# An Experimental Comparison of Performance Metrics for Event Detection Algorithms in NILM

Lucas Pereira
M-ITI / LARSYS and
Prsma.com
Funchal, Portugal
lucas.pereira@m-iti.org

Nuno Nunes
Técnico, U. Lisboa and
M-ITI / LARSYS
Funchal, Portugal
nunojnunes@tecnico.ulisboa.pt

***Abstract*—In this work, we analyze experimentally the behaviour of 23 performance metrics when applied to event detection algorithms in Non-Intrusive Load Monitoring, identifying relationships and clusters between the measures. Our results indicate that when applied to this type of problems, the performance metrics will evidence some considerable differences in behavior when compared to other more traditional machine learning problems. Our results also suggest that most of the differences in behavior are due to the naturally unbalanced nature of the event detection problem in which the number positive cases (True Positives and True Negatives) is much higher than the number of false cases (False Positives and False Negatives).**

## I. Introduction

One of the most interesting challenges of NILM research (besides the disaggregation in itself) is how to evaluate and report the performance of the several proposed approaches, in part due to the lack of a consensus regarding which performance metrics to use in each situation [1], [2]. For example, only recently there seems to be an agreement upon grouping performance metrics according to two main categories: i) event detection performance metrics (ED) designed to evaluate the NILM's ability to track the consumption over time; and ii) energy estimation metrics (EE), designed to characterize and evaluate the NILM disaggregated data against the actual ground-truth [3].

In this paper we extend the work from [4], in which the authors experimentally analyzed the behaviour of 18 performance metrics when applied to classification algorithms and found that, on the contrary to what one would initially expect, some metrics evidenced very high pairwise correlations.

Here instead, we focus on event detection algorithms. More concretely, we analyze the behavior of 23 event detection performance metrics when applied to evaluate five algorithms across four datasets. To this end, we first train and test five different algorithms across the different datasets by conducting a controlled parameter sweep on selected algorithm parameters. Then, for each model returned by the parameter sweep we compute the performance metric values. Once all the performance metrics are calculated we investigate the existence of correlations between the results obtained for each metric. More particularly, we study the existence of linear ($Pearson$) and non-linear ($Spearman$) pairwise correlations. Finally, after an initial correlation analysis, we further explore the metrics correlation by means of hierarchical clustering and *dendrograms*.

TABLE I
Event detection algorithms under evaluation

| Algorithm | Symbol |
|---|---|
| Expert Heuristic Detector [6] | $EHD$ |
| Log Likelihood Ratio Detector with Voting [7] | $LLR_{Vote}$ |
| Log Likelihood Ratio Detector with Maxima | $LLR_{Max}$ |
| Log Likelihood Detector with Maxima [8] | $LLD_{Max}$ |
| Log Likelihood Detector with Voting | $LLD_{Vote}$ |

The remaining of this paper is organized as follows: First we present and describe the algorithms, datasets and performance metrics that will serve as the base for this work. After this, we thoroughly describe our experimental design. We then move to the results and discussion. Finally we conclude this paper by presenting its research implications, limitations and providing an outline of future work.

## II. Algorithms, Datasets and Metrics

### A. Algorithms

According to the literature in event detection for NILM, the different approaches are grouped in three categories: i) expert heuristics; ii) probabilistic models; and iii) matched filters [5].

In this work we use five different algorithms for which the implementation details are available in the literature: one expert heuristic and four probabilistic detectors, which are summarized in table I. For additional details please refer to the respective publications.

### B. Datasets

At the time of this writing, the only dataset with labeled power events is BLUED [9]. Consequently, in order to proceed with this work, we had to manually label some of the already existing datasets. More precisely, we manually identified all the transitions with an absolute power change of at least 30 Watts, for one week of data from house 1 and house 2 of the UK-DALE dataset [10].

Table II summarizes the four datasets used in this experiment, where **P.E.** is the number of power events in the dataset, **Power Change (W)** is a summary of distribution of the power events in terms of mean, and the 25%, 50% and

75% percentiles, and **Elapsed Time (S)** is a summary of the difference in time between the power events in the same terms as the Power Change column.

### C. Performance Metrics

In this work we look at confusion matrix based metrics, Area Under Curve metrics, and domain specific metrics (i.e., performance metrics that where specifically created for event detection problems [5]).

*1) Confusion matrix based metrics:* As the name suggests, confusion matrix based metrics are derived from the values in the confusion matrix. In this work we selected 13 metrics: Accuracy ($A$), Error-rate ($E$), Precision ($P$), Recall ($R$), $F_{0.5}$, $F_1$, $F_2$, Standardized Mathews Correlation Coefficient ($SMCC$) [11], False Positive Rate ($FPR$), True Positive Percentage ($TPP$), False Positive Percentage ($FPP$), Precision-Recall Distance to Perfect Score ($DPS_{PR}$), TPR-FPR Distance to Perfect Score ($DPS_{Rate}$), and TPP-FPP Distance to Perfect Score ($DPS_{Perc}$).

*2) Area Under Curve metrics:* The Area Under the Receiver Operating Characteristic curve ($ROC-AUC$) is commonly used as a summary of two performance metrics ($R$ and $FPR$), and is traditionally calculated using the trapezoidal rule when evaluating scoring algorithms. However, since event detection is a discrete problem, the $AUC$ should not be measured by employing that rule given that the possible presence of outliers could lead to distorted results [12]. Instead, the non-parametric Wilcoxon statistic is used, as suggested by Hanley and Mcneil [11]. In this work, we selected four variations of the $ROC-AUC$, namely, the Wilcoxon based ROC-AUC ($WAUC$), the Wilcoxon based ROC-AUC Balanced ($WAUCB$), the Geometric Mean AUC ($GAUC$), and the Biased AUC ($BAUC$). A more detailed explanation of each metric can be found in [11].

*3) Domain specific metrics:* Domain specific metrics for event detection were first introduced in [5] motivated by the fact that metrics based solely on the confusion matrix implicitly assume that all power events are of equal importance. Here, we look at six DSMs, namely, the Total Power Change due to False Positives ($TPC_{FP}$), the Total Power Change due to False Negatives ($TPC_{FN}$), the Average Power Change due to False Positives ($APC_{FP}$), the Average Power Change due to False Negatives ($APC_{FN}$), the TPC Distance to Perfect Score ($DPS_{TPC}$), and the APC Distance to Perfect Score ($DPS_{APC}$).

## III. Method

### A. Training and Testing

In order to gain deeper insights on the nature and structure of the data that is generated by the event detection algorithms we first perform a parameter sweep on the selected algorithms.

In this particular case we decided to set the power threshold ($P_{thr}$) to 30 Watts, since this is the minimum power change for which there are labeled events in any of the four datasets. Also, we used the real power signal, since it is probably the most widely used power measurement in event detection literature.

Ultimately, in this work we train and test 47950 distinct event detection models. Each model is executed against the four datasets for a total of 109800 model-dataset pairs.

### B. Metrics Calculation

In this step we compute the performance metrics for each of the models returned by the parameter sweep. To do this, we first count the number true positives ($TP$), false positives ($FP$), true negatives ($TN$) and false negatives ($FN$) for each model. This is done by comparing the events triggered by each model with the true events in the corresponding dataset (i.e., the ground-truth). To accomplish this, we define a tolerance interval in which the detected events must fall in order to be considered correct detections. The detection interval is defined by equation 1 and is based on the ground-truth position (GT), and tolerance (Tol) value that was added to account for eventual ambiguity when defining exactly where an event occurs during the labeling process. This parameter was set to range between zero and three seconds with variable steps, as defined by the set $\tau$ in equation 2, where $F_s$ is the sampling rate of the dataset.

$$\Omega = [GT - Tol, GT + Tol] \tag{1}$$

$$\tau = \{0, 1, 5, 15, F_s, 1.5 \times F_s, 2 \times F_s, 2.5 \times_s, 3 \times F_s\} \tag{2}$$

Regarding the process of creating the confusion matrix, we developed an custom algorithm that given a list of detected events and another with the ground-truth data, works as follows:

For each ground-truth event, if there are detections that fall within the interval $\Omega$ given by equation 1, the event that is closer to the ground-truth position (in absolute distance) or the one that was detected first (in the case of equidistant detections) is considered a $TP$, whereas the others must be compared with the next ground-truth event. Otherwise, if no events are detected within the specified interval, a $FN$ is added. Next, the detected events that do not fall within any of the possible intervals $\Omega$ (one per each ground-truth event) are considered $FP$. Lastly, when all the detected and ground-truth events have been processed, the $TN$ are calculated by subtracting the $TP$, $FN$ and $FP$ from the number of samples in the dataset, i.e., all the positions where an event could have happened.

### C. Pairwise Correlations

In this step we compute the linear (*Pearson*) and rank (*Spearman*) pairwise correlations between the performance metrics. The former indicates the existence and direction of any linear relationships, whereas the later assesses the existence of monotonic relationships (results tend to change together but not necessarily at a linear rate).

In this particular case there are 27 metrics (including the $TP$, $FP$, $TN$ and $FN$), for a total of $(27 \times 26)/2 = 351$ unique pairwise correlations per coefficient.

Each model is evaluated 10 times in each of the four datasets, for a total of 200 correlation matrices per coefficient

TABLE II
SUMMARY OF THE ACTIVE POWER CHANGE AND ELAPSED TIME BETWEEN POWER EVENTS IN THE EVENT DETECTION DATASETS

| Dataset | P.E. | Power Change (W) | | | | Elapsed Time (S) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | 25% | 50% | 75% | Mean | 25% | 50% | 75% |
| UK-DALE H1 | 5440 | 268 | 48 | 100 | 273 | 111 | 4 | 7 | 28 |
| UK-DALE H2 | 2842 | 365 | 45 | 74 | 137 | 212 | 6 | 15 | 172 |
| BLUED PA | 887 | 274 | 84 | 116 | 582 | 690 | 18 | 294 | 892 |
| BLUED PB | 1562 | 351 | 40 | 170 | 428 | 383 | 7 | 35 | 83 |

(*10 tolerance values* $\times$ *5 algorithms* $\times$ *4 datasets*). From these matrices we compute a cross-dataset correlation matrix for each coefficient by averaging the correlations matrices of each dataset.

### D. Hierarchical Clustering

In this step we build clusters from the resulting cross-dataset correlation matrices using hierarchical clustering. To do so we first define the dissimilarity and linkage functions.

Regarding the former, we use the dissimilarity function that is defined in equation 3, where $D$ is the distance and $|C|$ is the absolute value of the correlation between metric pairs.

$$D = 1 - |C| \tag{3}$$

As for the latter, we use the average-group distance, whose distance between groups $A$ and $B$ is given by equation 4, where $d$ is a distance function (in our case the Euclidean distance) and $|A|$ and $|B|$ are the size of groups $A$ and $B$, respectively.

$$D_{AB} = \frac{1}{|A||B|} \times \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{4}$$

## IV. RESULTS AND DISCUSSION

Figure 1 shows the average rank and linear correlations of each metric across the four datasets. Metrics with average correlations above 0.5 are highlighted with a yellow background.

When examining the correlation results shown in figure 1, a first interesting observation is that the Average Power Change ($APC$) metrics do not correlate well with any of the other metrics. Furthermore, if we recall that $APC_{FP}$ and $APC_{FN}$ are just the $TPC_{FP}$ and $TPC_{FN}$ metrics normalized by the number of events it is possible to conclude that $APC$ metrics will evidence strong variations depending on the number and amplite of the power events. For example, if event detector $A$ fails to detect ($FN$) all the power events bellow 50 Watts but only fails to detect one event of 100 Watts, it will still have an $APC_{FN}$ of about 50 Watts. On the other hand, an event detector $B$ that only misses one event with 100 Watts will have an $APC_{FN}$ of 100 Watts. Hence, according to this metric algorithm $A$ is considered better than $B$ despite the fact that is misses more events.

Another general observation concerns to the relatively strong correlation (>0.5 in absolute value) between most of the other metrics in figure 1. The only exceptions to this trend are $F_1$, $F_2$, $DPS_{PR}$ and $SMCC$ than have an average correlation of only 0.46. This is particularly interesting since three out of the four metrics were designed to balance Precision and Recall ($F_1$, $F_2$ and $DPS_{PR}$), and still they do not correlate well with their "parent" metrics. For example, the $F_2$ metric does not have any pairwise correlation above 0.65, and perhaps even more surprising, it does not correlate at all (<0.5) with either $P$ or $R$.

Figure 2 shows the dendrograms obtained from the non-linear and linear pairwise correlations. A first general observation is the fact that linear and rank correlations return very similar clusters. In fact, the only difference is the absence of $DPS$ metrics in the linear correlation clusters, which is not necessarily surprising if we recall the quadratic nature of such metrics.

The dendrograms also reveals a very high correlation between the AUC metrics and $R$ (0.99). However, this is just a reflection of the fact that the $Specificity$ (or True Negative Rate - $TNR$ -) is always close to 1 since the number of $TN$ is much higher than the number of $FP$. Consequently, the $AUC$ metrics are only reflecting variations on $R$. Moreover, it is possible to observe that the $DPS_{Rate}$ metric is also very well correlated with the $AUC$ metrics in both coefficients. In this case this is a reflection of the fact that the $FPR$ is always close to 0, meaning that this metrics is also fully controlled by $R$.

A more specific observation concerns to the very strong (0.92) pairwise rank and linear correlations between $TPC_{FN}$ and $R$. A possible explanation for this is that most missed power events ($FN$) have a similar power change value (possibly near to the minimum power threshold), hence the strong linear and non-linear correlations. Similarly, if we consider the $TPC_{FP}$, it is possible to observe some relatively strong correlation (0.63) in the non-linear coefficient that is not followed by a strong linear correlation. Hence, it is expected that some $TPC_{FP}$ ranks will be relatively close to those obtained with the other metrics, in particular those that are derived from the $FP$. Still, the lack of a strong linear correlation is a good indicator that the amount of power change of the FP are heavily dependent on the dataset characteristics and not so much on the algorithm configuration.

## V. CONCLUSION

In this paper, study the relationship between 23 performance metrics when they are used to assess the performance of event detection algorithms. Overall, the major implications of this work are twofold: i) *uncovering important behaviors of the performance metrics when applied to the event detection*

| | | TP | FP | TN | FN | FPP | FPR | A | E | P | R | F05 | F1 | F2 | SMCC | DTPpr | DTPrate | DTPperc | WAUC | WAUCB | GAUC | BAUC | TPC_FN | TPC_FP | DTPtpc | APC_FN | APC_FP | DTPapc | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank | All | 0,56 | 0,67 | 0,67 | 0,56 | 0,67 | 0,67 | 0,63 | 0,63 | 0,58 | 0,56 | 0,55 | 0,44 | 0,39 | 0,42 | 0,41 | 0,56 | 0,62 | 0,56 | 0,56 | 0,56 | 0,56 | 0,58 | 0,62 | 0,26 | 0,29 | 0,14 | 0,24 | 0,51 |
| | No APC | 0,61 | 0,72 | 0,72 | 0,61 | 0,72 | 0,72 | 0,68 | 0,68 | 0,63 | 0,61 | 0,60 | 0,49 | 0,43 | 0,46 | 0,45 | 0,61 | 0,68 | 0,61 | 0,61 | 0,61 | 0,61 | 0,62 | 0,66 | 0,27 | | | | 0,60 |
| | No DSM | 0,62 | 0,74 | 0,74 | 0,62 | 0,74 | 0,74 | 0,69 | 0,69 | 0,65 | 0,62 | 0,62 | 0,50 | 0,45 | 0,48 | 0,47 | 0,62 | 0,69 | 0,62 | 0,62 | 0,62 | 0,62 | | | | | | | 0,62 |
| | Avg. | 0,59 | 0,71 | 0,71 | 0,59 | 0,71 | 0,71 | 0,66 | 0,66 | 0,62 | 0,59 | 0,59 | 0,47 | 0,42 | 0,45 | 0,44 | 0,59 | 0,66 | 0,59 | 0,59 | 0,59 | 0,59 | 0,60 | 0,64 | 0,26 | 0,29 | 0,14 | 0,24 | 0,57 |
| Linear | All | 0,55 | 0,67 | 0,67 | 0,55 | 0,67 | 0,67 | 0,65 | 0,65 | 0,58 | 0,55 | 0,56 | 0,49 | 0,44 | 0,44 | 0,43 | 0,52 | 0,6 | 0,55 | 0,56 | 0,54 | 0,55 | 0,54 | 0,6 | 0,47 | 0,29 | 0,21 | 0,23 | 0,52 |
| | No APC | 0,59 | 0,72 | 0,72 | 0,59 | 0,72 | 0,72 | 0,7 | 0,7 | 0,63 | 0,59 | 0,61 | 0,54 | 0,49 | 0,48 | 0,47 | 0,56 | 0,65 | 0,59 | 0,6 | 0,58 | 0,59 | 0,57 | 0,64 | 0,5 | | | | 0,6 |
| | No DSM | 0,6 | 0,72 | 0,72 | 0,6 | 0,72 | 0,72 | 0,71 | 0,71 | 0,64 | 0,6 | 0,63 | 0,55 | 0,51 | 0,5 | 0,48 | 0,57 | 0,65 | 0,6 | 0,6 | 0,59 | 0,6 | | | | | | | 0,62 |
| | Avg. | 0,58 | 0,7 | 0,7 | 0,58 | 0,7 | 0,7 | 0,68 | 0,68 | 0,61 | 0,58 | 0,6 | 0,52 | 0,48 | 0,47 | 0,46 | 0,55 | 0,63 | 0,58 | 0,58 | 0,57 | 0,58 | 0,55 | 0,62 | 0,48 | 0,29 | 0,21 | 0,23 | 0,58 |
| Rank + Linear | All | 0,55 | 0,67 | 0,67 | 0,55 | 0,67 | 0,67 | 0,64 | 0,64 | 0,58 | 0,55 | 0,55 | 0,46 | 0,41 | 0,43 | 0,42 | 0,54 | 0,61 | 0,55 | 0,56 | 0,55 | 0,55 | 0,56 | 0,61 | 0,36 | 0,29 | 0,17 | 0,23 | 0,51 |
| | No APC | 0,60 | 0,72 | 0,72 | 0,60 | 0,72 | 0,72 | 0,69 | 0,69 | 0,63 | 0,60 | 0,60 | 0,51 | 0,46 | 0,47 | 0,46 | 0,58 | 0,66 | 0,60 | 0,60 | 0,59 | 0,60 | 0,59 | 0,65 | 0,38 | | | | 0,60 |
| | No DSM | 0,61 | 0,73 | 0,73 | 0,61 | 0,73 | 0,73 | 0,70 | 0,70 | 0,64 | 0,61 | 0,62 | 0,52 | 0,48 | 0,49 | 0,47 | 0,59 | 0,67 | 0,61 | 0,61 | 0,60 | 0,61 | | | | | | | 0,62 |
| | Avg. | 0,58 | 0,70 | 0,70 | 0,58 | 0,70 | 0,70 | 0,67 | 0,67 | 0,61 | 0,58 | 0,59 | 0,49 | 0,45 | 0,46 | 0,45 | 0,57 | 0,64 | 0,58 | 0,58 | 0,58 | 0,58 | 0,57 | 0,63 | 0,37 | 0,29 | 0,17 | 0,23 | 0,57 |

Fig. 1. Rank and linear correlations averaged by metric for the four event detection datasets.

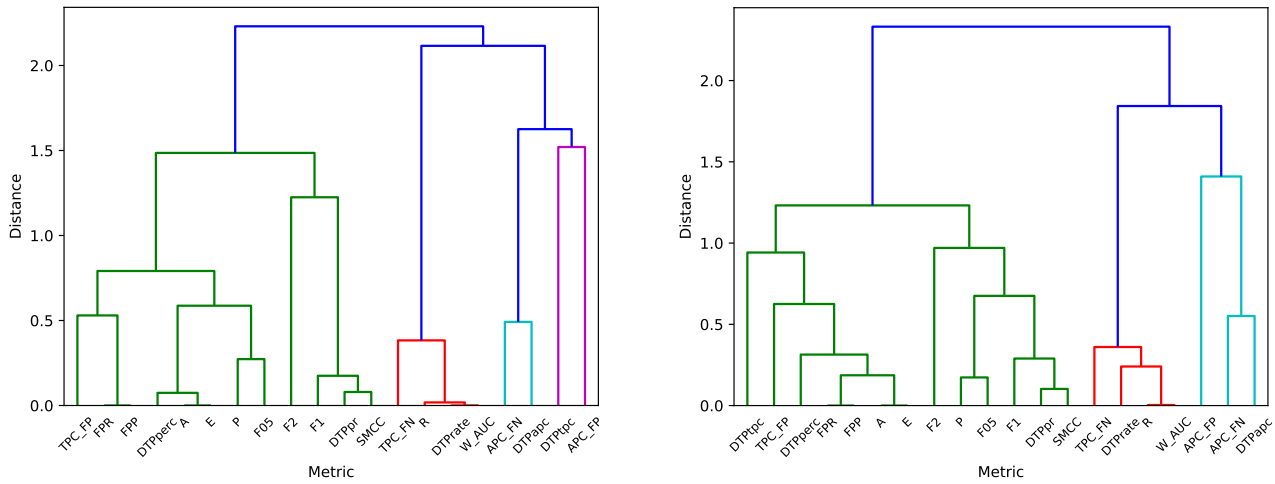


Fig. 2. Dendrograms showing rank (left) and linear (right) correlations of the performance metrics across datasets.

*problem*, and ii) *highlighting niches for which additional metrics should be studied and new ones created*.

For example, regarding the former, this work shows that the extremely unbalanced nature of the problem (towards $TN$) has several implications in the behavior of the performance metrics. As for the latter, this work also highlights the potential of $DSMs$ to unveil important characteristics of the underlying algorithms and datasets (e.g., $TPC_{FN}$ and $TPC_{FP}$). This can be helpful when choosing the most adequate metric or set of metrics to meet a specific goal.

## Acknowledgment

The authors would like to acknowledge Prof. Mario Bergés, for his valuable insights during the elaboration of this work.

## References

[1] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76–84, 2011.

[2] S. Makonin and F. Popowich, "Nonintrusive load monitoring (NILM) performance evaluation," *Energy Efficiency*, pp. 1–6, 2017.

[3] E. T. Mayhorn, G. P. Sullivan, J. M. Petersen, R. S. Butner, and E. M. Johnson, "Load Disaggregation Technologies: Real World and Laboratory Performance," Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep. PNNL-SA-116560, Sep. 2016.

[4] L. Pereira and N. J. Nunes, "A Comparison of Performance Metrics for Event Classification in Non-Intrusive Load Monitoring," in *Proceedings of the 2017 IEEE International Conference on Smart Grid Communications*. Dresden, Germany: IEEE, Oct. 2017.

[5] K. D. Anderson, M. E. Bergés, A. Ocneanu, D. Benitez, and J. M. F. Moura, "Event detection for Non Intrusive load monitoring," in *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, Oct. 2012, pp. 3312–3317.

[6] P. Meehan, C. McArdle, and S. Daniels, "An Efficient, Scalable Time-Frequency Method for Tracking Energy Usage of Domestic Appliances Using a Two-Step Classification Algorithm," *Energies*, vol. 7, no. 11, pp. 7041–7066, Oct. 2014.

[7] M. Berges, E. Goldman, H. Matthews, L. Soibelman, and K. Anderson, "User-Centered Nonintrusive Electricity Load Monitoring for Residential Buildings," *Journal of Computing in Civil Engineering*, vol. 25, no. 6, pp. 471–480, 2011.

[8] L. Pereira, "Developing and Evaluating a Probabilistic Event Detector for Non-Intrusive Load Monitoring," in *Proceedings of the fifth IFIP Conference on Sustainable Internet and ICT for Sustainability*. Funchal, Portugal: IEEE / IFIP, Dec. 2017.

[9] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges, "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research," in *2nd KDD Workshop on Data Mining Applications in Sustainability*, Beijing, China, Aug. 2012.

[10] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, no. 150007, 2015.

[11] H. Iba, Y. Hasegawa, and T. K. Paul, *Applied Genetic Programming and Machine Learning*. Boca Raton, FL, USA: CRC Press, Inc., 2009.

[12] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.