Appliance Classification using BiLSTM Neural Networks and Feature Extraction

Martha T. Correa-Delval, Hongjian Sun, Peter C. Matthews
Department of Engineering
Durham University, Durham, DH1 3LE, United Kingdom
Email: martha.t.correa-delval@durham.ac.uk

Jing Jiang
Department of Mathematics, Physics
and Electrical Engineering
Northumbria University

***Abstract*—One significant challenge in Non-Intrusive Load Monitoring (NILM) is to identify and classify active appliances used in a building. This research focuses on the classifying process, exploring different approaches for the feature extraction of the appliances' power load to improve the classification accuracy. In this paper, we present a new method - Spectral Entropy and Instantaneous Frequency-based Bidirectional Long Short Term Memory (SE-IF BiLSTM). It uses feature extraction from the power load to obtain information, such as instant frequency, spectral entropy, spectrogram, Mel spectrogram and signal variation, to feed BiLSTM Neural Network. We also test different options for the BiLSTM to decide the most optimal settings. This method improves the classification performance, achieving up to 98.57% classification accuracy.**

***Index Terms*—BILSTM, Appliance Classification, NILM**

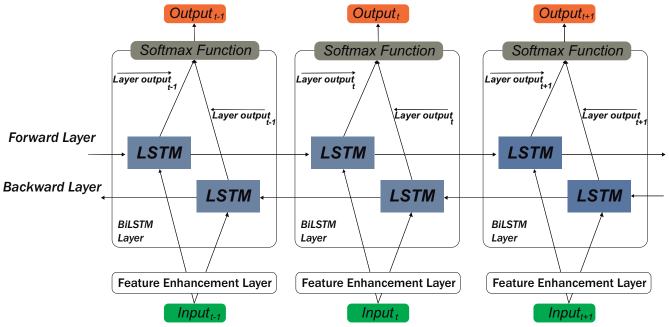

Fig. 1. Network Structure of the SE-IF BiLSTM where *t* stands for Timestep

I. Introduction

Carbon emission reduction and energy conservation are essential topics that usually go together nowadays. In recent years, the number of appliances used in homes and buildings has increased, and it is expected to keep increasing along with the power consumption, bringing with it the need of a proper way to optimise energy usage. According to [1], active energy data feedback to users can achieve up to 20% in energy savings. But to achieve these savings in both carbon emissions reduction and energy conservation, a way to analyse the energy consumption of the users is needed.

Non-Intrusive Load Monitoring (NILM) is a disaggregation process that analyses the power consumption of the appliances. It monitors which appliance(s) are being used over a period of time by analysing aggregated power without using any external hardware. This method allows users to see how much energy has been consumed by said appliances and creates a series of opportunities, such as reducing energy consumption, survey appliance usage behaviour or identifying faulty appliances [2]. There are various approaches for enabling appliance classification, such as Neural Network, Deep Learning, and Long Short-Term Memory [3], [4].

Artificial Neural Networks (ANN) such as Recurrent Neural Networks (RNN) is one of the main methods in the field of disaggregation because they can learn the pattern and behaviour of appliances in order to identify them in the aggregated power data [3]–[5]. Long-term dependency due to vanishing gradient is one of the problems of RNNs, where gradient information disappears or explodes as it is propagated back through time and can limit the RNN's memory. This happens because the greater the number of steps the network takes, the more previous information it will not retain and take into consideration. When analysing long sequences of data such as power consumption in a period of time, this issue is bound to rise [3].

Long short-term memory (LSTM) method aims to tackle this issue, helping the error gradient to flow back in time utilising a gated input, output and feedback loop within a memory cell. This way, the error is retained and carried through time steps for a much longer time, compared to typical RNNs. A wide variety of sequence tasks, including automatic speech recognition and machine translation, have used LSTMs with success [6].

In [3], a LSTM neural network was used to classify similar energy consumption appliances by emphasising the signal variation. A LSTM neural network in [4] used the spectrograms for appliance classification. In [7], LSTMs were tested against other methods such as decision trees and deep neural networks, by classifying general appliances in a household.

Bidirectional layers can improve the RNN's performance. Bidirectional RNNs have two parallel RNNs, one that reads the input sequence forwards and another one that reads it backwards, as seen in Figure 1. After that, the output from both forwards and backwards portions are combined by concatena-

This work is supported by the CONACyT-SENER (Mexican National Council for Science and Technology, along with the Secretary of Energy) energy sustainability programme.

tion or a sum [2]. It was reported that the deep learning-based models outperform conventional Auto Regressive Integrated Moving Average (ARIMA)-based models in forecasting time series and in particular for the long-term prediction problems [8] and it's also proven that BiLSTMs are more effective than LSTMs [9].

Despite BiLSTMs being proven more effective than LSTMs, they have not been widely used or tested for energy disaggregation purposes. Moreover, some other features of energy consumption data, such as spectral entropy, have been rarely examined and explored in the literature. It was believed these features also carry important information, e.g., [10], [11] used them for fault detection and diagnosis, and [12] used them for speech recognition.

Compared with existing research, the main contributions of this paper are as follows:

- A new method SE-IF BiLSTM is proposed to perform appliance classification. This method consists of a Bidirectional Long Short-Term Memory (BiLSTM) Neural Network and extraction of features including spectrogram frequency bands, Mel spectrogram, instantaneous frequency, spectral entropy and signal variation as inputs.
- Spectral entropy and instantaneous frequency are used for feature extraction of energy consumption data, and their effects for appliance classification are analysed.
- The new method SE-IF BiLSTM is proved to be very effective for classifying non-similar appliances and mixed appliances and can achieve up to 98.57% of accuracy in appliance classification in a common household environment.

II. Methodology

In this section, we propose the following features to be extracted and used as input to feed a BiLSTM. These features achieve better classification results.

A. BILSTM

The bidirectional LSTMs (BiLSTM) are an extension of the described LSTM models in which two LSTMs are applied to the input data. In the first round, an LSTM is applied on the input sequence (i.e., forward layer). In the second round, the reverse form of the input sequence is fed into the LSTM model (i.e., backward layer). Applying the LSTM twice leads to improved learning long-term dependencies and thus improved the accuracy of the model [9].

B. Mel Spectrogram

A spectrogram is a two-dimensional representation of the magnitude of a signal at various frequencies over time that shows the signal power at each frequency at a particular time as well as how it varies over time. This makes spectrogram an extremely useful tool for the frequency analysis of time-series data [4].

The Mel spectrograms use the Mel-frequency scale, a linear frequency interval of 1000 Hz or less and a log interval of 1000Hz or higher [13]. On Mel-frequency scale, spectrum is transformed into Mel-spectrum through Mel-filter banks of triangular overlapping windows. The Mel-filter bank is a critical band with various bandwidth on normal frequency scale and emphasizes information in the low frequency range by placing a large number of filters in low frequency bands than these of high frequency bands.

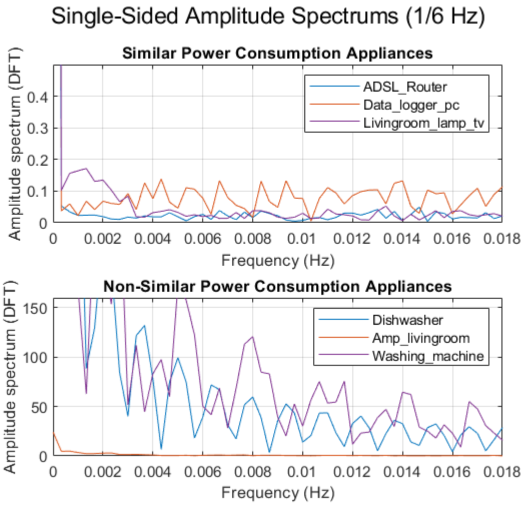

Fig. 2. Single-sided amplitude spectrums comparison of similar power consumption appliances and non-similar power consumption appliances from the UK DALE dataset [2].

It uses the Fourier transform to decompose the signal into its individual frequencies and the frequency's amplitude. Figure 2 shows a graph comparing the amplitude spectrums of a dishwasher and a battery charger. This feature contains useful information about the appliance and its behaviour to help classify it, such as the dominant frequencies.

C. Instantaneous Frequency

The instantaneous frequency is a useful characteristic for describing non-stationary signals. It is defined as:

$$f_{inst}(t) = \frac{1}{2\pi}\frac{d\phi}{dt} \tag{1}$$

where ϕ is the phase of the analytic signal of the input. It estimates the time-dependent frequency of a signal as the first moment of the power spectrogram [14].

D. Spectral Entropy

The spectral entropy (SE) of a signal is a measure of its spectral power distribution and it is based on the Shannon entropy [11]. The SE uses the signal's normalised power distribution in the frequency domain as a probability distribution, and then uses it to calculate its Shannon Entropy. For a signal *x(n)*, the power spectrum is *S(m)* = $\left|X(m)\right|^{2}$, where *X(m)* is the discrete Fourier transform of *x(n)*. To compute the instantaneous Spectral Entropy given a time-frequency power spectrogram *S(t,f)*, the probability distribution at time *t* is:

$$P(t,m) = \frac{S(t,m)}{\sum_{f} S(t,f)} \tag{2}$$

TABLE I
SE-IF BiLSTM OPTIONS

Control parameter	Method
Input	Spectrogram bands, mel spectrogram, instantaneous frequency, spectral entropy, Δp
Samples	500
Num. of epochs	1000
Mini batch size	32
Cost function	Cross Entropy
Optimization	Adam
Pre-processing	Z-Score
Weight initialization	He initialisation
Regularization	Dropout
Activation function	Softmax

Then the Spectral Entropy at time t can be given by:

$$H(t) = -\sum_{m=1}^{N} P(t,m)\log_2 P(t,m) \tag{3}$$

where N is the total frequency points.

E. Signal variation

This feature separates the original signal, which is low sample rate data (power consumption), using a reflection rate and subtracts one variant power signal (with a reflection rate of 0.1) from the other variant power signal (with a reflection rate of 0.01). This difference is denoted as Δp , which represents variation of the original signal. Its purpose is to emphasise the variation of multi-state and similar consumption appliances [3].

F. Neural Network Structure

Input Signal. The input is a selection of appliances from House 1 and House 2 from the UK DALE dataset [2]. The different testing groups consisted of Similar Appliances, Non Similar Appliances and Mixed Appliances.

Pre-processing Method. The closest the input data is to Gaussian distribution, the better the performance the model will have [3]. Z-Score is one of the most common normalisation methods. It uses the mean and standard deviation to normalise the input data, so it is a measure of how many standard deviations below or above the population mean a raw score is [15].

Weight Initialisation. It sets up the weights vector for all neurons of the Neural Network for the first time, just before the Neural Network training process starts. If the weights are not properly initialised, the forward pass can lead to the vanishing gradient. A common method used for weight initialisation is the He initialisation. In this method, the weights are initialised according to the size of the previous layer, helping to attain a global minimum of the cost function faster and more efficiently. While still being random, they have different ranges depending on the size of the previous layer, providing a controlled initialisation.

Activation Function. The activation function of a node defines the output of that node given an input or set of inputs [16]. The output unit activation function is the softmax equation, which is established as follows:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{4}$$

where x is the net input vector. The formula computes the exponential of the input parameter and the sum of exponential parameters of all existing values in the inputs. The output for the Softmax function is the ratio of the exponential of the parameter and the sum of exponential parameter.

Optimiser. Optimisers are used to change the attributes of the neural network such as weights and learning rate in order to reduce the losses. Adam optimiser is the one used in this Neural Network. It's been widely used due to its simple implementation, efficiency, little memory requirement, consistency when diagonally re-scaling the gradients and works well for problems with large data [17].

More details of the SE-IF BiLSTM can be found in Table I.

III. Performance Analysis Metrics

In this section, we explain the performance metrics used for evaluation: accuracy, recall, precision and F1 score. We also present K-Fold as the validation method used.

A. Performance Metrics

Used metrics for evaluation:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+FP+FN+TN},\\
\text{Recall} &= \frac{TP}{TP+FN},\\
\text{Precision} &= \frac{TP}{TP+FP},\\
\text{F1 Score} &= 2 * \frac{\text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}
\end{aligned} \tag{5}$$

where *TP* is True Positives for correctly predicted events, *FP* is False Positives for incorrectly predicted events, *TN* is True Negatives for correctly predicted non-events and *FN* is False Negatives for incorrectly predicted non-events. Recall is a ratio of the number of correct classifications to the total number of actual positive instances. Precision is a ratio of the number of correct classifications to the total number of predicted positive instances. Accuracy is a ratio of correct classification to the total test data. F1-Score is the harmonic average of Recall and Precision.

In these experiments, the following criteria are used: True Positives are when the network predicted a specific appliance (a monitor, for example) and it was that appliance. True Negatives are when the network didn't predict a monitor and it wasn't a monitor. False Positives are when the network predicted a monitor but it was another appliance. Finally, False Negatives are when the network predicted another appliance but it was a monitor.

TABLE II
TRAINING RESULTS COMPARISON

Category		Accuracy	Recall	Precision	F1 Score
Similar Apps	**SE-IF BiLSTM**	96.25%	96.78%	96.25%	96.29%
	Delta Power [3]	97.14%	97.14%	97.40%	97.14%
	Spectrogram Only [4]	84.71%	84.71%	84.17%	84.35%
Non-Similar Apps	**SE-IF BiLSTM**	98.57%	98.57%	98.70%	98.57%
	Delta Power [3]	57.14%	61.17%	57.14%	56.09%
	Spectrogram Only [4]	97.14%	97.40%	97.14%	97.05%
Mixed Apps	**SE-IF BiLSTM**	90%	90.46%	90%	89.97%
	Delta Power [3]	48.82%	58.55%	48.82%	48.51%
	Spectrogram Only [4]	85.88%	87.59%	85.88%	83.80%

B. Validation Method

K-Fold Cross-validation is a re-sampling procedure used to evaluate machine learning trained models on a limited data sample. This method uses one variable K that refers to the number of groups that a given data sample is to be split into.

The general procedure is as follows: Shuffle the dataset randomly and then split the dataset into K groups. For each unique group: Take the group as a hold out or testing data set, take the remaining groups as a training data set, fit a model on the training set and evaluate it on the test set and retain the evaluation score and discard the model. Finally, get the mean of K number of evaluation scores.

Each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. Each sample being used in the testing set one time and on the training set $K-1$ times [18].

IV. SIMULATIONS

A. Dataset Description

UK DALE is one of the very first UK based dataset published for energy disaggregation research. It contains both mains (aggregated power reading) and individual appliance power reading data for 5 houses. UK DALE dataset has several releases dated in which data are collected from 2012 to 2017. The power reading is collected at every 6 seconds (1/6 Hz), and some houses are also provided with a 16kHz voltage and current reading [2].

B. Neural Network Performance

Regarding this single-label multi-class classification task, a comparison among the SE-IF BiLSTM method and the methods proposed in [3] and [4] is shown in Table II.

Data preparation. Since the majority of the appliances are used for a very short time in the raw data, periods of approximately 45 minutes were made by placing several shorter periods of time in succession.

Benchmark methods. It has been difficult to correctly identify appliances in the past when their energy consumption is similar. The methods proposed in [3] and [4] were tested separately.

Experimental setup. Various appliances from the UK DALE dataset were used in these experiments. Appliances with similar consumption and non-similar energy consumption were chosen, because, by the nature of classification, the more similar the appliances are, the more difficult they are to distinguish. Among them, there are appliances with different behaviours and specific signatures, such as dishwashers and washing machines. There are, likewise, devices with similar spectrums and consumption, such as the various lamps inside a house, routers, modems and alarms.

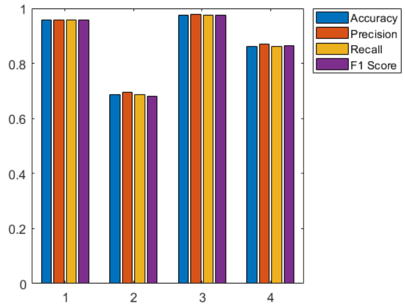

Fig. 3. Comparison of the accuracy, precision, recall and F1 Score results among LSTM and BiLSTM networks, using the features from the SE-IF BiLSTM method. 1: Non-similar appliances using BiLSTM, 2: Non-similar appliances using LSTM, 3: Similar appliances using BiLSTM, 4: Similar appliances using LSTM.

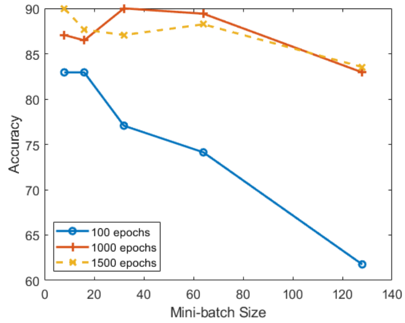

Fig. 4. Comparison of accuracy when testing different combinations of epochs and mini batch size using the SE-IF BiLSTM method

Experiment 1. The first experiment compared the use of the BiLSTM neural network against the single LSTM neural network method to verify its effectiveness. It was tested using similar energy consumption appliances and non-similar energy consumption appliances separately. Figure 3 shows the results, proving that the BiLSTM neural network is better suited for this problem than the single-layer LSTM neural network.

Experiment 2. This experiment compared different network options for epoch number and mini-batch size (mbs) to define the optimal ones for the SE-IF method, using the accuracy metric. Figure 4 shows the most relevant tests. The first test uses 100 epochs, the second test uses 1000 epochs and the third test uses 1500 epochs, all of them with mbs of 8, 16, 32, 64 and 128. It is observed that the largest the mbs is, the accuracy decreases, and that the largest the number of epochs, the accuracy increases. After these tests, it was determined that the best mini-batch size is 32 when the epochs are set to 1000. While the mini-batch size of 32 is not the best for every epoch configuration, it is the most accurate combination of epoch and mbs among all the tests. Therefore the SE-IF BiLSTM method uses 1000 epochs and a mini-batch size of 32 as training options.

Experiment 3. The third experiment used a selection of different appliances that have similar energy consumption. The method from [3] focused on identifying similar appliances so the most accurate results were achieved during this experiment by using this method, closely followed by the SE-IF BiLSTM method proposed in this paper achieving acceptable results. Results are shown in Table II.

Experiment 4. The fourth experiment used a selection of different appliances that don't have similar energy consumption. The SE-IF BiLSTM method achieved the best results out of the three methods tested, with the method from [3] dropping considerably in the performance metrics and method from [4] achieving similar results as the SE-IF BiLSTM method.

Experiment 5. The fifth experiment used all appliances used in the previous experiments at the same time, similar and non-similar. The SE-IF BiLSTM method achieved the best results out of the three methods tested, with method from [3] performing very low in the performance metrics and method from [4] achieving similar results as the SE-IF BiLSTM method.

Results analysis. Figure 3 shows that using BiLSTM achieves better results when compared to a single-layer LSTM. Figure 4 shows the best options for the BiLSTM network, being 1000 epochs and a mini-batch size of 32. Table II shows the comparison between the SE-IF method and the methods used in [3] and [4], where accuracy, recall, precision and F1-score were the main metrics used for comparison. Although the most accurate results were not achieved by the SE-IF method in the category of appliances with similar energy consumption, they were achieved in the other two categories (non-similar appliances and mixed appliances). The chosen features to evaluate in this method adjust to different situations, allowing it to be both a balanced and consistent method.

V. Conclusions

This paper presents an appliance classification method, i.e., SE-IF BiLSTM, for exploring appliances' features such as spectrograms, instantaneous frequency, spectral entropy and signal variation. This method has potential to be used for building or home energy consumption analysis, helping reduce the carbon emissions.

The SE-IF BiLSTM method proposed in this paper was tested against other existing methods shown in the literature, performing the best when analysing the non-similar appliances and mixed appliances categories, achieving 98.57% and 90% accuracy respectively. It was also very accurate when testing similar appliances, achieving 96.25% accuracy.

In future work, we will implement the use of aggregated data for the SE-IF BiLSTM to use on houses and/or buildings for appliance classification purposes.

References

[1] D. Vine, L. Buys, and P. Morris, "The effectiveness of energy feedback for conservation and peak demand: A literature review," *Open Journal of Energy Efficiency*, vol. 02, no. 01, pp. 7–15, 2013.

[2] 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.

[3] J. Kim, T.-T.-H. Le, and H. Kim, "Nonintrusive load monitoring based on advanced deep learning and novel signature," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1–22, 10 2017.

[4] J.-G. Kim and B. Lee, "Appliance classification by power signal analysis based on multi-feature combination multi-layer LSTM," *Energies*, vol. 12, no. 14, p. 2804, 2019.

[5] A. Tongta and K. Chooruang, "Long short-term memory (LSTM) neural networks applied to energy disaggregation," in *2020 8th International Electrical Engineering Congress (iEECON)*, pp. 1–4, 2020.

[6] Q. V. L. Ilya Sutskever, Oriol Vinyals, "Sequence to sequence learning with neural networks," *arXiv pre-print server*, 2014.

[7] R. Brito, M.-C. Wong, H. C. Zhang, M. G. Da Costa Junior, C.-S. Lam, and C.-K. Wong, "Instantaneous active and reactive load signature applied in non-intrusive load monitoring systems," *IET Smart Grid*, vol. 4, no. 1, pp. 121–133, 2021.

[8] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1394–1401, 2018.

[9] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3285–3292, 2019.

[10] V. Sharma and A. Parey, "A review of gear fault diagnosis using various condition indicators," *Procedia Engineering*, vol. 144, pp. 253–263, 2016.

[11] Y. N. Pan, J. Chen, and X. L. Li, "Spectral entropy: A complementary index for rolling element bearing performance degradation assessment," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 223, no. 5, pp. 1223–1231, 2009.

[12] Shen J., J. Hung, and L. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *ICSLP*, vol. 98, 1998.

[13] V. Tiwari, "MFCC and its applications in speaker recognition," *Int. J. Emerg. Technol.*, vol. 1, 01 2010.

[14] S. Pei and S. Huang, "Instantaneous frequency estimation by group delay attractors and instantaneous frequency attractors," in *22nd European Signal Processing Conference (EUSIPCO)*, pp. 471–475, 2014.

[15] S. A. McLeod, "Z-score: definition, calculation and interpretation," 2019. [Online] Available: https://www.simplypsychology.org/z-score.html (Date last accessed on Oct. 28th, 2020).

[16] C. M. Bishop, *Pattern recognition and machine learning*. Information science and statistics, New York, NY: Springer, 2006. Softcover published in 2016.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[18] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research 11*, 2010.