BAYESIAN-OPTIMIZED BIDIRECTIONAL LSTM REGRESSION MODEL FOR NON-INTRUSIVE LOAD MONITORING

Maria Kaselimi[1], Nikolaos Doulamis[1], Anastasios Doulamis[1], Athanasios Voulodimos[1,2] and Eftychios Protopapadakis[1]

[1] National Technical University of Athens, Athens 15773, Greece
[2] University of West Attica, Athens 12243, Greece

ABSTRACT

In this paper, a Bayesian-optimized bidirectional Long Short -Term Memory (LSTM) method for energy disaggregation, is introduced. Energy disaggregation, or Non-Intrusive Load Monitoring (NILM), is a process aiming to identify the individual contribution of appliances in the aggregate electricity load. The proposed model, Bayes-BiLSTM, is structured in a modular way to address multi-dimensionality issues that arise when the number of appliances increase. In addition, a non-causal model is introduced in order to tackle with inherent structure, characterizing the operation of multi-state appliances. Furthermore, a Bayesian-optimized framework is introduced to select the best configuration of the proposed regression model, thus increasing performance. Experimental results indicate the proposed method's superiority, compared to the current state-of-the-art.

Index Terms— Deep Learning, Recurrent Networks, LSTM, Bayesian Optimization, Energy Disaggregation

1. INTRODUCTION

A detailed monitoring of energy consumption at appliance-level should be performed to improve domestic energy efficiency, resulting in an overall picture of the Distributed Energy Resources (DER's) characteristics [1]-[3]. Then, recommendations are given to users to take action for energy savings [4]. There is a variety of sensors that can measure the power characteristics of an appliance, providing a reliable picture of the electricity load. However, such *intrusive methods* are expensive mainly due to the sensors' installation cost. On the contrary, *non-intrusive techniques* are able to estimate the appliances' energy consumption by exploiting the total (aggregate) energy signal as measured by the household smart meter interface. This is achieved by combining machine learning and signal processing tools. Using NILM estimates, users know the distribution of the measured aggregate power load signal per appliance [2].

Such a decomposition requires advanced signal processing tools since (i) the energy signatures of the appliances are not independent with each other, (ii) appliances arbitrarily switch ON/OFF, and (iii) each appliance has different operational models, which contribute quite differently to the total consumed power load [2].

1.1. Related Work

NILM approaches can be discriminated into methods using classification, clustering or regression schemes, or methods exploiting optimization techniques. Classification/clustering approaches are based on either supervised or unsupervised learning paradigms. Hart was the first to propose a method for disaggregating electrical loads through the clustering based on appliances' characteristics [5]. The limitation of clustering-based methods, is that they cannot predict the power load of an appliance since there are no supervised training samples. Other methods rely on Dynamic Time Warping (DTW) [6], matrix factorization [1], neuro-fuzzy modelling to handle uncertainties [7], appliance load modelling [8], [9] or even graph-based representations [10]. The drawback of all the aforementioned methods is that they handle the classification problem as one input-output linear or non-linear relationship. However, as the number of appliances (or the number of operations per appliance) increases, the targeted classes also increase exponentially, deteriorating the classification performance (the so-called *curse of dimensionality)* [11].

Other methods exploit optimization principles in which the optimal combination of appliances' signal has as a result an estimated aggregate signal close to the real total measured power. Hidden Markov Models (HMMs) and various extensions [12]-[14] or hybrid methods [15] are proposed. Again, the main limitation of these approaches is that as the appliances increase, the state space exponentially increases, making energy disaggregation impossible to be practically implemented. Recently, deep learning [16] has been investigated for energy disaggregation. Particularly, the works of [16], [18] investigate the use of Convolutional Neural Networks (CNNs). The main drawback of CNNs is that they have no recurrent properties, which is an important issue; appliance power load is a highly temporal dependent time series. To address this drawback, recurrent LSTM networks [19] have been introduced for NILM [20], [21].

1.2. Our Contribution

The proposed appliance-based, Bayesian-optimized BiLSTM regression model satisfies a set of crucial characteristics making it superior than the other previous methods. The arising limitations and drawbacks of these methods are addressed in this paper with the adoption of a modular, non-causal sequence-to-sequence regression model. The proposed model's specific features are summarized below:

Long Term Regression: NILM is often addressed, in the literature, as a classification problem, i.e., estimate the operational states of an appliance (e.g. ON/OFF or multi-state). However, such approaches have the drawback that significant information (power fluctuations) regarding the electricity load is lost. On the contrary, this work addresses NILM as a sequence-to-sequence regression problem, thus allowing to maintain all the necessary information. Additionally, existing long term dependencies should be accounted for, increasing regression performance.

Modularity: Usually, current approaches (e.g. HMM-based ones) handle energy disaggregation as a single input-output relationship. In this context, as the number of appliances increases, the complexity of model combinatorially increases. Such dimensionality issues are addressed in this work. In particular, the adopted procedure is conducted for each device separately with an appliance-based, modular and extensible model. Therefore, an increase of the expected number of detected appliances does not introduce additional complexity to the proposed model.

Optimization: Hyperparameters' tuning in a deep network is a major issue. In general, hyperparameters are not optimized and are assumed to be fixed throughout time. However, seasonal attributes affect appliance electricity loads, influencing energy disaggregation performance. Bayesian optimization strengthens model performance through the optimal hyperparameters selection, creating a unique optimal model, adaptable to each appliance's individual settings and seasonal variations.

Non-causality: Existing methods assume causal signal dependencies. However, the way that an appliance operates is often non-causal. In particular, the current state of a power load may have some dependence on future states (e.g. in a washing machine, pre-washing cycle always precedes washing cycle). In our approach, non-causality is achieved by modifying the conventional LSTM taking into account both previous and future states of electricity power load. Bidirectional recurrent regression deep models are thus adopted for NILM. Bidirectionality has been first introduced in [22] for shallow learning paradigms and especially for Recurrent Neural Networks (RNNs). The non-causality concept has been extended for LSTM networks [23] indicating promising results for speech recognition. To the authors' knowledge, there are no research works that investigate bidirectionality and long term dependence for energy disaggregation, incorporating a probabilistic Bayesian framework to optimally select the network parameters.

2. BIDIRECTIONAL SHORT-TERM RECURRENT REGRESSION FOR NILM

2.1. Notation and Problem Formulation

Let *M* be a set of all known household's appliances. Let $p(t_n)$ be the aggregate measured energy signal. Assuming discrete times $p(t_n) = p(nT) = p(n)$, where $T = t_n - t_{n-1}$ is the sampling interval. Let us now denote as $p_j(n)$ the active power load of j-th appliance out of *M* available. Furthermore, we assume that we have M independent models, each corresponding to an appliance, resulting in a modular framework; each time a new appliance is added, a new network is built. We can express $p(n)$ as:

$$p(n) = \sum_{j=1}^{M} p_j(n) + e(n) \tag{1}$$

where $e(n)$ denotes the additive noise of the measurements.

In an NILM modelling framework, the measurements $p_j(n)$ are not available, since there are no smart plugs installed. Instead, only $p(n)$ is given. Therefore, the problem is to estimate $p_j(n)$ from $p(n)$. Each appliance has a unique spectral signature. This is the main principle we exploit to decompose the aggregate signal $p(n)$ into its components $p_j(n)$. The spectral signatures of signal are actually derived as an integration of the values of the signal over time. Thus, in order to get the estimates $\hat{p}_j(n)$ of $p_j(n)$, we need to assemble measurements of the aggregate signal $p(n)$ over a time window K+1, thus, $\mathbf{p}(n) = [p(n) \quad \cdots \quad p(n-K)]^T$. Then, the values $p_j(n)$ can be expressed as an non-linear relationship of $\mathbf{p}(n)$. Therefore, we have that

$$p_j(n) = f(\mathbf{p}(n)) + e(n) = \hat{p}_j(n) + e(n) \tag{2}$$

2.2. Bidirectional Short-Term Recurrent Regression

One way to approximate the unknown relationship $f(\cdot)$ is through a feed-forward neural network [24], then

$$\hat{p}_j(n) = \mathbf{u}_j(\mathbf{n})^T \cdot \mathbf{v}_j \tag{3a}$$

$$\mathbf{u}_j(n) = \begin{bmatrix} u_{j,1}(n) \\ \vdots \\ u_{j,L}(n) \end{bmatrix} = \begin{bmatrix} tanH(\mathbf{w}_{j,1}^T \cdot \mathbf{p}(n)) \\ \vdots \\ tanH(\mathbf{w}_{j,L}^T \cdot \mathbf{p}(n)) \end{bmatrix} \tag{3b}$$

tanH refers to the hyperbolic tagent, while $\mathbf{w}_{j,i}$ =, *i*=1,…,*L*, are weights connecting the input and the *i*-th hidden neuron. Index *n* indicates the *n*-th time period. Vector **u**(n) is actually a state vector, gathering all hidden layer responses $u_{j,i}$. Index j refers to the j-th appliance in which the regressor is built. These non-linear transformations are linearly combined to provide the estimate of $\hat{p}_j(n)$, using a set of weights $\mathbf{v}_j$. In the following, we omit subscript j for simplicity, since we refer to a particular j-th regressor. Since the appliances randomly become dynamically active/inactive, the state vector $\mathbf{u}(\mathrm{n})$ depends on its previous values. This means that,

$$u_i(n) = g(\boldsymbol{w}_i^T \cdot \boldsymbol{p}(n) + \boldsymbol{r}_i^T \cdot \mathbf{u}(n-1)) \tag{4}$$

where $\mathbf{r}_i$ is a set of parameters that weigh the contribution of $\mathbf{u}(\mathrm{n}-1)$ to the current state values. Eq. (4) actually models a *recurrent regression of short range*. Since appliance power load dependencies are in fact non-causal signals, relations on both previous and future states should be taken into account. The idea is to split the model into two parts; *the* *forward* (relate the previous) *and the backward pass* (relate the future) [22]. That is,

$$u_i(n) = g(\mathbf{w}_i^T \cdot \mathbf{p}(n) + \vec{\mathbf{r}}_i^T \cdot \mathbf{u}(n-1) + \cdot \overleftarrow{\mathbf{r}}_i^T \cdot \mathbf{u}(n+1)) \quad (5)$$

3. BAYESIAN-OPTIMIZED BIDIRECTIONAL LSTM REGRESSION FOR NILM

3.1. Bidirectional LSTM for NILM

Household appliances follow repeated patterns span on long time periods, implying that short range dependency is not adequate. For instance, a washing machine follows several operational cycles (e.g. pre-washing, washing, drying, etc.) each related with each other in a long range. For this reason, a bidirectional LSTM network is adopted in this paper as the basic regression model for power load estimation. Each node is a memory cell that contains three different components (see Fig.1); (i) *the forget gate,* (ii) *the input gate and the input node, and* (iii) *the output gate.*

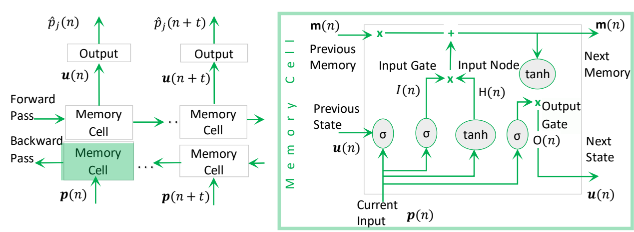


Fig. 1. Bidirectional long range recurrent regression model and the respective memory cell.

The forget gate: The purpose of this component is to throw the unnecessary information out of the memory cell. The output ranges between 0 and 1 -values close to 0 mean to dispose the incoming information, while values close to 1 indicate the worth-remembering information.

The input node/gate: The input node activates appropriately the respective state (true or false output from the "*tanH*" activation). Instead, t*he input gate* regulates whether the respective hidden state is "significant enough" on the regression model; sigmoid operation.

The output gate: This regulates whether the response of the current memory cell is "significant enough" to contribute to the next memory cell.

$$\{f(n), I(n), h(n), O(n)\} = \{\sigma, tanH\}$$
$$(\mathbf{w}^{T,\{f,I,h,O\}} \cdot \mathbf{p}(n) + \vec{\mathbf{r}}^{T,\{f,I,h,O\}} \cdot \mathbf{u}(n-1) + \overleftarrow{\mathbf{r}}^{T,\{f,I,h,O\}} \cdot \mathbf{u}(n+1)) \quad (6)$$

3.2. Bayesian Optimization

One critical aspect in our design is the selection of the configuration parameters of the proposed network (Bayes-BiLSTM). This paper, instead of applying the traditional manual-based tuning of the model parameters, adopts a probabilistic Bayesian framework through which the model configuration parameters are optimally tuned.

Let us assume that a certain number of configuration parameters is available, such as the number of memory cells, the learning rates, etc., denoted as $\boldsymbol{\pi}_i$. If we construct a set of Q different configurations, i.e., $D_{1:Q} = \{\boldsymbol{\pi}_1 \ \cdots \ \boldsymbol{\pi}_Q\}$, then, we can evaluate the error $E(\mathbf{p}, \mathbf{d}, \boldsymbol{\pi})$ that the network gives when (i) it receives as inputs the data p (i.e., a temporal-time series-collection of aggregate energy signals over a time window) (ii) the network output is compared against the desired (target) outputs d and (iii) a given $\boldsymbol{\pi}$ model configuration. In this context, we have omitted index n, since we refer to any time instance. We assume Mean Square Error (MSE). Let us denote as E_{min} the minimum across all Q configurations. Then, an improvement function is given:

$$I(\mathbf{p}, \mathbf{d}, \boldsymbol{\pi})) = \max\{0, E_{min} - E(\mathbf{p}, \mathbf{d}, \boldsymbol{\pi})\} \quad (7)$$

Assuming a probabilistic framework, we estimate

$$Expect(I(p, d, \pi)) = Expect(max\{0, E_{min} - E(p, d, \pi)\}) \quad (8)$$

Eq. (8) can be solved only if we know the probability distribution of the error function given a set of configurations, that is, $P(E|D_{1:Q})$. Exploiting the Bayesian rule we can express this probability as

$$P(E|D_{1:Q}) \propto P(D_{1:Q}|E)P(E) \quad (9)$$

Usually $P(E)$ follows a Gaussian distribution and $P(D_{1:Q}|E)$ is then expressed as a Gaussian process of mean $\mu(\boldsymbol{\pi})$ and standard deviation $\boldsymbol{\Sigma}$ [25]:

$$\boldsymbol{\Sigma} = \begin{bmatrix} k(\boldsymbol{\pi}_1, \boldsymbol{\pi}_1) & \cdots & k(\boldsymbol{\pi}_1, \boldsymbol{\pi}_Q) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{\pi}_Q, \boldsymbol{\pi}_1) & \cdots & k(\boldsymbol{\pi}_Q, \boldsymbol{\pi}_Q) \end{bmatrix} \quad (10)$$

where $k(\cdot)$ is a kernel function. The target of our optimization is to find out a new configuration $\boldsymbol{\pi}^* \equiv \boldsymbol{\pi}_{Q+1}$, which will further reduce the MSE or equivalently increase the improvement $I(\mathbf{p}, \mathbf{d}, \boldsymbol{\pi}^*)$. Then, for the new augmented set $D_{1:Q+1}$ that includes $\boldsymbol{\pi}^* \equiv \boldsymbol{\pi}_{Q+1}$, $P(D_{1:Q+1}|E)$ will again be a Gaussian process of standard deviation

$$\begin{bmatrix} \boldsymbol{\Sigma} & \mathrm{b} \\ \mathrm{b}^T & k(\boldsymbol{\pi}_{Q+1}, \boldsymbol{\pi}_{Q+1})) \end{bmatrix} \quad (11)$$

where $\mathrm{b} = [\mathrm{k}(\pi_{Q+1}, \pi_1) \ldots \mathrm{k}(\pi_{Q+1}, \pi_Q)]$. According to [25], it can be proven that the $P(E_{Q+1}|D_{1:Q}, \pi_{Q+1})$ is also a Gaussian with mean value and standard deviation related with previous variables. The new configuration $\boldsymbol{\pi}^*$ is estimated through Eq.(8), which is actually the integral of $I(\cdot)$ and $P(E_{Q+1}|D_{1:Q}, \pi_{Q+1})$, that is the probability that $I(\cdot)$ follows.

4. EXPERIMENTAL EVALUATION

4.1. Dataset description & experimental setup

The evaluation of the proposed method is conducted on the public AMPds dataset [26]. The AMPds contains electricity consumption data from a single house, in the greater area of Vancouver-Canada, including 21 individual circuits at one minute interval over a 2-year period. The selected appliances are of multi-state, making difficult its respective power load estimation (see Fig.2 with grey filled color). For each appliance, we have built a Bayes-BiLSTM. Bayesian optimization is used for optimally estimating the structure of each Bayes-BiLSTM.

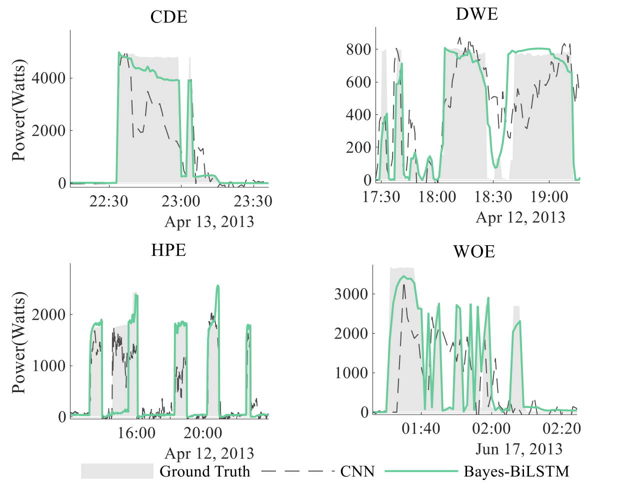
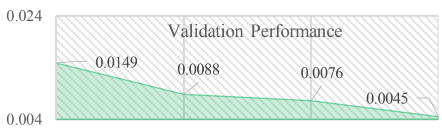


Fig. 2. Comparison of the proposed method (green solid line) with CNN (black dotted line) as well as ground truth (in gray).



Completed iterations	1	2	3	4
# of BiLSTM Layers	1	1	2	1
Hidden units	74	77	42/34	60
Learning Rate	0.093	0.0022	0.0438	0.0024

Fig. 3. Bayesian Optimization Results for CDE.

4.2. Performance Evaluation and Comparisons

Performance evaluation has been performed among Bayes-BiLSTM and other approaches, such as (i) CNNs [15], [18], (ii) unidirectional LSTMs [20], [21], (iii) combinatorial optimization (CO) [5] and (iv) a Factorial Hidden Markov Model (FHMM) [12]-[14]. The last two are benchmarked techniques in NILM toolkit (NILMTK) [27] used in power society for comparing various NILM methods.

Fig.2 shows signature identification examples for four selected appliances [clothes dryer (CDE), dishwasher (DWE), heat pump (HPE) and wall oven appliance (WOE)]. In this figure, we illustrate the performance of our proposed method compared to that of CNN as well as ground truth. As observed, our approach yields better performance in estimating not only ON/OFF states, but also more complicated energy patterns. This is important since it offers insight on whether a device is active as well as how it contributes to the total energy consumption. Table 1 presents the results of the comparison among the proposed method and other techniques based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Normalized RMSE (NRMS), which are commonly used metrics for the evaluation of energy disaggregation [2]. As observed, the proposed method outperforms the compared ones. Our proposed Bayes-BiLSTM as well as unidirectional LSTM generally perform best mainly due to their capability to effectively model long range dependencies. Between the two, the Bayes-BiLSTM attains the minimum error, since it is capable to model non-causal behavior and has been optimized using the Bayesian framework. Fig.3 shows the CDE's validation performance and the model's hyperparameters respectively, for four successive iterations. The final iteration performs best, as expected.

Table 1. Performance evaluation of the proposed method against other techniques for different objective metrics and appliances.

	MAE	RMSE	NRMS	MAE	RMSE	NRMS
Methods	*Appliance 1: CDE*			*Appliance 2: DWE*		
Bayes-BiLSTM	9.19	3.03	0.15	6.43	2.53	0.27
LSTM	25.35	5.03	0.43	24.04	4.90	1.00
CNN	34.42	5.87	0.48	32.67	5.72	1.24
CO [7], [16]	117.53	10.84	1.26	156.23	12.50	4.41
FHMM [15],[16]	129.57	11.38	0.90	313.68	17.71	4.44
Methods	*Appliance 3: HPE*			*Appliance 4: WOE*		
Bayes-BiLSTM	106.56	10.32	0.59	8.06	2.84	0.75
LSTM	161.74	12.72	0.55	15.82	3.97	0.89
CNN	158.68	12.60	0.46	23.30	4.83	0.88
CO [7], [16]	249.16	15.78	1.23	267.00	16.34	3.46
FHMM [15],[16]	121.69	11.03	1.14	49.38	7.03	2.89

In the following, we perform comparisons using the Estimated Energy Fraction Index (EEFI) and Actual Energy Fraction Index (AEFI) indicators defined as

$$EEFI(j) = \sqrt{\frac{\sum_n \hat{p}^j(n)}{\sum_n \sum_j \hat{p}^j(n)}} \text{ and } AEFI(j) = \sqrt{\frac{\sum_n p^j(n)}{\sum_n \sum_j \hat{p}^j(n)}} \quad (12)$$

where we recall that $\hat{p}^j(n)$ and $p^j(n)$ is the estimated and ground truth power load for the j-th appliance respectively. Fig.4 depicts the difference $DEFI(j) = |EEFI(j) - AEFI(j)|$ for the selected appliances over all compared methods, verifying that Bayes-BiLSTM yields the minimum value.

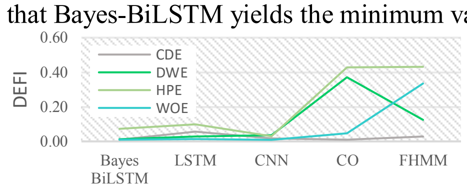


Fig. 4. Comparisons using DEFI indicator.

5. CONCLUSION

In this paper, we propose a Bayesian-optimized Bidirectional LSTM regression model for NILM. The Bayes-BiLSTM model introduces: (i) a modular approach in NILM, which addresses dimensionality issues arising in cases of large number of appliances; (ii) a non-causal modeling framework taking into account the inherent structure, which characterizes the operation of multi-state appliances; (iii) a Bayesian optimization process ensuring the creation of a best fitting configuration for each appliance. Experimental results indicate the superiority of the proposed method compared to the current state-of-the-art.

6. ACKNOWLEDGEMENTS

This research has received funding from the EU's H2020 research and innovation programme under grant agreement No768774.

7. REFERENCES

[1] A. Rahimpour, H. Qi, D. Fugate and T. Kuruganti, “Non-Intrusive Energy Disaggregation Using Non-Negative Matrix Factorization With Sum-to-k Constraint,” in *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4430-4441, Nov. 2017.

[2] A. Faustine, N. H. Mvungi, S. Kaijage, and K. Michael, “A Survey on Non-Intrusive Load Monitoring Methodies and Techniques for Energy Disaggregation Problem,” ArXiv170300785 Cs, Mar. 2017.

[3] N. D. Doulamis, A. D. Doulamis and E. Varvarigos, “Virtual Associations of Prosumers for Smart Energy Networks Under a Renewable Split Market,” in *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6069-6083, Nov. 2018.

[4] G. Sovero, N. C. Kemper and E. Kussul, “Load Signature Generator for Appliance State Sensing and Energy Disaggregation,” in *IEEE Sensors Journal*, vol. 17, no. 14, pp. 4587-4594, 15 July15, 2017.

[5] G. W. Hart, “Nonintrusive appliance load monitoring,” Proc. IEEE, vol. 80, no. 12, pp. 1870–1891, Dec. 1992.

[6] G. Elafoudi, L. Stankovic, and V. Stankovic, “Power disaggregation of domestic smart meter readings using dynamic time warping,” in *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2014, pp. 36–39.

[7] Y. Lin and M. Tsai, “Non-Intrusive Load Monitoring by Novel Neuro-Fuzzy Classification Considering Uncertainties,” *IEEE Trans. Smart Grid*, vol. 5, no. 5, pp. 2376–2384, Sep. 2014.

[8] H. Altrabalsi, V. Stankovic, J. Liao, and L. Stankovic, “Low-complexity energy disaggregation using appliance load modelling,” *AIMS Energy*, vol. 4, pp. 884–905, Jan. 2016.

[9] J. Liao, G. Elafoudi, L. Stankovic, and V. Stankovic, “Non-intrusive appliance load monitoring using low-resolution smart meter data,” in 2014 *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2014, pp. 535–540.

[10] K. He, L. Stankovic, J. Liao, and V. Stankovic, “Non-Intrusive Load Disaggregation Using Graph Signal Processing,” *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1739–1747, May 2018.

[11] N. Doulamis, and A. Doulamis. “Fast and adaptive deep fusion learning for detecting visual objects.” *European Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2012.

[12] J. Z. Kolter and T. S. Jaakkola, “Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation,” in *AISTATS*, 2012.

[13] R. Bonfigli, E. Principi, M. Fagiani, M. Severini, S. Squartini, and F. Piazza, “Non-intrusive load monitoring by using active and reactive power in additive Factorial Hidden Markov Models,” *Appl. Energy*, vol. 208, pp. 1590–1607, Dec. 2017.

[14] W. Kong, Z. Y. Dong, D. J. Hill, J. Ma, J. H. Zhao, and F. J. Luo, “A Hierarchical Hidden Markov Model Framework for Home Appliance Modeling,” *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3079–3090, Jul. 2018.

[15] L. Mauch and B. Yang, “A novel DNN-HMM-based approach for extracting single loads from aggregate power signals,” 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 2384-2388.

[16] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep Learning for Computer Vision: A Brief Review,” *Comput. Intell. Neurosci.*, vol. 2018, Article ID 7068349, pp. 1–13, Feb. 2018.

[17] J. Kelly and W. Knottenbelt, “Neural NILM: Deep Neural Networks Applied to Energy Disaggregation,” *Proc. 2nd ACM Int. Conf. Embed. Syst. Energy-Effic. Built Environ. - BuildSys* 15, pp. 55–64, 2015.

[18] W. He and Y. Chai, “An Empirical Study on Energy Disaggregation via Deep Learning,” *2nd Intern. Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016)*, 2016.

[19] S. Hochreiter, J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, 9 (8), pp. 1735-1780, 1997.

[20] L. Mauch and B. Yang, “A new approach for supervised power disaggregation by using a deep recurrent LSTM network,” in 2015 *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015, pp. 63–67.

[21] J. Kim, T.-T.-H. Le, and H. Kim, “Nonintrusive Load Monitoring Based on Advanced Deep Learning and Novel Signature,” *Computational Intelligence and Neuroscience*, 2017.

[22] M. Schuster, and K.K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, 45(11), 2673-2681, Nov. 1997.

[23] A. Graves, N. Jaitly, and A. Mohamed, “Hybrid speech recognition with Deep Bidirectional LSTM,” in 2013 *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 273–278.

[24] A.D. Doulamis, N.D. Doulamis, S.D., Kollias, “An adaptable neural-network model for recursive nonlinear traffic prediction and modeling of MPEG video sources,” *IEEE Transactions on Neural Networks*, 14 (1), pp. 150-166, 2003.

[25] R. Bardenet, and K. Balázs “Surrogating the surrogate: accelerating Gaussian-process-based global optimization with a mixture cross-entropy algorithm.” *27th International Conference on Machine Learning (ICML 2010) Omnipress*, 2010.

[26] S. Makonin, F. Popowich, L. Bartram, B. Gill and I. V. Bajić, “AMPds: A public dataset for load disaggregation and eco-feedback research.” *IEEE Electrical Power & Energy Conference (EPEC)*, 2013.

[27] N. Batra et al., “NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring,” *Proc. 5th Int. Conf. Future Energy Syst. - E-Energy* 14, pp. 265–276, 2014.