown in Table 1. Classification and load reconstruction was also performed on 15 min data and a summary of the produced results is presented in Table 5. Results

obtained when training on 15 min data are overall poorer, but not much so, than results with 1 min data, for all metrics, as expected.

Table 4. Mean values and standard deviations of assessment metrics, training and testing on the same household, using sequence-to-subsequence, 1 min data.

Area	House	MAE (μ, σ)	SAE (μ, σ)	Acc (μ, σ)	MR (μ, σ)
Austin	661	(45.45 W, 6.28 W)	(8.41%, 0.25%)	(89.09%, 1.51%)	(79.55%, 1.24%)
	1642	(55.13 W, 11.52 W)	(4.19%, 0.14%)	(88.87%, 0.74%)	(79.59%, 0.79%)
	4373	(82.33 W, 14.09 W)	(1.72%, 0.07%)	(89.57%, 1.11%)	(80.94%, 0.98%)
	4767	(210.6 W, 41.86 W)	(5.47%, 0.27%)	(58.19%, 3.57%)	(48.52%, 1.89%)
	4767-1	(219.1 W, 48.55 W)	(59.94%, 0.88%)	(46.84%, 2.59%)	(41.22%, 1.26%)
	4767-2	(490.2 W, 74.57 W)	(83.36%, 0.86%)	(30.98%, 2.10%)	(24.78%, 0.97%)
	6139	(103.2 W, 13.97 W)	(4.15%, 0.11%)	(70.36%, 1.47%)	(53.53%, 1.05%)
	8156	(82.33 W, 14.09 W)	(1.72%, 0.06%)	(89.57%, 1.10%)	(80.94%, 0.98%)
NY	27	(86.78 W, 19.74 W)	(3.47%, 0.08%)	(79.14%, 0.49%)	(74.38%, 1.14%)
	1222	(148.1 W, 28.17 W)	(5.73%, 0.07%)	(53.27%, 1.97%)	(45.11%, 1.47%)
	5679	(73.65 W, 6.75 W)	(16.22%, 0.72%)	(85.29%, 1.72%)	(79.51%, 1.18%)

Table 5. Assessment metrics, training and testing on the same household, using RF classifier and load reconstruction, 15 min data.

		Classification Metrics		Consumption Metrics	
Area	House	$Accuracy$	$F-$Score	Acc	MR
Austin	661	90.52%	69.16%	76.50%	62.84%
	1642	93.58%	83.17%	77.70%	65.16%
	4373	91.67%	81.27%	82.40%	71.49%
	4767	95.61%	86.04%	72.00%	50.49%
	6139	86.44%	50.93%	49.25%	44.98%
	8156	91.91%	70.93%	57.50%	45.41%
NY	27	95.83%	84.84%	85.60%	71.53%
	1222	94.47%	74.04%	70.66%	61.09%
	5679	97.80%	91.34%	88.77%	81.05%

Table 6 presents results produced from training and testing on the same household on 15 min data for the DNN-based, sequence-to-subsequence regression network. Using 15 min granularity data, as the available activation windows were fewer—approx. 7% less activation windows than in the 1 min data—training and testing on the same household produced poorer results. This is a result of the combination of fewer activation windows and lower data granularity. As discussed previously, DNN networks are more susceptible to the number of training samples than RF.

Table 6. Mean values and standard deviations of assessment metrics, training and testing on the same household, using sequence-to-subsequence, 15 min data.

Area	House	MAE (μ, σ)	SAE (μ, σ)	Acc (μ, σ)	MR (μ, σ)
Austin	661	(100.6 W, 4.52 W)	(22.67%, 0.14%)	(75.47%, 1.11%)	(56.65%, 0.87%)
	1642	(143.0 W, 6.81 W)	(23.66%, 0.75%)	(69.12%, 1.22%)	(58.12%, 0.88%)
	4373	(317.2 W, 23.61 W)	(71.81%, 2.49%)	(60.22%, 0.74%)	(23.41%, 1.44%)
	4767	(281.0 W, 21.57 W)	(23.15%, 0.18%)	(41.31%, 0.09%)	(31.06%, 0.54%)
	6139	(156.5 W, 3.98 W)	(18.34%, 1.41%)	(54.58%, 1.01%)	(33.33%, 0.25%)
	8156	(150.1 W, 4.77 W)	(28.42%, 1.29%)	(59.27%, 1.05%)	(35.62%, 0.09%)
NY	27	(95.57 W, 4.73 W)	(30.46%, 0.42%)	(75.48%, 0.97%)	(64.91%, 0.57%)
	1222	(173.0 W, 11.50 W)	(22.15%, 1.16%)	(29.58%, 0.92%)	(29.58%, 0.54%)
	5679	(169.8 W, 11.26 W)	(23.38%, 1.09%)	(63.25%, 1.53%)	(54.58%, 1.89%)

Table 7 demonstrates the loss incurred in classification and consumption estimation performance by reducing the granularity of meter readings from 1 min to 15 min. Loss was calculated as: $Loss = \left(1 - \frac{F-\text{score}_{15\text{min}}}{F-\text{score}_{1\text{min}}}\right)\%$ and $Loss = \left(1 - \frac{Acc_{15\text{min}}}{Acc_{1\text{min}}}\right)\%$ for classification and consumption estimation, respectively. As expected, the loss in $F-$score, i.e., accuracy of detecting EV load charging events, is very small except for House 6139, which had a relatively high noise metric—as per Table 1—and House 1222, which had insufficient training. Energy estimation using RF classification and load reconstruction is more robust to lower frequency data when compared to sequence-to-subsequence regression. As it was already stated, this can be the result of multiple factors, mainly due to insufficient activation windows when training the sequence-to-subsequence algorithm and also due to the fact that compared to the load reconstruction method using RF classification results, sequence-to-subsequence network does not have a priori knowledge of the EV charging level. Additionally, the loss in energy estimation performance for both approaches is correlated with the noisiness metrics as per Table 1, where Houses 6139, 8156 from Austin area and House 1222 from New York area, have relatively higher noisiness metrics and therefore greater losses in performance when using lower frequency data.

Table 7. Classification and regression granularity loss based on $F-$score and Acc metrics, respectively, using RF classifier and load reconstruction and sequence-to-subsequence (seq2subseq) network.

Area	House	Classification (RF)	Load Reconstruction (RF)	Regression (seq2subseq)
Austin	661	6.43%	7.32%	15.29%
	1642	−2.24%	8.52%	22.22%
	4373	6.3%	10.67%	32.77%
	4767	−0.09%	8.43%	29.61%
	6139	25.05%	36.25%	22.43%
	8156	6.93%	25.58%	33.83%
NY	27	−9.60%	−6.48%	4.62%
	1222	15.10%	20.16%	44.47%
	5679	5.26%	10.12%	25.84%

4.4.2. Generalisability Results

As with any real-world NILM scenario, the proposed solutions should be able to transfer knowledge from known houses to unknown ones that belong to the same area or, in this case, they use a similar EV charging load. This is essential, as the collection of metadata and/or labeled data for all households is a costly and time consuming process, and end-users are not always keen on sharing their personal information. Generalisability evaluation performs testing in houses, not included in the training set, that:

- belong to the same geographical area as the houses in the training set;
- use the same EV charging level as houses in the training set.

Table 8 presents results obtained by testing in the same area as training houses using the sequence-to-subsequence network. Each of the experiments was performed by testing on all houses in an area apart from one, which was kept for testing purposes. For example, results presented in Table 8 for Austin House 661, were obtained by training on the entire period for all Austin houses except House 661, and testing on the entire period of House 661. Results of Table 8 are compared against those of Table 4, in order to see the difference in performance when testing on an observed house vs. testing on an unseen house in a similar geographic area, namely Austin or NY. This is also captured via the $G_{\text{loss}}^{\text{reg}}$ and $G_{\text{loss}}^{\text{energy}}$ metrics of Equations (12) and (13), respectively, which indicate the MAE and Acc loss in performance. The two loss metrics are generally in agreement in terms of relative performance, except for the NY houses. The $G_{\text{loss}}^{\text{reg}}$ of House 27 is unusually high because it is the only house in NY with a 3.3 kW EV, and the regression network was trained on the other two houses with 6.6 kW EVs, and therefore the energy consumption is overestimated.

However, this is less pronounced in the $G_{\text{loss}}^{\text{energy}}$ metric. Overall, the performance loss is negligible across all metrics, except for the marginal drop in performance for Austin Houses 661, 4373 and 8156. This is captured by the positive G-loss for these values which are less than 15%. Houses 1222 and 5679 experience a more significant drop in performance, as captured by both $G_{\text{loss}}^{\text{reg}}$ and $G_{\text{loss}}^{\text{energy}}$, because they are both trained on House 27, which has about 50% more EV load charging events at 3.3 kW and therefore the energy consumption is underestimated for these two houses with 6.6 kW EVs.

Furthermore, as captured by the large negative G-loss values, Houses 4767-1 and 4767-2 now have significantly improved EV load estimation performance because the issue of insufficient training data previously encountered has been resolved with training on all other houses.

Table 8. Mean values of assessment metrics and $G-$loss, training on all houses of an area apart from one, and testing on the unseen house from the same area, using sequence-to-subsequence algorithm, 1 min data.

Area	House	MAE	SAE	*Acc*	MR	$G_{\text{loss}}^{\text{reg}}$	$G_{\text{loss}}^{\text{energy}}$
Austin	661	73.54 W	6.29%	81.79%	70.00%	61.8%	8.19%
	1642	52.26 W	1.02%	89.46%	80.84%	−5.21%	−0.66%
	4373	99.16 W	4.33%	87.86%	77.92%	20.44%	1.88%
	4767	198.8 W	29.92%	63.60%	40.05%	5.60%	−9.3%
	4767-1	72.47 W	9.09%	82.48%	71.29%	−66.92%	−76.09%
	4767-2	278.6 W	50.9%	58.92%	28.94%	−43.17%	−90.19%
	6139	95.68 W	6.94%	71.61%	54.56%	−7.29%	−1.78%
	8156	97.08 W	12.63%	76.31%	63.56%	17.92%	14.8%
NY	27	177.8 W	4.37%	63.62%	47.49%	104.9%	19.61%
	1222	189.3 W	58.42%	36.56%	34.14%	27.82%	31.37%
	5679	110.6 W	2.21%	76.56%	61.69%	50.17%	10.24%

Tables 9 and 10 present the results of generalisability tests for training on all houses with 3.3 kW loads regardless of geographical area, except House 1642, and testing on unseen House 1642, for both 1 and 15 min resolutions for the RF and sequence-to-subsequence DNN approaches, respectively. Comparing both Tables 4 and 6 with Table 10 for the regression network, and as indicated by $G_{\text{loss}}^{\text{reg}}$ and $G_{\text{loss}}^{\text{energy}}$, it can be seen that while the 1 min results are similar on observed and unseen scenarios, there is a significant improvement in performance for the 15 min results due the availability of additional training data from multiple houses. Similarly, comparing Tables 3 and 5 with Table 9, and as indicated by $G_{\text{loss}}^{\text{class}}$ and $G_{\text{loss}}^{\text{energy}}$, insignificant change in classification and energy estimation performance for both 1 min and 15 min resolutions is observed, since as it was discussed before, RF is less susceptible to the amount of training data. it can therefore be concluded that both the RF and regression network results are generalisable, without loss of performance, for similar EV charging levels.

Table 9. Mean values of performance and generalisation metrics, training on all houses that were charging on 3.3 kW, and testing on unseen House 1642, using RF classification and load reconstruction, with 1 min and 15 min granularity data.

		Classification Metrics			Consumption Metrics		
Granularity	House	*Accuracy*	$F-$Score	$G_{\text{loss}}^{\text{class}}$	*Acc*	MR	$G_{\text{loss}}^{\text{energy}}$
1 min	1642	95.17%	77.48%	4.76%	84.21%	72.31%	0.86%
15 min	1642	94.27%	78.36%	5.78%	78.38%	63.48%	−0.88%

Table 10. Mean values of assessment metrics and $G-$loss, training on all houses that were charging on 3.3 kW, and testing on unseen house 1642, using sequence-to-subsequence algorithm, with 1 min and 15 min granularity data.

Granularity	House	MAE	SAE	*Acc*	MR	$G_{\text{loss}}^{\text{reg}}$	$G_{\text{loss}}^{\text{energy}}$
1 min	1642	50.41 W	6.70%	89.83%	80.96%	−8.56%	−1.08%
15 min	1642	76.30 W	14.97%	84.60%	71.47%	−46.64%	−22.4%

4.4.3. Transferability Results

In this subsection, evaluation of cross-domain transferability to assess how robust a model is to training and testing on different geographical area and different EV charging levels, is aimed. Thus, transferability tests can be summarised as follows:

1. Testing on an unseen house in NY and training on all other houses from Austin, regardless of EV charging level;
2. Testing on an unseen house with and EV charge level of 6.6 kW and training on all houses with EV charge level of 3.3 kW, regardless of geographical area;
3. Testing on two unseen houses and training on a generic mix of houses from different areas and different EV charging levels.

Tables 11 and 12 show the outcome for transferability tests 1 and 2, for both 1- and 15 min resolutions, with the RF and sequence-to-subsequence network approach, respectively. House 5679 from NY with an EV charging level of 6.6 kW was tested on RF and DNN regression models trained with all Austin houses containing EVs with 3.3 kW charging level. The training set comprised Houses 661, 1642, 4373, 6139, and 8156.

Table 11. Mean values of performance and generalisation loss metrics for transferability tests 1 and 2, using RF classification and load reconstruction, with 1 min and 15 min data.

		Classification Metrics			Consumption Metrics		
Granularity	House	*Accuracy*	$F-$Score	$G_{\text{loss}}^{\text{class}}$	*Acc*	MR	$G_{\text{loss}}^{\text{energy}}$
1 min	5679	98.54%	86.30%	−11.38%	90.37%	82.96%	−7.32%
15 min	5679	95.45%	73.39%	6.35%	64.44%	57.41%	17.79%

Table 12. Mean values of performance and generalisation loss metrics for transferability tests 1 and 2, using sequence-to-subsequence algorithm, for 1 min and 15 min data.

Granularity	House	MAE	SAE	*Acc*	MR	$G_{\text{loss}}^{\text{reg}}$	$G_{\text{loss}}^{\text{energy}}$
1 min	5679	183.4 W	24.32%	61.15%	58.67%	149.0%	28.30%
15 min	5679	201.5 W	29.60%	56.92%	48.83%	18.67%	10.01%

Comparing Tables 3 and 5 with Table 11, a slight increase in classification and energy consumption estimation performance on 1 min data is observed, as indicated by $G_{\text{loss}}^{\text{class}}$ and $G_{\text{loss}}^{\text{energy}}$ which could be a result of more data available to train the RF classifier since House 5679 had relatively fewer EV charging hours. On the contrary, for 15 min data granularity, there is a larger drop in $G_{\text{loss}}^{\text{energy}}$ than $G_{\text{loss}}^{\text{class}}$, i.e., consumption vs. classification performance. This could indicate that a lower granularity of training data from different geographical area and EV charging level does impede transferability. Comparing Tables 4 and 6 with Table 12, and captured by $G_{\text{loss}}^{\text{energy}}$ than $G_{\text{loss}}^{\text{reg}}$, a drop in load estimation performance for both granularities is observed. The drop is more pronounced for 1 min granularity. Interestingly, from the regression network's output load reconstruction plots, it is observed that while the sequence-to-subsequence algorithm is correctly detecting EV charging events, it underestimates the EV load charging level since the network was trained on lower EV charge loads.

Similarly to Table 7, Table 13 demonstrates the loss introduced by using data with granularity of 15 min compared to 1 min. $F-$score and *Acc* metrics were used for classification and regression problems, respectively. As expected, the granularity loss is more pronounced during transferability than in the observed scenario—see House 5679 in Table 7. The regression network is less affected by reduced granularity when directly compared to RF classification and load reconstruction, as observed in Table 13.

Table 13. Classification and regression granularity loss based on $F-$score and *Acc* metrics, respectively, using RF classifier and load reconstruction and sequence-to-subsequence for transferability tests 1 and 2.

Area	House	Classification (RF)	Load Reconstruction (RF)	Regression (seq2subseq)
NY	5679	14.96%	28.69%	6.92%

Finally, a practical approach, as per transferability test 3, was taken whereby generic learning models were trained using a mix of houses across different geographic areas and contained different EV charging loads. The training set comprises Houses 661, 4373, 4767, 6139, 8156, 27, and 1222. Testing was performed on unseen Austin House 1642 and NY House 5679, with EV charge loads of 3.3 kW and 6.6 kW, respectively. Results are presented in Tables 14 and 15. As previously, results of Tables 3 and 5 were compared with Table 14 and a very small drop in classification $F-$score performance during transfer learning for House 1642 was observed. More importantly, consumption metrics *Acc* and MR were unaffected. This is not the case with House 5679 where both classification and consumption metrics are lower. These are also captured by the $G_{\text{loss}}^{\text{energy}}$ and $G_{\text{loss}}^{\text{class}}$ metrics, which are higher than the equivalent loss for transferability tests 1 and 2 for House 5679. This could be because there were more 3.3 kW than 6.6 kW EV charge loads in the training dataset, which makes the classifier more transferable or biased to 3.3 kW EV charge loads, as is the case with House 1642. Comparing Tables 4 and 6 with Table 15, significant drop in performance is observed when testing on House 5679. This could be the result of more houses that are charging in 3.3 kW in the training set compared to 6.6 kW. On the other hand, results for 15 min data on House 1642 are significantly improved, which is a result of more data available to the network during training process. The $G_{\text{loss}}^{\text{energy}}$ than $G_{\text{loss}}^{\text{reg}}$ loss metrics for this transferability test 3—as shown in Table 15—compared to transferability tests 1 and 2, as shown in Table 12, are relatively unchanged for 1 min granularity but there is less loss for 15 min granularity. This shows that the sequence to subsequence regression model performs equally well on all transferability tests.

Table 14. Mean values of performance and generalisation loss metrics, for transferability test 3, using RF classification and load reconstruction, for 1 min and 15 min data.

		Classification Metrics			Consumption Metrics		
Granularity	**House**	***Accuracy***	***F*−Score**	$G_{\text{loss}}^{\text{class}}$	***Acc***	**MR**	$G_{\text{loss}}^{\text{energy}}$
1 min	1642	95.15%	77.24%	5.05%	84.05%	71.91%	1.05%
	5679	98.65%	87.40%	9.35%	91.17%	84.40%	7.69%
15 min	1642	94.26%	78.19%	5.99%	78.40%	63.16%	−0.9%
	5679	95.57%	74.10%	18.87%	65.17%	58.23%	26.59%

Table 15. Mean values of performance and generalisation loss metrics, for transferability test 3, using sequence-to-subsequence algorithm, for 1 min and 15 min data.

Granularity	House	MAE	SAE	*Acc*	MR	$G_{\text{loss}}^{\text{reg}}$	$G_{\text{loss}}^{\text{energy}}$
1 min	1642	50.49 W	4.54%	89.81%	81.12%	−8.42%	−1.06%
	5679	178.4 W	26.94%	62.21%	49.21%	142.23%	27.06%
15 min	1642	78.66 W	11.69%	84.13%	71.15%	−44.99%	−21.72%
	5679	183.6 W	33.00%	60.75%	46.04%	8.13%	3.95%

This experiment demonstrates that if an adequate number of houses of a certain wattage level are included in the training set, then when testing on an unseen house that uses a same power level charger, the model is agnostic to the other EV power levels that are presented in the training set, and produces an accurate result.

Table 16 demonstrates the loss introduced by using data with granularity of 15 min compared to 1 min. $F-$score and *Acc* metrics were used for classification and regression problems, respectively.

Table 16. Classification and regression granularity loss based on $F-$score and *Acc* metrics, respectively, using RF classifier and load reconstruction and sequence-to-subsequence for transferability test 3.

Area	House	Classification (RF)	Load Reconstruction (RF)	Regression (seq2subseq)
AU	1642	−1.24%	6.72%	6.32%
NY	5679	15.22%	28.52%	2.35%

5. Discussion

During data processing and algorithm tuning, it was observed that, in the presence of houses with solar panels, it was better to extract EV load charge events without solar generation. EV load signatures have a distinctly high power levels, and therefore the drop in amplitude caused by solar generation is insufficient to completely obfuscate the EV signal. In the Dataport [24] houses considered in the study, EVs were connected to the grid in the evening and night hours, when solar generation is either very low or non-existent. This pattern agrees with the daily routines, as people tend to use their vehicles to commute to work during morning and afternoon—when solar generation is at its peak. It is therefore worth exploring the possibility of storage of energy produced during the daytime and use that energy later so as to charge EVs and help reducing grid peaks that usually occur in the late afternoon/evening.

$F-$score is more suitable to evaluate how accurately EV charging events are detected compared to metrics like *Accuracy*, especially for realistic, unbalanced testing datasets. Complementary metrics for measuring the accuracy in estimating the load consumption of the EV charging events are *Acc* and MR , whilst MAE and SAE can explain the performance of regression networks. Similarly, generalisation loss as a metric based on $F-$score, *Acc* and MAE, provide a good representation of performance loss of these measures due changes in granularity of the meter readings, as well as due to generalisability to unseen houses in a similar geographic area and with similar EV charging loads, and transferability to unseen houses in different geographic areas and with different EV charging loads.

The ensemble classification models were more robust to insufficient EV charging events for training than the DNN-based regression models. That is, the sequence-to-subsequence DNN is especially sensitive to the amount of training samples, which takes precedence over the noisiness level of the house due to interfering loads. Otherwise, it was observed that both ensemble-based classification and regression models accurately performed EV load estimation on the same house the models were trained on, as well as showed excellent generalisability performance when tested on unseen houses for similar EV charging levels in different geographic areas. During generalisability and transferability experiments, it

was observed that the regression network is less affected by lower granularity readings than the RF classification and load reconstruction approach. The proposed final recommendation for EV charging event detection, as well as accurate energy consumption for each charging event, is therefore a sequence-to-subsequence DNN, when plenty of training data are available in a mixed mode approach, with data from different geographic areas, and especially with a balanced number of EV charging load levels to avoid bias towards a particular EV charging level.

6. Conclusions

This paper demonstrated the feasibility of accurately estimating EV load charging consumption at scale by energy providers, using only smart meter measurements at resolutions of 1 min and 15 min. Specifically, several classifier approaches based on ensemble learning, including RF and gradient-boosting algorithms, together with regression networks based on deep learning sequence-to-subsequence architecture with conditional GAN were evaluated. The RF model of [26], initially proposed for classification of EV charging events, were replicated with a realistic train–test data split and unbalanced, realistic testing datasets, and supplemented with a proposed post-processing step for minimising false positives and load reconstruction. In addition, the regression network of [27], showing good disaggregation performance on typical household appliances, excluding EVs, was replicated for EV disaggregation. Evaluation was carried out for three scenarios: (i) training and testing on different portions of an observed house—observed scenario, (ii) generalisability across houses with similar geographic area and EV load charge, and (iii) cross-domain transferability in unseen houses from different geographic areas and different EV load charge levels. The merits of typically used classification, regression and NILM-specific energy consumption metrics were presented for all experiments and discussed, in conjunction with generalisation loss metrics and noisiness metrics, which are an indicator of unknown loads interfering with the EV load in the aggregate meter readings.

Further work is needed to evaluate cross domain transferability across distinct geographical areas, including Europe and Asia, as well as in datasets with different EV wattage levels. This is quite a challenging domain as EV load datasets are scarce in the literature, and useful next direction is to make more EV submetering or labelled datasets publicly available as EV roll-out is increasing. In addition, three-phased EV loads—that are common on continental Europe—should also be explored and adaptations of the proposed methodology should be performed. Lastly, as EV charging consumes a substantial amount of energy, and there is a degree of flexibility in charging times, unsupervised EV load forecasting should be researched, as it can facilitate novel DR solutions that avoid localised load shedding and high load appliances being cut off from the power supply via smart switches that are embedded in the smart grid.

Author Contributions: Conceptualisation, L.S.; methodology, software, validation, A.V. and B.G.; formal analysis, A.V. and B.G.; resources, A.V. and B.G.; writing—original draft preparation, A.V.; writing—review and editing, L.S. and V.S.; visualisation, A.V. and B.G.; supervision, L.S.; funding acquisition, V.S. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analysed in this study. These data can be found here by creating a free account: https://dataport.pecanstreet.org (accessed on 20 October 2021).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AC	Air Conditioning
ADA	AdaBoost
ANN	Artificial Neural Network
CO_2	Carbon dioxide
DNN	Deep Neural Network
DR	Demand Response
DT	Decision Trees
EV	Electric Vehicle
GAN	Generative Adversarial Network
GSP	Graph Signal Processing
HMM	Hidden Markov Model
HVAC	Heating, Ventilation and Air Conditioning
ICA	Independent Component Analysis
kNN	k-Nearest Neighbour
LGBM	Light Gradient-Boosting Machine
MAE	Mean Absolute Error
MR	Match Rate
MSE	Mean Squared Error
NILM	Non-Intrusive Load Monitoring
NM	Noisiness Metric
PC	Principal component
PCA	Principal component analysis
RF	Random Decision Forests
RMSE	Root Mean Square Error
SAE	Signal Aggregate Error
SGD	Stochastic Gradient Descent
SMETS2	Smart Meter Equipment Technical Specifications Version:2
SVM	Support Vector Machines
TECA	Total Energy Correctly Assigned
V2G	Vehicle-to-Grid
XGB	XGBoost

References

1. UN Climate Change Conference (COP26) at the SEC—Glasgow 2021. Available online: https://ukcop26.org (accessed on 22 January 2022).
2. Transport and Environment Statistics: Autumn 2021. Available online: https://www.gov.uk/government/statistics/transport-and-environment-statistics-autumn-2021/transport-and-environment-statistics-autumn-2021 (accessed on 14 February 2022).
3. IEA. Trends and Developments in Electric Vehicle Markets—Global EV Outlook 2021—Analysis-IEA. Available online: https://www.iea.org/reports/global-ev-outlook-2021/trends-and-developments-in-electric-vehicle-markets (accessed on 22 January 2022).
4. Global Electric Passenger Car Stock, 2010–2020—Charts—Data & Statistics. Available online: https://www.iea.org/data-and-statistics/charts/global-electric-passenger-car-stock-2010-2020 (accessed on 14 February 2022).
5. Alquthami, T.; Alsubaie, A.; Alkhraijah, M.; Alqahtani, K.; Alshahrani, S.; Anwar, M. Investigating the Impact of Electric Vehicles Demand on the Distribution Network. *Energies* **2022**, *15*, 1180. [CrossRef]
6. Leou, R.-C.; Su, C.-L.; Lu, C.-N. Stochastic Analyses of Electric Vehicle Charging Impacts on Distribution Network. *IEEE Trans. Power Syst.* **2013**, *29*, 1055–1063. [CrossRef]
7. Yang, J.; Long, X.; Pan, X.; Wu, F.; Zhan, X.; Lin, Y. Electric Vehicle Charging Load Forecasting Model Considering Road Network-Power Grid Information. In Proceedings of the 2019 International Conference on Technologies and Policies in Electric Power & Energy, Yogyakarta, Indonesia, 21–22 October 2019; pp. 1–5. [CrossRef]
8. Afzalan, M.; Jazizadeh, F. A Machine Learning Framework to Infer Time-of-Use of Flexible Loads: Resident Behavior Learning for Demand Response. *IEEE Access* **2020**, *8*, 111718–111730. [CrossRef]
9. Khosrojerdi, F.; Taheri, S.; Taheri, H.; Pouresmaeil, E. Integration of electric vehicles into a smart power grid: A technical review. In Proceedings of the 2016 IEEE Electrical Power and Energy Conference (EPEC), Ottawa, ON, Canada, 12–14 October 2016; pp. 1–6. [CrossRef]

10. Smart Metering Equipment Technical Specifications: Second Version. Available online: https://www.gov.uk/government/consultations/smart-metering-equipment-technical-specifications-second-version (accessed on 14 February 2022).
11. Zhao, B.; Ye, M.; Stankovic, L.; Stankovic, V. Non-intrusive load disaggregation solutions for very low-rate smart meter data. *Appl. Energy* **2020**, *268*, 114949. [CrossRef]
12. Huber, P.; Calatroni, A.; Rumsch, A.; Paice, A. Review on Deep Neural Networks Applied to Low-Frequency NILM. *Energies* **2021**, *14*, 2390. [CrossRef]
13. Angelis, G.F.; Timplalexis, C.; Krinidis, S.; Ioannidis, D.; Tzovaras, D. NILM Applications: Literature review of learning approaches, recent developments and challenges. *Energy Build.* **2022**, *13*, 111951. [CrossRef]
14. He, K.; Stankovic, L.; Liao, J.; Stankovic, V. Non-Intrusive Load Disaggregation Using Graph Signal Processing. *IEEE Trans. Smart Grid* **2016**, *9*, 1739–1747. [CrossRef]
15. Basu, K.; Debusschere, V.; Bacha, S.; Maulik, U.; Bondyopadhyay, S. Nonintrusive Load Monitoring: A Temporal Multilabel Classification Approach. *IEEE Trans. Ind. Inform.* **2014**, *11*, 262–270. [CrossRef]
16. Liao, J.; Elafoudi, G.; Stankovic, L.; Stankovic, V. Non-intrusive appliance load monitoring using low-resolution smart meter data. In Proceedings of the 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), Venice, Italy, 3–6 November 2014; pp. 535–540. [CrossRef]
17. Rehman, A.U.; Lie, T.T.; Vallès, B.; Tito, S.R. Low Complexity Non-Intrusive Load Disaggregation of Air Conditioning Unit and Electric Vehicle Charging. In Proceedings of the 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia), Chengdu, China, 21–24 May 2019; pp. 2607–2612. [CrossRef]
18. Parson, O.; Ghosh, S.; Weal, M.; Rogers, A. Non-intrusive load monitoring using prior models of general appliance types. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012; pp. 356–362.
19. Makonin, S.; Popowich, F.; Bajić, I.V.; Gill, B.; Bartram, L. Exploiting HMM Sparsity to Perform Online Real-Time Nonintrusive Load Monitoring. *IEEE Trans. Smart Grid* **2016**, *7*, 2575–2585. [CrossRef]
20. Murray, D.; Stankovic, L.; Stankovic, V.; Lulic, S.; Sladojevic, S. Transferability of Neural Network Approaches for Low-rate Energy Disaggregation. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8330–8334. [CrossRef]
21. Kolter, J.Z.; Johnson, M.J. REDD: A public data set for energy disaggregation research. In Proceedings of the SustKDD Workshop on Data Mining Applications in Sustainability, San Diego, CA, USA, 21 August 2011; pp. 59–62.
22. Kelly, J.; Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2015**, *2*, 150007. [CrossRef] [PubMed]
23. Murray, D.; Stankovic, L.; Stankovic, V. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Sci. Data* **2017**, *4*, 160122. [CrossRef] [PubMed]
24. Dataport—Pecan Street Inc. Available online: https://www.pecanstreet.org/dataport/ (accessed on 30 January 2022).
25. Amara-Ouali, Y.; Goude, Y.; Massart, P.; Poggi, J.-M.; Yan, H. A Review of Electric Vehicle Load Open Data and Models. *Energies* **2021**, *14*, 2233. [CrossRef]
26. Jaramillo, A.F.M.; Laverty, D.M.; del Rincón, J.M.; Hastings, J.; Morrow, D.J. Supervised Non-Intrusive Load Monitoring Algorithm for Electric Vehicle Identification. In Proceedings of the 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Dubrovnik, Croatia, 25–28 May 2020; pp. 1–6. [CrossRef]
27. Pan, Y.; Liu, K.; Shen, Z.; Cai, X.; Jia, Z. Sequence-To-Subsequence Learning With Conditional Gan For Power Disaggregation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3202–3206. [CrossRef]
28. Makonin, S.; Popowich, F. Nonintrusive load monitoring (NILM) performance evaluation. *Energy Effic.* **2015**, *8*, 809–814. [CrossRef]
29. Klemenjak, C.; Faustine, A.; Makonin, S.; Elmenreich, W. On metrics to assess the transferability of machine learning models in non-intrusive load monitoring. *arXiv* **2019**, arXiv:1912.06200.
30. Pereira, L.; Nunes, N. Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review. *WIREs Data Mining Knowl. Discov.* **2018**, *8*, e1265. [CrossRef]
31. Zhang, Z.; Son, J.H.; Li, Y.; Trayer, M.; Pi, Z.; Hwang, D.Y.; Moon, J.K. Training-free non-intrusive load monitoring of electric vehicle charging with low sampling rate. In Proceedings of the IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society, Dallas, TX, USA, 29 October–1 November 2014; pp. 5419–5425. [CrossRef]
32. Munshi, A.A.; Mohamed, Y.A.-R.I. Unsupervised Nonintrusive Extraction of Electrical Vehicle Charging Load Patterns. *IEEE Trans. Ind. Inform.* **2018**, *15*, 266–279. [CrossRef]
33. Zhao, B.; He, K.; Stankovic, L.; Stankovic, V. Improving Event-Based Non-Intrusive Load Monitoring Using Graph Signal Processing. *IEEE Access* **2018**, *6*, 53944–53959. [CrossRef]
34. Moreno Jaramillo, A.F.; Lopez-Lorente, J.; Laverty, D.M.; Martinez-del-Rincon, J.; Morrow, D.J.; Foley, A.M. Effective identification of distributed energy resources using smart meter net-demand data. *IET Smart Grid* **2022**. [CrossRef]
35. GitHub-DLZRMR/seq2subseq: Seq2subseq Method for NILM. Available online: https://github.com/DLZRMR/seq2subseq (accessed on 9 February 2022).
36. D'Incecco, M.; Squartini, S.; Zhong, M. Transfer Learning for Non-Intrusive Load Monitoring. *IEEE Trans. Smart Grid* **2019**, *11*, 1419–1429. [CrossRef]

37. Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Sutton, C. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 2604–2611.
38. Raileanu, L.E.; Stoffel, K. Theoretical comparison between the gini index and information gain criteria. *Ann. Math. Artif. Intell.* **2004**, *41*, 77–93. [CrossRef]