NILM: Multivariate DNN performance analysis with high frequency features

Camilo Mariño
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
Email: cmarino@fing.edu.uy

Elías Masquil
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
Email: emasquil@fing.edu.uy

Franco Marchesoni
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
Email: marchesoni@fing.edu.uy

Alicia Fernandez
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
Email: alicia@fing.edu.uy

Pablo Massaferro
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
Email: pmassaferro@fing.edu.uy

***Abstract*—In recent years we have seen deep neural networks (DNNs) appear in almost every signal processing problem. Non Intrusive Load Monitoring (NILM) was not an exception. A detailed evaluation of the supervised deep learning approach can provide powerful insights for future applications on the matter. In this work we improve a state of the art NILM system based on DNN, by including high frequency features and modifying the autoencoders' latent space dimension. Moreover, we introduce a novel dataset for evaluating NILM systems. This paper presents a contribution that adds relevant features as a multivariate input to the DNNs, based on high frequency measurements of the power. Furthermore, a thorough evaluation of the generalization capabilities of these models is presented, comparing results from public databases and those acquired locally in Latin America (LATAM), an underrepresented region on the NILM problem. The data and software generated are left of public access.**

***Index Terms*—NILM, DNN, Open Data, Deep Learning, Energy Disaggregation**

I. Introduction

Non Intrusive Load Monitoring (NILM) is a signal processing application, introduced by Hart [3] for estimating the individual electric consumption of a set of appliances by analysing the total power consumption of the house. This research area is in full development [2] [10] and is being addressed by our department in collaboration with the national electric power company (UTE).

The usage of deep neural networks (DNNs) to solve this problem was introduced by Kelly et al. [7] and achieved state-of-the-art results, improving the ones obtained by more traditional approaches [5] [12] [4]. In this work, we built over Kelly's solution, presenting modifications to those models in favour of using high frequency features. Furthermore, the models are adapted depending on the appliance of interest by changing the dimensions of the latent space of the autoencoders II-A.

The selection of the high frequency features is a result of a study of the appliance identification problem. This evaluation was carried out by the usage of Random Forest models and the Mutual Information criterium.

This work also presents an exhaustive evaluation of the performance of different DNNs over different datasets. As a result, a discussion on the generalization error of these models is presented.

Finally, we introduce a novel dataset for NILM releated research that was collected by the authors during this work. The dataset consists of high and low sampling-frequency voltage and current measurements of two local homes from Uruguay. This contribution takes special relevance in the context where NILM datasets are not abundant, being this one the first for the region. The location where NILM datasets are collected is relevant because of the different standards in the power supplies between regions and countries.

II. Proposed Solution

In this section we present our solution for NILM using DNNs. We built on top of the work of Kelly et al. [7], by including high frequency features as inputs to the DNNs and modifying the autoencoders' latent space dimension.

A. Models

Using as a baseline the architectures proposed in [7], we incorporated and evaluated some modifications. Out of the three architectures proposed in [7], we focus our work in two of them: *denoising autoencoder* and *start, stop and power regressor*. In Fig.1 a visual description of them is provided. Our modifications to those models are:

- Modify the first layer of the models to incorporate multivariate time series as input for adding high frequency features. Based on the results of experiments over a

[*] These authors contributed equally to this work.

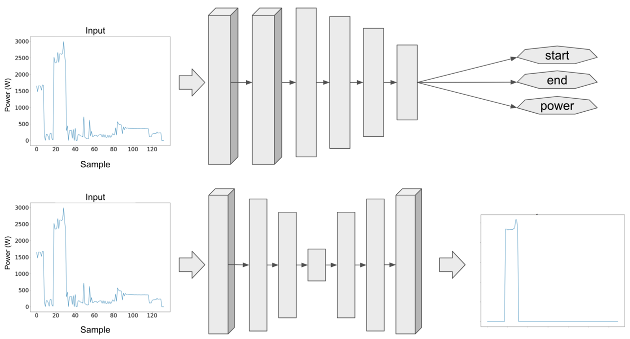

Fig. 1: Diagram of the implemented DNNs architectures. *Start, stop and power regressor* (top) and *Denoising autoencoders* (bottom). Convolution or upsampling layers are represented as 3D blocks. Fully-connected layers are represented as 2D blocks.

commercial current clamp [9] , we decided to add features based on samples up to $7kHz$. Up to the cited frequency, the clamp does not filter the signal, thus the signals may contain useful information.

- Increase the autoencoder number of layers and adjust its latent space dimension (code layer size) depending on the input size. Since the input size to the models depends on each appliance [7], it makes sense to adapt the latent space dimension as well. On another note, increasing the number of layers should increase the capacity of the model for solving the proposed problem.

B. Selection strategy

We implemented the following models:

- Baseline start, stop and power regressor
- Baseline denoising autoencoder
- Start, stop and power regressor with high frequency features
- Denoising autoencoder with high frequency features
- Denoising autoencoder with adjustable latent space dimension ("Custom")

A criteria must be established for selecting the best performing model for each of the appliances. We used the mean squared error as the training loss. As an evaluation metric we chose the Area Under the Curve (AUC) of the Receiving Operating Characteristic (ROC), as being able to classify whether an appliance was on or off was relevant to the task. Furthermore, after our initial experiments we found that high values of AUC were tied with visually appealing predictions. In order to transform the disaggregation problem into a classification one, for computing the ROC, binary labels were built from the dataset, as well as a criteria for considering the output of the models as a binary output. The definition of binary labels is described in Section IV-A and the binary output was obtained with a threshold for the maximum of the dissagregated power, i.e the output of the model.

We trained 7 models per appliance, one instance of each of the previously defined architectures plus two extra instances of the baselines trained using synthetic data as an addition to the training set. While in its strict sense those two extra instances are not different models, they are needed to evaluate the usage of synthetic data. For each instance of these 7 models we performed hyper-parameter optimization with a grid search strategy. In Table I we show the validation AUC for the 7 models for the microwave.

TABLE I: AUC for the 7 models for the microwave.

	Regressor	Autoencoder
Baseline	0.933	0.936
Synthetic Data	0.937	0.944
High Freq	0.927	**0.949**
"Custom"	-	0.932

TABLE II: Selected models.

	Selected model
Kettle	High frequency autoencoder
Fridge	High frequency regressor
Washing m.	High frequency regressor
Microwave	High frequency autoencoder
Dish washer	"Custom" autoencoder

The list of selected models, the ones with highest AUC, for each appliance is presented in Table II.

C. High Frequency Features

High frequency information was added to the model as a multivariate signal. Instead of having the univariate time series of the active power as an input, the modified input is a multivariate time series including the active power and other features derived from high sampling frequency measurements. For selecting which features to include as an input, we worked on a related but simpler problem: appliance identification. We used the Plug-Level Appliance Identification Dataset (PLAID) [1], which consist of more than 200 appliance instances from 11 appliance classes and more than 1000 records. Each record is the voltage and current measurement of one appliance at a sampling rate of 30kHz.

We calculated over 30 different features derived from the current and voltage, some of them being: statistical moments, features used in audio processing, VI trajectory image and power values [13]. Extracting the transient event was possible due to the records including the turning-on event for the appliances, which allowed us to compute the features for both the regime and the transient state separately.

This set of features was evaluated by the feature importance obtained by using a Random Forest (RF) classifier and the Mutual Information criterium (MI). Thus, the highest ranked features for appliance identification were selected to build the multivariate input for the NILM problem. In Fig. 2 transient feature importance is shown. Additional criteria for selecting these features were:

- Selected features should be both computable for transient and regime states (E.g. excluding the transient duration as a feature).

- Selected features should be one of the top 10 features by importance for both the transient and the regime by both importance criteria.

The features satisfying the criteria were **form factor of the current** and **phase shift between the fundamental component of the current and voltage**.

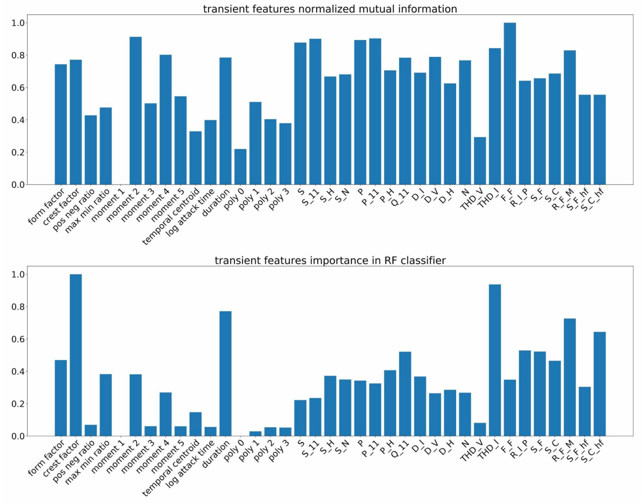

Fig. 2: Feature importance for the transient features. Evaluated by Random Forest (top) and Mutual Information (bottom)

Finally, the results over PLAID using different sets of features is presented in Table III. A 1-Nearest-Neighbor classifier was selected as a baseline. It is worth mentioning that our RF using all the features achieves state-of-the-art results, only being surpassed by 0.5% accuracy by the best result found in the literature [11].

TABLE III: Performance over different sets of features.

Samples	1074		1793	
Subset of features/Model	KNN	RF	KNN	RF
Transient	61.70	88.68±0.17	59.35	87.06±0.06
Steady state	75.88	87.24±0.28	66.76	84.23±0.25
Steady state + Transient	-	91.47±0.09	-	88.33±0.25
Steady state + VI	75.97	86.67±0.49	66.82	84.14±0.43
All	-	**92.79** ± 0.13	-	89.08±0.38

*VI means pixels of the VI image. The reported margin of error is the standard deviation between runs.

III. Local dataset

For evaluating the generalization capabilities of the proposed algorithms, we collected a local dataset for NILM in Latin America (LATAM), Uruguay. A complete explanation of the process, which included building custom meters, can be found in our previous work [9].

A. Data description

The dataset consists of electrical measurements from two houses for a time period of ≈3 months. The local utility installation has a nominal frequency of 50Hz and a nominal voltage of 230 V_{RMS}. Aggregated active power measurements were taken at a sampling rate of 14 kHz, and individual power measurements, at appliance level, were collected at a 1 minute sampling period. At the first house, we collected 7 individual measurements: fridge, electric water heater, microwave, air conditioner, general purpose plugs from the bedrooms, and washing machine. At the second house we collected 8 individual measurements: electric oven, electric water heater, two air conditioners, washing machine, fridge, kettle and dishwasher. It is worth to note that those appliances account for the major part of the house power usage.

B. Data release

The collected dataset can be used for further research in this exact topic and in other related areas. Electrical power standards varies between countries and regions, thus having datasets from different regions is important. What is more, LATAM is a region that is not represented in other popular datasets from the literature. A sample of the dataset can be found in the project's github repository as well as instructions on how to get access to the full dataset. [8]

IV. Experiments

To evaluate the results of the models we selected two sets of metrics. The first one, metrics for the regression problem, measure the reconstruction error between the target signal and the output from the model. The second one, classification metrics, evaluates the detection power of the models, in the sense of only considering if the appliance was on or not. While regression metrics can be affected if the prediction has a time shift respect to the target, classification metrics are not affected by time shifts, and only consider the appliance state in the time window of consideration.

The regression metrics are the Mean Absolute Error (MAE) and the Relative Error In Total Energy (REITE):

$$\text{REITE} = \frac{|\hat{E} - E|}{\max(\hat{E}, E)} \quad \text{MAE} = \frac{1}{N}\sum_{t} |\hat{y}_t - y_t| \tag{1}$$

where N is the number of samples considered. The classification metrics are the usual, precision, recall, accuracy, F1-score and AUC-ROC as mentioned in [14].

A. Training and evaluation data

The input to every model is a power time series or a multivariate time series including active power and the features derived in section II-C. The sampling period of those series is 6 seconds, matching the sampling rate of the training data IV-A, its length (window size) depends on the appliance.

Data preprocessing: The training dataset was composed only by data from UK-DALE dataset [6] . This data consist of measurements from 5 houses from the UK. Three of them also include high sampling frequency measurements. To match the sampling frequency used in the models, data from the local dataset was upsampled using a first order hold.

Dataset division: We split the dataset into non-overlapping subsets: training set, validation set, and 4 test sets. For each appliance, test set I consists of all the measurements from one of the 5 houses from UK-DALE. Test set II is formed with the last two weeks of data from the houses used during training. Test set III and IV are the data from the two local houses. From the rest of the data, training and validation sets were built. For building those two last sets, all the activations (i.e windows where the appliance was active as defined in [7]) were extracted using an activation-extracting function. An approximately equal number of non-activations (i.e periods between two activations) were also extracted. Finally, this activations dataset was split into the training set (80%) and validation set (20%).

Binary labels: The binary labels for each time window were True if the appliance activation was completely included in the window or False if the window consisted of a non-activation.

Synthetic data: As a form of data augmentation we created synthetic data. This process consisted in adding other appliances activations to a given appliance activation with some probability p. Being $p = 0.4$ the distraction probability value used.

The regressors output 3 values: the start time, the stop time of the appliance activation and the mean power consumed by it. The autoencoder's output is an univariate time series representing the power consumption of the appliance. The length of the autoencoders' output is the same as the input's length.

B. Evaluation

To evaluate the generalization capabilities of these models for NILM, four questions were formulated:

1) Do the models work in the context they were trained on? - Evaluation over test set II using "activations"
2) Are the models able to generalize to unseen appliances? - Evaluation over test set I using "activations"
3) Do the models work for a real use case? - Evaluation over test set I using "rolling windows"
4) How does the generalization error behave when the dataset became more diverse? - Evaluation using test set III and IV

The evaluation over "activations" considers windows with activations and non-activations, makes predictions for each window and compares them with the ground truth. This method is the one that was used for training and validation and consists in an approximately balanced evaluation. On the other hand, the "rolling windows" evaluation is closer to a real life application of NILM and results in an imbalanced evaluation. The process is the following: start by estimating the output for a rolling window with a stride of 1. I.e. estimate the first output, shift the window by 6 seconds, estimate the second output, and so on. Finally, for every time instant there will be `w=window_size` estimates, which are averaged and scaled by a factor $s_f = \frac{w}{w-2avg}$, being avg the average activation length for the application. This averaging and normalization are required because our models are trained to only detect full activations. After this process, the output time series is divided in non-overlapping windows which are compared with the ground truth.

V. Results

A positive answer to question 1 from Section IV-B can be found in the two rightmost columns of Table **??**, three out of the five best models in validation are still the best over test set II. The models perform well for appliances seen during training.

Then, in the two leftmost columns of Table **??** can be seen how the performance gets worse over unseen appliances, with a possible answer for question 2. Although the results remain acceptable, the generalization capability to newly data is limited.

Answer to question 3 can be found in tables V and VI with the results of the "rolling window" evaluation. The models achieve acceptable results in a real use case, especially in the classification problem.

TABLE V: AUCs scores using rolling window methodology.

	Test set II	Test set I
Kettle	1.000	0.998
Fridge	0.854	0.751
Washing m.	0.763	0.864
Microwave	0.973	0.956
Dishwasher	0.898	0.962

TABLE VI: Results across Test Set I of the selected models with rolling window methodology.

	Acc.	Prec.	Recall	F1	MAE	REITE
Kettle	0.987	0.686	0.967	0.802	22.48	0.609
Fridge	0.585	0.545	0.959	0.695	42.04	0.305
Washing m..	0.612	0.107	0.904	0.191	237.98	0.962
Microwave	0.835	0.019	0.964	0.038	58.48	0.173
Dish washer	0.961	0.679	0.743	0.710	45.00	0.639

Finally, an answer to question 4 can be found with the results over the local houses, see table VII. Performance gets much worse when evaluating over a diverse dataset containing appliances very different from different regions.

VI. Conclusions

High frequency features were studied in the appliance identification problem. By measuring feature importances, two of them were selected as the most relevant: form factor and phase shift. Modifications to the baseline architectures from [7] were successfully implemented to add that information and to make the autoencoder more flexible to the task. In this line, the selected models from our study were all a result of those modifications. The best performing models in validation for 4 of 5 appliances included high frequency features, and the other was the modified "Custom" autoencoder. Good results were obtained by evaluating the models over seen appliances, matching the validation results: adding high frequency features introduce relevant information to the task.

Regarding generalization, the studied models are affected when varying the evaluation datasets. While these models

TABLE VII: Results for the local dataset with rolling window methodology.

	House 1						House 2					
Appliance	Accuracy	Precision	Recall	F1	MAE (W)	REITE	Accuracy	Precision	Recall	F1	MAE (W)	REITE
Kettle	-	-	-	-	-	-	0.953	0.286	0.545	0.375	75	0.909
Fridge	0.759	0.781	0.959	0.861	143	0.507	0.918	0.990	0.926	0.957	96	0.071
Washing m.	0.071	0.057	1.000	0.107	793	0.995	0.323	0.300	1.000	0.462	691	0.971
Microwave	0.506	0.018	0.650	0.036	71	0.920	0.533	0.066	0.818	0.122	90	0.880
Dish washer	-	-	-	-	-	-	0.859	0.720	0.720	0.720	150	0.478

showed great performance over appliances seen during training (see Figure 3) the performance decreased as the datasets became more distinct. Results over appliances from the same database but different houses were worse than over known appliances. Finally, results over appliances from a completely different database were much worse than over the rest.

In addition to the experimental results we present a novel dataset for evaluation and research on NILM related topics that can be openly accessed.

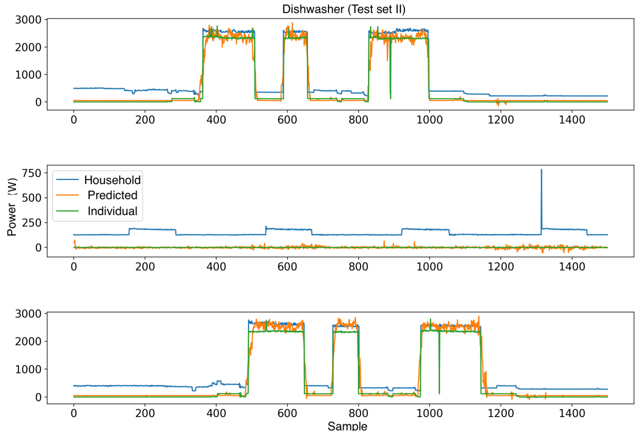

Fig. 3: Prediction for the dishwasher from test set II

Acknowledgments

The authors acknowledge the financial support provided by Uruguay's National Research and Innovation Agency (ANII), the Julio Ricaldoni Foundation and the National Administration of Power Plants and Electrical Transmissions (UTE). Also thank the support from Juan Pablo Maggi who generously shared his insights in the first stages of this project.

References

[1] Jingkun Gao, Suman Giri, Emre Can Kara, and Mario Bergés. Plaid: a public dataset of high-resoultion electrical appliance measurements for load identification research: demo abstract. In *proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 198–199. ACM, 2014.

[2] E. Gomes and L. Pereira. Pb-nilm: Pinball guided deep non-intrusive load monitoring. *IEEE Access*, 8:48386–48398, 2020.

[3] George William Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[4] Simon Henriet, Benoit Fuentes, Umut Şimşekli, and Gaël Richard. Matrix factorization for high frequency non intrusive load monitoring: Definitions and algorithms. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, pages 20–24, 2020.

[5] Jack Kelly, Nipun Batra, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani Srivastava. Nilmtk v0.2: A non-intrusive load monitoring toolkit for large scale data sets. In *The first ACM Workshop On Embedded Systems For Energy-Efficient Buildings at BuildSys 2014*, Memphis, USA, 2014.

[6] Jack Kelly and William Knottenbelt. The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes. *Scientific data*, 2:150007, 2015.

[7] Jack Kelly and William J. Knottenbelt. Neural NILM: deep neural networks applied to energy disaggregation. *CoRR*, abs/1507.06594, 2015.

[8] Franco Marchesoni-Acland, Camilo Mariño, and Elías Masquil. Nilm-uy. https://github.com/camilomarino/NILM-UY, 2021.

[9] Franco Marchesoni-Acland, Camilo Mariño, Elías Masquil, Pablo Masaferro, and Alicia Fernández. End-to-end nilm system using high frequency data and neural networks, 2020.

[10] Antonio Ruano, Alvaro Hernandez, Jesus Ureña, Maria Ruano, and Juan Garcia. Nilm techniques for intelligent home energy management and ambient assisted living: A review. *Energies*, 12(11):2203, 2019.

[11] Nasrin Sadeghianpourhamami, Joeri Ruyssinck, Dirk Deschrijver, Tom Dhaene, and Chris Develder. Comprehensive feature selection for appliance classification in nilm. *Energy and Buildings*, 151:98–106, 2017.

[12] Shikha Singh and Angshul Majumdar. Deep sparse coding for non–intrusive load monitoring. *IEEE Transactions on Smart Grid*, 9(5):4669–4678, 2017.

[13] K. Yumak and O. Usta. A controversial issue: Power components in nonsinusoidal single-phase systems. In *2011 7th International Conference on Electrical and Electronics Engineering (ELECO)*, pages I–157–I–161, Dec 2011.

[14] Michael Zeifman and Kurt Roth. Nonintrusive appliance load monitoring: Review and outlook. *IEEE transactions on Consumer Electronics*, 57(1):76–84, 2011.