This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Comparative Evaluation of Machine Learning Models and Input Feature Space for Non-intrusive Load Monitoring

Attique Ur Rehman, Tek Tjing Lie, Brice Vallès, and Shafiqur Rahman Tito

***Abstract*—Recent advancement in computational capabilities has accelerated the research and development of non-intrusive load disaggregation. Non-intrusive load monitoring (NILM) offers many promising applications in the context of energy efficiency and conservation. Load classification is a key component of NILM that relies on different artificial intelligence techniques, e.g., machine learning. This study employs different machine learning models for load classification and presents a comprehensive performance evaluation of the employed models along with their comparative analysis. Moreover, this study also analyzes the role of input feature space dimensionality in the context of classification performance. For the above purposes, an event-based NILM methodology is presented and comprehensive digital simulation studies are carried out on a low sampling real-world electricity load acquired from four different households. Based on the presented analysis, it is concluded that the presented methodology yields promising results and the employed machine learning models generalize well for the invisible diverse testing data. The multi-layer perceptron learning model based on the neural network approach emerges as the most promising classifier. Furthermore, it is also noted that it significantly facilitates the classification performance by reducing the input feature space dimensionality.**

***Index Terms*—Machine learning model, load feature, non-intrusive load monitoring (NILM), comparative evaluation.**

I. Introduction

With the fast development pace of the electronics market, the energy demand has risen exponentially in the last two decades. Further, the variability and forecasting uncertainty of energy consumption patterns make it difficult for the utilities to maintain the equilibrium between demand and supply. In this context, effective energy monitoring is essential for modern power systems. Energy monitoring offers many promising solutions for the grid stability, including but not limited to energy forecasting, demand-side management, and fault diagnosis [1]. One of the well-known techniques of efficient energy monitoring is load disaggregation, where an appliance- or circuit-level power profile has been extracted from an aggregated load power profile [2]. Load disaggregation, also referred to as energy disaggregation, can be broadly categorized into intrusive load monitoring (ILM) and non-intrusive load monitoring (NILM) techniques. ILM requires dedicated measurement devices to be installed with each appliance, which is simple but a cost-prohibitive method [3]. Alternatively, NILM is a non-intrusive and cost-efficient approach that collects the aggregated load measurements at a single-entry point and performs disaggregation via different software techniques. An NILM system comprises three components, i. e., data acquisition, feature extraction, and load classification.

Numerous research works have been done based on the initial concept of NILM [4]. Reference [5] have recently presented an state-of-the-art review of different NILM components. Data acquisition is the starting point of the NILM system, where data can be acquired either at low or high sampling rate. In this context, [6] and [7] present a comprehensive comparison of publicly available load disaggregation datasets. It is noted that most of these datasets, used for NILM evaluation, are based on high sampling rate. Subsequently, most of the available NILM literature is based on these highly sampled data [8]. Highly sampled data in NILM yield better energy disaggregation [9] with the larger number of appliance identifications [10] but at a cost of more complex hardware requirement, large storage demand, and huge capital investment [11].

Feature extraction is a process of transforming raw data into meaningful information. In the NILM domain, feature refers to a unique consumption pattern of an appliance, which is used for its identification. Numerous load features are proposed based on power, current, and voltage. However, active and reactive power are the most widely-used load features in the NILM domain [6], [12], [13].

To identify individual loads based on the extracted features, numerous artificial-intelligence-based techniques are adopted by the research community. In this context, machine learning (ML) is widely employed, such as the *k*-nearest neighbors (*k*-NN) model, which is successfully deployed to disaggregate the air conditioning unit and electric vehicle charging [14]. Likewise, the disaggregation of an air condi-

Manuscript received: October 29, 2020; accepted: March 19, 2021. Date of CrossCheck: March 19, 2021. Date of online publication: XX XX, XXXX.

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

A. U. Rehman (corresponding author) and T. T. Lie are with the Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland, New Zealand (e-mail: attique.rehman@aut.ac.nz; tek.lie@aut.ac.nz).

B. Vallès is with Brice Vallès Consulting, Auckland, New Zealand (e-mail: brice.valles@gmail.com).

S. R. Tito is with the School of Engineering and Trades, Manukau Institute of Technology, Auckland, New Zealand (e-mail: shafiqur.tito@manukau.ac.nz).

DOI: 10.35833/MPCE.2020.000741

tioning unit is also carried out using a support vector machine (SVM) in [15]. Further, the SVM and *k*-NN are used for load classification, where input features are extracted from active and reactive power, and power factor [16]. Other techniques like hidden Markov model (HMM) [17] and its variants [18], [19], and artificial neural network (ANN) [20]-[24] are also employed by numerous researchers towards load disaggregation.

In the existing literature, numerous studies present comparative analysis of different ML models. For example, [25]-[28] present a comprehensive review of different classification techniques along with their corresponding advantages and disadvantages. However, none of them are in the context of NILM. Most of the existing NILM studies are based on single or two to three ML models for the classification purposes of a given problem. To address this, [29] presents a comparative study of five different ML models in the context of NILM, which is, however, based on highly sampled data acquisition, i.e., a sampling rate of 30 kHz. It is observed that the existing literature is lagging in terms of providing a comprehensive comparative evaluation of different ML models in the context of NILM.

Further, as mentioned above, most of the available NILM studies are based on high data granularity. However, to realize the practical potential of NILM, studies need to be more focused on low-sampling NILM systems rather than high-sampling ones. Based on the lower data granularity, the low-sampling NILM system is not only a more viable option for the existing metering infrastructure [30], but also yields lower computational demands and costs. However, the existing NILM literature is limited in providing comprehensive insights in terms of low-sampling NILM systems.

To address the mentioned shortcomings, this paper is primarily intended to evaluate the performance of different ML models in the context of low data granularity based NILM system. Hence, we focus on 1/60 Hz data granularity, whose sampling rate is 60 times lower than 1 Hz, which is mostly used in the context of low-sampling NILM systems. Moreover, to further realize a practical load scenario, this paper is based on a recently released practical load database: New Zealand GREEN Grid database [31]. The contributions of this study are summarized as follows.

1) An event-based NILM methodology is presented for low-sampling practical load measurements.

2) A comprehensive performance evaluation of different ML models is presented in the context of low-sampling NILM system. For the above purpose, ten different ML models are employed.

3) A new performance metric is introduced in the context of NILM evaluation along with other well-known evaluation criteria.

4) A comparative evaluation of the employed ML models is presented in combination with different input features.

This study not only contributes to the existing state-of-the-art ML models in NILM applications but also facilitates future research in the mentioned domain. The reminder of this paper is organized as follows. Section II presents the detailed research methodology of NILM system. Section III presents the simulation details and the corresponding results and analysis. Section IV concludes this paper.

II. Research Methodology

This paper presents a low-sampling event-based NILM methodology, which comprises four key components, i.e., data acquisition/pre-processing, event detection, feature extraction, and load classification. Ten different supervised ML models, namely SVM, logistic regression (LR), decision tree (DT), random forest (RF), k-NN, Gaussian process (GP), multi-layer perceptron (MLP), naïve Bayes (NB), quadratic discriminant analysis (QDA), and stochastic gradient descent (SGD), are employed and evaluated in the context of NILM applications. Figure 1 presents the adopted flow of the research methodology. It also highlights the four key components of the event-based NILM system.

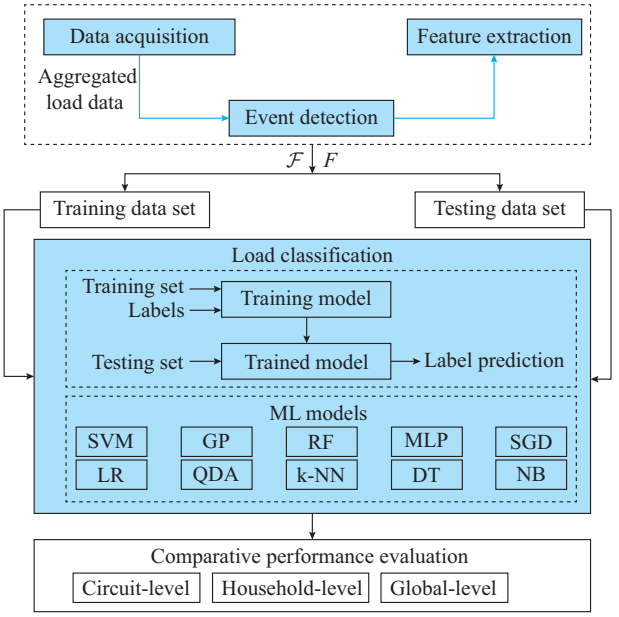

Fig. 1. Flow of research methodology.

In this paper, the methodology presented in Fig. 1 is primarily targeting the non-intrusive load inference of water heating (WH) load element at the circuit-level configuration. However, this methodology is also viable for the non-intrusive load inference of other load elements, even at the appliance-level configuration [14]. WH circuit is selected due to the attributes of the employed practical load database: data granularity and availability of the circuits. Due to the low data granularity of the employed database and the variations in circuit installation configuration, we choose to focus on WH, which is a high-consumption load element and has a dedicated standalone circuit installation configuration. Consequently, it is a more viable load element to be non-intrusively inferred under the given conditions [10], [30]. Moreover, the WH circuit is one of the main stakeholders in terms of electricity consumption in a residential sector [32]-[34]. More importantly, it is a flexible/interruptible load element [35]. These properties make WH as a high potential load element for many practical energy efficiency applications, e.g., demand response [34], [36] and power regulations [33].

A. Data Acquisition and Event Detection

In this study, load data are acquired from New Zealand GREEN Grid database [31]. This is the first database of this kind in New Zealand, where the data have been collected from 2014 to 2018. The database comprises load measurements of 45 households, where each household contains 1-minute (a sampling rate of 1/60 Hz) mean power data, in watt, available for individual circuits and main circuit (total incoming power). Further, each household has 6 circuits including the main circuit, where the installation configuration of individual circuit varies from household to household [37].

For simulation purposes, load data are acquired from four different households with dedicated WH circuit installed in their premises, where other individual circuits may vary. The details can be found in [37]. The acquired load data are pre-processed using the median filtering [38] technique prior to the event detection. In the event-based NILM system, event detection is a key component, where an event is defined as a transient portion of a signal that deviates from the prior steady state and lasts till the next steady state is achieved [39]. The events are an indication of variations triggered by turning-on/off of individual appliances/circuits within the aggregated load profile. In this context, event detection refers to a process of identifying these changes in the aggregated load data [40]. In this study, the mean absolute deviation sliding window (MAD-SW) [41] algorithm has been employed for event detection purposes. For the event detection simulations, the threshold value of 150 W is selected, and the window width ω and delay tolerance Δt are set empirically at 3 samples and 2 minutes, respectively. In terms of event detection, this paper aims to detect all the events within the input pre-processed aggregated load data, where at a later stage, non-intrusive load inference of WH circuit is of primary interest.

B. Feature Extraction and Reduction

Due to the low sampling rate, most of the waveform information, i.e., harmonic contents and reactive power, is lost except the active power information [30]. As this study is based on low data granularity, i.e., a sampling rate of 1/60 Hz, it uses the available mean power as an input variable for the feature extraction process. The extracted load features are related to different properties of the load events, i.e., geometrical, statistical, and power levels.

The extracted feature set, $\mathcal{F}$ [14], comprises five distinct load features and is given as:

$$\mathcal{F} = \left\{\tau_{width}, P_{p2p}, \sigma, \sigma^2, \mu\right\} \tag{1}$$

where $\tau_{width}, P_{p2p}, \sigma, \sigma^2,$ and μ are the transient width, peak-to-peak power magnitude, standard deviation, variance, and mean value of the event, respectively. These load features are computed for each detected event and the mathematical expressions of the features are given as:

$$\tau_{width} = \tau_{end} - \tau_{start} \tag{2}$$

$$P_{p2p} = P_{end} - P_{start} \tag{3}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \mu\right|^2} \tag{4}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \mu\right|^2 \tag{5}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{6}$$

where τ_{start} and τ_{end} are the indices of the starting time and ending time of the event, respectively; P_{start} and P_{end} are the power magnitudes at the starting time and ending time of the event, respectively; x_i is pre-processed active power values at time indices within the detected transient portion, i.e., event; and n is the total number of time indices that the transient portion lasts.

Another feature set F is also extracted using feature reduction, which is the process that features are intelligently grouped to reduce the feature space dimensionality. The feature set F is a combinatorial form of $\mathcal{F}$ that contains all the (features) information of $\mathcal{F}$. However, the feature space has been reduced, i.e., it is composed of three distinct features rather than five, as given in (7).

$$F = \left\{\mathcal{S}_{\mathcal{E}}, C_{Disp}, C_{var}\right\} \tag{7}$$

where $\mathcal{S}_{\mathcal{E}}$, C_{Disp}, and C_{var} are the slope, coefficient of dispersion, and coefficient of variation of the detected events, respectively, as given in (8)-(10).

$$\mathcal{S}_{\mathcal{E}} = \frac{P_{p2p}}{\tau_{width}} = \frac{P_{end} - P_{start}}{\tau_{end} - \tau_{start}} \tag{8}$$

$$C_{Disp} = \frac{\sigma^2}{\mu} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \mu\right|^2}{\frac{1}{n}\sum_{i=1}^{n}x_i} \tag{9}$$

$$C_{var} = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \mu\right|^2}}{\frac{1}{n}\sum_{i=1}^{n}x_i} \tag{10}$$

The extracted load feature sets, $\mathcal{F}$ and F, given in (1) and (7), respectively, are used as input features to the ML models used in this study.

C. ML Models

In the ML domain, no single model has superiority over others, and the quest is to identify the optimal model that provides the most accurate classification results under given conditions [26]. The simplest approach is to evaluate the accuracy performance of different ML models for a given problem and identify the one that yields the most accurate classification results. The ten ML models are selected due to their diverse working principles and different strengths and weaknesses. This provides an opportunity to evaluate distinct learning models and identify the most optimal one in the low-sampling NILM systems. Based on the available theoretical and empirical studies, Table I presents a detailed comparative analysis of the advantages and disadvantages of the employed ML models.

Furthermore, a brief methodological description of all the employed ML models is presented as follows.

TABLE I
COMPARISON OF EMPLOYED ML MODELS

ML model	Advantage	Disadvantage	Reference
SVM	Insensitive to data dimensionality, good generalization ability, versatile kernel selection	Higher complexity and memory requirements, rely on model parameters, poor interpretability	[26]-[28], [42]
LR	Parametric model, capability to handle nonlinearity	Multicollinearity issues, require large sample size	[28], [43]
DT	Good generalization ability, noise robustness, computationally faster, easy to interpret	Greedy construction process, overfitting issues, error propagation issue, prone to data dimensionality	[26], [28], [43], [44]
RF	Computationally faster, noise robustness, no parameter tuning, no over-fitting	The increasing number of trees slows down the model	[28], [44], [45]
k-NN	Suitable for multi-model classes, simplicity	Rely on k-value tuning, prone to noise/irrelevant features, dimensionality issue, higher memory requirement, poor interpretability	[26], [28], [43], [44]
GP	Probabilistic approach, good performance in practice	High computational cost	[46], [47]
MLP	Non-parametric, robust to noise and irrelevant features	Large training time, rely on input parameters, hard to interpret	[28], [43], [44], [48]
NB	No parameter tuning, robust to missing values, computationally faster, requires low memory	Prone to data dimensionality	[26], [28], [44]
QDA	Easily computed, work well in practice, no hyperparameter tuning	Long training time, complex operation	[42], [49]
SGD	Easy to implement, efficiency, faster convergence	Hyperparameter tuning required, sensitive to feature scaling	[42], [50]

1) SVM

SVM is a well-known classical supervised ML model based on a concept of a "margin", i.e., either side of a hyperplane that separates two data classes [26]. It is a widely used ML model and is considered as a must-try method due to its most accurate and robust technique among all the models [27]. Further, it establishes itself as a promising classifier for NILM applications [51].

2) LR

LR, also known as the logit model or maximum entropy classifier, is widely used for classification purposes. It is based on statistical models where a logistic curve is fitted to a dataset [44]. LR creates a logit variable comprising the natural log of the likelihoods that the class occurs. Later maximum likelihood estimation algorithm is employed to estimate the probabilities [44]. LR models have also proven themselves for numerous practical problems.

3) DT

DT is a powerful classification model that is simple to understand and easy to interpret. It is based on a recursive hierarchical structure comprising nodes (internal/leaf) and branches. Branches represent the decision rules, where internal and leaf nodes represent features (attributes) and outcomes, respectively.

4) RF

RF is based on a combination of DTs' prediction. Several DTs are trained and each DT votes for its preferred class. The class with a larger number of votes is taken as a final prediction. RF model is not only fast to be trained but also does not overfit regardless of the number of trees employed in combination [44].

5) k-NN

k-NN stores the complete training set and assigns an unlabeled data point to the class of its nearest neighbors. To attain the nearest neighbors for each data point, *k*-NN generally employs Euclidean distance to measure the distance between the data points [44].

6) GP

GP classifier is a generic supervised learning model designed to solve the problems of regression and classification. For classification purposes, the GP classifier implements the Gaussian processes to estimate the conditional probabilities from the given sample. In the given context, the two key approximation algorithms are Laplace and expectation-propagation [52], where further details on GP classifier can be found in [47]. The GP classifier is establishes in a wide range of domains including remote sensing image classification [46], electroencephalogram signal classification [53], and appearance-based gender classification [54].

7) MLP

MLP is the most widely-employed supervised learning model based on neural networks and has the capability to model complex functions [28]. MLP utilizes backpropagation for training purposes [42] and comprises three layers, i.e., input layer, hidden layer, and output layer. It is worth noting that any random classification problem can be learned even with one hidden layer, given that the hidden layer comprises enough units. Further details can be found in [42], [55].

8) NB

NB is a probabilistic learning model based on Bayes theorem for conditional probabilities. It builds and optimizes a function, given that all attributes in a database are independent. Generally, the maximum likelihood algorithm is used for the training of NB model [44].

9) QDA

QDA is a standard supervised classifier, which uses the Gaussian distribution to model the likelihood of each class and later employs the posterior distributions to classify the given testing data [56].

10) SGD

SGD classifier executes a plain SGD learning routine sup-

porting various loss functions and penalties for classification [42]. It is an efficient approach for discriminative learning of linear classifiers under convex loss functions like SVM and LR. SGD is established for large-scale and sparse ML problems [42].

D. Performance Evaluation Metrics

In this study, the employed ML models are comprehensively evaluated at three different levels: circuit level, household level, and global level, as depicted in Fig. 1. For the above-mentioned purposes, well-known performance metrics are used: recall (R), precision (P), f-score (F_s), and accuracy ($\mathcal{A}$). Moreover, the performance metric of Kappa index ($\mathcal{K}$) is also introduced in the context of NILM classification performance evaluation.

R is defined as the number of relevant items selected, while P is the number of relevant items within the selected items. R and P are mathematically given as in (11) and (12), respectively [7].

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{12}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively.

F_s is defined as the harmonic mean of R and P, mathematically defined as in (13) [7].

$$F_s = \left(\frac{P^{-1}+R^{-1}}{2}\right)^{-1} = 2\times\frac{PR}{P+R} \tag{13}$$

$\mathcal{A}$ is another performance metric used for the evaluation of classification models and is defined as the prediction fraction the model classifies correctly [57], given as in (14).

$$\mathcal{A} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{14}$$

where TN represents true negative.

The terminologies of TP, FP, FN, and TN are well explained in the form of a confusion matrix, given in Table II [58].

TABLE II
TABLE OF CONFUSION MATRIX

Model prediction	Ground-truth	
	Occurred	Not occurred
Detected	TP	FP
Not detected	FN	TN

Another performance metric introduced and employed in this study is the Kappa index $\mathcal{K}$. It is calculated using both the accuracy and expected accuracy, mathematically given as in (15) [59].

$$\mathcal{K} = \frac{\mathcal{A}-E_e}{1-E_e} \tag{15}$$

where the expected accuracy E_e is defined as the accuracy that any random classifier would be expected to attain based on the confusion matrix, as given in Table II. E_e is mathematically defined as in (16) [59].

$$E_e = \frac{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})+(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2} \tag{16}$$

$\mathcal{K} < \mathcal{A}$, however, $\mathcal{K}$ is the degree of agreement among two or more raters, so it is a more robust measure to evaluate the performance of ML model. Moreover, $\mathcal{K}$ of one ML model is directly comparable to that of another ML model employed for a similar classification task. Reference [60] assigns the labels in terms of agreement strength to different ranges of $\mathcal{K}$, as shown in Table III. The details of Table III are used as a benchmark. It is evident that the higher the $\mathcal{K}$ value, the better the agreement. Generally, $\mathcal{K} > 0.40$ is desirable [59].

TABLE III
DIFFERENT RANGES OF $\mathcal{K}$

$\mathcal{K}$	Label
Less than 0	Poor
0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

III. SIMULATIONS AND RESULTS

Comprehensive digital simulations are carried out based on the research methodologies presented in Section II. For the above-mentioned purpose, a desktop computer with Intel Core i7 (8700) processor and 32 GB RAM is used, where MATLAB R2018b and Python 3.6.7 are employed as simulation tools.

All the employed ML models are independently evaluated in combination with input features, $\mathcal{F}$ and F, as given in (1) and (7), respectively. Further, all the employed ML models are independently trained with 20-day load data from a single household and later tested on a diverse set of testing data that are not known in the training phase. This strategy aims at validating the robustness of the given classifiers and identifying the most optimal one for the given problem. Table IV presents the details of households in New Zealand GREEN Grid used for the training and testing purposes of the employed ML models along with the corresponding results in terms of event detection and feature extraction. It is worth noting that event detection simulation details are not within the scope of this study. However, further details can be found in [41].

Table V presents different learning model parameters adopted for simulation purposes. Further details of the presented parameters in Table V can be found in Scikit-Learn [42], which is an ML library for python programming language.

A. ML Simulations in Combination with $\mathcal{F}$

All the employed ML models are fed with the input feature set $\mathcal{F}$, and simulations are carried out according to the details presented in Fig. 1, and Tables IV and V.

Under the given conditions, Table VI presents the circuit-

TABLE IV
DATA ATTRIBUTES AND RESULTS OF NEW ZEALAND GREEN GRID

Data	Household ID	Data acquisition timeframe	No. of data samples	No. of detected events	$\mathcal{F}$	F
Training data	rf_2	May 11-May 30, 2014	28800	1504	1504×5	1504×3
Testing data	rf_2	July 1-10, 2014	14400	898	898×5	898×3
	rf_31	September 1-7, 2016	10080	166	166×5	166×3
	rf_36	June 21-27, 2017	10080	390	390×5	390×3
	rf_42	Janaury 7-13, 2017	10080	60	60×5	60×3

TABLE V
PARAMETERS OF EMPLOYED ML MODELS

ML model	Parameter detail
SVM	C = 1.0, kernel = 'rbf',
MLP	activation = 'relu', hidden_layer_sizes = (100,), solver = 'sgd'
DT	min_samples_leaf = 1, min_samples_split = 2, splitter = 'best'
RF	criterion = 'gini', min_samples_leaf = 1, min_samples_split = 2, n_estimators = 10
NB	priors = None, var_smoothing = 1e-09
GP	max_iter_predict = 100, multi_class = 'one_vs_rest'
LR	C = 1.0, max_iter = 100
k-NN	algorithm = 'auto', leaf_size = 30, p = 2, n_neighbors = 5, weights = 'uniform'
SGD	loss = 'hinge', penalty = 'l2'

level performance results of classifiers in terms of P, R, and F. The evaluation is based on the classification of four different classes, namely turning-on/off of WH and miscellaneous circuits, which are denoted as WH_{on}, WH_{off}, $Misc_{on}$, $Misc_{off}$, respectively. Furthermore, the weighted average performance of all circuits, which is denoted as W_{Avg}, is also included in Table VI. It is evident from the results presented in Table VI that all the employed classifiers generalize well for the entirely unknown testing data. It is observed that rf_2, as a testing data set, attains the best individual circuit-level inference performance by all the employed classifiers. It is anticipated because the testing data of rf_2 are not known in the training phase of the employed classifiers. However, the testing and training data belong to the same household with similar attributes like occupancy, size, the installation configuration of circuit, and usage pattern. In terms of diverse testing households, the worst circuit-level performance is recorded for rf_36. Further, it is worth noting that the WH circuit inference results presented as 0% for rf_31 in Table VI corresponds to the absence of WH circuit activity in reality, i.e., no ground-truth activity, at the given data acquisition timeframe. The absence of WH ground-truth activity is precisely predicted by all the employed classifiers.

As for circuit-level inference performance, it is also evident from Table VI that, in most cases, the MLP classifier based on the neural network outperforms other employed classifiers. The MLP classifier is followed by QDA, LR, SVM, and GP with marginal variations in terms of circuit-level inference performance. The DT model shows the worst circuit-level inference performance compared with other employed models under the given conditions. For further visualization purposes, Fig. 2 presents the circuit-level classification results of the MLP in the form of a normalized confusion matrix for all the testing households.

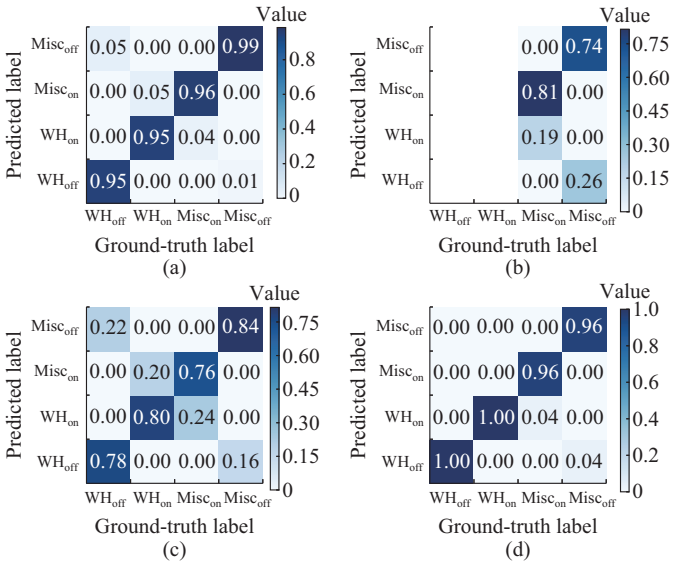

Fig. 2. Circuit-level classification results of MLP for different testing households. (a) rf_2. (b) rf_31. (c) rf_36. (d) rf_42.

Table VII presents the household-level performance of the employed classifiers in terms of $\mathcal{A}$ and $\mathcal{K}$. It is evident from the results presented in Table VII that the MLP and SGD classifiers outperform others for the rf_31 and rf_36. For rf_2, SVM and GP have an edge over the other models. For rf_42, the QDA outperforms all other employed ML models. This performance variation is expected due to the diverse nature of the employed ML models and diverse testing households. The least $\mathcal{A}$ and $\mathcal{K}$ (57.43% and 43.21%, respectively) are recorded for testing household rf_36 for the DT model.

In addition to the evaluation of ML models in terms of circuit-level and household-level, a global-level evaluation based on the entire set of testing households under consideration, is also carried out in this study. In this context, Fig. 3 presents a comparison of all employed ML models in the form of a box plot, to visualize different statistical parameters of the classification performance. The earlier analysis can be further validated from the results presented in Fig. 3, particularly from Fig. 3(b), where all the employed ML models attain the desired results of $\mathcal{K} > 0.4$ [59], [60] in terms of all statistical distribution, i.e., the minimum, maximum, median, and mean performances. Further, it is also evident from Fig. 3(b) that, the mean and median $\mathcal{K}$ performance of

TABLE VI
CIRCUIT-LEVEL INFERENCE PERFORMANCE COMPARISON OF ML MODELS IN COMBINATION WITH $\mathcal{F}$

House-hold ID	State	SVM			LR			DT			RF			k-NN			GP			MLP			NB			QDA			SGD		
		P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s
rf_2	WH_{off}	98	96	97	94	96	95	96	96	96	98	97	97	98	97	97	98	95	96	98	95	96	93	96	95	97	95	96	93	96	95
	WH_{on}	94	96	95	93	96	95	92	92	92	92	96	94	92	96	94	94	97	96	94	95	94	94	96	95	94	95	94	96	91	93
	$Misc_{on}$	98	96	97	97	96	97	95	95	95	98	95	96	98	95	96	98	96	97	97	96	96	97	96	97	97	96	96	94	97	96
	$Misc_{off}$	97	99	98	97	96	97	97	98	98	98	99	98	98	99	98	97	99	98	97	99	98	98	96	97	97	98	97	98	96	97
	W_{Avg}	97	97	97	96	96	96	96	96	96	97	97	97	97	97	97	97	97	97	96	96	96	96	96	96	96	96	96	95	95	95
rf_31	WH_{off}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	WH_{on}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	$Misc_{on}$	100	80	89	100	80	89	100	82	90	100	80	89	100	81	89	100	80	89	100	81	90	100	80	89	100	80	89	100	84	91
	$Misc_{off}$	100	69	82	100	71	83	100	57	73	100	66	79	100	64	78	100	67	80	100	74	85	100	67	80	100	67	80	100	69	82
	W_{Avg}	100	76	86	100	77	87	100	73	84	100	75	85	100	75	85	100	75	86	100	79	88	100	75	86	100	75	86	100	79	88
rf_36	WH_{off}	78	67	72	84	80	82	71	74	73	75	79	77	78	64	73	83	73	78	84	78	81	83	82	83	84	76	80	81	82	81
	WH_{on}	64	84	73	71	84	77	44	39	41	68	80	73	69	82	75	67	83	74	77	80	78	71	84	77	71	76	74	77	81	79
	$Misc_{on}$	76	56	65	70	65	72	44	49	47	75	61	67	77	62	69	76	58	66	79	76	77	80	62	70	74	69	72	80	76	77
	$Misc_{off}$	69	76	72	79	84	81	70	67	69	76	71	73	70	79	74	74	83	78	78	84	81	77	82	79	76	84	80	80	78	79
	W_{Avg}	72	71	70	79	78	78	57	57	57	73	73	73	74	73	73	75	74	74	79	79	79	78	77	77	76	76	76	79	79	79
rf_42	WH_{off}	83	100	91	83	100	91	50	100	67	50	100	67	50	100	67	83	100	91	83	100	91	62	100	77	83	100	91	56	100	71
	WH_{on}	62	100	77	62	100	77	38	60	46	50	80	62	50	100	67	62	100	77	83	100	91	62	100	77	100	100	100	62	100	77
	$Misc_{on}$	92	88	90	100	88	94	91	80	85	95	84	89	100	80	89	100	88	94	100	96	98	100	84	91	100	100	100	100	88	94
	$Misc_{off}$	100	88	94	100	96	98	100	80	89	100	80	89	100	80	89	100	96	98	100	96	98	96	88	92	100	96	98	100	84	91
	W_{Avg}	92	90	90	95	93	94	87	80	82	90	83	85	92	83	85	95	93	94	97	97	97	92	88	89	99	98	98	93	88	89

Note: all results are in percentage.

TABLE VII
PERFORMANCE OF HOUSEHOLD-LEVEL MODELS IN COMBINATION WITH $\mathcal{F}$

ML model	rf_2		rf_31		rf_36		rf_42	
	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)	$\mathcal{A}$(%)	$\mathcal{K}$ (%)	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)
SVM	96.99	95.91	75.90	58.36	70.76	61.03	90.00	84.87
LR	95.87	94.40	76.50	59.25	78.20	70.93	93.33	89.91
DT	95.54	93.94	73.49	54.43	57.43	43.21	80.00	70.73
RF	96.77	95.61	74.69	56.59	72.82	63.72	83.33	75.60
k-NN	96.65	95.46	74.69	56.46	72.82	63.77	83.33	76.00
GP	96.99	95.91	75.30	57.47	74.10	65.48	93.33	89.91
MLP	96.32	95.00	78.91	62.65	79.23	72.31	96.66	94.87
NB	96.10	94.71	75.30	57.47	77.43	69.90	88.33	82.64
QDA	96.21	94.85	75.30	57.47	76.15	68.21	98.33	97.41
SGD	95.43	93.79	78.91	62.29	79.23	72.29	88.33	82.78

the MLP and the median $\mathcal{K}$ performance of the QDA model lie in the almost perfect region.

B. ML Simulations in Combination with F

All the employed ML models are further evaluated in combination with the reduced number of features, i.e., F being an input feature set. This provides an opportunity to analyze the feature space dimensionality in the context of the performance of classification models. Table VIII presents the circuit-level performance results of all the employed ML models in combination with F. Under the given conditions, it is evident from the results presented in Table VIII that irrespective of whether the reduced feature space is regarded as an input to the ML models, all the employed ML models not only generalize well for the unknown diverse testing data, but also in some cases, attain better circuit-level inference results compared with the results presented in Table VI. For example, in the case of rf_36, a significant increase in DT circuit-level performance has been recorded, yielding a total

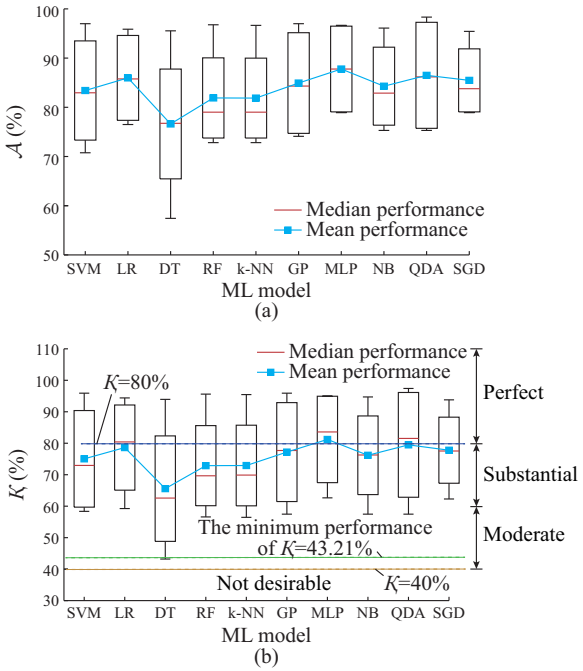

Fig. 3. Comparison of ML models in combination with $\mathcal{F}$. (a) $\mathcal{A}$. (b) $\mathcal{K}$.

of 12% improvement in the weighted average performance. As discussed in Table I, some of the employed ML models are prone to dimensionality issue; hence, it is expected that reducing the feature space dimensionality facilitates the corresponding ML models. Further, as mentioned earlier, the 0% WH circuit inference for rf_31 corresponds to the absence of ground-truth activity of the circuit.

TABLE VIII
CIRCUIT-LEVEL INFERENCE PERFORMANCE COMPARISON OF ML MODELS IN COMBINATION WITH F

Household ID	State	SVM			LR			DT			RF			k-NN			GP			MLP			NB			QDA			SGD		
		P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s	P	R	F_s
rf_2	WH_{off}	99	90	94	99	88	93	96	95	95	98	95	96	99	95	97	95	86	90	94	85	89	88	85	87	96	92	94	95	93	94
	WH_{on}	93	88	90	93	87	90	94	89	92	93	94	93	91	94	92	92	87	89	90	87	88	87	82	85	93	91	92	92	92	92
	$Misc_{on}$	93	96	94	92	96	94	94	96	95	96	96	96	96	95	95	92	95	94	92	94	93	90	93	91	94	96	95	95	95	95
	$Misc_{off}$	94	100	97	93	99	96	97	97	97	97	99	98	97	99	98	92	97	95	91	96	94	91	93	92	96	97	97	96	97	96
	W_{Avg}	95	94	94	94	94	94	95	95	95	96	96	96	96	96	96	93	93	93	92	92	92	89	89	89	95	95	95	95	95	95
rf_31	WH_{off}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	WH_{on}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	$Misc_{on}$	100	81	90	100	81	90	100	81	90	100	82	90	100	81	90	100	81	90	100	82	90	100	83	91	100	81	90	99	81	89
	$Misc_{off}$	100	76	86	100	76	86	100	60	75	100	64	78	100	72	84	100	72	84	100	72	84	100	74	85	100	71	83	100	72	84
	W_{Avg}	100	80	89	100	80	89	100	74	85	100	76	86	100	78	88	100	78	88	100	79	88	100	80	89	100	78	87	99	78	87
rf_36	WH_{off}	85	73	79	85	73	79	70	73	72	76	74	75	78	68	73	85	75	80	86	71	78	85	82	83	74	83	78	84	80	82
	WH_{on}	78	76	77	80	73	76	68	71	69	68	80	73	74	80	77	79	76	77	80	72	76	81	80	80	70	81	75	77	82	79
	$Misc_{on}$	76	79	77	75	82	78	69	66	68	75	62	68	78	71	74	76	80	78	74	82	78	80	81	80	77	65	71	80	76	78
	$Misc_{off}$	75	86	80	75	86	80	69	66	68	72	74	73	70	79	74	76	86	81	73	87	80	81	84	82	78	67	73	79	84	81
	W_{Avg}	79	78	78	79	78	78	69	69	69	73	73	72	75	75	75	79	79	79	78	78	78	82	82	82	75	74	74	80	80	80
rf_42	WH_{off}	71	100	83	71	100	83	38	100	56	50	80	62	56	100	71	71	100	83	71	100	83	56	100	71	71	100	83	71	100	83
	WH_{on}	83	100	91	83	100	91	56	100	71	50	80	62	57	80	67	83	100	91	83	100	91	71	100	83	62	100	77	83	100	91
	$Misc_{on}$	100	96	98	100	96	98	100	84	91	95	84	89	96	88	92	100	96	98	100	96	98	100	92	96	100	88	94	100	96	98
	$Misc_{off}$	100	92	96	100	92	96	100	68	81	95	84	89	100	84	91	100	92	96	100	92	96	100	84	91	100	92	96	100	92	96
	W_{Avg}	96	95	95	96	95	95	91	80	82	88	83	85	91	87	88	96	95	95	96	95	95	94	90	91	94	92	92	96	95	95

Note: all results are in percentage.

The employed ML models in combination with F are also evaluated at the household level. For the above-mentioned purposes, the $\mathcal{A}$ and $\mathcal{K}$ have been employed, and the extracted results are presented in Table IX.

TABLE IX
PERFORMANCE OF HOUSEHOLD-LEVEL MODELS IN COMBINATION WITH F

ML model	rf_2		rf_31		rf_36		rf_42	
	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)	$\mathcal{A}$ (%)	$\mathcal{K}$ (%)
SVM	94.43	92.40	79.51	63.58	78.20	70.96	95.00	92.37
LR	93.76	91.48	79.51	63.58	78.20	70.96	95.00	92.37
DT	95.10	93.32	74.09	55.44	69.23	58.94	80.00	71.65
RF	96.10	94.69	75.90	57.96	72.56	63.40	83.33	75.20
k-NN	95.87	94.39	78.31	61.73	74.61	66.17	86.66	80.16
GP	92.65	89.96	78.31	61.73	78.97	71.98	95.00	92.37
MLP	91.64	88.60	78.91	62.53	77.69	70.28	95.00	92.37
NB	89.30	85.42	80.12	64.29	81.53	75.38	90.00	85.12
QDA	94.87	93.02	77.71	60.81	74.35	65.75	91.66	87.50
SGD	94.76	92.88	78.31	61.46	80.25	73.67	95.00	92.37

It is evident from Table IX that for all the testing households, the employed ML models attain the promising results even when using the reduced feature set. It is also observed that similar to the results presented in Table VII, the performance of ML models varies from household to household. For household rf_2, the RF model outperforms others. The NB classifier unanimously attains the best performance for two testing households, i.e., rf_31 and rf_36.

ML models in combination with F, are also evaluated at the global level, where the corresponding comparative results in the form of a box plot are presented in Fig. 4 in terms of $\mathcal{A}$ and $\mathcal{K}$.

It is further validated from the results presented in Fig. 4 that no single model has a clear edge over others. Rather, it is observed that the ML models namely, SVM, LR, GP, NB, QDA, SGD, and MLP have marginal variations in terms of overall mean and median performances, as highlighted in Fig. 4(a). Furthermore, it is evident from Fig. 4(b) that in terms of the $\mathcal{K}$, the performance distributions of most models lie in the substantial region or above.

C. Comparative Analysis

To underline the influence of feature space dimensionality, a comparative evaluation of the employed ML models in combination with $\mathcal{F}$ and F is carried out. For the above purposes, the results presented in Figs. 3 and 4 are compared and analyzed. It is noted that in most cases, the feature space reduction facilitates the performance of models. Further, the least $\mathcal{K}$ attained by any employed ML model in

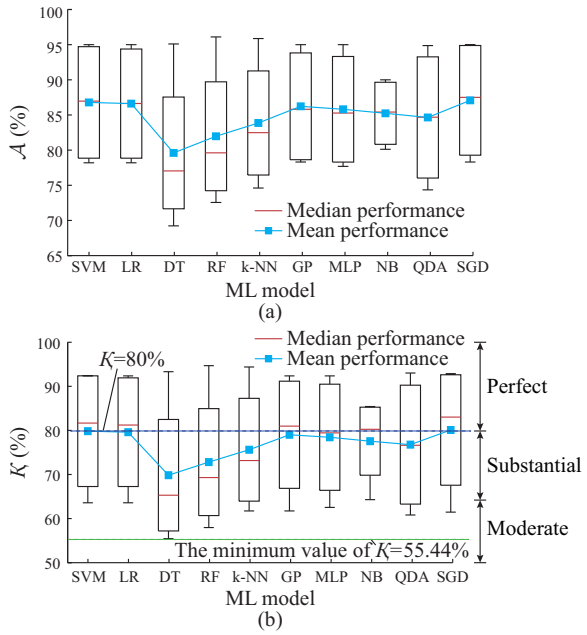

Fig. 4. Comparison of ML models in combination with F. (a) $\mathcal{A}$. (b) K.

combination with $\mathcal{F}$ is 43.21% (highlighted in Fig. 3(b)). However, the least K achieved by any employed learning model in combination with F is 55.44% (highlighted in Fig. 4(b)). This yields an overall improvement of 12.23% for the given ML model.

For further comparative analysis, the overall mean K, based on the entire set of testing households, is also extracted for each employed ML model in combination with $\mathcal{F}$ and F. Figure 5 presents the corresponding comparative analysis results in the form of a bar chart.

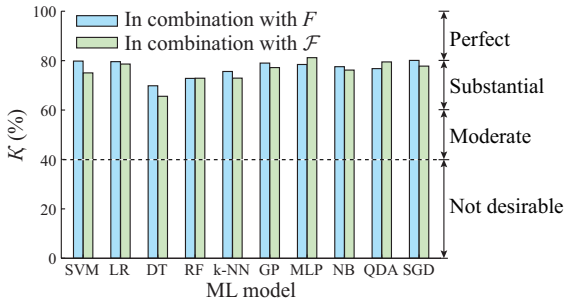

Fig. 5. Comparative evaluation of ML models in combination with $\mathcal{F}$ and F.

It is also evident from Fig. 5 that in most cases, except for MLP and QDA, reduced feature space facilitates the employed ML models in terms of classification performance. In the context of input features, it is also noted that for each ML model, the performance improvement margin varies. As all the employed ML models are different, they have their own advantages and disadvantages, as discussed in Table I. It is also noted that the dimensionality issues, which ML models are prone to, improve significantly with reduced feature space, e.g., DT classifier.

In terms of computational complexity, including time and space complexity, it is anticipated that reducing feature space dimensionality will facilitate the ML models. As feature space is directly proportional to the size of the input samples to ML models, consequently, there are fewer probabilities, weights, and distances to estimate, optimize, and compute, respectively. In this context, one of the key methodologies used is referred to as feature selection, which is a process to find the minimum subset of the most relevant features that retain the key information of the original set [61]. Feature selection methodologies are not within the scope of this paper. Our future research work will be extended to evaluate and underline the significance of feature selection towards more robust NILM development.

IV. Conclusion

This paper presents a comprehensive comparative performance evaluation study of ten diverse ML models in the context of low-sampling NILM applications. The employed ML models are also evaluated in combination with different input feature space. For the above-mentioned purposes, an event-based NILM approach is adopted and digital simulations are carried out on practical load measurements acquired from four different households of the New Zealand GREEN Grid database.

It is worth noting from the analysis that the selection of an optimal ML model is not a case of "one size fits all". In this context, for the given problem, i.e., low-sampling non-intrusive load inference, it is concluded that the MLP classifier based on the neural network outperforms other employed ML models for most of the cases. On the downside, the DT model attains the worst performance under the given conditions. It is also noted that for the given conditions, reducing the feature space dimensionality improves the performance of ML models in most cases.

Based on the presented study and corresponding analysis of the results, towards more robust NILM systems, the future research areas will be two-folded: ① explore ML: ensemble learning and deep learning techniques; and ② explore the feature engineering domain including feature selection methodologies.

References

[1] S. S. Hosseini, K. Agbossou, S. Kelouwani *et al.*, "Non-intrusive load monitoring through home energy management systems: a comprehensive review," *Renewable & Sustainable Energy Reviews*, vol. 79, pp. 1266-1274, Nov. 2017.

[2] N. Batra, R. Kukunuri, A. Pandey *et al.*, "Towards reproducible state-of-the-art energy disaggregation," in *Proceedings of the 6th ACM International Conference on Systems for Energy-efficient Buildings, Cities, and Transportation*, New York, USA, Nov. 2019, pp. 193-202.

[3] A. Gabaldon, R. Molina, A. Marin-Parra *et al.*, "Residential end-uses disaggregation and demand response evaluation using integral transforms," *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 1, pp. 91-104, Jan. 2017.

[4] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870-1891, Dec. 1992.

[5] A. Ruano, A. Hernandez, J. Urena *et al.*, "NILM techniques for intelligent home energy management and ambient assisted living: a review," *Energies*, vol. 12, no. 11, p. 2203, Jun. 2019.

[6] R. Bonfigli, S. Squartini, M. Fagiani *et al.*, "Unsupervised algorithms for non-intrusive load monitoring: an up-to-date overview," in *Proceedings of 15th International Conference on Environment and Electrical Engineering (EEEIC)*, Rome, Italy, Jun. 2015, pp. 1175-1180.

[7] A. Faustine, N. H. Mvungi, S. Kaijage *et al.* (2017, Mar.). A survey

on non-intrusive load monitoring methodies and techniques for energy disaggregation problem [Online]. Available: arXiv:1703.00785v3

[8] K. Basu, V. Debusschere, S. Bacha *et al.*, "Nonintrusive load monitoring: a temporal multilabel classification approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 262-270, Feb. 2015.

[9] A. Hernandez, A. Ruano, J. Urena *et al.*, "Applications of NILM techniques to energy management and assisted living," *IFAC Papersonline*, vol. 52, no. 11, pp. 164-171, Aug. 2019.

[10] K. C. Armel, A. Gupta, G. Shrimali *et al.*, "Is disaggregation the holy grail of energy efficiency? The case of electricity," *Energy Policy*, vol. 52, no. C, pp. 213-234, Jan. 2013.

[11] M. Sun, F. M. Nakoty, Q. Liu *et al.*, "Non-intrusive load monitoring system framework and load disaggregation algorithms: a survey," in *Proceedings of 2019 International Conference on Advanced Mechatronic Systems (ICAMechS)*, Kusatsu, Japan, Aug. 2019, pp. 284-288.

[12] J. Yu, Y. Gao, Y. Wu *et al.*, "Non-intrusive load disaggregation by linear classifier group considering multi-feature integration," *Applied Sciences-Basel*, vol. 9, no. 17, p. 3558, Sept. 2019.

[13] S. M. Tabatabaei, S. Dick, and W. Xu, "Toward non-intrusive load monitoring via multi-label classification," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 26-40, Jan. 2017.

[14] A. U. Rehman, T. T. Lie, B. Vallès *et al.*, "Low complexity non-intrusive load disaggregation of air conditioning unit and electric vehicle charging," in *Proceedings of 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, Chengdu, China, May 2019, pp. 2607-2612.

[15] S. Su, Y. Yan, H. Lu *et al.*, "Non-intrusive load monitoring of air conditioning using low-resolution smart meter data," in *Proceedings of 2016 IEEE International Conference on Power System Technology (POWERCON)*, Wollongong, Australia, Sept. 2016, pp. 1-5.

[16] M. Figueiredo, A. de Almeida, and B. Ribeiro, "Home electrical signal disaggregation for non-intrusive load monitoring (NILM) systems," *Neurocomputing*, vol. 96, pp. 66-73, Nov. 2012.

[17] T. Zia, D. Bruckner, and A. Zaidi, "A hidden Markov model based procedure for identifying household electric loads," in *Proceedings of IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society*, Melbourne, Australia, Nov. 2011, pp. 3218-3223.

[18] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," *Artificial Intelligence and Statistics*, vol. 2012, pp. 1472-1482, Sept. 2012.

[19] H. Kim, M. Marwah, M. Arlitt *et al.*, "Unsupervised disaggregation of low frequency power measurements," in *Proceedings of the 2011 SIAM International Conference on Data Mining*, Arizona, USA, Apr. 2011, pp. 747-758.

[20] J. Cho, Z. Hu, and M. Sartipi, "Non-intrusive A/C load disaggregation using deep learning," in *Proceedings of 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, Denver, USA, Apr. 2018, pp. 1-5 .

[21] J. Kelly and W. Knottenbelt, "Neural NILM: deep neural networks applied to energy disaggregation," in *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, Seoul, South Korea, Nov. 2015, pp. 55-64.

[22] L. Mauch and B. Yang, "A new approach for supervised power disaggregation by using a deep recurrent LSTM network," in *Proceedings of 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Orlando, USA, Dec. 2015, pp. 63-67.

[23] L. De Baets, C. Develder, T. Dhaene *et al.*, "Detection of unidentified appliances in non-intrusive load monitoring using siamese neural networks," *International Journal of Electrical Power & Energy Systems,* vol. 104, pp. 645-653, Jan. 2019.

[24] Y. Lin and Y. Hu, "Electrical energy management based on a hybrid artificial neural network-particle swarm optimization-integrated two-stage non-intrusive load monitoring process in smart homes," *Processes*, vol. 6, no. 12, p. 236, Dec. 2018.

[25] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159-190, Nov. 2006.

[26] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Emerging Artificial Intelligence Applications in Computer Engineering*, vol. 160, no. 3, pp. 249-268, Oct. 2007.

[27] X. D. Wu, V. Kumar, J. R. Quinlan *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, Jan. 2008.

[28] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *Proceedings of 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, Mar. 2016, pp. 1310-1315.

[29] M. Azaza and F. Wallin, "Evaluation of classification methodologies and Features selection from smart meter data," in *Proceedings of the 9th International Conference on Applied Energy*, Cardiff, UK, Aug. 2017, pp. 2250-2256.

[30] K. Basu, V. Debusschere, S. Bacha *et al.*, "A generic data driven approach for low sampling load disaggregation," *Sustainable Energy Grids & Networks*, vol. 9, pp. 118-127, Mar. 2017.

[31] B. Anderson, D. Eyers, R. Ford *et al.* (2018, Nov.). New Zealand GREEN Grid household electricity demand study 2014-2018 [Online]. Available: http://reshare.ukdataservice.ac.uk/853334/

[32] Electricity Authority. (2018, Nov.). Electricity in New Zealand, 2018 [Online]. Available: https://www.ea.govt.nz/about-us/media-and-publications/electricity-new-zealand/

[33] Y. Yang, Z. Mi, X. Zheng *et al.*, "Accommodation of curtailed wind power by electric water heaters based on a new hybrid prediction approach," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 3, pp. 525-537, May 2019.

[34] M. Wu, Y. Bao, J. Zhang *et al.*, "Multi-objective optimization for electric water heater using mixed integer linear programming," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 5, pp. 1256-1266, Sept. 2019.

[35] Z. M. Haider, K. K. Mehmood, M. K. Rafique *et al.*, "Water-filling algorithm based approach for management of responsive residential loads," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 1, pp. 118-131, Jan. 2018.

[36] M. Pipattanasomporn, M. Kuzlu, S. Rahman *et al.*, "Load profiles of selected major household appliances and their demand response opportunities," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 742-750, Mar. 2014.

[37] B. Anderson, D. Eyers, R. Ford *et al.* (2018, Nov.). NZ GREEN Grid household electricity demand study: 1 minute electricity power (version 1.0) Centre for Sustainability, University of Otago, Duned [Online]. Available: http://www.otago.ac.nz/centre-sustainability/

[38] M. Liu, J. Yong, X. Wang *et al.*, "A new event detection technique for residential load monitoring," in *Proceedings of 2018 18th International Conference on Harmonics and Quality of Power (ICHQP)*, Ljubljana, Slovenia, May 2018, pp. 1-6.

[39] B. Wild, K. S. Barsim, and B. Yang, "A new unsupervised event detector for non-intrusive load monitoring," in *Proceedings of 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Orlando, USA, Dec. 2015, pp. 73-77.

[40] L. Pereira, "NILMPEds: a performance evaluation dataset for event detection algorithms in non-intrusive load monitoring," *Data*, vol. 4, no. 3, p. 127, Sept. 2019.

[41] A. U. Rehman, T. T. Lie, B. Valles *et al.*, "Event-detection algorithms for low sampling nonintrusive load monitoring systems based on low complexity statistical features," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 3, pp. 751-759, Mar. 2020.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.

[43] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5-6, pp. 352-359, Oct. 2002.

[44] A. C. Lorena, L. F. O. Jacintho, M. F. Siqueira *et al.*, "Comparing machine learning classifiers in potential distribution modelling," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5268-5275, May 2011.

[45] W. Yan, "Application of random forest to aircraft engine fault diagnosis," in *Proceedings of the Multiconference on Computational Engineering in Systems Applications*, Beijing, China, Oct. 2006, pp. 468-475.

[46] P. Morales-Alvarez, A. Pérez-Suay, R. Molina *et al.*, "Remote sensing image classification with large-scale Gaussian processes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1103-1114, Oct. 2017.

[47] T. N. A. Nguyen, A. Bouzerdoum, and S. L. Phung, "A scalable hierarchical Gaussian process classifier," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 3042-3057, Jun. 2019.

[48] A. Subasi and E. Ercelebi, "Classification of EEG signals using neural network and logistic regression," *Computer Methods Programs Biomed*, vol. 78, no. 2, pp. 87-99, May 2005.

[49] A. Starzacher and B. Rinner, "Evaluating KNN, LDA and QDA classification for embedded online feature fusion," in *Proceedings of 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, Sydney, Australia, Dec. 2008, pp. 85-90.

[50] R. G. Wijnhoven and P. de With, "Fast training of object detection using stochastic gradient descent," in *Proceedings of 2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, Aug.

2010, pp. 424-427.
[51] A. Zoha, A. Gluhak, M. A. Imran *et al.*, "Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey," *Sensors (Basel)*, vol. 12, no. 12, pp. 16838-66, Dec. 2012.
[52] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 186-197, Jan. 2010.
[53] B. Wang, F. Wan, P. U. Mak *et al.*, "EEG signals classification for brain computer interfaces based on Gaussian process classifier," in *Proceedings of 2009 7th International Conference on Information, Communications and Signal Processing (ICICS)*, Macau, China, Dec. 2009, pp. 1-5.
[54] H. C. Kim, D. Kim, Z. Ghahramani *et al.*, "Appearance-based gender classification with Gaussian processes," *Pattern Recognition Letters*, vol. 27, no. 6, pp. 618-626, Apr. 2006.
[55] I. D. Longstaff and J. F. Cross, "A pattern recognition approach to understanding the multi-layer perception," *Pattern Recognition Letters*, vol. 5, no. 5, pp. 315-319, 1987.
[56] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1277-1305, Jun. 2007.
[57] J. Alcala, J. Urena, A. Hernandez *et al.*, "Event-based energy disaggregation algorithm for activity monitoring from a single-point sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 10, pp. 2615-2626, Oct. 2017.
[58] M. Aiad and P. H. Lee, "Unsupervised approach for load disaggregation with devices interactions," *Energy and Buildings*, vol. 116, pp. 96-103, Mar. 2016.
[59] Y. Sakiyama, H. Yuki, T. Moriya *et al.*, "Predicting human liver microsomal stability with machine learning techniques," *Journal of Molecular Graphics & Modelling*, vol. 26, no. 6, pp. 907-915, Feb. 2008.
[60] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159-174, Mar. 1977.
[61] N. Ghadimi, A. Akbarimajd, H. Shayeghi *et al.*, "Two stage forecast engine with feature selection technique and improved meta-heuristic algorithm for electricity load forecasting," *Energy*, vol. 161, pp. 130-142, Oct. 2018.

Attique Ur Rehman received the B.E. degree in electrical engineering from Air University, Islamabad, Pakistan, in 2009, and the M.Sc. degree in electrical engineering, information technology, and computer engineering from RWTH Aachen University, Aachen, Germany, in 2013. He is currently pursuing his Ph. D. degree from Auckland University of Technology (AUT), Auckland, New Zealand. He is a recipient of the Callaghan Innovation R&D Student-Fellowship Grant and AUT Doctoral Fee Scholarship. His research interests include the integration of ICT and power systems, applications of artificial intelligence, and load disaggregation.

Tek Tjing Lie received the M.S. and Ph.D. degrees in electrical engineering from Michigan State University, East Lansing, USA, in 1988 and 1992, respectively. He is a Senior Member of IEEE. Currently, he is working as a Professor and an Acting Head of the School of Engineering, Computer, and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand. He is also a Director of the Center for Energy & Power Engineering. He has regularly attracted external funding for projects with industry and supervises several postgraduate students on a wide range of research topics. His research interests include power system planning, operation and control, deregulated electric power markets, energy management, renewable energy, and smart grids.

Brice Vallès received the Ph.D. degree in fluid mechanics from Norwegian University of Science and Technology, Trondheim, Norway, in 2001. Since 2010, he worked in different roles at UniServices, Inland Revenue, and Genesis Energy Limited, Auckland, New Zealand. His research interests include sequential data assimilation to update simulation models in real-time, business development, strategy, planning, and marketing.

Shafiqur Rahman Tito received the Ph.D. degree in electrical engineering from Auckland University of Technology (AUT), Auckland, New Zealand, in 2016. He also worked as a Program Leader of Electrical Engineering at the International College of Auckland, Auckland, New Zealand. Currently, he is working as a Lecturer at Manukau Institute of Technology, Auckland, New Zealand. He was a recipient of the AUT Vice-Chancellor Doctoral Scholarship. His research interests include energy optimization and hybrid renewable energy systems.