

# *Electronics Letters*

## Special issue Call for Papers

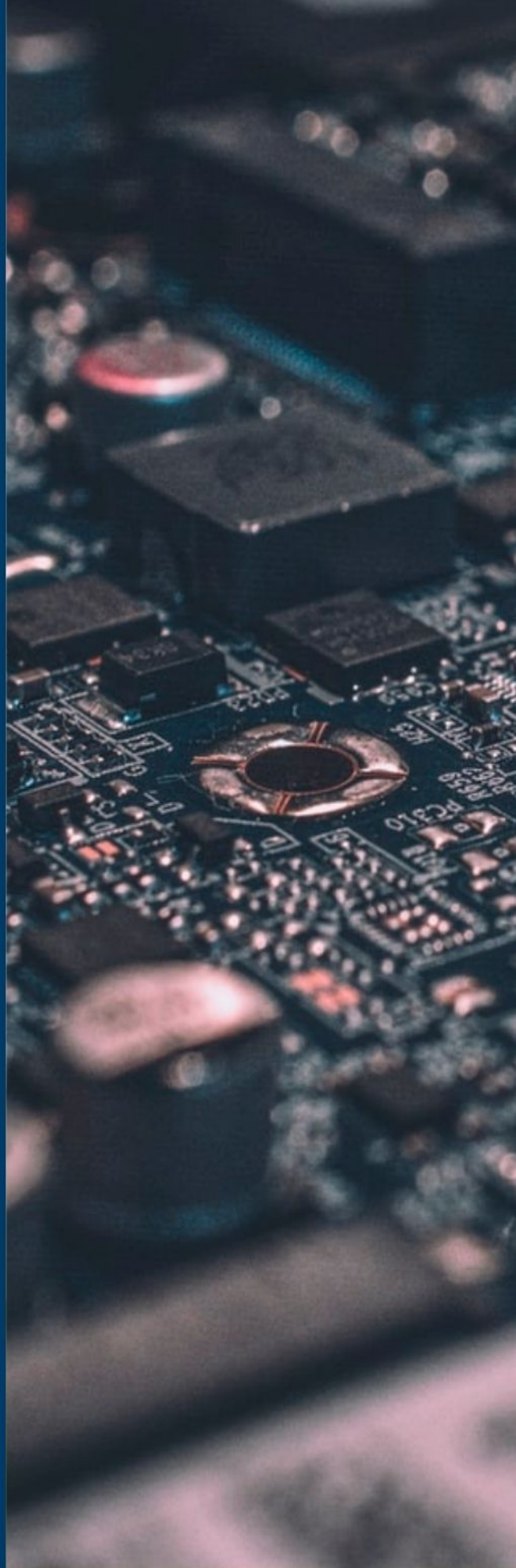
---

**Be Seen. Be Cited.  
Submit your work to a new  
IET special issue**

Connect with researchers and experts in your field and share knowledge.


Be part of the latest research trends, faster.

[Read more](#)



The Institution of  
Engineering and Technology

# Reliable assessment of uncertainty for appliance recognition in NILM using conformal prediction

Lorin Werthen-Brabants,  Tom Dhaene, and Dirk Deschrijver

IDLab, Ghent University - imec, Ghent, Belgium

 E-mail: lorin.werthenbrabants@ugent.be

A primary task of Non-intrusive Load Monitoring (NILM) is the identification of appliances that are switched on or off. However, state-of-the-art machine learning methods such as deep learning do not express uncertainty of their predictions. Especially in cases where appliances are confused, it is desirable that an NILM system can suggest multiple possible predictions to the end-user, including its confidence and credibility of any given prediction. This can be achieved using conformal prediction, being an effective way to quantify uncertainty of a given machine learning model. In this work, conformal prediction is introduced for NILM and applied to a neural network. The approach is explained and supported by several examples.

**Introduction:** Identifying appliances in NILM using machine learning methods, such as deep learning, relies on finding the class that has the highest probability of being the true class of the observation [1–10]. However, in cases where appliances are confused, other classes can also be likely, given the outputs of the model. Many existing appliance classification models do not express uncertainty of their predictions, which means there is less information supplied to the end-user or decision system. Although some approaches can inherently express uncertainty on their predictions [11–14], they do not readily return a set of possible appliances, and the confidence and credibility of the model are difficult to interpret. Entropy is used as a way of quantifying uncertainty, but only provides a crude notion of how much uncertainty a model exhibits.

**Inductive Conformal prediction (ICP)** [15, 16] only requires an extra holdout calibration dataset with switching events that have not been trained on, and a measure of conformity (such as the probability of a class).

Rather than only returning a single (most likely) label of an appliance class, using conformal prediction, the model can return any combination of possible classes, or no classes. This conformal prediction can also be tuned with a parameter  $\epsilon$  for the desired error rate, where an error corresponds to the true class not being present in the prediction set.

In this work, conformal prediction is applied to appliance recognition in NILM and it is shown how conformal prediction can adequately give uncertainty estimates by making use of prediction sets, confidence and credibility. These different concepts surrounding conformal prediction are explained in the context of appliance recognition, and applications are discussed.

**Inductive conformal prediction:** Conformal prediction [15, 17] can be classified in two types: *transductive* and *inductive* [16], with the former being more computationally expensive than the latter, as the underlying algorithm needs to be trained for every inference. With modern methods, such as neural networks, this is infeasible. Instead, inductive conformal prediction (ICP) is computationally efficient but requires a hold-out calibration dataset  $\mathcal{D}_{\text{calibration}}$ . Note that this can be applied to any appliance classification model that outputs scores. As many of these models are based on neural networks, the assumption is made that the model  $f$  outputs a probability  $P(y = \kappa | \mathbf{x})$  for any given class  $\kappa$ . The process is as follows:

1. Train the model on training dataset  $\mathcal{D}_{\text{train}}$ . The output of the resulting model  $f$  can easily be transformed to a measure of nonconformity as follows:  $\alpha_i = 1 - f(\mathbf{x}_i, y_i) = 1 - P(y = y_i | \mathbf{x}_i)$ .
2. Assign a nonconformity score  $\alpha_i$  to every sample in the calibration dataset  $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{calibration}}$ . The resulting nonconformity scores form set  $\mathcal{A}$ .
3. For a new sample  $\mathbf{x}'$ , perform the following: For every possible class  $\kappa \in \mathcal{K}$ , compute  $\alpha^{(\kappa)} = 1 - f(\mathbf{x}', \kappa)$ .

4. Calculate the  $p$ -value for every class  $\kappa$  by finding the proportion of nonconformity values  $\alpha_i$  that are greater than or equal to the current nonconformity value in question,  $\alpha^{(\kappa)}$ . This can be done as follows:

$$p_{\alpha^{(\kappa)}} = \frac{|\{\alpha_i \in \mathcal{A} | \alpha_i \geq \alpha^{(\kappa)}\}|}{|\mathcal{A}| + 1}. \quad (1)$$

The resulting  $p$ -values for every  $\kappa$  can be used for multiple purposes:

- Prediction set** Given an error-rate  $\epsilon$ , a set of classes  $C_\epsilon(\mathbf{x})$  can be returned for which  $p \geq \epsilon$ , also called the *prediction set*. The probability that the new sample is among the highest  $\epsilon\%$  of nonconformity scores should be low. Therefore, any  $p < \epsilon$  should be rejected. The prediction set should always contain the correct class, with the aforementioned error rate  $\epsilon$ , where an error indicates that the prediction set does not contain the class. The higher  $\epsilon$  is set, the smaller the prediction sets become and vice-versa. Usually  $\epsilon = 0.01$  or  $\epsilon = 0.05$  are used, to signify 99% and 95% confidence levels.
- Predicted class** The *predicted class* of the model is the class with the highest  $p$ -value, as this means there are many samples in  $\mathcal{D}_{\text{calibration}}$  with a higher nonconformity score than the current sample. The predicted class corresponds to the class with the highest output probability given by the model. This is not necessarily the case, however, when using variants such as label-conditional ICP [18], where another class with a lower output probability may have a higher  $p$ -value. This means the F1-score and accuracy of a model are unaffected by the use of ICP.
- Confidence** Using the second highest  $p$ -value  $p_x^{(2)}$  of the prediction,  $1 - p_x^{(2)}$  is the *confidence* of the prediction of the classifier. In other words, the confidence expresses the probability that the prediction set is a singleton  $\{\kappa\}$ . If this value is high, then no other candidates are likely to be in the prediction set, and the predicted value is almost certainly the only possible prediction.
- Credibility** The value of the highest  $p$ -value  $p_x^{(1)}$  is the *credibility* of the classifier in its prediction. The higher the value, the more likely it is that the prediction is correct. It must be noted that this value also denotes the highest error rate  $\epsilon$  for which the prediction set would be empty, as all candidates would be rejected at that value.

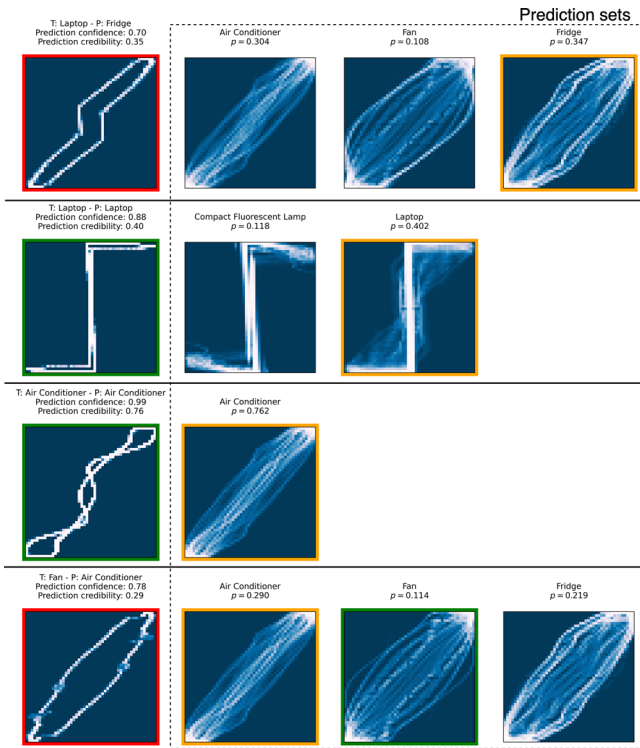
## Discussion:

**Experimental setup:** The performance of ICP is assessed on the PLAID dataset [1, 19], with a modified version of the Convolutional Neural Network (CNN) model proposed by De Baets et al. [2] as presented in [11]. The time series in PLAID are transformed into VI trajectories as described in [2, 20]. The basic structure of the CNN [11] is as follows: it takes as input a  $50 \times 50$  VI trajectory, normalized to contain values between 0 and 1. Next, a convolutional layer with 20 filters of size 5, a pooling layer, another convolutional layer with 20 filters of size 5, another pooling layer, a layer which flattens the resulting features and an output layer with  $K$  nodes are added, where  $K$  denotes the amount of classes available in a given training set.

The following tests are performed using a leave-one-house-out cross-validation. This means per fold a house is left out for the test set, and the remaining 54 houses are split into train (40 houses), validation (7 houses) and calibration (7 houses) sets.

The accuracy obtained with the model is 0.71, AUC ROC is 0.94 and F1-score is 0.68.

**Qualitative results:** Figure 1 shows concrete examples for ICP in NILM applied to PLAID with error rate  $\epsilon = 0.05$ , meaning in only 5% of the cases will the prediction set not contain the actual, ground truth class. The four samples were chosen at random from all test datasets in the



**Fig. 1** A visualization of four conformal predictions of a random subset of PLAID. On the left hand side the input VI trajectory can be seen. Above each trajectory is some context information: The true and predicted class (denoted with  $T$  and  $P$ ), the confidence in the prediction, and the credibility of the prediction. All predictions have their associated  $p$ -value included. A green border denotes a correct prediction, a red border a wrong prediction. On the right hand side the prediction sets can be seen with mean VI trajectories from the training data. The predicted, most likely class has an orange border, while a green border shows the correct class. If no green border is shown in the prediction set, the set does not contain the correct class.

leave-one-house-out cross-validation scheme. A discussion for every one of the four examples follows.

1. The first example shows a wrong prediction, with the prediction not being present in the prediction set. Here the confidence is high (0.70), but the credibility low (0.35). The confidence value indicates there is a high chance that the predicted output (fridge) is the only possible output. But the low credibility indicates the sample is not representative of the training data.
2. The second example is a correct prediction, albeit with low credibility. Analogous to the first example, we can conclude that according to the output probabilities in  $\mathcal{D}_{\text{calibration}}$ , the sample is not very representative of the input data, even though a correct prediction was given.
3. The third example shows a successful prediction with high confidence and credibility. A high credibility indicates how probable it is that the prediction is correct, given the possible predictions, with the confidence stating there is a low chance the prediction set is not a singleton.
4. The final example also has low credibility, meaning uncertainty that the sample is part of the training data. This time, however, the correct label is present in the prediction set.

**Conformal error ratio:** To quantify the improvement of error detection using ICP, the conformal error ratio can be calculated as follows:

$$\frac{P(y \neq \hat{y} | |C_\epsilon(\mathbf{x})| = k)}{P(y \neq \hat{y})}. \quad (2)$$

In other words, for a given efficiency  $\epsilon$  and ICP prediction set size  $k$ , calculate the ratio between the errors found in predictions with this set size and the general error rate ( $1 - \text{accuracy}$ ). When choosing  $\epsilon = 0.05$ , it is seen that the error ratio for  $k = 1$  is 0.17, whereas the error ratio for

$k > 1$  is 1.48. This shows that using ICP can greatly decrease the error rate when a singleton set is returned, while any set containing more than one class has a larger probability of containing an error. The classifier trained in this work produces a singleton set for 36.70% of the samples in the test sets.

**Practical notes on implementation for NILM:** The aforementioned confidence, credibility, most likely class and prediction set produced by ICP can all be used to supply more information to the end user. Assuming the end-user of the NILM system can get feedback and correct false predictions, the following list summarizes how such a system can make use of these concepts:

- The *prediction sets* can be used in cases where many classes exist, and the implementer of the NILM system wants to give the user an informed choice of possible appliances detected. This reduces the cognitive load on the user, as the application type will very likely be present in the presented set, according to error rate  $\epsilon$ .
- The *predicted class* can be utilized as it would normally be in an appliance recognition system, being the most likely class as predicted by the classifier.
- If the credibility is high enough, a threshold on *confidence* can be established to determine whether to request verification from the end-user for a prediction. Confidence serves as a metric that informs the user whether the prediction set is a singleton, meaning that a high confidence prediction is likely to be the only possible outcome.
- *Credibility* is the lowest error rate  $\epsilon$  for which the prediction set would be empty, with an error defined as the absence of the true class in the prediction set. Setting a threshold on credibility equivalent to setting the error rate  $\epsilon$ .
- By subtracting the two highest  $p$ -values  $s(\mathbf{x}) = p_{\mathbf{x}}^{(1)} - p_{\mathbf{x}}^{(2)}$  for a given sample  $\mathbf{x}$ , the implementer can use the resulting information  $I = 1 - s(\mathbf{x})$  as a measure of *informativeness* or uncertainty about the sample  $\mathbf{x}$ . The highest values of  $I$  contain the least amount of information and can be good candidates for further training in an active learning scenario [21–23].

These quantities can be very informative to the end-user of a NILM system to assess and correct an otherwise black-box model. However, due to the necessity of a calibration set, a sufficiently large dataset needs to be captured.

**Conclusion and future work:** This work motivates the utility and potential conformal prediction could provide for NILM systems. It requires minimal or no changes to a machine learning pipeline, and provides more information to supply to an end-user. Examples in a NILM context are discussed, and benefits to calibration are showcased. In future work, conformal prediction could be evaluated as part of a decision system. Also, label-conditional and online inductive conformal prediction could be explored, as NILM systems are imbalanced and ever changing, requiring frequent updates.

**Author contributions:** Lorin Werthen-Brabants: Conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing - original draft, writing - review and editing. Tom Dhaene: Funding acquisition, supervision, validation, writing - review and editing. Dirk Deschrijver: Funding acquisition, supervision, validation, writing - review and editing.

**Acknowledgements:** This research received funding from the Flemish Government under the ‘‘Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen’’ programme.

**Conflict of interest statement:** The authors have declared no conflict of interest.

**Data availability statement:** The dataset used in this study is openly available at [https://figshare.com/articles/dataset/PLAID\\_-\\_A\\_Voltage\\_and\\_Current\\_Measurement\\_Dataset\\_for\\_Plug\\_Load\\_Appliance\\_Identification\\_in\\_Households/10084619](https://figshare.com/articles/dataset/PLAID_-_A_Voltage_and_Current_Measurement_Dataset_for_Plug_Load_Appliance_Identification_in_Households/10084619).

Copyright © 2023 Wiley-VCH GmbH

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 28 March 2023 Accepted: 21 June 2023

doi: 10.1049/ell2.12860

## References

- 1 Medico, R., et al.: A voltage and current measurement dataset for plug load appliance identification in households. *Sci. Data* **7**(1), 49 (2020). <https://doi.org/10.1038/s41597-020-0389-7>
- 2 De Baets, L., et al.: Appliance classification using VI trajectories and convolutional neural networks. *Energy Build.* **158**, 32–36 (2018). <https://doi.org/10.1016/j.enbuild.2017.09.087>
- 3 Giri, S., Bergés, M.: An energy estimation framework for event-based methods in Non-Intrusive Load Monitoring. *Energy Convers. Manage.* **90**, 488–498 (2015). <https://doi.org/10.1016/j.enconman.2014.11.047>
- 4 Reinhardt, A., et al.: On the accuracy of appliance identification based on distributed load metering data. In: *2012 Sustainable Internet and ICT for Sustainability (SustainIT)*, pp. 1–9. International Federation for Information Processing, Laxenburg, Austria (2012)
- 5 Ruano, A., et al.: NILM techniques for intelligent home energy management and ambient assisted living: A review. *Energies* **12**(11), 2203 (2019). <https://doi.org/10.3390/en12112203>
- 6 Zhang, C., et al.: Sequence-to-point learning with neural networks for non-intrusive load monitoring. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 pp. 2604–2611 (2018)
- 7 Medeiros, A.P., et al.: Event classification in non-intrusive load monitoring using convolutional neural network. In: *2019 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America)*, pp. 1–6. IEEE, Piscataway (2019)
- 8 Khodayar, M., Wang, J., Wang, Z.: Energy disaggregation via deep temporal dictionary learning. *IEEE Trans. Neural Networks Learn. Syst.* **31**(5), 1696–1709 (2020). <https://doi.org/10.1109/TNNLS.2019.2921952>
- 9 Jiang, J., et al.: Deep learning-based energy disaggregation and on/off detection of household appliances. *ACM Trans. Knowl. Discovery Data* **15**(3), 50:1–50:21 (2021). <https://doi.org/10.1145/3441300>
- 10 Himeur, Y., et al.: Effective non-intrusive load monitoring of buildings based on a novel multi-descriptor fusion with dimensionality reduction. *Appl. Energy* **279**, 115872 (2020). <https://doi.org/10.1016/j.apenergy.2020.115872>
- 11 Werthen-Brabants, L., Dhaene, T., Deschrijver, D.: Uncertainty quantification for appliance recognition in non-intrusive load monitoring using Bayesian deep learning. *Energy Build.* **270**, 112282 (2022). <https://doi.org/10.1016/j.enbuild.2022.112282>
- 12 Bansal, V., et al.: “I do not know”: Quantifying uncertainty in neural network based approaches for non-intrusive load monitoring. In: *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 79–88. ACM, New York (2022)
- 13 Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds) *Proceedings of the 33rd International Conference on Machine Learning*, vol. **48**, pp. 1050–1059. PMLR, New York (2016)
- 14 Liu, Y., et al.: A home energy management system incorporating data-driven uncertainty-aware user preference. *Appl. Energy* **326**, 119911 (2022). <https://doi.org/10.1016/j.apenergy.2022.119911>
- 15 Shafer, G., Vovk, V.: A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9**(3), 371–421 (2008)
- 16 Papadopoulos, H.: Inductive Conformal Prediction: Theory and Application to Neural Networks. IntechOpen (2008)
- 17 Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **110**(3), 457–506 (2021). <https://doi.org/10.1007/s10994-021-05946-3>
- 18 Vovk, V.: Conditional validity of inductive conformal predictors. In: *Asian Conference on Machine Learning*, pp. 475–490. PMLR, New York (2012)
- 19 Gao, J., et al. PLAID: A public dataset of high-resolution electrical appliance measurements for load identification research. In: *BuildSys 2014 - Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pp. 198–199. ACM Press, New York (2014)
- 20 Hassan, T., Javed, F., Arshad, N.: An Empirical Investigation of V-I Trajectory Based Load Signatures for Non-Intrusive Load Monitoring. *IEEE Trans. Smart Grid* **5**(2), 870–878 (2014). <https://doi.org/10.1109/TSG.2013.2271282>
- 21 Settles, B.: Active learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **6**(1), 1–114 (2012). <https://doi.org/10.2200/S00429ED1V01Y201207AIM018>
- 22 Matiz, S., Barner, K.E.: Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification. *Patt. Recogn.* **90**, 172–182 (2019)
- 23 Ho, S.S., Wechsler, H.: Query by transduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(9), 1557–1571 (2008)